BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 17 2007, pages 2247–2255
doi:10.1093/bioinformatics/btm320

Gene expression

Penalized and weighted *K*-means for clustering with scattered objects and prior information in high-throughput biological data

George C. Tseng*
Department of Biostatistics, University of Pittsburgh, Pittsburgh, USA

Received on February 27, 2007; revised on May 11, 2007; accepted on June 8, 2007
Advance Access publication June 27, 2007
Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: Cluster analysis is one of the most important data mining tools for investigating high-throughput biological data. The existence of many scattered objects that should not be clustered has been found to hinder performance of most traditional clustering algorithms in such a high-dimensional complex situation. Very often, additional prior knowledge from databases or previous experiments is also available in the analysis. Excluding scattered objects and incorporating existing prior information are desirable to enhance the clustering performance.

Results: In this article, a class of loss functions is proposed for cluster analysis and applied in high-throughput genomic and proteomic data. Two major extensions from *K*-means are involved: penalization and weighting. The additive penalty term is used to allow a set of scattered objects without being clustered. Weights are introduced to account for prior information of preferred or prohibited cluster patterns to be identified. Their relationship with the classification likelihood of Gaussian mixture models is explored. Incorporation of good prior information is also shown to improve the global optimization issue in clustering. Applications of the proposed method on simulated data as well as high-throughput data sets from tandem mass spectrometry (MS/MS) and microarray experiments are presented. Our results demonstrate its superior performance over most existing methods and its computational simplicity and extensibility in the application of large complex biological data sets.

Availability: http://www.pitt.edu/∼ctseng/research/software.html

Contact: ctseng@pitt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Cluster analysis has been widely applied in various unsupervised data mining problems including microarray analysis, sequence analysis, image segmentation and marketing research. The general problem considers clustering data $X=\{x_1, \dots, x_n\}$ into k clusters, where each object x_i is d-dimensional and k is estimated a priori. The large variety of available methods in the literature seems like a jungle, e.g. Bishop (1995), Gordon (1999), Grabmeier and Rudolph (2002), Jain and Dubes (1988), Jobson (1992), Kaufman and Rousseeuw (1990), McLachlan and Basford (1987), Ripley (1996) and Spaeth (1984), to mention only a few. Although tremendous efforts have been made to benchmark clustering algorithms, it is generally agreed that clustering evaluation measures and performance of each method heavily depend on the original application problem (Messatfa and Zait, 1997; Milligan and Cooper, 1985). In this article, we focus our discussion and comparison on high-throughput biological data, especially in the analysis of microarray expression profiles and mass spectrometry proteomic data. In the literature, classical clustering methods, including hierarchical clustering (Eisen *et al*., 1998), *K*-means and its variants, self-organizing maps (Conrads *et al*., 2003; Tamayo *et al*., 1999) and mixture model approaches (McLachlan *et al*., 2002; Yeung *et al*., 2001), as well as modern techniques, such as gene shaving (Hastie *et al*., 2000), a graph-theoretical method CLICK (Sharan *et al*., 2003) and tight clustering (Tseng and Wong, 2005), have been widely applied. A recent comparative study for gene clustering in expression profiles (Thalamuthu *et al*., 2006) suggests that clustering methods allowing scattered objects not being clustered, with explicit or implicit model assumptions, and with resampling evaluations seem to perform better.

Many clustering methods are based on global optimization of a criterion that measures compatibility of the clustering result to the data. *K*-means and mixture Gaussian model-based clustering are examples of this category. In the following paragraphs, we will elucidate the connections between the two methods and introduce the motivation for our proposed method. In statistical literature, clustering is often obtained through likelihood-based inference including the mixture maximum-likelihood (ML) approach and the classification maximum-likelihood (CML) approach (Celeux and Govaert 1992; Ganesalingam 1989). In the ML approach, the R^d-valued vectors $x_1, \dots, x_n$ are sampled from a mixture of densities, $f(x)=\sum_{j=1}^{k}\pi_j f(x|\theta_j)$, where π_j is the probability that the data is generated from cluster j and each $f(x|\theta_j)$ is the density for cluster j from the same parametric family with parameter θ_j. The log-likelihood to be maximized is

$$L=\log\left\{\prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j f\left(x_i|\theta_j\right)\right\}$$

and clustering is obtained by the assignment of each x_i to the cluster with greatest posterior probability. For more details refer to Fraley and Raftery (2002) and McLachlan *et al*. (2002).

*To whom correspondence should be addressed.

In the CML approach, the partition $C = \{C_i, \ldots, C_k\}$, where C_j's are disjoint subsets of $X = \{x_i, \ldots, x_n\}$, is considered as an unknown parameter and is directly pursued in the optimization. This criterion samples data of $n_1, \ldots, n_k$ observations in each cluster, where n_j's are fixed and unknown and $\sum n_j = n$. Following the convention of Celeux and Govaert (1992), it takes the form

$$C(C,\theta) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \log f\left(x_i \mid \theta_j\right).$$

Throughout this article we restrict f to be Gaussian distributed with $\theta_j = (\mu_j, \Sigma_j)$ and the CML criterion becomes

$$C(C,\theta) = \log f(X \mid C,\theta) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \log f(x_i \mid \mu_j, \Sigma_j) \qquad (1)$$

where $f(x_i \mid \mu_j, \Sigma_j)$ is the Gaussian distribution.

In addition to likelihood-based inference, many clustering methods have utilized heuristic global optimization criteria. *K*-means (Hartigan and Wong, 1979) is an effective clustering algorithm in this category and is applied in many applications due to its simplicity. In the *K*-means criterion, objects are assigned to clusters so that the within cluster sum of squared distance is minimized. It will be shown later (Example 1 in Section 2.1) that *K*-means is actually a simplified form of the CML sampling scheme under the Gaussian assumption. In this article, we propose a class of loss functions extended from *K*-means, namely penalized and weighted *K*-means (PW-*K*-means). A penalty term is added to allow clustering with scattered objects not being clustered and a weight term is introduced to incorporate prior information.

In the analysis of high-throughput biological data, the importance of allowing a set of scattered objects in cluster analysis has been demonstrated (Dasgupta and Raftery, 1998; Fraley and Raftery, 2002; Thalamuthu *et al.*, 2006; Tseng and Wong, 2005). The resulting clustering assignment is represented as $C = \{C_1, \ldots, C_k, S\}$, where clusters C_j are disjoint subsets of X, S is the set of scattered objects and $X = (\cup_{j=1}^{k} C_j) \cup S$. In Section 2.1, it will be shown that the penalty term in PW-*K*-means automatically assigns outlying objects to the S set and the formulation is equivalent to the CML model with a noise set uniformly distributed over the space. For prior information incorporation, most existing methods applied to high-throughput biological data ignored such information in the process of clustering except for a few recent papers: Hanisch *et al.* (2002), Cheng *et al.* (2004), Pan (2006) and Huang and Pan (2006). Compared to these approaches, PW-*K*-means provides a more general and flexible formulation to incorporate prior information in cluster formation. It should be noted that the prior information incorporation we discuss here is different from the well-established field, semi-supervised machine learning (Basu *et al.*, 2004; Pan *et al.*, 2006; Segal *et al.*, 2003). In semi-supervised machine learning, a subset of objects come with 'known' class labels (supervised) while the remaining objects have unknown class labels (unsupervised). The algorithm determines whether the unlabeled objects should be assigned to existing classes or a new cluster(s) should be formed. In prior information incorporation for clustering, the problem is more from Bayesian perspective. Groups of objects (e.g. genes of the same functional annotation) are suggested a priori that they 'likely' but not always cluster together in the data. This is especially true in microarray analysis because gene functional annotations often contain many errors or genes of the same function may not reflect coregulation in the data. The prior information should be used to enhance, instead of force, cluster formation.

The article is structured as the following. In Section 2.1, formulation of PW-*K*-means is described. Insights and properties of the method are discussed. Three examples of *K*-means and *K*-medoids, penalized *K*-means (P-*K*-means) without the weighting term, and an explicit formulation of PW-*K*means for gene clustering in microarray data are then presented. Section 2.2 discusses computational issues of the method including implementation and parameter selection. In Section 3, two sets of simulated data, an application to tandem mass spectrometry data, and an application to gene clustering of microarray data are explored to demonstrate our proposed method. Section 4 contains conclusion and discussion.

2 METHODS

2.1 Formulation of general PW-*K*-means

A general class of loss function extended from *K*-means is proposed for clustering purposes:

$$W(C; k,\lambda) = \sum_{j=1}^{k} \sum_{x_i \in C_j} w(x_i; \mathrm{P}) \cdot d(x_i, C_j) + \lambda |S| \qquad (2)$$

where $w(\cdot;\cdot)$ is a function of the weighting factor, P is the prior information available, $d(x,C_j)$ calculates the dispersion of point x in cluster C_j, $|\cdot|$ represents size of the set and λ is a tuning parameter representing the degree of penalty of each noise point. The weighting factor $w(\cdot;\cdot)$ is used to incorporate prior knowledge of preferred or prohibited patterns of cluster selections. Minimizing Equation (2) with given weight function $w(\cdot;\cdot)$, distance measure $d(\cdot;\cdot)$, k and λ produces a clustering solution. We denote by $C^*(k,\lambda) = \{C_1^*(k,\lambda), \ldots, C_k^*(k,\lambda), S^*(k,\lambda)\}$ the minimizer of $W(C;k,\lambda)$. Proposition 1 shows several useful properties of this formulation. In particular, λ is inversely related to the number of scattered objects (i.e. $|S|$). This is a desirable property to control tightness of the resulting clusters in practice. Proof of Proposition 1 is left to Supplementary Material.

Proposition 1. (a) Similar to *K*-means, if $k_1 > k_2$, then $W(C^*(k_1, \lambda); k_1, \lambda) \le W(C^*(k_2, \lambda); k_2, \lambda)$. (b) If $\lambda_1 > \lambda_2$, then $|S^*(k, \lambda_1)| \le |S^*(k, \lambda_2)|$. (c) If $\lambda_1 > \lambda_2$, then $W(C^*(k, \lambda_1); k, \lambda_1) > W(C^*(k, \lambda_2); k, \lambda_2)$.

In the remaining of this subsection, three specific examples are illustrated:

Example 1: *K*-means and *K*-medoids. The *K*-means algorithm has been widely used in various applications of clustering (Hartigan and Wong, 1979). The loss function to be minimized is:

$$W_{K-\mathrm{means}}(C; k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \bar{x}^{(j)} \right\|^2 \qquad (3)$$

where $\bar{x}^{(j)}$ is the center of cluster C_j and $\|\cdot\|$ is the usual Euclidean distance. It is easily seen that *K*-means is a simple special case of Equation (2) when $w(x) = 1$, $\lambda \to \infty$ and $d(x_i, C_j) = ||x_i - \bar{x}^{(j)}||^2$. When the data is not in Euclidean space or a certain metric other than Euclidean distance is preferred in the application content, *K*-medoids is often used instead of *K*-means. The criterion is similar to *K*-means except that $d(x_i, C_j)$ is replaced by $||x_i - \tilde{x}^{(j)}||^2$, where $\tilde{x}^{(j)} \in C_j$ is the median point such that sum of squared distances of all points in C_j to this point is minimized.

It is known that the *K*-means algorithm is equivalent to maximizing the classification likelihood under a mixture of identical spherical Gaussian distributions (Proposition 1 in Celeux and Govaert, 1992); namely when $\Sigma_j = \sigma_0^2 I$ $(j = 1, \ldots, k)$, maximizing (1) is equivalent to minimizing (3).

Example 2: P-*K*-means. In this example we discuss a version of penalized *K*-means:

$$W_{\mathrm{P}}(C; k,\lambda_0) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \bar{x}^{(j)} \right\|^2 + \eta^2 \cdot \lambda_0 |S| \qquad (4)$$

where λ_0 is a tuning parameter, $\eta = H/\sqrt[d]{k}$, H is defined as the average of all pairwise distances in the data and d is the dimensionality of the data. The purpose of H and $\sqrt[d]{k}$ is to avoid the scaling problem of the penalty term. Under this formulation, the selection of λ_0 is invariant under data scaling and different k. In contrast to *K*-means above, P-*K*-means provides flexibility of not assigning all points into clusters and allows a set of scattered points, S. Following Tseng and Wong (2005), scattered points are defined as noises that do not tightly share common patterns with any of the clusters in the data. For clustering problems in complex data such as gene clustering in expression profiles, ignoring scattered genes has been found to dilute identified patterns, make more false positives and even distort cluster formation and interpretation.

We can similarly find relationships between penalized *K*-means and classification likelihood. If the scattered points in S are uniformly distributed in the hyperspace V (i.e. generated from a homogeneous Poisson process), then the C_1-CML criterion becomes

$$f(X|C,\theta) = \prod_{j=1}^{k} \prod_{x_i \in C_j} f\left(x_i|\mu_j,\Sigma_j\right) \cdot \prod_{x_i \in S} \frac{1}{|V|} \qquad (5)$$

where $|V|$ is the hypervolume of V (see a similar model in Fraley and Raftery, 2002). Assume $\Sigma_j = \sigma_0^2 I$. We find maximizing (5) is equivalent to minimizing (4) if $\lambda_0 = 2\sigma_0^2 \cdot (1/\eta)^2 \cdot \log|V|$. This relationship provides a good guidance for the selection of λ_0.

Example 3: PW-*K*-means. In this example we further extend (4) to consider prior information incorporation by adding weights in the cost function:

$$W_{\mathrm{PW}}(C; k,\lambda_0) = \sum_{j=1}^{k} \sum_{x_i \in C_j} w(x_i; \mathrm{P}) \cdot \left\| x_i - \bar{x}^{(j)} \right\|^2 + \eta^2 \cdot \lambda_0 |S| \qquad (6)$$

where P represents the prior information available to enhance clustering and $\eta = H/\sqrt[d]{k}$ as in P-*K*-means. Very often P includes prior knowledge of multiple groups of objects that are likely to be clustered together. In gene clustering of expression profile, some groups of genes may be known to be co-regulated in biological pathways from a particular biological database or previous experiments. In Supplementary Table 1 and Supplementary Figure 1, e.g. a list of 104 annotated genes functioning in different cell cycle periods are obtained from Spellman *et al*. (1998). Suppose information of p pathways are known and each pathway contains $n_1, \ldots, n_p$ genes. We denote by $P_m^{(l)}$ the vector of expression pattern of gene m in pathway l $(l = 1, \ldots, p;\ m = 1, \ldots, n_l)$ and $\mathrm{P} = ((P_1^{(1)}, \ldots, P_{n_1}^{(1)}), \ldots, (P_1^{(p)}, \ldots, P_{n_p}^{(p)}))$. The following transformed logistic function is proposed as the weight function for this type of prior information:

$$w_{\mathrm{pw}}(x; \mathrm{P}) = \alpha + (1 - \alpha) \cdot \frac{1 - e^{-\tau \cdot h(x;\mathrm{P})}}{1 + e^{-\tau \cdot h(x;\mathrm{P})}} \qquad (7)$$

where α and τ are tuning parameters and $h(x_i; \mathrm{P}) = \min_l (1/\eta) \cdot (\Sigma_{m=1}^{n_l} ||x_i - P_m^{(l)}||/n_l)$. Intuitively, when the expression vector of gene i (i.e. x_i) is close to the neighborhood of one of the p pathways in P, $h(x_i;\mathrm{P})$ is small and consequently $w_{\mathrm{pw}}(x;\mathrm{P})$ shrinks to a smaller value, which will force the gene not to be abandoned to the noise set S but to form a cluster in the neighborhood instead. In Supplementary Figure 2, the relationship of the weight function and tuning parameters α and τ are plotted. The parameters α and τ are similar to hyper-parameters in the specification of Bayesian priors, which often reflect the confidence of the prior information. In our experience, $\alpha \in [0.1, 0.4]$ and $\tau \in [2, 3]$ are recommended for most microarray data analysis and the results are quite robust to different selections of α and τ. The weight function can be modified to suit different structures of prior information and to serve for different purposes. For example, the trimmed mean can be used in $h(\cdot;\cdot)$ to accommodate the problem of possible outliers and errors in the prior information P.

2.2 Computation and implementation

2.2.1 Implementation For implementation we modified the *K*-means routine in a published C clustering library (De Hoon *et al*., 2004). The optimization is performed through a classification EM algorithm with multiple numbers (denoted by M) of independent random initial values to search for the global optimum of the loss function. Depending on the complexity and structure of the data, a suitably large M is needed to obtain the global optimum solution with high probability. In the simulatioins, we will explore this problem in simulated data and will further show that good prior information in PW-*K*-means helps to alleviate computational difficulty.

2.2.2 Parameter estimation In the formulation presented above, the number of clusters k and λ (or λ_0) are both considered known a priori. In cluster analysis, estimation of k received tremendous attention in the literature and still remains a difficult problem in the analyses of most real data sets. Milligan and Cooper (1985) conducted a comprehensive evaluation for more than 30 methods. None showed superior performance in all situations. Some methods outperform the others under certain data distribution assumptions but may perform poorly in other settings.

In this article, we follow the prediction-based resampling method proposed by Tibshirani and Walther (2005) (Breckenridge, 1989; Dudoit and Fridlyand, 2002) for selecting k and λ. Given k and λ the whole data set is first split into two equal parts: the first is the training set X_{tr} and the second is the testing set X_{te}. The main idea involves three steps: (a) cluster the training data X_{tr}, (b) cluster the testing data X_{te} and (c) measure how well the training set clustering result predicts co-memberships in the testing data. The correct parameter selection should generate consistent clustering results in training and testing data and produce a good prediction in step (c).

We denote by $C(X_{\mathrm{tr}};k,\lambda)$ the clustering operation on the training data. Following the convention in Tibshirani and Walther (2005), we denote by $D[C(X_{\mathrm{tr}};k,\lambda), X_{\mathrm{te}}]$ the $(n/2) \times (n/2)$ co-membership matrix in the testing data X_{te} judged by the clustering result from training data, $C(X_{\mathrm{tr}};k,\lambda)$. The nearest centroid criterion is used for such judgment, i.e. each point in the testing data is assigned to the nearest cluster centroid of $C(X_{\mathrm{tr}};k,\lambda)$. For any pair of points i and i' in the testing data, the i–i'-th element of the co-membership matrix $D[C(X_{\mathrm{tr}}; k,\lambda),X_{\mathrm{te}}]_{ii'}$ will take the value 1 if both i and i' fall into the same cluster under $C(X_{\mathrm{tr}};k,\lambda)$ judgment and zero otherwise. We denote by $(C_1^{\mathrm{te}}, \ldots, C_k^{\mathrm{te}}, C_{k+1}^{\mathrm{te}} = S^{\mathrm{te}})$ the resulting cluster indexes from clustering the test data in step (b) such that $X_{\mathrm{te}} = \cup_{j=1}^{k+1} C_j^{\mathrm{te}}$ and $n_1, \ldots, n_{k+1}$ are the number of observations in each cluster. The prediction strength of the training and testing data split is defined as

$$ps(k,\lambda) = \min_{1 \le j \le k+1} \frac{\sum_{i \ne i' \in C_j^{\mathrm{te}}} I(D[C(X_{\mathrm{tr}}; k,\lambda),X_{\mathrm{te}}]_{ii'} = 1)}{n_j(n_j - 1)} \qquad (8)$$

where $I(\cdot)$ is the indicator function which equals 1 if the statement is true and 0 otherwise. Intuitively, we compute for each cluster in the test data, the proportion of all pairs of objects that are also assigned in the same cluster by the training cluster centroids judgment. We repeat

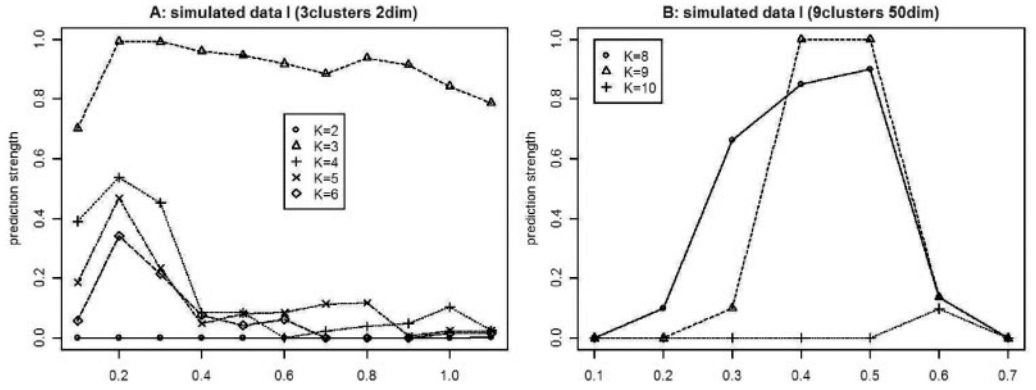

Fig. 1. Averaged prediction strength of different k and λ_0 in both simulated data for parameter selection.

the independent samplings for the training and testing data (10 times in this article) and the averaged prediction strength $\overline{ps}$ is reported. Normally $(k^*,\lambda^*)=\arg\max\overline{ps}(k,\lambda)$ is used for final clustering in practice. However, in the context of gene clustering in microarray, we may want to select as large k and as small λ as possible with reasonably high prediction strength ($\overline{ps}>0.6$ or 0.7) so that many important tight cluster patterns are retrieved.

3 RESULTS

3.1 Simulation

We perform two simulations to evaluate P-*K*-means and PW-*K*-means. The first data contain $k=3$ clusters normally distributed in 2D space. A number of $n=50$ points is simulated for each cluster from $N((-10,10)^T,\sigma^2 I)$, $N((0,-10)^T,\sigma^2 I)$ and $N((10,10)^T,\sigma^2 I)$ $(\sigma=2)$. Another $m=50$ noise points is uniformly generated in $[-20,20]\times[-20,20]$ with the restriction that they are at least three SDs from the centers. Under this restriction, no confusion of clustered points and noise points should exist. For the second data set, we apply a hierarchical log-normal model proposed by Thalamuthu *et al.* (2006) to simulate gene cluster structure in microarray data. Supplementary Figure 3 shows a heatmap presentation of the data with 392 clustered genes of 9 clusters and 392 noise genes in 50 samples/dimensions.

The prediction strength method in the Section 2.2 is used to estimate k and λ_0 in P-*K*-means. Figure 1A shows the averaged prediction strength of 10 independent repeated resamplings for various k and λ_0 in the first data set ($M=100\,000$). Clearly $k=3$ and $\lambda_0=0.2$–0.3 give the highest prediction strength. In Supplementary Figure 4 clustering results of P-*K*-means with $k=3$ and $\lambda_0=0.1$, 0.2 and 0.8 (Supplementary Fig. 4B, C and D) as well as the original true clustering structure (Supplementary Fig. 4A) are presented. P-*K*-means with $\lambda_0=0.2$ gives clustering identical to the underlying truth. When a small $\lambda_0=0.1$ (Fig. B) is selected, a few true cluster points are identified as noises. Conversely when a too large $\lambda_0=0.8$ is selected, many true noises are grouped into the clusters. The result demonstrates the inverse relationship of λ_0 and the size of the noise set in Proposition 1. Figure 1B shows the result of the estimation of K and λ_0 for the second simulation ($M=100\,000$). The parameters are correctly estimated ($K=9$ and $\lambda_0=0.4$–0.5) and the underlying true clustering structure is recovered under this parameter selection.

In the above results, large M is applied to guarantee correct global optimization. Next, we investigate effects of various M in the optimization of clustering. We apply the adjusted Rand index (ARI) (Hubert and Arabie, 1985) to evaluate the clustering results. ARI calculates the similarity of a clustering result to the underlying true clustering structure. It takes a maximum value 1 if the clustering result is identical to the underlying truth and takes expected value 0 if the clustering result is obtained from random partitioning. Figure 2 shows the means and standard errors of ARI over 30 independent tries under different numbers of M for the two simulation data sets ($\lambda_0=0.2$ for the first data set and $\lambda_0=0.4$ for the second). We found that around $M=10^5$–10^6 independent random initial values are needed to guarantee the global optimum for P-*K*-means with high probability (triangles) in both sets of simulated data. Suppose 'good prior' information P containing three clustered points randomly selected from each cluster respectively is available and is applied to PW-*K*-means. The M needed to acquire global optimum becomes $\sim 10^3$-fold less than P-*K*-means in the first simulated data set. Similar result is found in the second simulated data set where good prior information reduces the M needed by $\sim 10^4$-fold. This analysis suggests that good prior information for PW-*K*-means helps to alleviate the local minimum issue in implementation. We also test a prior information randomly selected from the data ('bad prior') for PW-*K*-means. The result is similar to P-*K*-means as if no prior information is given in the first simulation while the result of 'bad prior' seems to slightly improve P-*K*-means in the second simulation, possibly because the weights have smoothed the surface of the cost function in the higher dimensional and complex situation.

3.2 Clustering MS/MS spectra

Mass spectrometry is an important experimental technique in proteomic research which involves large-scale study of proteins in a tissue. In this example, we discuss mining of data sets from tandem mass spectrometry (MS/MS). Each peptide is a sequence of amino acids (AA) composed of 20 possible choices (AA_i, $1\le i\le 20$: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y) at each location. Protonated peptides are randomly dissociated in the gas phase under low energy collision-induced dissociation (CID). For example, a peptide sequence 'AAIANAPR' has a certain probability to dissociate at the

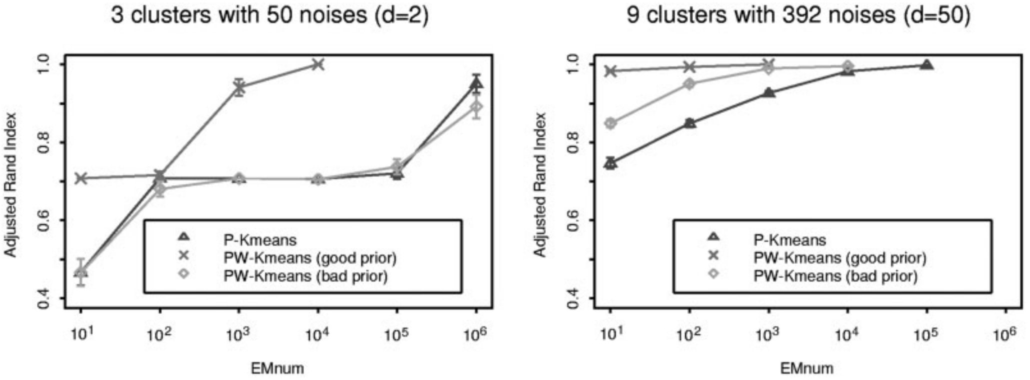

Fig. 2. Adjusted Rand index for different selection of M in P-*K*-means (triangle), PW-*K*-means with good prior (cross) and PW-K-means with bad prior (diamond).

'I-A' position, breaking into two sub-sequences, 'AAI' and 'ANAPR'. Similarly it has a different probability to break at the 'N-A' position, resulting in 'AAIAN' and 'APR'. In the experiment, the output fragmentation intensities are measured and are assumed to be proportional to dissociation probabilities. The intensities are all normalized between 0 and 1. We denote by x_{ijk} the fragmentation intensity at position 'AA_j-AA_k' of peptide i ($i = 1, \ldots, 28330$, $1 \le j$, $k \le 20$, $0 \le x_{ijk} \le 1$). For example, in peptide 'AAIANAPR'' we observe intensities at six dissociation positions {(A-A), (A-I), (I-A), (A-N), (N-A) (A-P)} to be {0.15, 0.11, 0.69, 0.49, 0.76 and 0.51} [(P-R) is not observed in the experiment] and the remaining 394 out of $20 \times 20 = 400$ possible fragmentation positions are not observable (missing). In a given set of peptides X, we observe an empirical distribution of fragmentation intensities $\{x_{ijk}: i \in X\}$ at each fragmentation position 'AA_j-AA_k'. For convenience of pattern visualization, a 'quantile map' in Supplementary Figure 5 is introduced to visualize the distribution. The intensities of 5%, 15%, ..., 95% percentiles are shown in 10 concentric ring areas from outside to inside. The magnitude of the intensity is represented in gradient gray (or color) scale. Supplementary Figure 5 demonstrates three types of distributions with the density plot in the middle panel and their corresponding quantile maps on the left.

It has been reported recently that different peptide sequence contents and chemical backgrounds, which we will call motifs hereafter, may contribute to different fragmentation patterns (Huang *et al.*, 2005). For example, in the upper left plot of Figure 3 a quantile map of 674 peptides having one positive charge, ending in R and containing P in the middle is shown. The map clearly shows that the dissociation probabilities at most positions of 'D-X' and 'E-X' (X represents any AA) are significantly higher than other positions. In the lower left plot another motif of 2182 peptides having two positive charges, ending in R and containing P in the middle shows a different fragmentation pattern. 'X-P' positions are easier to break in this motif.

Learning the patterns of dissociation strength under different motifs is essential to improve protein identification models and algorithms in the analysis of MS/MS data. A natural question then arises: is it possible to cluster the full 28 830 peptides into clusters of similar fragmentation patterns and in turn 'learn' the underlying motifs contributing to these fragmentation patterns? Since each peptide is only 8–20 AA long, only 1.8% (7/400) to 4.8% (19/400) of fragmentation intensities are observed for each peptide. Calculating distances between any two peptides is usually impossible and most clustering methods become inapplicable to such a data set. On the other hand, the distance from a peptide to a set of peptides is still computable; in fact, we can calculate the distance of a peptide to the center of a peptide set where the center is the simple dimension-wise average ignoring missing values. We let $\{x_i = x_{ijk}: 1 \le j,k \le 20\}$ and define the distance between a peptide x_i and a set of peptides $Y = \{y_i\}$ to be

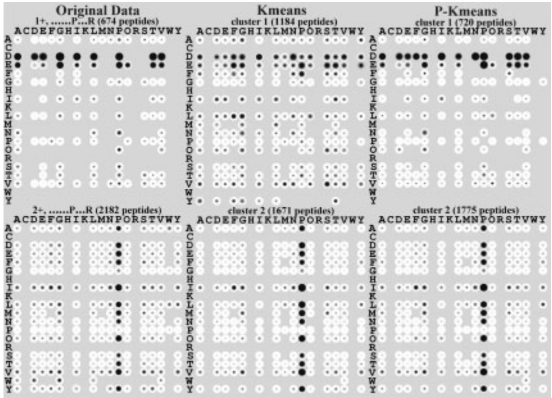

Fig. 3. Left: patterns of quantile maps for two sets of peptides are shown: 'charge 1, P in the middle and end with R' and 'charge 2, P in the middle and end with R'. Middle: clustering result of traditional *K*-means. Right: clustering result of P-*K*-means.

$$d(x_i, Y) = \frac{400}{\left(\begin{matrix}\text{Number of}\\ \{j,k : x_{ijk} \,\&\, \bar{y}_{jk}\ \text{not missing}\}\end{matrix}\right)} \sum_{j,k:x_{ijk} \,\&\, \bar{y}_{jk}\ \text{not missing}} (x_{ijk} - \bar{y}_{jk})^2$$

where

$$\bar{y}_{jk} = \frac{1}{\text{Number of } \{i : y_{ijk} \text{ not missing}\}} \sum_{i:y_{ijk}\ \text{not missing}} y_{ijk}$$

Note that the fraction in the above distance definition is to adjust for different numbers of missingness. With the above distance definition, the cost function in *K*-means Equantion (3) and P-*K*-means Equation (4) can be used for clustering

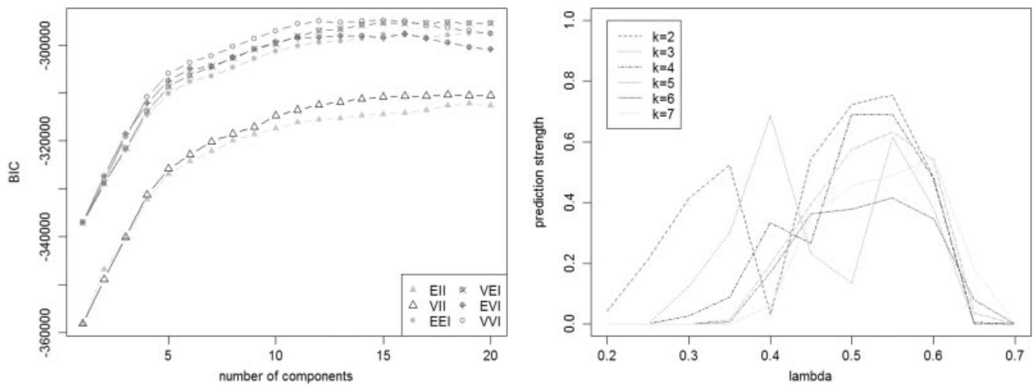

Fig. 4. BIC and prediction strength to estimate K and λ.

peptides. For demonstration purposes in this article, we only pooled peptides of the two sets shown in the left panel of Figure 3 and applied *K*-means and P-*K*-means clustering respectively to the combined 2856 peptides. In the *K*-means result at the middle panel, we find cluster 1 is identified to have a similar fragmentation pattern as the charge 1 set but with diluted information in ‘D-X’ and ‘E-X’ and contaminated information in many positions including ‘L-X’ and ‘V-X’ by the many more peptides included (1184 versus 674). P-*K*means, on the other hand, almost perfectly recovers the two patterns on the right panel. The result shows the importance of excluding scattered peptides in clustering and superior performance of P-*K*-means. For details of how P-*K*-means was applied to cluster the complete data set of 28 830 peptides and how a data mining scheme of integrated P-*K*-means and regression tree was further developed for learning sequence motifs, see Huang *et al.* (2007).

3.3 Clustering genes in expression profile

An expression profile of the yeast cell cycle from Spellman *et al.* (1998) is analyzed to evaluate performance of clustering results by different methods. Seventy-seven samples of yeast cells synchronized under various time points of different cell cycle stages were applied to cDNA microarray to analyze patterns of expression changes during the cell cycle. Expression data of 6179 genes are examined of which 1663 genes are used for gene clustering after pre-filtering procedures (deleting genes with missing value more than 20% or SD at log-2 scale less than 0.4). Missing values are imputed and each gene vector is standardized to mean 0 and SD 1 so that expression patterns of differential expression levels become comparable. The data matrix is then input to *K*-means and P-*K*-means for gene clustering. Other popular clustering methods are also evaluated.

In Figure 4, Bayesian information criterion (BIC) for Gaussian mixture model (mclust) is used to estimate K (left panel) and prediction strength introduced in Section 2.2 is used to determine K and λ (right panel). It is easily seen that the parameter selection in this real data is not as clear as in the simulated examples. BIC can hardly determine a suitable selection of K. The best prediction strength appears at $K=2$ and $\lambda=0.55$, which does not seem a reasonable selection in this example. If we try to select as many tight clusters as possible with reasonably high prediction strength, we might settle with $K=5$ and λ=0.5. The ambiguous parameter selection is commonly seen in almost any real-world microarray data set.

Genes with similar expression patterns often imply co-regulated biological functions. A common application of gene clustering is to predict functional annotation for novel (unannotated) genes. For a functional annotation F, we suppose a total genome size of G ($G=1663$ in this example) genes are analyzed among which $D(F)$ genes are known to be categorized in F from a given biological database. We use the hypergeometric distribution (Tavazoie *et al.*, 1999) to assess the probability of observing at least $d(F)$ genes belonging to F in a cluster C. We denote by $n(C)$ the total number of genes in cluster C. The resulting P-value is given by

$$P(G,D(F),n(C),d(F)) = 1 - \sum_{i=1}^{d(F)-1} \frac{\binom{D(F)}{i}\binom{G-D(F)}{n(C)-i}}{\binom{G}{n(C)}}.$$

On average we should observe $d(F) \sim D(F) \cdot n(C)/G$ when the cluster C is not enriched in functional category F. Intuitively, an unusually large $d(F)$ which corresponds to a small P-value implies that the majority of the genes in this cluster belong to the functional category F.

In general, it is difficult to evaluate the performance of a gene clustering in a real-world data set due to lack of underlying truth or gold standard. Biological databases such as Gene Ontology and KEGG can serve for this purpose, however, information from this famous yeast cell cycle data may have been incorporated to the databases and the usage of the databases may cause tautological bias. Here, we cite 104 cell cycle regulated genes (see Supplementary Table 1) from Spellman *et al.* (1998) that were verified by traditional experimental methods and treat them as the gold standard. Mapping them to the 1663-gene data set and discarding duplicate genes, 87 genes belonging to six disjoint functional categories were obtained: F_1(M/G1), F_2(late G1 SCB regulated), F_3(late G1 MCB regulated), F_4(S-phase), F_5(S/G2-phase) and F_6(G2/M-phase). The remaining 1576 unannotated genes are denoted by F_7. Since estimating the number of clusters k in gene clustering is usually difficult, we performed clustering with $k=5,\ldots,20$ for each method. The resulting clusters are denoted by C_{ki}, where $k=5,\ldots,20$ and $i=1,\ldots,k$ for a given clustering method. For convenience, we denote by

d_{kij} the number of genes in cluster i and functional category j when k is number of clusters and $P_{kij} = P(G, D(F_j), n(C_{ki}), d_{kij})$ are the corresponding P-values. A contingency table of d_{5ij} $(i = 1, \ldots, 6;\ j = 1, \ldots, 7)$ in P-K-means clustering $(k = 5)$ and functional category distributions is demonstrated in Table 1 with their corresponding P-values (P_{5ij}).

In practice, P-values should be adjusted for multiple comparisons in any clustering result inference. Here, we take a different evaluation criterion used in Thalamuthu *et al.* (2006) (Wu *et al.*, 2002) to compare performance of different clustering methods. To avoid sensitivity of estimation of k in the following evaluation, we pool all clustering results C_{ki} $(k = 5-20,\ i = 1-k)$ to draw a prediction-accuracy plot. Given a P-value threshold δ, we assign all genes in the cluster C_{ki} to functional annotation F_j when the corresponding P-value P_{kij} is less than δ. Thus among $n(C_{ki})$ annotation predictions made, d_{kij} genes are verified as correct by the biological database. We define 'predictions made' $= \Sigma_{k,i,j:P_{kij}<\delta}\, n(C_{ki})$, 'verified predictions' $= \Sigma_{k,i,j:P_{kij}<\delta} d_{kij}$, and 'accuracy' = 'verified predictions'/'predictions made'. For example, in Table 1, if $\delta = 0.01$, then the 'predictions made' $= (42 + 98 + 98) = 238$ and 'accuracy' $= (10 + 4 + 23)/238 = 15.55\%$. Figure 5 shows the prediction-accuracy plot with 'predictions made' on the x-axis and 'accuracy' on the y-axis. For each clustering method, varying the threshold δ produces a different number of 'predictions made' and corresponding 'accuracy', thus generating a curve. $\delta = 1\text{E}-4,\ 1\text{E}-5, \ldots, 1\text{E}-20$ was used in the figure and, in general, a more stringent threshold (smaller δ) corresponds to fewer number of predictions made and better accuracy. The horizontal dotted line shows the average accuracy under random annotation assignment: $87/1663 = 5.23\%$. Methods producing curves on the above have better performance. P-values were not calculated and no annotation prediction was made for genes in the noise set. Note, in this approach a gene may be predicted multiple times because multiple clustering results (different k) were empirically assessed together. In Figure 5, we have selected seven methods that are popularly applied in the literature of microarray analysis: hierarchical clustering, SOM, K-means, CLICK, GIMM, model-based (mclust) and tight clustering. Description of the methods and detailed implementation are provided in Supplementary Material. The result shows that P-K-means outperforms K-means due to its ability to leave scattered genes unclustered. Intuitively, P-K-means is almost equivalent to the simplified form of identical spherical covariance structure in mclust except that P-K-means follows CML sampling scheme and mclust is an ML approach. It is found that mclust also has better performance than K-means and is similar to P-K-means ($\lambda = 0.4$ and 0.35). Tight clustering has similar performance to P-K-means with $\lambda = 0.35$. It should, however, be noted that tight clustering relies on resampling assessment and the computation is much heavier than P-K-means. P-K-means with $\lambda = 0.3$ generates the tightest clusters with highest accuracy. The number of predictions made is, however, much fewer than the other methods. All other methods seem to provide worse prediction accuracy.

Table 1. Contingency table of P-K-means

	Counts (P-values)						
	F_1	F_2	F_3	F_4	F_5	F_6	F_7
C_1	0(1)	0(1)	0(1)	0(1)	0(1)	0(1)	58
C_2	1(0.35)	0(1)	0(1)	0(1)	0(1)	**10(8E−13)**	31
C_3	0(1)	0(1)	0(1)	0(1)	0(1)	0(1)	101
C_4	0(1)	0(1)	0(1)	0(1)	0(1)	0(1)	39
C_5	0(1)	**4(4E−3)**	**23(8E−13)**	0(1)	0(1)	0(1)	71
S	16	8	5	8	7	5	1276

P-K-means ($k = 5$ and $\lambda = 0.3$) clusters and functional annotation distributions and their corresponding P-values (in parentheses) are presented. Functional categories enriched in the clusters ($P < 0.01$) are in boldface. The eight histone genes are left in the noise set S (shaded in gray).

Finally, we examine the usefulness of PW-K-means. We notice from Supplementary Figure 1 that the eight genes in F_4 share a tightly co-regulated pattern. In fact, it is widely known that the eight genes are core histone genes required for chromatin assembly and chromosome function, and their modifications play central roles in a variety of chromosomal processes during S phase. Unfortunately in Table 1 we find that all eight genes are left in the noise set S in the P-K-means clustering result. To demonstrate the ability of PW-K-means to incorporate prior information, we randomly select three F_4 genes as the prior information P. As shown in Table 2 for the PW-Kmeans result ($\alpha = 0.4$ and $\tau = 2$), we were able to recover and identify cluster C_3 containing all the eight histone genes. In Table 3, we compare P-K-means and PW-K-means with various α and τ and show the frequencies of the eight histone genes (F_4) being assigned to scattered gene set in 100 trials with independent starting seeds. We find that P-K-means usually assign the eight histone genes to the scattered set. In PW-K-means, smaller α and smaller τ give stronger prior information to increase the probabilities of identifying these important genes in the clusters.

4 DISCUSSION

In this article, we propose a general class of penalized and weighted K-means for clustering. The method allows a noise set without being clustered and provides the flexibility for prior information incorporation through the weights. We have shown the success and flexibility of this method in high-throughput biological data through simulations, MS/MS and microarray applications. In the literature, many different sampling types of the Gaussian mixture model-based approach have been developed and applied. Different forms of the Bayesian hierarchical models are also proposed for clustering and efficient Markov chain Monte Carlo (MCMC) or Gibbs sampling techniques are used to simulate the posterior distributions (Jain and Neal, 2004; Medvedovic and Sivaganesan, 2002). However, probably none except for Pan (2006) and Huang and Pan (2006) have demonstrated the use of prior information in clustering. Pan (2006) applied a stratified model-based approach to incorporate gene annotation as prior information. The annotation information, however, was not fully utilized to improve the formation of clusters. The weights in our PW-K-means provides a more general and flexible alternative. Huang and Pan (2006) incorporated prior information in a modified shrinkage distance

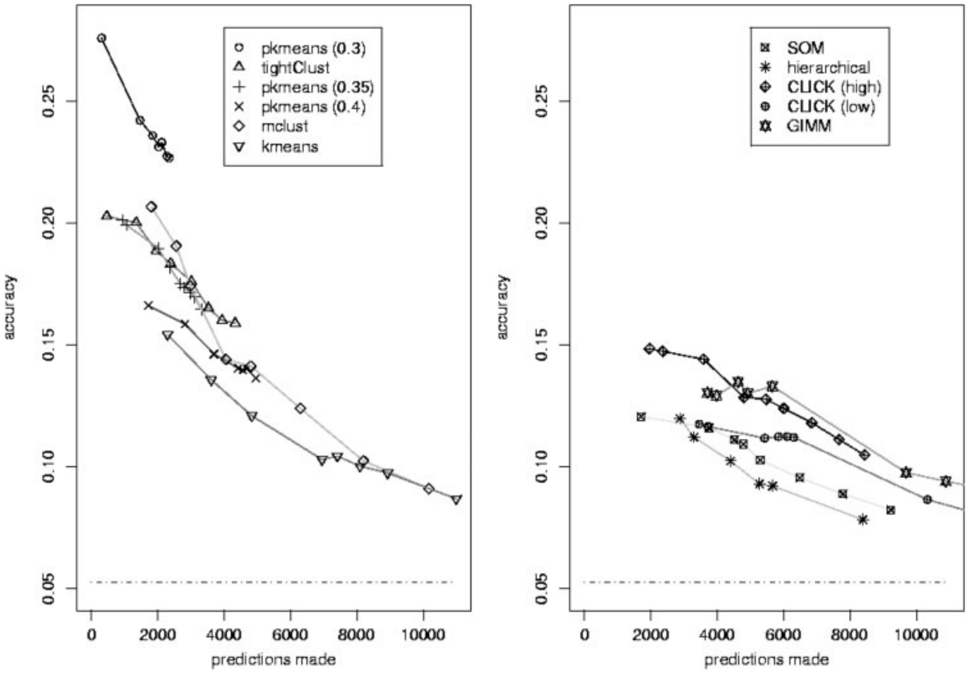

Fig. 5. Prediction-accuracy curves of *K*-means, P-*K*-means, tight clustering, hierarchical clustering, SOM, model-based clustering (mclust), CLICK and GIMM. Curves above have better accuracy and performance. P-Kmeans improves traditional *K*-means and performs among the best compared to existing popular methods. 'CLICK (high)' pools results from larger homogeneous parameters in the CLICK algorithm and 'CLICK (low)' pools results from smaller ones. For GIMM, the hierarchical tree was cut and only clusters with more than 20 genes are considered a cluster for desired *K* (see Supplementary Material for detailed implementation of each

for clustering. The distance of any two genes with the same known gene annotation is shrunken to a fixed smaller proportion. On the other hand, our proposed weight function shrinks all gene distances in the spatial neighborhood of a clustered set of same annotated genes and is expected to be more general and effective.

K-means is known to be a fast algorithm among many clustering methods. As an extension from *K*-means, PW-*K*-means only adds limited extra computation in calculating the weight function and the penalty term. Like all optimization-based clustering methods, global optimization of PW-*K*-means is NP-hard. Multiple random initial values and other stochastic methods have been proposed for such purposes (Biernacki *et al*., 2003). In general as the number of clusters and noises increase and the data become more complex, computation becomes heavier to approach the global optimization with high probability. In this article, we showed that good prior information reduces the computation efforts required for global optimization.

In the example of MS/MS (Section 3.2), we showed that PW-*K*-means was capable of clustering a large data set with a high percentage of missing values (>95%), to which most other clustering methods could not be applied. In the microarray example in Section 3.3, we demonstrated that the clustering result of P-*K*-means had one of the best prediction accuracies. In our experience, P-*K*-means usually has comparable performance to the resampling-based tight clustering method (Tseng and Wong, 2005) while P-*K*-means has the advantage of fast computation and the flexibility of choosing different λ to easily control the tightness of clusters. On the other hand, PW-*K*-means can suffer from local optimization results and tight clustering usually provides a more stable solution through the resampling evaluation. As a result, these two methods are complementary to serve different data mining purposes in different types of data sets.

Table 2. Contingency table of PW-*K*-means

	Counts (*P*-values)						
	F_1	F_2	F_3	F_4	F_5	F_6	F_7
C_1	0(1)	0(1)	0(1)	0(1)	0(1)	0(1)	112
C_2	2(0.42)	**4(1E−2)**	**22(4E-12)**	0(1)	0(1)	0(1)	111
C_3	0(1)	**4(2E−3)**	3(.18)	**8(4E−11)**	1(0.32)	0(1)	72
C_4	0(1)	0(1)	0(1)	0(1)	0(1)	0(1)	158
C_5	1(.45)	0(1)	0(1)	0(1)	1(.22)	**10(2E−13)**	45
S	14	4	3	0	5	5	1078

PW-*K*-means ($k=5$, $\lambda=0.3$, $\alpha=0.4$ and $\tau=2$) clusters and functional annotation distributions and their corresponding *P*-values (in parentheses) are presented. Functional categories enriched in clusters ($P<0.01$) are in boldface. The eight histone genes are clustered in C_3 (shaded in gray).

In the cell cycle data, we have shown that both BIC and prediction strength cannot give very clear parameter estimation as in the simulation examples. In tight clustering (Tseng and Wong, 2005), the algorithm attempts to sequentially retrieve tight patterns from the data, which makes the estimation of *K* a secondary task. This provides certain degree of robustness while the input parameter of the method still needs approximate knowledge of *K* for the method to perform well. An alternative direction may be to take clustering results of multiple *K* and λ to

Table 3. Frequency of the eight histone genes (F_4) being assigned to scattered gene set in 100 trials with independent starting seeds in P-*K*-means and PW-*K*-means (*M*: number of independent random initial values to obtain global optimum)

P-*K*-means ($\lambda = 0.3$)	$M = 1$				$M = 100$			
	95				73			
PW-*K*-means ($\lambda = 0.3$)	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 5$	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 5$
$\alpha = 0.2$	0	0	0	93	0	0	0	58
$\alpha = 0.4$	0	22	71	93	0	0	30	67
$\alpha = 0.6$	81	90	92	93	54	50	57	65
$\alpha = 0.8$	90	92	93	94	63	73	72	64

show a sequential multi-resolution outlook and different degree of tightness of the cluster patterns (Supplementary Fig. 6). In general the parameter estimation (especially estimation of *K*) is still an open, although old, question in the field and the solution seems to be very data and goal dependent. More research efforts are still needed in the future.

Several issues for PW-*K*-means can be further pursued in future research. The current formulation does not consider the covariance structure of clusters such as variable dispersion (SD) and non-spherical cluster structures. The weight function in PW-*K*-means needs further development and improvement for different types of data. For example, we have encountered some applications with prior information of cluster locations (centroids) or information that some points are unlikely to be noise points. We expect further development of PW-*K*-means will help better data mining and analyses of many high-throughput biological data sets.

ACKNOWLEDGEMENTS

G.C.T. is supported by grant 1 KL2 RR024154-01. The author is indebted to E. Feingold for valuable discussion and comments. The author wishes to thank the editor and the reviewers for many insightful suggestions and comments.

Conflict of Interest: none declared.

REFERENCES

Basu,S. *et al.* (2004) A probabilistic framework for semi-supervised clustering. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 59–68.

Biernacki,C. *et al.* (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Stat. Data Anal.*, **41**, 561–575.

Bishop,C. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK.

Breckenridge,J.N. (1989) Replicating cluster analysis: method, consistency, and validity. *Multivariate Behav. Res.*, **24**, 147–161.

Celeux,G. and Govaert,G. (1992) A classification EM algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal.*, **14**, 315–332.

Cheng,J. *et al.* (2004) A knowledge-based clustering algorithm driven by Gene Ontol-ogy. *J. Biopharm. Stat.*, **14**, 687–700.

Conrads,T.P. *et al.* (2003) Cancer diagnosis using proteomic patterns. *Expert Rev. Mol. Diagn.*, **3**, 411–420.

Dasgupta,A. and Raftery,A.E. (1998) Detecting features in spatial point processes with clutter via model-based clustering. *J. Am. Stat. Assoc.*, **93**, 294–302.

De Hoon,M.J.L. *et al.* (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.

Dudoit,S. and Fridlyand,J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.*, **3**, 0036.1–0036.21.

Fraley,C. and Raftery,A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.

Ganesalingam,S. (1989) Classification and mixture approach to clustering via maximum likelihood. *Appl. Stat.*, **38**, 455–566.

Gordon,A.D. (1999) *Classification*. 2nd ed. Chapman & Hall/CRC, Boca Raton, USA.

Grabmeier,J. and Rudolph,A. (2002) Techniques of cluster algorithms in data mining. *Data Mining Knowl. Discov.*, **6**, 303–360.

Hanisch,D. *et al.* (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**, 145–154.

Hartigan,J.A. and Wong,M.A. (1979) A K-means clustering algorithm. *Appl. Stat.*, **28**, 100–108.

Hastie,T. *et al.* (2000) Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**, 0003.1–0003.21.

Huang,D. and Pan,W. (2006) Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics*, **22**, 1259–1268.

Huang,Y. *et al.* (2005) Statistical characterization of charge state and residue dependence of low energy CID peptide dissociation patterns. *Anal. Chem.*, **77**, 5800–5813.

Huang,Y. *et al.* (2007) A data mining scheme for identifying peptide structural motifs responsible for different MS/MS fragmentation intensity patterns. *Journal of Proteomic Research*. (in revision).

Hubert,J. and Arabie,P. (1985) Comparing partitions. *J. Classific.*, **2**, 193–218.

Jain,A.K. and Dubes,R.C. (1988) *Algorithms for Clustering Data*. Wiley, New York.

Jain,S. and Neal,R.M. (2004) A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *J. Comput. Graph. Stat.*, **13**, 158–182.

Jobson,J.D. (1992) *Applied Multivariate Data Analysis*. Springer Bd.I and II, New York.

Kaufman,L. and Rousseeuw,P.J. (1990) *Finding Groups in Data*. Wiley, New York.

McLachlan,G.J. and Basford,K.E. (1987) *Mixture Models*. Marcel Dekker, New York.

McLachlan,G.J. *et al.* (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

Medvedovic,M. and Sivaganesan,S. (2002) Bayesian infinite mixture model-based clustering of gene expression profiles. *Bioinformatics*, **18**, 1194–1206.

Messatfa,H. and Zait,M. (1997) A comparative study of clustering methods. *Future Generation Comput. Syst.*, **13**, 149–159.

Milligan,G.W. and Cooper,M.C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179.

Pan,W. (2006) Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, **22**, 795–801.

Pan,W. *et al.* (2006) Semi-supervised learning via penalized mixture model with application to microarray sample classification. *Bioinformatics*, **22**, 2388–2395.

Ripley,B.D. (1996) *Pattern Recognition and Neural Network*. Cambridge University Press, Oxford, UK.

Segal,E. *et al.* (2003) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, **19**, i264–i271.

Sharan,R. *et al.* (2003) CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics*, **19**, 1787–1799.

Spaeth,H. (1984) *Cluster Analysis Algorithms*. Ellis Horwood Limited, Chicester.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Tamayo,P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *PNAS.*, **96**, 2907–2912.

Tavazoie,S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Thalamuthu,A. *et al.* (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**, 2405–2412.

Tibshirani,R. and Walther,G. (2005) Cluster validation by prediction strength. *J. Comput. Graph. Stat.*, **14**, 511–528.

Tseng,G.C. and Wong,W.H. (2005) Tight clustering : a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, **61**, 10–16.

Wu,L.F. *et al.* (2002) Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nat. Genet.*, **31**, 255–265.

Yeung,K.Y. *et al.* (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.