

Analysis for Time Course Microarray Experiments

吳漢銘

hmwu@stat.sinica.edu.tw
<http://idv.sinica.edu.tw/hmwu>



Outlines

2/57

- Time Series Microarray Experiments
- Overview of Analyzing Software

■ Some Issues

- P-values
- Multiple Hypothesis Testing
- Permutation Test
- Gene Set Enrichment Analysis

■ SAM: Significance Analysis of Microarrays

- Algorithm
- Interpretation

Differential Expressed Genes

■ STEM: Short Time-Series Expression Miner

- Algorithm
- Example

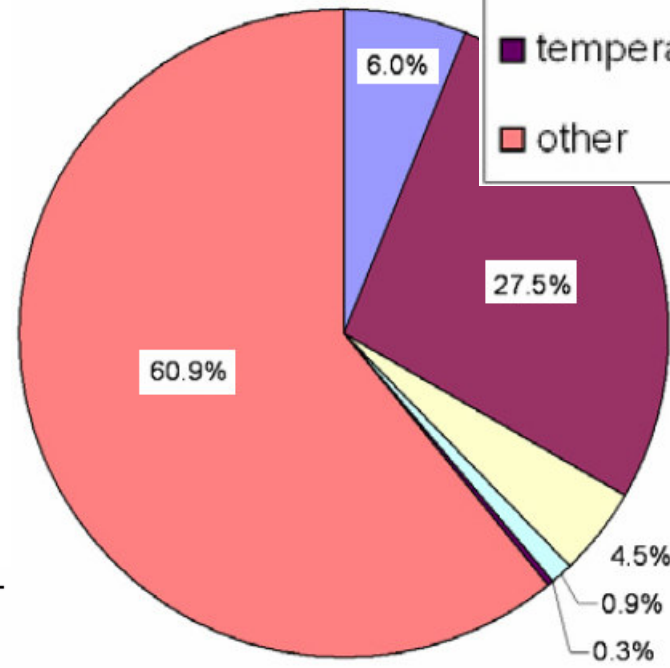
Clustering

Time Series Microarray Experiments

3/57

Study dynamic biological process

- Cell cycle (Spellman et al., 1998, *Mol Bio Cell*)
- Response to temperature changes (Gasch et al, 2000)
- Developmental studies (Arbeitman et al., 2002, *Science*)
- Immune response (Guillemin et al., 2002, *PNAS*)



Distribution of Microarray Data Sets in the Gene Expression Omnibus



Distribution of microarray experiments by type. Summary of the 786 microarray datasets for human, mouse, rat, and yeast in the Gene Expression Omnibus as of August 2005.

Source: Ernst and Bar-Joseph. 2006, *BMC Bioinformatics*.

Time Course Experiments

Home | About TAIR | Sitemap | Contact | Help | Order | Login

Search | Tools | Arabidopsis Info | News | Links | FTP | Stocks

Gene Search

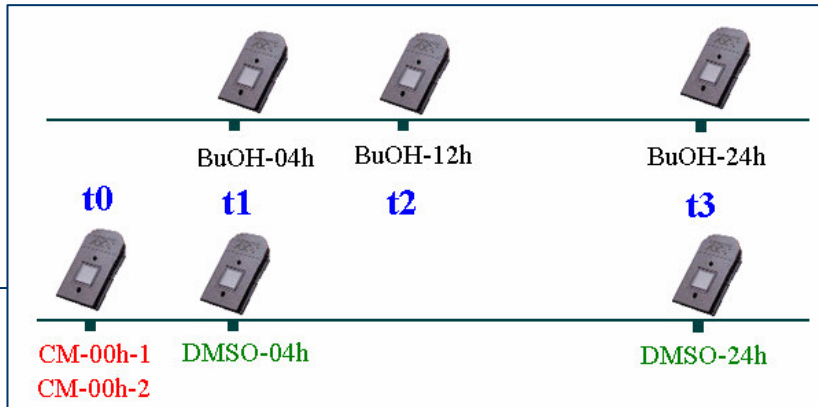
Experiment: AtGenExpress: Heat stress time course

Experiment Summary | Samples | Slides & Datasets | Array Design | View All

Submission Number ME00339
TAIR Accession ExpressionSet:1007967124
Author(s) Lutz Nover, Pascal von Koskull-Döring
Organization(s) AtGenExpress
Experimental Variables root, shoot, time, cultured cell, heat
Variable Type Plant Material
Experiment Category abiotic treatment, tissue comparison
Experiment Goals response to heat
Description This experiment studies the effect of heat in shoot and root tissue of 16 days old Arabidopsis Columbia-0 ecotype and in cell culture at two samples from each time point, which were taken in duplicate in a time course (0.25h, 0.5h, 1h, 3h of heat stress and at heat stress). Control samples for the cell non-treated cells (0h; 3h; 6h; 12h; 24h at expression data of the samples were generated using the ATH1 chip from Affymetrix. The study is funded by the DFG.

Data Counts Number of Slides: 94 Number of Replicates: 47

General comments or questions: curator@ncgr.org
 Seed or DNA stock questions (donations, availability): abrc@arabidopsis.org



Root (r)
Shoot (s)
Cell Culture (c)

	0	.25	.5	1	3	4	6	12	24
Treatment	c	c	c	c	c	c	c	c	c
	r	r	r	r	r	r	r	r	r
	s	s	s	s	s	s	s	s	s
Control	c				c		c	c	c
	r	r	r	r	r	r	r	r	r
	s	s	s	s	s	s	s	s	s

3 x 8 x 2 = 48
23 x 2 = 46



Short Time Series Microarray Experiments

5/57

- About 80 % of microarray time series experiments are short:
 - 3-8 time points.
 - Cost of microarray.
 - limited availability of biological material.

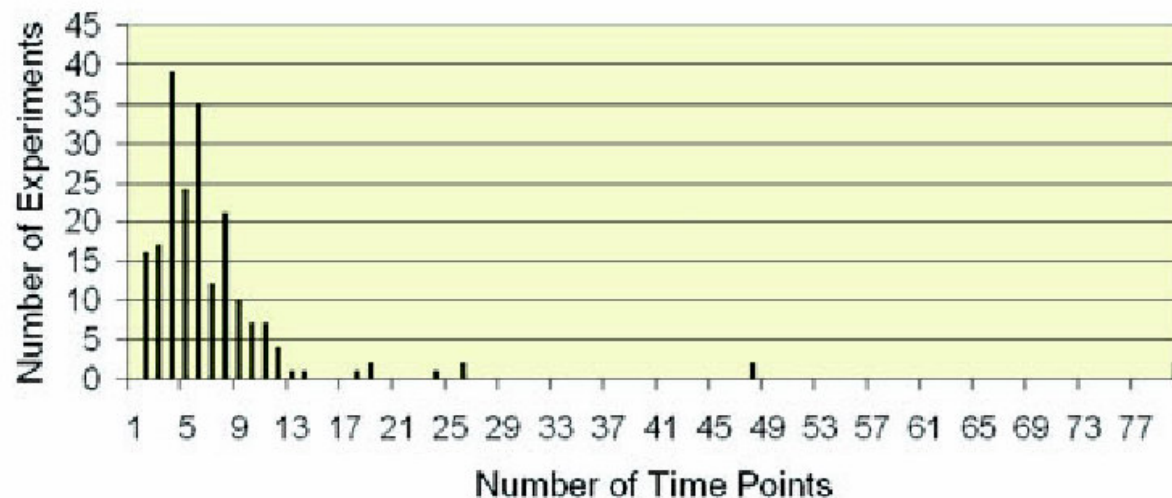


Fig. 1. Distribution of lengths of times series in the SMD as of June 2004.

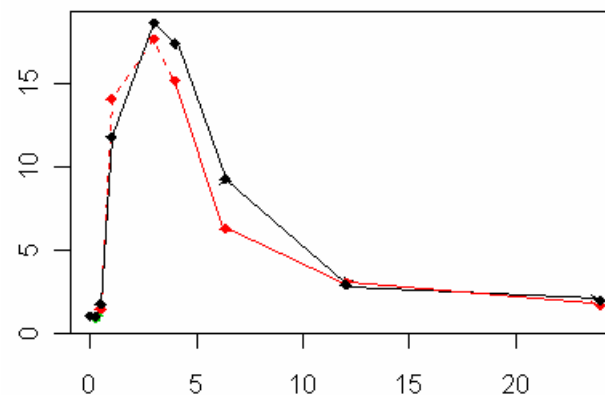
NOTE: SMD (June, 2004): ~170 published papers, ~30% are time series.

Source: Ernst et al., 2005, *Bioinformatics*.

Analyzing Software

Software for *Static* Gene Expression Data

- Do not take advantage of the **sequential information** in time series data.
- Popular clustering: hierarchical clustering, kmeans clustering, self-organizing maps.
 - ignore the temporal dependency among successive time points.
 - random permute the order of time points, the results would not change.

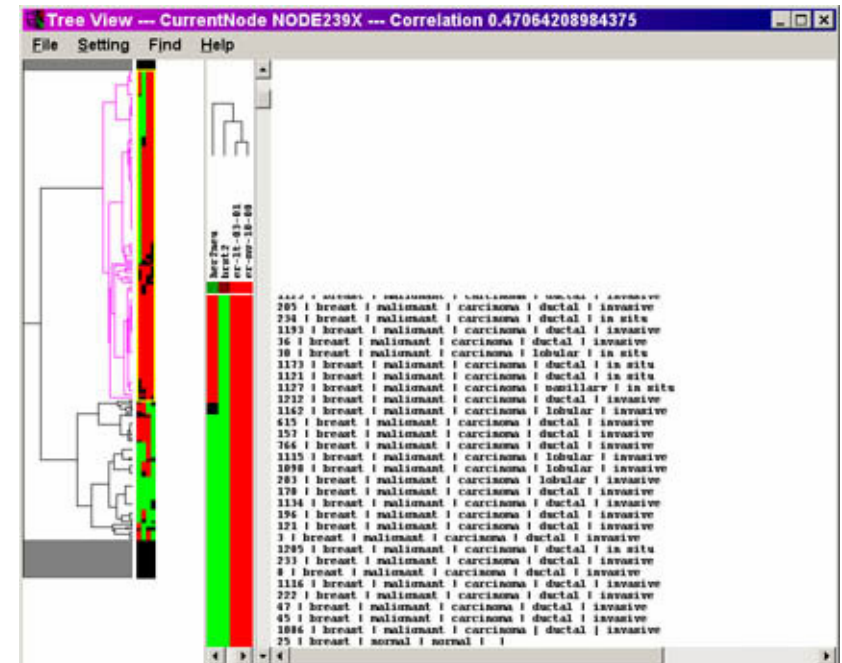
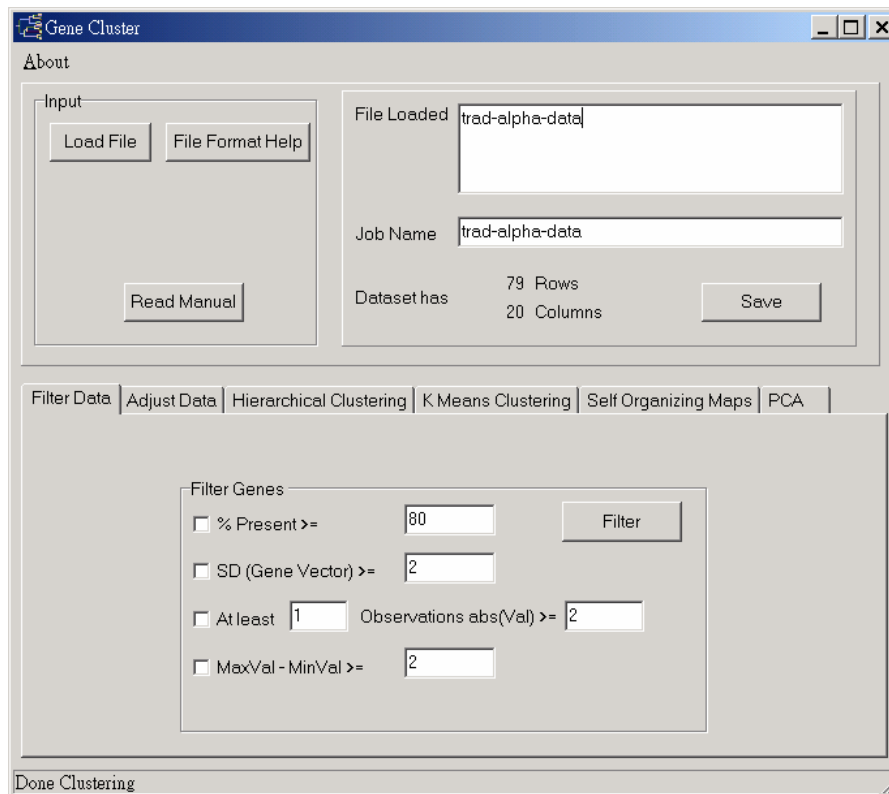


Software for *Time Series* Gene Expression Data

Software for *Static* Gene Expression Data (Clustering and Visualization)

7/57

Cluster & TreeView Eisen et al., 1998, *PNAS*



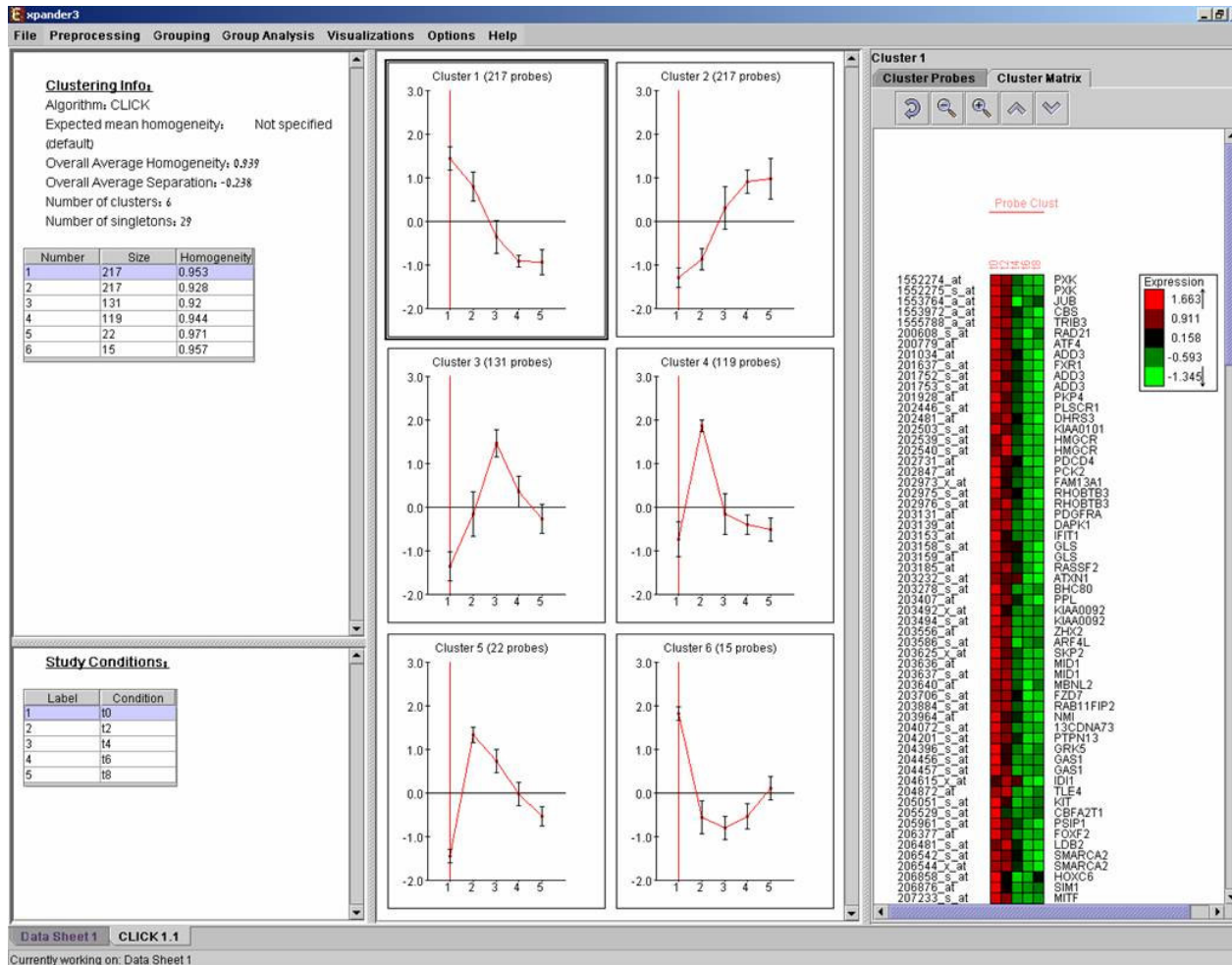
<http://rana.lbl.gov/EisenSoftware.htm>

Software for *Static* Gene Expression Data (Clustering and Visualization)

8/57

EXPANDER

(EXpression Analyzer and DisplayER) Shamir et al., 2005, *BMC Bioinformatics*

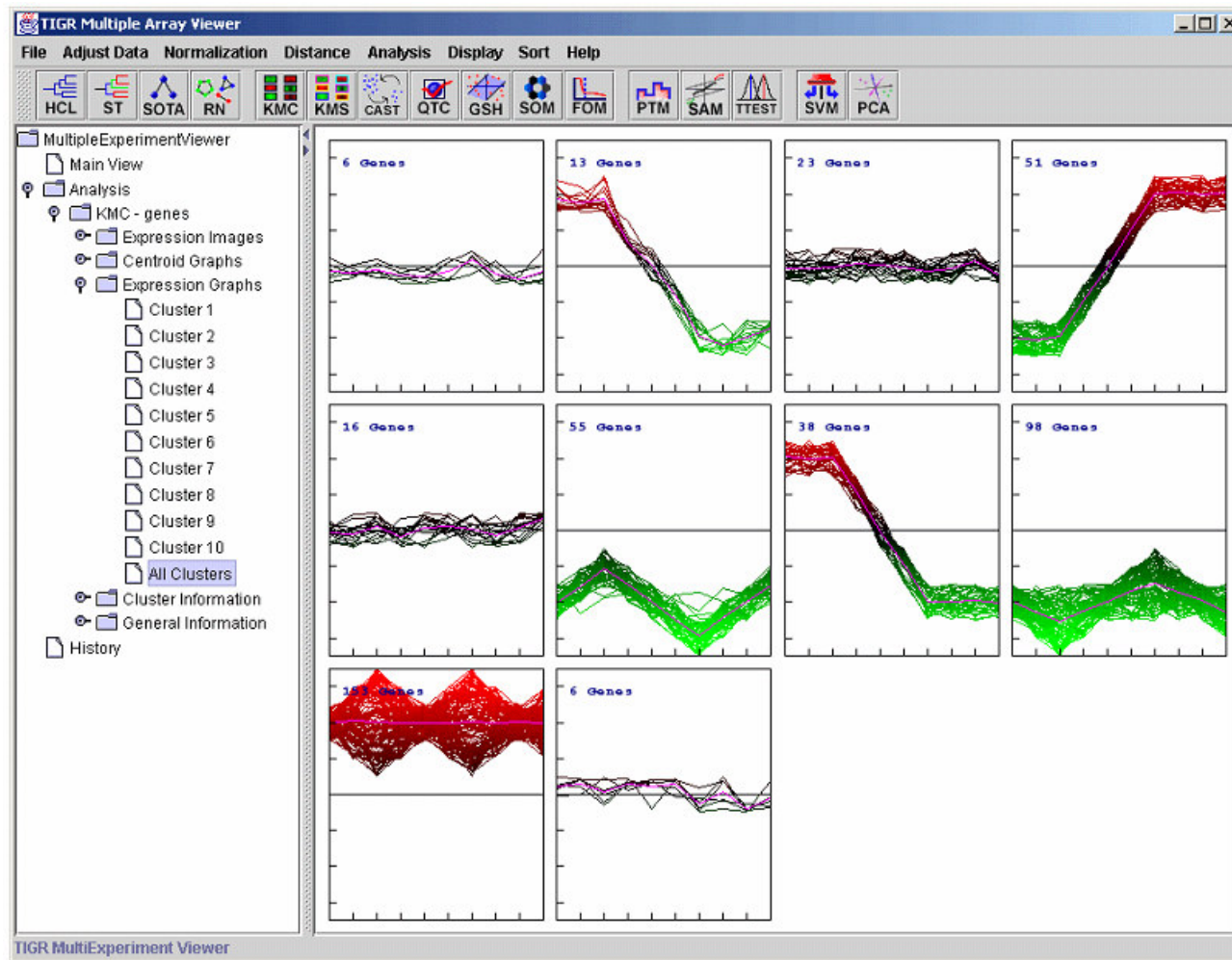


<http://www.cs.tau.ac.il/~rshamir/expander/expander.html>

Software for *Static* Gene Expression Data (General Purpose)

9/57

TM4: **MeV**
(MultiExperiment Viewer) Saeed et al., 2003, *Biotechniques*



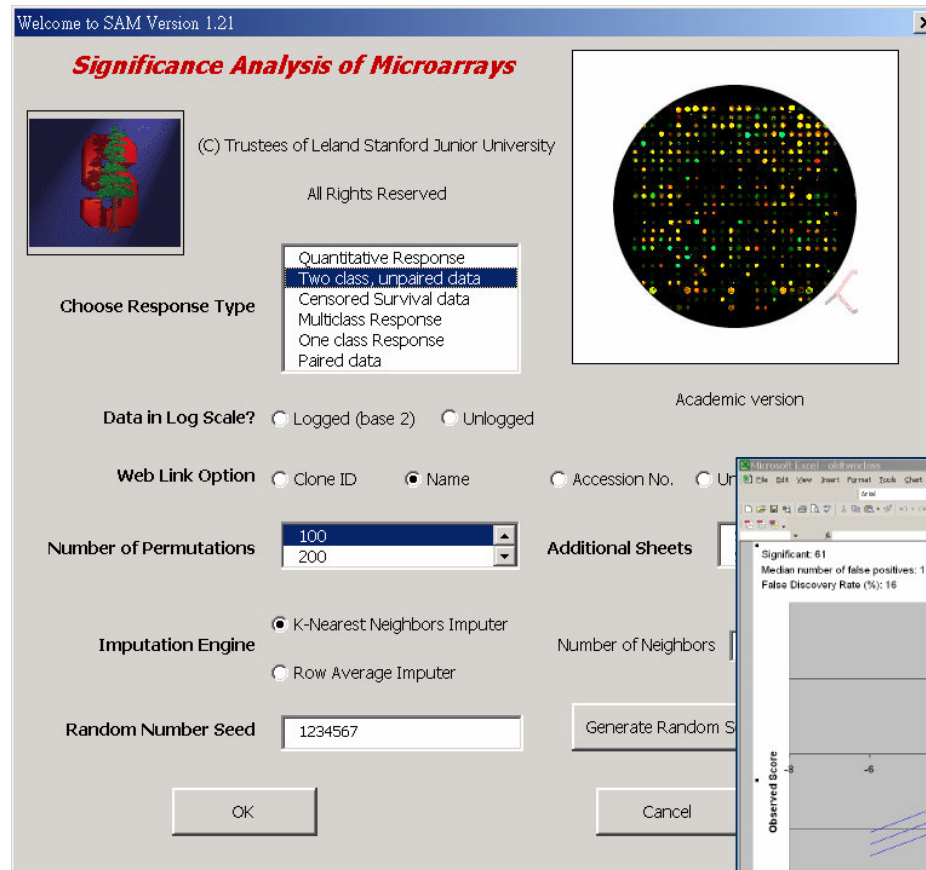
<http://www.tm4.org/mev.html>

Software for *Static / Time Series* Gene Expression Data (Differentially Expressed Genes)

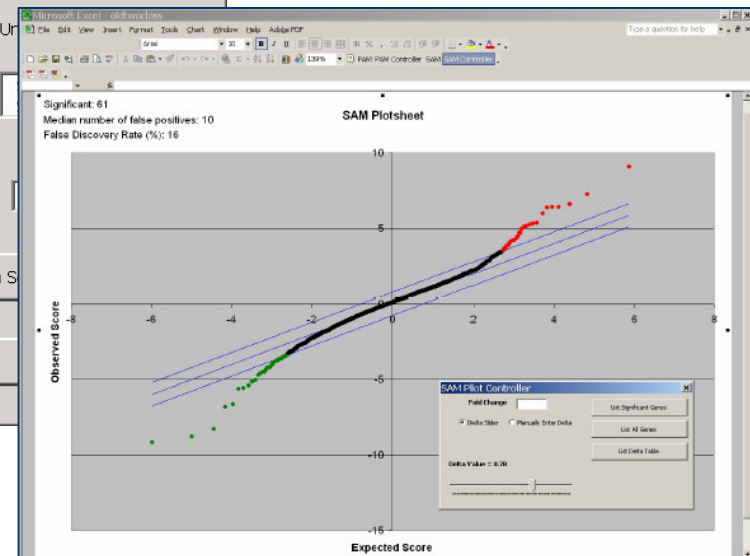
10/57

SAM:

Significance Analysis of Microarrays, Detect differentially expressed gene in time series data. (Tusher et al., 2001, *PNAS*)



SAM plot



<http://www-stat.stanford.edu/~tibs/SAM/>

Software for *Time Series* Gene Expression Data (Differential Expressed Genes)

11/57

EDGE:
Extraction of Differential Gene Expression
(Leek et al., 2006, *Bioinformatics*)

The left screenshot shows the main menu of the EDGE software. It features the 'edge' logo and version information (Version 0.1.47, Created by Storey Lab). Under 'Select Your Option', there is a list of actions: Load/Save Expression Data and Covariates, Impute Missing Data, View Covariates, Transform Data, Display Boxplots, Display Hierarchical Clustering, Display Eigengenes and Eigenarrays, and Identify Differentially Expressed Genes. A 'GO' button is at the bottom of this menu. A welcome message is displayed at the bottom of the window.

The right screenshot shows the '297 Genes Called Significant' table. The table has three columns: Gene Name, P-Value, and Q-value. Below the table are several buttons for further analysis: 'ACCESS PUBMED FOR SELECTED GENE', 'Q-PLOT', 'P-VALUE HISTOGRAM', 'CLUSTER SIGNIFICANT GENES', 'SAVE RESULTS', and 'DONE'. There are also input fields for 'UID' and 'Accession Number'. A 'RECALCULATE' button is located below the 'Optional Q-Value Arguments' section. At the bottom, a text box displays the estimated overall proportion of non-differentially expressed genes (p0 = 0.6062).

Gene Name	P-Value	Q-value
gene 2621	1.1e-05	0.004041
gene 1413	1.1e-05	0.004041
gene 543	1.1e-05	0.004041
gene 1315	1.1e-05	0.004041
gene 2217	1.1e-05	0.004041
gene 1087	4.2e-05	0.012628
gene 542	5.3e-05	0.012628
gene 2954	5.3e-05	0.012628
gene 571	6.3e-05	0.012858
gene 1730	7.4e-05	0.012858

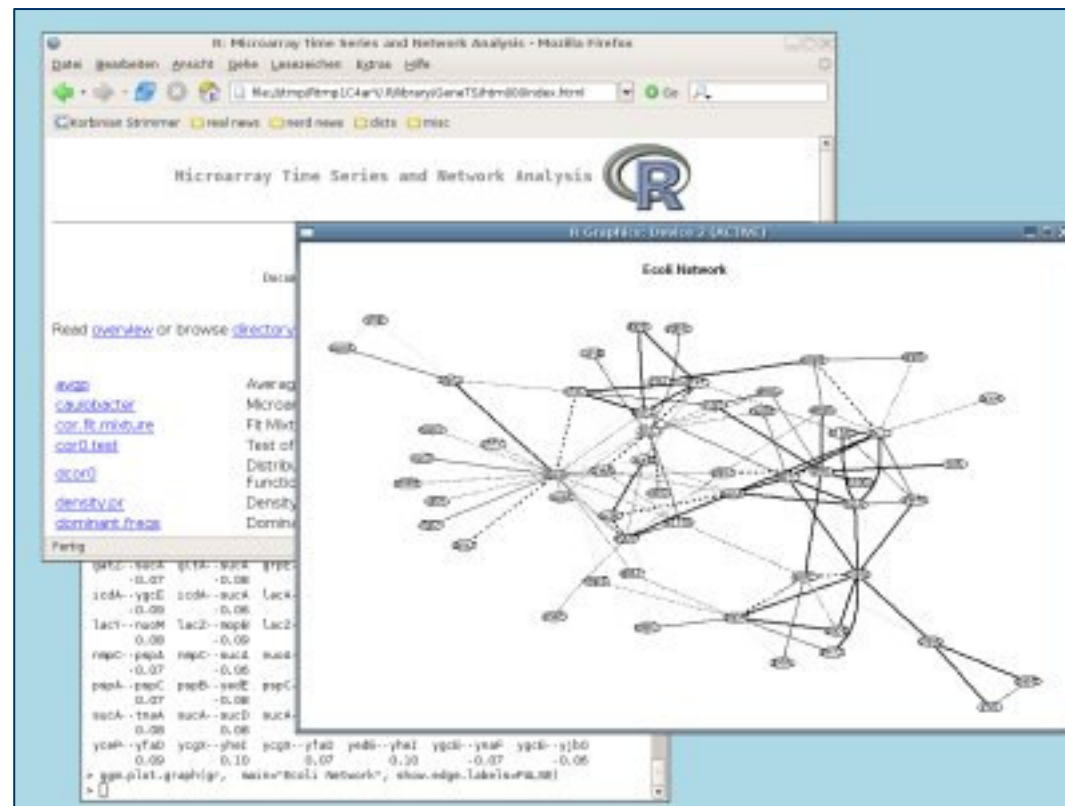
<http://www.biostat.washington.edu/software/jstorey/edge/>

Timecourse differential expression method: Storey JD, Xiao W, Leek JT, Tompkins RG, and Davis RW. (2005) Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences*, 102: 12837-12842.

Software for *Time Series* Gene Expression Data (Differential Expressed Genes and Networks)

12/57

R package, **GeneTS**:
Microarray Time Series and Network Analysis. Detect periodically expressed gene. (Wichert et al., 2004, *Bioinformatics*)



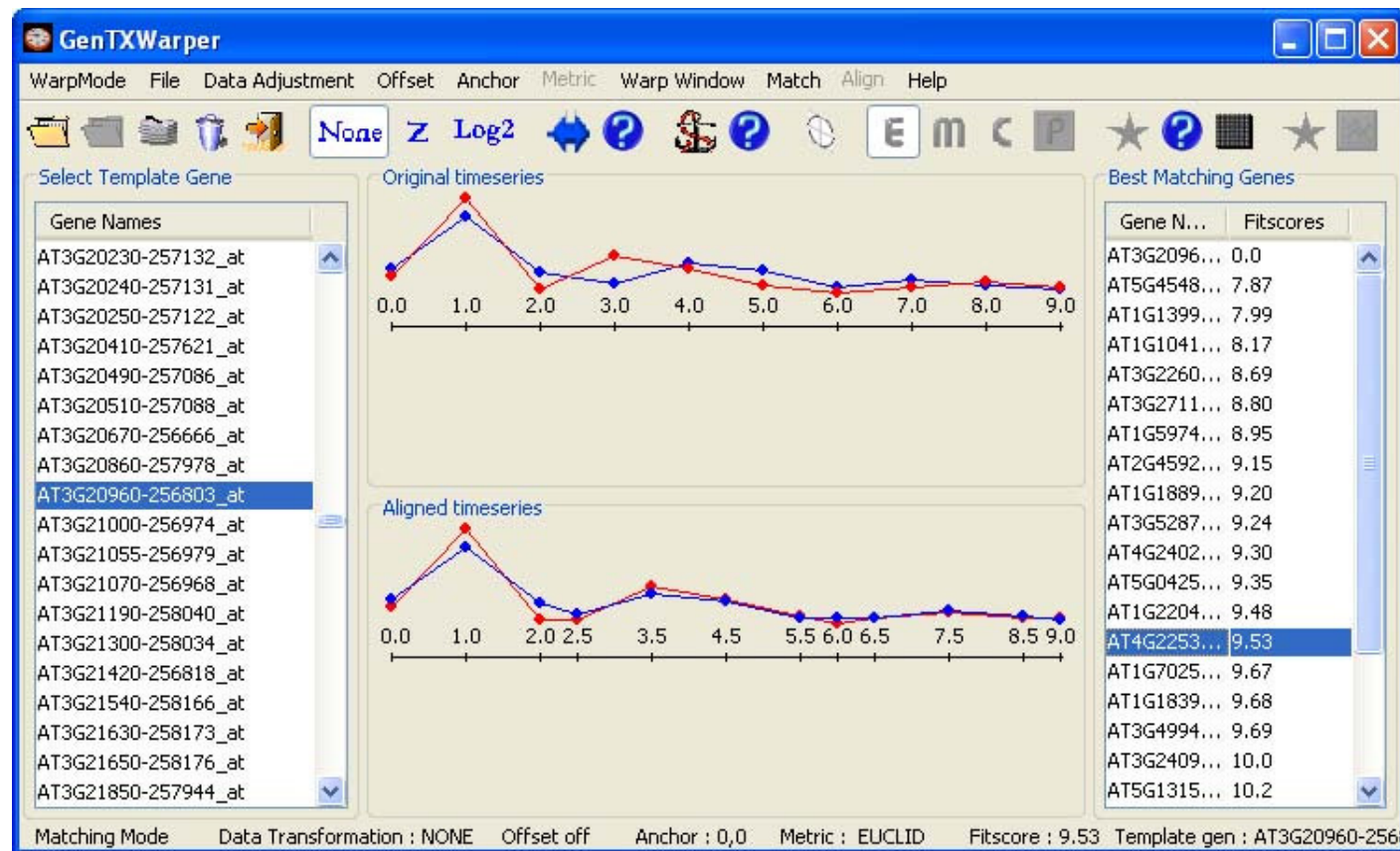
<http://www.strimmerlab.org/software/genets/>

Software for *Time Series* Gene Expression Data (Visualization)

13/57

GenT χ Warper:

Mining of gene expression time series with dynamic time warping techniques
(Criel and Tshiporkova, 2005, *Bioinformatics*)

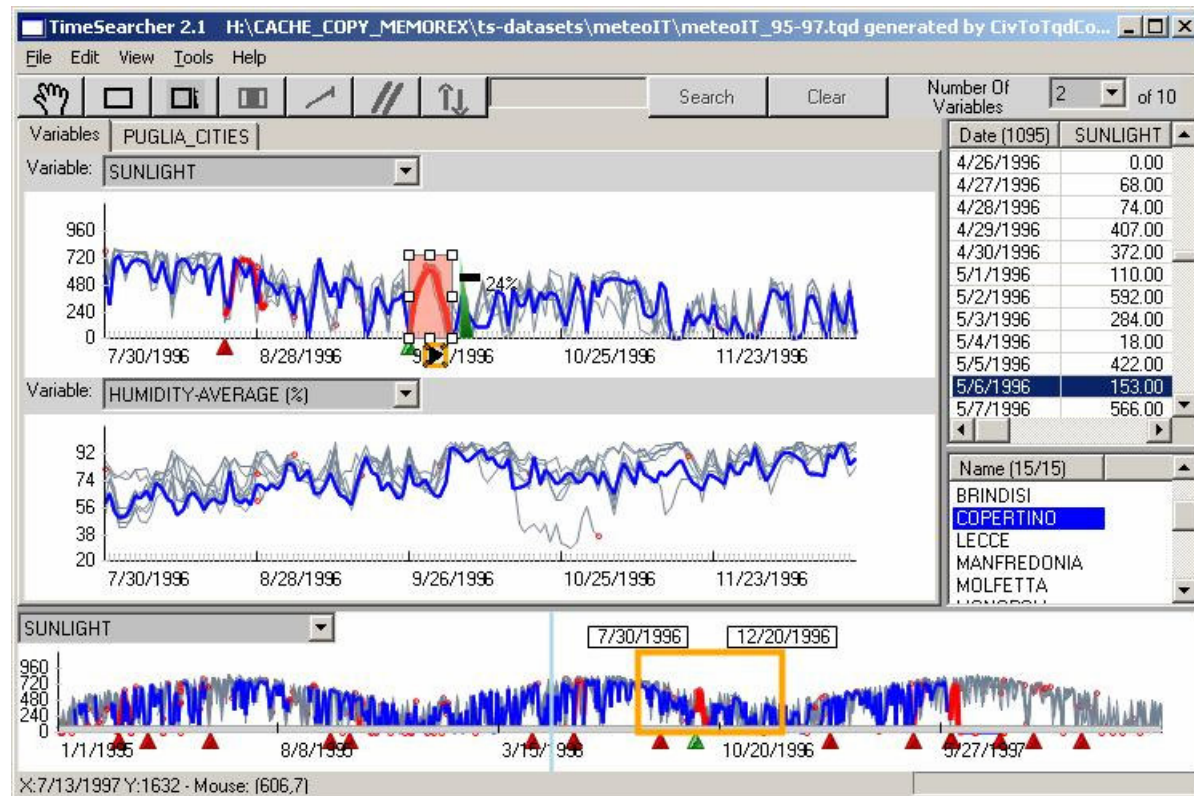


<http://www.psb.ugent.be/cbd/papers/gentxwarper/>

Software for *Time Series* Gene Expression Data (Visualization)

14/57

TimeSearcher:
Visual Exploration of Time-Series Data
(Hochheiser et al, 2003)



<http://www.cs.umd.edu/hcil/timesearcher/>

ORIOGEN:

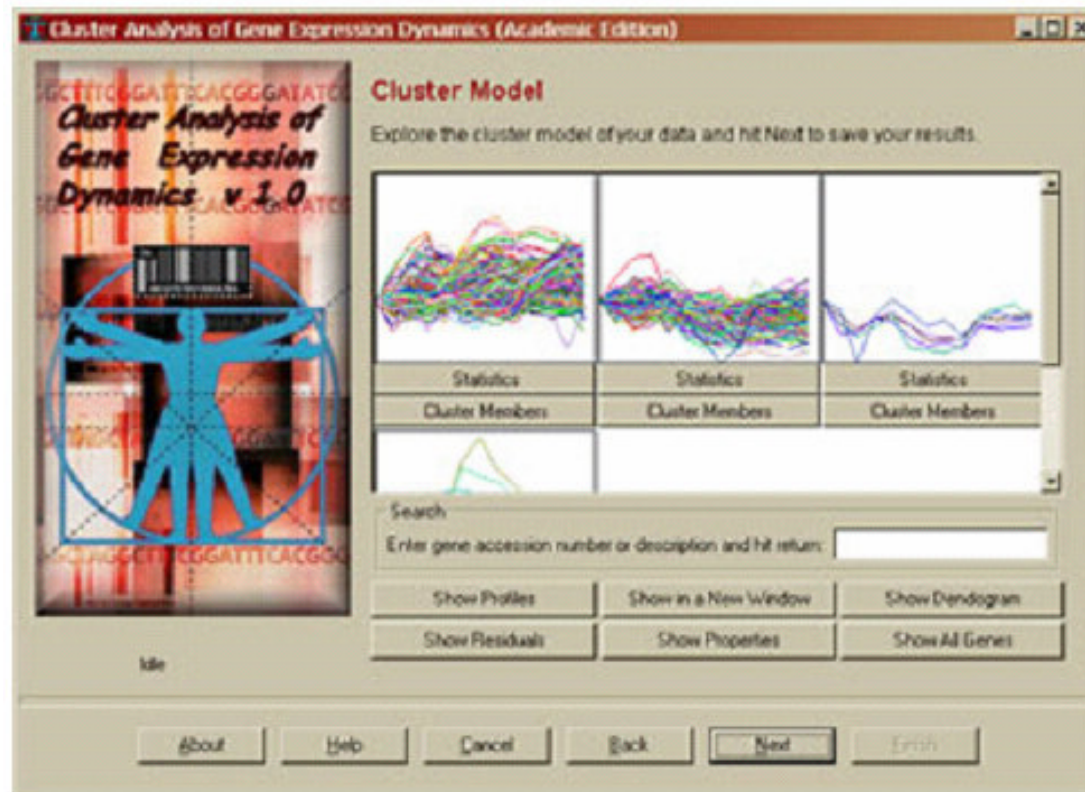
Order Restricted Inference for Ordered Gene Expression clustering for time series.
(Peddada et al., 2005, *Bioinformatics*) <http://dir.niehs.nih.gov/dirbb/oriogen1/index.cfm>

Software for *Time Series* Gene Expression Data (Visualization and Clustering)

15/57

CAGED:

Cluster analysis of gene expression dynamics based on autoregressive equations
(Ramoni et al., 2002, *PNAS*)



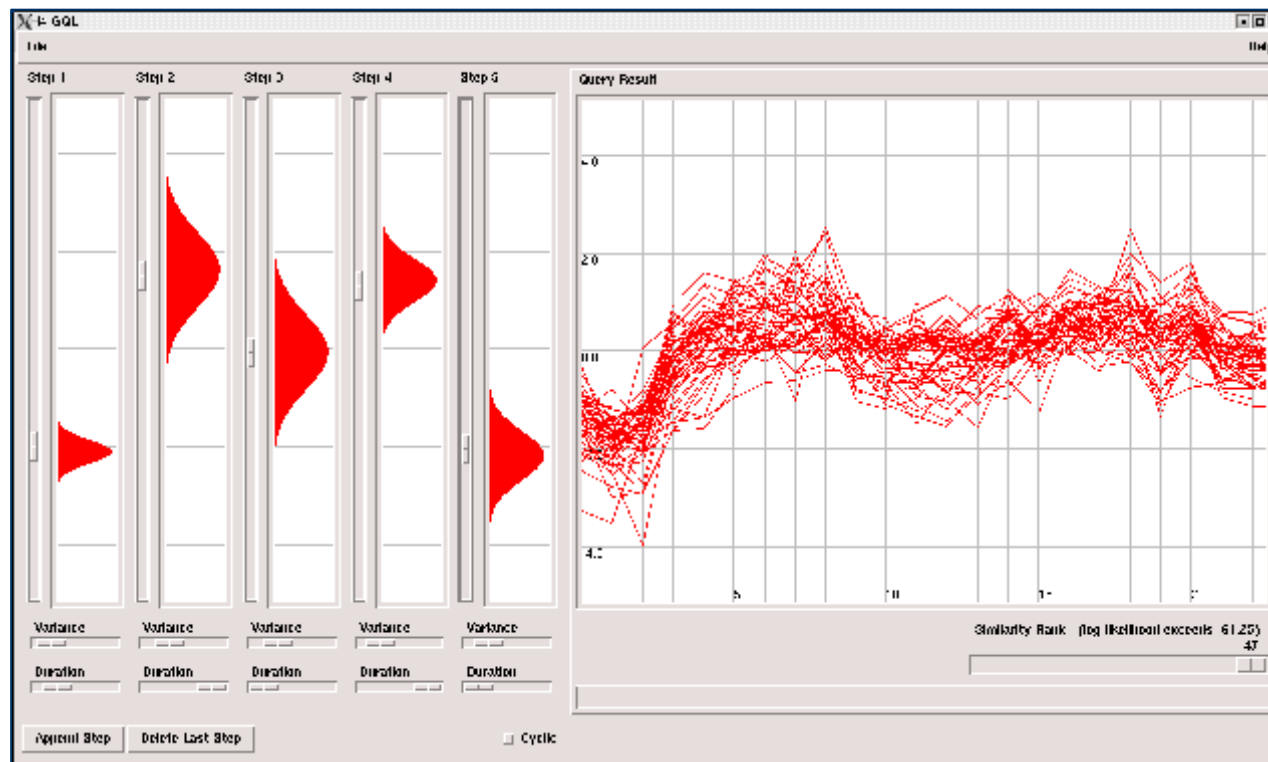
<http://genomethods.org/caged/>

Software for *Time Series* Gene Expression Data (Visualization and Clustering)

16/57

GQL:

The Graphical Query Language: A GHMM-based tool for querying and clustering Gene-Expression time-course data (Costa et al., 2005, *Bioinformatics*)



<http://www.ghmm.org/gql>

Software for *Short Time Series Gene Expression Data* (clustering and visualization)

17/57

STEM:

Short Time-series Expression Miner
(Ernst and Bar-Joseph. 2006, *BMC Bioinformatics*.)

The screenshot displays the STEM software interface, which is organized into four main sections:

- 1. Expression Data Info:** Includes a text field for the Data File (g27_1.txt), a Browse... button, and buttons for View Data File and Repeat Data... Below this are radio buttons for normalization: Log normalize data, Normalize data (selected), and No normalization/add 0. A checkbox for Spot IDs included in the data file is also present.
- 2. Gene Annotation Info:** Features dropdown menus for Gene Annotation Source and Cross Reference Source, both set to Human (EBI). It includes text fields for Gene Annotation File (gene_association.goa_human.gz) and Cross Reference File (human.xrefs.gz), each with a Browse... button. At the bottom, there are checkboxes for Annotations, Cross References, and Ontology, along with a Download the latest... button.
- 3. Options:** Contains a Clustering Method dropdown set to STEM Clustering Method. It also has input fields for Maximum Number of Model Profiles (50) and Maximum Unit Change in Model Profiles between Time Points (2), each with up/down arrows and a button. An Advanced Options... button is located below.
- 4. Execute:** A large yellow Execute button.

At the bottom of the main window, there is a copyright notice: © 2004, Carnegie Mellon University. All Rights Reserved.

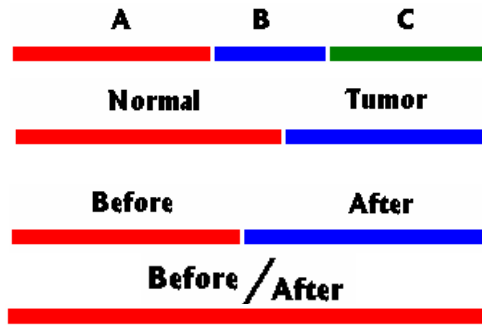
An inset window titled "All Profiles (2)" is overlaid on the right side. It displays a grid of 41 small line graphs, each representing a gene profile. The profiles are ordered by significance. Below the grid, there are buttons for Filtered Gene List, Main Gene Table, Order Profiles By..., Order Clusters By..., and Compare... The title of the inset window is "Clusters ordered based on number of genes and profiles ordered by significance (default)".

<http://www.cs.cmu.edu/~jernst/stem/>

Some Issues

- **The p-values**
- **Multiple Testing Corrections**
- **Permutation Test**
- **Correlation Coefficient**
- **Gene Set Enrichment Analysis**

Finding Differentially Expressed Genes



→ More than two samples

→ Two-sample (independent)

→ Paired-sample (dependent)

Cy 5: treatment

Cy 3: control

MA Table	exp01	exp02	exp03	exp04	exp05	exp...	exp P
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58

MA Table	exp01	exp02	exp03	exp04	exp05	exp...	exp P
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58
gene003	0.87	0.25	-0.17	0.18	-0.13		-1.02
gene004	1.57	1.03	1.22	0.31	0.16		-1.02
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44
gene006	0.04	-0.12	0.31	0.16	0.17		0.08
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63
gene013	-0.16	0.29	0.17	-0.28	-0.02		-0.04
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40
gene022	-0.11	0.13	0.41	0.60	0.23		0.19
gene...							
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66

p-values

0.067
0.052
0.013 *
0.016 *
0.112
0.017 *
0.059
0.063
0.516
-0.009 *
0.068
0.030 *
0.002 *
0.423
0.084
0.048
0.018 *
0.538
0.053
0.074
0.764
0.423
0.723

Significance Level: alpha

H0: no differential expressed.
 ■ The test is significant = Reject H0

- p : probability of observing your data under the assumption that the null hypothesis is true.
- p : probability that you will be in error if you reject the null hypothesis.
- p : probability of **false positives** (Reject H0 | H0 true).

Microarray Data Matrix

gene001	-0.48	-0.42	0.87	0.92	0.67	-0.35
⋮						
gene022	-0.11	0.13	0.41	0.60	0.23	0.19

The p -values for detecting DE genes

20/57

The **p-value** is the probability that a gene's expression level are different between the two groups **due to chance**.

False Positive = (Reject **H₀** | **H₀** true)
= concluding that a gene is differentially expressed when in fact it is not.

Decision Rule

- Reject H_0 if P is less than alpha.
- $P < 0.05$ commonly used. (Reject **H₀**, the test is significant)
- The lower the p-value, the more significant the difference between the groups.

Type I Error (alpha): calling genes as differentially expressed when they are NOT

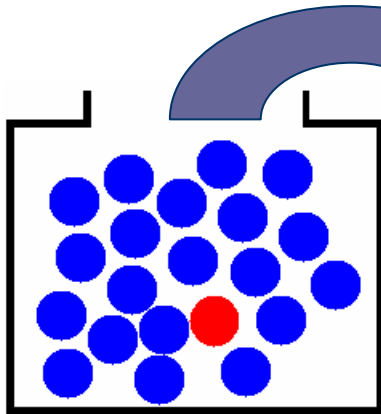
Type II Error: NOT calling genes as differentially expressed when they ARE

Hypothesis Testing		Truth	
		H ₀	H ₁
Decision	Reject H ₀	Type I Error (alpha) (false positive)	Right Decision (true positive)
	Don't Reject H ₀	Right Decision	Type II Error (beta)

$$\text{Power} = 1 - \beta.$$

Multiple Hypothesis Correction

21/57



Population

20 marbles:

19: 

1: 

- What are the odds of randomly sampling the red marble by chance? **It is 1 out of 20.**
- Sample a single marble (and put it back) 20 times. **Have a much higher chance to sample the red marble.**

This is exactly what happens when testing several thousand genes at the same time.

Imagine that the **red** marble is a **false positive gene**: the chance that false positives are going to be sampled is higher the more genes you apply a statistical test on.

Multiplicity of Testing

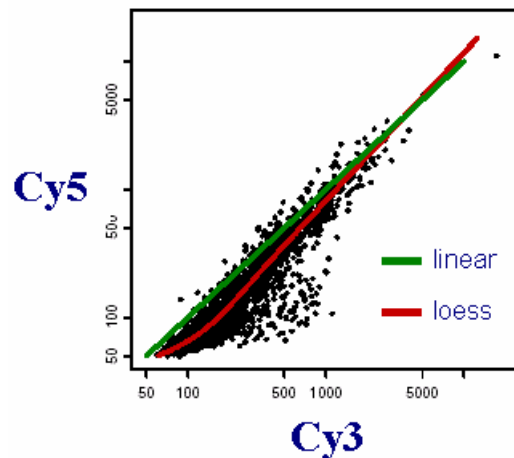
X: false positive gene

$$\begin{aligned} P(X \geq 1) \\ &= 1 - P(X = 0) \\ &= 1 - 0.95^n \end{aligned}$$

Number of genes tested (N)	False positives incidence	Probability of calling 1 or more false positives by chance ($100(1-0.95^N)$)
1	1/20	5%
2	1/10	10%
20	1	64%
100	5	99.4%

Multiplicity of Testing (for detecting DE genes)

22/57



- Label reference sample with **Cy3** and **Cy5**:
 - No genes are DE.
 - Differences are experimental error.
- **p-value=0.01**: each gene would have a 1% chance of having a p-value of less than 0.01, and thus be significant at the 1% level.

Example: 10000 genes

- Expect to find 100 significant genes at the 1% level.
- Expect to find 10 genes with a p-value less than 0.001.
- Expect to find 1 gene with p-value less than 0.0001

Question:

Truly differentially expressed?, or a false positive results (because we are analyzing a large number of genes?)?

Types of Error Control

23/57

- Multiple testing correction **adjusts the p-value** for each gene to keep the **overall error rate** (or false positive rate) to less than or equal to the user-specified p-value cutoff or error rate individual.

Multiple Testing

	# Reject H_0	# not Reject H_0	
# true H_{0j}	V	U	m_0
# true H_{1j}	S	T	m_1
	R	$m - R$	m

V : false positives = Type I errors

T : false negatives = Type II errors

PCER: Per-comparison error rate

PFER: Per-family error rate

PWER: Family-wise error rate

FDR: False discovery rate

Type One Errors Rates

$$\text{PCER} = \frac{E[\mathbf{V}]}{m}$$

$$\text{PFER} = E[\mathbf{V}]$$

$$\text{FWER} = P(\mathbf{V} \geq 1)$$

$$\text{FDR} = E\left(\frac{\mathbf{V}}{\mathbf{R}} \mid \mathbf{R} > 0\right) \Pr(\mathbf{R} > 0)$$

Power = Reject the false null hypothesis

$$\text{Any-pair Power} = P(\mathbf{S} \geq 1)$$

$$\text{Per-pair Power} = \frac{E[\mathbf{S}]}{m_1}$$


$$\text{All-pair Power} = P(\mathbf{S} = m_1)$$

Multiple Testing Corrections

24/57

Test Type	Type of Error control	Genes identified by chance after correction
Bonferroni	Family-wise error rate	If error rate equals 0.05, expects 0.05 genes to be significant by chance
Bonferroni Step-down		
Westfall and Young permutation		
Benjamini and Hochberg	False Discovery Rate	If error rate equals 0.05, 5% of genes considered statistically significant (that pass the restriction after correction) will be identified by chance (false positives).

most stringent
More false negatives



More false positives
least stringent

Bonferroni Correction

25/57

- The p-value of each gene is multiplied by the number of genes in the gene list.
- If the corrected p-value is still below the error rate, the gene will be significant:
 - Corrected p-value = p-value * n < 0.05.
 - If testing 1000 genes at a time, the highest accepted individual uncorrected p-value is 0.00005, making the correction very stringent.
- With a Family-wise error rate of 0.05 (i.e., the probability of at least one error in the family), the expected number of false positives will be 0.05.



Bonferroni, Carlo Emilio
(1892-1960)

- Italian mathematician
- Bonferroni correction (1935-36)
- Bonferroni's Inequality

$$P\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n P(E_i)$$

Benjamini and Hochberg FDR

26/57

- This correction is the least stringent of all 4 options, and therefore tolerates more false positives.
- There will be also less false negative genes.
- The error rate is a proportion of the number of called genes.
- FDR: Overall proportion of false positives relative to the total number of genes declared significant.

$$\text{Corrected P-value} = p\text{-value} * (n / R_i) < 0.05$$

Let $n=1000$, error rate= 0.05

Gene name	p-value (from largest to smallest)	Rank	Correction	Is gene significant after correction?
A	0.1	1000	No correction	$0.1 > 0.05 \rightarrow$ No
B	0.06	999	$1000/999 * 0.06 = 0.06006$	$0.06006 > 0.05 \rightarrow$ No
C	0.04	998...	$1000/998 * 0.04 = 0.04008$	$0.04008 < 0.05 \rightarrow$ Yes



The Permutation Test

27/57

- The **permutation test** is a test where the null hypothesis allows to reduce the inference to a **randomization problem**.

Randomization test

- Works of R.A. Fisher and E.J.G. Pitman in the 1930s.
- Possible to ascribe a probability distribution to the difference in the outcome possible **under null hypothesis**.
- The outcome data are analyzed many times
 - once for each acceptable assignment that could have been possible under **H0**
 - and then compared with the observed result,
 - without dependence on additional distributional or model-based assumptions.

Ref: Mansmann, U. (2002), Practical microarray analysis: resampling and the Bootstraap. Heidelberg.

The Permutation Test (conti.)

28/57

Coexpression of genes

H_0 : Gene 1 and Gene 2 are not correlated.

Test statistic T:

Pearson (or Spearman) correlation coefficient, calculate t_{obs}

Randomization: Under H_0 it is possible to permute the values observed for Gene 2. There are $n!$ possibilities.

p-value: $p = P(T \geq t_{obs} \mid H_0) \approx \frac{\#\{T^* \geq t_{obs}\}}{n!}$

Data

Gene1	Gene2
g_1^1	g_1^2
\vdots	\vdots
g_n^1	g_n^2



$g_{(1)}^1$	$g_{(1)}^2$
\vdots	\vdots
$g_{(m)}^1$	$g_{(m)}^2$

Random Permutation for group labels

Gene 1	Gene 2	Group	Group
1.4482	1.0709	1	2
0.4850	0.9324	1	1
1.1331	1.2379	1	4
		\vdots	\vdots
0.8015	0.6765	2	1
		\vdots	\vdots
1.3726	1.2373	3	4
		\vdots	\vdots
1.1030	1.735	4	2
0.5148	1.0015	4	3



Perform a Permutation Test (general):

1. Analyze the problem, choice of null hypothesis
2. Choice of test statistic **T**
3. Calculate the value of the test statistic for the observed data: **tobs**
4. Apply the randomization principle and look at all possible permutations, this gives the distribution of the test statistic **T** under H_0 .
5. Calculation of *p-value*:

$$p = P(T \geq t_{obs} \mid H_0) \approx \frac{\#\{t^* \geq t_{obs}\}}{\# \text{ permutations}}$$

Correlation Coefficient and Distance

29/57

Cov	x1	x2	x3	x4	x p
x1	1.00	0.48	0.10	-0.10	-0.28
x2	0.48	1.00	0.41	0.22	-0.23
x3	0.10	0.41	1.00	0.36	-0.05
x4	-0.10	0.22	0.36	1.00	0.10
x p	-0.28	-0.23	-0.05	0.10	1.00

Proximity Matrix

Pearson Correlation Coefficient

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Data Matrix

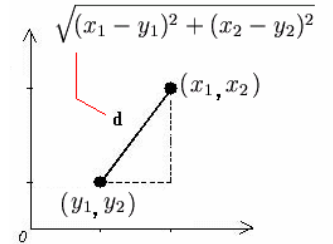
	x	y				
Data	x1	x2	x3	x4	...	x p
subject01	-0.48	-0.42	0.87	0.92		-0.18
subject02	-0.39	-0.58	1.08	1.21		-0.33
subject03	0.87	0.25	-0.17	0.18		-0.44
subject04	1.57	1.03	1.22	0.31		-0.49
subject05	-1.15	-0.86	1.21	1.62		0.16
subject06	0.04	-0.12	0.31	0.16		-0.06
subject07	2.95	0.45	-0.40	-0.66		-0.38
subject08	-1.22	-0.74	1.34	1.50		0.29
subject09	-0.73	-1.06	-0.79	-0.02		0.44
subject10	-0.58	-0.40	0.13	0.58		0.02
subject11	-0.50	-0.42	0.66	1.05		0.06
subject12	-0.86	-0.29	0.42	0.46		0.10
subject13	-0.16	0.29	0.17	-0.28		-0.55
subject14	-0.36	-0.03	-0.03	-0.08		-0.25
subject15	-0.72	-0.85	0.54	1.04		0.24
subject16	-0.78	-0.52	0.26	0.20		0.48
subject17	0.60	-0.55	0.41	0.45		-0.66
⋮						
subject n	-2.29	-0.64	0.77	1.60		0.55
mean	0.07	-0.04	0.44	0.31	...	-0.21

Euclidean Distance

$$x = (x_1, x_2, \dots, x_n)$$

$$y = (y_1, y_2, \dots, y_n)$$

$$d_{xy} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$



Pearson Correlation coefficient: great success in computational biology, especially in clustering algorithm.

Advantage

it can group together genes with **similar** expression profiles even if their units of change are different.

Disadvantage

The Correlation Coefficient can take negative values and does not satisfy the **triangle inequality** and thus **not a metric**.

Correlation Coefficient and Distance

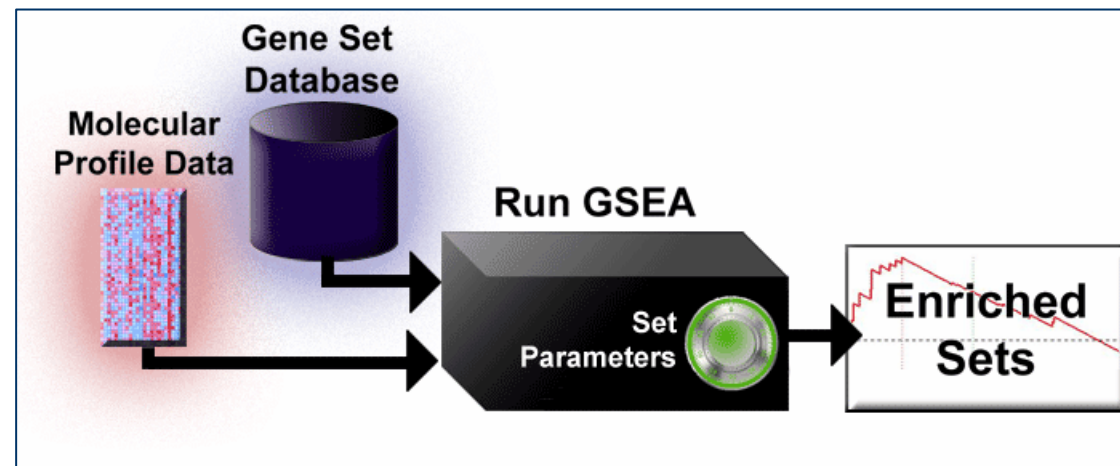
30/57

- Use $d = 1 - r$: (Eisen *et al.* 1998) $d_{rs} = 1 - c_{rs}$
 - still not a metric, does not satisfy the triangle inequality.
- Generalized version of the triangle inequality:
 - $g_m(x, z) \leq 2(g_m(x, y) + g_m(y, z)) \rightarrow$ a transitive measure.
 - When using the correlation coefficient two highly dissimilar profiles can't be very similar to a third profile.
- The standard transformation from a similarity matrix C to a distance matrix D is given by $d_{rs} = (c_{rr} - 2c_{rs} + c_{ss})^{1/2}$.
- Other transformations
(Chatfield and Collins 1980, Section 10.2)

Gene Set Enrichment Analysis

31/57

- **Single-gene analysis** may **miss important effects** on metabolic pathways, transcriptional programs and stress response.
- Study same biological system, **little overlap** statistically significant genes.
- Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows *statistically significant*, concordant differences between two biological states (e.g. phenotypes).



GSEA-p
Molecular Signature Database (MSigDB)

Source: <http://www.broad.mit.edu/gsea/>

Subramanian, Tamayo, et al. (2005), Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102, 15545-15550

SAM

32/57

SAM assigns a score to each gene in a microarray experiment based upon its change in gene expression relative to the standard deviation of repeated measurements.

- **SAM plot**: the number of observed genes versus the expected number. This visualizes the outlier genes that are most dramatically regulated.
- **False discovery rate**: is the percent of genes that are expected to be identified by chance.
- **q-value**: the lowest false discovery rate at which a gene is described as significantly regulated.

SAM does not do any normalization!

The screenshot shows the SAM Plot Control dialog box overlaid on an Excel spreadsheet. The dialog box is titled "Significance Analysis of Microarrays" and includes the following settings:

- Choose Response Type:** Two class, unpaired data (selected)
- Data in Log Scale?:** Logged (base 2) (selected)
- Web Link Option:** Name (selected)
- Number of Permutations:** 100 (selected)
- Imputation Engine:** K-Nearest Neighbors Imputer (selected)
- Random Number Seed:** 1234567

The background Excel spreadsheet shows a table with columns A through J and rows 1 through 35. The data in the spreadsheet is as follows:

	A	B	C	D	E	F	G	H	I	J
1			1	1	2	2	1	1	2	2
2	AFFX-Biol	100001	7.64252	-0.50242	-1.95964	10.1298	-10.77	-4.47036	-7.65613	7.58627
3	AFFX-Biol	100002	38.1083	4.86575	7.87245	-13.5974	-9.79556	-13.4659	-8.91639	-5.07128
4	AFFX-Biol	100003	21.1568	5.96949	3.20649	-4.74098	-3.70624	-12.351	-10.1714	0.63687
5	AFFX-Biol	100004	187.22	-23.8126	16.7677	14.1087	-99.7636	-89.1146	-10.9241	5.51881
6	AFFX-Biol	100005	64.135	53.612	1.97359	81.4896	-61.0625	-55.0031	-21.5555	-63.589
7	AFFX-Biol	100006	43.2501	39.4						
8	AFFX-Biol	100007	38.7908	191						
9	AFFX-Cre	100008	676.819	48						
10	AFFX-Cre	100009	731.028	559						
11	AFFX-Biol	100010	-45.0362	18.9						
12	AFFX-Biol	100011	9.83463	-23.1						
13	AFFX-Biol	100012	-6.23839	1.8						
14	AFFX-Biol	100013	-76.144	-13.8						
15	AFFX-Biol	100014	-9.927	-10.8						
16	AFFX-Biol	100015	-13.4207	-10.9						
17	AFFX-Biol	100016	5.39054	6.4						
18	AFFX-Cre	100017	-4.37465	-9.7						
19	AFFX-Cre	100018	4.7197	-26.8						
20	hum_alu	100019	221.097	886						
21	AFFX-Dap	100020	-20.7535	-12.1						
22	AFFX-Dap	100021	18.6053	-13						
23	AFFX-Dap	100022	12.0019	8.4						
24	AFFX-Lys	100023	-8.29982	-0.2						
25	AFFX-Lys	100024	-20.6604	-15.9						
26	AFFX-Lys	100025	5.48261	-14						
27	AFFX-Phe	100026	-3.1287	-3.2						
28	AFFX-Phe	100027	-19.0192	-35.1						
29	AFFX-Phe	100028	14.2111	-20.8						
30	AFFX-Thr	100029	-24.7369	-13.9						
31	AFFX-Thr	100030	4.9279	-0.6						
32	AFFX-Thr	100031	11.3658	-26.1						
33	AFFX-Trpr	100032	16.9344	10.4						
34	AFFX-Trpr	100033	24.6975	-39						
35	AFFX-Trpr	100034	-5.37853	-35.8						

Tusher VG, Tibshirani R, Chu G.(2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* 98(9):5116-21.

SAM: Significance Analysis of Microarrays
<http://www-stat.stanford.edu/~tibs/SAM/>

SAM: Response Type

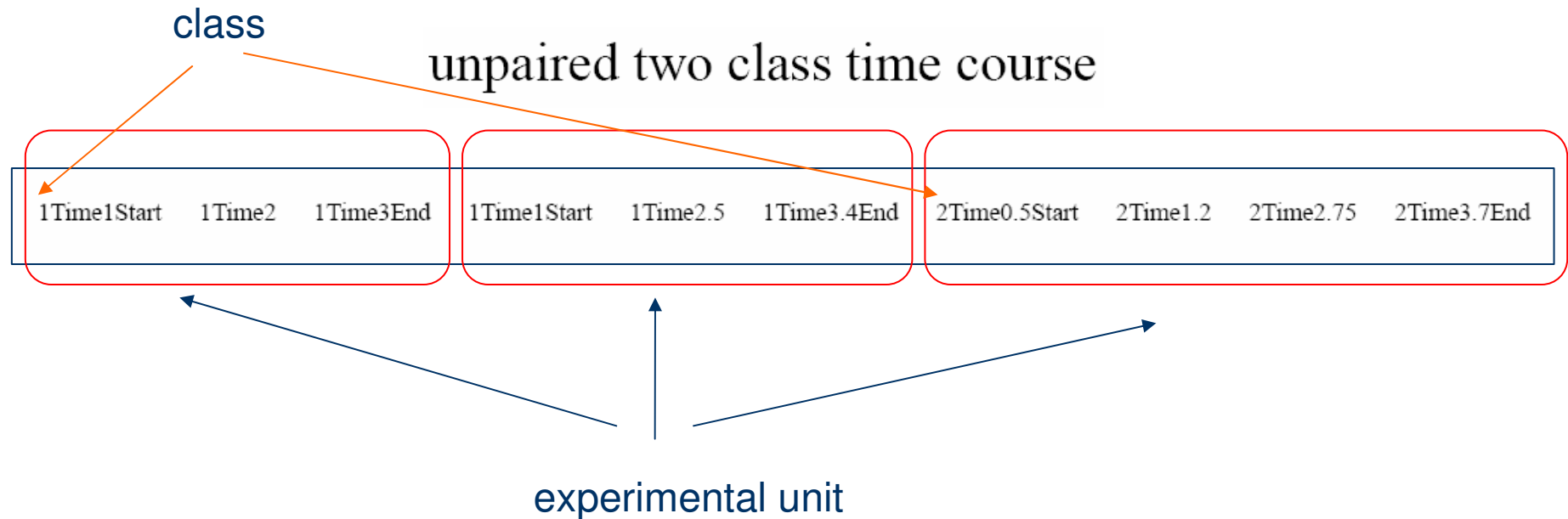
33/57

Response type	Coding
Quantitative	Real number eg 27.4 or -45.34
Two class (unpaired)	Integer 1, 2
Multiclass	Integer 1, 2, 3, ...
Paired	Integer -1, 1, -2, 2, etc. eg - means Before treatment, + means after treatment -1 is paired with 1, -2 is paired with 2, etc.
Survival data	(Time, status) pair like (50,1) or (120,0) First number is survival time, second is status (1=died, 0=censored)
One class	Integer, every entry equal to 1
Time course, two class (unpaired)	(1 or 2)Time(t)[Start or End]
Time course, two class (paired)	(-1 or 1 or -2 or 2 etc)Time(t)[Start or End]
Time course, one class	1Time(t)[Start or End]
Pattern discovery	eigengenes, where k is one of 1,2,... number of arrays

SAM Users guide and technical document

SAM: Time Series

34/57

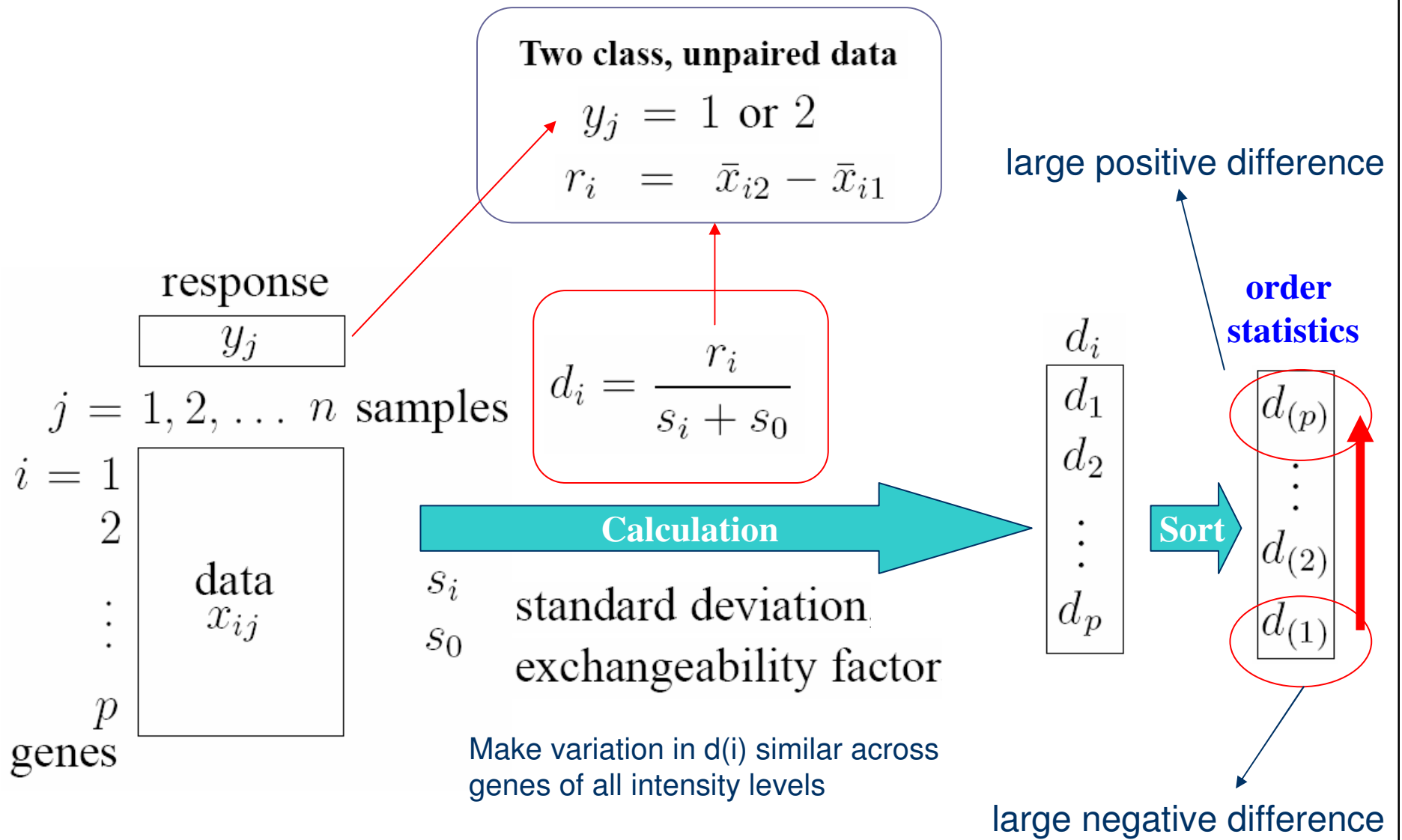


- Paired data time courses: class label is -1, or 1 or -2 or 2.
- One class time courses: class label is a 1.

NOTE: SAM summarizes each time course by a *slope* or a *signed area*, and then treats the summarized data in the same way as it treats two class, one class, or a two-class paired design.

SAM: Significance Analysis of Microarrays

35/57

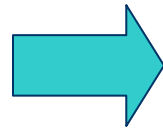


SAM: Expected Test Statistics

response

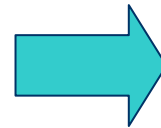
$$y_j$$

$$1, 1, \dots, 2, \dots, 2$$



Permutation

$$1, 2, 1, 2, 1, \dots, 1$$



$$r_i^* = \bar{x}_{i2}^* - \bar{x}_{i1}^*$$

$$d_i^* = \frac{r_i^*}{s_i^* + s_0^*}$$



$$d_{(p)}^{*b}$$

⋮

$$d_{(2)}^{*b}$$

$$d_{(1)}^{*b}$$

$$b = 1, 2, \dots, B$$

$$\bar{d}_{(i)} = (1/B) \sum_b d_{(i)}^{*b}$$



expected order statistics

$$\bar{d}_{(p)}$$

⋮

$$\bar{d}_{(2)}$$

$$\bar{d}_{(1)}$$

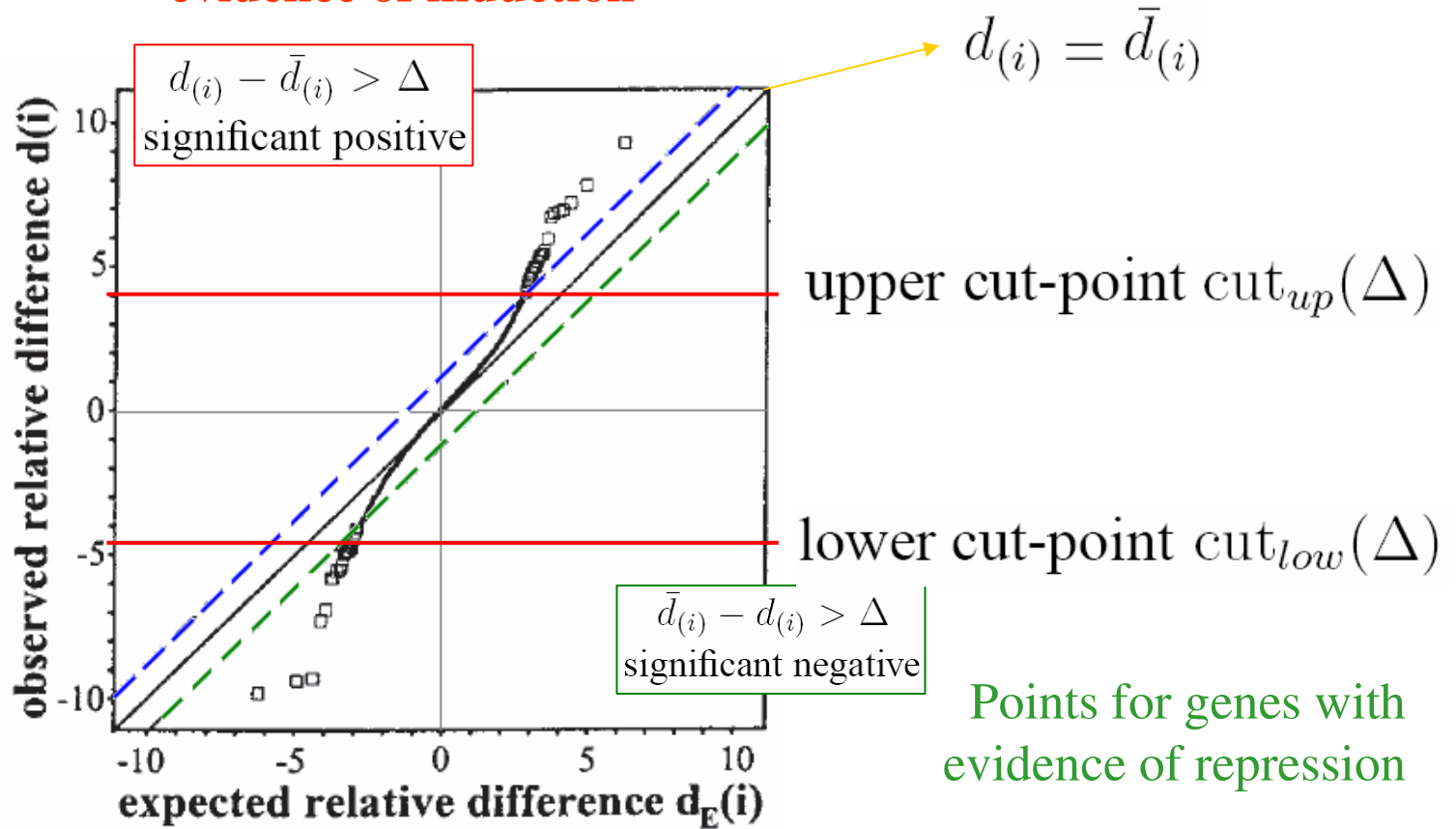
SAM Plot

Points for genes with evidence of induction

$d_{(p)}$
⋮
 $d_{(2)}$
 $d_{(1)}$

vs

$\bar{d}_{(p)}$
⋮
 $\bar{d}_{(2)}$
 $\bar{d}_{(1)}$



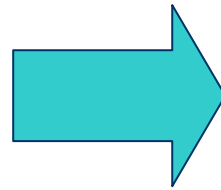
Points for genes with evidence of repression

Estimating FDR for a Selected Δ

38/57

$$\begin{matrix} d_{(p)}^{*b} \\ \vdots \\ d_{(2)}^{*b} \\ d_{(1)}^{*b} \end{matrix}$$

$$b = 1, 2, \dots, B$$



number of falsely called genes,

fall above $\text{cut}_{up}(\Delta)$
or
fall below $\text{cut}_{low}(\Delta)$

False Discovery Rate (FDR) =

[median (or 90th percentile) of the number of falsely called genes]

[the number of genes called significant]

← in the original data.

The q-value of a gene is the false discovery rate for the gene list that includes that gene and all genes that are more significant. It is computed by finding the smallest value of $\hat{\Delta}$ for which the gene is called significant, and then is the FDR corresponding to $\hat{\Delta}$.

John D. Storey (2002) A direct approach to false discovery rates, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (3), 479–498.

- The **q-value** gives the scientist a hypothesis testing error measure for each observed statistic with respect to **pFDR**.
- The **p-value** accomplishes the same goal with respect to the **type I error**, and the **adjusted p-value** with respect to **FWER**.

Interpretation of Results for Time Series Data by SAM

39/57

SAM Summarize each time course by a **slope** (least squares slope of expression vs time), or a **signed area**.

For two class unpaired data:

Slope: summarizes each time series by a slope.

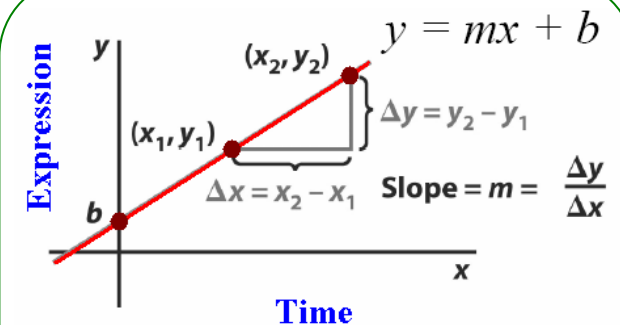
- Compare slopes across the two groups.
- Useful for finding genes with a consistent increase or decrease over time.

Signed area: the time course profile is shifted so that it is zero at the first time point.

- Counting positive area above the line and negative below the line.
- Compares the areas across the groups.
- Useful for finding genes that rise and then level off or come back down to their baseline.

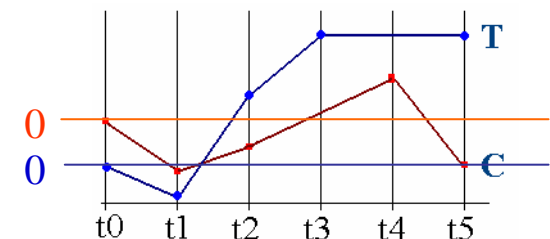
$$r_i = \bar{x}_{i2} - \bar{x}_{i1}$$

$$d_i = \frac{r_i}{s_i + s_0}$$



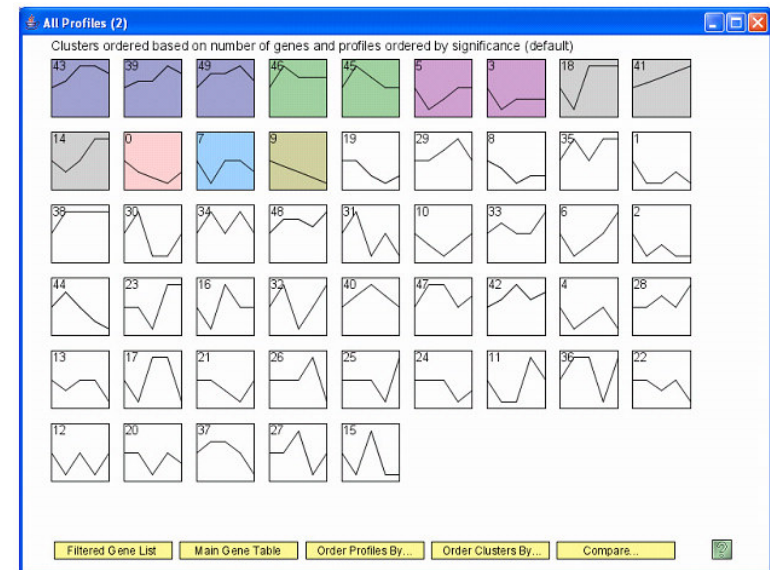
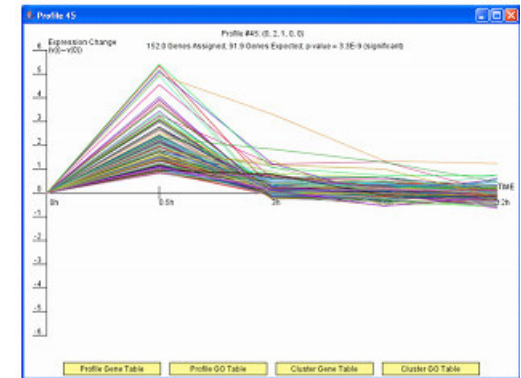
$$m = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{D}$$

$$D = n\sum(x_i^2) - (\sum x_i)^2$$



Purpose: Identifying Significant Expression Patterns (Clustering Short Time Series Gene Expression Data)

- Unique challenges
 - Thousands of genes are being profiled simultaneously while the number of time points is few.
 - Many genes will have the same expression pattern just by random chance.
 - Generally require the estimation of many parameters and are less appropriate for short time series data.
 - Do not differentiate between real and random patterns.



STEM: 4 Steps

41/57

1. Selecting Model Profiles

- select a set of distinct and representative temporal expression profiles (Model Profiles), selected independent of the data.

2. Assigning Genes to Model Profiles

- Assign each gene passing the filtering criteria to the model profile that most closely matches the gene's expression profile as determined by the correlation coefficient.

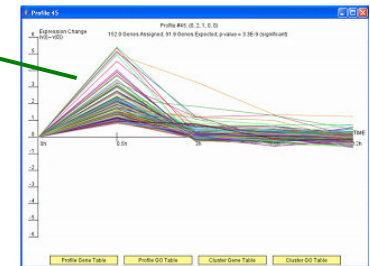
3. Identifying Significant Model Profiles

- Algorithm can determine which profiles have a statistically significant higher number of genes assigned using a permutation test.

4. Grouping Significant Profiles

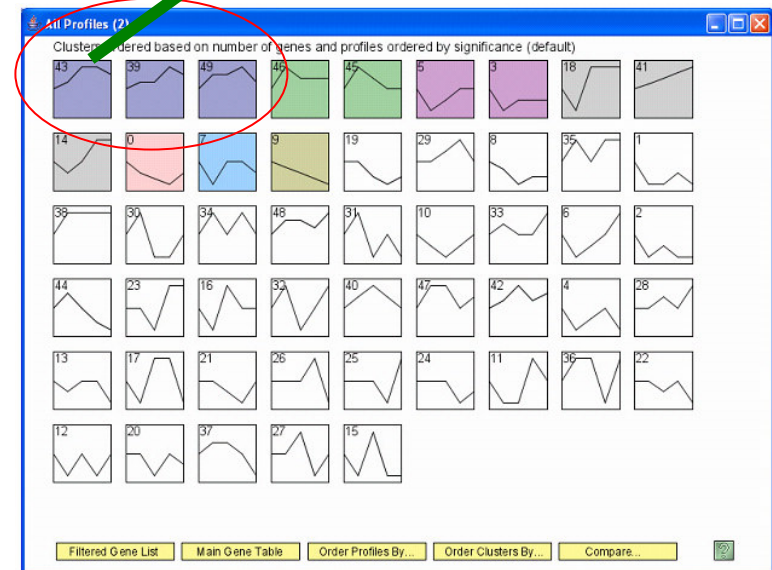
- Significant model profiles can be grouped based on similarity to form clusters of significant profiles.

Genes



Cluster

Model Profile



1. Selecting Model Profiles (pre-defined)

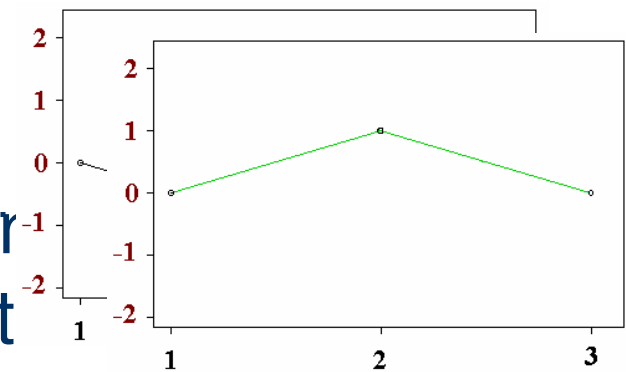
42/57

- Expression values (*log ratios*), where the ratios are with respect to the expression of the first time point (always be 0).

A parameter c : controls the amount of change a gene can exhibit between successive time points.

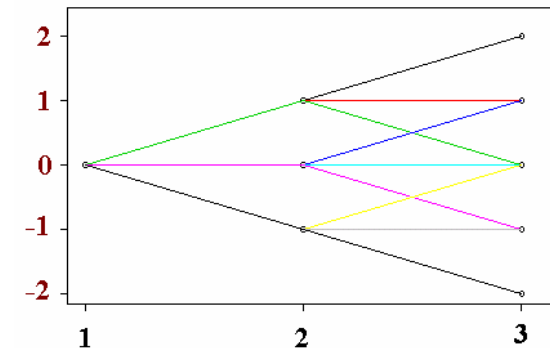
- Select a set of model expressions which are distinct from one another.
- $c=2$: a gene can go up either one or two units, stay the same, or go down one or two units.

- n time points, $\rightarrow (2c + 1)^{n-1}$ distinct profiles.
 - 5 time points and $c = 1$, would result in 81 model profiles.
 - 6 time points and $c = 2$, would result in 3125 model profiles.
- Select m representative profiles (a subset of profiles)
 - Greedy approximation algorithm (see Ernst et al., 2005, *Bioinformatics*).



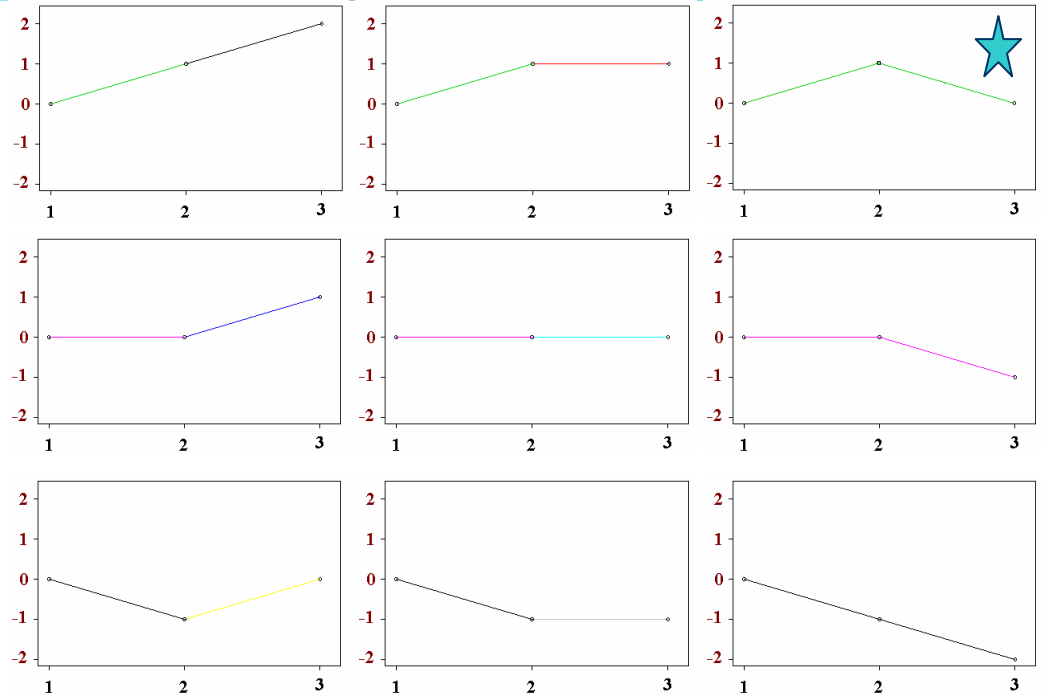
Example:

$$c = 1, n = 3 \Rightarrow Np = 9$$



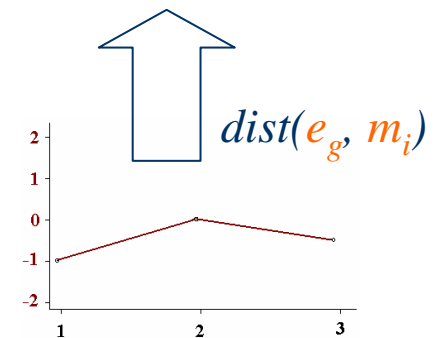
2. Assigning Genes to Model Profiles

- Given a set m of model profiles and a set of genes G , each gene g in G is assigned to a model expression profiles m_i in m such that $dist(e_g, m_i)$ is the minimum over all m_i in m .
 - e_g is the temporal expression profile for gene g .
 - Ties: assign g to all of these profiles (h), weights $1/h$.
- $T(m_i)$: The number of genes assigned to each model profile.



Example:
 $c = 1, n = 3 \Rightarrow Np = 9$

	A	B	C	D	E
1	SPOT	Gene Symbol	0h	0.5h	3h
2	1	ZFX	-1.027	0.158	-0.569
3	2	ZNF133	0.183	-0.068	-0.134
4	3	USP2	-0.67	-0.709	-0.347
5	4	DSCR1L1	-0.923	-0.51	-0.718
6	5	WNT5A	-0.471	-0.264	-0.269
7	6	VHL	-0.327	-0.378	-0.229
8	7	TCF3	-0.021	0.129	-0.209
9	8	TCN2	-0.492	-0.41	-0.306
10	9	TIMP1	-0.111	0.351	0.168
11	10	SERPINA7	-0.468	-0.488	-0.199
12	11	THBD	-1.013	-0.895	-0.743
13	12	EPHA2	0.13	0.313	0.645

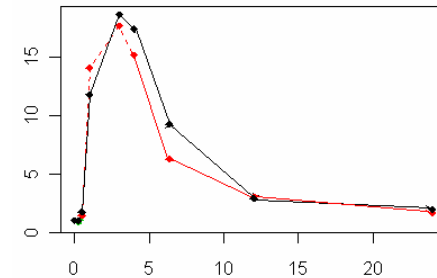


3. Identifying Significant Model Profiles

44/57

Identify model profiles that are significantly enriched for genes.

- Null hypothesis: the data are *memoryless*.
 - i.e., the probability of observing a value at any time point is **independent of past and future values**.
 - Under null hypothesis: any profile we observe is a result of *random fluctuation* in the measured values for genes assigned to that profile.
 - Model profiles that represent true biological function deviate significantly from null hypothesis since many more genes than expected by random change are assigned to them.
- **Permutation Test**: permutation is used to quantify the **expected number** of genes that would have been assigned to each profile if the data were generated at random.

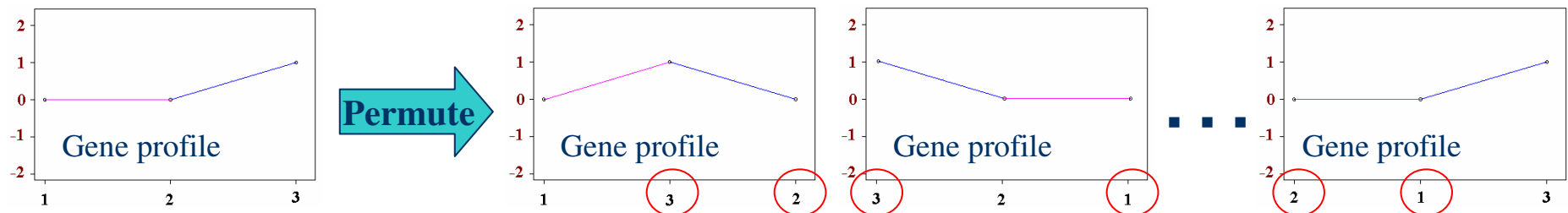


3. Identifying Significant Model Profiles

(conti.)

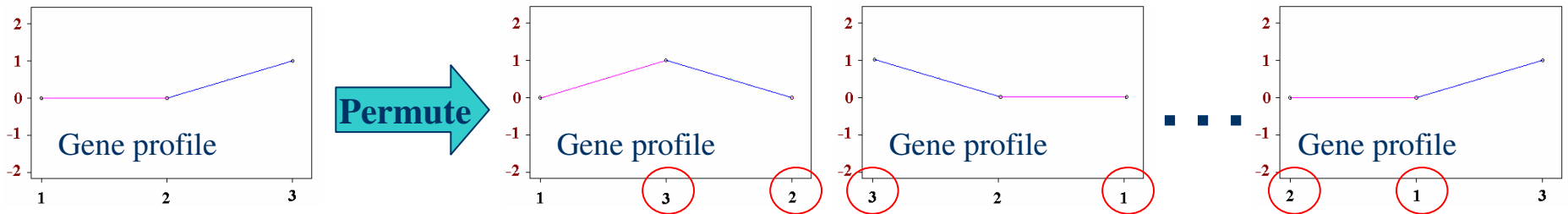
45/57

- Under the null hypothesis, the **order** of the observed values is **random**.
 - as each point is **independent** of any other point.
 - permutations are expected to result in profiles that are **similar** to the null distribution.
- Since there are n time points, each gene has $n!$ possible permutations (can be computed for small n).
- For each possible permutation, assign genes to their closet model profile.
 - Let s_{ij} be the number of genes assigned to model profile i in permutation j .
 - Set $S_i = \sum_j s_{ij}$, then $E_i = S_i/n!$ is the **expected number** of genes for each profile model if the data were indeed generated according to the null hypothesis.



3. Identifying Significant Model Profiles (conti.)

46/57



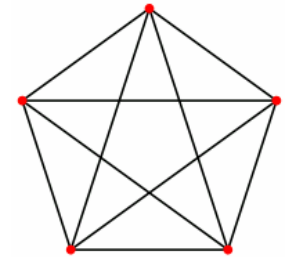
- **Assume:** The number of genes in each profile is distributed as a Binomial with parameters $|G|$ and $E_i/|G|$.
 - Thus the p-value of seeing $T(m_i)$ genes assigned to profile m_i is $P(X \geq T(m_i))$, where $X \sim \text{Binomial}(|G|, E_i/|G|)$.
- **Bonferroni Correction:** consider the number of genes assigned to m_i to statistically be significant if $P(X \geq T(m_i)) < \alpha/m$.

4. Grouping Significant Profiles

47/57

Graph theoretic problem

- **Graph (V, E) :**
 - V : the set of significant model profiles.
 - E : the set of edges.
- Two profiles v_1, v_2 in V are connected with an edge iff $dist(v_1, v_2) < \delta$.
- **Cliques** in this graph correspond to sets of significant profiles which are all similar to one another.
- **To identify large cliques of profiles which are all very similar to each other.**



a clique of size 5

Greedy algorithm: to partition the graph into cliques and thus to group significant profiles.

- Cluster for a significant profile $C_i = \{p_i\}$,
- Initial $C_i = \{p_i\}$, look for a profile p_j such that p_j is the closet profile to p_i that is not already included in C_i .
 - If $dist(p_j, p_k) \leq \delta$ for all profiles p_k in C_i , add p_j to C_i and repeat process,
 - otherwise stop and declare C_i as the cluster for p_i .
- After obtaining clusters for all significant profiles, select the cluster with **largest number of genes** (by counting the number of genes in each of the profiles that are included in this cluster), remove all profiles in that cluster and repeat the above process.
- The algorithm terminates when all profiles have been assigned to clusters.

Example by STEM

48/57

- Data: immune response data from Guillemin et al. (2000, *PNAS*)
- Use human cDNA microarray to study the gene expression profile of gastric AGS cells infected with various strains of *Helicobacter pylori*.
 - *H. pylori* is one of the most abundant human pathogenic bacteria.
 - Cy3 (for the reference), Cy5 (for the experimental sample)
- Analyze data from the response of the wild-type G27 strain.



The screenshot shows the PNAS website interface. At the top, there is a navigation bar with links for 'Info for Authors', 'Editorial Board', 'About', 'Subscribe', 'Advertise', 'Contact', and 'Site Map'. The PNAS logo is prominently displayed on the right. Below the navigation bar, there are buttons for 'Current Issue', 'Archives', and 'Online Submission', along with a search box containing 'GO' and 'advanced search >>'. The main content area includes the text 'Institution: Life Science Library, Academia Sinica' and a 'Sign In as Member / Individual' link. The article title is 'Cag pathogenicity island-specific responses of gastric epithelial cells to *Helicobacter pylori* infection' by Karen Guillemin, Nina R. Salama, Lucy S. Tompkins, and Stanley Falkow. The publication date is 'Published online before print October 31, 2002, 10.1073/pnas.182558799' and the journal information is 'PNAS | November 12, 2002 | vol. 99 | no. 23 | 15136-15141'. The article is categorized under 'Microbiology'. The authors' affiliations are listed as 'Departments of *Microbiology and Immunology, and §Medicine, Stanford University School of Medicine, Stanford, CA 94305'. The article was contributed by Stanley Falkow and approved on September 13, 2002.

- Two replicates on the same biological sample in which time series data were collected at 5 time points: 0, 0.5, 3, 6, 12 hours.
- Select 2243 genes from 24192 array probes.
- Set $m=50$ model profiles and $c=2$.

STEM Interface

STEM: Short Time-series Expression Miner

1. Expression Data Info:
 Data File: g27_1.txt
 Log normalize data Normalize data No normalization/add 0
 Spot IDs included in the data file

2. Gene Annotation Info:
 Gene Annotation Source: Human (EBI)
 Cross Reference Source: Human (EBI)
 Gene Annotation File: gene_association.goa_human.gz
 Cross Reference File: human.xrefs.gz
 Download the latest: Annotations Cross References Ontology

3. Options:
 Clustering Method: STEM Clustering Method
 Maximum Number of Model Profiles: 50
 Maximum Unit Change in Model Profiles between Time Points: 2

4. Execute:

© 2004, Carnegie Mellon University. All Rights Reserved

	A	B	C	D	E	F	G
1	SPOT	Gene Symbol	0h	0.5h	3h	6h	12h
2	1	ZFX	-0.027	0.158	0.169	0.193	-0.165
3	2	ZNF133	0.183	-0.068	-0.134	-0.252	0.177
4	3	USP2	-0.67	-0.709	-0.347	-0.779	-0.403
5	4	DSCR1L1	-0.923	-0.51	-0.718	-0.512	-0.668
6	5	WNT5A	-0.471	-0.264	-0.269	-0.154	-0.254
7	6	VHL	-0.327	-0.378	-0.229	-0.264	-0.072
8	7	TCF3	-0.021	0.129	-0.209	-0.245	0.036
9	8	TCN2	-0.492	-0.41	-0.306	-0.494	-0.273
10	9	TIMP1	-0.111	0.351	0.168	0.129	-0.293
11	10	SERPINA7	-0.468	-0.488	-0.199	-0.144	-0.185
12	11	THBD	-1.013	-0.895	-0.743	-0.601	-0.543
13	12	EPHA2	0.13	0.313	0.645	-0.155	0.28

Advanced Options

Filtering | Model Profiles | Clustering Profiles | Gene Annotations | GO Analysis

Maximum Number of Missing Values: 0
 Minimum Correlation between Repeats: 0
 Minimum Absolute Expression Change: 0.8

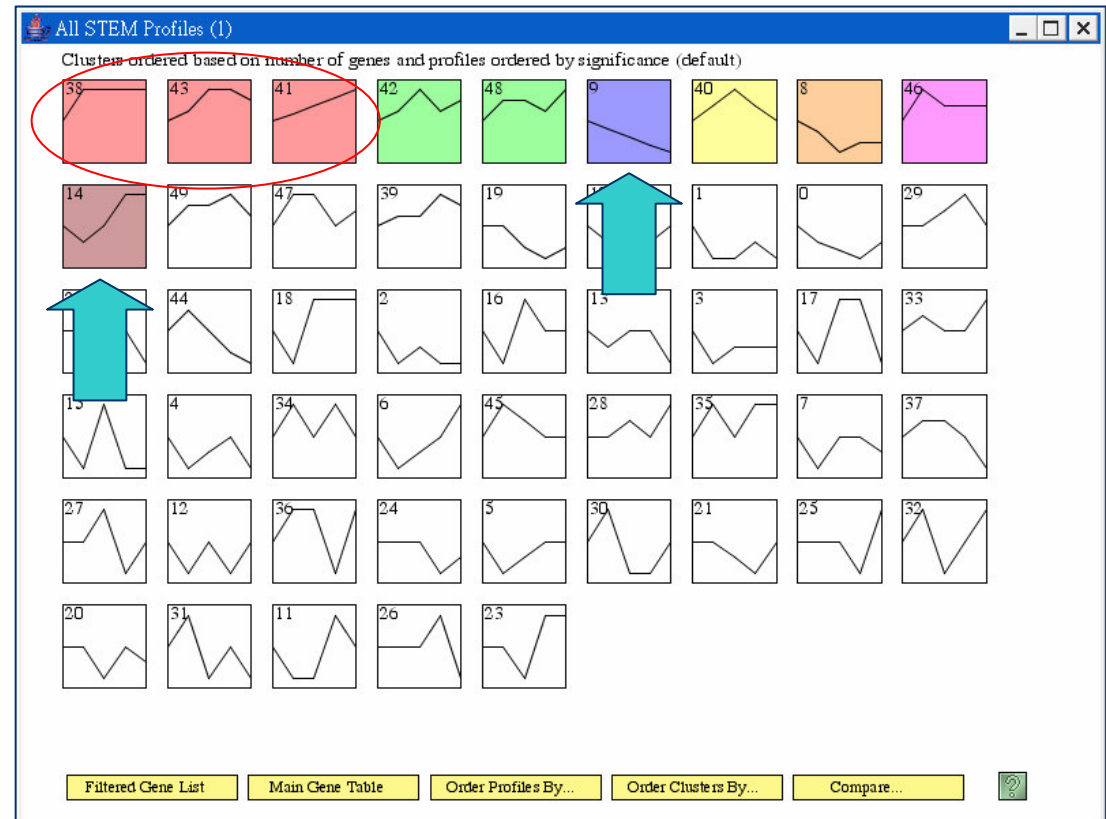
Change should be based on: Maximum - Minimum Difference from 0

Pre-filtered Gene File:

Example: Clustering Results

50/57

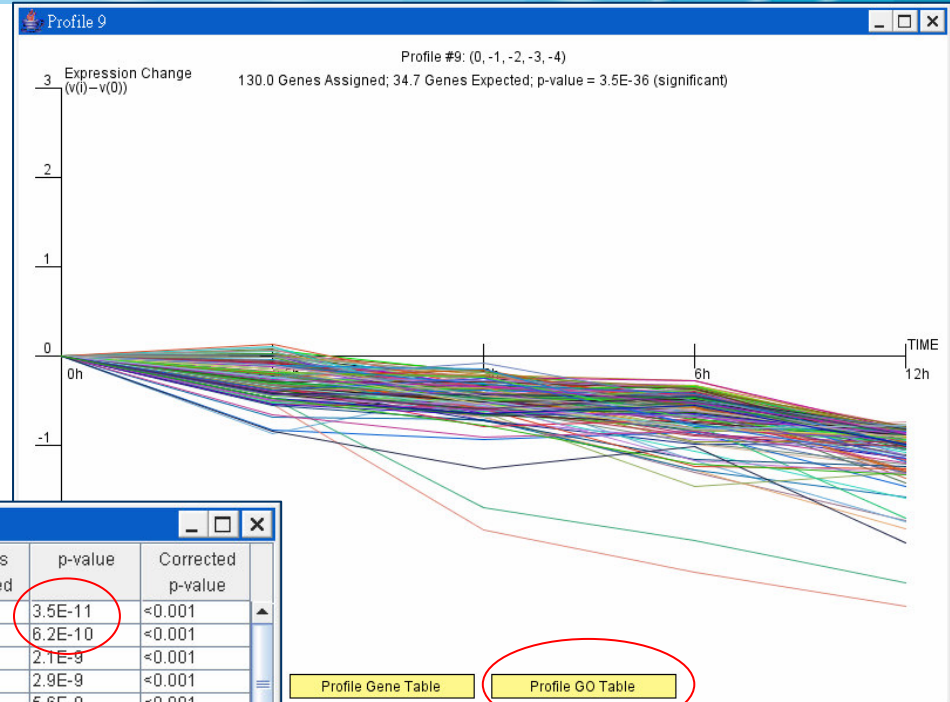
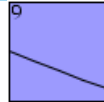
- Colored profiles are significant.
- Profiles with the same shade belong to the same cluster.
- $\text{Corr}=0.7 \rightarrow \delta = 0.3$ in grouping method.
- one: 3 profiles,
one: 2 profiles,
five: single profiles.



Four of the 10 significant model profiles were significantly enriched for GO categories.

Example: GO Interpretation

- Profile 9 (0, -1, -2, -3, -4): 131 down-regulated genes during the entire experiment duration.
- This profile was significantly enriched for cell-cycle genes ($p\text{-value} < 10^{-10}$).



GO Results for Profile 9 based on the actual number of genes assigned to the profile

Category ID	Category Name	#Genes Category	#Genes Assigned	#Genes Expected	#Genes Enriched	p-value	Corrected p-value
GO:0007049	cell cycle	432	19.0	2.7	+16.3	3.5E-11	<0.001
GO:0006259	DNA metabolism	344	16.0	2.2	+13.8	6.2E-10	<0.001
GO:0006260	DNA replication	110	10.0	0.7	+9.3	2.1E-9	<0.001
GO:0006139	nucleobase, nucleoside, nucleotide and nuc...	1490	31.0	9.5	+21.5	2.9E-9	<0.001
GO:0000074	regulation of progression through cell cycle	293	14.0	1.9	+12.1	5.6E-9	<0.001
GO:0051726	regulation of cell cycle	294	14.0	1.9	+12.1	5.8E-9	<0.001
GO:0006261	DNA-dependent DNA replication	49	7.0	0.3	+6.7	2.5E-8	<0.001
GO:0005634	nucleus	1667	31.0	10.6	+20.4	3.9E-8	<0.001
GO:0044238	primary metabolism	3112	43.0	19.8	+23.2	2.8E-7	<0.001
GO:0006281	DNA repair	141	9.0	0.9	+8.1	3.0E-7	<0.001
GO:0043283	biopolymer metabolism	1295	25.0	8.2	+16.8	5.3E-7	<0.001
GO:0006974	response to DNA damage stimulus	159	9.0	1.0	+8.0	8.2E-7	<0.001
GO:0044237	cellular metabolism	3175	42.0	20.2	+21.8	1.4E-6	<0.001
GO:0009719	response to endogenous stimulus	170	9.0	1.1	+7.9	1.4E-6	<0.001
GO:0050875	cellular physiological process	4335	51.0	27.6	+23.4	2.1E-6	<0.001
GO:0008152	metabolism	3379	43.0	21.5	+21.5	2.7E-6	<0.001
GO:0043231	intracellular membrane-bound organelle	2476	35.0	15.7	+19.3	3.3E-6	<0.001
GO:0043227	membrane-bound organelle	2477	35.0	15.7	+19.3	3.3E-6	<0.001
GO:0044424	intracellular part	3287	42.0	20.9	+21.1	3.4E-6	<0.001
GO:0005622	intracellular	3450	43.0	21.9	+21.1	4.7E-6	<0.001
GO:0043229	intracellular organelle	2840	37.0	18.1	+18.9	1.1E-5	0.002
GO:0048015	phosphoinositide-mediated signaling	48	5.0	0.3	+4.7	1.3E-5	0.002
GO:0044464	cell part	4598	50.0	29.2	+20.8	2.8E-5	0.010
GO:0051301	cell division	97	6.0	0.6	+5.4	3.6E-5	0.012
GO:0016779	nucleotidyltransferase activity	61	5.0	0.4	+4.6	4.3E-5	0.012

Click for GO Results Based on the Profile's Expected Size Save Table

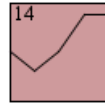
Profile GO Table

Many of the cycling genes in this profile are known transcription factors, which could contribute to repression of cell-cycle genes, and ultimately, the cell cycle.

Example: GO Interpretation

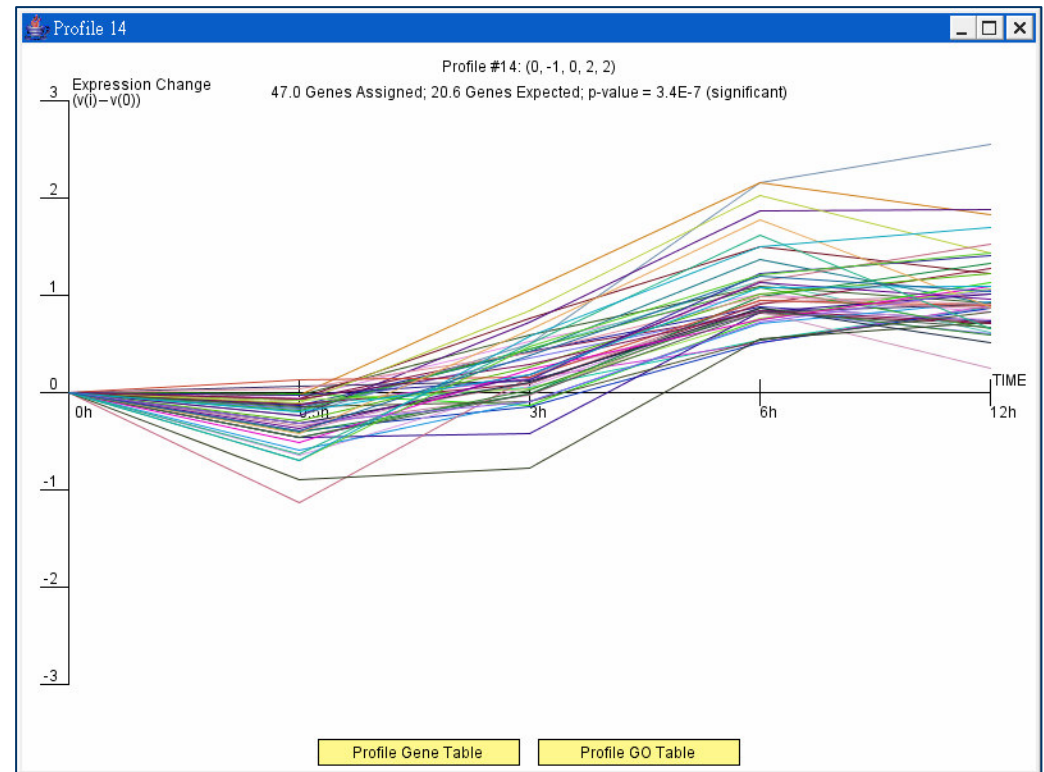
52/57

- Profile 14 (0, -1, 0, 2, 2) contained 49 genes.



- GO analysis indicates that many of these genes were relevant to **cell structure** and annotated as belonging to the categories

- cytoskeleton ($p=9 \times 10^{-5}$),
- extracellular matrix (9×10^{-4}),
- membrane (2×10^{-6}).



Profile GO Table

Structural elongation of cells is a known **phenotypical response to pathogens**, and thus the enrichment of such genes in up-regulated expression profiles is consistent with this biological response.

STEM: Other Functionalities

53/57

■ Bidirectional Integration

- Determine for a given **model profile** what **GO terms** are significantly enriched.
- Determine for a given **GO category** what **model profiles** were most enriched for genes in that category.

■ Comparing Data Sets

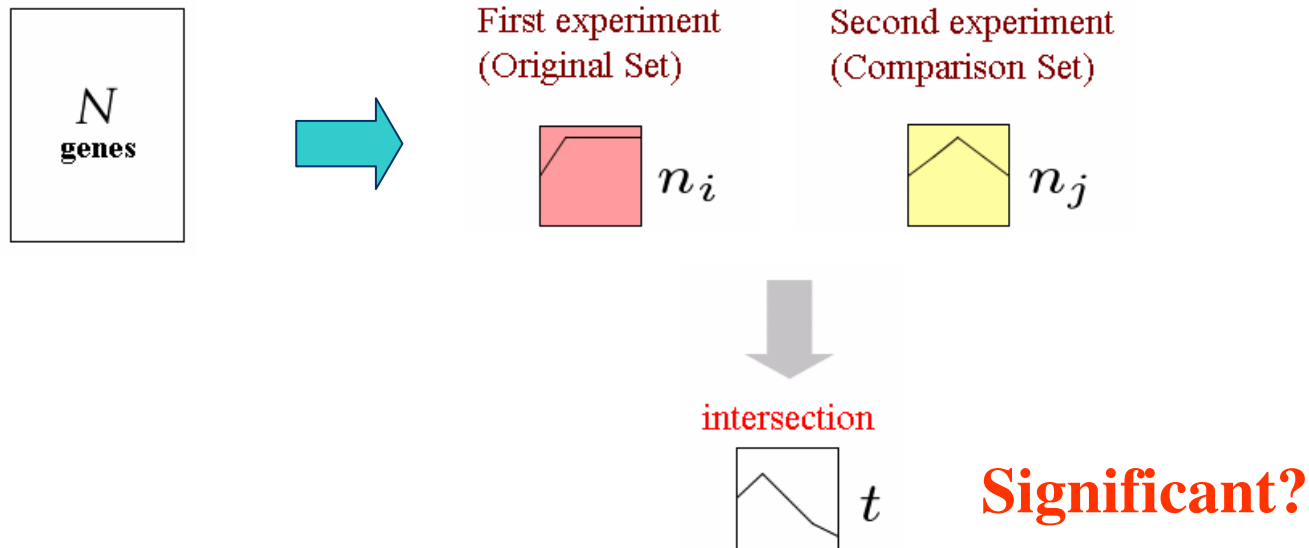
- For a set of genes which had temporal response X in experiment A, which significant responses did they have in experiment B?
- use **hypergeometric distribution** to compute the significance of overlap between gene sets of model profiles of two experiments.

Example

- Compare the temporal response of gene infected with a **wildtype pathogen** to those infected with a knockout mutant version of the pathogen (Guillemin, PNAS, 2002).
- The response of genes when exposed to a certain chemical substance to their response when not exposed. (Jorgensen et al., *Cell Cycle*, 2004)

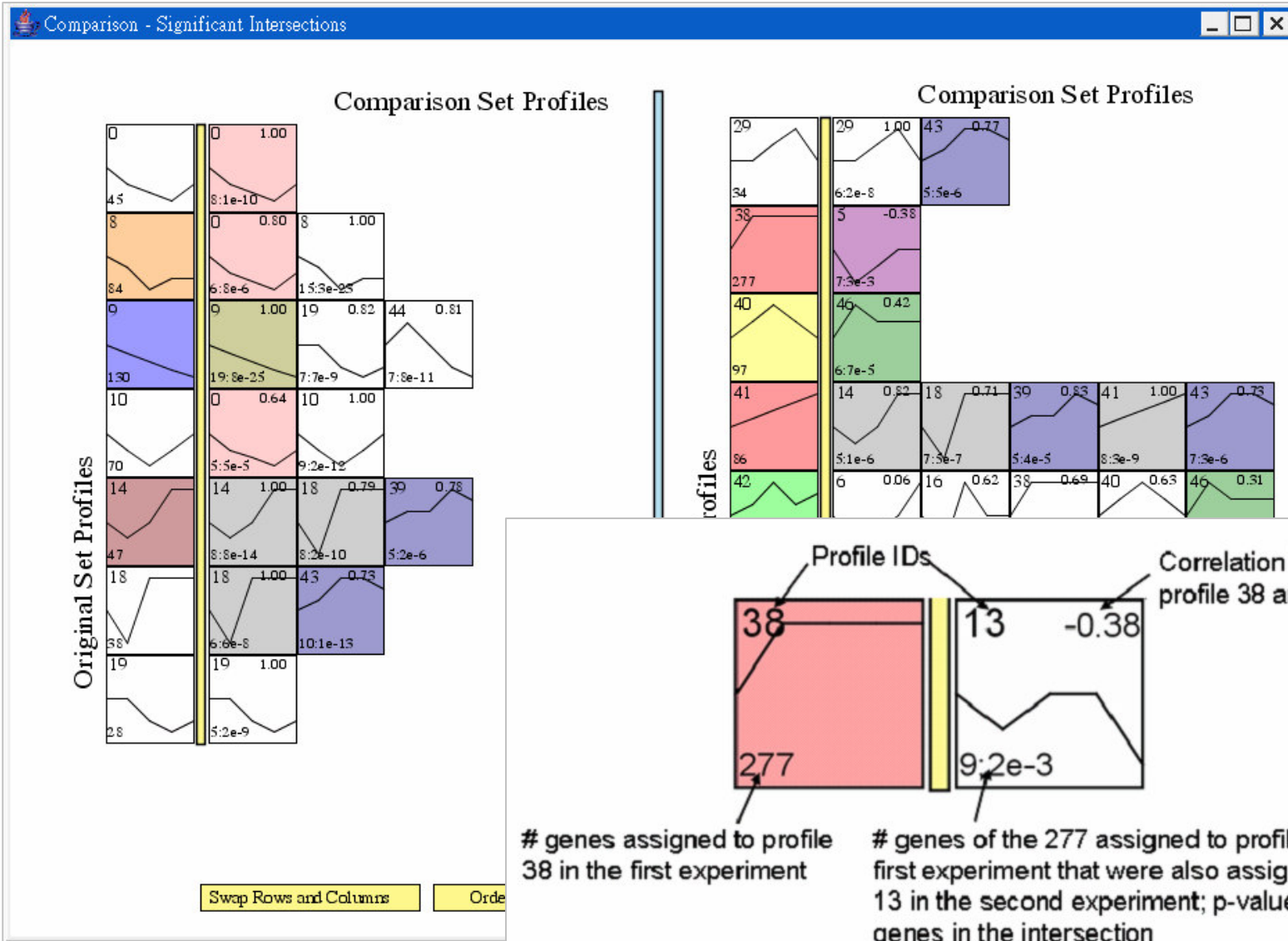
STEM: Comparing Data Sets

54/57



$$\begin{aligned} & p\text{-value} \\ &= P(\text{seeing } t \text{ or more genes}) \\ &= \sum_{m=t}^{\min(n_i, n_j)} \frac{\binom{n_j}{m} \binom{N-n_j}{n_i-m}}{\binom{N}{n_i}} \end{aligned}$$

STEM: Comparing Data Sets (conti.)



- Time Series Microarray Experiments
- Overview of Analyzing Software

- Some Issues

- P-values
- Multiple Hypothesis Testing
- Permutation Test
- Gene Set Enrichment Analysis

- SAM: Significance Analysis of Microarrays

- Algorithm
- Interpretation

Differential Expressed Genes

- STEM: Short Time-Series Expression Miner

- Algorithm
- Example

Clustering

Questions?

57/57

Reference: <http://idv.sinica.edu.tw/hmwu/SMDA/TimeCourse/index.htm>

Hank's Talks: Statistical Microarray Data Analysis - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

← 上一頁 → 搜尋 我的最愛 媒體

網址(D) <http://idv.sinica.edu.tw/hmwu/CourseSMDA/index.htm> 連結 »

Google 開始 4461 已擱載 設定

Welcome To Hank's Homepage!

Home | Experience | Research | Publication | Course | Talks | Software | Links | Updated 2007/04/09

Talks >> Statistical Microarray Data Analysis

2007

[2006/04/21] 3. Design and Analysis for Time Course Microarray Experiments
國防醫學院 生命科學所

[2006/04/12] 2. Microarray Data Analysis: [Finding Differential Expressed Genes](#) (57pages, 2.741MB)
Case Demo using affylmGUI: <http://bioinf.wehi.edu.au/affylmGUI/>
國立臺灣大學 資訊所, Course: [生物資訊之統計與計算方法](#)

[2007/03/29] 1. Microarray Data Analysis: Data Preprocessing for Affymetrix GeneChip Data
[\[Data Preprocessing for Affymetrix GeneChip Data\]](#) (69pages, 5.77MB)
[\[Gene Filtering, Missing Values Imputation\]](#) (13pages, 680KB)
國立臺灣大學 資訊所, Course: [生物資訊之統計與計算方法](#)

2006

[2006/11/07] 6. Microarray Data Analysis

[2006/07/19] 5. Microarray Data Analysis (ii): Clustering & visualization
[上課講義] [軟體下載: GAP | Tutorial] [參考文獻] [Homework, Data]
[Homework 參考資料: Plots_Rcode, PCA&MDS_Rcode, Clustering_Rcode, testdata, link1]
國立陽明大學生物資訊研究所, 95學年度暑期「生物資訊與系統生物學學分班」[上課講義]
Course: [系統生物學實驗](#)

[2006/05/25] 3. Microarray Data Analysis:

完成 網際網路

Thank You!

吳漢銘

hmwu@stat.sinica.edu.tw
<http://idv.sinica.edu.tw/hmwu>



中央研究院 統計科學研究所
Institute of Statistical Science, Academia Sinica