Information criterion-based clustering with order-restricted candidate profiles in short time-course microarray experiments

Tianqing Liu  
*Northeast Normal University*

Nan Lin  
*Washington University School of Medicine in St. Louis*

Ningzhong Shi  
*Northeast Normal University*

Baoxue Zhang  
*Northeast Normal University*

Follow this and additional works at: [https://digitalcommons.wustl.edu/open_access_pubs](https://digitalcommons.wustl.edu/open_access_pubs)

Part of the Medicine and Health Sciences Commons

Please let us know how this document benefits you.

**Recommended Citation**

Liu, Tianqing; Lin, Nan; Shi, Ningzhong; and Zhang, Baoxue, "Information criterion-based clustering with order-restricted candidate profiles in short time-course microarray experiments." BMC Bioinformatics. 10, 146. (2009).

[https://digitalcommons.wustl.edu/open_access_pubs/345](https://digitalcommons.wustl.edu/open_access_pubs/345)

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
Methodology article

Information criterion-based clustering with order-restricted candidate profiles in short time-course microarray experiments

Tianqing Liu¹, Nan Lin², Ningzhong Shi¹ and Baoxue Zhang*¹

Address: ¹Key Laboratory for Applied Statistics of MOE and School of Mathematics and Statistics, Northeast Normal University, Changchun, PR China and ²Department of Mathematics, Washington University in St Louis, St Louis, USA

E-mail: Tianqing Liu - tianqingliu@gmail.com; Nan Lin - nlin@math.wustl.edu; Ningzhong Shi - shinz@nenu.edu.cn; Baoxue Zhang* - bxzhang@nenu.edu.cn

*Corresponding author

Published: 15 May 2009 Received: 4 February 2009


Accepted: 15 May 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/146

© 2009 Liu et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Time-course microarray experiments produce vector gene expression profiles across a series of time points. Clustering genes based on these profiles is important in discovering functional related and co-regulated genes. Early developed clustering algorithms do not take advantage of the ordering in a time-course study, explicit use of which should allow more sensitive detection of genes that display a consistent pattern over time. Peddada et al. [1] proposed a clustering algorithm that can incorporate the temporal ordering using order-restricted statistical inference. This algorithm is, however, very time-consuming and hence inapplicable to most microarray experiments that contain a large number of genes. Its computational burden also imposes difficulty to assess the clustering reliability, which is a very important measure when clustering noisy microarray data.

Results: We propose a computationally efficient information criterion-based clustering algorithm, called ORICC, that also takes account of the ordering in time-course microarray experiments by embedding the order-restricted inference into a model selection framework. Genes are assigned to the profile which they best match determined by a newly proposed information criterion for order-restricted inference. In addition, we also developed a bootstrap procedure to assess ORICC's clustering reliability for every gene. Simulation studies show that the ORICC method is robust, always gives better clustering accuracy than Peddada's method and saves hundreds of times computational time. Under some scenarios, its accuracy is also better than some other existing clustering methods for short time-course microarray data, such as STEM [2] and Wang et al. [3]. It is also computationally much faster than Wang et al. [3].

Conclusion: Our ORICC algorithm, which takes advantage of the temporal ordering in time-course microarray experiments, provides good clustering accuracy and is meanwhile much faster than Peddada's method. Moreover, the clustering reliability for each gene can also be assessed, which is unavailable in Peddada’s method. In a real data example, the ORICC algorithm identifies new and interesting genes that previous analyses failed to reveal.
Background
The development of microarray technology provides a powerful analytical tool for large scale genomic research. Its ability to simultaneously study thousands of genes under a multitude of conditions presents a huge challenge to comprehend and interpret the resulting mass of data. An important application of microarray technology is to study the dynamic patterns of gene expression across a series of time points and find gene clusters within which genes share similar patterns. The premise is that genes sharing similar expression profiles might be functionally related or co-regulated. Therefore, microarray data may provide insights into gene-gene interactions, gene function and pathway identification. Examples of such studies include response to temperature changes and other stress conditions [4], immune response [5], developmental studies [6], and various systems in the cell [7]. Early clustering analysis of microarray data were mostly on static microarray experiments, such as hierarchical clustering [8], the $k$-nearest neighbors method [9,10], and other correlation-based methods [11,12]. These methods are not designed for time-course microarray data and can not effectively utilize the temporal information. Many clustering algorithms for time-course microarray data have been developed afterwards. Most of them view observed temporal gene expression profiles coming from underlying smooth curves and cluster genes based on estimated expression profiles obtained from nonparametric smoothing [2,13-25]. While these algorithms work well for relatively long time series data, they are not appropriate for short time-course microarray data often taken on a small number of sparse time points. Generally, these algorithms tend to overfit the data when the number of time points is small [2]. A few clustering methods were also proposed specifically for short time-course microarray data, such as [1-3,26,27]. Among them, Peddada et al. [1] proposed an interesting idea of using order-restricted inference in clustering short time-course microarray data. A set of candidate expression profiles is first defined by inequality constraints among expression levels at different time points, i.e. up and downs. This strategy is less restrictive than those that define profiles via pre-specified expression levels because only the general shape of the profile is needed. Each candidate profile then represents a potential gene cluster. For short time-course data, the total number of candidate profiles is generally not large. Peddada's method then assigns genes to the cluster represented by its best matched candidate profile determined by some order-restricted statistical inference procedure. Two genes fall into the same cluster if their best matched candidate profiles are the same. This profile matching clustering strategy is different from most unsupervised clustering where a representation of a cluster is often calculated only after the cluster is formed. This method requires no smoothing of the expression profiles and was shown to discover more functionally related genes when applied to a breast cancer cell-line data in [28]. However, Peddada's method is computationally very costly when the number of genes is large. According to our experience on a workstation with a 2.30 GHz AMD Athlon(tm) 64 × 2 Dual Core 4400+ processor and a 2.00 GB memory, it took at least 72 hours to run Peddada’s method (implemented in Matlab) for the breast cancer cell line data containing about 1900 genes.

In this article, from a different perspective to the order-restricted inference, we propose a new order-restricted information criterion-based clustering (ORICC) algorithm, which is computationally much more efficient than Peddada’s method. Our method selects and clusters genes using the ideas of model selection for order-restricted inference, where estimation makes use of inequalities that define the candidate profiles. The first step is to define candidate profiles and express them in terms of inequalities between the expected gene expression levels at various time points. For a given candidate profile, we estimate the mean expression level at different time points of each gene using the order-restricted maximum likelihood [29]. The best fitting profile for a given gene is then selected using an information criterion for order-restricted inference. Due to the simplicity of our algorithm, the analysis of the breast cancer cell-line data in [28] can now be done in a few minutes using our method.

Results and discussion
Inequality profiles
Suppose that a time-course microarray experiment includes $T$ time points, and at each time point there are $M$ arrays each with $G$ genes. Denote $y_{gt}$ the expression measurement of gene $g$ at time point $t$ on the $i$th array. Suppose that the unknown true mean expression level of gene $g$ at time $t$ is $\mu_{g}$, i.e. $\mu_{gt} = E(y_{gt})$ for all $i$. A candidate gene expression profile of gene $g$ is then given by inequalities between the components of $\mu_{g} = (\mu_{g1}, \mu_{g2}, ..., \mu_{gT})^{T}$. In the following, we define some typical inequality profiles, and we drop the subscript $g$ for simplicity.

- $C_{0}$ profile:
  \[ C_{0} = \{ \mu \in R^{T} : \mu_{1} = \mu_{2} = \cdots = \mu_{T} \}. \tag{1} \]

- $C_{\perp}$ profile:
  \[ C_{\perp} = \{ \mu \in R^{T} : \mu_{1} \perp \mu_{2} \perp \cdots \perp \mu_{T} \}, \tag{2} \]
where $\mu_{i} \perp \mu_{j}$ means that there is no defined inequality constraint between $\mu_{i}$ and $\mu_{j}$. 

$C_{0}$ is the usual full profile, and $C_{\perp}$ is a profile in which we only require $\mu_{1}$ to be the only profile.

\[ y_{gt} \text{ is assumed to have a normal distribution with mean } \mu_{gt} \text{ and variance } \sigma_{g}^{2}, \]
\[ y_{g} = (y_{g1}, y_{g2}, ..., y_{gT})^{T} \text{ follows a multivariate normal distribution with mean } \mu_{g} \text{ and variance } \Sigma \text{.} \]

\[ \mu_{gt} = \mu_{g1} + \mu_{g2} + \cdots + \mu_{gT} \]

\[ \mu_{g1} \leq \mu_{g2} \leq \cdots \leq \mu_{gT} \]

\[ \mu_{gt} = \mu_{g1} + \mu_{g2} + \cdots + \mu_{gT} \]

\[ \mu_{g1} \leq \mu_{g2} \leq \cdots \leq \mu_{gT} \]

\[ \mu_{g1} \perp \mu_{g2} \perp \cdots \perp \mu_{gT} \]
Monotone increasing profile (simple order) :
\[ C_\uparrow = \{ \mu \in R^T : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_T \} \] (3)
(with at least one strict inequality). Similarly, a monotone decreasing profile \( C_\downarrow \) is given by replacing \( \leq \) by \( \geq \) in (3).

Up-down profile with maximum at \( i \) (umbrella order) :
\[ C_{\uparrow \downarrow} = \{ \mu \in R^T : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_i \geq \mu_{i+1} \geq \cdots \geq \mu_T \} \] (4)
(with at least one strict inequality among \( \mu_1 \leq \mu_2 \leq \cdots \leq \mu_i \) and one among \( \mu_i \geq \mu_{i+1} \geq \cdots \geq \mu_T \)). Genes satisfying this profile have mean expression values non-decreasing in time up to time point \( i \) and non-increasing thereafter. One may similarly define a down-up profile \( C_{\downarrow \uparrow} \).

Cyclical profile with minima at \( 1, j \), and \( T \) and maxima at \( i \) and \( k \) :
\[ C_{\uparrow \downarrow} = \{ \mu \in R^T : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_T \} \] (5)
(with at least one strict inequality among each monotone sub-profile). Cyclical profiles may be important in relatively long time-course experiments where the mean expression value could oscillate.

Incomplete inequality profiles :
\[ C_1 = \{ \mu \in R^T : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_i \geq \mu_{i+1} \leq \cdots \leq \mu_T \} \] (6)
(with at least one strict inequality among each monotone sub-profile). Profiles (6) are useful when the investigator is unable to specify inequalities between certain means.

Information-criterion based clustering using order-restricted maximum likelihood

Our procedure seeks to match a gene’s true profile, estimated from the observed data, to one of a specified set of candidate profiles. Provided the relationship of a gene’s mean expression levels at different time points is defined by a given candidate profile, we first obtain the order-restricted maximum likelihood estimates (MLE) of the gene’s mean expression levels at all time points. Details for simple order and umbrella order constraints are given in the Methods section. A general discussion of order-restricted MLE can be found in [29]. Some specific examples can also be found in [1]. Peddada’s method then carried out a bootstrap-based likelihood ratio test to decide a gene’s best matched profile, which needs to repeat computing the order-restricted MLEs at least hundreds of times for each gene. We replace this very time-consuming bootstrap procedure by computing an information criterion function instead. The best matched profile is given by the one with the smallest information criterion function value. Information criterion functions, such as Akaike information criterion (AIC) [30] and Bayesian information criterion (BIC) [31], are widely used in selecting a model from a set of potential models. In our context, each candidate profile can be viewed as a potential model for a gene’s expression pattern, and hence a properly defined information criterion function can be used to decide the gene’s best matched profile. However, the widely used AIC and BIC are designed for models with a fixed number of parameters, whereas the inequality constraints do not really specify any parameters explicitly. Some information criteria for order-restricted inference have been proposed recently [32,33], but they only apply to simple order constraints (3).

We have recently proposed a new order-restricted information criterion (ORIC) function for general inequality profiles, which we show to always select the correct profile when the sample size is large enough (unpublished manuscript). Suppose that there are \( d \) candidate profiles of interest. Let \( I(\lambda) \) denote the maximum log-likelihood under the \( \lambda \)-th candidate profile, and \( v_1(\lambda) \) and \( v_2(\lambda) \) are the number of \( \geq \) and \( \leq \) specified in the profile, \( \lambda = 1, \ldots, d \). Then, the ORIC function is
\[ ORIC(\lambda) = -2I(\lambda) \cdot \log(T \cdot M), \] (7)
where \( M \) is the number of replicated arrays and
\[ p(\lambda) = v_1(\lambda) + \sum_{i=0}^{v_2(\lambda)} \frac{1}{1+i}. \] (8)
The ORIC function is similar to AIC and BIC in essence with \( p(\lambda) \) representing the model complexity. That is, the more the inequality constraints in a profile, the more complex it is as a model. And a profile with a smaller ORIC value is regarded as a better match to the gene’s expression pattern. In the following, we describe our ORIC-based clustering (ORICC) algorithm (one-stage ORICC) and its computationally faster variant (two-stage ORICC).

One-stage ORICC

Step 1. Pre-specify a collection of candidate profiles, \( \{ C_1, \ldots, C_d \} \). To prevent genes with very little changes over time matched to these profiles, we also include \( C_0 \) defined in (1) into the collection.

Step 2. Compute \( p(\lambda) \) in (8) for all candidate profiles.

Step 3. For gene \( g \), compute the order-restricted MLE \( \{ \mu_{g1}, \ldots, \mu_{gT} \} \) of \( \{ \mu_{11}, \ldots, \mu_{1T} \} \) and the maximum log-likelihood \( I(\lambda) \) under each candidate profile \( C_\lambda, \lambda = 0, 1, \ldots, d \).

Step 4. For gene \( g \), compute the information criterion function \( ORIC(\lambda) \) in (7) for all \( \lambda = 0, 1, \ldots, d \). The best
matched profile is then selected as that corresponds to 
\[ \hat{\lambda} = \arg \min_{\lambda \in \mathcal{A}} \text{ORIC}(\lambda), \]
and gene \( g \) is assigned to the \( \hat{\lambda} \)th cluster if \( \hat{\lambda} \neq 0 \).

Step 5. Repeat Steps 3 and 4 for every gene.

Although our one-stage ORICC algorithm is hundreds of times faster than Peddada’s method, performing Step 3 for all genes can still cost a lot of computational time even when only a moderate number of candidate profiles are considered because the number of genes is generally huge. This issue is more imminent for relatively longer time course microarray studies as more candidate profiles usually need to be considered. Next, we propose a computationally more efficient two-stage algorithm by adding a pre-screening stage.

Two-stage ORICC

Step 1. Pre-specify a collection of candidate profiles, \( \{C_1, \ldots, C_d\} \). Here, we also add \( C_0 \) and \( C_{\perp} \) defined in (1) and (2) into the collection for screening purpose.

Step 2. Compute \( p(\hat{\lambda}) \) in (8) for all candidate profiles.

Step 3. For gene \( g \), compute the order-restricted MLE \( (\hat{\mu}_{g1}, \ldots, \hat{\mu}_{gT}) \) of \( (\mu_{g1}, \ldots, \mu_{gT}) \) and the maximum log-likelihood \( I(\hat{\lambda}) \) under profiles \( C_0 \) and \( C_{\perp} \).

Step 4. For gene \( g \), compute \( \text{ORIC}(\hat{\lambda}) \) in (7) for \( \lambda = 0, \perp \). Exclude gene \( g \) for further consideration if \( \text{ORIC}(0) < \text{ORIC}(\perp) \).

Step 5. Repeat Steps 3 and 4 for every gene. Denote the set of remained genes by \( S \).

Step 6. Run Steps 3–5 in the one-stage ORICC algorithm for genes in \( S \) considering only \( \{C_1, \ldots, C_d\} \) as candidate profiles.

In the one-stage algorithm, the ORIC function is evaluated for every gene under every candidate profile, whereas the two-stage algorithm first screens out genes that show no significant changes over time by comparing between two profiles \( C_0 \) and \( C_{\perp} \) and then applies the one-stage algorithm to a much smaller set of remained genes. As a result, the two stage algorithm is usually much faster and report tighter clusters with less genes in them.

Filtering genes with small expression levels

Some genes selected by the ORICC algorithm may have small mean expression levels at every time point. Such genes may not be of interest to some investigators. Peddada et al. [1] suggested a simple step to remove them. We include it here for completeness.

Let

\[ v_g = \frac{1}{T} \sum_{t=1}^{T} (\hat{\mu}_{gt} - \bar{\mu})^2, \quad (9) \]

where \( \bar{\mu} = \frac{1}{T} \sum_{t=1}^{T} \hat{\mu}_{gt} \). Large values of \( v_g \) indicate that the mean expression of gene \( g \) is high for at least one time point. Arrange the genes that selected by the ORICC algorithm in descending order of \( v_g \) and retain the top \( R \) genes.

Assessing the reliability of the oricc results

Microarray data are often noisy and hence it is important to assess the reliability of the clustering results. Among the recently developed methods for assessing clustering reliability [34-38], we adopt a general bootstrap framework proposed by Kerr and Churchill [36], in which the clustering procedure is first applied to the original data and then to a large number of bootstrap samples obtained from perturbing the original data. Since Peddada’s method is computationally so costly, it is impossible to put it into this framework. By contrast, our ORICC method is computationally very efficient, and we can easily embed it into this general framework.

In time-course microarray studies, we can use the following analysis of variance (ANOVA) model to account for sources of variation in microarray data.

\[ y_{gti} = \mu + \alpha_{gi} + \beta_i + \gamma_t + \omega_{gti} + \phi_{gi} + \psi_{ui} + \epsilon_{gti}, \quad (10) \]

where \( y_{gti} \) is the relative expression measurement from array \( i \), time point \( t \), and gene \( g \) on appropriate scale (typically the log scale). The terms \( \beta_i, \gamma_t, \) and \( \psi_{ui} \) account for all effects that are not gene-specific. We assume that the error terms \( \epsilon_{gti} \) are independent with mean 0 and variance \( \sigma^2 \) but do not make any other distributional assumption. The bootstrap assessment is then done as follows.

Step 1. Estimate model (10), which can be done straightforwardly in any statistics software, such as SAS [39] and R [40].

Step 2. Generate \( B \) bootstrap samples by

\[ y_{gti}^* = \bar{\mu} + \tilde{\alpha}_g + \tilde{\beta}_i + \tilde{\gamma}_t + \tilde{\omega}_{gti} + \tilde{\phi}_{gi} + \tilde{\psi}_{ui} + \epsilon_{gti}^*, \]

where a \( ^* \) over a term means the estimate from the original model fit in Step 1, and \( \epsilon_{gti}^* \) are drawn with
replacement from the studentized residuals of the original model fit.

Step 3. Repeat the ORICC algorithm for each bootstrap sample.

Now, the original clustering is accompanied by a collection of bootstrap clusterings, which can be regarded as a sample of clusterings that are close to the original clustering in space of all possible clusterings. When the level of noise in the original data is low, the bootstrap clusterings tend to be more like the original clustering. Then we can calculate a reliability measure for each gene by counting the proportion in the bootstrap clusterings it is attached to the same profile as in the original clustering. The larger the measure, the more reliable the gene’s clustering membership.

**Simulation studies**

In this section, we use Monte Carlo simulation to examine the performance of the ORICC method and compare it with other clustering methods for short time-course microarray data, including Peddada’s method, STEM [2], and Wang’s method [3].

The STEM algorithm works by assigning genes to a predefined set of model profiles that capture the potential distinct patterns that can be expected from the microarray experiment. Each gene is then assigned to the closest model profile in certain distance measure, e.g. correlation, and genes assigned to the same model profile consist a cluster. Significant profiles/clusters are next determined by hypotheses tests. As a result, genes in insignificant clusters are usually not reported. Wang’s method represents each gene’s temporal profile by a polynomial model and estimates the model using a Bayesian approach. A heuristic search strategy [23] is then applied to obtain clusters by repeatedly merging models to improve marginal likelihood.

All simulations were carried out on a workstation with a 2.30 GHz AMD Athlon(tm) 64 × 2 Dual Core 4400+ processor and a 2.00 GB memory. Peddada’s method, Wang’s method and the one-stage ORICC algorithm are implemented in R [40], whereas the STEM software is in JAVA written by its author.

**Simulation 1**

In the first simulation study, we consider ten inequality profiles (C1–C10) plus a flat pattern (C0) to represent a total number of eleven clusters. We set the number of time points as T = 6. These numbers are set to be similar to those in the real data set analyzed in the next section. The eleven profiles are specified as follows. For compactness, we drop μ ∈ RT and the phrase ‘with a strict inequality’ when defining the profiles. True values of μ = (μ1, μ2, μ3, μ4, μ5, μ6) in each profile are also given.

\[ C0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6, \quad \mu = (0,0,0,0,0,0); \]

\[ C1: \mu_1 \geq \mu_2 \geq \mu_3 \geq \mu_4 \geq \mu_5 \geq \mu_6, \quad \mu = (0,-0.5,-1,-1.5,-2,-2.5); \]

\[ C2: \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4 \leq \mu_5 \leq \mu_6, \quad \mu = (0,0.5,1,1.5,2,2.5); \]

\[ C3: \mu_1 \leq \mu_2 \geq \mu_3 \geq \mu_4 \geq \mu_5 \geq \mu_6, \quad \mu = (0,0.5,0,0.5,1,1.5); \]

\[ C4: \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4 \geq \mu_5 \geq \mu_6, \quad \mu = (0,0.5,0,0.5,1,1.5); \]

\[ C5: \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4 \leq \mu_5 \geq \mu_6, \quad \mu = (0,0.5,1,1.5,1,0.5); \]

\[ C6: \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4 \leq \mu_5 \leq \mu_6, \quad \mu = (0,0.5,1,1.5,2,1.5); \]

\[ C7: \mu_1 \geq \mu_2 \geq \mu_3 \leq \mu_4 \leq \mu_5 \leq \mu_6, \quad \mu = (0,-0.5,0,0.5,1,1.5); \]

\[ C8: \mu_1 \geq \mu_2 \geq \mu_3 \leq \mu_4 \leq \mu_5 \leq \mu_6, \quad \mu = (0,-0.5,-1,-0.5,0,0.5); \]

\[ C9: \mu_1 \geq \mu_2 \geq \mu_3 \geq \mu_4 \leq \mu_5 \leq \mu_6, \quad \mu = (0,-0.5,-1,-1.5,0,0.5); \]

\[ C10: \mu_1 \geq \mu_2 \geq \mu_3 \geq \mu_4 \geq \mu_5 \leq \mu_6, \quad \mu = (0,-0.5,-1,-1.5,-2,-1.5); \]

Figure 1 shows the inequality profiles C1–C10. We generated a data set with 200 genes from each profile.

![Figure 1](http://www.biomedcentral.com/1471-2105/10/146)

**Figure 1**

Ten inequality profiles in Simulations 1 and 2.
At each time point $t$, we generated $M$ replicates for each gene’s expression level from normal distributions with means $\mu_t$ and variance $\sigma^2$. To assess the effect of the data variability and replicates on the clustering results, we varied the variance $\sigma^2$ from 0.2 to 1.2 (by an incremental of 0.2) and the number of replicates $M$ from 2 to 10.

Next, we clustered the simulated data using Peddada’s method and the one-stage ORICC algorithm, considering ten candidate inequality profiles $C_1$–$C_{10}$. For Peddada’s method, we set the number of bootstrap replications as 200 and the significance level of the bootstrap based test as 0.025. Peddada et al. originally proposed to use significance level 0.0025, and we have observed that a large number of “non-flat” genes will be clustered to $C_0$ using this choice. Meanwhile, using the common significance level of 0.05 tends to cluster many genes from $C_0$ to other “non-flat” profiles. Using significance level of 0.025 offers a good compromise between the two kinds of false clustering.

Let $g_i$ denote the number of genes with true profile $C_i$ and correctly clustered to profile $C_i$, $i=0, 1, \ldots, 10$. The overall error rate and the false positive rate are then given by $1 - \sum_{i=0}^{10} g_i/(11 \times 200)$ and $1 - g_0 = 200$, respectively. Let $\gamma_i$ denote the number of genes with true profile $C_i$.

**Figure 2**
Simulation 1: The overall error rate of Peddada’s method and the one-stage ORICC algorithm. The horizontal axis represents the number of replicates, and the vertical axis represents the overall error rate. Dashed lines are for the one-stage ORICC algorithm, and solid lines are for Peddada’s method.
\( i = 1, \ldots, 10 \) and clustered to profile \( C_0 \). The **false negative rate** is then given by \( \sum_{i=1}^{10} r_i^+ / (10 \times 200) \).

We then use the overall error rate, the false positive rate and the false negative rate to evaluate the accuracy of the two algorithms. Simulation results are summarized in Figures 2, 3 and 4. Figure 2 and Figure 4 show that the overall error rate and the false negative rate of the one-stage ORICC algorithm are always better than those of Peddada’s method. Figure 3 shows that the false positive rate of the one-stage ORICC algorithm is better than that of Peddada’s method in most cases. The one-stage ORICC algorithm not only provides good clustering accuracy but also is much faster than Peddada’s method. For example, when \( \sigma^2 = 1 \) and \( M = 5 \), the run time for Peddada’s method and one-stage ORICC algorithm is 2979.29 seconds versus 25.55 seconds.

**Simulation 2**

In the second simulation study, we consider the same set of inequality profiles and simulate the data in the same way as in Simulation 1, but we fix the number of replicates \( M \) to be 5. To study the effect of the true cluster size to the clustering accuracy, we consider different cluster sizes, 50, 100, 150 and 200. Meanwhile, we also vary the variance \( \sigma^2 \) from 0.2 to 3.0

---

**Figure 3**

**Simulation 1:** The false positive rate of Peddada’s method and the one-stage ORICC algorithm. The horizontal axis represents the number of replicates, and the vertical axis represents false positive rate. Dashed lines are for the one-stage ORICC algorithm, and solid lines are for Peddada’s method.
with an incremental of 0.4. Then we cluster the simulated
data set using methods including Peddada's method, Wang's
method, STEM and the one-stage ORICC algorithm.

For Peddada's method and the one-stage ORICC algo-
rithm, we consider eleven candidate inequality profiles
$C_0$–$C_{10}$. For Wang's method, we set the prior hyper-
parameters $(\alpha_1, \alpha_2)$ in the gamma prior distribution
Gamma$(\alpha_1, \alpha_2)$ as $(2, 2)$. For STEM, we assume 50
possible profiles and use the recommended default
settings in the package. To be consistent, we did not
filter out any genes in any of these analyses. Then we use
Rand's C statistic [41] to evaluate the similarity between
the true cluster assignment and the clustering results of
different methods. Rand's C statistic is defined as
follows. Given a pair of clusterings $C$ and $C'$ of the
same $N$ objects, arbitrarily number the clusters in each
clustering and let $n_{ij}$ be the number of objects
simultaneously in the $i$th cluster of $C$ and the $j$th cluster
of $C'$. Then, Rand's C statistic is given by

$$
\text{Rand's C} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} n_{ij}^2 - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} n_{ij}^2 + \frac{1}{2} \sum_{i=1}^{N} n_i^2 + \frac{1}{2} \sum_{j=1}^{N} n_j^2}{\binom{N}{2}}
$$

Figure 4

Simulation 1: The false negative rate for Peddada's method and the one-stage ORICC algorithm. The horizontal
axis represents the number of replicates, and the vertical axis represents the false negative rate. Dashed lines are for the one-
stage ORICC algorithm, and solid lines are for Peddada's method.
Figure 6 shows the simulated eleven clusters when $s$ is relatively small and the other methods the best, and Peddada's method and the one-stage ORICC algorithm perform the worst. Wang's method keeps eight clusters. In particular, Wang's method did not cluster genes from the flat profile $C_0$ into one cluster but assign them into different clusters.

In this simulation, Peddada's method, Wang's method and the one-stage ORICC method are implemented in R, whereas the STEM software is in JAVA written by its author. So, we can only compare the computational efficiency of the first three methods and the one-stage ORICC method is much faster than the other two. For example, when $\sigma^2 = 3$ and the cluster size is 200, the run time for Peddada's method, Wang's method and the one-stage ORICC algorithm is 3073.37 seconds, 10303.9 seconds and 24.72 seconds, respectively.

### Simulation 3

In the third simulation, we examine the robustness of the ORICC algorithm. We consider eleven inequality profiles $(C_0-C_{10})$ plus a cyclical profile $(C_{\Lambda\Lambda})$ to represent a total number of twelve clusters. We set the number of time points as $T = 6$. The eleven inequality profiles $C_0-C_{10}$ and the true values of $\mu = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6)$ in each profile are the same as in Simulation 1.

The cyclical profile $C_{\Lambda\Lambda}$ and the true value of $\mu = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6)$ in $C_{\Lambda\Lambda}$ are given as follows:

$$C_{\Lambda\Lambda}: \mu_1 \leq \mu_2 \geq \mu_3 \leq \mu_4 \geq \mu_5 \leq \mu_6, \quad \mu = (-0.5, 0.5, -0.5, 0.5, -0.5, 0.5).$$

We generate a data set with 200 genes from each profile of $C_0-C_{10}$ and $200 \times r$ genes from cyclical profile $C_{\Lambda\Lambda}$. At each time point $t$, we generated 5 replicates for each gene's expression level from normal distributions with means $\mu_t$ and variance $\sigma^2$. To study the robustness of the one-stage ORICC algorithm, we consider different cluster sizes, $200 \times r, r = 1, 2, \ldots, 10$, for the cyclical profile $C_{\Lambda\Lambda}$. Meanwhile, we also vary the variance $\sigma^2$ from 0.2 to 3.0 with an increment of 0.4. Then we cluster the simulated data set using the one-stage ORICC algorithm. For the one-stage ORICC algorithm, we consider eleven candidate inequality profiles $C_0 - C_{10}$. Note that the cyclical profile $C_{\Lambda\Lambda}$ is not included in the candidate profiles.

Then we use Rand's C statistic to evaluate the similarity between the true cluster assignment and the clustering results from ORICC. Figure 11 gives Rand's C statistic from the one-stage ORICC algorithm for different $\sigma^2$ and cluster sizes of the cyclical profile $C_{\Lambda\Lambda}$. It shows that the precision of the one-stage ORICC algorithm as measured by Rand's C statistic increases for decreasing variance, and decreases when increasing the cluster size of the cyclical profile $C_{\Lambda\Lambda}$. The cluster precision is however...
always greater than 80%, thus suggesting that the one-stage ORICC algorithm is very stable.

Simulation 4
In the fourth simulation, we further examine the robustness of the ORICC algorithm. We consider the scenario where a true profile is not explicitly included in the candidate profiles but maybe viewed as a special case of a more flexible candidate profile. Let the true inequality profile be

\[ C_{\alpha} : \mu_1 \leq \mu_2 \geq \mu_3 \geq \mu_4 \leq \mu_5, \quad \mu = (-0.5, 1, -0.5, 0.5, -0.5, 0.5). \]
We generate a data set containing 2000 genes from this profile. At each time point \( t \), we generate \( M \) replicates for each gene’s expression level from normal distributions with means \( \mu_t \) and variance 0.5. Then, we consider candidate profiles being \( C_1, C_2, C_3 \) and \( C_4 \) plus the profile \( C_\perp \) and cluster the simulated data using the one-stage ORICC algorithm. Note that the set of candidate profiles does not contain the true one \( C_\wedge \), but \( C_\perp \) may be viewed as a special case of \( C_\wedge \). Let \( \gamma_\wedge \) denote the proportion of genes clustered to the profile \( C_\wedge \) and define the detection error as \( 1 - \gamma_\wedge \). The simulation results are summarized in Figure 12. It shows that the detection error decreases.
rapidly as the number of replicates $M$ increases. With 5 replicates, the detection error is below 30%, and with 10 replicates, it goes down to below 5%. This indicates that it is quite safe to apply the ORICC algorithm even if some true profile is missing from the candidate profiles but a more comprehensive profile is considered.

**Application to breast cancer cell line data**

Next, we apply the ORICC algorithm to log-transformed relative expression data from a breast cancer cell line microarray study in [28]. The same data set was also analyzed by Peddada *et al.* [1]. The experiment was done as follows. First, the MCF-7 breast cancer cell line was treated

---

**Figure 8**

*Simulation 2: Temporal profiles for clusters from Peddada’s method.*
with 17β-estradiol or ethanol (vehicle control). Then, samples were harvested at 1, 4, 12, 24, 36 and 48 hours after treatment. At each time point, \( M = 8 \) replicate arrays were prepared with each array consisting of \( G = 1901 \) genes. Similar as in [1], we assumed for each gene that the variance of the log relative expression was homoscedastic over time, and consider the following 10 candidate profiles for clustering: monotone decreasing, \( C_1 \); monotone increasing, \( C_2 \); four up-down profiles with maxima at 4, 12, 24, 36 hours, \( C_3 - C_6 \), respectively; and 4 down-up profiles with minima at 4, 12, 24, 36 hours, \( C_7 - C_{10} \), respectively. Genes matched to these profiles will be regarded estrogen responsive.

The original analysis in [28] used a simple confidence interval approach [42], and identified 105 genes that

Figure 9
Simulation 2: Temporal profiles for clusters from Wang’s method.
demonstrated estrogen responsive expression. From the one-stage ORICC analysis, we finally had 981 genes in the 10 clusters, with 68 in $C_1$, 24 in $C_2$, 76 in $C_3$, 44 in $C_4$, 97 in $C_5$, 72 in $C_6$, 35 in $C_7$, 98 in $C_8$, 409 in $C_9$, and 58 in $C_{10}$. Due to limitation of space, we only present the top 50 genes ranked by the filtering criterion in (9) (Additional file 1). The last column in Additional file 1 indicates whether the gene was previously identified in [28]. Clustering reliability for each gene is also attached. Note that these 50 genes come from only nine of the 10 clusters, with none from $C_2$. Figure 13 presents the estimated profiles of these 50 genes from the order-restricted MLE. Of the 105 genes identified in [28], 44 are among our top 50, 82 are among our top 100, 94 were among our top 150, and 101 were among our top 200. Most of the 44 genes in the top 50 selected in common are involved in cell cycle progression and DNA replication reflecting the known sensitivity of MCF-7 cells to estrogen. Among the six genes identified by our ORICC algorithm but not in [28] (denoted by dashed lines in Figure 13), two have quite high clustering reliability. The methylmalonyl Coenzyme A mutase (Clone ID 35468 in $C_9$) has reliability 0.9967, and the deoxythymidylate kinase (Clone ID 489092 in $C_5$) has reliability 0.7800. Both genes are known specific to the metabolic process, and hence are very likely responsive to the metabolism of estrogen when overdosed estrogen are supplied to the cell. For example, the methylmalonyl Coenzyme A mutase could be involved in the breakdown of estradiol into smaller metabolic fragments. However, this gene was not reported in the top 50 list by Peddada et al. [1]. Estimated profiles in Figure 13 suggests this gene matches very well with the candidate profile $C_9$. An interesting phenomenon about the deoxythymidylate kinase is that this gene actually corresponds to two spots on the microarray chips (Clone IDs 489092 and 248008).
The original analysis in [28] was only able to identify one of them (Clone ID 248008) but not the other, whereas our method identifies both in the same cluster with high clustering reliability.

In [1], Peddada et al. reported that 39 genes in the top 50 were in common with the 105 genes selected in [28]. Our ORICC result has a good overlap with Peddada's result, with 32 genes in our top 50 also in the top 50 list reported in [1]. Peddada et al. discussed in details several newly identified genes in their top 50, such as replication factors C4 and C5. Our ORICC analysis also cluster these genes into the same clusters, but they are not in our top 50 list due to relatively low clustering reliability, with replication factor C4 ranked 75 and replication factor C5 ranked 148.

We further applied STEM and Wang's method on the breast cancer cell line data. Table 1 reports Rand's C statistics among results from four clustering methods. It shows that the three profile matching algorithms have results more alike each other, while the unsupervised Wang's method is

**Figure 13**
Breast cancer cell line data: Estimated profiles of the top 50 genes. Curves are given by the order-restricted MLE of mean log expression ratios. Dashed lines indicate newly identified genes.

**Table 1: Rand's C statistics among four clustering methods in the breast cancer cell line example.**

<table>
<thead>
<tr>
<th></th>
<th>ORICC</th>
<th>Peddada</th>
<th>Wang</th>
<th>STEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORICC</td>
<td>1.0000</td>
<td>0.7767</td>
<td>0.6313</td>
<td>0.7142</td>
</tr>
<tr>
<td>Peddada</td>
<td>0.7767</td>
<td>1.0000</td>
<td>0.5948</td>
<td>0.7694</td>
</tr>
<tr>
<td>Wang</td>
<td>0.6313</td>
<td>0.5948</td>
<td>1.0000</td>
<td>0.6025</td>
</tr>
<tr>
<td>STEM</td>
<td>0.7142</td>
<td>0.7694</td>
<td>0.6025</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

A larger value indicates more overlap between two clustering results.
less similar to the rest. This observation can also be seen in the temporal profiles of the clusters given by four different methods (Figures 14, 15, 16 and 17). While ORICC and Peddada’s method give ten clusters plus the ‘flat’ cluster, Wang’s method identifies seven clusters, and STEM reports twelve significant clusters.

Discussion

In time-course microarray experiments, the ability to exploit the temporal ordering information may be especially valuable because genes whose expression levels change over time may be involved in the same cellular process or belong to the same regulatory pathway. Making use of the valuable ordering information can improve inference. Our proposed ORICC algorithm utilizes the temporal ordering information in clustering time-course microarray data using order-restricted maximum likelihood, while most existing clustering methods either cannot incorporate the temporal information or require long time series to perform reliable nonparametric smoothing.

Figure 14
Breast cancer cell line data: Temporal profiles of clusters from the ORICC analysis. Curves are given by connecting the observed log expression ratios at different time points.
e.g. spline smoothing, though most time-course microarray data are short time series. In our method, the temporal ordering information is exploited through a set of pre-defined candidate expression profiles given by inequality constraints among the mean expression levels at different time points. By viewing each candidate profile as a potential model for the data and let each profile represent a cluster, we transform the clustering problem into a model selection problem. Using an ORIC function, we decide the best matched profile for each gene and hence determine the gene’s clustering membership. Peddada et al. [1] instead performed a likelihood based test to decide the best matched profile. However, a bootstrap procedure is needed to decide the threshold
value on the test statistic, which makes Peddada’s method computationally very intensive.

In many situations, field researchers can have good ideas on defining the inequality profiles. For example, when studying the gene expression patterns for disease onset. It is easy to postulate that gene expressions tend to go up before the disease onset and then go down after certain treatment is given. So, the inequality constraints allow an easy adoption of a prior knowledge into the analysis, whereas existing methods usually can not take such information into consideration. In addition, when inequality constraints are given, the order-restricted MLE has some optimal properties and universally dominates the unrestricted MLE [43]. Moreover, the candidate profiles are defined only based on ranks instead of the numerical value of the mean expression levels, hence our model specification is robust to small perturbation in the data. This feature is especially valuable in microarray studies since it is well known that microarray data are quite noisy. Furthermore, the rank-based specification is often closer to the real meaning of ‘coexpression’ that refers to two genes’ expression levels changing in the same direction instead of with the same magnitude.

In this paper, we present our algorithm under the context of clustering time-course microarray data. Actually, it can be applied to data from any experiment with ordered treatment or conditions, such as dose-response microarray experiments where the dose levels provide the ordering.

Our current ORICC algorithm is based on order-restricted MLE for gene expressions with a constant variance through time. It can be generalized to handle situations where the variances change or are subject to order restrictions themselves. In such situations, the estimation of mean expression levels outlined in this paper can be modified according to the approach in [44]. However, it remains a subject for future investigation to modify the definition of the model complexity $p(\lambda)$ (8) accordingly.

**Conclusion**

We developed a new clustering algorithm, ORICC, for short time-course microarray data, by taking a model selection approach in order-restricted statistical inference. Our method clusters genes into clusters represented by candidate profiles defined through inequalities among mean expression levels at different time points. A newly proposed information criterion function is used to determined the cluster assignment. Compared with a previous clustering method by Peddada et al. [1] that also uses order-restricted inference, our method is computationally much more efficient and provides an assessment of clustering reliability.
Simulation studies indicate that the ORICC algorithm possesses good clustering accuracy when a moderate number of replicate arrays are available, and competes well with other existing clustering methods, such as [2] and [3]. Real data applications also indicate that our method can identify interesting genes that some correlation-based methods have failed to identify.

**Methods**

**Order-restricted maximum likelihood estimation**

Here, we briefly present the order-restricted MLE under simple order (3) and umbrella order constraints (4), which are needed for ORICC analysis in our simulation and real data example. For more general results, we refer to [29,45]. Suppose that $y_{ti}$’s independent observations from normal distributions with unknown means $\mu$ and variances $\sigma^2$ for $t = 1, ..., T$ and $i = 1, ..., n_t$. Then the data log-likelihood is

$$l(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \ln n_t - \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \frac{(y_{ti} - \mu_t)^2}{\sigma^2}.$$  \hspace{1cm} (11)

Where $n = \sum_{t=1}^{T} n_t$, $\mu = (\mu_1, ..., \mu_T)^T$ and $\sigma^2 = (\sigma^2_1, ..., \sigma^2_T)^T$. Let $\hat{\mu}_t = \hat{\mu}_t$ and $\hat{\sigma}^2 = \hat{\sigma}^2$.

When assuming the variances $\sigma^2$ are known, we have the order-restricted MLE of $\mu$ as

1. $\mu_1 \leq \mu_2 \leq ... \leq \mu_T$:  
   $$\hat{\mu}_t = \max_{a \leq b \leq T} \hat{\mu}_{a,b}, \quad t = 1, 2, ..., T;$$

2. $\mu_1 \geq \mu_2 \geq ... \geq \mu_T$:  
   $$\hat{\mu}_t = \min_{a \leq b \leq T} \hat{\mu}_{a,b}, \quad t = 1, 2, ..., T;$$

3. $\mu_1 \leq ... \leq \mu_h \geq ... \geq \mu_T$:  
   $$\hat{\mu}_h = \max_{a,b; a \leq b \leq h} \hat{\mu}_{a,b};$$
   $$\hat{\mu}_t = \min \{\hat{\mu}_h, \min_{a; a \leq b \leq h-1} \hat{\mu}_{a,b}\}, \quad t = 1, 2, ..., h-1;$$
   $$\hat{\mu}_t = \min \{\hat{\mu}_h, \max_{a; a \leq b \leq h} \hat{\mu}_{a,b}\}, \quad t = h+1, ..., T;$$

4. $\mu_1 \geq ... \geq \mu_h \leq ... \leq \mu_T$:  
   $$\hat{\mu}_h = \min_{a,b; a \leq b \leq h} \hat{\mu}_{a,b};$$
   $$\hat{\mu}_t = \max \{\hat{\mu}_h, \min_{a; a \leq b \leq h-1} \hat{\mu}_{a,b}\}, \quad t = 1, 2, ..., h-1;$$
   $$\hat{\mu}_t = \max \{\hat{\mu}_h, \max_{a; a \leq b \leq h} \hat{\mu}_{a,b}\}, \quad t = h+1, ..., T;$$

The maximum log-likelihood under the $\lambda$-th candidate profile is then obtained by plugging in the corresponding order-restricted MLE into (11).

If the variances $\sigma^2$ are unknown, we need to impose the assumption $\sigma^2_1 = \sigma^2_2 = ... = \sigma^2_T = \hat{\sigma}^2$. In microarray data, this assumption is reasonable after the data are properly normalized. Now, the data log-likelihood is

$$l(\mu, \hat{\sigma}^2) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \frac{(y_{ti} - \hat{\mu}_t)^2}{\hat{\sigma}^2}.$$  \hspace{1cm} (12)

Again, the maximum log-likelihood for the $\lambda$-th candidate profile is then obtained by plugging in the corresponding order-restricted MLE into (12).

**Availability and requirements**

We have implemented ORICC in an R program, which can be downloaded from http://www.math.wustl.edu/~nlin/oricc/.

**Authors’ contributions**

NL, NS and BZ had the initial idea and initiated the study. NL and TL conducted the data analyses, created all tables and figures, and prepared the manuscript under the supervision of BZ. All authors read and approved the final manuscript.

**Additional material**

**Additional File 1**

*Gene clusters from the ORICC analysis of the breast cancer cell line data.* This table presents gene clusters given by the ORICC algorithm using ten candidate inequality profiles for the breast cancer cell line microarray data in [28]. The fourth column gives the rank of the gene according to the filtering criterion (9), and numbers in the parentheses are clustering reliability computed based on 300 bootstrap samples. The last column shows whether the gene was identified by the original analysis in [28].

Click here for file [http://www.biomedcentral.com/content/supplementary/1471-2105-10-146-S1.pdf]

**Acknowledgements**

The authors gratefully appreciate the editor and the referees for their valuable comments and suggestions. The authors are partly supported by the National Science Foundation of China (No.10871037) and the Science Foundation for Young Teachers of Northeast Normal University (No.20050107).
References


Publish with BioMed Central and every scientist can read your work free of charge

“BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime.”

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp