JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 10, 2008
© Mary Ann Liebert, Inc.
Pp. 1315–1327
DOI: 10.1089/cmb.2008.0043

A Bayesian Nonparametric Approach for Comparing Clustering Structures in EST Libraries

ANTONIO LIJOI,[1] RAMSÉS H. MENA,[2] and IGOR PRÜNSTER[3]

ABSTRACT

Inference for Expressed Sequence Tags (ESTs) data is considered. We focus on evaluating the redundancy of a cDNA library and, more importantly, on comparing different libraries on the basis of their clustering structure. The numerical results we achieve allow us to assess the effect of an error correction procedure for EST data and to study the compatibility of single EST libraries with respect to merged ones. The proposed method is based on a Bayesian nonparametric approach that allows to understand the clustering mechanism that generates the observed data. As specific nonparametric model we use the two parameter Poisson–Dirichlet (PD) process. The PD process represents a tractable nonparametric prior which is a natural candidate for modeling data arising from discrete distributions. It allows prediction and testing in order to analyze the clustering structure featured by the data. We show how a full Bayesian analysis can be performed and describe the corresponding computational algorithm.

Key words: Bayesian nonparametrics, clustering, EST analysis, species sampling, two-parameter Poisson–Dirichlet process.

1. INTRODUCTION

Classical species sampling problems have recently gained renewed interest due to their importance in genomic applications. In such inferential problems, one is interested in the species composition of a certain population containing an unknown number of species and only a sample drawn from it is available. Specifically, a sample of size n, $X_1, \dots, X_n$, will exhibit $K_n \in \{1, \dots, n\}$ distinct species with frequencies $(N_1, \dots, N_{K_n})$, where clearly $\sum_{i=1}^{K_n} N_i = n$. Given such a sample, interest lies in estimating the number of new species to be observed in an additional sample of size m and in

[1] Department of Economics and Quantitative Methods, University of Pavia, Pavia, and Institute of Applied Mathematics and Computer Science (IMATI), National Research Council (CNR), Milan, Italy.

[2] Institute of Applied Mathematics and Systems, National Autonomous University of Mexico, Mexico City, Mexico.

[3] Department of Statistics and Applied Mathematics and ICER, University of Turin, Torino, and Carlo Alberto College, Moncalieri, Italy.

determining the decay of the discovery probability as a function of the sample size m. Estimators for such quantities were first provided in Good (1953) and Good and Toulmin (1956), whereas among recent contributions we mention Mao (2004, 2007), Wang et al. (2005), and Lijoi et al. (2007a). When samples from different but somehow related populations are available it is often also fundamental to test whether the two samples can be explained by the same model or, in other terms, are compatible in a suitable sense. This happens, for instance, when analyzing ecological data about species diversity in different geographical regions.

There is a variety of experimental platforms that give rise to data for which these issues are relevant. In this paper, we focus on the analysis of Expressed Sequence Tags (ESTs), which naturally falls within such a framework. Indeed, the available information about a certain cDNA library, which contains a large and unknown number of unique genes, is represented by an ESTs basic sample of size n, each EST identifying a specific gene or species. Since repetitions are common on these experiments, the sample consists of $K_n \leq n$ distinct genes with frequencies, or expression levels, $(N_1, \ldots, N_{K_n})$. Starting from these data one needs statistical methods for assessing some features of the whole library. The above mentioned problems then take on the interpretation of: predicting the number of new genes that will arise from further sequencing, which provides a measure of redundancy of the library; comparing different libraries, either from the same organism under different biological conditions, such as cancer versus normal, or from different parts of the same organism, in order to establish, which library yields more information so to optimize the sequencing procedure.

Before outlining the model we are going to exploit, we briefly explain how ESTs arise and which particular EST datasets we will focus on. ESTs are created by partially sequencing the $5'$ and/or the $3'$ ends of randomly isolated gene transcripts that have been converted into cDNA (Adams et al., 1991). Analysis of ESTs constitute a cost-effective tool in genomic technologies. Their public access through dbEST, a division of the National Center for Biotechnology Information that collects and stores information of EST data, provides researchers with elements for identification, discovery and characterization of organisms. It also constitutes the basis for other gene expression profiling such as cDNA microarrays. cDNA libraries typically contain many expressed mRNAs corresponding to the same gene, hence ESTs derived from these mRNAs might be redundant. This leads to the need of bioinformatics methods to compare, cluster and annotate EST data. Of particular interest is the transcript abundance: this can be obtained through EST clustering procedures and it allows to identify the abundance of mRNA species in the cDNA library. The gene cluster profile, underlying cDNA libraries, describes the gene diversity of an organism and constitutes an appealing source of genomic information. Therefore, the development of suitable computational and statistical methodologies to analyze such data is of critical importance.

EST datasets have limitations when used as means to identify genome content since they only represent a small portion of a coding sequence and their annotation, processing and assembly are prone to several kinds of errors. Although many of these are efficiently addressed, others are difficult to avoid, such as those arising from the imperfect nature of the enzymes used in the construction of cDNA libraries.

In this paper we use four cDNA libraries from *Arabidopsis thaliana*, previously prepared and studied by Wang et al. (2004, 2005). This organism constitutes a model for understanding several biological phenomena in plant sciences. Two libraries, namely green silique and flower bud, consist of reverse sequenced ESTs ($3'$); other two libraries, 2–6 weeks above-ground organs (ABGR) and root, arise from forward sequenced ESTs ($5'$). A more exhaustive description of the data, as well as their availability, can be found in Wang et al. (2004, 2005).

1.1. Bayesian nonparametric methods

Applications of Bayesian methods have recently exploited very general families of discrete nonparametric priors within complex hierarchical mixture models for density estimation and semiparametric regression (Müller and Quintana, 2004). However, very little has been done when the data are actually generated by a discrete probability distribution: in this case it would be appropriate to model the data according to a discrete nonparametric prior. Such an argument obviously applies to species sampling problems and, in particular, to EST analysis. Our inference approach is model–based and we assume that the EST data arise from some (unknown) discrete probability distribution $\tilde{P}$. We take a nonparametric Bayesian perspective and complete the model with a prior distribution for the random probability measure $\tilde{P}$. That is, we treat

$\tilde{P}$ as an infinite–dimensional random element. The use of such probability distributions on probability distributions is characteristic for nonparametric Bayesian inference.

Among different proposals of priors, a convenient choice is represented by the two parameter Poisson–Dirichlet process (Pitman, 1995). Such a random probability measure, denoted by $\tilde{P}$, can be defined as

$$\tilde{P} = \sum_{j=1}^{\infty} \tilde{p}_j \, \delta_{X_j^*}$$

where $\delta_{X_i^*}$ is the point mass at X_i^*, the random weights $\tilde{p}_j$'s are independent from the X_i^*'s and the X_i^*'s are i.i.d. from a continuous distribution P_0. Moreover, the $\tilde{p}_j$'s admit a stick–breaking representation

$$\tilde{p}_j = V_j \prod_{i=1}^{j-1} (1 - V_i) \qquad \text{with} \quad V_j \overset{\text{ind}}{\sim} \text{Beta}(1 - \sigma, \theta + j\,\sigma)$$

and $\sigma \in (0, 1)$, $\theta > -\sigma$, having set by convention $\prod_{i=1}^{0} := 0$. A useful and accessible introduction to this, and to more general species sampling, priors can be found in Pitman (1996). In the sequel, the two parameter Poisson–Dirichlet process will be denoted as PD(σ, θ). The above structure is clearly appropriate to model data related to the detection of species: since P_0 is continuous, the X_i^*'s are distinct and denote different species labels and $\tilde{p}_i$ can be seen as the random proportion with which the species X_i^* is present in the population. Our main focus will be on comparing clustering structures of samples sequenced from different libraries: since such clustering structures heavily depend on the parameters σ and θ (Pitman, 2006), it is important to specify a prior also for (σ, θ). The desired comparison among populations will then be carried out by computing Bayes factors.

Other interesting inferential applications of the two parameter PD pocess can be found in Ishwaran and James (2001), Teh et al. (2006), and Teh (2006). It is worth noting that the popular Dirichlet process (Ferguson, 1973) can be seen as a member of such a family of priors and corresponds to the case where $\sigma \to 0$. Finally, the PD(σ, θ) process belongs to a wide and tractable class of random probability measures, introduced in Gnedin and Pitman (2005), which are said of *Gibbs–type*. See also Griffiths and Spano (2007) for a study of their age–ordered frequencies.

1.2. Outline

In Section 2, we provide a description of the nonparametric model which is used to fit the data. Subsection 2.1 introduces the framework within which hypothesis testing for comparing clustering structures is carried out. Subsection 2.2 adapts the estimators of Lijoi et al. (2007a) to this framework and describes a Blackwell–MacQueen sampling scheme used for computing them. Section 3 is devoted to the application to real EST data. In Subsection 3.1, we evaluate the effect of the ISO error correction procedure of Wang et al. (2004), whereas in Subsection 3.2 we study the compatibility of individual EST libraries with respect to merged ones. Subsection 3.3 reports results of a sensitivity analysis. Finally, Section 4 contains some concluding remarks. A complete and exhaustive description of the numerical output, on which our predictions rely, is provided as Supplementary Material. (See online supplementary material at *www.liebertonline.com*.)

2. THE BAYESIAN NONPARAMETRIC MODEL

We assume that the EST data form an exchangeable sequence $(X_n)_{n \geq 1}$. By de Finetti's representation theorem an infinitely exchangeable sequence can be characterized by a hierarchical model, with the X_n's as a random sample from some distribution $\tilde{P}$ and a prior on $\tilde{P}$. Within the parametric Bayesian framework, the distribution $\tilde{P}$ is assumed to belong to some some parametric class and the model is completed with a prior on the parameters. Nonparametric Bayesian inference is less restrictive by allowing $\tilde{P}$ to vary within a larger class and assuming a nonparametric prior for $\tilde{P}$. We use the two parameter Poisson–Dirichlet process PD(σ, θ) (Pitman, 1995). This is equivalent to assuming that

$$X_i | \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$$

$$\tilde{P} | (\sigma, \theta) \sim \text{PD}(\sigma, \theta). \tag{1}$$

Note that the two parameter PD selects discrete distributions (almost surely), which is a desirable feature in this context, in contrast to situations where one has to model continuous data. Moreover, in order to carry out a full Bayesian analysis and, in particular, to develop a testing procedure, we specify a hyperprior $\pi_{\sigma,\theta} = \pi_\sigma \times \pi_\theta$ for its parameters. This differs from the setup of Lijoi et al. (2007a) where an empirical Bayes specification for (sigma,theta) is adopted: such an approach cannot be pursued here since it does not allow to perform a test which compares different EST libraries.

When analyzing EST libraries, one is interested in the number of distinct genes and their expression levels. This naturally leads us to consider the partition structure induced by model (1): a sample of n EST data yields $K_n \in \{1, \dots, n\}$ distinct gene species with corresponding frequencies $\boldsymbol{N} = (N_1, \dots, N_{K_n})$ such that $\sum_{j=1}^{K_n} N_j = n$. Given (σ, θ), the probability distribution for K_n and the frequencies $\boldsymbol{N}$ induced by Equation (1) coincides with Pitman's sampling formula (Pitman, 1995), which is of the form

$$Pr[K_n = k, \boldsymbol{N} = \boldsymbol{n} | (\sigma, \theta)] = \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{(\theta + 1)_{n-1}} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1} \tag{2}$$

with $\boldsymbol{n} = (n_1, \dots, n_k)$ and $(a)_n = a(a+1) \cdots (a+n-1)$ being the ascending factorial with $(a)_0 \equiv 1$. The partition distribution (2) represents a generalization of the popular Ewens' sampling formula (Ewens, 1972) which corresponds to the partition structure induced by the Dirichlet process and is recovered by letting $\sigma \to 0$. See Arratia et al. (2003) for a stimulating account. In order to obtain the joint distribution of K_n and $\boldsymbol{N}$, we simply have to marginalize Equation (2) with respect to (σ, θ) leading to

$$Pr[K_n = k, \boldsymbol{N} = \boldsymbol{n}] = \int_0^1 \int_0^\infty \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{(\theta + 1)_{n-1}} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1} \, \pi_\sigma(\mathrm{d}\sigma) \, \pi_\theta(\mathrm{d}\theta) \tag{3}$$

When used for predictive purposes, we will see in Section 2.2 that Equation (3) can be interpreted as the prior distribution on the clustering structure of an EST sequence.

2.1. *Hypothesis tests for comparing libraries*

We consider the issue of pairwise comparison between different libraries. The main factor driving the comparison is the kind of clustering present in the different libraries: two different libraries are considered equivalent if they give rise to similar groupings of the observations. This, in turn, entails that they produce a similar number of distinct genes when sampling from such libraries and can then be considered equivalent in terms of redundancy. In this setting, a reasonable measure of redundancy is the proportion of genes, detected in further sampling, which coincide with genes that have been already observed in the basic sample. Note that our goal is not to assess the homogeneity by means of a test for equality of the distributions of the data. Such a task seems not achievable in this context where: (i) data are categorical; (ii) a very small portion of the population is observed; (iii) the supports of the distributions are unknown as well as the number of support points. Moreover, interest relies in redundancy rather than in the labels which identify the single genes. For this reason a sensible approach for the assessment of homogeneity versus heterogeneity of two cDNA libraries can be traced back to a comparison of the values of the distribution parameters responsible for the particular grouping that is observed. Within the PD(σ, θ) model we are adopting, this basically reduces to comparing the parameters (σ, θ) corresponding to the libraries. Suppose there are N_1 and N_2 data from the first and second library, respectively. The model one can refer to consists in taking

$$X_i^{(j)} \,|\, (\tilde{P}_1, \tilde{P}_2) \overset{\text{ind}}{\sim} \tilde{P}_j \qquad j = 1, 2$$

$$X_i^{(j)} \,|\, (\tilde{P}_1, \tilde{P}_2) \overset{\text{iid}}{\sim} \tilde{P}_j \qquad i = 1, \dots, N_j$$

with independent PD priors $\tilde{P}_j|(\sigma_j, \theta_j) \sim \mathrm{PD}(\sigma_j, \theta_j)$, for $j = 1, 2$, and a hyperprior π for $(\sigma_1, \theta_1, \sigma_2, \theta_2)$. To facilitate the desired inference we write the hyperprior as a mixture

$$\pi(\mathrm{d}\sigma_1, \mathrm{d}\theta_1, \mathrm{d}\sigma_2, \mathrm{d}\theta_2) = \lambda_0\, \pi_0(\mathrm{d}\sigma, \mathrm{d}\theta)\, \mathbb{I}_{(\sigma_1,\theta_1)=(\sigma_2,\theta_2)} + (1-\lambda_0)\pi_1(\mathrm{d}\sigma_1, \mathrm{d}\theta_1)\, \pi_2(\mathrm{d}\sigma_2, \mathrm{d}\theta_2)\mathbb{I}_{(\sigma_1,\theta_1)\neq(\sigma_2,\theta_2)}, \quad (4)$$

where $\mathbb{I}_A$ is 1 whenever A is true and 0 otherwise. The value of λ_0 is an indication of how much, *a priori*, we believe that the two libraries are equivalent. When there is equivalence, the distribution of the vector $(\sigma_1, \theta_1, \sigma_2, \theta_2)$ degenerates on a two dimensional space. On the other hand, when $(\sigma_1, \theta_1) \neq (\sigma_2, \theta_2)$ we specify independent priors for (σ_1, θ_1) and (σ_2, θ_2). Finally, we suppose that $\pi_j(\mathrm{d}\sigma_j, \mathrm{d}\theta_j) = \pi_{j,\sigma}(\mathrm{d}\sigma_j)\pi_{j,\theta}(\mathrm{d}\theta_j)$, for $j = 1, 2$, and $\pi_0(\mathrm{d}\sigma, \mathrm{d}\theta) = \pi_\sigma(\mathrm{d}\sigma)\, \pi_\theta(\mathrm{d}\theta)$. Within this framework we are, then, going to establish a suitable decision rule for testing

$$H_0 : (\sigma_1, \theta_1) = (\sigma_2, \theta_2) \qquad \text{vs} \qquad H_1 : (\sigma_1, \theta_1) \neq (\sigma_2, \theta_2)$$

where (σ_1, θ_1) and (σ_2, θ_2) are the parameters corresponding to the two libraries. Formally, the problem can be given a statistical answer by resorting to a test based on the use of Bayes factors. In order to describe the Bayes factor, set $\Pi^{(i)}(\sigma_i, \theta_i) = \Pr[K_n^{(i)} = k_i,\ \boldsymbol{N}_i = \boldsymbol{n}_i|(\sigma_i, \theta_i)]$ as defined in Equation (2). Hence, the Bayes factor is

$$\mathrm{BF}_{01} = \frac{\int_0^1 \int_0^\infty \Pi^{(1)}(\sigma, \theta)\, \Pi^{(2)}(\sigma, \theta)\, \pi_\sigma(\mathrm{d}\sigma)\pi_\theta(\mathrm{d}\theta)}{\prod_{j=1}^2 \int_0^1 \int_0^\infty \Pi^{(j)}(\sigma_j, \theta_j)\, \pi_{j,\sigma}(\mathrm{d}\sigma_j)\, \pi_{j,\theta}(\mathrm{d}\theta_j)} \quad (5)$$

We will use $2\log(\mathrm{BF}_{01})$ to establish whether H_0 must be rejected or not. See Kass and Raftery (1995) for a discussion of Bayes factors and the indications of thresholds. An alternative test we consider is based on the idea that σ is the main parameter being responsible for the specific grouping of the data that have been observed. Such a claim is motivated by the asymptotic behavior of K_n as n diverges: indeed, K_n grows at a rate n^σ (Pitman, 2006, Theorem 3.8). See also Lijoi et al. (2007b) for further considerations on this point. Hence, it is also important to verify that a rejection of the null hypothesis is not solely due to differences in θ. To this end, we evaluate a Bayes factor for testing $H_0 : \sigma_1 = \sigma_2$ vs $H_1 : \sigma_1 \neq \sigma_2$. In order to do so, we slightly change the prior specification of the model and in place of Equation (4) we have

$$\pi(\mathrm{d}\sigma_1, \mathrm{d}\theta_1, \mathrm{d}\sigma_2, \mathrm{d}\theta_2) = \pi_{1,\theta}(\mathrm{d}\theta_1)\pi_{2,\theta}(\mathrm{d}\theta_2)\ \left\{\lambda_0\pi_0(\mathrm{d}\sigma_1)\mathbb{I}_{\sigma_1=\sigma_2} + (1-\lambda_0)\pi_{1,\sigma}(\mathrm{d}\sigma_1)\pi_{2,\sigma}(\mathrm{d}\sigma_2)\mathbb{I}_{\sigma_1\neq\sigma_2}\right\} \quad (6)$$

Hence, the Bayes factor is in this case

$$\mathrm{BF}'_{01} = \frac{\int_0^1 \int_0^\infty \int_0^\infty \prod_{j=1}^2 \Pi^{(j)}(\sigma, \theta_j)\, \pi_{j,\theta}(\mathrm{d}\theta_j)\, \pi(\mathrm{d}\sigma)}{\prod_{j=1}^2 \int_0^1 \int_0^\infty \Pi^{(j)}(\sigma_j, \theta_j)\, \pi_{j,\sigma}(\mathrm{d}\sigma_j)\, \pi_{j,\theta}(\mathrm{d}\theta_j)} \quad (7)$$

2.2. *Bayesian nonparametric estimators and a Blackwell–MacQueen sampling scheme*

In order to interpret the testing results and to provide a comprehensive analysis of EST data, it is useful to combine the Bayes factors with the Bayesian nonparametric estimators derived in Lijoi et al. (2007a). Here, we briefly recall and adapt them to the case where hyperpriors on (σ, θ) are specified.

Having observed a sample of size n, the probability distribution of detecting j new genes in a future sample of size m, denoted by $P_m^{\boldsymbol{X}}(j)$, is the main tool for deriving estimators of: (i) the number of new genes in an additional sample and (ii) the probability of discovering a new gene at the $(n+m+1)$th draw. To derive an expression for $P_m^{\boldsymbol{X}}(j)$, consider first $P_m^{(k,n)}(j|\sigma, \theta)$, which denotes the probability of recording j new tags in a future sample of size m, conditional on data with k distinct tags in the basic sample of size n, and conditional on σ and θ. Indeed, we have

$$P_m^{(k,n)}(j|\sigma, \theta) = \frac{(\theta+1)_{n-1}}{(\theta+1)_{m+n-1}}\, \frac{\prod_{i=k}^{k+j-1}(\theta+i\sigma)}{\sigma^j}\, \frac{1}{j!} \sum_{i=0}^{j} (-1)^i \binom{j}{i} (n-(i+k)\sigma)_m, \quad (8)$$

and, additionally,

$$\pi_{\sigma,\theta}(\mathrm{d}\sigma, \mathrm{d}\theta|\boldsymbol{X}) \propto \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k}(1-\sigma)_{n_j-1}\, \pi_\sigma(\mathrm{d}\sigma)\pi_\theta(\mathrm{d}\theta).$$

Thus, the desired distribution is given by

$$P_m^{\boldsymbol{X}}(j) = \int_0^1 \int_0^\infty P_m^{(k,n)}(j|\sigma,\theta)\, \pi_{\sigma,\theta}(\mathrm{d}\sigma, \mathrm{d}\theta|\boldsymbol{X}).$$

The expected number of new genes observed in a future sample of size m, given (σ, θ), is

$$\hat{E}_m^{(k,n)}(\sigma,\theta) = \sum_{j=1}^{m} j\, \frac{1}{j!} \sum_{i=0}^{j} (-1)^i \binom{j}{i} \frac{(k+\theta/\sigma)_j}{(\theta+n)_m} (n-(i+k)\sigma)_m$$

and, hence, the Bayes estimator for the expected number of new genes is

$$\hat{E}_m^{\boldsymbol{X}} = \int_0^1 \int_0^\infty \hat{E}_m^{(k,n)}(\sigma,\theta)\, \pi_{\sigma,\theta}(\mathrm{d}\sigma, \mathrm{d}\theta|\boldsymbol{X}). \tag{9}$$

Then, the discovery probability can be estimated by

$$\hat{D}_m^{\boldsymbol{X}} = \int_0^1 \int_0^\infty \frac{\theta + \left[k + \hat{E}_m^{(k,n)}(\sigma,\theta)\right]\sigma}{\theta+n+m}\, \pi_{\sigma,\theta}(\mathrm{d}\sigma, \mathrm{d}\theta|\boldsymbol{X}). \tag{10}$$

The highest posterior density intervals corresponding to Equations (9) and (10) can be derived in a quite straightforward way from Equation (8).

In contrast to the case of fixed (σ, θ), where the exact estimators are easily computed as in Lijoi et al. (2007a, 2007b), in such a hierarchical setup the quantities we have been describing cannot be easily evaluated for large values of n and m. Hence, we outline a suitable algorithm for achieving this task.

The main idea for computing integrals in Equations (9) and (10) consists in applying a generalization of the well-known Blackwell–MacQueen Pólya urn scheme (Pitman, 1996). The implementation turns out to be straightforward given the simple form of the predictive distributions associated to the PD(σ, θ) prior. Note that, when sampling (σ, θ) from the posterior one needs to implement a Gibbs sampling step. The full conditionals are as follows

$$\pi_\sigma(\mathrm{d}\sigma|\boldsymbol{X},\theta) = \pi_\sigma(\mathrm{d}\sigma|K_n = k, \boldsymbol{N} = \boldsymbol{n}) \propto \prod_{i=1}^{k-1}(\theta+i\sigma) \prod_{j=1}^{k}(1-\sigma)_{n_j-1}\, \pi_\sigma(\mathrm{d}\sigma) \tag{11}$$

$$\pi_\theta(\mathrm{d}\theta|\boldsymbol{X},\sigma) = \pi_\theta(\mathrm{d}\theta|K_n = k) \propto \frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1}}\, \pi_\theta(\mathrm{d}\theta)$$

In computational terms it is extremely convenient to rewrite $\prod_{j=1}^{k}(1-\sigma)_{n_j-1}$ in Equation (11) as $\prod_{l=1}^{l^*}[(1-\sigma)_{l-1}]^{r_l}$ where, for $l = 1, 2, \ldots, l^*$, r_l denotes the number of genes with expression level l, i.e., $r_l = \sum_{i=1}^{k} \mathbb{I}_{n_i=l}$, where l^* is the maximum level of expression. The computational advantage is due to the fact that typically l^* is much smaller than the number of distinct genes k.

We agree to denote by $X_1^*, \ldots, X_k^*$ the k labels identifying the distinct genes observed in the basic sample $\boldsymbol{X} = (X_1, \ldots, X_n)$. Moreover, recall that n_l is the frequency of X_l^* in $X_1, \ldots, X_n$. Having this in mind, the algorithm works as follows:

(1) Generate N pairs of (σ, θ) values. In order to do so, fix an initial value σ_0 and, then, sample θ_0 from $\pi_\theta(\mathrm{d}\theta|\boldsymbol{X}, \sigma_0)$. Then, at iteration $i \in \{1, \ldots, N\}$
   (1a) Sample σ_i from $\pi_\sigma(\mathrm{d}\sigma|\boldsymbol{X}, \theta_{i-1})$
   (1b) Sample θ_i from $\pi_\sigma(\mathrm{d}\theta|\boldsymbol{X}, \sigma_i)$

(2) Correspondingly to each pair (σ_i, θ_i) in (1), simulate a sample $X^{(i)}_{n+1}, \dots, X^{(i)}_{n+m}$ by resorting to a Blackwell–MacQueen urn scheme which generates $X^{(i)}_{n+r}$, given the data $X_1, \dots, X_n$ and the previously sampled values $X^{(i)}_{n+1}, \dots, X^{(i)}_{n+r-1}$, for any $r = 1, \dots, m$ as

$$X^{(i)}_{n+r} = \begin{cases} \text{new} & \text{with probab. } (\theta + (k + j_{r-1})\sigma)/(\theta + n + r - 1) \\ X^*_l & \text{with probab. } (n_l + m_{l,r-1} - \sigma)/(\theta + n + r - 1) \quad l = 1, \dots, k \\ X^*_{l,r-1} & \text{with probab. } (q_{l,r-1} - \sigma)/(\theta + n + r - 1) \quad l = 1, \dots, j_{r-1} \end{cases}$$

where $X^*_{1,r-1}, \dots, X^*_{j_{r-1},r-1}$ are the new genes, not coinciding with any of $X^*_1, \dots, X^*_k$, detected in $X_{n+1}, \dots, X_{n+r-1}$; $m_{l,r-1}$ is the number of observations in $X_{n+1}, \dots, X_{n+r-1}$ that coincide with X^*_l; and, finally, $q_{l,r-1}$ is the number of observations $X_{n+1}, \dots, X_{n+r-1}$ coinciding with $X^*_{l,r-1}$.

Hence, after a burn-in period of size N_0, the output of the algorithm is a collection of future scenarios $\{(X^{(i)}_{n+1}, \dots, X^{(i)}_{n+m}) : i = N_0, \dots, N\}$ which will be used in order to evaluate the main quantities we are interested in for inferential purposes. Letting $j^{(i)}_m$ denote the number of new distinct genes observed in $X^{(i)}_{n+1}, \dots, X^{(i)}_{n+m}$, the estimator (9) is evaluated as $\hat{E}^{\boldsymbol{X}}_m \approx \frac{1}{N-N_0} \sum_{i=N_0+1}^{N} j^{(i)}_m$ whereas the discovery probability is approximated by

$$\hat{D}^{\boldsymbol{X}}_m \approx \frac{1}{N - N_0} \sum_{i=N_0+1}^{N} \frac{\theta_i + \left[k + j^{(i)}_m\right] \sigma_i}{\theta_i + n + m}.$$

3. CLUSTERING STRUCTURE OF EST DATA

3.1. Specification of the model parameters

The implementation of the model in Equation (1) requires the specification of the prior distribution for (σ, θ), i.e., one needs to assess π_σ and π_θ. We have taken σ to be distributed according to a beta(a^*, b^*) law discretized over the grid $\{0.01, 0.02, \dots, 0.99\}$, whereas a Poi$(\hat\theta)$ has been specified for θ. The choices of π_σ, as a discretization of a continuous distribution, and of π_θ supported by $\mathbb{N}$ are motivated by the desire on one hand to calculate the exact Bayes factors and on the other to sample from the exact posterior quantities when computing the estimators. The hyperparameters (a^*, b^*) and $\hat\theta$ are elicited by suitably centering the priors on the empirical Bayes specification set forth in Lijoi et al. (2007a), where no prior is introduced for (σ, θ). Such a specification leads to a pair $(\hat\sigma, \hat\theta)$ defined as

$$(\hat\sigma, \hat\theta) = \arg\max_{(\sigma,\theta)} \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{(\theta + 1)_{n-1}} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1}.$$

Hence, we have that $\mathbb{E}[\theta] = \hat\theta$, whereas (a^*, b^*) are fixed in a way that the mode of π_σ is $\hat\sigma$ and that $\mathrm{Var}(\sigma) = 1/27$. Such a value of the variance is the largest compatible with the requirement that the mode is $\hat\sigma$. This specification is set for all the libraries to be considered and the resulting hyperparameters are reported in Table 1. The fact that, for all EST samples, $\hat\sigma$ is far away from 0, which corresponds to the Dirichlet case, witnesses the advisability of resorting to its two parameter extension.

Specifically, when comparing two different libraries, the denominator of the Bayes factors in Equations (5) and (7) is evaluated by choosing $\pi_{j,\sigma}$ and $\pi_{j,\theta}$ as above, for $j = 1, 2$. Moreover we need to define the distributions π_0 corresponding to the null hypothesis in Equations (6) and (4). To this end, let N_1 and N_2 denote as before the sizes of the basic sample from library 1 and library 2, respectively. For $\pi_0 = \pi_{\theta_0}\pi_{\sigma_0}$ in Equation (6) we set: π_{θ_0} is Poi(θ_0), with $\theta_0 := (N_1\hat\theta_1 + N_2\hat\theta_2)/(N_1 + N_2)$; π_{σ_0} is a discretized beta distribution with mode $\sigma_0 = (N_1\hat\sigma_1 + N_2\hat\sigma_2)/(N_1 + N_2)$ and variance 1/27. When performing a test involving only σ, we use as a prior π_0 in Equation (4) the previous discretized beta π_{σ_0}.

Table 1. $(\hat{\sigma}, \hat{\theta})$ Represent the Empirical Bayes Specification of (σ, θ)

Library	$\hat{\sigma}$	$\hat{\theta}$	a^*	b^*
Silique	0.43	1186	3.03	2.69
ABGR before ISO	0.66	409	3.10	1.08
ABGR after ISO	0.59	471	3.33	1.62
Root before ISO	0.66	536	3.10	1.08
Root after ISO	0.59	599	3.33	1.62
Flower Bud	0.64	267	3.19	1.23
ABGR and Root before	0.6	891	3.31	1.54
ABGR and Root after	0.51	1048	3.29	2.20
Silique and Flower bud	0.45	1210	3.11	2.58

The hyperprior for θ is then a Poi$(\hat{\theta})$ and for σ it is a Beta(a^*, b^*), where (a^*, b^*) are fixed so to have the mode in $\hat{\sigma}$ and the variance equal to $1/27$.

3.2. *ISO error correction*

As mentioned above, EST data play a crucial role in gene annotation and inference of the number of expressed genes in the transcriptome of an organism. However, a major problem for predicting the discovery of new genes is due to the EST clustering error: this affects the basic sample on which predictions are based. As pointed out and thoroughly discussed in Wang et al. (2004), errors from different sources can bias the number of observed genes upward by 35–40%. Such a problem is especially relevant for 5′ ESTs such as the ABGR and Root data considered here, whereas errors are less frequent for 3′ ESTs such as Silique and Flowerbud data. It is to be noted that for 5′ ESTs, the false separation error (to be understood as insufficient overlap (ISO) between ESTs from the same gene) can cause up to 80% of all the clustering errors: hence, ESTs present a higher number of distinct genes than they actually should. Wang et al. (2004) proposed a method, termed ISO error correction, for overcoming this problem. Given this procedure corrects a large portion of the transcripts errors, the data, after its application, can be considered "good" data.

Here, we face the problem of establishing how much inferences are affected by the ISO error. We perform such an analysis by comparing EST data of the same library before and after having applied the ISO correction procedure: it is clear that, if the clustering structure of the data before and after ISO correction is compatible in a suitable sense, also inferences based on these data will be different but compatible with the hypothesis that the underlying model is the same. We first consider the estimates of the expected number of new genes. From Tables 2–5 of the Supplementary Material a steady change in the estimates after the ISO correction is apparent. Indeed, the estimate of the expected number of new genes decreases by a percentage ranging between 15% and 20%. The same phenomenon can be observed with reference to the discovery probabilities. Table 2 here reports the expected number of new genes and the discovery probabilities (both with corresponding highest posterior density intervals) for the ABGR data before and after ISO correction at selected values. More exhaustive results for this case can be found in Tables 2 and 3 of the Supplementary Material. Still for the ABGR data, Figure 1 displays our Bayesian estimates for $\hat{E}_m^X$'s. For comparison purposes, we also report a plot of the corresponding Good–Toulmin frequentist estimator (Good and Toulmin, 1956): as it is well–known that such an estimator features reliable predictions only up to $m = n$. In contrast, for the Bayesian nonparametric estimator the relative dimension of m with respect to n is not an issue.

Since, in the analysis of EST data, prediction is required also for future samples significantly larger than the basic sample, recently there have been various proposals of alternative frequentist estimators which partially overcome this drawback allowing prediction up to $m = 2n$ (Mao, 2007; Lijoi et al., 2007c). In particular, Wang et al. (2005) report the estimates of $\hat{E}_m^X$, before and after the ISO correction, for additional samples of size at most equal to $2n$. They also point out that their method underestimates the actual expected number of new genes. By comparing our estimates with those in Wang et al. (2005) it is

TABLE 2. EXPECTED NUMBER OF NEW GENES AND DISCOVERY PROBABILITIES FOR DIFFERENT SIZES OF THE ADDITIONAL SAMPLE m COMPUTED FOR THE ABGR LIBRARY BEFORE AND AFTER ISO ERROR CORRECTION OF THE BASIC SAMPLE

m	$\hat{E}_m^{\boldsymbol{X}}$	*HPD(95%)*	$\hat{D}_m^{\boldsymbol{X}}$	*HPD(95%)*
ABGR before ISO correction				
1000	383	(338, 414)	.3738	(.3698, .3767)
5000	1765	(1640, 1903)	.3218	(.3144, .3299)
10,000	3267	(3061, 3521)	.2836	(.2752, .2940)
15,000	4619	(4208, 4903)	.2588	(.2460, .2677)
20,000	5870	(5384, 6271)	.2409	(.2287, .2511)
25,000	7045	(6461, 7568)	.2272	(.2148, .2383)
30,000	8156	(7473, 8797)	.2161	(.2036, .2278)
ABGR after ISO correction				
1000	334	(301, 367)	.3230	(.3203, .3257)
5000	1508	(1377, 1618)	.2700	(.2631, .2758)
10,000	2758	(2588, 2982)	.2324	(.2262, .2406)
15,000	3856	(3557, 4143)	.2083	(.2000, .2163)
20,000	4853	(4520, 5239)	.1911	(.1836, .1999)
25,000	5774	(5383, 6261)	.1780	(.1706, .1872)
30,000	6640	(6182, 7249)	.1676	(.1601, .1776)

The sizes of the basic samples are $n = 5811$ with $j = 3116$ distinct genes before ISO correction and $n = 5812$ with $j = 2883$ after ISO correction.

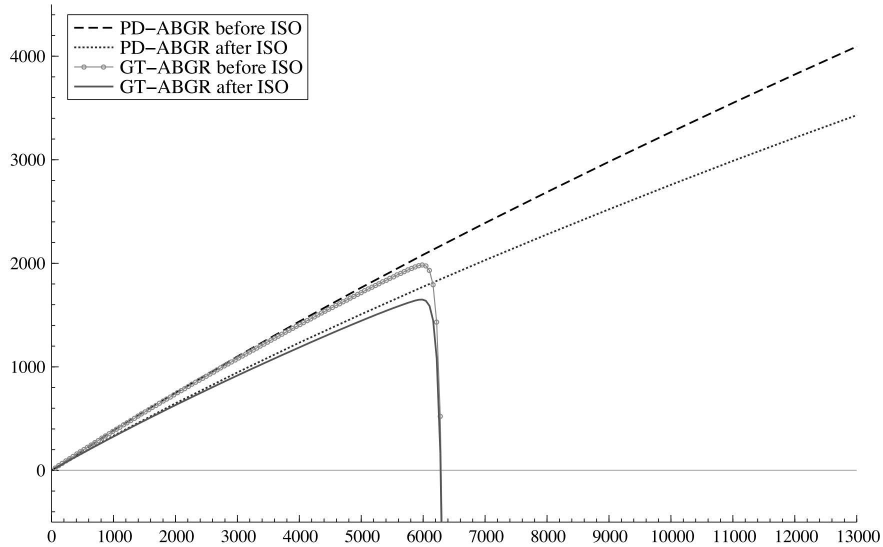

FIG. 1. Expected number of new genes, $\hat{E}_m^{\boldsymbol{X}}$, for the ABGR libraries before and after ISO correction. The plot shows the estimates using the Poisson-Dirichlet (PD) with beta-Poisson prior and the estimates derived from the Good and Toulmin (GT) estimator.

TABLE 3. BAYES FACTORS (REPORTED AS 2 ln BF) FOR TESTING THE CLUSTERING STRUCTURE OF (A) LIBRARIES BEFORE AND AFTER ISO ERROR CORRECTION AND (B) MERGED VERSUS INDIVIDUAL LIBRARIES

	Beta-Poisson prior		*Uniform*	
Test	H_0^C	H_0^S	H_0^C	H_0^S
Test for ISO error correction				
ABGR (ISO) vs. ABGR	−11.162	−15.492	−6.058	−5.279
Root (ISO) vs. Root	−11.105	−14.771	−6.446	−3.790
ABGR and Root (ISO) vs. ABGR and Root	−35.392	−47.788	−30.681	−19.461
Test for merging of libraries				
ABGR and Root (ISO) vs. ABGR (ISO)	−37.564	−19.490	−28.862	−4.078
ABGR and Root (ISO) vs. Root (ISO)	−10.857	−19.854	−5.824	−3.658
Silique and Flower bud vs. Silique	1.477	3.300	6.272	4.725
Silique and Flower bud vs. Flower bud	−159.899	−168.379	−127.286	−77.492

The 2 ln BF are computed under both the Beta–Poisson and the uniform prior specifications. The null–hypotheses are specified as $H_0^C : (\sigma, \theta) = (\sigma_0, \theta_0)$ and $H_0^S : \sigma = \sigma_0$.

apparent that our estimators do not incur in such a problem. Besides, it is to be noted that we can consider additional samples of any size m.

Turning back to the problem of establishing the impact of the ISO correction on the clustering structure, we compute log–Bayes factors for data before and after ISO error correction in Table 3. From Table 3, the relevance of the ISO procedure is apparent. Indeed, in the three comparisons (1) ABGR after ISO versus ABGR before ISO, (2) Root after ISO versus Root before ISO, and (3) ABGR and Root after versus ABGR and Root before, the log Bayes factor supports a neat evidence against the null hypothesis. Hence, the clustering structure before and after the ISO correction differs significantly. Such an evidence highlights that without the ISO correction wrong inferences can be drawn from EST data and it should always be performed before trying to draw conclusions from the data. In the following subsection we deal with the problem of merging of libraries: in doing this we only consider after ISO correction data.

3.3. *Merging of libraries*

The machinery for hypothesis testing we have employed in the previous subsection will now be used in order to assess the effect of merging of different libraries. In other terms, one might be interested in evaluating whether it is equivalent to examine individual libraries prepared from different tissues of the same organism or to analyze directly merged libraries. Once again, such an equivalence will be judged according to the corresponding clustering structures of the data. An analysis of merging should obviously involve libraries with the same data structure; that is, we consider merging of libraries having both either 3′ or 5′ ESTs. Hence, with reference to our datasets, we consider the individual ABGR and Root samples (5′ ESTs) and compare them with the union of the two samples. The same comparison is pursued with Silique and Flower bud samples (3′ ESTs).

The merged sample for the ABGR and Root library contains 11,529 ESTs with 5243 distinct genes, whereas for the merged sample of the Silique and Flower bud library we have 17,784 ESTs with 6595 distinct genes. It is worth noting that the number of distinct genes in the merged sample is smaller than the sum of the distinct genes within individual samples since the libraries have co–expressed genes. For instance, the individual samples of the ABGR and of the Root libraries contain 2883 and 3126 distinct genes, respectively: if no genes were co–expressed, the merged sample would have contained 6009 distinct genes, whereas it exhibits just 5243.

The compatibility between libraries is investigated by resorting to the Bayes factors described in Equations (5) and (7). The results are reported in Table 3. The values of the log Bayes factors provide strong evidence of incompatibility between the ABGR and Root libraries with the merged library. The same conclusion can be reached when comparing the Flower bud library with the merged of Silique and Flower bud library. On the other hand, the log Bayes factor arising when one compares the Silique with the

TABLE 4. EXPECTED NUMBER OF NEW GENES AND DISCOVERY PROBABILITIES AT $n+m$ EQUAL TO THE SIZE OF THE MERGED LIBRARIES WITH BETA–POISSON PRIOR

Basic sample	n	j	m	$\hat{E}_m^{\mathbf{X}}$	*HPD(95%)*	$n+m$	$j+\hat{E}_m^{\mathbf{X}}$	$\hat{D}_m^{\mathbf{X}}$	*HPD(95%)*
ABGR after ISO	5812	2883	5717	1700	(1573, 1813)	11,529	4583	0.2633	(0.2570, 0.2689)
Root after ISO	5891	3126	5638	1835	(1686, 2096)	11,529	4961	0.2898	(0.2824, 0.3026)
ABGR and Root after ISO	11,529	5243	0			11,529	5243	0.3192	
Silique	12,330	5093	5454	1230	(1116, 1443)	17,784	6323	0.2058	(0.2032, 0.2108)
Flower Bud	5503	2564	12,281	3222	(2981, 3462)	17,784	5786	0.2204	(0.2118, 0.2290)
Silique and Flower Bud	17,784	6595	0			17,784	6595	0.2341	

merged Silique and Flower bud library suggests positive evidence in favor of compatibility. This finding could be possibly explained by the fact that Silique accounts for more than two-thirds of the merged library.

Having established the incompatibility of the libraries, it is now important to interpret the information conveyed by the Bayes factors. Indeed, there are two main sources of incompatibility between two different libraries: the first one is due to a too small amount of co–expressed genes and the second one is, on the contrary, a too large amount of co–expressed genes in the two libraries. It is clearly essential to understand which of the two causes has led to the actual rejection of the null hypothesis, since they lead to different conclusions about the benefits of merging. Indeed, if the incompatibility is due to a very small amount of shared genes it seems reasonable to analyze the libraries separately. Vice versa, when a large amount of distinct genes are shared by the two libraries, it would be appropriate to work with the merged library since the separate analysis of individual libraries would yield a waste of expensive efforts. Hence, in order to complete the analysis, we now aim at identifying the source of incompatibility for the two cases under investigation. To achieve this goal, we resort to our estimators for evaluating the expected number of new genes. In particular, we need to compare the number of distinct genes between samples of the same size. Since the merged library is of size $N_1 + N_2$, for library 1 one needs to estimate the number of new distinct genes in an additional sample of size N_2. The overall estimate of the number of distinct genes in a sample of size $N_1 + N_2$ from library 1 is, then, obtained as a sum of the actual number observed in the basic sample of size N_1 and the estimated number in the additional sample of size N_2 which is evaluated by means of Equation (9). The same procedure is adopted for library 2. The results are reported in Table 4.

For the ABGR and Root libraries one obtains an estimated number of distinct genes (for a sample of size $N_1 + N_2 = 11{,}529$) equal to 4583 and 4961, respectively. Since the merged library exhibited 5243 distinct genes, we conclude that the libraries are incompatible because too few genes are co–expressed and a separate individual analysis is advisable. As for the Silique and Flower bud libraries, on a global sample of size $N_1 + N_2 = 17{,}784$, one estimates the number of distinct genes as 6323 and 5786, respectively. The actual number of distinct genes in the merged library is 6595 thus pointing out that the two libraries have too few genes in common. However, unlike the previous case, only the Flower bud displays a too small number of distinct genes to be compatible with the merged library. Hence, in both cases we get to the conclusion of incompatibility because of the too small number of co–expressed genes.

3.4. Sensitivity analysis

In order to check the sensitivity of the results with respect to the choice of the priors, we also consider an alternative specification, which does not make use of the information conveyed by the empirical Bayes estimates $(\hat{\sigma}, \hat{\theta})$. Hence, we choose for σ and θ discrete uniform priors with support points $\{0.01, 0.02, \ldots, 0.99\}$ and $\{0, 1, \ldots, 2000\}$, respectively. Consequently, the point estimates for the expected number of new genes and for the discovery probability do not depend on $(\hat{\sigma}, \hat{\theta})$. Moreover, when testing the null hypothesis of compatibility between two different libraries, one does not need to deal with the issue of properly centering the prior under the null assumption since it is uniform as well.

Tables 2–10 of the Supplementary Material report the estimates for $\hat{E}_m^{\mathbf{X}}$ and $\hat{D}_m^{\mathbf{X}}$ corresponding to both, the beta–Poisson and the uniform, prior specifications. It is apparent that the influence of the different prior is almost negligible thus providing convincing support for the robustness of the proposed method. As for the comparison of libraries and the effect of ISO correction, the numerical output we have obtained under the uniform priors is reported in Table 3 here (see also Table 11 of the Supplementary Material): one notices that the results basically replicate those obtained under the beta–Poisson prior.

4. CONCLUSION

The present paper has proposed a full Bayesian nonparametric analysis for problems of species sampling where one is interested in (1) testing the compatibility of clustering structures featured by samples taken from different populations; (2) estimating the number of new distinct species to be observed in an additional sample; and (3) evaluating the discovery probability. The specific application to EST data considered in the paper naturally fits into such a framework. However, we emphasize that the methods apply to any inferential problem with data arising from discrete distributions with a large and unknown number of support points. An important aspect of the proposed methodology is the robustness of the inferences with respect to the prior specification: indeed choices of both “informative" and “non–informative" priors lead to the same conclusions. Finally, exact computation of the Bayes factors is straightforward and the simulation algorithm we have adopted in order to obtain predictions is simple to implement since it arises as a generalization of the Blackwell–MacQueen urn scheme. Software is available upon request from the authors.

ACKNOWLEDGMENTS

We gratefully acknowledge the hospitality of the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, where important parts of this work were carried out during the INI Programme Bayesian Nonparametric Regression. A.L. and I.P. are partially supported by MiUR (grants 2006/134525 and 2006/133449, respectively). R.H.M. was partially supported by CONACyT (grant 50160) and also gratefully acknowledges support by ICER for visiting the University of Turin.

DISCLOSURE STATEMENT

No conflicting financial interests exist.

REFERENCES

Adams, M., Kelley, J., Gocayne, J., et al. 1991. Complementary DNA sequencing: expressed sequence tags and Human Genome Project. *Science* 252, 1651–1656.

Arratia, R., Barbour, A., and Tavaré, S. 2003. *Logarithmic Combinatorial Structures: A Probabilistic Approach*. EMS Monographs in Mathematics, Zurich.

Ewens, W.J. 1972. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3, 87–112.

Good, I.J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* 40, 237–264.

Ferguson, T.S. 1973. A Bayesian analysis of some nonparametric problems. *Ann. Statist.* 1, 209–230.

Gnedin, A., and Pitman, J. 1956. Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S. Peterburg. Otdel. Mat. Inst. Steklov.* 325, 83–102.

Griffiths, R.C., and Spanò, D. 2007. Record indices and age-ordered frequencies in exchangeable Gibbs partitions. *Electr. J. Probab.* 12, 1101–1130.

Good, I.J., and Toulmin, G.H. 1956. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* 43, 45–63.

Ishwaran, H., and James, L.F. 2001. Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Assoc.* 96, 161–173.

Kass, R., and Raftery, A. 1995. Bayes factors. *J. Am. Statist. Assoc.* 90, 773–795.

Lijoi, A., Mena, R.H., and Prünster, I. 2007a. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* 94, 769–786.

Lijoi, A., Mena, R.H., and Prünster, I. 2007b. Controlling the reinforcement in Bayesian nonparametric mixture models. *J. R. Statist. Soc. Series B* 69, 715–740.

Lijoi, A., Mena, R.H., and Prünster, I. 2007c. A Bayesian nonparametric method for prediction in EST analysis. *BMC Bioinform.* 8, 339.

Mao, C.X. 2004. Prediction of the conditional probability of discovering a new class. *J. Am. Statist. Assoc.* 99, 1108–1118.

Mao, C.X. 2007. Estimating species accumulation curves and diversity indices. *Statistica Sinica* 17, 761–775.

Müller, P., and Quintana, F.A. 2004. Nonparametric Bayesian data analysis. *Statist. Sci.* 19, 95–110.

Pitman, J. 1995. Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* 102, 145–158.

Pitman, J. 1996. Some developments of the Blackwell-MacQueen urn scheme, 245–267. In: Ferguson, T.S., Shapley, L.S., and MacQueen, J.B., eds. *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell, Lecture Notes, Monograph Series, Volume 30*. IMS, Hayward.

Pitman, J. 2006. Combinatorial stochastic processes. *Lect. Notes Math.* 1875, Springer, Berlin.

Teh, Y.W. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. *Proc. Annu. Mtg. Assoc. Comput. Ling.*, 44.

Teh, Y.W., Jordan, M.I., Beal, M.J., et al. 2006. Hierarchical Dirichlet processes. *J. Am. Statist. Assoc.* 101, 1566–1581.

Wang, J.P.Z., Lindsay, B.G., Cui, L., et al 2005. Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries *BMC Bioinform.* 6, 300.

Wang, J.P.Z, Lindsay, B.G., Leebens–Mack, J., et al. 2004. EST clustering error evaluation and correction *Bioinformatics* 20, 2973–2984.

Address reprint requests to:
Dr. Igor Prünster
Department of Statistics and Applied Mathematics and ICER
University of Turin
Piazza Arbarello 8
10122 Torino, Italy

E-mail: igor@econ.unito.it