INTERNATIONAL CENTRE FOR ECONOMIC RESEARCH

ICER

WORKING PAPER SERIES

Antonio Lijoi, Ramsés H. Mena and Igor Prünster

A BAYESIAN NONPARAMETRIC METHOD FOR PREDICTION IN *EST* ANALYSIS

Working Paper no. 16/2007
April 2007

APPLIED MATHEMATICS AND QUANTITATIVE METHODS
WORKING PAPER SERIES

A Bayesian nonparametric method for prediction in EST analysis

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

[1] Dipartimento Economia Politica e Metodi Quantitativi, Università degli Studi di Pavia, via San Felice 5, 27100 Pavia, and CNR-IMATI, Milano, Italy. *e-mail*: lijoi@unipv.it.

[2] IIMAS, National Autonomous University of Mexico, Mexico City, A.P. 20-726, Mexico. *e-mail*: rasmes@sigma.iimas.unam.mx.

[3] Dipartimento di Statistica e Matematica Applicata, ICER and Collegio Carlo Alberto, Università degli Studi di Torino, Piazza Arbarello 8, 10122 Torino, Italy. *e-mail*: igor@econ.unito.it

Motivation: Expressed sequence tags (ESTs) analyses are a fundamental tool for gene identification in organisms. Given a preliminary EST sample from a certain library, several statistical prediction problems arise. In particular, it is of interest to estimate how many new genes can be detected in a future EST sample of given size and also to determine the gene discovery rate: these estimates represent the basis for deciding whether to proceed sequencing the library and, in case of a positive decision, a guideline for selecting the size of the new sample. Such information is also useful for establishing sequencing efficiency in experimental design and for measuring the degree of redundancy of an EST library.

Results: In this work we propose a Bayesian nonparametric approach for tackling statistical problems related to EST surveys. In particular, we provide estimates for: a) the coverage, defined as the proportion of unique genes in the library represented in the given sample of reads; b) the number of new unique genes to be observed in a future sample; c) the discovery rate of new genes as a function of the future sample size. The Bayesian nonparametric model we adopt conveys, in a statistically rigorous way, the available information into prediction. Our proposal has appealing properties over frequentist nonparametric methods, which become unstable when prediction is required for large future samples. EST libraries studied in Susko and Roger (2004), with frequentist methods, are analyzed in detail.

1 Introduction

Expressed Sequence Tags (ESTs) are generated by partially sequencing randomly isolated gene transcripts that have been converted into cDNA. From their introduction Adams et al. (1991), ESTs have played an important role in the identification, discovery and characterization of organisms as they provide an attractive and efficient alternative to full genome sequencing. The resulting transcript sequences and their corresponding abundances are the main focus of interest providing the identification and level of expression of genes. Despite the novel advances in sequencing technology Emrich et al. (2007), EST projects aimed at library construction and sequencing still incur in big expenses and therefore suitable cost-effectiveness thresholds must be established. This suggests that there is the need for assessing the relative redundancy of various libraries prepared from the same organism in order to detect which one yields new genes at a higher rate. Indeed, there are ‘normalization’ protocols which aim at making the frequencies of genes in the library more uniform typically improving the discovery rate. However, performing such protocols is expensive. Hence, the decision, whether to proceed with sequencing of a non–normalized library or to resort to a normalization procedure, has to balance carefully the involved costs: such a decision is necessarily based on statistical estimates of the coverage of the given sample, of the expected number of new genes in a future sample and on the future discovery rate. Note that ideally one would like to sequence the smallest possible portion of the library and, based on the outcome, predict the tentative future sequencing well beyond the size of the given dataset.

The main statistical issues to be faced, once an initial sample of EST is available, are as follows:

a) *Coverage*: Coverage can be seen as the proportion of genes in the library represented in the initial sample or, equivalently, the probability that a new read will not produce a new gene. The coverage estimate provides a first description of redundancy of the library.

b) *Expected number of new genes*: Having observed an initial sample of size n generated from the cDNA library and estimated its coverage, prediction of outcomes of further reads is in order. The first question to answer is: ‘How many new unique genes are expected to be detected in an additional EST dataset of targeted size m?’ Such estimates provide, then, an overall measure of redundancy of the library with reference to a further EST survey.

c) *Discovery rate*: In addition to the expected number of genes in a future sample of size m, it is also important to establish the rate at which the probability of discovering a new gene decays as more and more reads are recorded. In other words, interest lies in determining the probability that the $n+m+1$–th read leads to a new gene, given the observed initial sample of size n and regardless of the experimental outcome yielded by the m intermediate draws. The availability of the discovery rate as a function of the size of the future sample m, represents then a pointwise predictive measure of the evolution of redundancy as the sequencing ideally proceeds.

Note that the combination of the measures under b) and c) provide a natural guideline for selecting the size of a future sample m. Supposing the targeted number of new genes is j, the expected number of new genes allows to select the minimum survey size $\bar{m}$ which leads to j new genes. Then, one can resort to the discovery rate: in case it is relatively low around $\bar{m}$, it may be convenient to reduce the size of the future sample in a way that the discovery rate does not fall below a threshold suggested by the problem at issue. On the other hand, if the discovery rate around $\bar{m}$ is still relatively high, one may decide to enlarge the survey size. Moreover, the information conveyed by b) and c) is useful in comparing libraries and, again, it is worth considering these estimates together. Indeed, suppose we have to compare two libraries and that, for a fixed size m of the additional sample, library 1 yields a larger expected number of new genes but a lower discovery rate in comparison with library 2. If the sample size m is increased to $m+m'$, for m' sufficiently large, the comparison between the two libraries can lead to different conclusions in the sense that a larger number of new genes is predicted for library 2. This happens because library 1 features a lower discovery rate, which implies that, within the additional m' draws, the expected number of new genes is lower for library 1. With reference to ‘normalization’ protocols, this means that the decision whether to carry it out or not should depend also on the foreseen sample size. For instance, the normalized *Mastigamoeba balamuthi* data we analyze exhibit a higher discovery rate, with respect to the non–normalized one, for small m. But, since the discovery rate has a faster decay, it appears that, already for moderately large m, the effect of the ‘normalization’ is exhausted producing fewer number of new genes.

The three questions raised above can be seen as particular instances of classical species sampling problems: indeed, in the present context each species takes on the meaning of gene and the population is given by the library. Species problems appear in a variety of different applied situations such as astronomy, ecology, linguistics, machine learning, population biology. We now briefly recall well–known estimation methods which have recently been applied to EST data and then outline the key ideas of our Bayesian nonparametric approach.

1.1 Estimation methods

The main frequentist tools, that are useful for inference on the cDNA library properties described in the previous section, are based on the theory set forth in Good (1953) and Good and Toulmin (1956), where nonparametric estimators for the sample coverage and the expected number of new species to be detected in a future sample of size m given the initial sample are provided. The estimator of the sample coverage Good (1953) coincides with the proportion of distinct species represented by at least two units in the sample. Good attributes the original idea to Turing and this explains why it is usually referred to as Turing estimator. The popular Good–Toulmin estimator for the number of new species to be observed in a future sample is derived in Good and Toulmin (1956) and, as a by–product, they also are able to evaluate the discovery probability. Recently, the interest in species sampling problems has remarkably grown, mainly due to their importance in genomics. Indeed, Mao (2004) studies various properties of the Good–Toulmin estimator and shows that it can be also viewed as a non–parametric empirical Bayes estimator. In Susko and Roger (2004), the authors suggest a parametric variation of the Good–Toulmin estimator. An alternative to it is presented in Wang et al. (2005), where the detection of ESTs from each gene in EST sequencing is modeled by means of a Poisson process whose intensity is governed by some unknown distribution. It is to be noted that all frequentist nonparametric approaches lead to reliable estimates for the number of new genes in an additional sample only if its size is not too large. For instance, if the size of the additional survey m is larger than the initial sample n, it is well–known that the Good–Toulmin predictor can become a monotone decreasing function of m: this leads to the paradox of predicting fewer new genes by enlarging the additional sample size m. Even the nonparametric alternative proposed in Wang et al. (2005) yields reliable results only when $m \leq 2n$. This fact is also outlined in Mao (2007). Hence, one needs to resort to a parametric framework if one wishes to predict the number of new genes for large m. As we will see, the relative dimension of m with respect to n is not an issue in a Bayesian nonparametric framework, and the expected number of new genes that will be discovered in m further reads is monotone increasing with respect to m.

The application of Bayesian methods in this area of research is, to the authors' knowledge, quite modest even if the Bayesian learning scheme is very well suited for making predictions with EST data. An early contribution, based on a model for sampling from a finite population, is provided by Hill (1979) where posterior estimates of the coverage are obtained. However, computational problems do not allow, in this approach, a direct and effective evaluation of the expected number of new species in a future sample. Recently, Lijoi et al. (2007) have proposed new Bayesian nonparametric estimators for the problems a)–c) mentioned above. The prior distribution they employ is induced by a family of exchangeable *Gibbs* random partitions. See Pitman (2006) for an interesting review of recent advances and applications of the theory of Gibbs random partitions. Their application to a Bayesian inferential framework is very useful since they provide a general scheme which encompasses some of the most notable nonparametric priors such as the Dirichlet and the two parameter Poisson–Dirichlet process. In this paper, we adapt the general formulas derived in Lijoi et al. (2007) to the case in which the prior distribution is the two parameter Poisson–Dirichlet process prior. It will be seen that the expressions we obtain can be evaluated exactly and do not need for any supplementary simulation scheme. Moreover, such a Bayesian approach does not incur in any problem for large values of m since all possible behaviors of future EST data are incorporated in the probabilistic model.

1.2 Outline of the paper

The outline of the paper is as follows. In Section 2 we present the four EST datasets we analyze together with the results arising from the application of our Bayesian nonparametric approach. In Section 3 we describe Pitman's sampling formula and explain why it constitutes a natural framework in which EST sequences can be embedded. Then, the resulting estimators are provided and the empirical Bayes approach for tuning the prior parameters is discussed. Section 4 contains some concluding remarks.

2 EST Datasets and results

The datasets we analyze consist of ESTs samples obtained from cDNA libraries from two different organisms: the amitochondriate protist *Mastigamoeba balamuthi* (non-normalized and normalized libraries, where the normalized library was prepared from the non-normalized library) and *Naegleria gruberi* libraries, prepared from cells grown under different culture conditions, aerobic and anaerobic. These data sets have been previously analyzed in Susko and Roger (2004), where a full account of their preparation is detailed. It is worth mentioning that our approach assumes full-length cDNA clones and high quality sequence reads. Therefore, possible errors associated with the clustering procedure are not considered. For the statistical identification and evaluation of types of clustering errors one may incur in EST sequencing, the reader is referred to Wang et al. (2004).

Specifically, each EST survey consists of n reads with k unique genes and corresponding frequencies $n_1, \ldots, n_k$, i.e. n_i is the number of tags displaying the i–th gene in the initial sample of size n. Clearly,

$\sum_{i=1}^{k} n_i = n$. The reads can equivalently be clustered according to their level of expression, that is

$$r_l \equiv \sum_{i=1}^{k} I\{n_i = l\}, \qquad \text{for} \qquad l = 1, 2, \ldots, s \tag{1}$$

where $I(A)$ is an indicator of A: $I(A) = 1$ if A is true and 0 otherwise. Note that s represents the maximum level of expression among unique genes in the sample and that the number of positive r_l's is typically smaller than s.

Below the four EST samples are summarized using the compact notation set in (1). For example, the survey of the *naeglaria* aerobic library produces $n = 959$ reads with $k = 473$ unique genes, which are clustered into 17 levels of expression $1, 2, \ldots, 12, 16, 17, 18, 27, 55$. For the first level we have $r_1 = 346$, meaning that 346 genes appear just once, that is $n_1 = n_2 = \cdots = n_{346} = 1$. For the second level $r_2 = 57$ implies that 57 genes appear twice and, hence, $n_{347} = n_{348} = \cdots = n_{403} = 2$ and so on up to $r_{55} = 1$, which means that 1 gene is represented 55 times yielding $n_{473} = 55$.

Data: EST surveys information clustered into levels of expression. Source: Susko and Roger (2004)

Library	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**27**	**55**	k	n
N. Aer.	346	57	19	12	9	5	4	2	4	5	4	1	0	0	0	1	1	1	1	1	473	959
N. Anaer.	491	72	30	9	13	5	3	1	2	0	1	0	1	3	0	0	0	0	0	0	631	969
M. Non–n.	378	33	21	9	6	1	3	1	1	1	0	0	1	0	5	0	0	0	0	0	460	715
M. Norm.	200	21	14	4	3	3	1	0	1	0	0	0	0	1	0	0	0	0	0	0	248	363

Table 1: Non–normalized and normalized *Mastigamoeba* libraries: the first column provides the size of the additional sample in % of the size of the initial sample, the second the actual size of the additional survey, the third presents the expected number of new genes and the fourth the discovery probability. The estimates in the third and fourth column are accompanied by the 95% highest posterior density intervals.

$\%n$	m	Expected number of new genes in a additional sample of size m	Probability of discovering a new gene at the $(n+m+1)$–th read
Mastigamoeba non-normalized			
50	358	$180 \in (158\,, 204)$	$0.481 \in (0.466\,, 0.498)$
100	715	$346 \in (312\,, 382)$	$0.452 \in (0.434\,, 0.470)$
150	1072	$503 \in (458\,, 550)$	$0.430 \in (0.411\,, 0.449)$
200	1430	$654 \in (599\,, 711)$	$0.412 \in (0.393\,, 0.433)$
250	1788	$799 \in (734\,, 866)$	$0.398 \in (0.379\,, 0.419)$
300	2145	$939 \in (865\,, 1015)$	$0.386 \in (0.367\,, 0.407)$
Mastigamoeba normalized			
50	182	$94 \in (79\,, 111)$	$0.493 \in (0.475\,, 0.512)$
100	363	$180 \in (156\,, 206)$	$0.456 \in (0.434\,, 0.479)$
150	544	$260 \in (229\,, 293)$	$0.428 \in (0.406\,, 0.452)$
200	726	$336 \in (299\,, 375)$	$0.406 \in (0.384\,, 0.430)$
250	908	$408 \in (365\,, 453)$	$0.389 \in (0.366\,, 0.412)$
300	1089	$477 \in (428\,, 528)$	$0.374 \in (0.351\,, 0.398)$

Table 2: Aerobic and anaerobic libraries: the first column provides the size of the additional sample in % of the size of the initial sample, the second the actual size of the additional survey, the third presents the expected number of new genes and the fourth the discovery probability. The estimates in the third and fourth column are accompanied by the 95% highest posterior density intervals.

$\%n$	m	Expected number of new genes in an additional sample of size m	Probability of discovering a new gene at the $(n+m+1)$–th read
		Naegleria aerobic	
50	480	$162 \in (138\,,\,188)$	$0.318 \in (0.307\,,\,0.329)$
100	959	$307 \in (271\,,\,345)$	$0.290 \in (0.277\,,\,0.303)$
150	1438	$441 \in (394\,,\,488)$	$0.270 \in (0.257\,,\,0.282)$
200	1918	$566 \in (510\,,\,624)$	$0.254 \in (0.241\,,\,0.267)$
250	2398	$685 \in (619\,,\,751)$	$0.242 \in (0.229\,,\,0.255)$
300	2877	$798 \in (725\,,\,873)$	$0.231 \in (0.219\,,\,0.244)$
		Naegleria anaerobic	
50	484	$231 \in (206\,,\,258)$	$0.450 \in (0.440\,,\,0.461)$
100	969	$440 \in (402\,,\,478)$	$0.412 \in (0.400\,,\,0.424)$
150	1454	$632 \in (583\,,\,683)$	$0.384 \in (0.371\,,\,0.397)$
200	1938	$812 \in (753\,,\,873)$	$0.362 \in (0.349\,,\,0.375)$
250	2422	$983 \in (915\,,\,1053)$	$0.344 \in (0.332\,,\,0.357)$
300	2907	$1146 \in (1069\,,\,1225)$	$0.330 \in (0.317\,,\,0.342)$

We applied the Bayesian nonparametric method detailed in Section 3 to these datsets and obtained the following results. Denote the unknown proportion of genes (in the whole library) belonging to the i–th class is denoted by p_i. Then, the coverage of the initial sample of size n is given by

$$C = \sum_{i:n_i>0} p_i, \tag{2}$$

which is precisely the proportion of unique genes represented in the initial sample. Our estimates for the coverage are 0.47 and 0.45 for the non–normalized ($n = 715$) and normalized ($n = 363$) *Mastigamoeba*, respectively. This means that, by virtue of the ‘normalization’, an initial sample of about half the size produces almost the same coverage. Moreover, we get 0.64 and 0.49 for the aerobic ($n = 959$) and anaerobic ($n = 969$) *Naegleria*, respectively: clearly, the sequencing for the tissue cultured aerobically is more effective reaching a remarkably higher coverage with an initial sample of the same size. It is worth noting that our results for the coverage match exactly the ones obtained in Susko and Roger (2004), where the frequentist estimator described in Good (1953) was exploited.

Turning attention to predicting the outcomes of future sequencing for the libraries at issue, we focus on the expected number of new genes in an additional sample of size m and on the discovery rate. The first index provides an overall measure of redundancy with respect to the additional sample of size m, whereas the discovery rate predicts the trend at which the discovery probability decays as more and more reads come in. Adopting a Bayesian nonparametric approach these quantities can be estimated rigorously and exactly since such an approach is naturally designed for prediction. In contrast note that, as already anticipated, the Good–Toulmin estimator becomes highly variable and unstable if the size of the additional sample m is larger than the size of the initial sample n. In particular, the Good–Toulmin estimator often produces negative values as estimates for the number of new genes if $m \in (n, 2n)$ and almost always behaves badly for $m > 2n$. Such a phenomenon can be seen in Figure 1 for the two *Naegleria* libraries. In order to overcome these problems, frequentist methods typically give up the flexibility of the nonparametric approach and resort to parametric models, whose fit can be a delicate issue. For instance, Susko and Roger (2004) resort to an approximated version of the Good–Toulmin estimator which assumes a parametric model for the expression levels r_l.

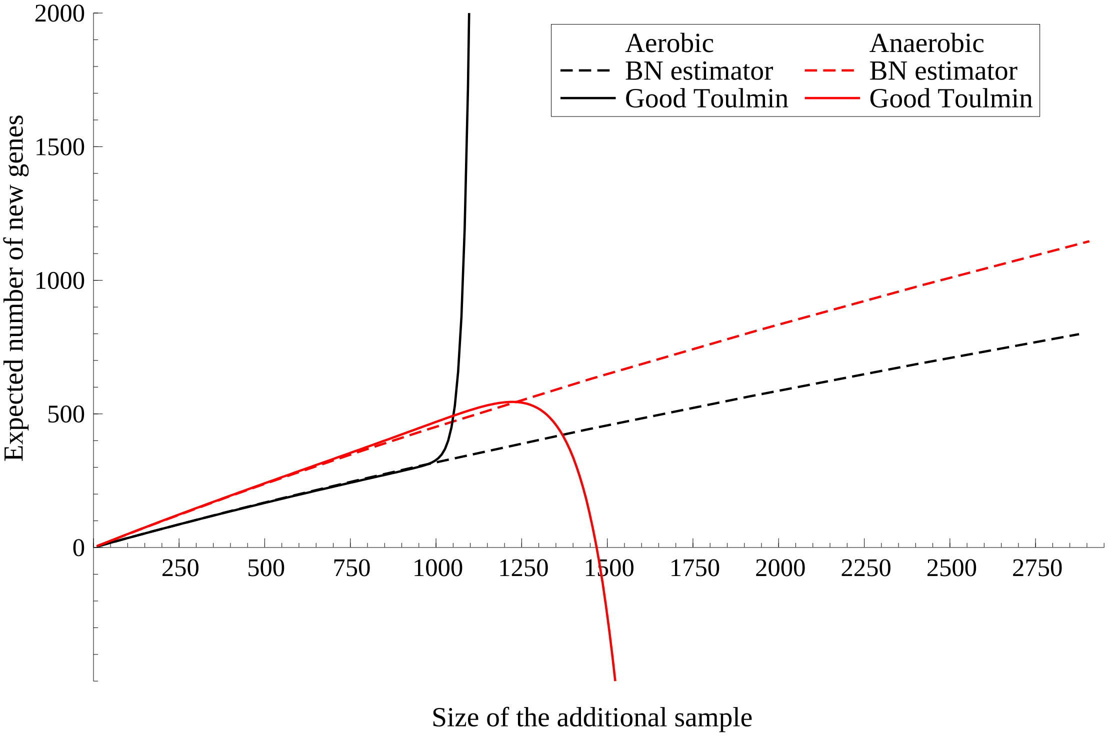

Figure 1: Expected number of new genes in an additional sample for the *Naegleria gruberi* aerobic and anaerobic libraries arising from the application of the Good–Toulmin estimator and of the Bayesian nonparametric estimator.

In order to give a complete picture, it is important to accompany our point estimates by the 95% highest posterior density intervals, which represent the Bayesian counterpart to frequentist confidence intervals (see Bernardo and Smith, 1994). In Tables 1 and 2 the results arising by the application of the Bayesian nonparametric method are displayed.

As for the *Mastigamoeba* libraries, an interesting phenomenon takes place: the survey of the normalized library has achieved almost the same coverage (0.45) as the non–normalized library (0.47), but considering an additional sample it exhibits a significantly faster decay in the discovery rate. Figure 2 compares the discovery rate for the two libraries. It is worth pointing out that our estimates predict that the discovery rates associated to both libraries coincide for $m = 125$ yielding a discovery probability of 0.508. For larger m the non–normalized exhibits a higher discovery rate. This implies that at some point also the estimates for the expected number of new genes in the additional sample will coincide: indeed, this is estimated to happen for $m = 270$, for which 137 new genes are predicted to be identified from both libraries. Hence, for $m > 270$ the expected number of new genes is systematically higher for the non–normalized library. For instance, if $m = 1089$, just 477 new genes are expected for the normalized library and 510 for the non–normalized. Taking m larger, at some point even the highest posterior density intervals will not overlap anymore. Such a behavior hints toward the fact that, in deciding whether to perform a 'normalization' protocol, the sizes of the samples to be drawn from the libraries is a variable to be taken into account.

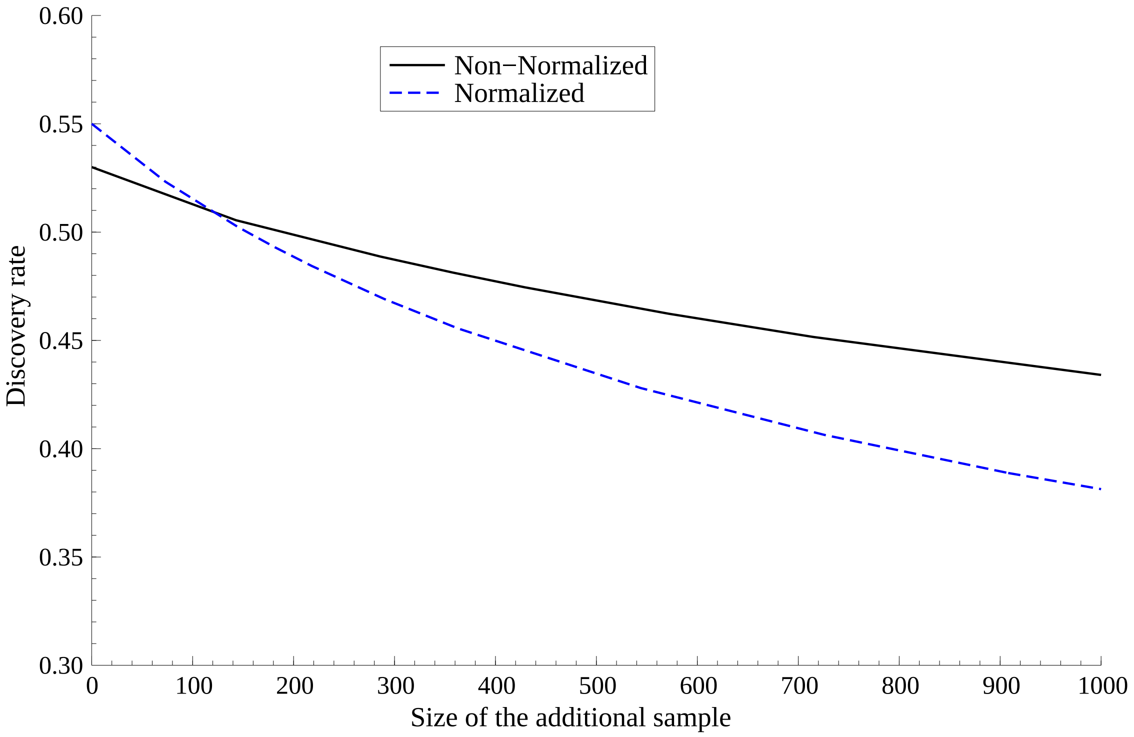

Figure 2: Bayesian nonparametric estimates of the discovery rate associated to the non–normalized and normalized *Mastigamoeba* libraries.

As for the *Naegleria* libraries the behavior is apparent in the sense that the anaerobic library systematically produces more new genes and the discovery probability is sensibly higher at the considered levels of m. Note that the aerobic library presents a slightly slower decay rate but an extremely large m is required for matching the expected number of genes of the anaerobic one. Figure 3 displays the estimated decay rate of the discovery probability for both libraries with the corresponding 95% highest posterior density intervals.

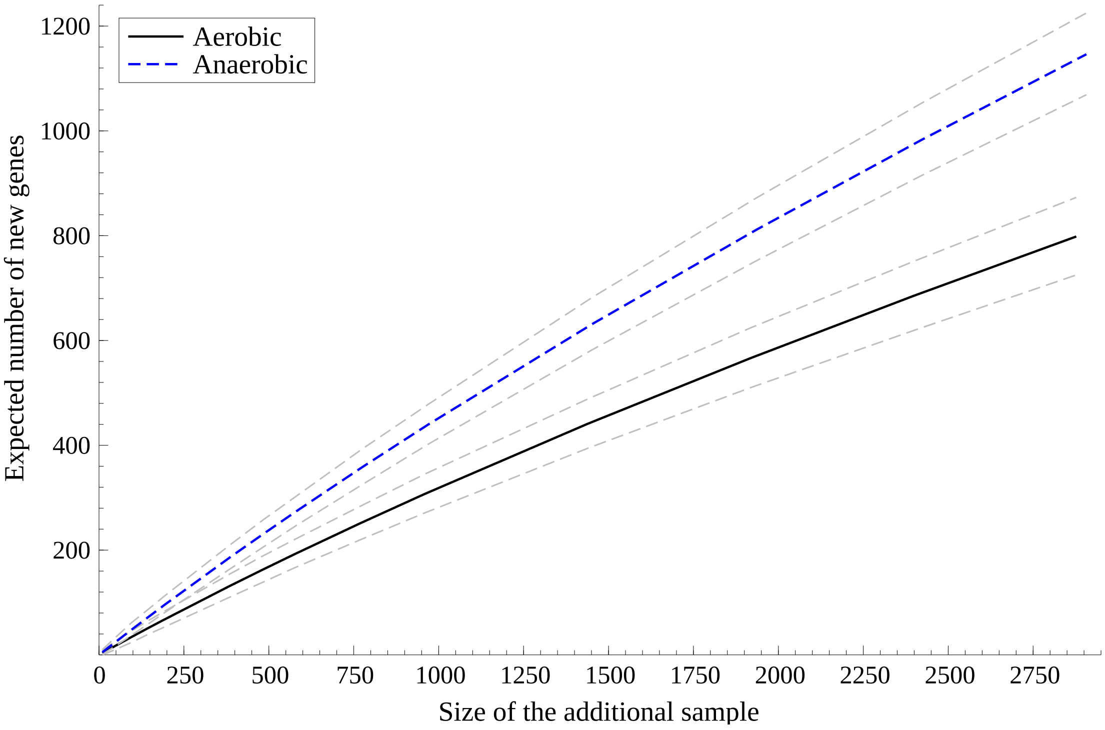

Figure 3: Expected number of new genes in an additional sample and corresponding 95 % highest posterior density intervals for the *Naegleria gruberi* aerobic and anaerobic libraries arising from the application of the Bayesian nonparametric method.

3 A Bayesian nonparametric approach

The primary aim of the Bayesian approach to inference is prediction and Bayesian methods are tailored for conveying the available information into prediction. In particular, for EST sequencing, the main problem of frequentist methods is represented by difficulty of incorporating not yet observed unique genes into the model. This can then produce unpleasant behaviors of estimators such as the one exhibited by the Good–Toulmin estimator discussed before. In contrast, the Bayesian nonparametric approach naturally incorporates the fact that further sequencing will feature new unique genes and leads to consistent predictions.

In our framework we are going to consider a sample of n EST data yielding K_n distinct gene species with corresponding frequencies $\boldsymbol{N} = (N_1, \dots, N_{K_n})$. Clearly $K_n \in \{1, \dots, n\}$ and $\sum_{j=1}^{K_n} N_j = n$. Our basic model is the so–called Pitman's sampling formula (Pitman, 1995) which consists in a probability distribution for K_n and the frequencies $\boldsymbol{N}$ of the form

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

where $\sigma \in (0,1)$, $\theta > 0$, $\boldsymbol{n} = (n_1, \dots, n_k)$ and $(a)_n = a(a+1)\cdots(a+n-1)$ is the ascending factorial with $(a)_0 \equiv 1$. Formula (3) is a generalization of the famous Ewens' sampling formula Ewens (1972) which can be recovered by letting σ tend to zero and it represents a fundamental tool in modern probability theory. See Pitman (2006). Recently, it has found many interesting applications for bacterial taxonomy Gyllenberg and Koski (2001), clustering of microarray gene expression data Zhaohui (2006), mixture models Ishwaran and James (2001), linguistics Teh (2006), among others.

In a Bayesian nonparametric setting, one alternatively obtains model (3) by selecting the two parameter Poisson–Dirichlet process as a prior for the genes proportions within the library. This clearly makes the sequence of tags *exchangeable*, thus implying that the order of appearance of the tags does not influence probability assessments. Such an assumption, which constitutes the Bayesian analog of the frequentist assumption of independent and identically distributed data, is clearly reasonable in the context of EST sequences. Note that we implicitly assume that the sequence of tags can be extended to infinity. However, the size of the library represents an upper bound for the number of unique genes that will be observed and it is always finite, thus implying that all the estimates we are going to obtain will be finite.

As mentioned before, the Bayesian nonparametric approach has the advantage of yielding in a straightforward way predictive distributions for future observations given the data. Considering Pitman's sampling formula, the probability of detecting a new gene from a future observation, given a sample of n tags containing k distinct genes, is

$$(\theta + k\sigma)/(\theta + n) \tag{4}$$

whereas the probability of re-observing the j–th unique gene coincides with

$$(n_j - \sigma)/(\theta + n) \qquad j = 1, \ldots, k. \tag{5}$$

See Pitman (2006). Hence, the coverage coincides with

$$1 - (\theta + k\sigma)/(\theta + n). \tag{6}$$

As it has already pointed out, in the analysis of ESTs one is also interested in evaluating: (i) the expected number of new genes that will be recorded in a further sample of size m and (ii) the discovery probability, which is the probability of observing a new gene in the $(n + m + 1)$–th draw, given the initial sample of size n. The basis for deriving estimators for these quantities is represented by the distribution of the number of new genes to be observed in an additional sample given the initial sample. Such a posterior probability, which can be seen as the predictive distribution for the outcome of additional m reads, is given by

$$P_m^{(k,n)}(j) = \frac{(\theta+1)_{n-1}}{(\theta+1)_{m+n-1}} \frac{\prod_{i=k}^{k+j-1}(\theta+i\sigma)}{\sigma^j} \frac{1}{j!} \sum_{i=0}^{j} (-1)^i \binom{j}{i} (n-(i+k)\sigma)_m \tag{7}$$

See Lijoi et al. (2007) for details on its derivation. From (7) Bayes estimators (under quadratic loss function) for both the expected number of new genes and the dicovery probability have been obtained, within general Gibbs random partition models, in Lijoi et al. (2007). The expected number of new genes observed in a future sample of size m coincides with

$$E_m^{(k,n)} = \sum_{j=1}^{m} j \frac{(k+\theta/\sigma)_j}{(\theta+n)_m} \frac{1}{j!} \sum_{i=0}^{j} (-1)^i \binom{j}{i} (n-(i+k)\,\sigma)_m \tag{8}$$

and the discovery probability turns out to be equal to

$$\hat{D}_m^{(k,n)} = \frac{\theta + \left[k + E_m^{(k,n)}\right] \sigma}{\theta+n+m}. \tag{9}$$

Moreover, the highest posterior density intervals can be derived in a quite straightforward way from (7). The only point left to discuss concerns the specification of the parameters (σ, θ). In order to avoid subjective inputs in the model, (σ, θ) is fixed according to an empirical Bayes rule which consists in choosing σ and θ that maximize (3) corresponding to the observed sample $(k, n_1, \ldots, n_k)$, i.e.

$$(\hat{\sigma}, \hat{\theta}) = \arg\max_{(\sigma,\theta)} \frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}. \tag{10}$$

Figure 4 provides with the contour plots corresponding to the two *Naegleria gruberi* datasets: the parameters maximizing (3) turned out to be $(\hat{\sigma}, \hat{\theta}) = (0.67,\ 46.3)$ for the aerobic case and $(\hat{\sigma}, \hat{\theta}) = (0.66,\ 155.5)$ for the anaerobic case. On the other hand, for the two *Mastigamoeba balamuthi* datasets (normalized and non–normalized) (10) yields $(\hat{\sigma}, \hat{\theta}) = (0.77,\ 46)$ and $(\hat{\sigma}, \hat{\theta}) = (0.7,\ 57)$, respectively. These parameters have then be used for computing the estimators (8) and (9) for the 4 datasets, whose results are reported in Section 2.

At this point it is worth pointing out how the structure of the data influences the choice of the parameters (σ, θ). Indeed, the value of θ is linked to the number of distinct genes observed in the n–sample: the larger k/n the larger $\hat{\theta}$. On the other hand, the value of σ is determined by the configuration of the frequencies $n_1, \ldots, n_k$. Moreover, one may note that, for a given value of θ, the expected number of new genes in (8) is an increasing function of σ: as σ increases one expects that a larger number of new genes is going to be observed in a further m–sample. This is also confirmed by the behavior of $E_m^{(k,n)}$ as σ varies. Indeed, Figure 3 suggests an almost linear increase of $E_m^{(k,n)}$, as a function of m, and accordance with linearity is higher the closer σ is to 1. In contrast, when σ is low and close to 0 the function is concave and $E_m^{(k,n)}$ increases at a lower rate as σ increases.

4 Conclusions

In this paper we have presented a Bayesian nonparametric approach, which relies on Pitman's sampling formula, for prediction problems arising in sequencing of EST libraries. This provides a fully probabilistic model which conveys, in a statistically rigorous way, the available information into prediction. No parametric assumption is made and the prior is fixed using an empirical Bayes approach, thus leaving no room for subjective input. The resulting estimators are applied to four EST libraries and lead to interesting and coherent predictions of the outcome of additional sequencing. The arising information is of great value for researchers providing guidelines in: establishing the quality of a certain library; deciding whether to perform a normalization protocol; choosing whether to proceed with sequencing from a certain library; determining the size of an additional EST survey etc.

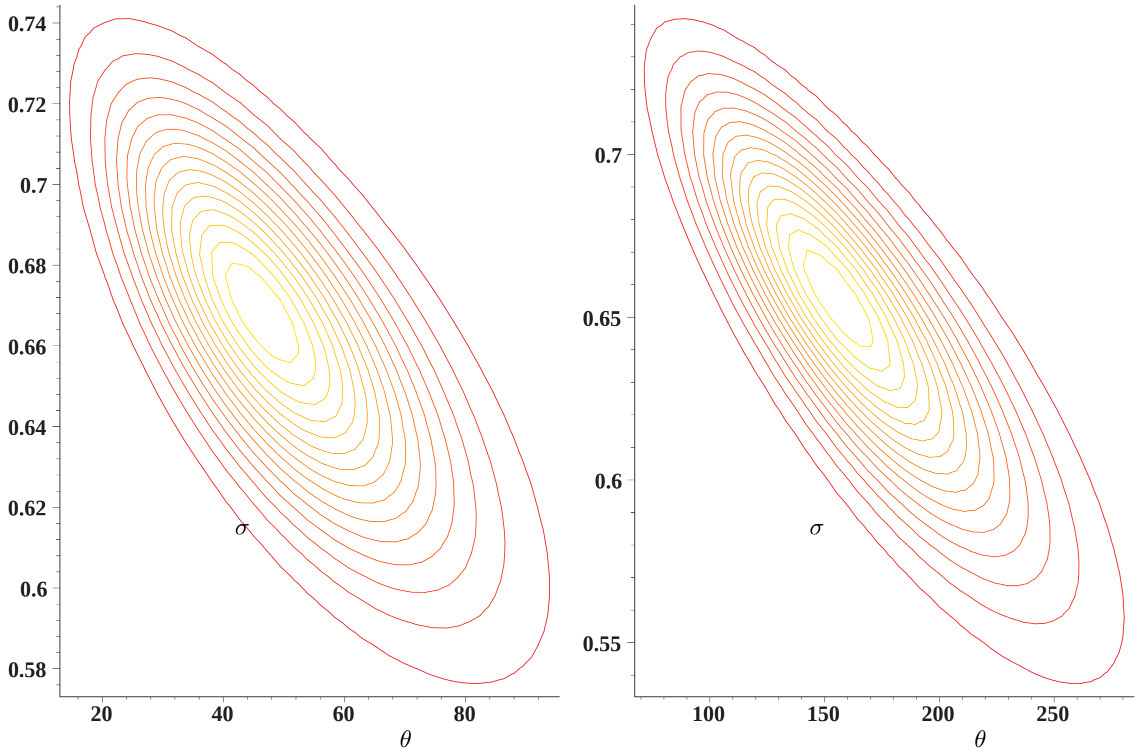

Figure 4: Contour plots of Pitman's sampling formula corresponding to the two *Naegleria gruberi* datasets: aerobic (right) and anaerobic (left)

It is important to remark that our Bayesian nonparametric approach does not incur in problems usually exhibited by frequentist methods. In particular, no ad–hoc adjustments or introduction of parametric components is necessary for predicting future reads if their number is larger than the initial survey. Finally, it is worth remarking that the estimators presented here can be easily adapted to take into account joint data from multiple libraries leading to Bayesian analogs of the estimators set forth in Susko and Roger (2004).

Acknowledgements

The authors are grateful to E. Susko who kindly provided us the EST data analyzed in the paper. A. Lijoi and I. Prünster are partially supported by the Italian Ministry of University and Research, grants 2006/134525 and 2006/133449, respectively. R.H. Mena acknowledges support from CONACYT–México, grant J50160-F.

References

Adams, M., Kelley, J., Gocayne, J., Mark, D., Polymeropoulos, M., Xiao, H., Merril, C., Wu, A., Olde, B., Moreno, R., Kerlavage, A., McCombe, W. and Venter, J. (1991) Complementary DNA

Sequencing: Expressed Sequence Tags and Human Genome Project. *Science*, **252**, 1651–1656.

Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian theory.* Wiley, Chichester.

Emrich, S., Barbazuk, W., Li, L. and Schnable, P. (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* **17**, 69–73.

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

Gyllenberg, M. and Koski, T. (2001) Probabilistic models for bacterial taxonomy *Int. Statist. Review*, **69** 249–276.

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

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.

Hill, B.M. (1979) Posterior moments of the number of species in a finite population and the posterior probability of finding a new species. *J. Amer. Statist. Assoc.* **74** 668–673.

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

Lijoi, A., Mena, R. and Prünster, I. (2007) Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, in press. Available at: `http://web.econ.unito.it/igor/species.pdf`

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

Mao, C.X. (2007) Estimating species accumulation curves and diversity indices. *Statistica Sinica*, in press.

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

Pitman, J. (2006) *Combinatorial Stochastic Processes.* Ecole dEté de Probabilités de Saint-Flour XXXII 2002. Lecture Notes in Mathematics **1875**. Springer, Berlin.

Susko, E. and Roger, A.J. (2004) Estimating and comparing the rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys *Bioinformatics*, **20**, 2279–2287.

Teh, Y.W. (2006) A hierarchical Bayesian language model based on Pitman-Yor processes *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, **44**.

Wang, J.P.Z, Lindsay, B.G., Cui, L., Wall, P.K., Marion, J., Zhang, J., dePamphilis, C.W. (2005) Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries *BMC Bioinformatics*, **6**:300

Wang, J.P.Z, Lindsay, B.G., Leebens-Mack, J., Cui, L., Wall, K., Miller, W.C., dePamphilis, C.W. (2004) EST clustering error evaluation and correction *Bioinformatics*, **20**, 2973–2984.

Zhaohui, S. (2006) Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics*, **22**, 1988–1997.

Appendix

Here we briefly describe how the estimators in (8) and (9) are derived by simplifying the expressions provided in Lijoi et al. (2007). In particular, one finds out that $E_m^{(k,n)} = \sum_{j=0}^m j\, P_m^{(k,n)}(j)$, where the $P_m^{(k,n)}(j)$'s are displayed in (7) and can be deduced from equation (8) in Lijoi et al. (2007). The further simplification yielding the expression of $E_m^{(k,n)}$ in (8) is obtained by observing that $(\theta+1)_{n-1}/(\theta+1)_{n+m-1} = 1/(\theta+n)_m$ and

$$\frac{\prod_{i=k}^{k+j-1}(\theta+i\sigma)}{\sigma^j} = (k+\theta/\sigma)_j$$

As far as the determination of (9), note that

$$\hat{D}_m^{(k,n)} = \sum_{j=0}^{m} P_1^{(k+j,m+n)}(1)\, P_m^{(k,n)}(j)$$

where $P_1^{(k+j,m+n)}(1)$ is the probability of observing a new gene at the $(n+m+1)$–th draw given the in the previous sample, of size $n+m$, there have been detected $k+j$ distinct genes. Hence, by virtue of the prediction structure associated with the two parameter Poisson–Dirichlet process as outlined in Section 3, one has $P_1^{(k+j,m+n)}(1) = (\theta+(k+j)\sigma)/(\theta+n+m)$. From this one deduces

$$\hat{D}_m^{(k,n)} = \sum_{j=0}^{m} \frac{\theta+(k+j)\sigma}{\theta+n+m}\, P_m^{(k,n)}(j) = \frac{\theta+k\sigma}{\theta+n+m}\sum_{j=0}^{m} P_m^{(k,n)}(j) + \frac{\sigma}{\theta+n+m}\sum_{j=0}^{m} j\, P_m^{(k,n)}(j)$$

and one obtains the expression in (9) since $\sum_{j=0}^m P_m^{(k,n)}(j) = 1$.