

Measuring similarities between gene expression profiles through new data transformations

Presented by Dragana Veljkovic

Kim K., S. Zhang, et al. (2007) Measuring similarities between gene expression profiles through new data transformations. BMC Bioinformatics 8(1):29

Outline

- Traditional metrics
- Poisson modeling of SAGE data
- Distances in the lifted space
- Results

Standard distances

● Pearson correlation coefficient

- Overly sensitive to the shape of expression curve
- Variables X and Y , with means X_m and Y_m and standard deviations S_X and S_Y respectively

$$r = \frac{\sum_{i=1}^n (X - X_m)(Y - Y_m)}{(n-1)S_X S_Y}$$

● Euclidian distance

- Mainly considers magnitude change

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

More standard distances

● Spearman Rank Correlation

- Cannot distinguish real differences from random errors

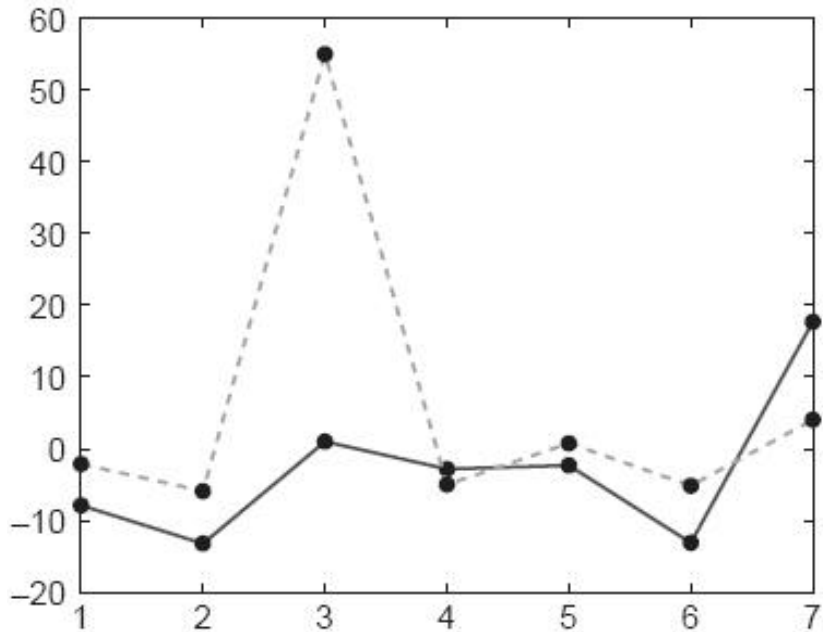
$$r = 1 - \frac{6 \sum_{i=1}^n (\text{rank}(x_i) - \text{rank}(y_i))^2}{n(n^2 - 1)}$$

● Sign of the first-order differences

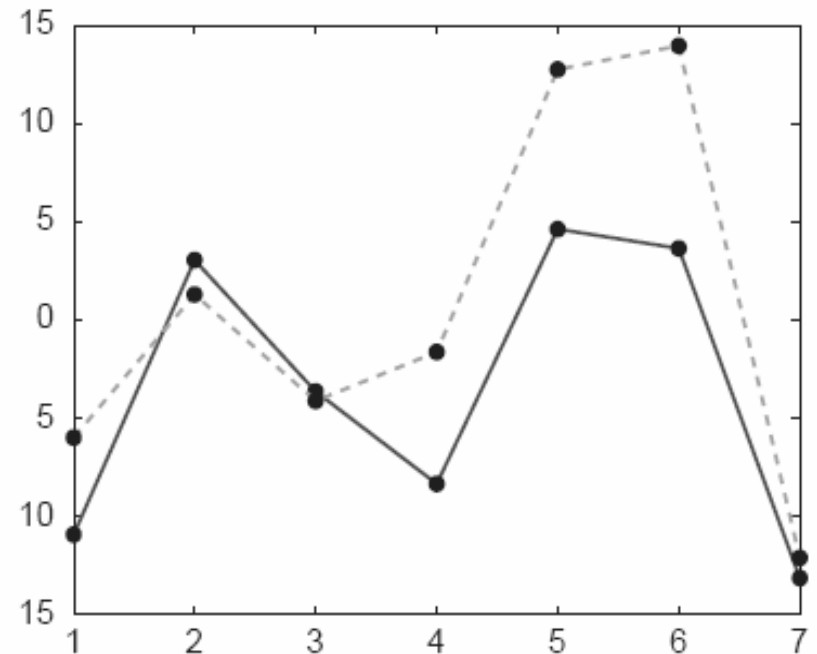
- Loses information

$$r = \frac{1}{n} \sum_{i=1}^{n-1} \text{sgn}((x_{i+1} - x_i)(y_{i+1} - y_i))$$

Distance comparison



The SRC for the two profiles is 0.93 whereas the Pearson correlation is 0.3.



The SRC for the two profiles is 0.93 whereas sign of first order differences yields a similarity of 0.3.

SAGE data

- Serial Analysis of Gene Expression
- Produces a snapshot of the messenger RNA population in a sample of interest
- The output of SAGE is a list of short sequence tags and the number of times it is observed, called a SAGE library
- Tag counts are approximately Poisson distributed
- Comparison to DNA micorarray
 - SAGE is a sequence-based sampling technique.
 - Observations are not based on hybridization, which result in more qualitative, analog values.
 - mRNA sequences do not need to be known a priori, so genes or gene variants which are not known can be discovered.
 - Microarray experiments are much cheaper to perform, large-scale studies do not typically use SAGE.

Data model

- $Y_i(t)$ count of tag i in library t
- $\mathbf{Y}_i = (Y_i(1), \dots, Y_i(T))$ vector of counts of tag i over a total of T libraries
- $Y_i(t)$ is Poisson distributed with mean $\lambda_{it} = \mu_i(t) \alpha_i$ where:
 - α_i defines magnitude
 - $\mu_i = (\mu_i(1), \dots, \mu_i(T))$ defines shape
- Parameters are estimated as:

$$\mathbf{q}_i = \sum_{t=1}^T Y_i(t) \quad \mathbf{l}_i(t) = \frac{Y_i(t)}{\mathbf{q}_i} \quad \text{and} \quad \sum_{t=1}^T \mathbf{l}_i(t) = 1$$

Likelihood estimates

- Take a cluster consisting of tags $1, 2, \dots, m$ corresponding to libraries $\mathbf{Y}_1, \dots, \mathbf{Y}_m$
- Maximum likelihood estimates of \bullet and $\square_1, \dots, \square_m$ are

$$\hat{\mathbf{q}}_i = \sum_{t=1}^T Y_i(t) \quad \text{and} \quad \hat{\mathbf{I}}(t) = \frac{\sum_{i=1}^m Y_i(t)}{\sum_{i=1}^m \hat{\mathbf{q}}_i} = \frac{\sum_{i=1}^m Y_i(t)}{\sum_{i=1}^m \sum_{t=1}^T Y_i(t)}$$

- Common shape parameter $\bullet = (\bullet(1), \dots, \bullet(T))$
- Joint likelihood function estimates the probability of tags $1, 2, \dots, m$ belonging to the same cluster

Cluster estimates

- Evaluate how a particular tag fits in a cluster
- Log-likelihood function estimate

$$L_j = -\log(Y_j | \hat{\mathbf{q}}_j) = \sum_{t=1}^T (\hat{\mathbf{I}}(t)\hat{\mathbf{q}}_j - Y_j(t) \log(\hat{\mathbf{I}}(t)\hat{\mathbf{q}}_j) + \log(Y_j(t)!))$$

- Chi-square static estimate

$$D_j = \sum_{t=1}^T \frac{(Y_j(t) - \hat{\mathbf{I}}(t)\hat{\mathbf{q}}_j)^2}{\hat{\mathbf{I}}(t)\hat{\mathbf{q}}_j}$$

PoissonC and PoissonL

1. All SAGE tags are assigned at random to K sets. λ_j is estimated for each tag
2. Set cluster center $\bullet(k,0)$ for each cluster.
Current iteration $i=0$
3. In the i^{th} iteration, assign each tag j to cluster with the closest center
 - PoissonC uses Chi-square distance to estimate fit
 - PoissonL uses log-likelihood
4. Set new cluster center $\bullet(k,i+1)$
5. Go to step 3 until convergence

Lifting methods

- By projecting to an abstract space, previously hidden properties can be revealed
- Similar to kernel functions in dimension reduction methods
- Set of column vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ defines the transformation of the gene expression profile $\mathbf{Y}_i = (Y_i(1), \dots, Y_i(T))$ as
$$\mathbf{Z}_i = (Z_{i1}, \dots, Z_{in}) = \mathbf{Y}_i(\mathbf{e}_1, \dots, \mathbf{e}_n)$$
- Column vectors can be
 - Eigenspace of a parametric covariance matrix
 - Set of principal component vectors based on the empirical covariance matrix (PCAChisq)
 - Vector that compute the difference between each pair of original vector elements (TransChisq)

TransChisq

- Considers mutual differences of the original vector components
- Gene with expression profile $\mathbf{Y}_i = (Y_i(1), \dots, Y_i(T))$ is transformed to vector \mathbf{Z}_i of dimension $T(T-1)/2$ with components of the form $Y_i(t_1) - Y_i(t_2)$ for $t_1 = 1, \dots, T-1$ and $t_2 = (t_1+1), \dots, T$
- Mean and variance of the projected data can be computed as:
 - $E(Y_i(t_1) - Y_i(t_2)) = (\mu_i(t_1) - \mu_i(t_2)) \square_i$
 - $\text{Var}(Y_i(t_1) - Y_i(t_2)) = (\sigma_i(t_1) + \sigma_i(t_2)) \square_i$
- For a cluster consisting of tags $1, 2, \dots, m$ cluster dispersion can be computed as

$$S_j = \sum_i \sum_{t_1, t_2} \frac{((Y_i(t_1) - Y_i(t_2)) - E(Y_i(t_1) - Y_i(t_2))))^2}{\text{Var}(Y_i(t_1) - Y_i(t_2))}$$

Generalized metric

- According to Poisson model, mean and variance of the projected data are

$$E(Z_{it}) = E(\mathbf{Y}_i)\mathbf{e}_t = (\mathbf{I}_i(1)\mathbf{q}_i, \dots, \mathbf{I}_i(T)\mathbf{q}_i)\mathbf{e}_t$$

$$\text{Var}(Z_{it}) = (\mathbf{I}_i(1)\mathbf{q}_i, \dots, \mathbf{I}_i(T)\mathbf{q}_i)\mathbf{e}_t^2$$

- For cluster consisting of tags $1, \dots, m$ cluster dispersion can be calculated as:

$$S = \sum_i \sum_{t_1=1}^T \frac{(Z_{it} - E(Z_{it}))^2}{\text{Var}(Z_{it})}$$

Simulated data

Table 2. Five dimensional simulation dataset with Normal distributions ($\sigma^2 = 3\mu$).

Group ID	Mean parameters of the Normal distributions (μ)					
Group A	a1 ~ a3	1	1	1	15	150
Group B	b1 ~ b6	15	1	1	1	150
Group C	c1 ~ c4	10	30	30	60	10
	c5 ~ c6	100	300	300	600	100
Group D	d1 ~ d7	200	70	70	10	10
	d8 ~ d9	2000	700	700	100	100
Group E	e1 ~ e5	210	120	10	10	10
	e6 ~ e7	2100	1200	100	100	100
Group F	f1 ~ f3	5	50	5	5	5
	f4 ~ f6	5	75	5	5	5
	F7 ~ f9	5	100	5	5	5
	f10 ~ f11	50	500	50	50	50
	f12 ~ f13	50	750	50	50	50
	f14 ~ f15	50	1000	50	50	50

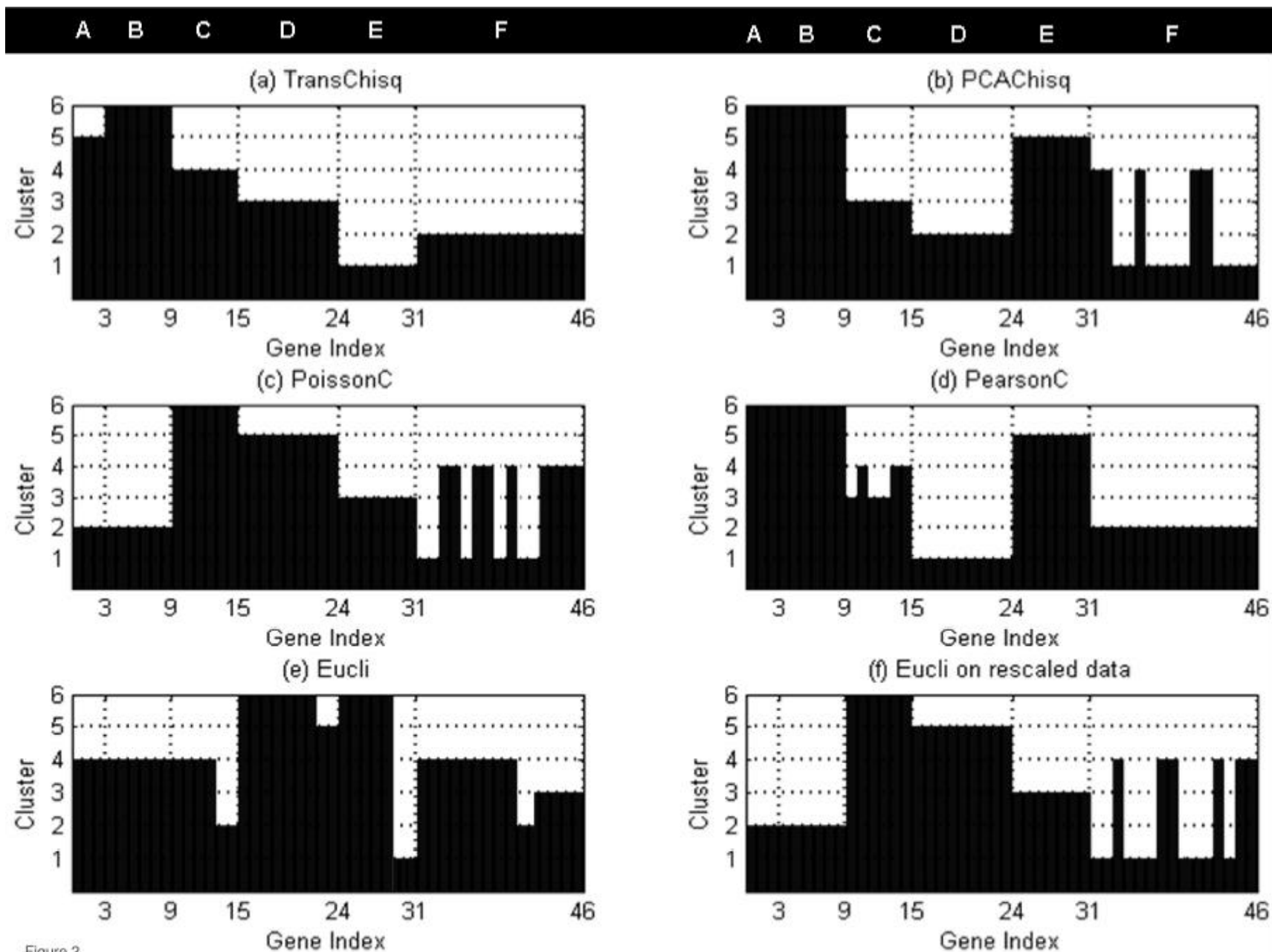


Figure 2

Experimental maize gene expression data

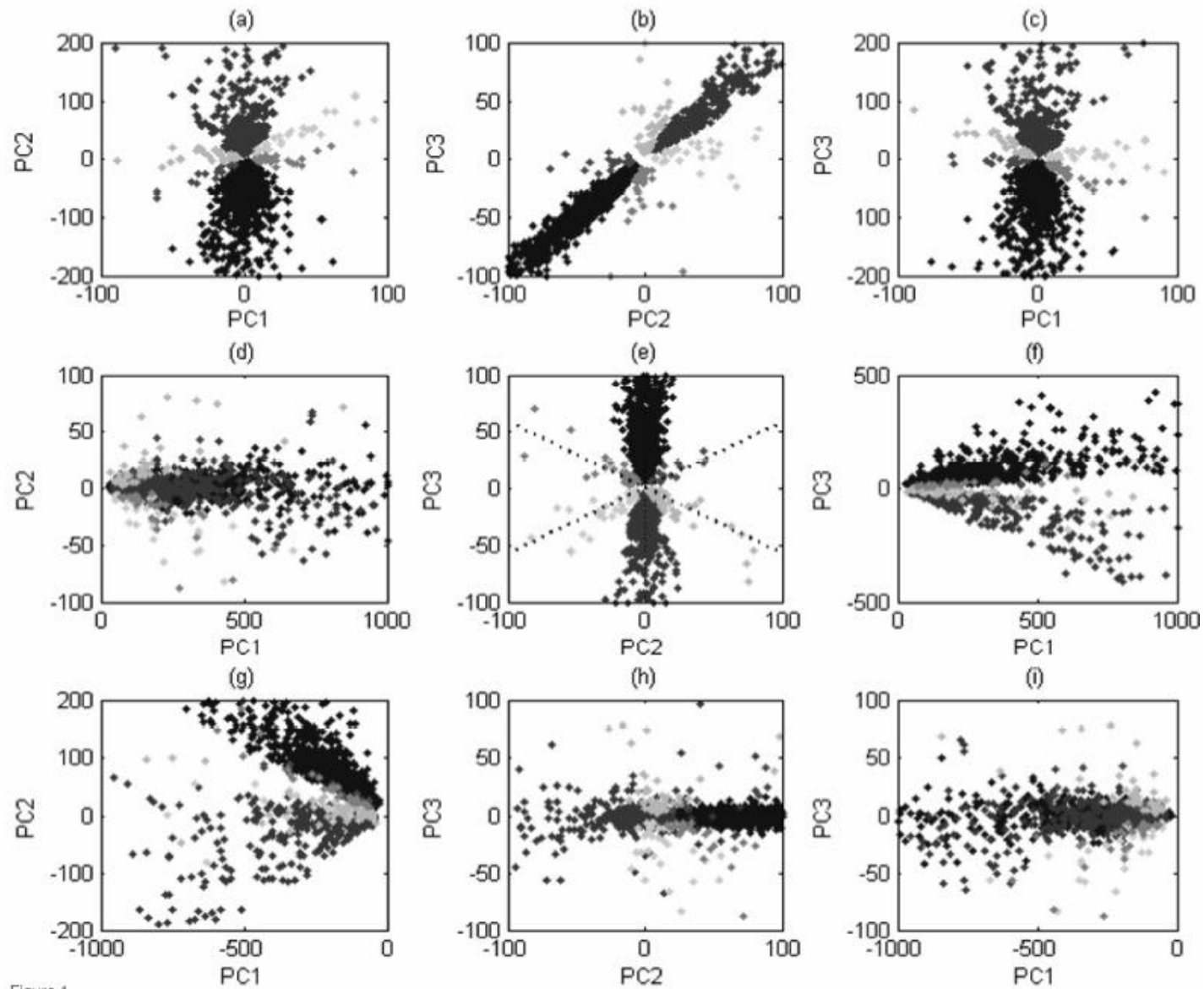


FIGURE 1

Microarray yeast sporulation gene expression data

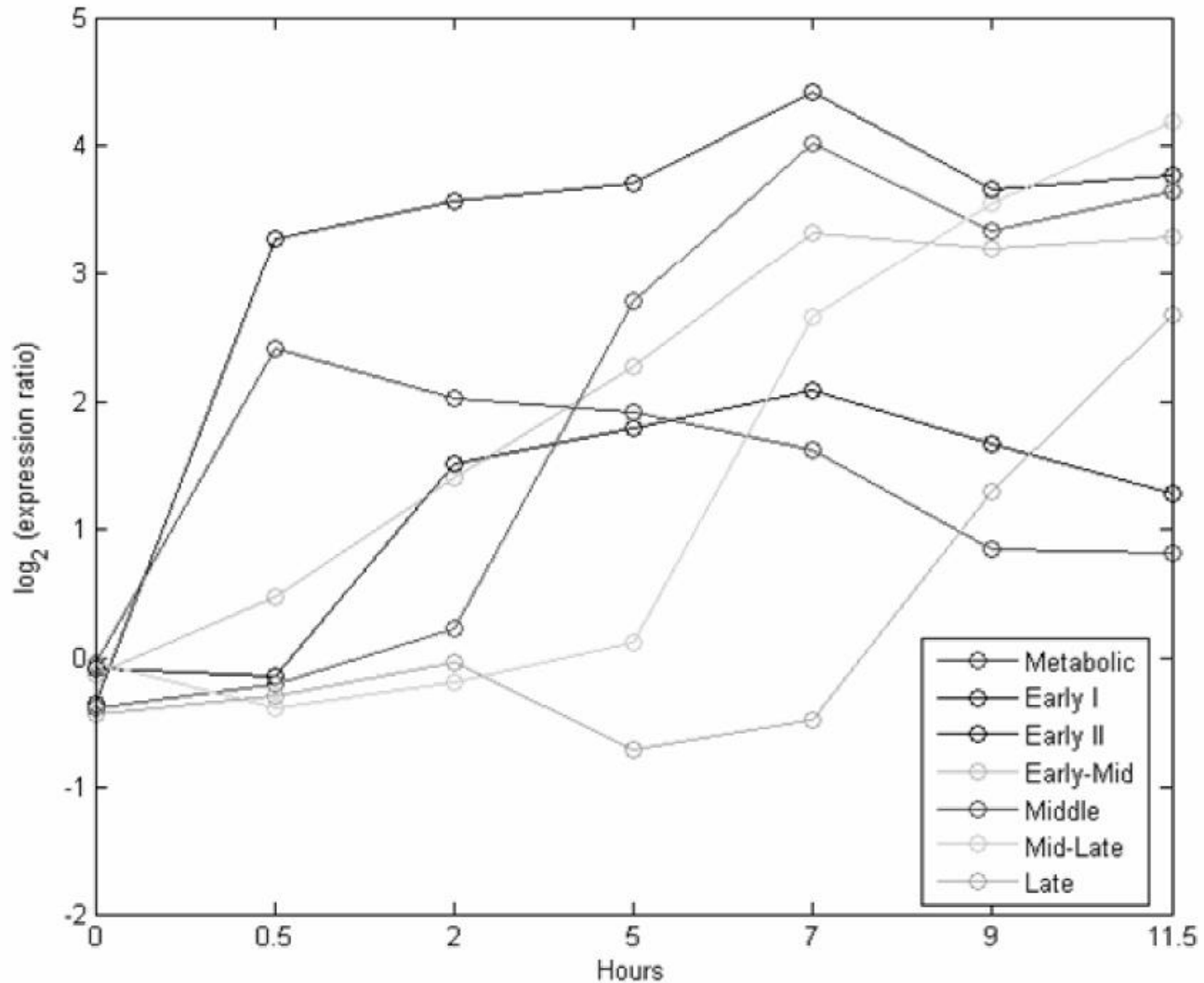


Figure 4

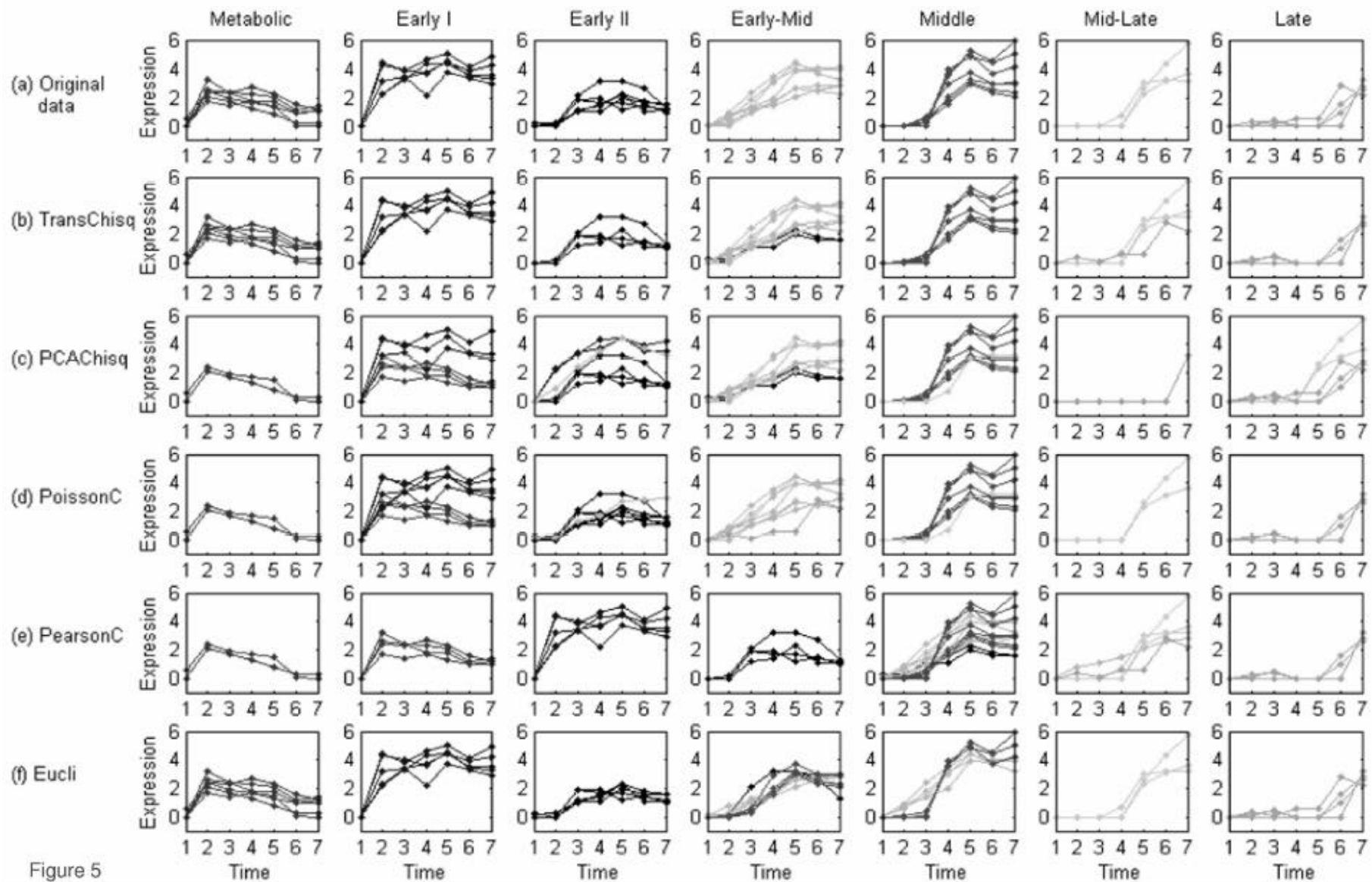


Figure 5

References

- Kim, K., S. Zhang, et al. (2007). "Measuring similarities between gene expression profiles through new data transformations." *BMC Bioinformatics* 8(1): 29.
- Balasubramaniyan, R., E. Hullermeier, et al. (2005). "Clustering of gene expression data using a local shape-based similarity measure." *Bioinformatics* 21(7): 1069-77.
- Cai, L., H. Huang, et al. (2004). "Clustering analysis of SAGE data using a Poisson approach." *Genome Biology* 5(7): R51.