Statistical

Microarray Data Analysis

Analysis for Time Course Microarray Experiments

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」
Course: 系統生物學實驗

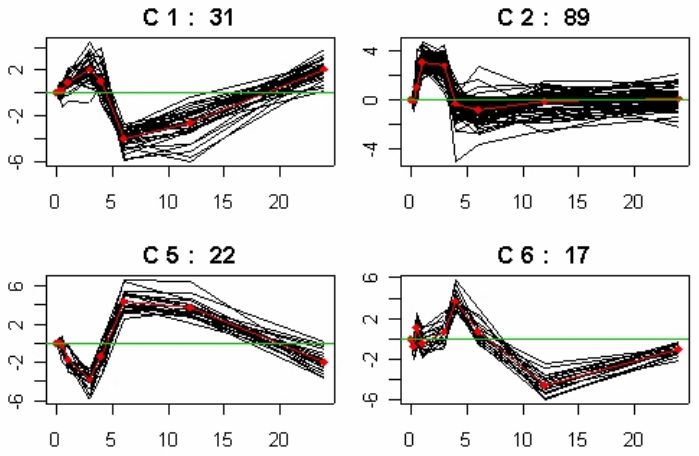
2006年7月21日

吳漢銘
hmwu@stat.sinica.edu.tw
http://www.sinica.edu.tw/~hmwu

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

Outlines

- Overview of Analyzing Software
- Analysis of Short Time-series Expression: STEM (Short Time-series Expression Miner)
- Gene Ontology (GO) and Gene Set Enrichment Analysis
- P-values and Multiple Hypothesis Testing
- STEM
  - Identifying Significant Model Profiles
  - (Permutation Test)
  - Grouping Significant Profiles
- Example
- Other functionalities
- Software Practice



Design Examples

(1)

tair

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

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

Gene Search

Experiment: AtGenExpress: Heat stress time course

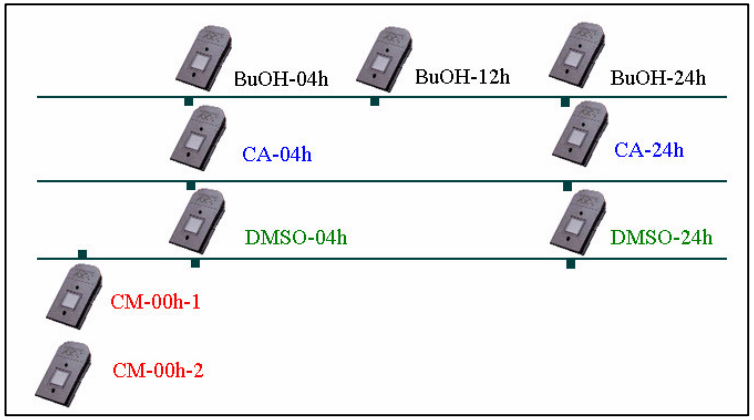
Experiment Summary | Samples | Slides & Datasets | Array Design | View All

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 Environment
Experiment Category	abiotic treatment , tissue comparison
Experiment Goals	response to heat
Description	This experiment studies the effect of heat stress on the expression of genes in shoot and root tissue of 16 days old Arabidopsis thaliana seedlings of Columbia-0 ecotype and in cell culture at different time points. We collected two samples from each time point, which are serving as replicates. Samples were taken in duplicate in a time course during heat stress and recovery (0.25h, 0.5h, 1h, 3h of heat stress and after 1h, 3h, 9h, 21h recovery after 3h heat stress). Contr non-treated cells (expression data of the ATH1 chip fron which is funded by
Data Counts	Number of Slides 94

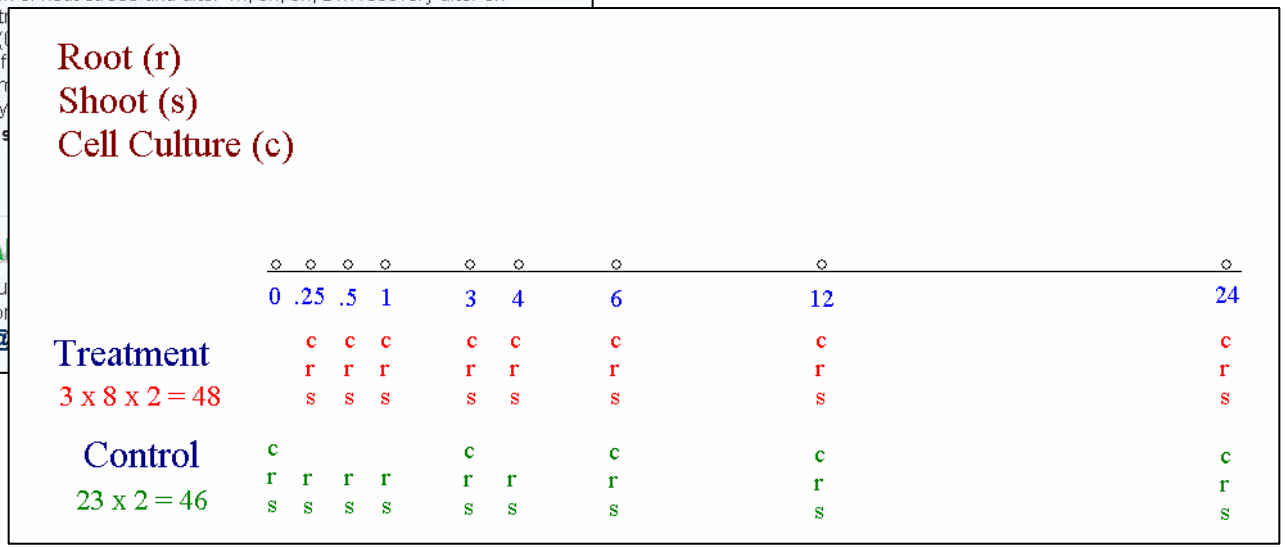
NCGR

General comments or qu
Seed or DNA stock questio
abrc@

(2)



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



Time Series Microarray Experiments

- Study dynamic biological process
  - Cell cycle (Spellman et al., 1998, *Mol Bio Cell*)
  - Developmental studies (Arbeitman et al., 2002, *Science*)
  - Immune response (Guillemin et al., 2002, *PNAS*)
- About 80 % of microarray time series experiments are short (3-8 time points)
  - Cost of microarray
  - limited availability of biological material:

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

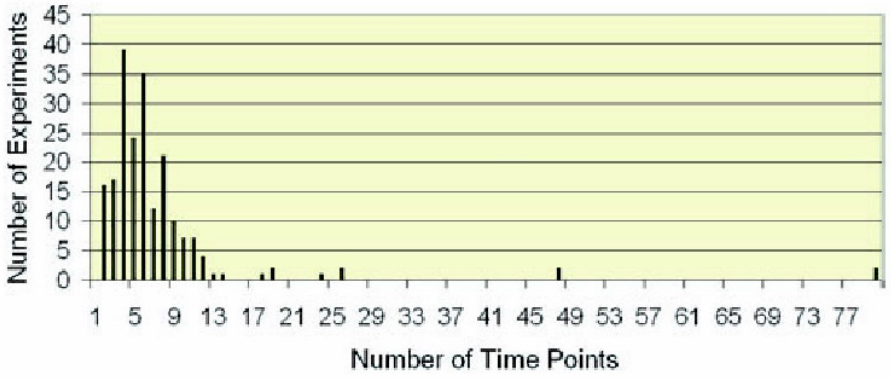
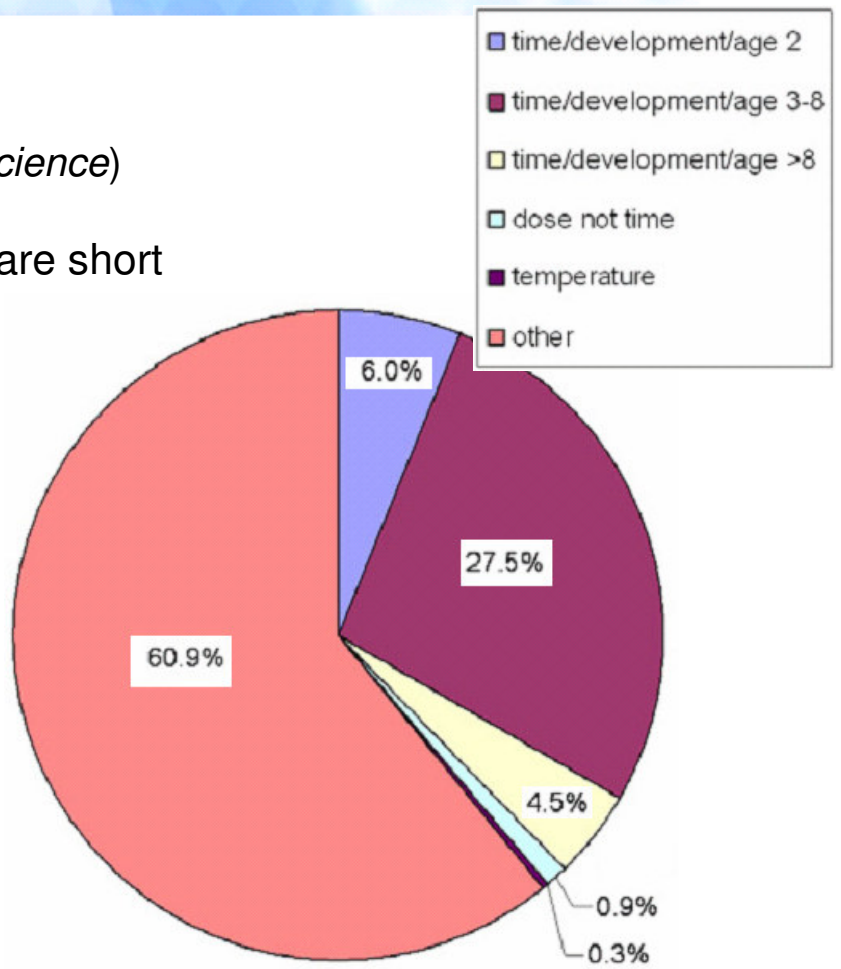


Fig. 1. Distribution of lengths of times series in the SMD as of June 2004.

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



Distribution of Microarray Data Sets in the Gene Expression Omnibus

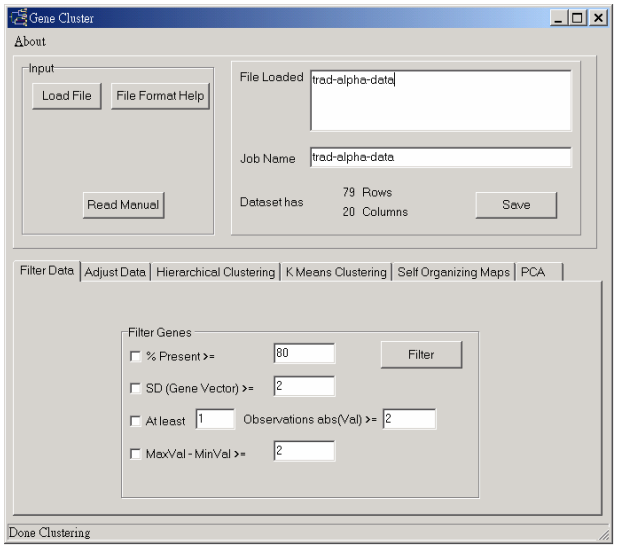
Distribution of microarray experiments by type. Summary of the 786 microarray datasets for human, mouse, rat, and yeast in the Gene Expression Omnibus as of August 2005.

Source: Ernst et al., 2005, Bioinformatics

Analyzing Software (General)

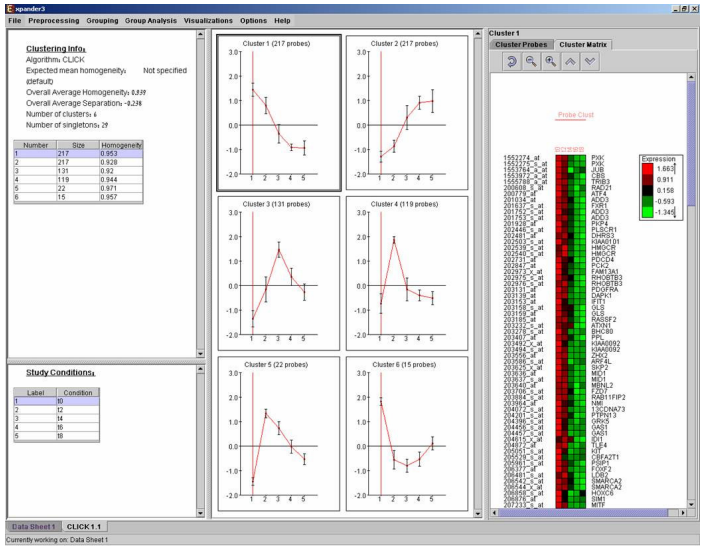
- Do not take advantage of the sequential information in time series data
- Popular clustering: hierarchical clustering, kmeans clustering, self-organizing maps. (ignore the temporal dependebcy among successive time points.)(random permute the order of time points, the results would not change)

Cluster
Eisen et al., 1998, *PNAS*

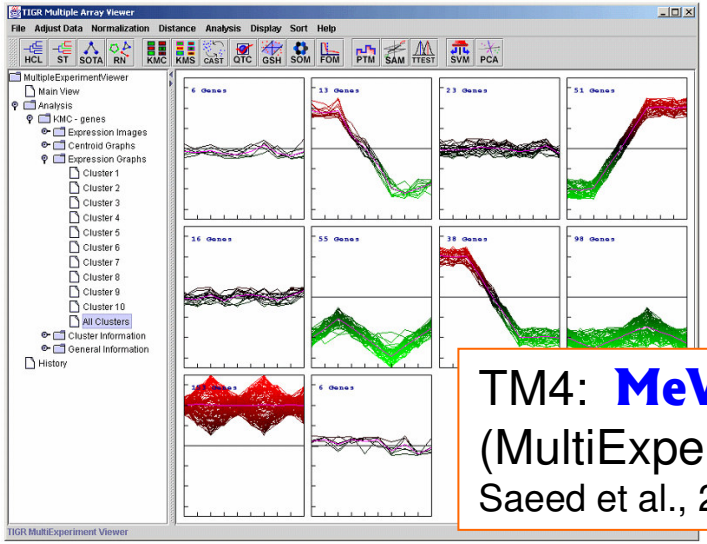


http://rana.lbl.gov/EisenSoftware.htm

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



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



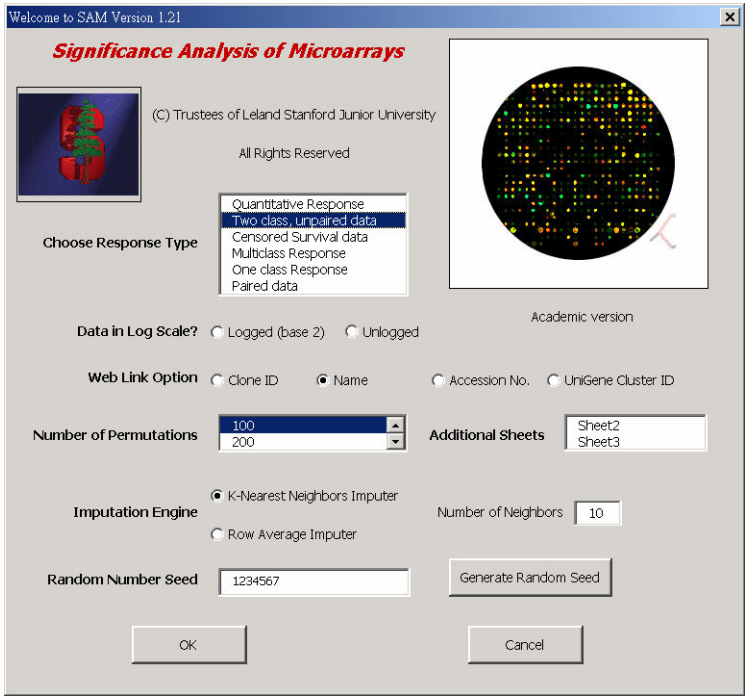
TM4: **MeV**
(MultiExperiment Viewer)
Saeed et al., 2003, *Biotechniques*

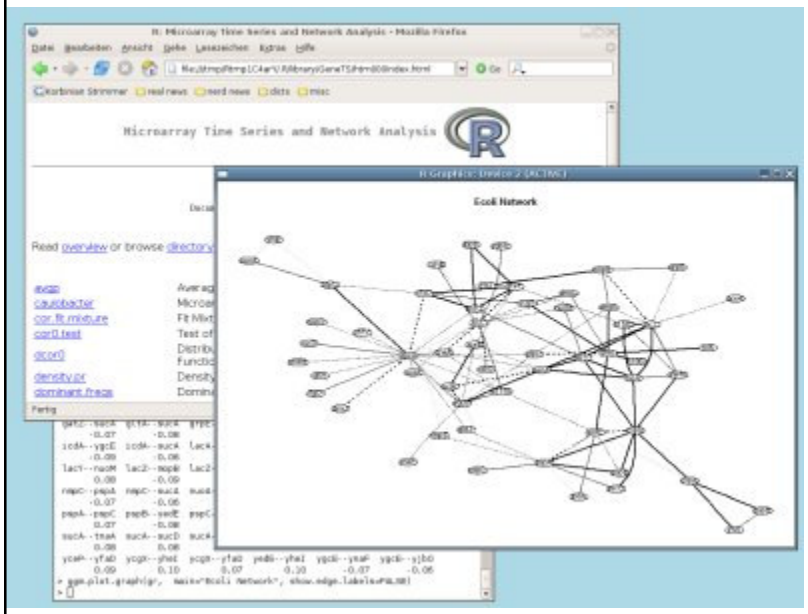
http://www.tm4.org/mev.html

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

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

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





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

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

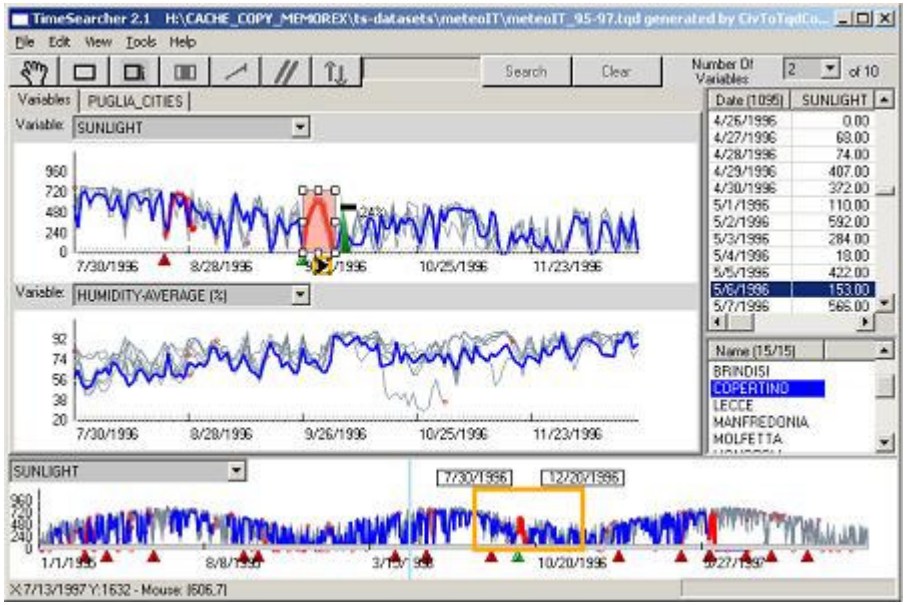
EDGE:
Extraction of Differential Gene Expression
(Leek et al., 2006, *Bioinformatics*)

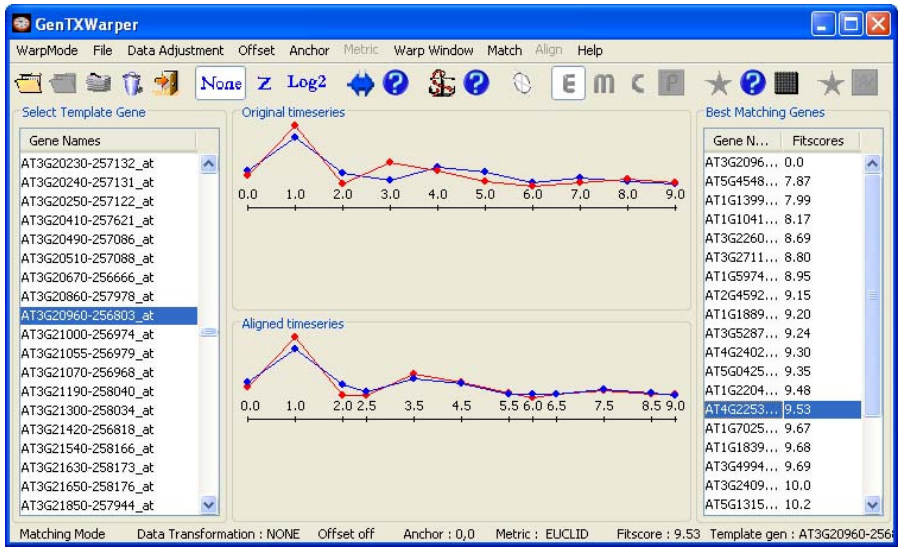
http://www.biostat.washington.edu/software/jstorey/edge/

Software Designed for Time Series Gene Expression Data (Visualization)

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

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





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

GenT χ Warper:
Mining of gene expression time series with dynamic time warping techniques
(Criel and Tsiporkova, 2005, *Bioinformatics*)

ORIOGEN:
Order Restricted Inference for Ordered Gene ExpressioN clustering for time series.
(Peddada et al., 2005, *Bioinformatics*)

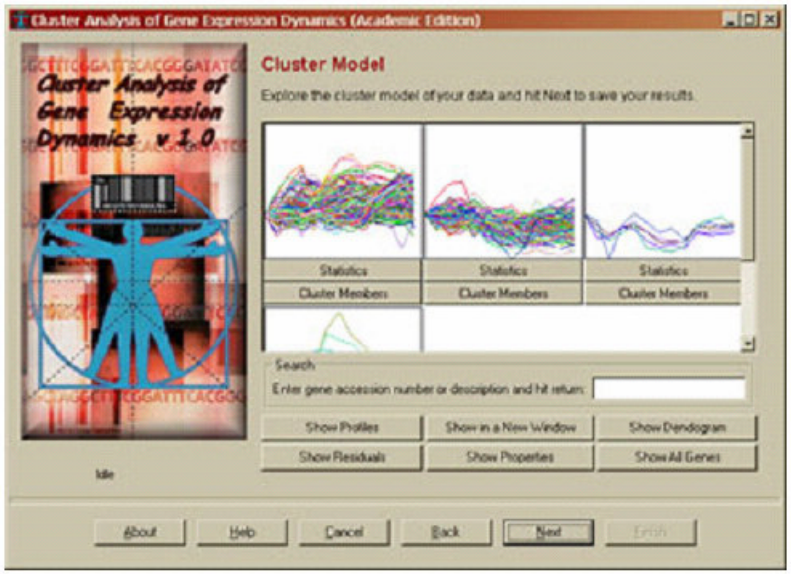
http://dir.niehs.nih.gov/dirbb/oriogen/index.cfm

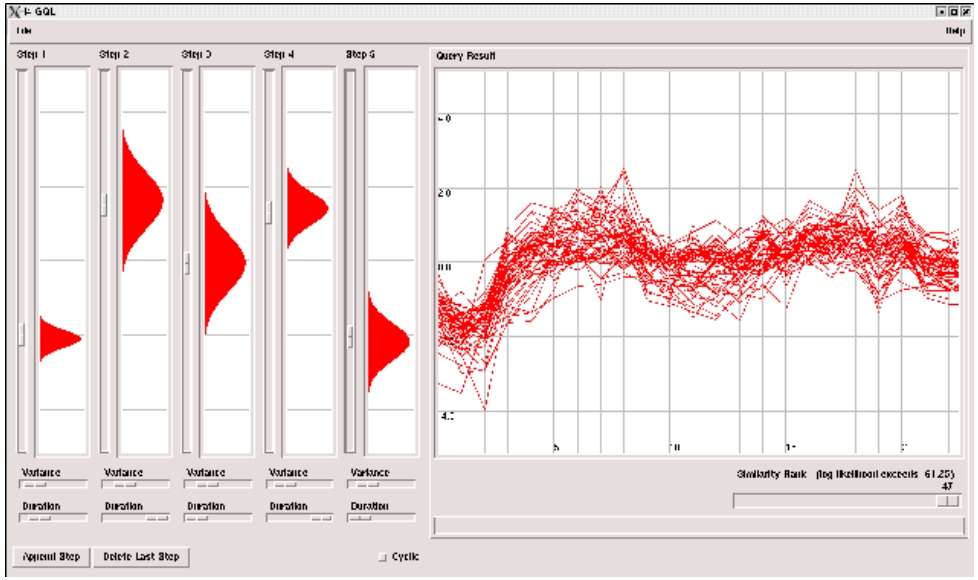
Software Designed for Time Series Gene Expression Data (Visualization and Clustering)

Gene expression time series analysis software primarily designed for longer time series.

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

http://genomethods.org/caged/



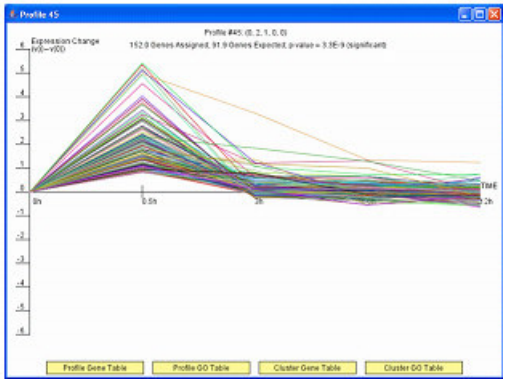


http://www.ghmm.org/gql

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

Software Designed for *Short* Time Series Gene Expression Data

- Unique challenges and opportunities inherent in short time series gene expression data.
  - 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:
Short Time-series Expression Miner
(Ernst and Bar-Joseph. 2006, *BMC Bioinformatics*.)

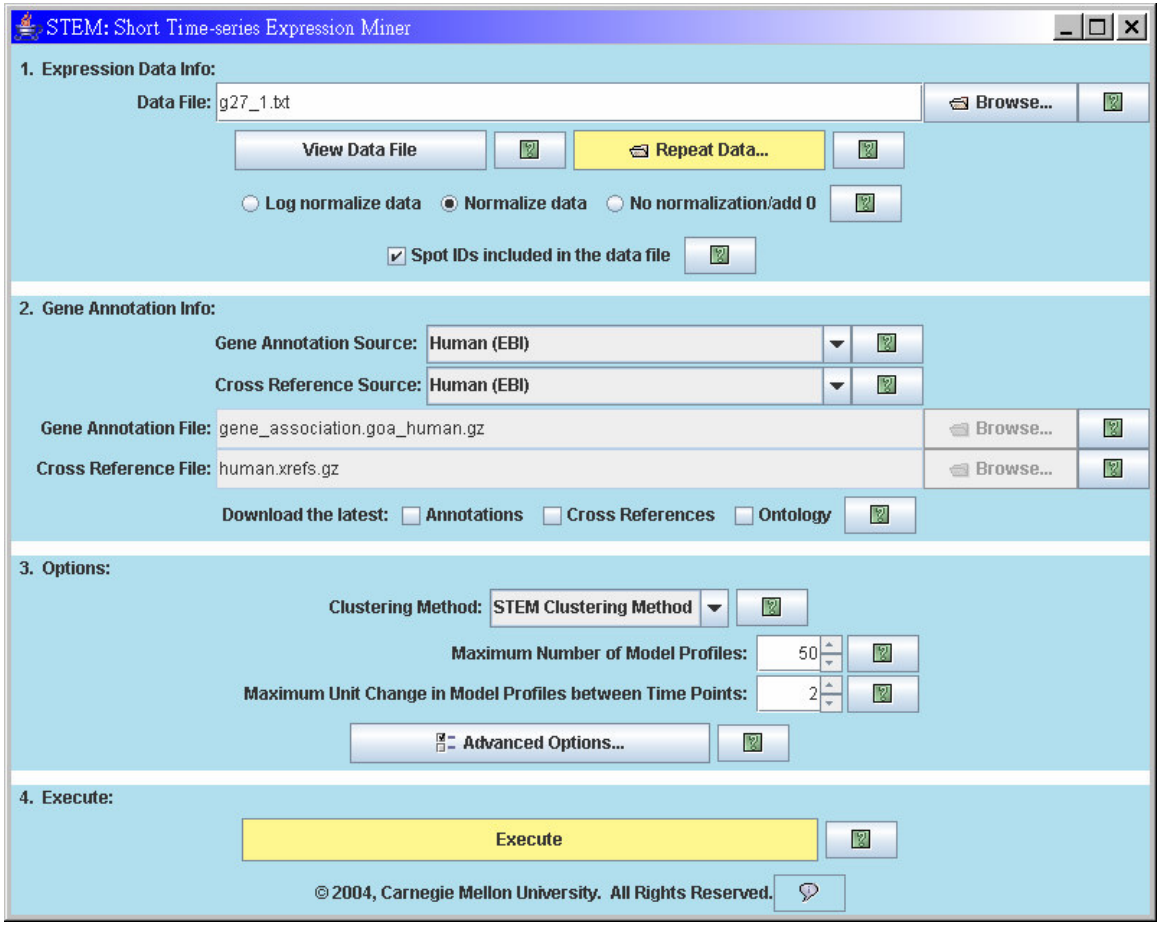
http://www.cs.cmu.edu/~jernst/stem/

Software: STEM

- Java-designed,
  - Used library: Piccolo toolkit (Bederson et al, 2004)
- Function: Clustering, Visualization and Comparison

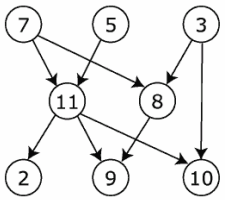
- Determine and visualize the behavior of genes belonging to a given GO category, identifying which temporal expression profiles were enriched for genes in that category.

- External: gene ontology and gene annotation from GO (http://www.geneontology.org) or EBI (http://www.ebi.ac.uk/GOA)

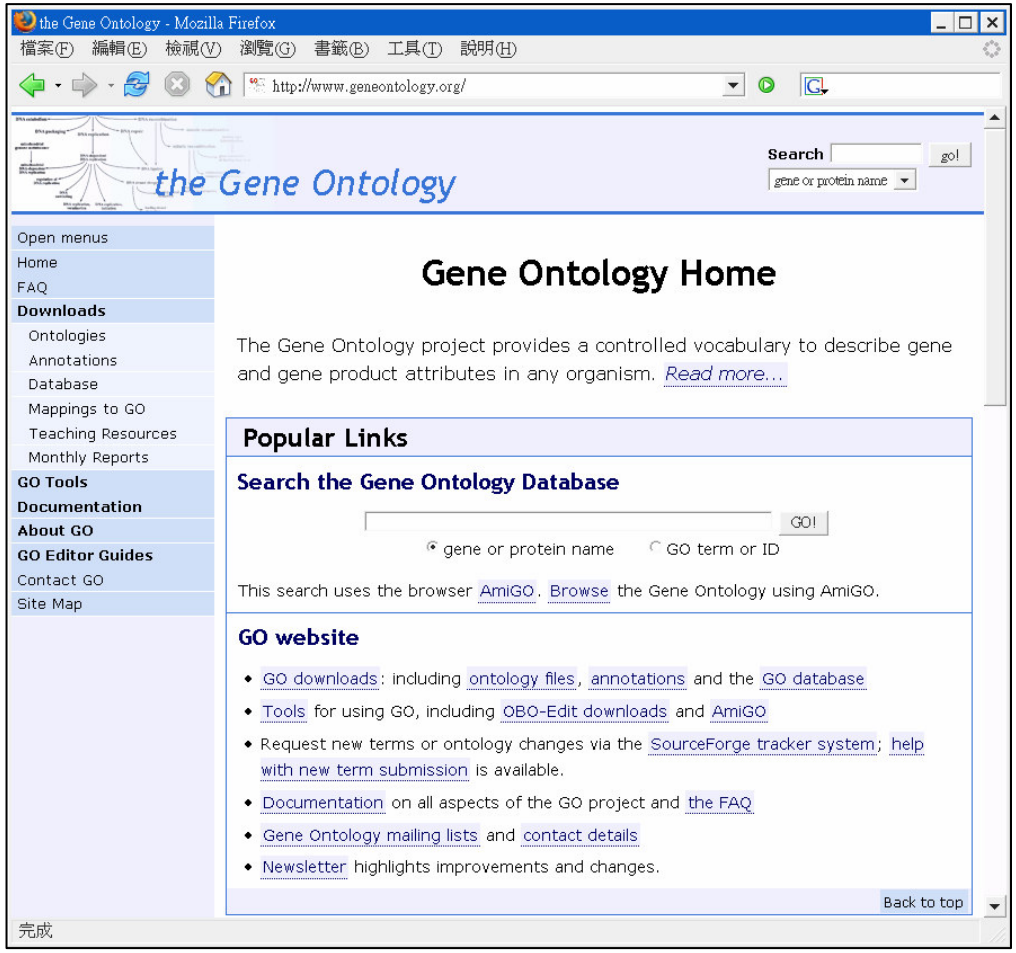


Gene Ontology (GO)

- The go is a structured vocabulary for describing biological processes, cellular components and molecular functions of gene products.
- The ontology is a hierarchy of terms organized as a directed acyclic graph.



- GO term annotations of gene products is available for many organisms.
- Official 15 column gene annotation format.

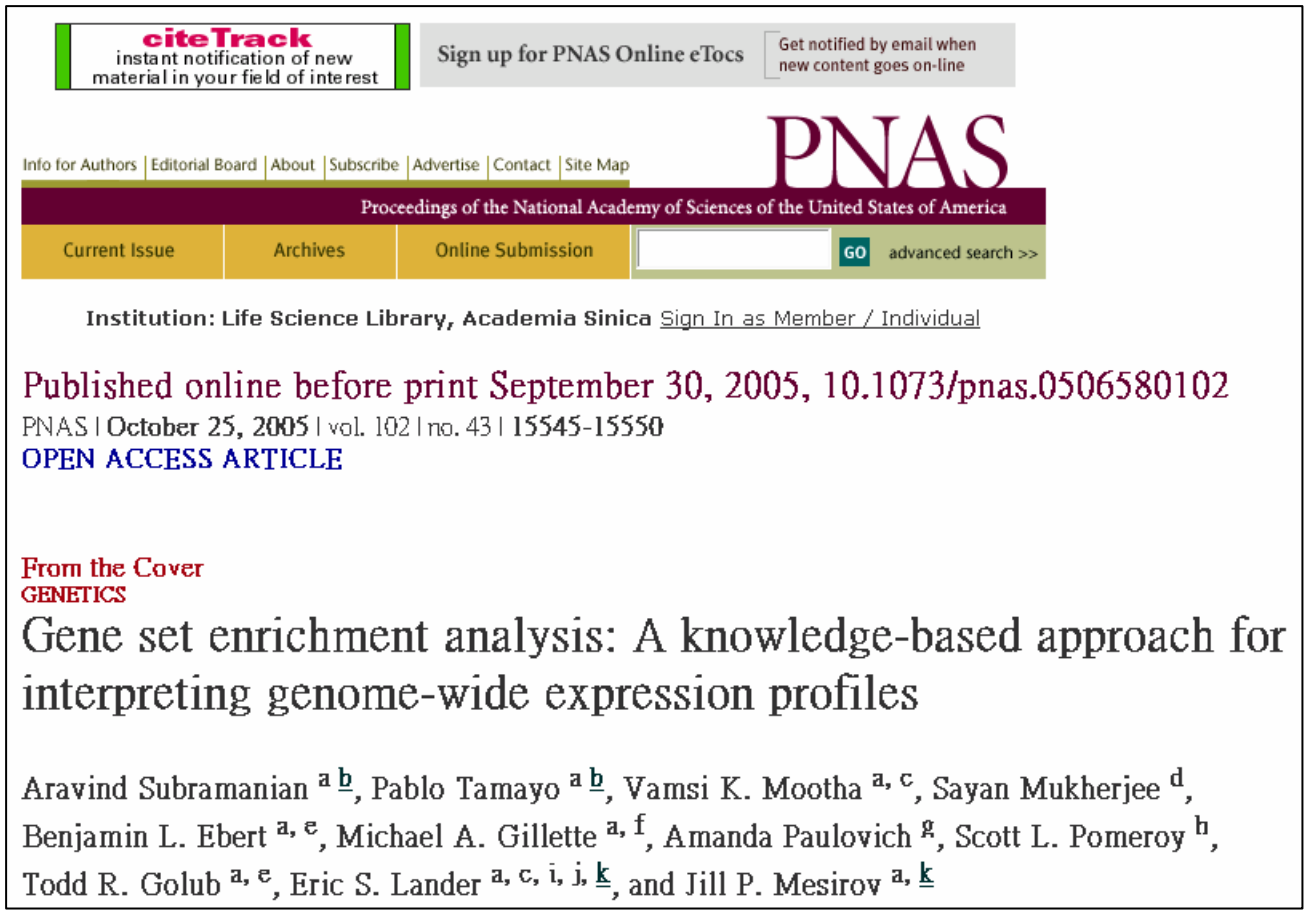


http://www.geneontology.org

A popular approach to gain biological insights from a set of identified genes of interest is to determine which GO terms annotations are overrepresented among the genes in the set.

Gene Set Enrichment Analysis

- 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).



Published online before print September 30, 2005, 10.1073/pnas.0506580102
PNAS | October 25, 2005 | vol. 102 | no. 43 | 15545-15550
OPEN ACCESS ARTICLE

From the Cover
GENETICS

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian [a b], Pablo Tamayo [a b], Vamsi K. Mootha [a, c], Sayan Mukherjee [d], Benjamin L. Ebert [a, e], Michael A. Gillette [a, f], Amanda Paulovich [g], Scott L. Pomeroy [h], Todd R. Golub [a, e], Eric S. Lander [a, c, i, j, k], and Jill P. Mesirov [a, k]

The *p*-values (for detecting DE genes)

H0: no differential expressed.
- The test is significant = Reject **H0**
- False Positive = (Reject **H0** | **H0** true)
  = concluding that a gene is differentially expressed when in fact it is not.

- *p* is the probability of observing your data under the assumption that the null hypothesis is true.
- *p* is the probability that you will be in error if you reject the null hypothesis.
- *p* represents the probability of **false positives** (Reject $\mathbf{H_0}$ | $\mathbf{H_0}$ true).

p=0.03 indicates that you would have only a 3% chance of drawing the sample being tested if the null hypothesis was actually true.

Decision Rule

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

P is *not* the probability that the null hypothesis is true!

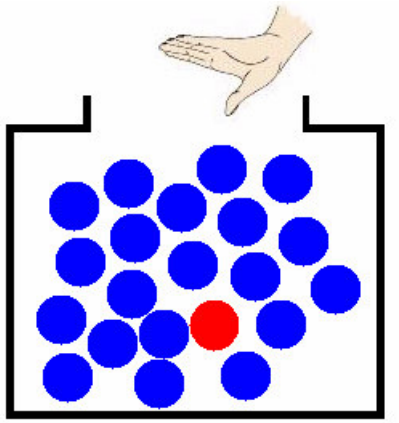
$$\text{Power} = 1 - \beta.$$

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	
		H0	H1
Decision	Reject H0	Type I Error (alpha) (false positive)	Right Decision (true positive)
	Don't Reject H0	Right Decision	Type II Error (beta)

Multiple Hypothesis Correction

- As many GO categories are being tested simultaneously, it is necessary to correct p-values using a multiple hypothesis correction.



Imagine a box with 20 marbles: 19 are blue and 1 is red.
What are the odds of randomly sampling the red marble by chance?
It is 1 out of 20.

Now let's say that you get to sample a single marble (and put it back into the box) 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.

X: false positive gene

P(X>=1)

= 1-P(X=0)

= 1- 0.95^n

Multiplicity of Testing

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)

- There is a serious consequence of performing statistical tests on many genes in parallel, which is known as multiplicity of p-values.
- Take a large supply of reference sample, label it with Cy3 and Cy5: no genes are differentially expressed: all measured differences in expression are experimental error.
  - By the very definition of a p-value, 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.
  - Because there are 10000 genes on this imaginary microarray, we would expect to find 100 significant genes at this level.
  - Similarly, we would expect to find 10 genes with a p-value less than 0.001, and 1 gene with p-value less than 0.0001
  - The p-value is the probability that a gene's expression level are different between the two groups due to chance.

Question:

1. How do we know that the genes that appear to be differentially expressed are truly differentially expressed and are not just artifact introduced because we are analyzing a large number of genes?
2. Is this gene truly differentially expressed, or could it be a false positive results?

Types of Error Control

- 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*

Type One Errors Rates

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

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

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

$$\text{FDR} = E[\frac{V}{R}] \quad \text{if } R > 0$$

Power = Reject the false null hypothesis

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

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

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

Multiple Testing Corrections

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).

- The more stringent a multiple testing correction, the less false positive genes are allowed.
- The trade-off of a stringent multiple testing correction is that the rate of *false negatives* (genes that are called non-significant when they are) is very high.
- FWER is the overall probability of false positive in all tests.
  - Very conservative
  - False positives not tolerated
- False discovery error rate allows a percentage of called genes to be false positives.

Bonferroni Correction

- 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 un-corrected 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

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

Benjamini and Hochberg FDR

- 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 correction becomes more stringent as the p-value decreases, similarly as the Bonferroni Step-down correction.
- This method provides a good alternative to Family-wise error rate methods.
- 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.

Corrected P-value= p-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 ➔ No
B	0.06	999	1000/999*0.06 = 0.06006	0.06006 > 0.05 ➔ No
C	0.04	998…	1000/998*0.04 = 0.04008	0.04008 < 0.05 ➔ Yes

Clustering Short Time Series Gene Expression Data (STEM)

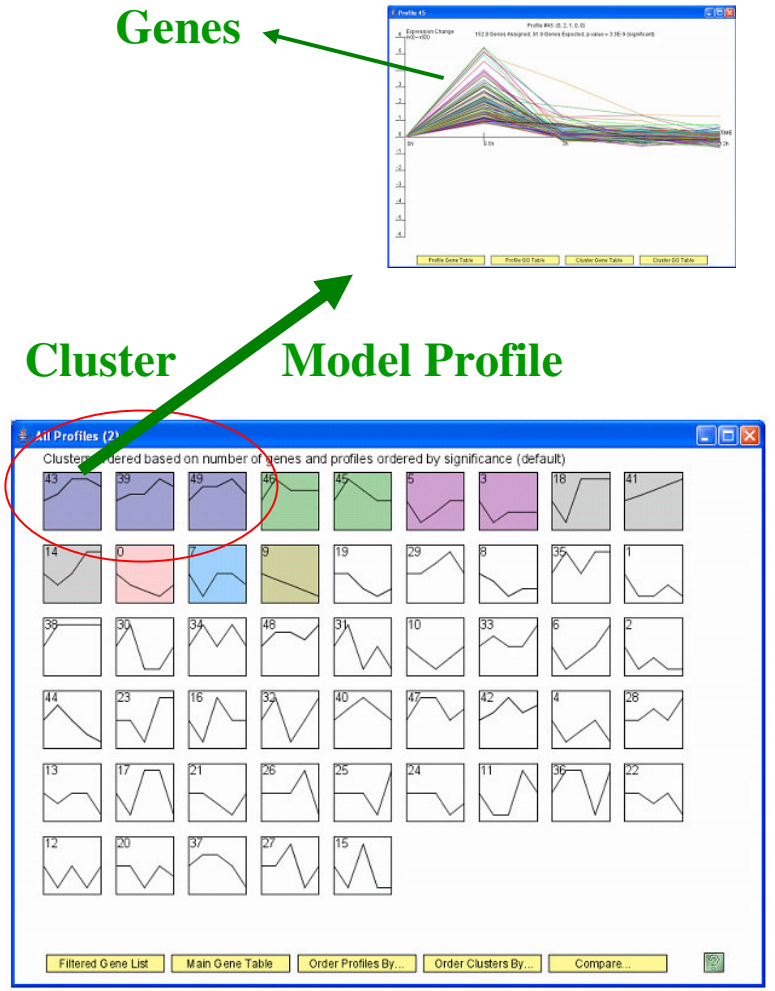
Purpose: Identifying Significant Expression Patterns

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.



Selecting Model Profiles (pre-defined)

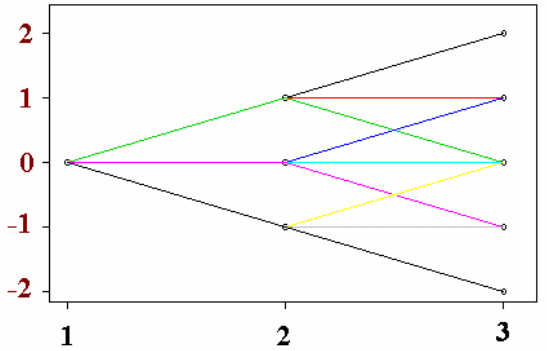
- Select a set of model expression profiles all of which are distinct from one another.
- Expression values (*log ratios*), where the ratios are with respect to the expression of the first time point.
- The first value always be 0.

- A parameter c: controls the amount of change a gene can exhibit between successive time points.
- 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, ➔ $(2c+1)^{n-1}$ 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) (see Ernst et al., 2005, *Bioinformatics*).

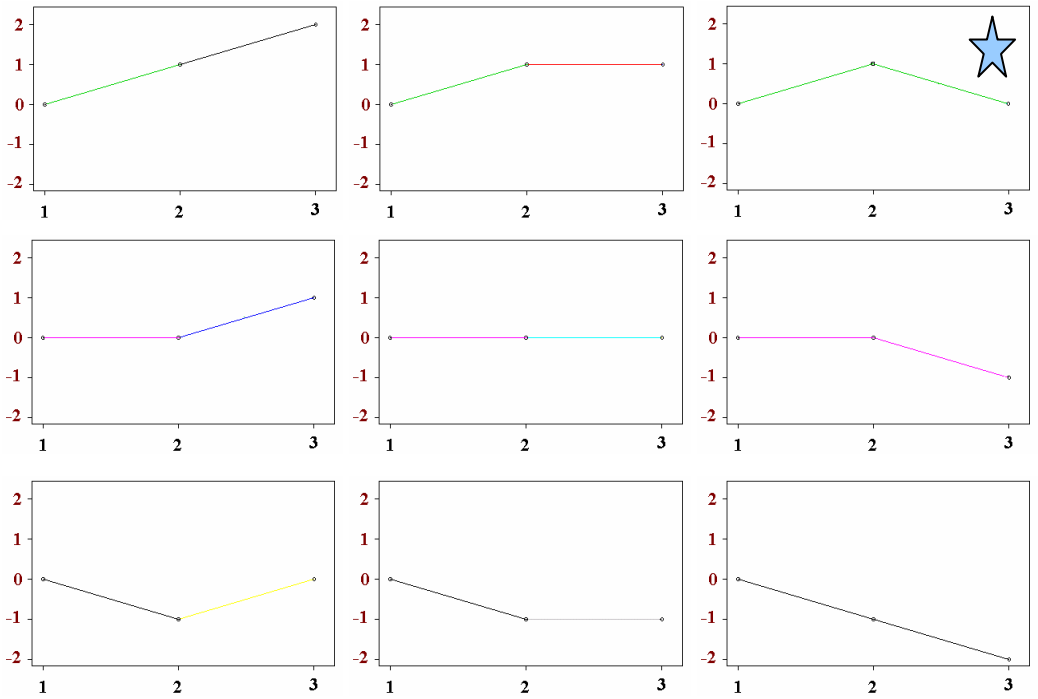
Example:

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



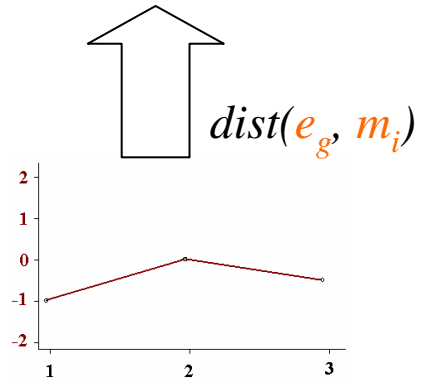
Assigning Genes to Model Profiles

- 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.

	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



Identifying Significant Model Profiles

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 results of *random fluctuation* in the measured values for genes assigned to that profile.
- **Permutation Test**: permutation is used to quantify the expected number of gene that would have been assigned to each profile if the data were generated at random.

Identifying Significant Model Profiles (conti.)

- Under the null hypothesis, the order of the observed values is random.
  - as each point is independent of any other point.
  - thus 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 $Si=\sum_j s_{ij}$, then $E_i=Si/n!$ is the expected number of genes for each profile model if the data were indeed generated according to the null hypothesis.
- **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 seening $T(m_i)$ genes assigned to profile m_i is $P(X>= T(m_i))$, where $X\sim Binomial(|G|, |E_i/G|)$.
- **Bonferroni Correction**: consider the number of genes assigned to m_i to statistically be significant if $P(X>= T(m_i)) < \alpha /m$.

The Permutation Test

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

- The process of randomizations makes it possible to ascribe a probability distribution to the difference in the outcome possible under null hypothesis.

$$p = \mathrm{P}(\mathrm{T} \geq \mathrm{t}_{\mathrm{obs}} \mid \mathrm{H}_0) \approx \frac{\#\{\mathrm{t}^* \geq \mathrm{t}_{\mathrm{obs}}\}}{\#\,\text{permutations}}$$

- The outcome data are analyzed many times (once for each acceptable assignment that

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

The Permutation Test (conti.)

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

Data

Gene1	Gene2
g_1^1	g_1^2
⋮	⋮
g_n^1	g_n^2

⬇

$g_{(1)}^1$	$g_{(1)}^2$
⋮	⋮
$g_{(n)}^1$	$g_{(n)}^2$

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
		⋮		⋮
0.8015	0.6765	2		1
		⋮		⋮
1.3726	1.2373	3		4
		⋮		⋮
1.1030	1.735	4		2
0.5148	1.0015	4		3

The permutation test allows determining the statistical significance of the score for every gene.

Correlation Coefficient and Distance

Cov	x1	x2	x3	x4		xp
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
xp	-0.28	-0.23	-0.05	0.10		1.00

Proximity Matrix

Data Matrix x y

Data	x1	x2	x3	x4	...	xp
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

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

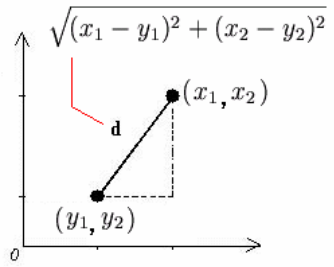
$$y = (y_1, y_2, \cdots, y_n)$$

Pearson Correlation Coefficient

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

Euclidean Distance

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



- The standard transformation from a similarity matrix C to a distance matrix D is given by $d_{rs} = (c_{rr} - 2c_{rs} + c_{ss})^{1/2}$.
- (Eisen *et al.* 1998) $d_{rs} = 1 - c_{rs}$
- Other transformations (Chatfield and Collins 1980, Section 10.2)

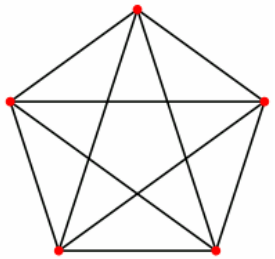
Correlation Coefficient

- **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.
- Use $d = 1 - r$:
  - still not a metric, does not satisfy the triangle inequality.
- Generalized version of the triangle inequality:
  - $g_m(x, z) <= 2(g_m(x, y)+g_m(y, z))$ ➔ a transitive measure.
  - When using the correlation coefficient two highly dissimilar profiles can't be very similar to a third profile.

Grouping Significant Profiles

Use Graph Theory

- **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 egde *iff* $dist(v_1,v_2) < \delta$.
- **Cliques** in this graph correspond to sets of significant profiles which are all similar to one another.



a clique of size 5

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

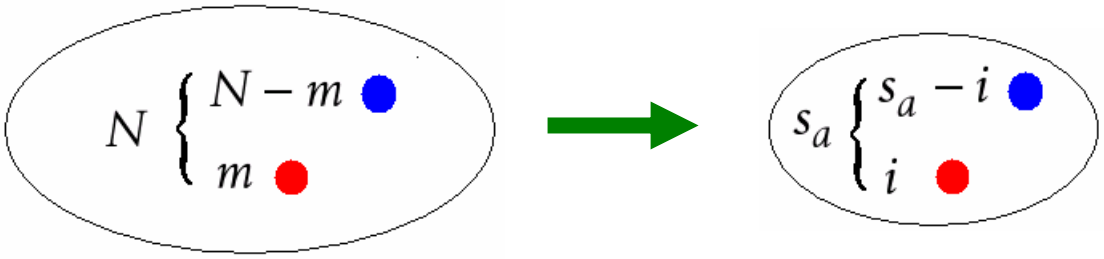
- Cluster for a significant profile $C_i = \{p_i\}$,
- Initial $C_i = \{p_i\}$, look for a profile p_j such that p_j is the closet profile to p_i that is not already included in C_i.
  - If $dist(p_j, p_k) <= \delta$delta for all profiles p_k in C_i, add p_j to C_i and repeat process,
  - otherwise stop 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.

Actual Size Gene Set Enrichments (STEM)

The enrichment is computed using the *hypergeometric distribution* (超幾何分配) based on the *actual* number of genes in the set of interest.



$$p(X=i;N,m,s_a)=\frac{\binom{m}{i}\binom{N-m}{s_a-i}}{\binom{N}{s_a}}$$

- N: the total number of unique genes on the microarray
- m: the total number of genes that are in the GO category of interest.
- S_a: the number of gene's assigned to profile ***r***.

The p-value of seeing v or more genes in the intersection of the category of interest and profile ***r*** can be computed as:

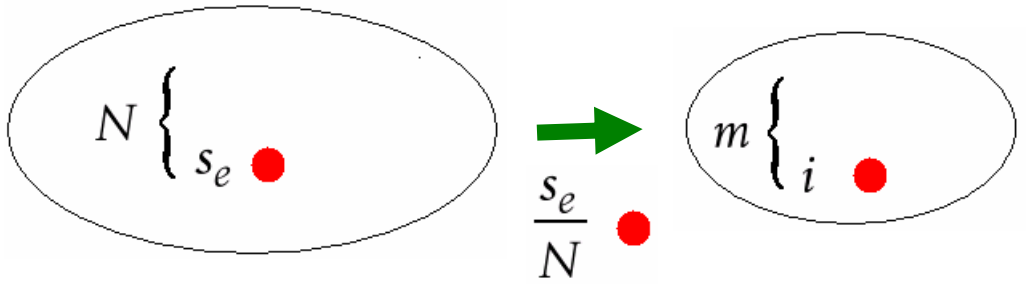
$$p(X\geq v;N,m,s_a)=\sum_{i=v}^{\min(m,s_a)}\frac{\binom{m}{i}\binom{N-m}{s_a-i}}{\binom{N}{s_a}}$$

Advantage: provides a means to externally validate a clustering algorithm, since the enrichment calculate makes no assumptions about how a set of genes was produced.

Expected Size Gene Set Enrichments (STEM)

A GO category enrichment p-value based on a profile's expected size.

The enrichment is computed using the *binomial distribution* (二項分配) based on the *expected* number of genes in the set of interest.



$$p(X = i ; m, \frac{s_e}{N}) = \binom{m}{i}\left(\frac{s_e}{N}\right)^i\left(1-\frac{s_e}{N}\right)^{m-i}$$

- N: the total number of unique genes on the microarray
- m: the total number of genes that are in the GO category of interest.
- S_e: the expected size of profile ***r***.

The p-value of seeing more than v genes belonging to both the category and profile r can be computed as:

$$p(X \geq v ; m, \frac{s_e}{N}) = \sum_{i=v}^{m}\binom{m}{i}\left(\frac{s_e}{N}\right)^i\left(1-\frac{s_e}{N}\right)^{m-i}$$

Advantage occurs: in the case in which the genes of multiple independent processes happen to have the same temporal expression pattern.

Example

- Data: immune response data from Guillemin et al. (2000, *PNAS*)
- Use hmuan 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.

Info for Authors | Editorial Board | About | Subscribe | Advertise | Contact | Site Map

PNAS

Proceedings of the National Academy of Sciences of the United States of America

Current Issue | Archives | Online Submission | GO | advanced search >>

Institution: Life Science Library, Academia Sinica Sign In as Member / Individual

Published online before print October 31, 2002, 10.1073/pnas.182558799

PNAS | November 12, 2002 | vol. 99 | no. 23 | 15136-15141

Microbiology

Cag pathogenicity island-specific responses of gastric epithelial cells to *Helicobacter pylori* infection

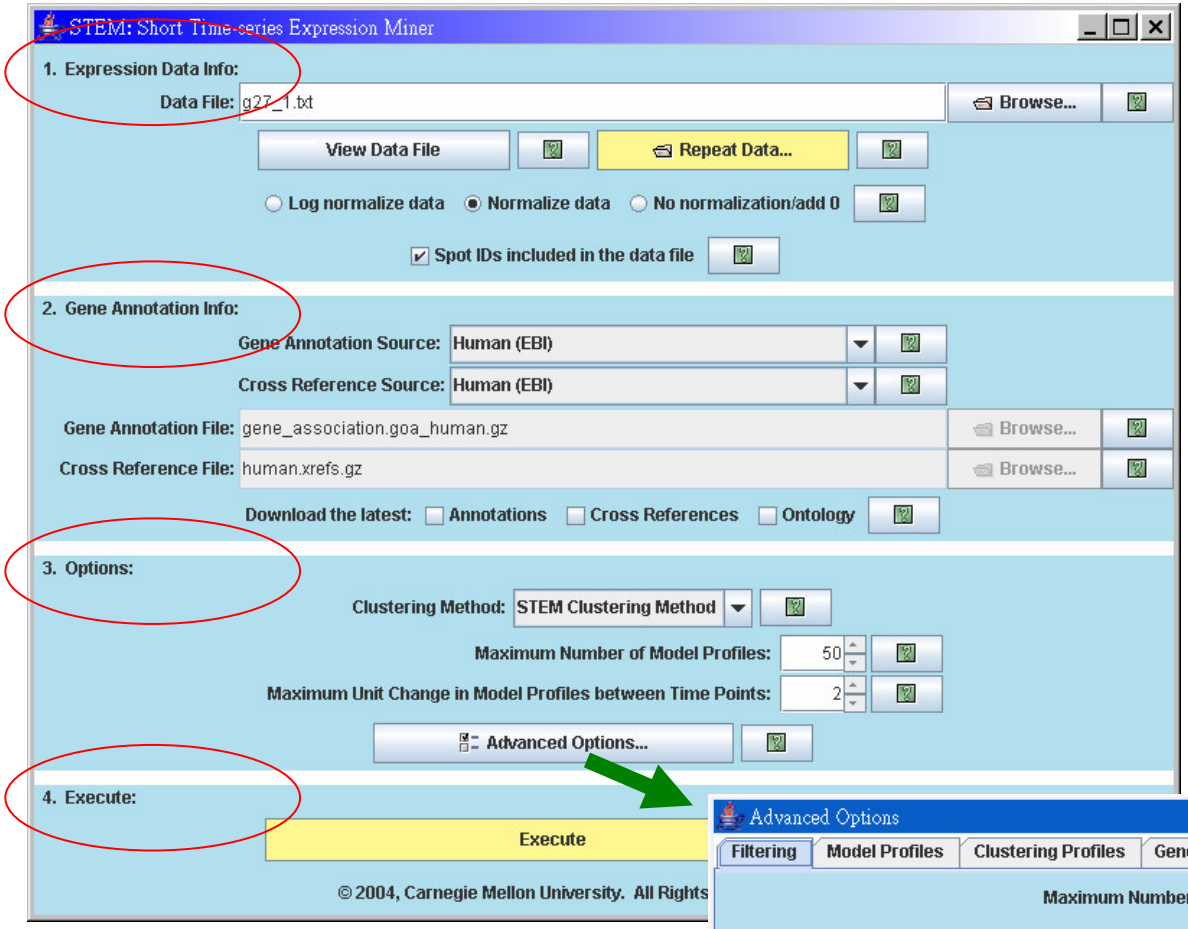
Karen Guillemin *†, Nina R. Salama *‡, Lucy S. Tompkins *§, and Stanley Falkow *

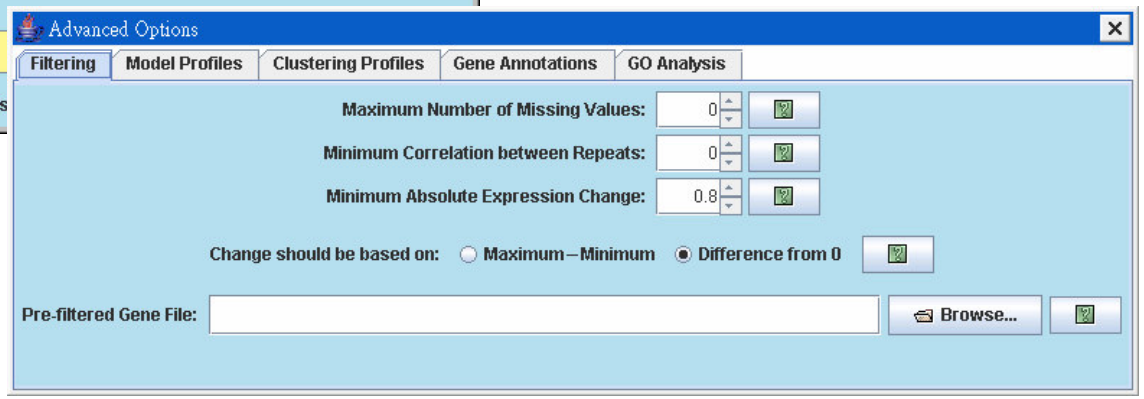
Departments of *Microbiology and Immunology, and§ Medicine, Stanford University School of Medicine, Stanford, CA 94305

Contributed by Stanley Falkow and approved 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

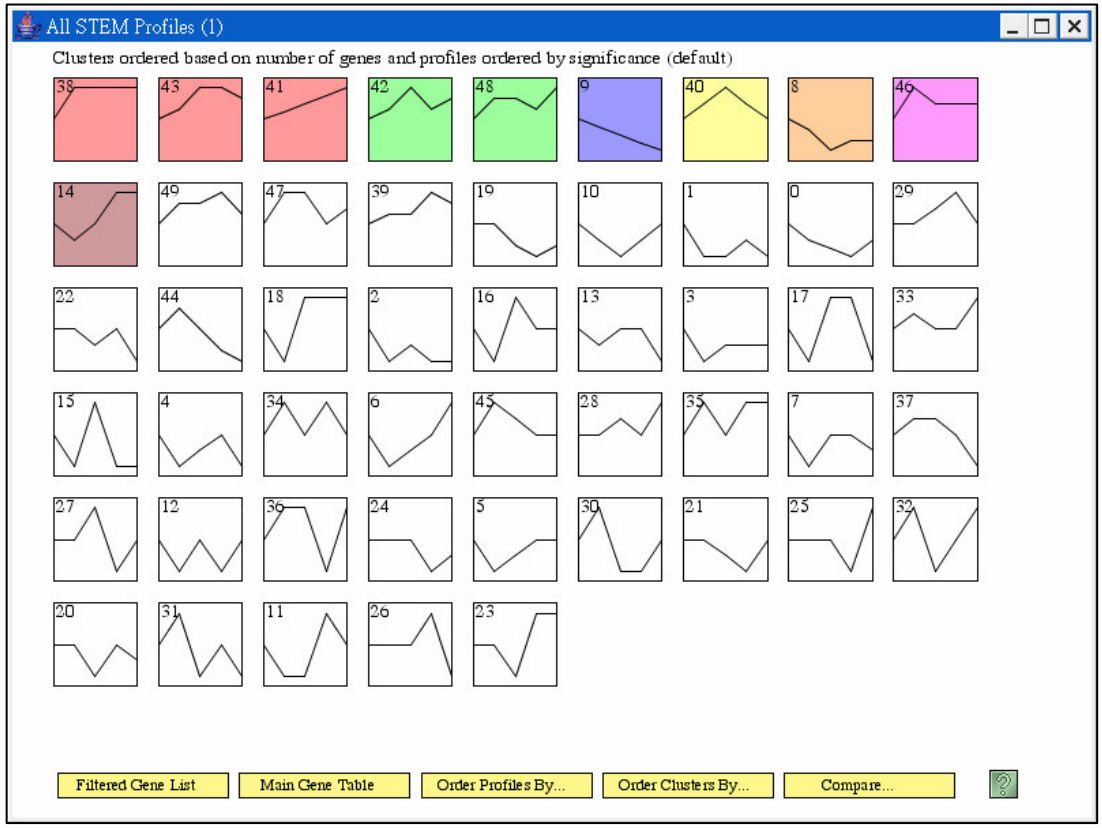




	A	B	C	D	E	F	G
1	SPOT	Gene Symbol	0h	0.5h	3h	6h	12h
2	1	ZFX	-0.027	0.158	0.169	0.193	-0.165
3	2	ZNF133	0.183	-0.068	-0.134	-0.252	0.177
4	3	USP2	-0.67	-0.709	-0.347	-0.779	-0.403
5	4	DSCR1L1	-0.923	-0.51	-0.718	-0.512	-0.668
6	5	WNT5A	-0.471	-0.264	-0.269	-0.154	-0.254
7	6	VHL	-0.327	-0.378	-0.229	-0.264	-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

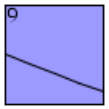
Clustering Results

- Colored profiles are significant.
- Profiles with the same shade belong to the same cluster.
- Corr=0.7➔ δ =0.3 in grouping method.
- one: 3 profiles,
  one: 2 profiles,
  five: single profiles.

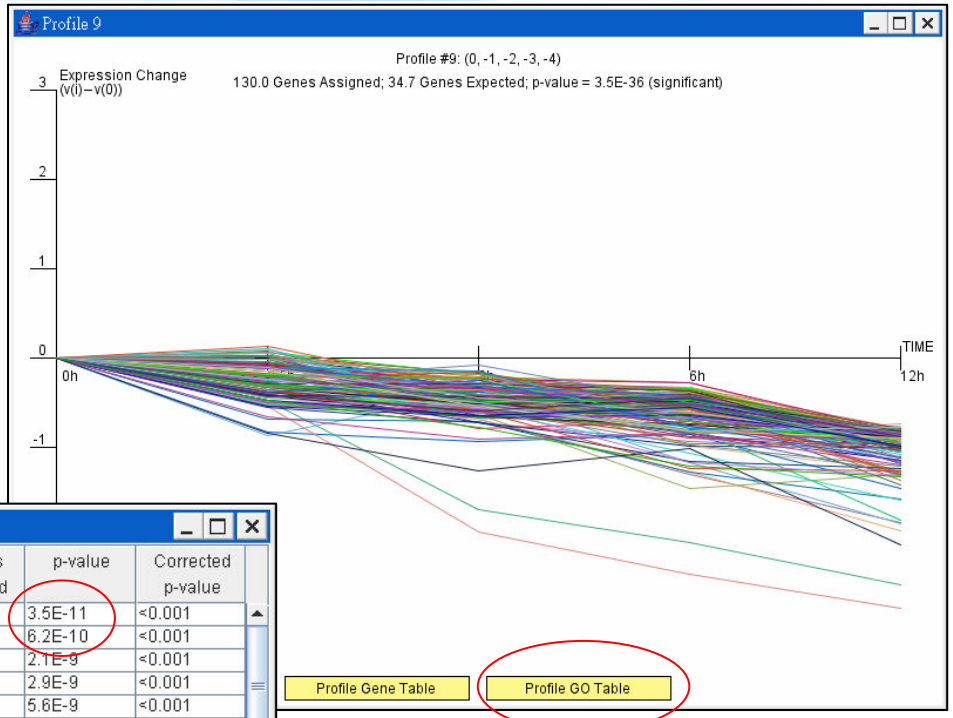


Four of the 10 significant model profiles were significantly enriched for GO categories. Two of these profiles were assigned to the cluster containing three profiles while the others remained separate.

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-value < 10^{-10}).



Profile GO Table

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

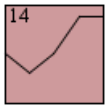
Category ID	Category Name	#Genes Category	#Genes Assigned	#Genes Expected	#Genes Enriched	p-value	Corrected p-value
GO:0007049	cell cycle	432	19.0	2.7	+16.3	3.5E-11	<0.001
GO:0006259	DNA metabolism	344	16.0	2.2	+13.8	6.2E-10	<0.001
GO:0006260	DNA replication	110	10.0	0.7	+9.3	2.1E-9	<0.001
GO:0006139	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

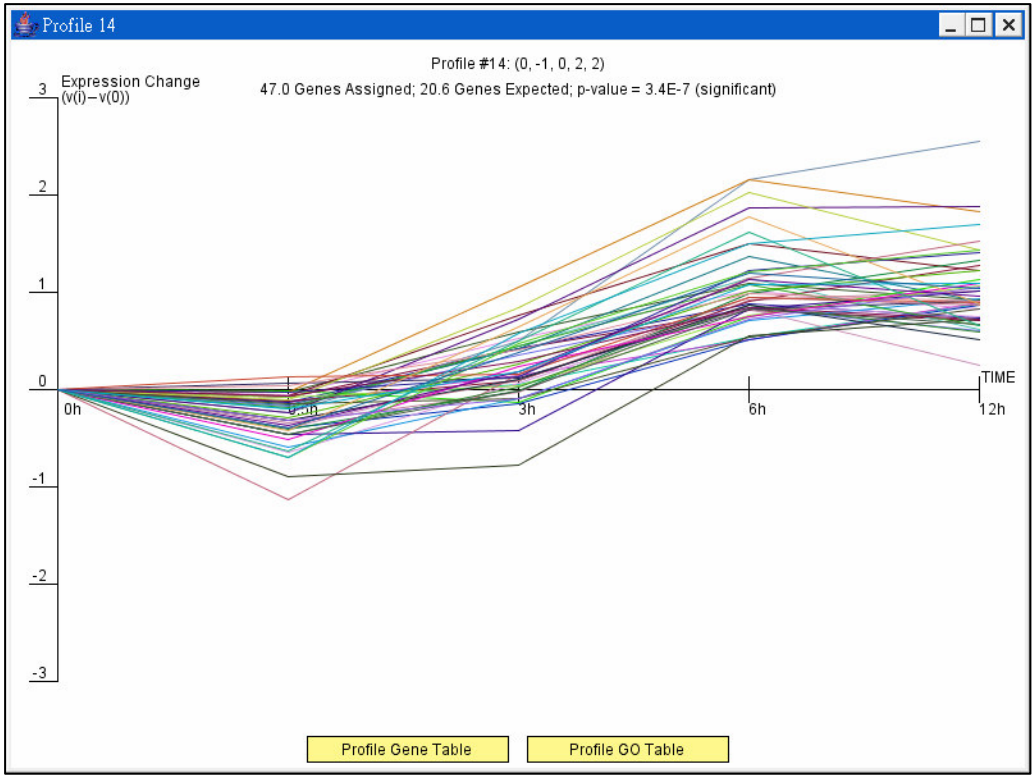
Many of the cycling genes in this profile are known transcription factors, which could contribute to repression of cell-cycle genes, and ultimately, the cell cycle.

GO Interpretation

- 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 ($p=9x\ 10^{-5}$),
  - extracellular matrix ($9x\ 10^{-4}$),
  - membrance ($2x\ 10^{-6}$).



Profile GO Table

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

Other Functionalities of STEM

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

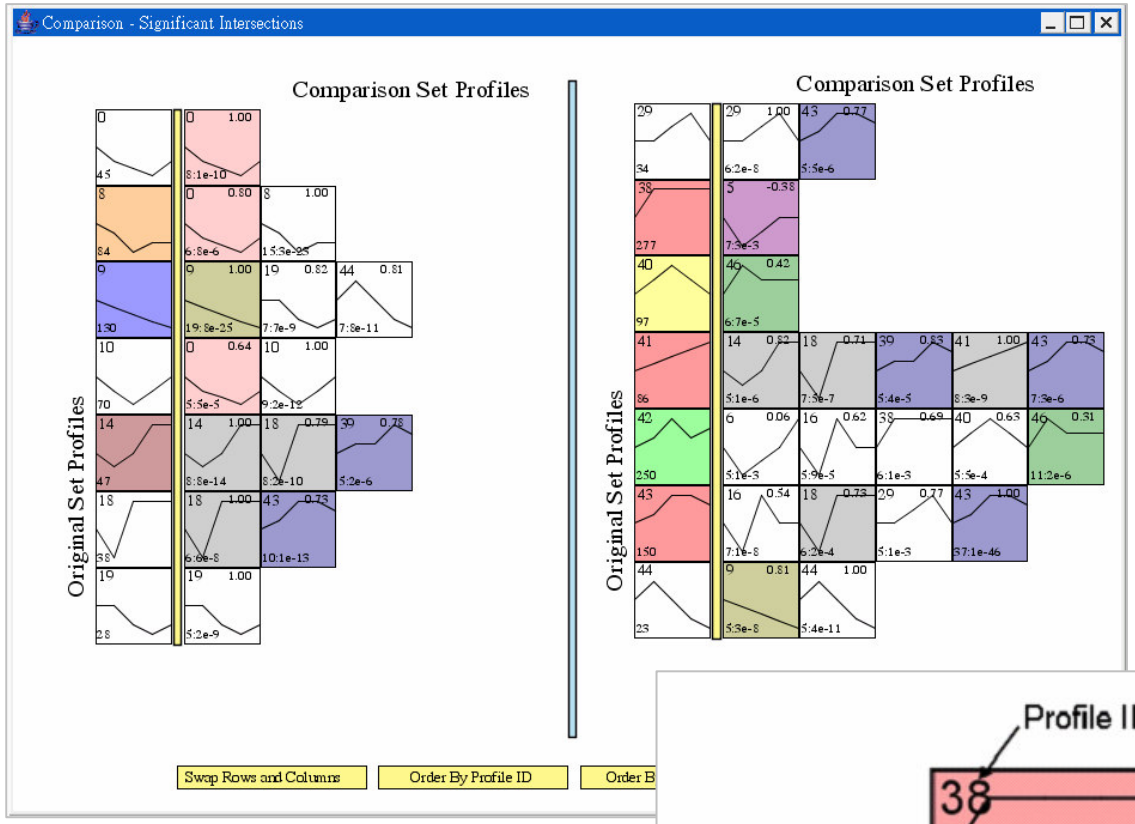
- For a set of genes which had temporal response X in experiment A, which significant responses did they have in experiment B?
- use hypergeometric distribution to compute the significance of overlap between gene sets of model profiles of two experiments

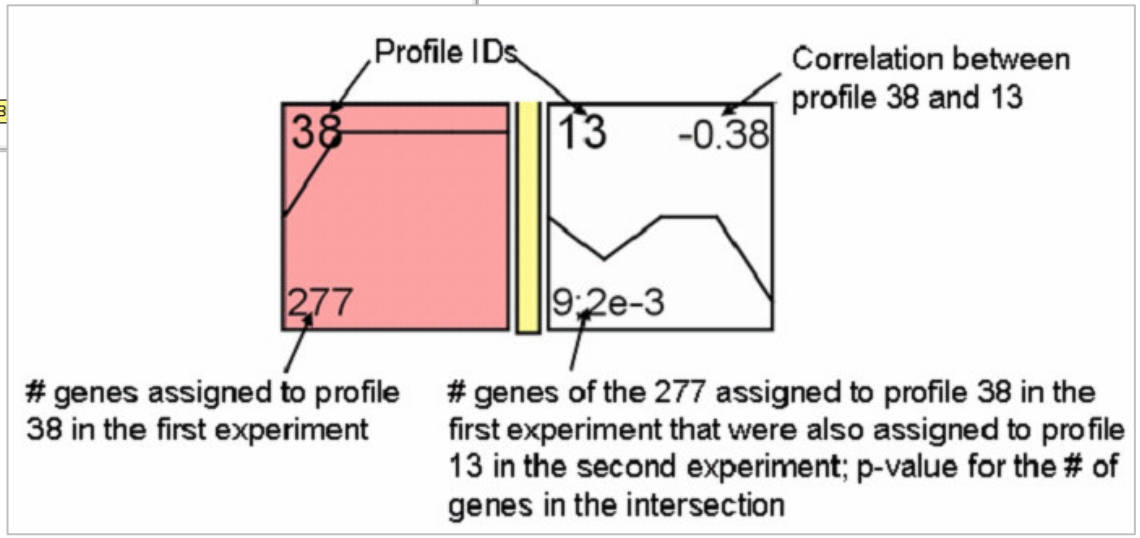
Example

- Compare the temporal response of gene infected with a wildtype pathogen to those infected woth 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)

Comparing Data Sets

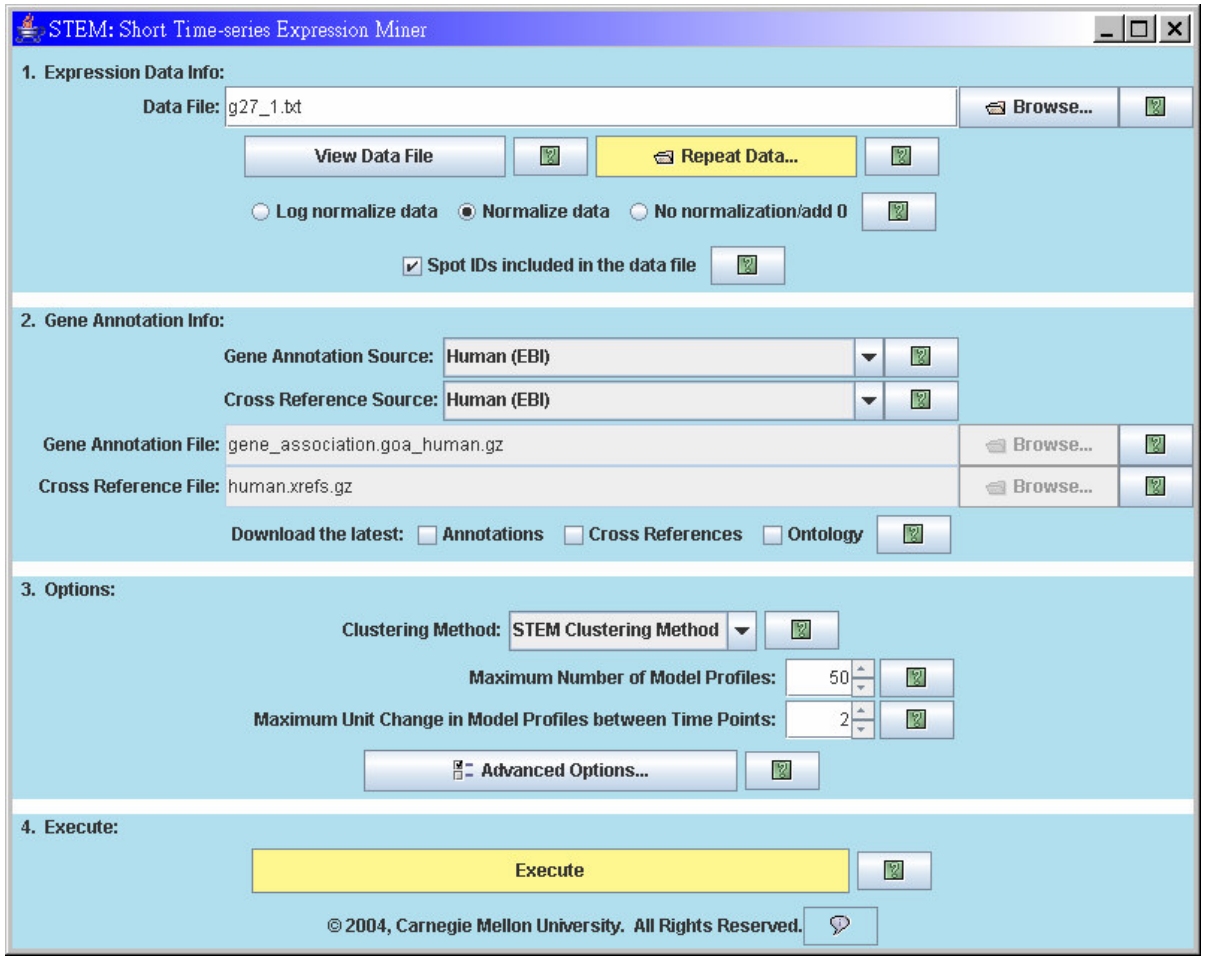


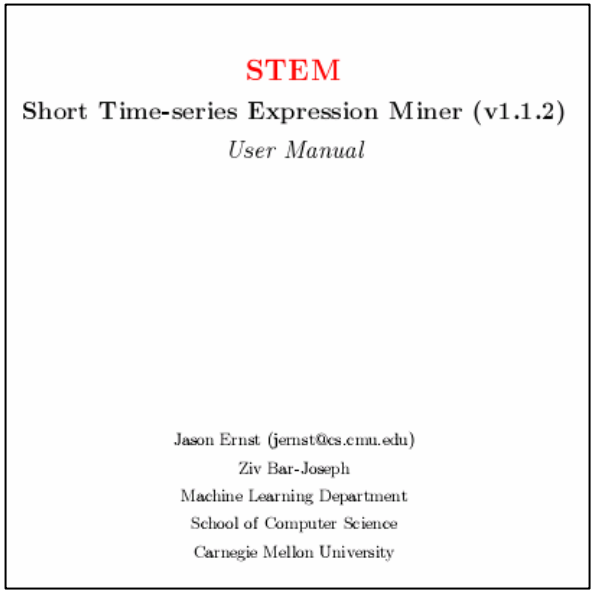


Software Practice

stemGuilleminSample.cmd

```
"C:\Program Files\Java\jre1.5.0_04\bin\java.exe" -mx1024M -ms512M -jar stem.jar -d defaultsGuilleminSample.txt
pause
```





STEM

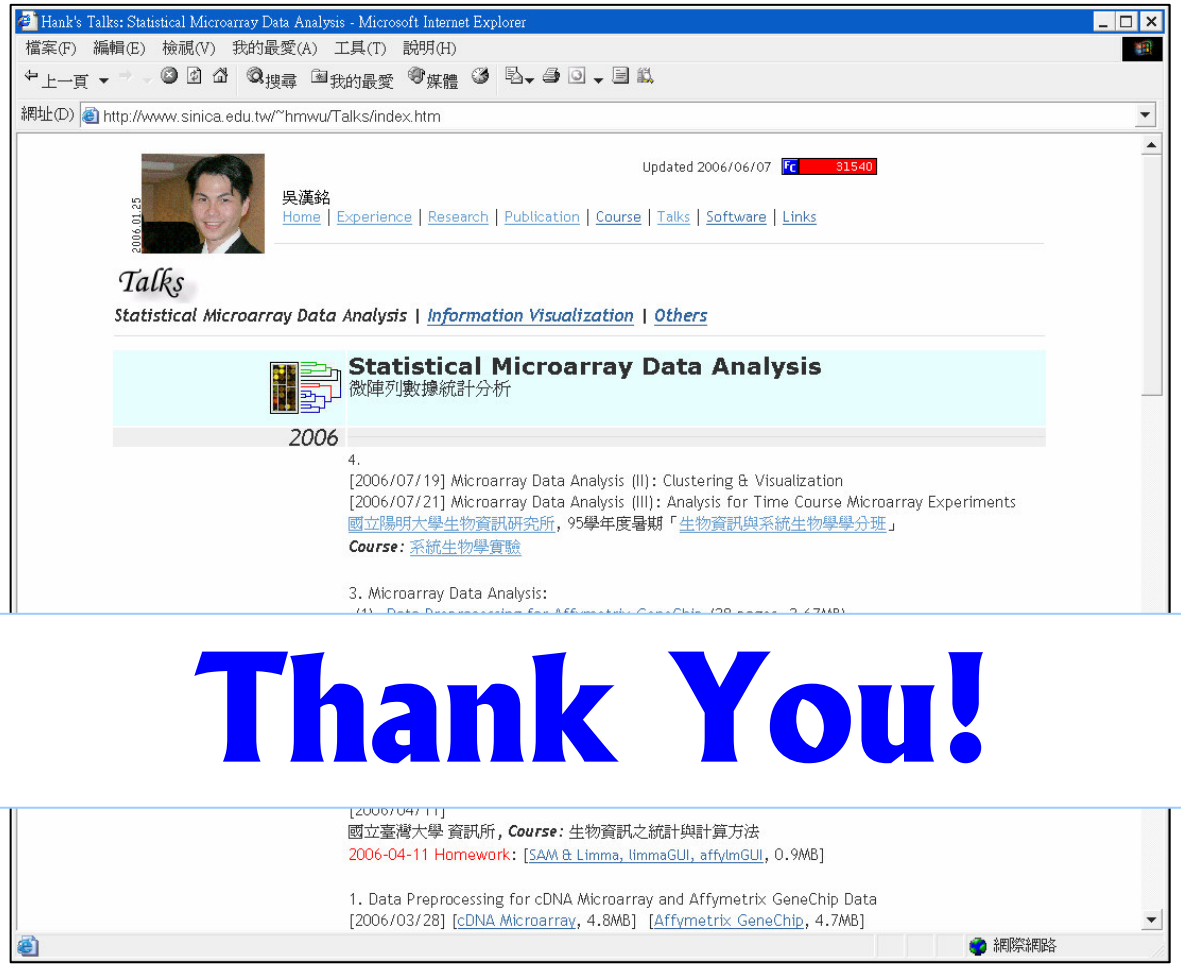
Short Time-series Expression Miner (v1.1.2)

User Manual

Jason Ernst (jernst@cs.cmu.edu)
Ziv Bar-Joseph
Machine Learning Department
School of Computer Science
Carnegie Mellon University

Questions?

Reference: http://www.sinica.edu.tw/~hmwu/MADA/TimeCourse/index.htm



Thank You!

吳漢銘
hmwu@stat.sinica.edu.tw
http://www.sinica.edu.tw/~hmwu
中央研究院 統計科學研究所
Institute of Statistical Science, Academia Sinica