Clustering High-Dimensional Data: The Expression of E-cadherin, CD44 and p53 Molecules in Lip Cancer

Dudak Kanserinde E-cadherin, CD44 ve p53 Molekülleri: Yüksek Boyutlu Verilerin Kümelenmesi

Kyriaki KITIKIDOU,[a]
Aris NTOMOUCHTSIS,[b]
Chrisoula TSOMPANIDOU,[b]
Konstantinos VAHTSEVANOS[b]

[a]Dimokritos University of Thrace,
[b]Cancer Hospital Theagenion, Greece

Geliş Tarihi/*Received:* 30.11.2009
Kabul Tarihi/*Accepted:* 01.02.2010

Yazışma Adresi/*Correspondence:*
Kyriaki KITIKIDOU
Pandazidou 193, 68200,
Orestiada, Greece
kkitikid@fmenr.duth.gr

ABSTRACT Objective: Clustering techniques can determine which expression patterns are important and which genes contribute to such patterns. We evaluate performance on data from a lip carcinoma study in Greece. Lip carcinoma is one of the most common malignant oral and maxillofacial tumours and in advanced clinical stages has a poor prognosis. E-cadherin, CD44 and p53 molecules are associated with cellular adhesion. **Material and Methods:** To prepare for clustering, we divided each of the median normalized gene expression values by the range of that gene. Next, we set our prior parameters and we performed the final inference using pooled sets of Markov chain Monte Carlo (MCMC) runs. After pooling the chains, we grouped the data into clusters and selected E-cadherin, CD44 and p53 molecules using the marginal median model as cut off. The selection of a small set of genes is advantageous here. A small number of selected genes is appealing to biologists because they constitute a manageable set of candidates on which further studies can be performed. **Results:** E-cadherin, CD44 and p53 molecules were selected as discriminatory. Results highlight the fact that clustering method has successfully selected genes that are biologically consistent with current research and that provide strong biological validation of the cluster configuration suggested. **Conclusion:** A clustering method that takes advantage of known substructure in the data when simultaneously clustering high-dimensional data with an unknown number of clusters, and selecting the best discriminating variables for those clusters implies the opportunity to handle bigger datasets. When analyzing real data, clustering has found three genes that agree with current biological research and literature and that provide biological validation of the cluster configuration. Overall, clustering can provide biologists with both useful and manageable information for further experimental research.

Key Words: Bayesian clustering; bayesian variable selection; carcinoma, cluster analysis; clustering high-dimensional; reversible-jump markov chain monte carlo; squamous cell

ÖZET Amaç: Kümeleme teknikleri, hangi ekspresyon modelinin önemli olduğunu ve hangi genlerin bu oluşumlara katkısı olduğunu belirleyebilir. Bu çalışmamızda Yunanistan' da dudak kanseri ile ilgili bir çalışmaya ait verilerin performansı değerlendirilecektir. Dudak kanseri, oral ve maksillofasiyal tümörlerin en sık görülen sebeplerinden biridir ve ileri evreleri kötü prognoza sahiptir. E-cadherin, CD44 ve p53 molekülleri hücre adhezyonu ile ilişkilidir. **Gereç ve Yöntemler:** Kümelemeye hazırlık olarak medyanı normalleştirilmiş gen değerlerinin her birini, o genin değer aralıklarına göre bölümlendirdik. Ardından önsel değerlerimizi belirledik ve havuzlanmış Markov zinciri Monte Carlo değerleri ile sonuçlara ulaştık. Serilerin havuzlanmasından sonra verileri kümeler halinde grupladık ve marjinal medyan modelleri kesim değeri alınarak E-cadherin, CD44 ve p53 moleküllerini seçtik. Burada küçük bir gen grubunun seçilmesi avantaj sağlamaktadır. Bir kaç tane genin seçilmesi, biyologlara cazip gelmektedir çünkü bu genler ileride gerçekleştirilebilecek çalışmalar için kullanımı kolay bir set özelliği taşımaktadırlar. **Bulgular:** E-cadherin, CD44 ve p53 molekülleri ayırt edici olarak seçildiler. Elde edilen bulgular kümeleme metodunun mevcut çalışma ile biyolojik olarak uyumluluk gösteren genleri başarıyla seçtiğini ve önerilen küme yapılandırmasının güçlü biyolojik doğrulama sağladığını öne çıkarmaktadırlar. **Sonuç:** Verilerdeki bilinen altyapıyı kullanan kümeleme metodu, yüksek boyutlu verileri sayısı bilinmeyen kümelerle eş zamanlı olarak grupladığında ve bu kümeler için en iyi ayırt edici değişkenleri seçildiğinde, daha büyük veri setlerini elde etme fırsatı anlamına gelmektedir. Gerçek veri analizinde, kümeleme analizi sonucu mevcut biyolojik araştırma ile daha önce yayınlanan sonuçlarla uyumlu üç gen bulunmuş ve bu da küme oluşumunun biyolojik teyidini sağlamıştır. Sonuç olarak, kümeleme daha sonra yapılabilecek deneysel çalışmalar için biyologlara kullanımı kolay bilgi sağlayabilmektedir.

Anahtar Kelimeler: Bayesci kümeleme; bayesci değişken seçimi; karsinoma; kümeleme analizi; yüksek boyutlu kümeleme; ters-atlamalı markov zinciri monte carlo; skuamöz hücreli

Turkiye Klinikleri J Biostat 2010;2(1):42-8

In biostatistics, often the number of *p* variables far exceeds the number of samples. These scripts lead to the problem of dimensionality,[1] that substantially means that the data appear rare across the *p* - dimensional interval, and the ordinary asymptotic hypotheses tests usually are not in effect. A lot of researchers consequently have turned into Bayesian techniques for the analysis of these high-dimensional data to search for differentially expressed genes.[2,3]

A lot of studies that they include microarrays have the substructure innate in the data. This happens, for example, with the designed experiments that group the data within treatments. Recently, Bayesian methods have been presented in the literature that propose an approach in the discovery of genes in designed experiments. [4-7] Efron et al[4] developed a nonparametric approach for the microarray analysis that uses transformation methods in order to estimate the null distribution of summary statistics on the expression of genes.

The non-parametric transformation methods, however, can be contradictory with a restricted number of counterparts by group. Ibrahim et al[6] present a parametric model of mixture of two components which combines a point mass at a threshold value with a component of normal distribution. This method applies to only two groups. Newton et al[5] also applied a model of mixture of two components for the gene expression, supposing that the components approach gamma distributions. Their original method applied to two groups and was recently prolonged to compare the differentially expressed genes when considering multiple groups.[7] This method works well in identifying models of differential expression, but it requires enumeration of all the possible models or a certain external justification to reduce the models.

Other methods for the microarray data without substructure use cluster analysis. In these studies, the aim is classifying the individuals based on their values of expression of the gene. To cluster the individuals effectively, researchers must reduce the number of expression values because including large numbers of uninformative variables can remarkably interfere with recovering of the true structure of the clusters.[8,9] In consequence, the methods of clustering for microarrays must comprise all the information in the data in order to select the values of expression of the gene that guide the clusters. The Bayesian variable selection techniques applied to clustering offer a complete method to select the most informative genes (variables) and recover the structure of the clusters.

Two recent techniques combine the Bayesian techniques of selection of models based on the model clustering. The first technique that is described in Tadesse et al[9] imports a new Bayesian approach in the clustering of high dimensional data. This process estimates jointly the cluster models in the data and selects the variables that determine the best models via the use of Markov chain Monte Carlo (MCMC) method. The second approach that is described in Raftery and Dean[10] uses the same mixture model approach with the model selection approach that is led from Bayes factors and a search algorithm. This algorithm is simplified with the utilization of the BIC to approximate Bayes factors. Both methods recover simultaneously the structure of clusters in the data and select the individual variables that determine better the structure of clusters.

Swartz et al[11] extended the work of Tadese et al[9] in the modeling of data with a known substructure, such as the structure that is imposed by an experimental drawing. They jointly cluster the data and select discriminatory variables, so their method determines which experimental treatments are important, and also which genes have the most differentiating expression values affected by the treatments. With substantial approximation of the within group covariance, their approach facilitates the clustering without it upsets the groups that are determined by the experimenter. This extension applies to any data with substructure, and more specifically to microarrays that are used in the preclinical medical research, where frequently the differentiating genes are more interesting than the clusters that they determine.

MATERIAL AND METHODS

In this article we apply the method of Swartz et al[11] to an experimental design of a microarray dataset from a lip carcinogenesis study. 33 cases with lip carcinoma (squamous cell carcinoma, SCC) were studied, either with low or high histological grade of carcinoma invasion and presentation of positive lymph nodes. Thus, by design we have four groups of patients: low invasion, control; high invasion, control; low invasion, SCC with metastasis; high invasion, SCC with metastasis. The original microarrays consisted of 270 genes. The objective of this study is to discover a small subset of genes that is connected to cervical lymph node metastasis and that can be investigated further using biological trials.

The developed Bayesian method for mixture models that simultaneously cluster the data and select discriminatory variables consists of the following:[9,11]

Let $\mathbf{X} = [\mathbf{x_1},\mathbf{x_2},...\mathbf{x_n})$ denote n-independent p-dimensional observations from G underlying subpopulations. Clustering the n samples can be modeled as a mixture of the G subpopulation models:

$$f(\mathbf{x}_i|\mathbf{w},\theta) = \sum_{k=1}^{G} w_k f(\mathbf{x}_i|\theta_k) \qquad (2.1)$$

where $f(\mathbf{x}_i|\mathbf{w},\theta)$ is the density for the observation from the kth subpopulation and $\mathbf{w}$ is the vector of nonnegative component weights w_k that sum up to 1, and $\boldsymbol{\theta}$ denotes the distribution parameters. The model is completed with a latent vector $\mathbf{y}=(y_1,y_2,...y_n)$ with elements indicating to which subpopulation component each observation belongs to. If the y_i's are iid, with $p(y_i = k) = w_k$, and define the subpopulation distributions to be multivariate normal with mean vector $\boldsymbol{\mu}_k$ and variance matrix Σ^k, then each sample i, can be modeled conditional on y_i, as

$$(\mathbf{x}_i|y_i = k,\mathbf{w},\theta) \sim N(\mu_k,\Sigma k) \qquad (2.2)$$

In order to account for substructure in the data, the covariance structure of the data is considered. If there are known subgroups within the data, as in a designed experiment, there will be within-group covariance and between-group covariance. The within-group and between-group covariances will obviously be different. The original method described by Tadesse et al[9] treats the covariance between the individuals as the same, regardless of whether the individuals are in the same group or in different groups. Therefore Swartz et al[11] improve the method by accounting for this within-group and between-group difference in covariance. One way to do this is to construct a formal Bayesian model using blocked covariance matrices in the likelihood and/or priors that adequately reflect the within- and between-group variance structures. This approach, however, requires at least a $p \times p$ blocked covariance matrix and introduces a large number of parameters, especially in scenarios where $p \gg n$, bringing instability into the model. To avoid this, the within-group covariance structure is approximated and the cluster allocation which reflects subgroups in the data is modified.

Here the structure on the data via the definition of the cluster allocation vector, $\mathbf{y}$, is imposed. This vector now has elements indicating subgroups, that is, blocks of observations, rather than individual observations. Thus, all individuals in a given subgroup will be always assigned to the same cluster. When clustering the data, the original subgroups may collapse into bigger groups but they cannot be further divided into smaller groups.

In order to do variable selection a latent indicator to select the discriminatory gene expression values that best cluster the data is employed. Let $\boldsymbol{\gamma}$ be such an indicator vector, where $\gamma_j=1$ if the jth expression level (variable) contributes to differentiating the clusters and $\gamma_j=0$ if the jth variable is nondiscriminatory. This generates a likelihood that is a product of the mixture model (2.1) and a single multivariate normal distribution that models the nondiscriminating variables. Parameters $(\boldsymbol{\gamma})$ and $(\boldsymbol{\gamma}^c)$ are used to index the discriminating variables and those that do not discriminate, respectively.

Recall that $p(y_i = k)=w_k$. In the likelihood calculation there is a need to compute the exponent of

the term corresponding to the weights w_k based on the number of subgroups belonging to cluster k (denoted m_k), rather than on the number of individuals in cluster k (denoted n_k). The likelihood function is as follows:

$$L(G,\gamma,\mathbf{w},\mu,\Sigma,\eta,\Omega|\mathbf{x},\mathbf{y}) = (2\pi)^{\cdots} \times \exp\left\{-\frac{1}{2}\sum_{i=1}^{\cdots}\left(x_{i(\gamma)}-\eta_{(\gamma)}\right)'\Omega_{(\gamma)}^{-1}\left(x_{i(\gamma)}-\eta_{(\gamma)}\right)\right\} \times \prod_{k=1}^{\cdots}(2\pi)^{\cdots}\left|\Sigma_{k(\gamma)}\right|^{\cdots} \times \exp\left\{-\frac{1}{2}\sum_{i\in C_k}\left(x_{i(\gamma)}-\mu_{k(\gamma)}\right)'\Sigma_{k(\gamma)}^{-1}\left(x_{i(\gamma)}-\mu_{k(\gamma)}\right)\right\} \quad (2.3)$$

In the Equation (2.3), C_k denotes the kth mixture component, μ_k denotes its mean, and $\boldsymbol{\eta}$ the mean of the nondiscriminatory distribution. Likewise, Σk and Ω denotes the variance-covariance matrices. Notice that this likelihood depends on n, the total number of samples, n_k, the number of samples allocated to cluster k and also on m_k, the total number of subgroups allocated to component k, unlike the likelihood of Tadesse et al,[9] which is only a function of n and n_k.

The indicator variables are modeled as independent Bernoulli random variables, with common probability parameter φ. The φ is elicited as the expected proportion of the variables that will be discriminating a priori. A natural prior for the number of clusters, G, is a truncated Poisson, with rate parameter λ:

$$P(G=g)=\frac{e^{-\lambda}\lambda^{g}/g!}{1-\left[e^{-\lambda}(\lambda+1)+\sum_{\cdots}^{\cdots}e^{-\lambda}\lambda^{i}/i!\right]},\ g=2,\ \ldots,\ G\max. \quad (2.4)$$

For the vector of component weights, we use a symmetric Dirichlet prior, $\mathbf{w}|G \sim Dirichlet(a,\ldots,a)$.

For the component means and variances, as well as the mean and variance of the non discriminating variables, the usual conjugate priors are used.

$$\mu_{k(\gamma)}|\Sigma_{k(\gamma)},G \sim N\left(\mu_{0(\gamma)},h_0\Sigma_{k(\gamma)}\right),\quad \eta_{(\gamma)}|\Omega_{(\gamma)} \sim N\left(\mu_{0(\gamma)},h_1\Omega_{(\gamma)}\right),\quad \Sigma_{k(\gamma)},G \sim IW\left(\delta,\mathbf{Q}_{0(\gamma)}\right),\quad \Omega_{(\gamma)} \sim IW\left(\delta,\mathbf{Q}_{0(\gamma)}\right) \quad (2.5)$$

Here, $IW(\delta,\mathbf{Q}_{0(\gamma)})$ denotes the inverse-Wishart distribution, with shape parameter $\delta = n - p_\gamma + 1$, dimension p_γ, degrees of freedom n, and mean $\mathbf{Q}_{1(\gamma)/\delta-2}$. Also, as in Tadesse et al,[9] $\delta = 3$ is used to denote an uninformative prior and define $\mathbf{Q}_1=1/\kappa_1\mathbf{I}_{p\times p}$ and $\mathbf{Q}_0=1/\kappa_0\mathbf{I}_{p\times p}$, where κ_1 and κ_0 are defined respectively as proportional to the upper and lower deciles of the $n-1$ nonzero eigenvalues of cov($\mathbf{X}$). These choices follow the guidelines given by Tadesse et al.[9] matrix. Some sensitivity to the parameter choices is typical of any model-based clustering method. For the mean parameters, each element of $\boldsymbol{\mu}_0$ was set to the midpoint of the range of the variable, and h_0 and h_1 were chosen arbitrarily large, between 10 and 1 000, for flat priors. For more details on regarding the hyper-prior parameters.[9] $\pi_{\mathbf{w}|G},\pi_{\mu_{0(\gamma)}|\eta_{(\gamma)}},h_0,h_1,\delta,\mathbf{Q}_{0(\gamma)},\mathbf{Q}_{1(\gamma)}$

The mean and variance parameters were expertly integrated out in Tadesse et al.,[9] and the modification described above, is constant with respect to these parameters, and therefore does not change the integration calculations. Thus, even after accounting for substructure, it is only necessary to update the parameters ($\boldsymbol{\gamma}$,$\mathbf{w}$, $\mathbf{y}$, G). The simulation from the posterior is done by using a hybrid Gibbs sampler and Metropolis–Hastings algorithm that iterates sampling from the following distributions:

$$f(\mathbf{y}|G,\mathbf{w},\gamma,\mathbf{X}) \propto f(\mathbf{X},\mathbf{y}|G,\mathbf{w},\gamma) \quad (2.6)$$

$$f(\gamma|G,\mathbf{w},\mathbf{y},\mathbf{X}) \propto f(\mathbf{X},\mathbf{y}|G,\mathbf{w},\gamma)p(\gamma|G) \quad (2.7)$$

$$\mathbf{w}|G,\gamma,\mathbf{y},\mathbf{X} \sim Dirichlet(a-\kappa_1,\ldots,a-\kappa_G) \quad (2.8)$$

The vector is updated via (2.7) using the Metropolis search algorithm that has now become quite standard in variable selection.[9,12]

At a single iteration the vector $\boldsymbol{\gamma}$ is updated either by swapping two of its elements or by randomly selecting one element and changing its value from 0 to 1 or 1 to 0. The cluster allocation vector $\mathbf{y}$ is updated element by element using a Gibbs sampling strategy via Equation (2.6). According to our modified model, each element of $\mathbf{y}$ corresponds to an experimental group. The full conditional probability that the ith experimental group is in the kth cluster is therefore calculated as

$$f(y_i=k|\mathbf{X},\mathbf{y}_{(-i)},\gamma,\mathbf{w},G) \propto f(\mathbf{X},y_i=k,\mathbf{y}_{(-i)}|G,\mathbf{w},\gamma) \quad (2.9)$$

Here, $\mathbf{y}_{(-i)}$ is the standard notation denoting the vector of cluster assignments for all subgroups except the ith subgroup.

The weights are updated by Gibbs sampler via Equation (2.8). The calculations are simplified by sampling independent gamma random variables with common scale and shape parameters (a+ n_1, ..., $a + n_G$), and scaling the random variates to sum to 1. As in the original model formulation of Tadesse et al[9] the number of clusters, G, is unknown and updated using reversible jump Markov chain Monte Carlo (RJMCMC) technology.[13,14] The RJMCMC construction updates G using a split/merge cluster move, and a birth/death move as in Tadesse et al.[9] However, to calculate the acceptance ratio, the new likelihood (2.3) is used, and accounts for using the experimental subgroups as items to be clustered.

In order to make inference from the posterior samples, first the method proposed by Stephens[15] is used to resolve cluster identifiability. Once the clusters are suitably relabeled to be consistent across all iterations, frequency approximations are calculated to the posterior probabilities since the quantities of interest are multinomial or binomial random variables. From experience, these frequency estimates are more robust to correlation that may be present in the Markov chain than calculating the marginal posterior probabilities—especially when analyzing real data.[16] For inference on cluster memberships, the most probable number of clusters is conditioned and counts how many iterations each experimental group appear in each cluster. For inference on the variables, the numbers of iterations that each variable is selected are counted and divide that by the total number of iterations kept after burn in. For the simulations below, similar distributions using the posterior probability calculations detailed in Tadesse et al.[9] and modified frequency approximations are found.

RESULTS

Recall that our data consist of four groups of patients: low invasion, control; high invasion, control; low invasion, SCC with metastasis; high invasion, SCC with metastasis, and the original microarrays consisted of 270 genes. For preprocessing, data were normalized by using the global median method. That is, for each array, the expression value of each gene was divided by the median expression value of the expressions on the array. Here we apply our method to these 270 expression values, with the purpose of further refining the gene discovery.

To prepare for clustering, we divided each of the 270 median normalized gene expression values by the range of that gene. Next, prior parameters were set as follows: Parameters $\kappa_0 = 0.0232$ and $\kappa_1 = 0.0972$ were chosen, proportional to the first and last decile of the nonzero eigenvalues as our covariance parameters. The prior means for both cluster mixtures and the nondiscriminating distributions were set as $\mu_0 = \frac{1}{2}(\max(x_j) + \min(x_j))$. The symmetric Dirichlet distribution prior parameter was set as $\alpha = 1$, the truncated Poisson distribution prior rate parameter as $\lambda = 5$; and the prior probability for the Bernoulli distribution $p\varphi = 10$. Here the prior covariance matrices $\Omega_{\alpha=1}$ and $\Omega_{\alpha=0}$ were defined to be diagonal matrices with diagonal elements equal to the variances of each gene. Incorporating empirical variances has been shown to improve variable selection.[17] Two MCMC chains were run. Both chains were run for 1 000 000 iterations, using the last 60 000 iterations for inference and the rest were considered burn-in. The first chain started with 100 randomly selected genes, and using each subgroup as an initial cluster. The second chain started with 50 randomly selected genes, and using two clusters: low invasion control with SCC metastasis, and high invasion control with SCC metastasis patients.

The final inference was performed using the pooled sets of samples from the two MCMC chains. After pooling the chains, the data were grouped into two clusters and selected 17 genes using the marginal median model as cutoff. Cluster membership probabilities for each patient are reported in Table 1. These clearly separate control patients from the others.

For comparison, we applied FDR multiple testing correction Benjamini and Hochberg[18] to the p-values for each gene. This is a standard method

TABLE 1: Real data: Probability of cluster memberships.

Patients group	p (member of cluster 1)	p (member of cluster 2)
low invasion control	0.6044	0.3956
low invasion control	0.6044	0.3956
low invasion control	0.6044	0.3956
low invasion SCC with metastasis	0.2270	0.7730
low invasion SCC with metastasis	0.2270	0.7730
low invasion SCC with metastasis	0.2270	0.7730
high invasion control	0.8340	0.1660
high invasion control	0.8340	0.1660
high invasion control	0.8340	0.1660
high invasion SCC with metastasis	0.1673	0.8327
high invasion SCC with metastasis	0.1673	0.8327
high invasion SCC with metastasis	0.1673	0.8327

commonly used in microarray analysis. Three of the 17 identified genes, E-cadherin, CD44 and p53 were also selected by the FDR method. A level of 0.1 detected a larger number of genes, and included the three genes we identified. The selection of a small set of genes is advantageous here. A small number of selected genes is appealing to biologists because they constitute a manageable set of candidates on which further studies can be performed via biological assays. Of course, if necessary, more genes can be selected by the Bayesian method by lowering the threshold of the 50% median model that was used.

Lip carcinoma is one of the most common malignant tumors in oral and maxillofacial region. In advanced stages with regional metastasis it has a poor prognosis. E-cadherin and CD44 molecules play a role in cell-to-cell adhesion; p53 is associated with cellular proliferation and cell death.[19-25]

Since these patients had a poor prognosis, it seems that decreased E-cadherin and CD44 expression and over expression of p53 in cancerous tissue correlates with this outcome in lip carcinoma patients. Detection of the expression of these proteins is useful to confirm the risk for cervical lymph node metastasis; further studies are encouraged to reveal the detail mechanisms in formation of lymph node metastatic focus.

The description above highlights the fact that the three selected genes, E-cadherin, CD44 and p53, are biologically consistent with current research and that provide strong biological validation of the cluster configuration suggested.

DISCUSSION

A method that takes advantage of known substructure in the data when simultaneously clustering high-dimensional data with an unknown number of clusters, in order to select the best discriminating variables for those clusters, was applied. Given the structure of designed experiments, breaking the basic experimental structure would have no interpretation with regard to the experiment. This method approximates stronger within design group covariance by defining the cluster member indicator vector $\mathbf{y}$ to assign all members of a design group to the same cluster. The approach is similar to the idea of forcing the elements of the original vector $\mathbf{y}$, indexed over individuals rather than subgroups, into subsets where all entries in the same subset have the same value. In this approach the likelihood is adjusted to compute the proper probability that corresponds with the reduced variation. Additionally, by jointly finding structure in the data and selecting variables, here genes, we answer the researchers' questions of, first, whether the design groups affect the subjects differently and, second, which genes define those differences.

The true correlation of gene expression values is quite complex, and modeling this correlation structure is an interesting research question in its own right. The underlying covariate selection mechanism used for the selection of the discriminating variables has been shown to be effective in analyzing correlated covariates in studies with genetic markers, which is simpler to model than gene expression correlation.[17,26]

When analyzing real data, three genes, E-cadherin, CD44 and p53, which agree with current biological research and literature and that provide biological validation of the cluster configuration were found. Overall, the method applied can provide biologists with both useful and manageable information for further experimental research.

REFERENCES

1. Scott DW. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley Series in Probability and Mathematical Statistic: Applied Probability and Statistics. A Wiley-Interscience Publication. 1st ed. New York: John Wiley&Sons, Inc; 1992. p.137.
2. Tadesse M, Sha N, Kim S, Vanucci M. Identification of Biomarkers in Classification and Clustering of High-Throughout Data. In: Do K, Müller P, eds. Bayesian Inference for Gene Expression and Proteomics.1st ed. New York: Cambridge University Press; 2006. p.97-115.
3. Sebastiani P, Ramoni M, Kohane IS. Bayesian clustering on gene expression dynamics. In: Parmigiani G, Garrett E, Irizarry A, Zeger S, eds. The Analysis of Gene Expression Data: Methods and Software. New York: Springer; 2003. p.409-24.
4. Efron B, Tibshirani R, Storey J, Tusher V. Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 2001;96(456):1151-60.
5. Newton M, Kendziorski C, Richmod C, Blattner F, Tsui K. On differential variability of expression ratio: Improving statistical inference about gene expression changes from microarray data. J Comput Biol 2001;8(1):37-52.
6. Ibrahim J, Chen M, Gray R. Bayesian models for gene expression with DNA microarray data. J Am Stat Assoc 2002;97(457):88-99.
7. Kendziorski C, Newton M, Lan H, Gould M. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Stat Med 2003;22(24): 3899-914.
8. Brusco M, Cradit J. A variable selection heuristic for k-means clustering. Psychometrika 2001;66(2):249-70.
9. Tadesse M, Sha N, Vannucci M. Bayesian variable selection in clustering high dimensional data. J Am Stat Assoc 2005;100(470): 602-17.
10. Raftery A, Dean N. Variable selection for model-based clustering. J Am Stat Assoc 2006; 101(473):168-78.
11. Swartz M, Mo Q, Murphy M, Lupton J, Turner N, Young Hong M, et al. Bayesian variable selection in clustering high-dimensional data with substructure. JAES 2008; 13(4):407-23.
12. Sha N, Vannucci M, Tadesse MG, Brown PJ, Dragoni I, Davies N, et al. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. Biometrics 2004;60(3):812-9.
13. Green P. Reversible-jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 1995;82(4):711-32.
14. Richardson S, Green P. On Bayesian analysis of mixtures with an unknown number of components (with discussion). J R Stat Soc B 1997;59(4):731-92.
15. Stephens M. Dealing with label switching in mixture models. J R Stat Soc B 2000;62(4): 795-809.
16. Kim S, Tadesse M, Vannucci M. Variable selection in clustering via Dirichlet process mixture models. Biometrika 2006;93(4):877-93.
17. Swartz M, Kimmel M, Mueller P, Amos C. Stochastic search gene suggestion: a Bayesian hierarchical model for gene mapping. Biometrics 2006;62(2):495-503.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B 1995;57(1):289-300.
19. Bülbül Başkan E, Balaban Adım Ş, Sarıcaoğlu H, Tunalı Ş, Tokgöz N. [The evaluation of p53 protein expression in squamous cell carcinoma: A retrospective study]. Turkiye Klinikleri J Dermatol 2003;13(4):215-20.
20. Carinci F, Lo Muzio L, Piattelli A, Rubini C, Palmieri A, Stabellini G, et al. Genetic portrait of mild and severe lingual dysplasia. Oral Oncology 2005;41(4):365-74.
21. Diniz-Freitas M, García-Caballero T, Antúnez-López J, Gándara-Rey J, García – García A. Reduced E-cadherin expression is an indicator of unfavourable prognosis in oral squamous cell carcinoma. Oral Oncology 2006;42(2):190-200.
22. Han Ö, Arık D, Seçkin S. [P53 mutation and nm23 expression in laryngeal squamous cell carcinoma and relationship between histological prognostic parameters]. Turkiye Klinikleri J Med Sci 2005;25(3):348-53.
23. Lopes FF, da Costa Miguel MC, Pereira AL, da Cruz MC, de Almeida Freitas R, Pinto LP, et al. Changes in immunoexpression of E-cadherin and beta-catenin in oral squamous cell carcinoma with and without nodal metastasis. Ann Diagn Pathol 2009;13(1):22-9.
24. Molinolo AA, Amornphimoltham P, Squarize CH, Castilho RM, Patel V, Gutkind JS. Dysregulated molecular networks in head and neck carcinogenesis. Oral Oncol 2009;45(4-5):324-34.
25. Tanaka N, Odajima T, Ogi K, Ikeda T, Satoh M. Expression of E-cadherin, α-catenin, and β-catenin in the process of lymph node metastasis in oral squamous cell carcinoma. Br J Cancer 2003;89(3):557-63.
26. Swartz MD, Shete S. The null distribution of stochastic search gene suggestion: A Bayesian approach to gene mapping. BMC Proc 2007;1(Suppl 1): S113.