RESEARCH Open Access

A nonparametric Bayesian approach for clustering bisulfate-based DNA methylation profiles

Lin Zhang[1], Jia Meng[2], Hui Liu[1], Yufei Huang[2,3]*

From IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS) 2011
San Antonio, TX, USA. 4-6 December 2011

Abstract

Background: DNA methylation occurs in the context of a CpG dinucleotide. It is an important epigenetic modification, which can be inherited through cell division. The two major types of methylation include hypomethylation and hypermethylation. Unique methylation patterns have been shown to exist in diseases including various types of cancer. DNA methylation analysis promises to become a powerful tool in cancer diagnosis, treatment and prognostication. Large-scale methylation arrays are now available for studying methylation genome-wide. The Illumina methylation platform simultaneously measures cytosine methylation at more than 1500 CpG sites associated with over 800 cancer-related genes. Cluster analysis is often used to identify DNA methylation subgroups for prognosis and diagnosis. However, due to the unique non-Gaussian characteristics, traditional clustering methods may not be appropriate for DNA and methylation data, and the determination of optimal cluster number is still problematic.

Method: A Dirichlet process beta mixture model (DPBMM) is proposed that models the DNA methylation expressions as an infinite number of beta mixture distribution. The model allows automatic learning of the relevant parameters such as the cluster mixing proportion, the parameters of beta distribution for each cluster, and especially the number of potential clusters. Since the model is high dimensional and analytically intractable, we proposed a Gibbs sampling "no-gaps" solution for computing the posterior distributions, hence the estimates of the parameters.

Result: The proposed algorithm was tested on simulated data as well as methylation data from 55 Glioblastoma multiform (GBM) brain tissue samples. To reduce the computational burden due to the high data dimensionality, a dimension reduction method is adopted. The two GBM clusters yielded by DPBMM are based on data of different number of loci (P-value < 0.1), while hierarchical clustering cannot yield statistically significant clusters.

Background

DNA methylation profiles has become an alternative molecular footprint for classification. It occurs in the context of a CpG dinucleotide. It is an important epigenetic modification, which can be inherited through cell division. In this chemical modification of the cytosine nucleotide, the 5-carbon position is enzymatically modified by the addition of a methyl group such that cytosines can occur in a methylated or unmethylated state. CpG islands are usually not methylated in normal tissues but frequently become hypermethylated in cancer [1]. This hypermethylation is associated with gene silencing [2] and plays an important role in the inactivation of tumor suppressor genes. Most CpGs or CpG regions have been found to have a bimodal distribution of methylation profiles, either hypomethylated or hypermethylated [3]. Unique methylation patterns have been shown to exist in diseases including various types of cancer [4]. DNA methylation analysis promises to become a powerful tool in cancer diagnosis, with possible applications to the choice of treatment and prognostication. The high throughput methylation profiling technology has been developed to survey methylation

* Correspondence: yufei.huang@utsa.edu
[2]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249, USA
Full list of author information is available at the end of the article

status of more than 1500 CpG sites for a large collection of cancer genes and been specifically targeting. Studying how the methylation profiles can be used to distinguish different subtypes of the tumor has been a focus in current cancer research. But most existing algorithms working on methylation data are from sequence level. The exact levels of methylation expression are not fully considered yet.

To this end, clustering analysis is often used to identify methylation subgroups that are distinct from one another in data [5,6]. However, the DNA methylation data presents unique challenges. First, it is not appropriate to cluster DNA methylation expressions using traditional clustering methods. The traditional k-means clustering algorithms are based on Gaussian Mixture Model (GMM) assumptions. In GMM, the individual data points are assumed to follow multivariate Gaussian distribution and thus the distance between two points can be evaluated by Euclidean distance conveniently. However, since "beta" values from DNA methylation array represent the percentage of the methylated alleles and are between 0[1], traditional GMM is no longer appropriate. Instead, a mixture of the beta distribution [7,8] would be a more accurate model. Second, a model selection process is often needed in clustering to determine the number of clusters, making the clustering analysis more complicated. A predefined number of clusters (or model) is required in the mixture distribution based methods (such as k-means). Since different number of clusters will yield different clustering results, a model selection process is desirable to determine the best number of clusters. The model selection is very different problem, whose optimal solution is of exponential complexity. The popular suboptimal solutions have been proposed that include minimum description length (MDL) and Bayesian information criterion (BIC). Although computationally efficient, these methods would fail when clusters are not well separated. The recent proposed nonparametric Bayesian methods including Dirichlet process (DP) provide an avenue that can lead to a better solution.

In a response to the aforementioned limitations, we proposed here a nonparametric Dirichlet process beta mixture model (DPBMM) method for clustering DNA methylation expression profiles produced by Illumina Infinium Beadchip. DPBMM makes use of Dirichlet process mixture to place a prior [9] on cluster assignment, thus enables automatic determination of the optimal number of clusters. To perform the analytical intractable learning, an algorithm based on Gibbs sampling and "no-gap" sampling is developed to effectively infer all the relevant variables. The proposed DPBMM method builds an infinite beta mixture model to describe DNA methylation data, which is different from the finite beta mixture model in [8]. We present a simulation study comparing its properties to RPMM (Recursively partitioned mixture model) employing BIC (Bayesian information criterior) in [8]. The results demonstrated the better performance of our proposed method. Finally, we applied the DPBMM to the methylation array obtained from 55 Glioblastoma Multiform (GBM) brain tissue samples.

Methods

Problem formulation

Model DNA methylation profiles with beta mixture distribution

For a two-color hybridization based array such as Illumina Infinium array, the measurements are associated with the percentage of the methylated alleles, which is called the "beta" values because it can be described by a mixture of beta distributions [7,10]. Since the distribution of "beta" values shows bimodalities [11], the beta distribution component in the mixture model should be convex, which means the beta distribution component should be equipped with large parameters, shown in Figure 1.

Consider the problem of clustering n independent DNA methylation samples, let $X = \{X_1, X_2, ..., X_n\}$ be the DNA methylation expressions for n samples. For each sample i, $X_i = \{x_{i1}, x_{i2}, ..., x_{iL}\}$ be a vector of L continuous outcomes falling between zero and one. Suppose there exists a total K clusters and sample i belongs to cluster class $c_i \in \{1, ..., K\}$. Conditional on class membership say k, each outcome x_{il} could be viewed as an independent identically distributed variable from a beta distribution with α_{kl} and β_{kl}

$$f(x_{il}|\alpha_{kl}, \beta_{kl}, c_i = k) = \frac{x_{il}^{\alpha_{kl}-1}(1-x_{il})^{\beta_{kl}-1}}{B(\alpha_{kl}, \beta_{kl})} \tag{1}$$

where $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ stands for the Beta function. Then, DNA methylation sample X_i can be modeled by (2).

$$p(X_i|\theta) = \sum_{k=1}^{K} \pi_k \prod_{l=1}^{L} \frac{x_{il}^{\alpha_{kl}-1}(1-x_{il})^{\beta_{kl}-1}}{B(\alpha_{kl}, \beta_{kl})} \tag{2}$$

where $\theta_l = \{\alpha_{kl}, \beta_{kl}, \forall l\}$. With the limitation of large parameters for beta distribution component, $\alpha_{kl} >1$ and $\beta_{kl} >1$. Note that due to clustering in samples, θ_l and θ_i for $i \neq l$ may be equal, π_k's represent the cluster proportion and $\sum_{k=1}^{K} \pi_k = 1$. Now, in reality, the total cluster number K is not known *a priori*. We discuss next a model based on Dirichlet process to address this difficulty.

Dirichlet process mixture model

The Dirichlet process is an nonparametric extension of the original Dirichlet distribution. Let x_i be a random sample from a distribution f with parameters θ_i. In a Bayesian formulation, the model for parameter θ_i can be defined as

$$\begin{aligned} x_i|\theta_i &\sim f(\theta_i) \\ \theta_i|G &\sim G \end{aligned} \tag{3}$$

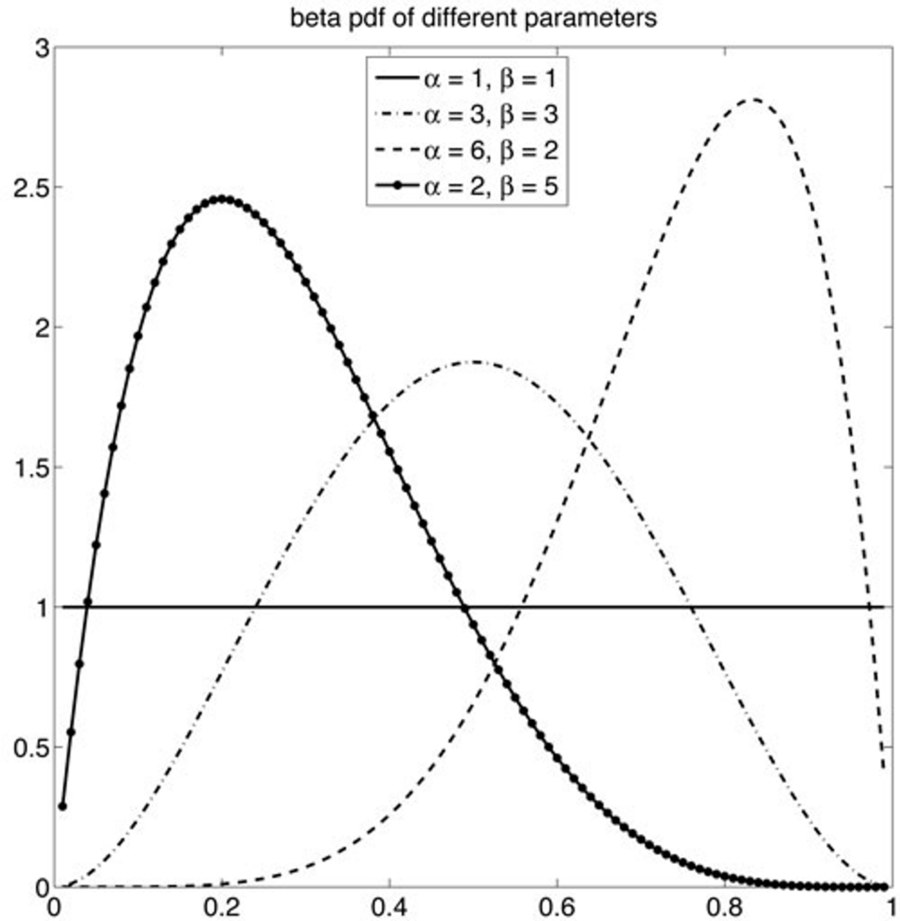

Figure 1 Examples of beta distributions. Beta densities with large hyperparameters ($\alpha > 1$, $\beta > 1$) are unimodal.

where G is the prior distribution of θ_i. It is not always realistic to assume that G is of a known form and the nonparametric Bayesian models including the Dirichlet process (DP) is proposed to address this problem. Now, instead of defining a parametric form for G, G is assumed to be a draw from a Dirichlet process with a base distribution G_0 and a precision parameter τ [12]. The model for the Bayesian estimation is also built in Figure 2 following the principles of graphical models. It can also be written as (4) with a DP prior.

$$\begin{aligned} X_i|\theta_i &\sim f(\theta_i) \\ \theta_i|G &\sim G \\ G|\tau, G_0 &\sim DP(\tau, G_0) \end{aligned} \tag{4}$$

where G_0 is such that $E[G] = G_0$ and has a parametric form, τ measures the strength of belief in G_0. The DP of mixtures (DPM) are proposed to model the clustering effect in data. Compared with other clustering models, DPM is very attractive because it allows the cluster number K to be *a priori* ∞ and learned from the data. To capture the clustering natural of DNA methylation samples, a beta mixture model with infinite classes can be built with DPM. Let $\theta_i = \{\alpha_i, \beta_i\}$ be the set of parameters for each sample and note that some of them may be equal. In DPM models, each θ_i is marginally sampled from G_0, and with positive probability some of the θ_i are identical due to the discreteness of the random measure G. Therefore the new value of θ_i can either be one of the θ_l's$(l \neq i)$, or θ_i could be a new draw from G_0. Let K in (2) be ∞, we assume a DPBMM for DNA methylation array.

Inference

Let $\Phi = \{\Phi_1, \Phi_2, ..., \Phi_K\}$ denote the set of distinct θ_i's, where K is the number of distinct elements of $\theta_1, ..., \theta_m$. Let $s = \{s_1, ..., s_m\}$ denote cluster assignment vector, that means, $s_i = l$ if and only if $\theta_i = \varphi_l$. Then $\theta = \{\theta_i : i = 1, ..., m\}$ can be reparameterized as $\{\varphi_1, ..., \varphi_k, s_1, ..., s_m\}$. Let n_i, $i = 1, ..., K$ be the number of elements s_l equal to i. Let subscript "*-i*" stands for all the variables except the i-th one. The goal from a Bayesian perspective is to calculate the

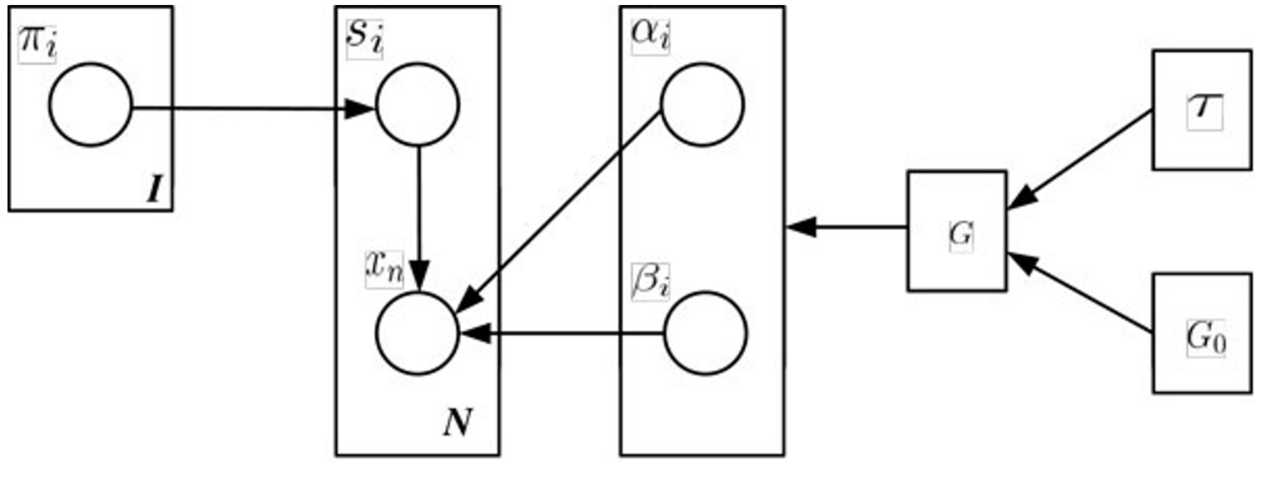

Figure 2 Graphical model. The model for the Bayesian estimation is built following the principles of graphical model.

posterior distribution of the known parameters {Θ, *π*, *τ*}. However, the analytical expression is intractable and we instead develop a Gibbs sampling solution to obtain random samples from the posterior distribution. The key for Gibbs sampling is to derive the conditional posterior distributions of the unknown parameters. Due to the constrains on α and β, we first re-parameterize α as L_α by $\alpha = exp(|L_\alpha|)$ and β as L_β by $\beta = exp(|L_\beta|)$. Thus, we only need to sample in the range of (-∞, ∞) for L_α and L_β. Then the transformed α >1 and β >1. Thus, we can specify G_0 as $G_0(\alpha, \beta) = \mathcal{N}(0, \sigma_\alpha^2)\mathcal{N}(0, \sigma_\beta^2)$, where $\mathcal{N}(\mu, \sigma^2)$ represents the Gaussian distribution with mean μ and variance σ^2 [13]. The prior distribution of the cluster proportion π is the Dirichlet distribution

$$\pi \sim Dir(n_1 + \tau/K, ..., n_K + \tau/K). \tag{5}$$

There are some useful expression of a Dirichlet process, such as Chinese Restaurant Process(CRP) [14,15], Stick-breaking construction [16], Polya Urn formulation [17,18], etc... Blackwell showed that Dirichlet process are discrete as they consist of countably infinite point probability masses [19]. Escobar and West [20] first showed that Markov Chain Monte Carlo (MCMC) techniques, specifically Gibbs sampling, could be used for posterior density estimation if the Blackwell-MacQueen Polya Urn formulation of Dirichlet process is used. Based on the generalized Polya urn scheme, the conditional prior distributions $s_i|s_1, ..., s_{i-1}$, $i = 1, ..., n$ and $\theta_i|\theta_{-i}$ have the following forms as (6) and (7).

$$\begin{gathered} P(s_1 = 1) = 1 \\ \begin{cases} P(s_i = l|s_1, ..., s_{i-1}) = \frac{n_{-i,l}}{(\tau+i-1)}, l = 1, ..., k_i \\ P(s_i = k_i + 1|s_1, ..., s_{i-1}) = \frac{\tau}{(\tau+i-1)} \end{cases} \end{gathered} \tag{6}$$

and,

$$\begin{aligned} &\theta_1 \sim G_0(\theta_1) \\ &\theta_i|\theta_1 ... \theta_{i-1} \sim \frac{\tau}{\tau+i-1} G_0(\theta_i) + \sum_{l=1}^{K} n_l \frac{1}{\tau+i-1} \delta_{\Phi_l}(\theta_i), \text{ for } i \geq 1. \end{aligned} \tag{7}$$

Then the conditional posterior distribution for sampling θ_i has the form

$$\begin{aligned} p(\theta_i|\theta_{-i}, s_{-i}, X) &\propto q_{i,0} G_i(\theta_i) + \sum_{l=1, l \neq i}^{n-1} q_{i,l} \delta_{\theta_l}(\theta_i) \\ &= q_{i,0} G_i(\theta_i) + \sum_{l=1}^{K} n_{-i,l} q_{i,l} \delta_{\Phi_l}(\theta_i). \end{aligned} \tag{8}$$

Thus the conditional posterior distribution for sampling Φ_i has the form

$$\begin{aligned} &p(\Phi_i|\Phi_{-i}, s, X, \pi) \\ &\quad \propto p(X_{m:s_m=s_i}|\Phi, s, \pi) p(\Phi_i|\Phi_{-i}, s, \pi) \\ &\quad = G_0 \prod_{m:s_m=s_i} \prod_{l=1}^{L} \frac{x_{ml}^{\alpha_{kl}-1}(1 - x_{ml})^{\beta_{kl}-1}}{B(\alpha_{kl}, \beta_{kl})} \end{aligned} \tag{9}$$

It is obvious that G_0 is not conjugate with *f*, so the integral $q_{i,0}$ cannot be evaluated analytically and drawing samples from G_i is also extremely challenging [21]. To overcome the difficulty, we adopt the "no-gaps" algorithm proposed in [22] to enable sampling from (8).

As to τ, it is useful to choose a weakly informative prior in many applications. If τ is assigned a gamma prior, its posterior becomes a simple function of *K*, then samples are easily drawn via an auxiliary variable method. For the convenience of sampling, we adopt the $\tau \sim Gamma(a, b)$ as the prior [9,20].

The final Gibbs sampling steps can be summarized by the following steps:

Gibbs sampling for DPBMM

Iterate the following steps and for the t-th iteration:

1. For each sample i, re-sample s_i according to (6) if $n_{s_i} > 1$. In this case $k_{-i} = K$. If $n_{s_i} = 1$, then with probability 1 - $1/K$ leave s_i unchanged. With probability $1/K$ rearrange s such that $s_i = K$, then re-sample s_i according to (6). But in this case $k_{-i} = K - 1$.
2. For $i = 1, ..., K$, the posterior distribution for Φ_i has the form as (9).
For $i = K + 1, ..., n$, both prior and posterior distribution for Φ_i are G_0.
3. Sample π following (5) with $n_k = \sum_{i=1}^{n} \delta(s_i, k)$.

4. Based on Step 1, we can get the value of K, then sample $\tau|K, n$ where $\tau \sim Gamma(a, b)$.

Due to the large number of parameters, the initial values for parameters σ_α^2 and σ_β^2 should be chosen carefully.

Results

Test on simulated data

We conducted simulations to test our proposed method. For the first case, the simulated data set is generated based on the model described in (2) with $K = 4$. The simulated dataset consists of 100 samples, each having 200 continuous response lying in the unit interval. The occurring probability of each cluster is set to {0.2, 0.3, 0.2, 0.3}. For each cluster, parameters L_α, L_β related to beta distribution in the model are generated randomly from Gaussian distributions with zero means and different variances. In order to systematically evaluate the clustering performance, the F metric that combines BCubed overall precision and recall [23] was implemented as suggested in [24]. Let $\{c\}$ represent the real cluster label of samples and $\{s\}$ represent the cluster assignment by clustering method, the correctness of the relation between sample i and i' is defined as $Ct(i, i')$ based on $\{c\}$ and $\{s\}$.

$$Ct(i,i') = \begin{cases} 1 & \text{iff } c_i = c_{i'} \leftrightarrow s_i = s_{i'}; \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The overall precision P and recall R are defined as

$$\begin{aligned} R &= Avg_i[Avg_{i'.c_i=c_{i'}}[Ct(i,i')]] \\ P &= Avg_i[Avg_{i'.s_i=s_{i'}}[Ct(i,i')]] \end{aligned} \tag{11}$$

F metrics is used to evaluate the clustering result by combining P and R metrics.

$$\mathcal{F}(R,P) = \frac{1}{0.5/P + (1-0.5)/R} \tag{12}$$

Figure 3(a) illustrates the sampled number of clusters in each Gibbs sampling iteration for one time of DPBMM clustering. After 300 iterations of “burn-in” stage, the number of clusters stay at four. The uncovered cluster proportion is {0.19, 0.31, 0.19, 0.31}. Figure 3(b-d) show that for 2000 times of DPBMM clustering, F metric can come to one for most times.

For our second case, we used two simulated data set from [8]. The data set of Case I consists of 100 subjects, which mimics the real methylation data. Each subject has 1413 continuous responses lying in the unit interval. Each subject was a member of five classes, each cluster occurring with 0.2 probability. The clusters were defined by beta-distribution parameters for each of 1413 methylation loci that were autosomal and passed quality-assurance, obtained by fitting a beta model on each locus to one of the five data sets from our normal data: adult blood, newborn blood, placenta, lung/pleura, and everything else. The data set of Case II considered 100 subjects from four clusters. We compare the performance with RPMM method proposed in [8], with the same dimension reduction method employed. We order all the loci with respect to variance, and the J most variable loci are considered in the clustering algorithm. Table 1 and Table 2 summarizes the number of classes found with RPMM and with our proposed DPBMM for both Case I and Case II. For the cases considered, DPBMM obtained the correct K with *a priori* ∞ directly while the RPMM fitted finite mixture models for a range of possible values and chose the correct K by BIC statistic. The F metric vs. recall curve of $J \in$ {25, 50} loci for case I is shown in Figure 4(a). The histogram of F metric results with $J = 50$ is shown in Figure 4(b). The F metric vs. recall curve of different $J \in$ {5, 10} loci for Case II is shown in Figure 4(c). The histogram of F metric results with $J = 10$ is shown in Figure 4(d). For the above two cases, the more the number of loci are considered in the clustering, the better clustering performance we can get.

Test on real data

We then applied our proposed DPBMM clustering on the GBM methylation microarray dataset in The Cancer Genome Atlas (TCGA). This dataset consists of 74 patients assayed on Illumina HumanMethylation450 array. Samples for DPBMM clustering analysis were selected to have clinical annotations. At last, 55 patients were left for consideration. Twenty-seven patients were alive at the time of last follow up, whereas twenty-eight patients experienced disease progression since last follow-up. The median follow up time was 198 days (range, 2-953 days). Each sample includes up to 485,577 CpG dinucleotides spanning gene-associated elements as well as intergenic regions. The associated detection P-value reported together with the methylation expression data is used as a

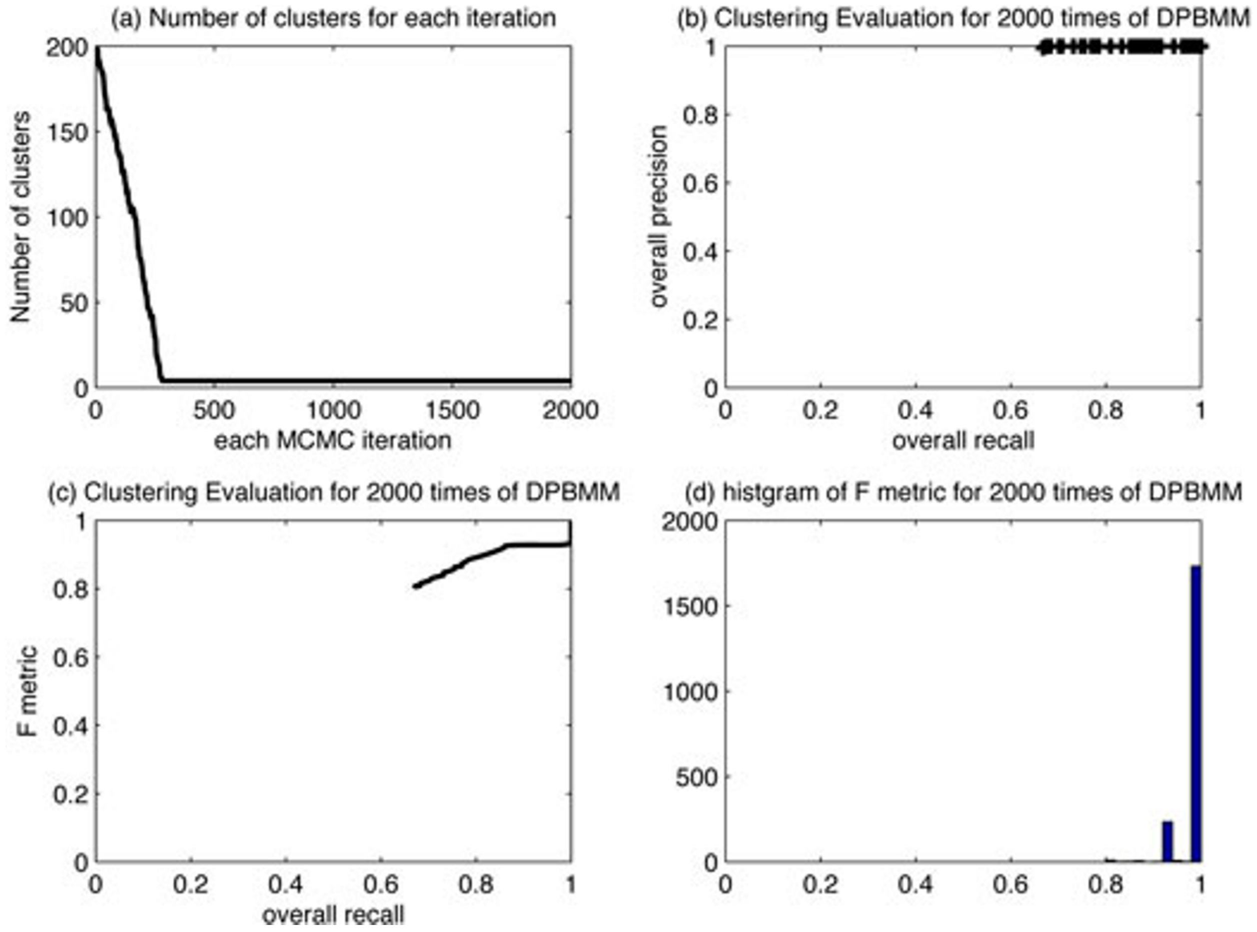

Figure 3 Clustering evaluation on simulation data set. The result is based on the simulated data with 4 dimensions. Figure 3(a) shows the number of clusters *k* in 2000 MCMC iterations. Figure 3(b) shows the overall precision vs. overall recall for 2000 times of DPBMM. The overall precision almost always stays at 1. Figure 3(c) shows the F metric vs. recall curve. Figure 3(d) shows the histogram of F metric results for 2000 times of DPBMM clustering.

quality control measure of probe performance. Following the probe excluding method in [25], the probes with detection P-values >0.01 in >10% of the samples are excluded from further consideration.

Since the small sample, large dimensional property of methylation array, many loci in the data set have low variance and may not contribute to clustering. it is safer only to consider loci that change significantly [26]. Thus, those loci with low variance across all 55 samples were removed from the data sets which is also used by [8]. This also made the DPBMM clustering process computationally more tractable. In this paper, we only consider $J \in \{1, 2, ..., 20\}$ most variable loci for DPBMM clustering method since the number of samples is only 55. The selected top 20 variable loci are listed in Table S1 (see Additional file 1). DPBMM yields two clusters from the data for most *J*. Kaplan-Meier survival analysis are carried out based on the clustering results, and the P-values of Kaplan-Meier confidence for $J \in \{1, 2, ..., 20\}$ are shown in Table S2 (see Additional file 2). Among these, J = 11

Table 1 Number of classes obtained for RPMM and DPBMM applied to simulated data (Case I: 5 classes).

Method	J	Median	Mean	SD
RPMM	25	8	7.7	2.0
	50	5	5.6	1.32
DPBMM	25	5	5.16	0.93
	50	5	5.29	1.43

Table 2 Number of classes obtained for RPMM and DPBMM applied to simulated data (Case II: 4 classes).

Method	J	Median	Mean	SD
RPMM	5	2	2.0	0.10
	10	2	2.4	2.38
DPBMM	5	7	6.9	1.04
	10	4	4.09	1.60

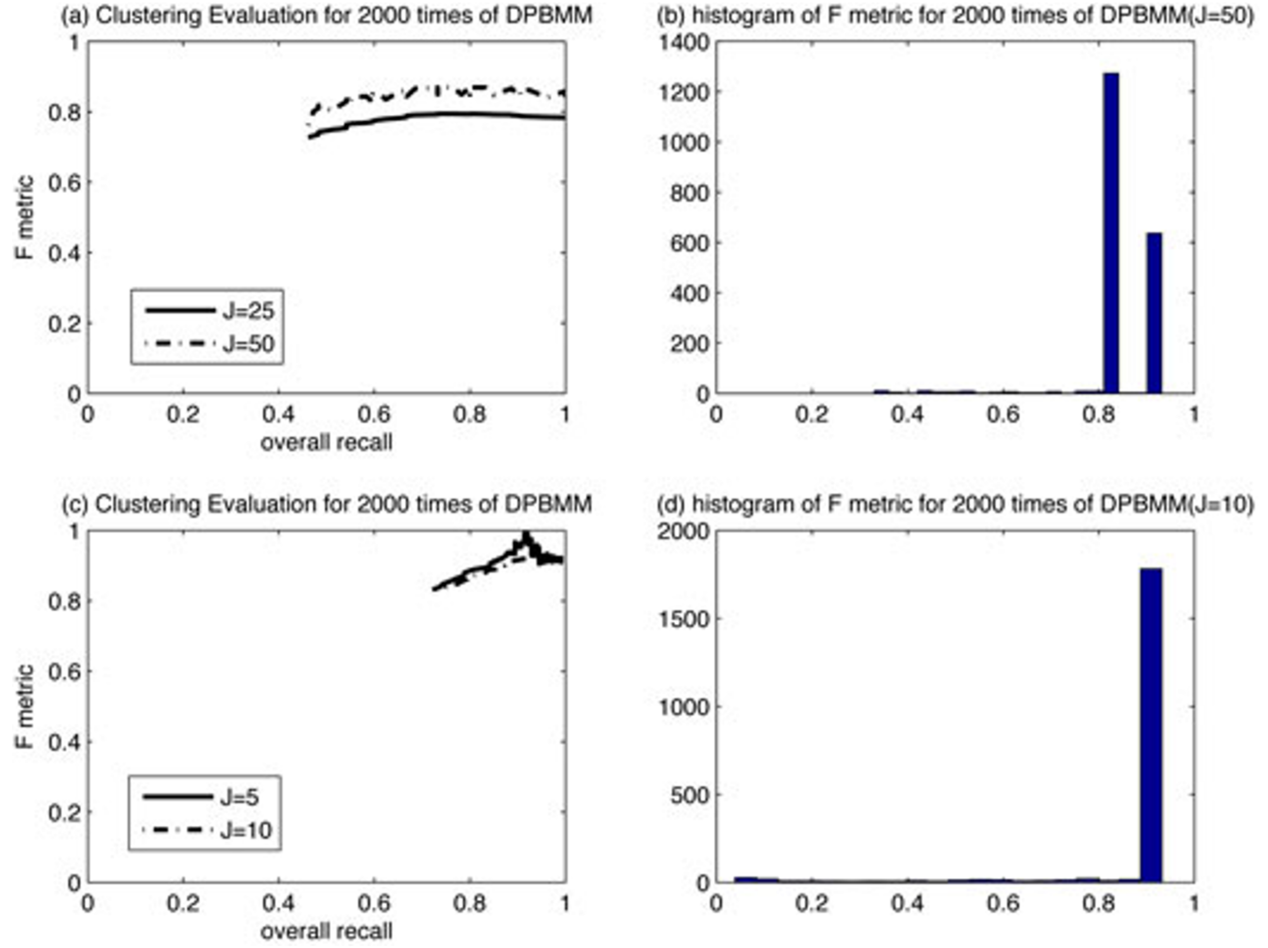

Figure 4 Clustering evaluation based on different *J*. Figure 4(a) shows the F metric vs. recall curve of $J \in \{25, 50\}$ loci for case I. Figure 4(b) shows the histogram of F metric results with $J = 50$. Figure 4(c) shows the F metric vs. recall curve of different $J \in \{5, 10\}$ loci for Case II. Figure 4(d) shows the histogram of F metric results with $J = 10$.

gives the best P-value of 0.03. And the heatmap plot of $J = 11$ is shown in Figure 5, the Kaplan-Meier overall survival curve is shown in Figure 6. When $J = 11$, the clusters in GBM methylation array uncovered by DPBMM are statistically significant (P-value < 0.1). We also analyzed the survival of the two clusters uncovered by hierarchical clustering, but the clusters yielded are not statistically significant (P-value > 0.1).

The computation time is always an issue for Gibbs sampling methods. Our simulation is carried out on a Linux based high-performance computer cluster. Each processing core is equipped with 2GB RAM. Figure 7 displays the computation time resulting from the real data study described before. The more loci considered for clustering, the more time the algorithm takes.

Discussion

We discuss next a few distinct features of DPBMM. First, in accordance with the fact that “beta” values in DNA methylation array data fall in the range of zero to one, we assume mixtures of beta distribution for the data. It can provide more flexible shapes, thus can describe data of various types. This is different from traditional Gaussian mixture model based clustering methods such as K-means. Second, since most existing methods can not determine the number of clusters automatically, we adopted a Dirichlet process prior for cluster assignment. Thus, we get a non-conjugate Dirichlet process beta mixture model, whose parameters are hard to estimate. A Gibbs sampling and “no-gap” sampling solution is developed to overcome this difficulty. This is different from traditional parametric methods, whose result also relies on a model parameter, which is usually determined in a model selection process.

The limitation of the proposed methods are mainly as follows. First, the algorithm is based on Gibbs sampling, which is somewhat a resource-heavy MCMC method,

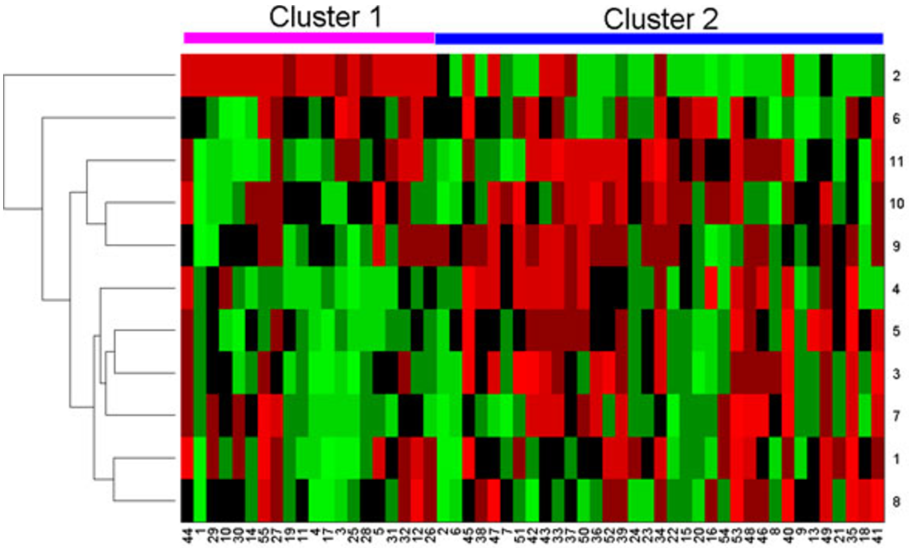

Figure 5 Estimated clustering structure based on DPBMM and Hierarchical clustering. 55 samples from TCGA are separated into two clusters on the basis of Illumina methylation expression array. The samples (columns) are arranged according to the estimated clusters by DPBMM while the locus (rows) according to hierarchical clustering.

therefore, the computation time is still heavy. Second, the model is computationally too slow to apply to methylation data of genome scale. We need to reduce the dimensionality to keep DPBMM computationally affordable.

In the future, it would be interesting to develop more effective dimension reduction method for DPBMM. It would also be interesting to integrate the information from different data sources such as gene expression and copy numbers variation into one model for cluster analysis.

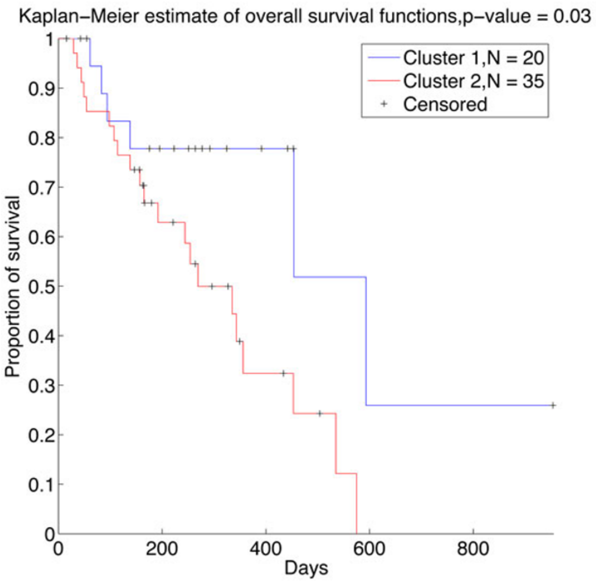

Figure 6 Kaplan-Meier estimate of survival analysis based on uncovered structure of DPBMM method (J = 11). The figure shows the survival functions of the two clusters obtained based on the top 11 variable locus (P-value = 0.03) by DPBMM, which is more significant than the corresponding result of hierarchical clustering (P-value = 0.51).

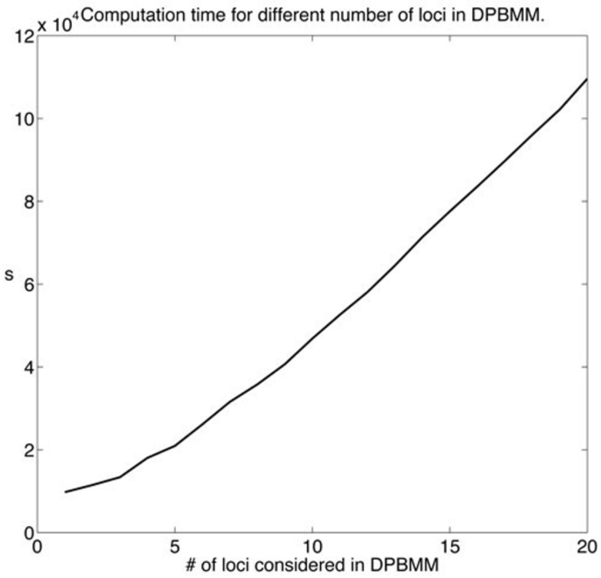

Figure 7 The computation time resulting from the real data study for *J* ⌊ {1, 2, ..., 20}. The figure shows the computation time resulting from the real data study for *J* ∈ {1, 2, ..., 20}. It is carried out on a Linux based high-performance computer cluster. Each processing core is equipped with 2GB RAM. With the number of loci considered for DPBMM clustering, the computation time increases.

Conclusions

An infinite Dirichlet process beta mixture model was proposed to unveil the latent cluster structure from Illumina Infinium methylation profiles. By utilizing a Dirichlet process prior for cluster assignment, the number of clusters is determined. A Gibbs sampling and "no-gaps" sampling solution was developed to infer the relevant parameters automatically. The effectiveness and validity of the model and the proposed Gibbs sampler were evaluated on simulated data and on real data. The results demonstrated that DPBMM could yield the cluster structure automatically with better accuracy.

Availability

MATLAB code is available at https://sites.google.com/site/bdpmmmethy/home.

Additional material

Additional file 1: Top 20 variable loci (ranked by variance through samples) selected from the methylation profiles of the 55 GBM samples.

Additional file 2: The number of uncovered clusters and P-value of overall survival analysis for *J* ⎣ {1, 2, ..., 20}. P-value is used to test the Kaplan-Meier confidence.

Acknowledgements

Based on "Clustering DNA methylation expressions using nonparametric beta mixture model", by Lin Zhang, Jia Meng, Hui Liu and Yufei Huang which appeared in *Genomic Signal Processing and Statistics (GENSIPS), 2011 IEEE International Workshop on*. © 2011 IEEE [27].

The work of L. Zhang is supported by "the Fundamental Research Funds for the Central Universities" (2010QNA50). The work of H. Liu is supported by "the Fundamental Research Funds for the Central Universities" (2010QNA47). The work of Y. Huang is supported by Qatar National Research Fund (09-874-3-235).

This article has been published as part of *BMC Genomics* Volume 13 Supplement 6, 2012: Selected articles from the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS) 2011. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S6.

Author details

[1]School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou, 221116, China. [2]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249, USA. [3]Department of Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA.

Authors' contributions

LZ, JM, and YH conceived the idea. LZ, JM, and YH worked out the detailed algorithms and derivations. LZ, JM and HL implemented the algorithm and performed the testing. LZ, JM, HL, and YH wrote the paper.

Competing interests

The authors declare that they have no competing interests.

Published: 26 October 2012

References

1. Graff J, Herman J, Myöhänen S, Baylin S, Vertino P: **Mapping patterns of CpG island methylation in normal and neoplastic cells implicates both upstream and downstream regions inde novo methylation.** *Journal of Biological Chemistry* 1997, **272(35)**:22322.
2. Jones P, Laird P: **Cancer-epigenetics comes of age.** *Nature genetics* 1999, **21(2)**:163-167.
3. Esteller M: **CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future.** *Oncogene* 2002, **21(35)**:5427-5440.
4. Jones P, Baylin S: **The fundamental role of epigenetic events in cancer.** *Nature reviews genetics* 2002, **3(6)**:415-428.
5. Shen L, Kondo Y, Guo Y, Zhang J, Zhang L, Ahmed S, Shu J, Chen X, Waterland R, Issa J: **Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters.** *PLoS genetics* 2007, **3(10)**:2023-2036.
6. Siegmund K, Laird P, Laird-Offringa I: **A comparison of cluster analysis methods using DNA methylation data.** *Bioinformatics* 2004, **20(12)**:1896.
7. Ji Y, Wu C, Liu P, Wang J, Coombes K: **Applications of beta-mixture models in bioinformatics.** *Bioinformatics* 2005, **21(9)**:2118.
8. Houseman E, Christensen B, Yeh R, Marsit C, Karagas M, Wrensch M, Nelson H, Wiemels J, Zheng S, Wiencke J, *et al*: **Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions.** *BMC Bioinformatics* 2008, **9**:365.
9. Sudderth E, Adviser-Freeman W, Adviser-Willsky A: **Graphical models for visual object recognition and tracking.** *PhD thesis* Massachusetts Institute of Technology; 2006.
10. Kuan P, Wang S, Zhou X, Chu H: **A statistical framework for Illumina DNA methylation arrays.** *Bioinformatics* 2010, **26(22)**:2849.
11. Elango YSV N: **DNA methylation and structural and functional bimodality of vertebrate promoters.** *Molecular Biology and Evolution* 2008, **25(8)**:1602-1608.
12. Murugiah S: **Bayesian nonparametric clustering based on Dirichlet processes.** *PhD thesis* University College London; 2010.
13. Gelman A: *Bayesian Data Analysis* Boca Raton, FL: Chapman and Hall/CRC; 2004.
14. Pitman J: *Combinatorial stochastic processes, Volume 1875* Springer-Verlag; 2006.
15. Teh Y, Jordan M, Beal M, Blei D: **Hierarchical Dirichlet processes.** *Journal of the American Statistical Association* 2006, **101(476)**:1566-1581.
16. Sethuraman J: **A constructive definition of Dirichlet priors.** *Statistica Sinica* 1994, **4**:639-650.
17. Blackwell D, MacQueen J: **Ferguson distributions via Pólya urn schemes.** *The annals of statistics* 1973, **1(2)**:353-355.
18. Paddock S, Ruggeri F, Lavine M, West M: **Randomized Polya tree models for nonparametric Bayesian inference.** *Statistica Sinica* 2003, **13(2)**:443-460.
19. Pitman J: **Some developments of the Blackwell-MacQueen urn scheme.** *Lecture Notes-Monograph Series* 1996, 245-267.
20. Escobar M, West M: **Bayesian density estimation and inference using mixtures.** *Journal of the american statistical association* 1995, 577-588.
21. Tang Y, Ghosal S, Roy A: **Nonparametric Bayesian estimation of positive false discovery rates.** *Biometrics* 2007, **63(4)**:1126-1134.
22. MacEachern S, Muller P: **Estimating mixture of Dirichlet process models.** *Journal of Computational and Graphical Statistics* 1998, 223-238.
23. Van Rijsbergen C: **Foundation of evaluation.** *Journal of Documentation* 1993, **30(4)**:365-373.
24. Amigó E, Gonzalo J, Artiles J, Verdejo F: **A comparison of extrinsic clustering evaluation metrics based on formal constraints.** *Information Retrieval* 2009, **12(4)**:461-486.
25. Hernandez-Vargas H, Lambert M, Le Calvez-Kelm F, Gouysse G, McKay-Chopin S, Tavtigian S, Scoazec J, Herceg Z: **Hepatocellular carcinoma displays distinct DNA methylation signatures with potential as clinical predictors.** *PLoS One* 2010, **5(3)**:e9749.
26. Dougherty E: **Small sample issues for microarray-based classification.** *Comparative and Functional Genomics* 2001, **2**:28-34.
27. Zhang L, Meng J, Liu H, Huang Y: **Clustering DNA methylation expressions using nonparametric beta mixture model.** *Genomic Signal Processing and Statistics (GENSIPS), 2011 IEEE International Workshop on: 4-6 December 2011* 2011, 170-173.

doi:10.1186/1471-2164-13-S6-S20
Cite this article as: Zhang *et al.*: A nonparametric Bayesian approach for clustering bisulfate-based DNA methylation profiles. *BMC Genomics* 2012 13(Suppl 6):S20.