Statistical Applications in Genetics and Molecular Biology

Volume 6, *Issue* 1 2007 *Article* 30

T-BAPS: A Bayesian Statistical Tool for Comparison of Microbial Communities Using Terminal-restriction Fragment Length Polymorphism (T-RFLP) Data

Jing Tang, *University of Helsinki*
Jinglun Tao, *Ocean Research Institute, The University of Tokyo*
Hidetoshi Urakawa, *Ocean Research Institute, The University of Tokyo*
Jukka Corander, *Åbo Akademi University*

Recommended Citation:
Tang, Jing; Tao, Jinglun; Urakawa, Hidetoshi; and Corander, Jukka (2007) "T-BAPS: A Bayesian Statistical Tool for Comparison of Microbial Communities Using Terminal-restriction Fragment Length Polymorphism (T-RFLP) Data," *Statistical Applications in Genetics and Molecular Biology*: Vol. 6: Iss. 1, Article 30.
DOI: 10.2202/1544-6115.1303

T-BAPS: A Bayesian Statistical Tool for Comparison of Microbial Communities Using Terminal-restriction Fragment Length Polymorphism (T-RFLP) Data

Jing Tang, Jinglun Tao, Hidetoshi Urakawa, and Jukka Corander

Abstract

The investigation of microbial communities is an essential part of the study of the biosphere. Flexible molecular fingerprinting tools such as terminal-restriction fragment length polymorphism (T-RFLP) analysis are often applied in the studies to enable the characterization of the microbial population. However, such data have so far been primarily analyzed using conventional clustering methods. Here we introduce a Bayesian model-based method for the purpose of comparing microbial communities using T-RFLP data. Such datasets have in general several challenging features, e.g. sparseness, missing values and structurally zero-valued observations. These features are taken into account by developing a Bayesian latent class mixture model for the observations in our framework. To make inferences under the model we use a recent Markov chain Monte Carlo (MCMC) -based method for the Bayesian model selection. To assess the introduced method we analyze both simulated and real datasets. The simulations show that our approach compares preferably to standard statistical clustering tools, such as k-means, hierarchical clustering, and Autoclass. The developed tool is freely available as a software package T-BAPS at http://www.abo.fi/fak/mnf/mate/jc/software/t-baps.html.

KEYWORDS: Bayesian modelling, Markov chain Monte Carlo, microbial communities, T-RFLP, unsupervised classification

Author Notes: We thank Andrew Thomas for his valuable help in programming OpenBUGS. We thank Kazuo Wanami and Haruo Ando for organizing water sampling in the Keihin Canal. We also thank Kazuhiro Kogure for supporting and encouraging the study. This work was financially supported by the Academy of Finland and the Centre of Population Genetic Analyses, University of Helsinki, Finland.

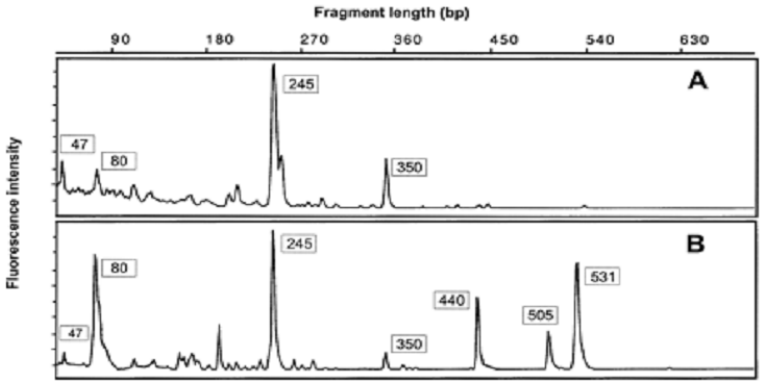
1 Introduction

Microbial communities can be found in various environments including those that exist in marine sediment, soil and human bodies. It is recognized that microbial communities respond to many factors that affect the underlying ecosystems. The study of microbial communities may thus improve our understanding of their role in ecological functions and human health. For example, analyzes of microbial composition in marine sediments help to discover the transfer of nutrients and energy within the whole marine environment (Urakawa et al., 2005). In medical research, comparison of temporal changes of microbiota in human organs will enable development of novel pharmaceuticals (Wang et al., 2004). In agriculture, the impact of transgenic crops on natural ecosystems can be assessed by examining structural changes in soil bacterial communities (Park et al., 2006).

Microbial communities are composed of multiple bacterial species and often exhibit complex structure. Traditional techniques used for microbial characterization include bacterial isolation and cultivation. These culture-based techniques, despite of their significant contribution to our current knowledge of microbes as a whole, provide very limited information for the assessment of microbial diversity. Alternatively, DNA-based techniques, such as 16S rDNA clone library analysis, can further reveal microbial diversity on a molecular level. However, this technique is rather time-consuming and labor-intensive and therefore has limited use for comparative studies involving a large number of samples (Abdo et al., 2006).

A molecular fingerprinting technique, terminal restriction fragment length polymorphism (T-RFLP), has become increasingly popular in the studies of complex microbial communities (Marsh, 1999). As its name suggests, T-RFLP analysis extracts a certain nucleotide sequence from a PCR amplified marker, and measures the length polymorphism of terminal restriction fragments (T-RFs) by using a DNA sequencer. The output electropherogram is a profile that consists of a series of peaks each of which indicates the relative abundance of a particular T-RF length (Figure 1). The presence or absence of a fragment length and its relative abundance is assumed to indicate the richness of a certain bacterial phylotype. Without the necessity of knowing the actual sequence information, T-RFLP provides a rapid, robust and reproducible method for estimating bacterial diversity and community composition on the molecular level. Figure 1 presents an electropherogram and a section of its standardized data from a T-RFLP analysis from two marine sediment samples.

Further processing of fragment profiles usually generates a data matrix



	Fragment length(bp)								
	47	80	245	350	440	505	531	629	630
Sample A	0.2	0.02	0.5	0.18	0.01	0	0	0	0
Sample B	0.1	0.12	0.4	0.05	0.06	0.03	0.09	0	0

Figure 1: T-RFLP profiles and data of two marine sediment samples. The electropherogram shows the lengths of terminal fragments present, measured in base pairs. The height of each peak is measured in fluorescent intensities. The peak patterns are converted to a table of feature vectors where the relative abundance of each fragment length is calculated.

where rows correspond to samples and columns represent relative intensities of fragment lengths. Samples with peak patterns dissimilar to each other will be assigned into different clusters, by which the mechanisms that drive the diversity can be further investigated. Conventional methods, such as k-means and hierarchical clustering, have been used to provide simplistic solutions. However, these methods cannot easily provide uncertainty measures due to their deterministic nature. Recent approaches used for T-RFLP data clustering include principle component analysis (PCA) and discriminant analysis (DA) (Wang et al., 2004; Park et al., 2006). By reducing data dimensionality using PCA, the distances in T-RFLP patterns can be visualized. PCA is known to work well with data sets that are approximately normally distributed. However, when examining non-normal and heavily skewed T-RFLP data, PCA tends to obscure underlying patterns that might be important in differentiating microbial communities (Rees et al., 2004).

Model-based clustering methods have been available for quite some time

DOI: 10.2202/1544-6115.1303

(see Bock, 1996), but they have not yet been in an as wide-spread use as the standard algorithmic methods available in most statistical software packages. Nevertheless, the popularization of Bayesian computational algorithms, known typically as Markov chain Monte Carlo (MCMC) methods, has lead to an increasing awareness of the potential hidden in the model-based approach. For recent examples of applications within molecular biology, see Zhou and Wakefield (2006) and Dai and Lieu (2006). However, these unsupervised classification approaches are more tailored for the analysis of data from microarray experiments. In a majority of the existing Bayesian unsupervised classification methods, the data are either assumed to be inherently discrete or pre-processed into a discrete format (see, e.g. Corander et al., 2007; Gyllenberg et al., 1997). Although the discretization enables the use of standard reference choices of the likelihood and priors in the classification model, for sparse data sets it also leads to a considerable loss of information, which might be essential in differentiating the microbial communities. On the other hand, standard continuous unsupervised classifiers based on Gaussian mixture models, do not either reflect well the characteristics of T-RFLP data, and are thus likely to fail to produce sensible clustering results.

Clustering T-RFLP data is demanding primarily for two reasons. Firstly, the number of variables, e.g. the number of unique fragment lengths, often exceeds the number of samples. The excessive number of explanatory variables poses significant challenges in terms of modeling and computation. Secondly, for each variable, structurally zero-valued observations will always be present. The zero-inflated feature of T-RFLP data tends to invalidate the assumptions underlying many Gaussian mixture models. Gaussian mixture models could therefore lead to biased inference when used in clustering of T-RFLP data.

In this article, we develop a Bayesian method to tackle the above-mentioned difficulties. The mixed zero and positive values of each variable are explicitly modeled by a mixture of discrete and continuous distributions. Our classification framework is built according to the latent class concept (see, e.g. Duda et al., 2000), and formulated in the OpenBUGS environment (Thomas et al., 2006). We consider models for different predefined number of clusters separately, and apply the recently introduced BICM criterion for model selection (Raftery et al., 2006). The optimal cluster number, as well as the cluster-specific parameters can be jointly inferred from their posterior distributions. The developed methodology is implemented in a freely downloadable user-friendly software package (T-BAPS) enabling biologists to investigate their data sets.

The paper is organized as follows. In Section 2 we introduce notation and describe the mathematical details of the model and the suggested inference

procedure. In Section 3, we test the method on simulated data and compare it with a number of previously suggested approaches. In section 4, we analyze two real T-RFLP data sets, and some discussion is provided in the final section.

2 Bayesian model-based unsupervised classification

2.1 Latent class model for a predefined number of clusters

2.1.1 Modeling framework and likelihood function

We consider N samples and J fragment lengths present in a T-RFLP data set. Let y_{ij} denote, for sample i, the relative abundance of the j^{th} fragment length. Our model aims at partition of the data into clusters. To simplify the modeling process, we assume that a permutation of data ordering, in either sample-wise or variable-wise, does not affect its marginal probability. In our model, each data point y_{ij} is supposed to be independently drawn from an unknown component k of a finite mixture of K distributions, each of which is associated with a density f_{jk} having a set of cluster-specific parameters ϕ_{jk}, such that

$$p(y_{ij}|\pi,\phi_j)=\sum_{k=1}^{K} p(y_{ij}|Z_i=k,\phi_{jk},\pi)p(Z_i=k|\pi), i=1,\ldots,N, j=1,\ldots,J. \tag{1}$$

Here Z_i is an unobserved partition label indicating the underlying cluster to which the i^{th} sample belongs; $\pi=(\pi_1,\ldots,\pi_K)$ is a probability vector, such that $\pi_k \geq 0$ and $\sum_{k=1}^{K}\pi_k=1$. We assume that the partition labels are independent observations from a generalized Bernoulli distribution with parameter π

$$p(Z_i=k|\pi)=\pi_k.$$

It is further assumed that conditional on Z, $y_{1j},\ldots,y_{nj}$ are independently sampled from the cluster-conditional densities such that

$$p(y_{ij}|Z_i=k,\ \phi_{jk},\pi)=f(y_{ij};\phi_{jk}).$$

Therefore (1) is simplified as

$$p(y_{ij}|\pi,\phi_j)=\sum_{k=1}^{K}\pi_k f(y_{ij};\phi_{jk}), i=1,\ldots,N, j=1,\ldots,J. \tag{2}$$

DOI: 10.2202/1544-6115.1303

To define the cluster-specific densities in (2), we in particular take into account the zero-inflated feature of the data. We assume that y_{ij} is either a zero value sampled from a Bernoulli distribution, or a positive value drawn from a Gaussian distribution. It follows that

$$p(y_{ij}|Z_i = k, \pi, \phi) = \lambda_{kj} I(y_{ij} = 0) + (1 - \lambda_{kj}) f_g(y_{ij}; \mu_{kj}, \tau_{kj}) I(y_{ij} \neq 0),$$

where λ_{kj} is the weight of zero-inflatedness at variable j for cluster k; $I(\cdot)$ is a boolean indicator; μ_{kj} and τ_{kj} are the mean and precision parameters of the Gaussian density f_g. With the extra layer added to the model hierarchy, the cluster-specific parameter ϕ represents (λ, μ, τ).

Assuming exchangeability on both data dimensions, the likelihood function for parameter (Z, π, ϕ) given the observed data is then compactly expressed as

$$p(\mathbf{y}|Z, \pi, \phi) = \prod_{i=1}^{N} \prod_{j=1}^{J} p(y_{ij}|Z, \pi, \phi).$$

2.1.2 Prior settings

We are primarily interested in the number K of clusters and the partition labels Z supported by the data $\mathbf{y}$. In a hierarchical Bayesian setting, it is computationally convenient to assume first that K is predefined. The joint distribution of the remaining parameters and the data can be specified as a product of conditional distributions

$$p(Z, \pi, \phi, \mathbf{y}) = p(\pi, \phi) p(Z|\pi, \phi) p(\mathbf{y}|Z, \pi, \phi).$$

A Bayesian approach requires specifying the prior distribution $p(\pi, \phi)$. The marginal probabilities for the partition labels Z can then be inferred from the posterior distribution $p(\pi, \phi|\mathbf{y})$ using Bayes' theorem:

$$p(Z_i = k|\mathbf{y}) = \int \frac{\pi_k f(y_i; \phi_k)}{\sum_{l=1}^{K} \pi_l f(y_i; \phi_l)} p(\pi, \phi|\mathbf{y}) d\pi d\phi. \tag{3}$$

Assuming mutual independence, the prior $p(\pi, \phi)$ can be further factorized as

$$p(\pi, \phi) = p(\pi) p(\phi) = p(\pi) p(\lambda) p(\mu) p(\tau). \tag{4}$$

We now consider the specification of priors for each parameter in (4). We assign a symmetric Dirichlet prior to π as

$$\pi_k \sim Dirichlet(\alpha, \ldots, \alpha), \alpha > 0, k = 1, ..., K. \tag{5}$$

Throughout the paper we are using $\alpha = 1$. This prior setting indicates a symmetric assumption of all the clusters being equally probable *a priori*.

For the Bernoulli-Gaussian component weight λ we define a Beta prior

$$\lambda_{kj} \sim Beta(\beta_{bin}, \beta_{gauss}). \tag{6}$$

The choice of β_{bin} and β_{gauss} reflects prior knowledge about the level of zero-inflatedness of the data. Based on prior experience on similar data, if an analyst expects that the probability of observing zero value in the data is larger than 0.5, then he could incorporate this information by letting $\beta_{bin} > \beta_{gauss}$. For convenience of presentation, hereafter we use a uniform distribution between 0 and 1 by choosing $\beta_{bin} = \beta_{gauss} = 1$. This setting implies a non-informative prior on λ.

To allow for a flexible inference while avoiding an over-parametrization of the model, the prior distributions of the mean μ and precision τ in Gaussian components were carefully considered. In particular, we add additional hierarchy on μ and τ separately, such that

$$\mu_{kj} \sim Gaussian(a_{kj}, b_{kj}), \tag{7}$$

and

$$\tau_{kj} \sim Gamma(r, s). \tag{8}$$

If there is no prior information that distinguishes the clusters, the posterior distribution of $p(\pi, \phi)$ will be similarly symmetric. Consequently, the marginal classification probabilities given in (3) tend to be identical for every sample, thus leading to an unidentifiable partition. To avoid the identification problem related to latent class models, a common response is to impose constraints on the parameter space, such that the symmetry of the data likelihood can be broken down (Stephens, 2000). The constrains were here made to the prior setting of hyperparameters a and b according to the predefined K and the observed range of the data. We impose the restriction by ordering a_{kj} to be evenly distributed along the range of observations by

$$a_{kj} = min(y_j) + \frac{k-1}{K-1} R_j, \tag{9}$$

where R_j is the length of the non-zero data interval for variable j. The precision hyper-parameters b_{kj} are chosen independently as

$$b_{kj} = \sqrt{K}/R_j^2, \tag{10}$$

according to the suggestion of Richardson and Green (1997).

DOI: 10.2202/1544-6115.1303

As the amount of information in a T-RFLP data set is typically limited, we choose to restrict further the precision parameters τ_{kj} to be independent of clusters and variables. We use a vague Gamma prior by taking $r = s = 0.01$ in (8), to express the belief that the τ_{kj} are similar and that the prior uncertainty is high.

2.1.3 Posterior inference

Using Bayes' theorem, the posterior distribution of model parameters is proportional to the product of the prior probability and the likelihood

$$p(\pi, \lambda, \mu, \tau|\mathbf{y}) \propto p(\mathbf{y}|\lambda, \mu, \tau)p(\pi, \lambda, \mu, \tau).$$

For our model, the exact form of the posterior distribution is in practice intractable, so we will rely on MCMC computation that approximates the posterior by drawing samples from the distribution. We use OpenBUGS (Thomas et al., 2006) environment to perform the necessary sampling procedure. Using a Gibbs sampler, OpenBUGS can construct a Markov chain whose stationary distribution is the posterior distribution of model parameters. In particular, we are interested in the inference of cluster-specific parameters $\phi = (\lambda, \mu, \tau)$ and clustering labels Z. The algorithm that combines the posterior inference and model selection will be outlined in section 2.3.

2.2 Model selection in a range of number of clusters

In practical analysis of T-RFLP data, the number K of clusters is typically unknown. Note that our model formulation in section 2.1 requires that K is predefined. Therefore, we need to explore a range of models in which K is allowed to vary. The optimal number of clusters will be determined by the model that best fits to the data. In this section, we describe a model selection criterion by which the fitness of multiple competitive models can be evaluated. Furthermore, rather than giving a single best model, we would like to consider the model uncertainty as well.

In general, a Bayesian approach can provide a formal comparison of model structures with a varying number of parameters. The key quantity is integrated likelihood of the data, which is defined in the current context as

$$p(\mathbf{y}|K) = \int p(\mathbf{y}|\pi, \phi, K)p(\pi, \phi|K)d\pi d\phi.$$

The integrated likelihood enables the comparison of K's by taking simultaneously into account the complexity and the ability to predict the observed data.

In the case of two models being under comparison, it is common to use the Bayes factor, defined as the ratio of integrated likelihoods for the two models. However, the estimation of the integrated likelihood is an extremely difficult problem in general. For detailed discussions in this issue, see Spiegelhalter et al. (2002).

Recently, an efficient novel method for model selection has been developed by Raftery et al. (2006). They provided a BICM criterion that can be calculated from the posterior distribution through an MCMC simulation. The BICM can be interpreted as an approximation to the log integrated likelihood of the data and therefore can be used as a measure of model fitness.

By applying the BICM criterion, we provide a model selection method for our latent class models. With specification of a range of initial cluster numbers *a priori*, a BICM score for each predefined K can be calculated from the posterior MCMC simulation. The BICM is defined by

$$\log\hat{\pi}_{\mathrm{BICM}}(\mathbf{y})|_K = \bar{l} - s_l^2(\log(n) - 1), \tag{11}$$

where n is the size of the data; $\bar{l}$ and s_l^2 are the sample mean and variance of the log-likelihood $\log p(\mathbf{y}|\pi, \phi)$. In the OpenBUGS environment, summary of the log-likelihood is particularly straightforward since it is automatically calculated in the posterior simulation.

2.3 Sampling algorithm

In the present context, we can examine the partition Z and also the cluster-specific parameters (λ, μ, τ) by comparing the integrated likelihood of the data under various values of initial Ks using the BICM approximation. This procedure has been automated in our software package T-BAPS using the following steps:

1. Predefine the number of clusters K and determine the hyperparameters (a, b) according to (9) and (10) separately.
2. Generate initial values according to (5), (6), (7) and (8).
3. Generate a posterior sample for all the parameters using the Gibbs sampler algorithm in the OpenBUGS environment.
4. Calculate the BICM for the model with the specified initial K, according to (11).
5. Repeat the procedure from step 1 to step 4 for a range of Ks and compare their BICM scores.

DOI: 10.2202/1544-6115.1303

The optimal cluster numbers can be determined by the partition result from the model with the highest BICM score. Note that for some of initial Ks, not all the clusters in the model are assigned data points after the MCMC has reached a stable state, and thus the actual number of non-empty clusters may be smaller than the initially predefined value of K. However, this has no practical consequence for the analysis and does not affect the estimation of the BICM. The above analysis allows for the identification of the optimal clustering solution, which can be investigated using the graphical tools implemented in the T-BAPS package.

3 Analysis of Simulated Data

In order to assess our model performance, we first used simulated data to compare our method with other popular approaches to clustering. The data was generated by a mixture model corresponding to $K = 4$ clusters, and the sample size from each cluster was 25, resulting in a total of $N = 100$ observations. The number of variables J, corresponding to the number of fragment sizes in real T-RFLP data, was set equal to 100. The simulated data set hereby consists of 100 observations and 100 variables.

The data generating process basically involves two steps. Firstly, we generate a binary data set using Bernoulli distributions. To appropriately mimic the excessive zero-values in real T-RFLP data sets, we did sample λ_{kj}, the probability of observing zero at variable j in cluster k, from a Uniform distribution bounded by [0.5 1]. Zero-values for each variable y_{kj} are then sampled according to a Bernoulli distribution with λ_{kj}. Secondly, the remaining values in the binary data are replaced with continuous values sampled from a Gaussian distribution, with the mean μ_{kj} was randomly generated from the $Uniform(0, 1)$ distribution and the precision $\tau_{kj} = 100$ invariantly for k and j. Finally, any negative values generated by the normal distributions are replaced by zero, in order to satisfy the non-negativeness of T-RFLP data.

The MCMC simulation was ran for 90,000 iterations, after the first 10,000 iterations were discarded as a burn-in period. The chain was further thinned at every 10th iteration to reduce sample autocorrelation. The total time needed for running the whole procedure was approximately 96 hours on a 2.8GHz Pentium 4 PC.

The BICM scores on the range of initial Ks from 2 to 20 are summarized in Figure 2. From the figure it is seen that the model with the initial value of $K = 13$ received the highest BICM score. This model provided an optimum number $K = 4$ of real (non-empty) clusters, indicating that our method can

identify the correct cluster number.

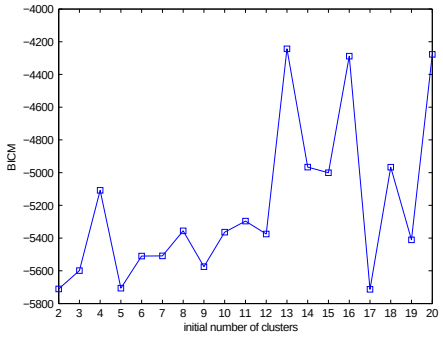


Figure 2: BICM scores for our latent class models with initial number of clusters K from 2 to 20. The highest BICM was found at the model with initial $K = 13$. The partition result from this model consists of four non-empty clusters, which exactly matches the true data generating process.

For a direct comparison with standard clustering tools, we focus here on the similarity of the data partition with the true clusters. We used k-means, hierarchical clustering and Autoclass (Cheeseman et al., 1988) for comparison. For k-means and hierarchical clustering, we conditioned the analysis on the true value $K = 4$. For Autoclass clustering, we used the Single Normal CN Model. The Adjusted Rand index (RI) values (Hubert and Arabie, 1985) for comparing the true data partition with those obtained by the considered methods are listed in Table 1.

As can be clearly seen from the table, our method is relatively superior compared to the standard tools, which is not surprising as it takes into account the specific features of T-RFLP data. However, k-means clustering is able to find a partition quite close to the truth with the maximum RI value of 0.9214, but it must be noted that the average Rand Index from its 100 repetitions is 0.6121, indicating that there is also a high probability of misclassifications. In fact, the lowest RI value in this comparison happened to be found by the k-means method. Furthermore, without the knowledge of true number of clusters, k-means and hierarchical clustering would hardly provide sufficient confidence in the prediction results.

As an alternative unsupervised clustering method, Autoclass determines automatically the number of clusters using a EM algorithm. However, the experiment result showed that Autoclass tended to produce many more clusters than were present in the underlying data. It suggested 20 clusters for the simulated data, leading to a low RI (0.2046). One possible explanation may

DOI: 10.2202/1544-6115.1303

Table 1: Classification of the simulated data with $K = 4$ partitions. The adjusted rand index by k-means shows a range identified by running the algorithm 100 times on the data. The hierarchical clustering is based on the Cosine distance metrics.

Methods	Adjusted rand index
K-means	[0.1690 0.9214]
Hierarchical clustering centroid	0.6601
Hierarchical clustering complete	0.3162
Autoclass	0.2046
Our method	1

be that the discrete-continuous mixture of the data brings a significant violation to Autoclass's underlying Gaussian assumptions. With the complex data structure, Autoclass found better fit when using a larger number of component distributions (Li and Biswas, 2002).

4 Applications to two real T-RFLP data sets

4.1 The Japanese marine sediment data

4.1.1 Materials and Methods

The first real data set contains samples from marine sediments. The sediment samples were collected from three bay areas in the Japanese coast line including Tokyo Bay, Suruga Bay and Sagami Bay, as described previously in Urakawa et al. (2005). Each sample was collected by slices of the surface sediment at a certain depth, as denoted in Table 2. The bacterial composition was determined using the T-RFLP technique, and represented as terminal fragments in a T-RFLP profile. The corresponding data matrix was analyzed by the T-BAPS software for a comparison of microbial diversity in these bay areas.

4.1.2 Data and Results

A total of 24 samples and 255 unique fragment lengths ranging from 35 to 674 bp (base pairs) were identified by the T-RFLP analysis. Figure 4(a) shows the percentage of zero-values observed in the data. This figure illustrates the

Depth(cm)	0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20
Tokyo Bay	S1	S2	S3	S4	S5					
Sagami Bay	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
Suruga Bay	S16	S17	S18	S19	S20	S21	S22	S23	S24	

Table 2: Sampling sites of the 24 sediment samples. Each sample was extracted at a certain depth level of the sediment at one of the three bay areas.

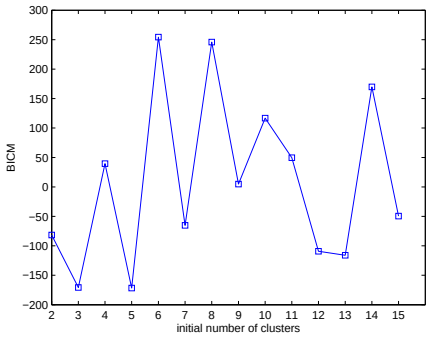
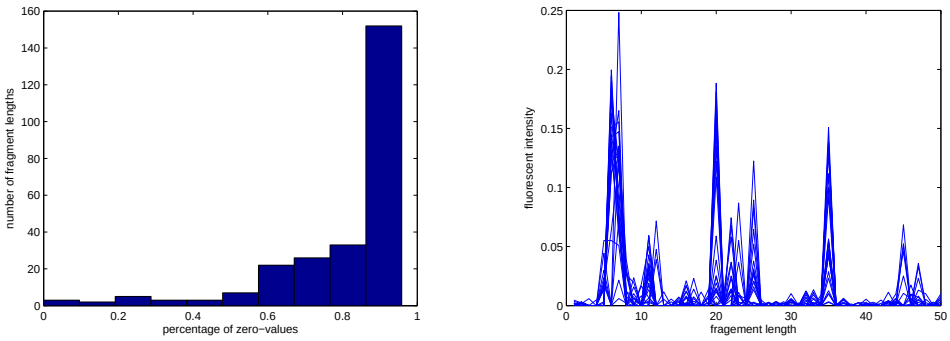


Figure 3: BICM scores for models with different initial Ks. The best model was found with initial $K = 6$. The resulting partition identified two clusters, one of which contains the Tokyo Bay and the Sagami Bay samples and the other of which contains the Suruga Bay samples.

striking feature of zero-inflatedness in a typical T-RFLP data set. The difficulties in inferring the clusters are further weighted by the limited variability presented in the profiles. As can be seen in Figure 4(b), there are no obvious clusters.

We tested our method by considering models with initial Ks from 2 to 15. For each model, the MCMC sampler was run for 100,000 iterations, discarding a burn-in period of 50,000 iterations. Further, thinning by storage of every 10^{th} sample led to a collection of 5,000 approximately independent samples. The BICM score was then calculated on these thinned samples. As shown in Figure 3, the best model was identified with an initial $K = 6$. The resulting partition showed that samples from the Tokyo Bay and the Sagami Bay formed a cluster that is separate from the samples from the Suruga Bay. This geographical structure resolved here indicates that site-specific differences may have a greater influence on the microbial communities than depth-specific differences. This was also supported by previous analysis of respiration quinone profiling of the microbial communities from the same areas (Urakawa et al., 2005)

DOI: 10.2202/1544-6115.1303



(a) (b)

Figure 4: (a) Distribution of percentages of zero-values in the marine sediment data. The majority of fragment lengths contains zero-values with a proportion of larger than 0.5. This can be seen as a typical T-RFLP data set where the zero-inflatedness is commonly observed for most of the fragment lengths. (b) The peak patterns of the 24 marine sediment T-RFLP samples. This figure shows only the fluorescent intensity for the first 50 fragment lengths. Each curve represents a single sample.

4.2 The Keihin Canal data

4.2.1 Materials and Methods

The objective of this study is elucidation of the influence of wastewater from a sewage treatment plant on the microbial communities in the Keihin Canal at the Tokyo Bay. As shown in Figure 5, 18 water samples from the surface and bottom water were taken from 10 sites along the Keihin canal, starting from a sewage pipe S1, to the canal's outlet into the Tokyo Bay S10 (Table 3). At sites S1 and S2 only the surface water was successfully collected. From S3 to S10, both the surface and the bottom water were sampled.

It is expected that shift is present in microbial communities from S1 to S10. Sampling sites S8 to S10 are located in Tokyo Bay which are influenced by higher salinity and chlorophyll concentration compared to the waste water. A vertical salinity gradient was obvious in the canal and a depth-related stratification in the oxygen concentration was also observed. Therefore, we also expect differences of microbial communities between surface and bottom

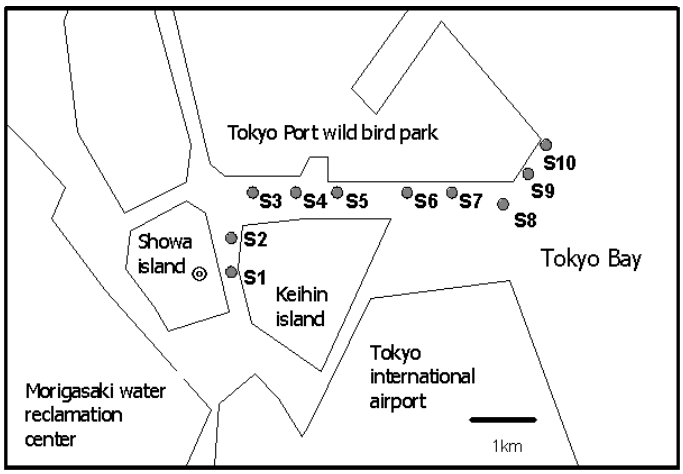


Figure 5: The sampling sites of the Keihin Canal T-RFLP data.

Sampling sites	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Surface	1	2	3	4	5	6	7	8	9	10
Bottom			11	12	13	14	15	16	17	18

Table 3: The sampling notations for the Keihin Canal T-RFLP Data. Note that no samples were collected from the bottom water of sites S1 and S2.

waters.

Using T-RFLP analysis, 450 fragment lengths from 50 bp to 500 bp were identified. We removed 380 fragment lengths which were not observed in any of the 18 samples and thus kept only 71 informative fragment lengths for data clustering. Figure 6 shows the histogram of zero-proportions.

Clustering analysis using the T-BAPS software was implemented by choosing initial number of clusters K from 2 to 10. For each initial K, the MCMC simulation was run for 100,000 iterations after a burn-in period of 50,000 iterations. The summaries of posterior distribution were obtained by further thinning in every 10th of the iterations.

4.2.2 Results

To check that the Markov chain after the burn-in period was long enough to make accurate posterior inferences, we calculated the MC error for the model parameters. For each predefined initial K, the MC error is less than 3% of the standard deviation, indicating that the accuracy of the posterior estimate of

DOI: 10.2202/1544-6115.1303

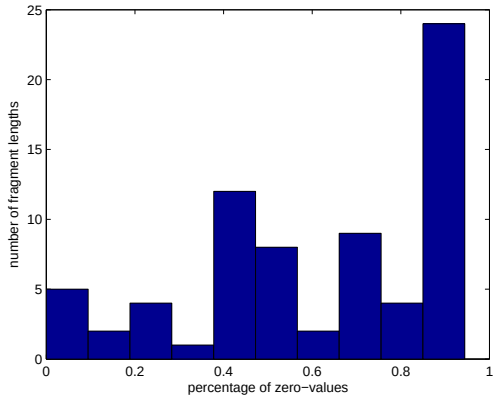


Figure 6: Distribution of percentages of zero-values in the Keihin Canal data.

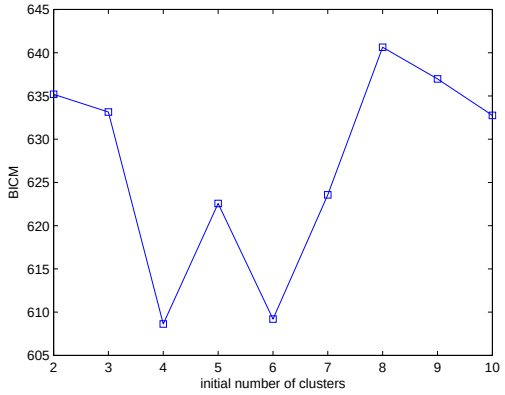


Figure 7: BICM plots for models with initial K from 2 to 10.

deviance was rather satisfactory. Model selection showed that the model with an initial value of $K = 8$ received the highest BICM score $\log\pi_{\text{BICM}} = 640.62$, resulting in an optimum partition of 3 clusters (Figure 7). In fact, the same partition was uniformly identified by models with K from 4 to 10. An alternative partition of two clusters was found with $K = 2$ and 3 . Due to the uncertain nature of MCMC simulations, we report both of the partition results in Table 4.

The optimal clustering result confirms our expectation that the microbial communities in the surface and bottom water belong to distinct community types. Although the posterior density of cluster-specific parameters, such as cluster means μ and zero-inflatedness weights λ, may provide further insights

on characterizing the microbial diversity, we do not report these here in more details.

We further compared the introduced method with the clustering obtained by a binary discretization of the data and consequent use of the Bayesian approach implemented in the BAPS 4 software (see, e.g. Corander and Tang, 2007). The partition given by the BAPS algorithm is the same as Partition C reported in Table 4. To evaluate the information loss due to the data discretization, we compared the BICM scores for partition A and C, conditional on the best model with $K = 8$. This can be achieved by fixing the allocation variable Z in our model according to a partition. The difference of BICM scores between these two partitions is 25.011. Recalling that BICM is an approximation of the integrated data likelihood, this difference might serve as a numerical measure of how much information is lost by the data discretization.

Partition A $\text{BICM}\|_{K=8} = 640.62$	Cluster 1: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} Cluster 2: {11, 12, 13, 14, 15, 16, 18} Cluster 3: {17}
Partition B $\text{BICM}\|_{K=2} = 635.18$	Cluster 1: {11, 12, 13, 14, 15, 16, 17, 18} Cluster 2: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
Partition C $\text{BICM}\|_{K=8} = 615.61$	Cluster 1: {11, 12, 13, 14, 15, 16, 17, 18} Cluster 2: {2} Cluster 3: {1, 3, 4, 5, 6, 7, 8, 9, 10}

Table 4: Three partitions of the Keihin Canal T-RFLP Data. Partition A and B were discovered by our method. Utilizing the method of Corander and Tang (2007) by data discretization produced Partition C. The partition uncertainty can be measured by the differences between the corresponding BICM values.

5 Discussion

To identify microbial communities that are similar to each other using T-RFLP data, it is common to apply some statistical cluster analysis tools. However, the standard clustering methods cannot be adjusted for the specific characteristics of such data, and consequently, they neglect a considerable amount of existing biological information. This is particularly important in the current context, as the sample sizes in T-RFLP data sets are often fairly small. In our example analyses we have illustrated how the hierarchical Bayesian modelling framework can be efficiently utilized to focus on the biologically

DOI: 10.2202/1544-6115.1303

relevant questions and to utilize the existing knowledge about the features of the considered data.

Choosing appropriate identifiability constraints is quite difficult in the mixture modelling framework and some authors have argued that whatever constrains are used, they may not always remove the label-switching problem efficiently. However, in our simulation studies and further real data analyses, the effect of label-switching was examined by the trace and history of the MCMC chains. These investigations showed that our current constrains work very satisfactorily. Some criteria for checking whether label-switching is a problem in a particular application, are based on examining whether the posterior of the parameter of interest is unimodal. Moreover, estimation, clustering and posterior inference can be viewed in a decision-theoretic manner, for details, see Stephens (2000).

Although we did not stress this aspect in our example analyses, it is important to notice a specific advantage of the model-based methods over the standard algorithmic approaches. Namely, a statistical model also provides a probabilistic quantification of the uncertainty related to the cluster parameters and to the allocation of an observation to any particular cluster. Observations having similar posterior probabilities of allocation to alternative clusters should be deemed uncertain, and thus be given special attention in the investigation of the results. However, in the considered analyses the posterior probabilities related to the data partition were quite stable and were not further considered in the presentation of the results.

For biologists working with the microbial communities using T-RFLP as a molecular tool, it is ideal that appropriate statistical tools are available for doing the analyses. Therefore, we have implemented our method into a user-friendly software package (T-BAPS) utilizing internally the OpenBugs environment for the MCMC simulations. Our package is freely downloadable from the URL http://www.abo.fi/fak/mnf/mate/jc/software/t-baps.html. It has built-in graphics for exploring the clustering solutions and it enables an intuitive investigation of the estimated posterior clustering uncertainties.

References

Abdo, Z., Schuette, U., Bent, S., Williams, C., Forney, L., and Joyce, P. (2006). Statistical methods for characterizing diversity of microbial communities by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes. *Environmental Microbiology*, 8(5):929–938.

Bock, H. (1996). Probabilistic models in cluster analysis. *Comp. Stat. Data Analysis*, 23:5–28.

Cheeseman, P., Kelly, J., and Self, M. (1988). Autoclass: A bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning*, pages 54–64. Morgan Kaufmann Publishers.

Corander, J., Gyllenberg, M., and Koski, T. (2007). Random partition models and exchangeability for Bayesian identification of population structure. *Bulletin of Mathematical Biology*, 69:797–815.

Corander, J. and Tang, J. (2007). Bayesian analysis of population structure based on linked molecular information. *Mathematical Biosciences*, 205:19–31.

Dai, J. and Lieu, L. (2006). Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology*, 5(1):article 6.

Duda, R., Hart, P., and Stork, D. (2000). *Pattern classification and scene analysis.* New York: Wiley.

Gyllenberg, M., Koski, T., and Verlaan, M. (1997). Classification of binary vectors by stochastic complexity. *J. Multiv. Analysis.*, 63:47–72.

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

Li, C. and Biswas, G. (2002). Unsupervised learning with mixed numeric and nominal data. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):673–690.

Marsh, T. (1999). Terminal restriction fragment length polymorphism (T-RFLP): an emerging method for characterizing diversity among homologous populations of amplification products. *Curr Opin Microbiol.*, 2(3):323–7.

Park, S., Ku, Y., Seo, M., Kim, D., and Yeon, J. (2006). Principal component analysis and discriminant analysis (PCA-DA) for discriminating profiles of terminal restriction fragment length polymorphism (T-RFLP) in soil bacterial communities. *Soil Biology and Biochemistry*, 38:2344–2349.

Raftery, A., Newton, M., Satagopan, J., and Krivitsky, P. (2006). Estimating the integrated likelihood via posterior simulation using the harmonic mean

identity. In Bernardo, J., editor, *Bayesian Statistics 8*, pages 1–45. Oxford University Press.

Rees, G., Baldwin, D., Watson, G., Perryman, S., and Nielsen, D. (2004). Ordination and significance testing of microbial community composition derived from terminal restriction fragment length polymorphisms: application of multivariate statistics. *Antonie van Leeuwenhoek*, 86:339–347.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. B*, 59:731–792.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Roy. Statist. Soc*, 64:583–640.

Stephens, M. (2000). Dealing with label-switching in mixture models. *J. Roy. Statist. M.*, 62(4):795–809.

Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006). Making BUGS open. *R News*, 6:12–17.

Urakawa, H., Yoshida, T., Nishimura, M., and Ohwada, K. (2005). Characterization of depth-related changes and site-specific differences of microbial communities in marine sediments using quinone profiles. *Fisheries Science*, 71:174–182.

Wang, M., Ahrne, S., Antonsson, M., and Molin, G. (2004). T-RFLP combined with principle component analysis and 16s rRNA gene sequencing: an effective strategy for comparison of fecal microbiota in infants of different ages. *Journal of Microbiological Methods*, 59:53–69.

Zhou, C. and Wakefield, J. (2006). A Bayesian mixture model for partitioning gene expression data. *Biometrics*, 62:515–525.