LLNL-TR-680718

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Technical Report: Algorithm and Implementation for Quasispecies Abundance Inference with Confidence Intervals from Metagenomic Sequence Data

K. McLoughlin

January 11, 2016

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-TR-680718

Technical Report: Algorithm and Implementation for Quasispecies Abundance Inference with Confidence Intervals from Metagenomic Sequence Data

DHS Bioforensics Program

IAA No.: HSHQPM-13-X-00219

Principal Investigator and Correspondent

Kevin McLoughlin

Lawrence Livermore National Laboratory (LLNL), Livermore, CA

925-423-5486, mcloughlin2@llnl.gov

Submission date: July 7, 2014

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

1 Introduction

This report describes the design and implementation of an algorithm for estimating relative microbial abundances, together with confidence limits, using data from metagenomic DNA sequencing. For the background behind this project and a detailed discussion of our modeling approach for metagenomic data, we refer the reader to our earlier technical report, dated March 4, 2014. Briefly, we described a fully Bayesian generative model for paired-end sequence read data, incorporating the effects of the relative abundances, the distribution of sequence fragment lengths, fragment position bias, sequencing errors and variations between the sampled genomes and the nearest reference genomes. A distinctive feature of our modeling approach is the use of a Chinese restaurant process (CRP) to describe the selection of genomes to be sampled, and thus the relative abundances. The CRP component is desirable for fitting abundances to reads that may map ambiguously to multiple targets, because it naturally leads to sparse solutions that select the best representative from each set of nearly equivalent genomes.

Since the model described in our earlier report deals comprehensively with many factors that affect metagenomic DNA sequencing, it is necessarily very complex. For our initial fitting algorithm, we decided to work with a greatly simplified model that focuses on the abundances only. We also began to question whether a full Bayesian approach was needed for the fragment length, position bias and sequencing error components of the model. The rationale for Bayesian modeling of abundances is clear: in order to provide confidence intervals for the abundance estimates, we need to fit a distribution rather than a point estimate. However, maximum likelihood estimates (MLEs) for the length distribution, bias and sequence error parameters are sufficient for our purposes; they are only used internally, so confidence intervals need not be reported. MLE estimation of these parameters will greatly simplify the implementation effort.

2 A simplified model for generating sequence reads

To focus our efforts toward an algorithm for fitting relative abundances, we developed a simplified generative model incorporating some additional assumptions:

1. The data consist of N single-end reads, rather than read pairs. Therefore, we are no longer concerned with fragment length.
2. The true target sequences are all in the reference target set, so we can ignore evolution of the target from the seed sequences.
3. We can ignore position bias when we derive the target sampling probability τ_t.
4. The position bias π_t^p (the probability of starting a read at position p given the target sequence t) is based on a zero-order Markov model, rather than the more complex third-order model

used in eXpress.

5. For now, we ignore sequencing errors.

As in our previous model, the relative abundances ρ_t of the target sequences are determined by a Chinese restaurant process. These are generated by a stick-breaking construction. We draw breakpoints u_t, with $0 \leq u_t \leq 1$, randomly from a $\mathrm{Beta}(\alpha_0, \beta_0)$ distribution, and then generate the abundances by:

$$\rho_t = u_t \prod_{j<t} (1 - u_j) \tag{1}$$

Under assumption (3), the probability τ_t for sampling a read from target t depends only on the relative abundance ρ_t, the target length $l(t)$, and the read length l:

$$\tau_t \propto \rho_t(l(t) - l + 1) = \rho_t \tilde{l}(t)$$

where $\tilde{l}(t)$ is the effective length of t, i.e. the number of positions at which a read could begin.

Under assumption (4), the position bias π_t^p is derived from the probabilities $\Phi_{x,b}$ of base b occurring at offset x relative to the read start position p, in a 21-base window centered at p:

$$\pi_t^p \propto \prod_{x=-10}^{10} \Phi_{x,t[p+x]}$$

where $t[p]$ denotes the base at position p in target t. Both τ_t and π_t^p are normalized so that the probabilities sum to 1 over targets and positions within the target, respectively.

In this simplified model, the variables t_n and p_n for read n, with $n = 1 \ldots N$, are hidden. The observed data are the read sequences r_n. Because sequencing errors and mutations are not included in the model, read sequence r_n is identical to the length l segment of target t_n beginning at position p_n. However, because the reference database includes sets of similar genomes from closely related organisms, and because genomes may contain duplicated elements, a read will frequently match multiple segments from different targets. This ambiguity makes inference of target abundances from read alignment data a nontrivial problem.

We define y_n^{tp} as the probability of observing read sequence r_n, given that it was derived from position p in target t: $y_n^{tp} = \mathbb{P}(r_n | t_n = t, p_n = p)$. For our simplified model where errors and mutations are absent, $y_n^{tp} = 1$ if $r_n = t[p : p + l - 1]$ and 0 otherwise. Here the notation $t[a : b]$ refers to the subsequence of target t between positions a and b inclusive.

Figure 1 shows a graphical representation of the simplified model, showing the conditional dependency relationships between the major random variables and parameters.

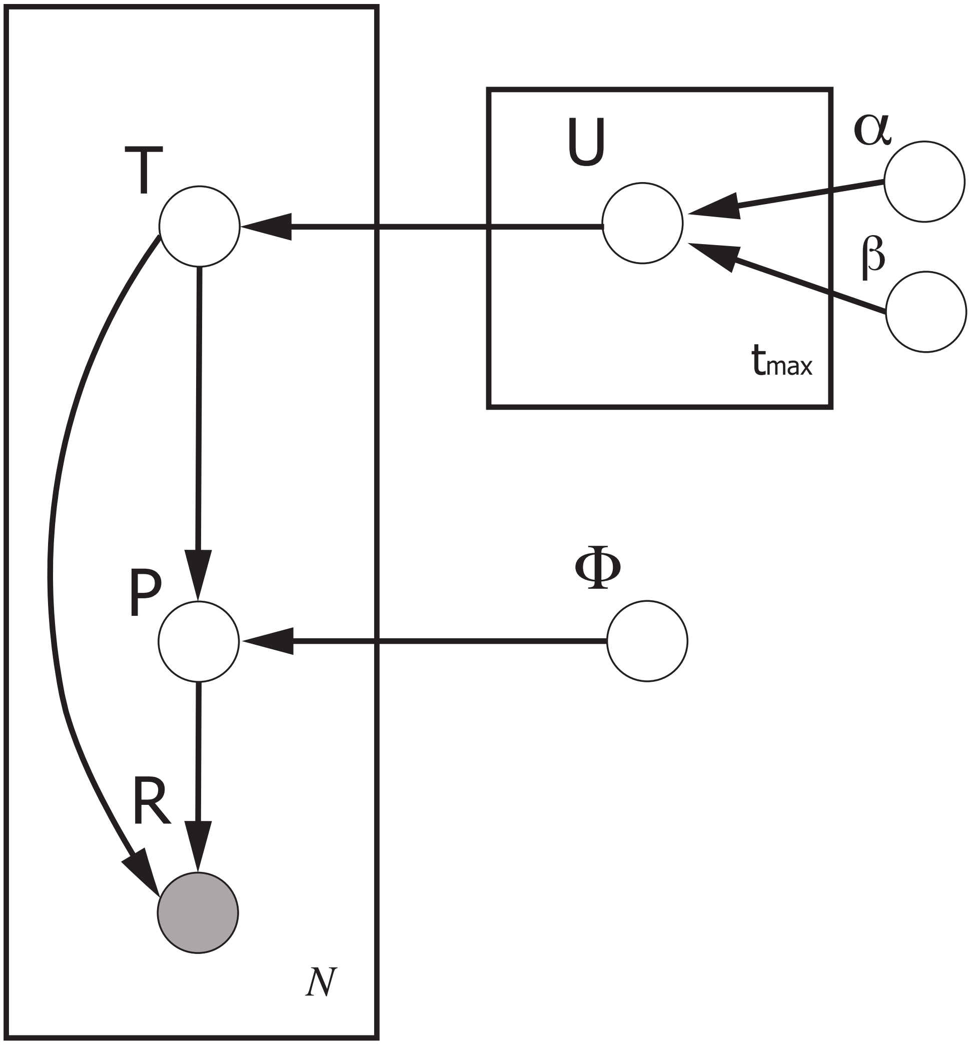

Figure 1: Graphical representation of simplified generative model for metagenomic data

3 Variational Bayes algorithm

Under the simplified model, the complete data likelihood is

$$\begin{aligned}\mathcal{L}(\mathbf{r},\mathbf{p},\mathbf{t},\mathbf{u}|\alpha_0,\beta_0,\Phi) &= \prod_{t=1}^{t_{\max}} \mathbb{P}\left(U_t = u_t|\alpha_0,\beta_0\right) \prod_{n=1}^{N} \mathbb{P}\left(R_n = r_n|T_n = t_n, P_n = p_n\right) \\ &\quad \cdot \mathbb{P}\left(P_n = p_n|T_n = t_n, \boldsymbol{\Phi}\right) \mathbb{P}\left(T_n = t_n|\mathbf{u}\right) \\ &= \prod_{t=1}^{t_{\max}} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} u_t^{\alpha_0-1}(1-u_t)^{\beta_0-1} \prod_{n=1}^{N} y_n^{t_n p_n} \pi_{t_n}^{p_n} \tau_{t_n}\end{aligned}$$

where

$$\begin{aligned}\tau_{t_n} &= \frac{\tilde{l}(t_n)\cdot u_{t_n}\prod_{j<t_n}(1-u_j)}{\sum_{t=1}^{t_{\max}} \tilde{l}(t)\cdot u_t \prod_{k<t}(1-u_k)} \\ \pi_{t_n}^{p_n} &= \frac{\prod_{x=-10}^{10} \Phi_{x,t_n[p_n+x]}}{\sum_{q=1}^{\tilde{l}(t_n)} \prod_{x=-10}^{10} \Phi_{x,t_n[q+x]}}\end{aligned}$$

In the above, normalization constants are included, but may be omitted in the subsequent discussion. The complete data log likelihood is then:

$$\begin{aligned}\log\mathcal{L}(\mathbf{r},\mathbf{p},\mathbf{t},\mathbf{u}|\alpha_0,\beta_0,\Phi) &= \sum_{t=1}^{t_{\max}} (\alpha_0-1)\log u_t + (\beta_0-1)\log(1-u_t) \\ &+ \sum_{n=1}^{N} \left\{ \log u_{t_n} + \sum_{j<t_n} \log(1-u_j) + \log y_n^{t_n p_n} + \log \tilde{l}(t_n) + \log \pi_{t_n}^{p_n} \right\} + \log Z\end{aligned}$$

where Z incorporates all the normalization factors.

To derive the variational Bayes algorithm for this model, we will need to express the complete data log likelihood in the form of an exponential family model. To do this, we want to replace indexing over hidden and observed variables with dot products of parameters and indicator variables. We introduce the following indicator variables:

$t_n^t \equiv 1$ iff $t_n = t$

$p_n^p \equiv 1$ iff $p_n = p$

We then write the terms of the complete data log likelihood using indicator variables:

$$\log \mathcal{L} = \sum_{t=1}^{t_{\max}} (\alpha_0 - 1)\log u_t + (\beta_0 - 1)\log(1-u_t) \\ + \sum_{n=1}^{N}\sum_{t=1}^{t_{\max}} \left\{ t_n^t \left[\log \tilde{l}(t) + \log u_t + \sum_{j<t}(1-u_j) + \sum_{p=1}^{\tilde{l}(t)} p_n^p (\log y_n^{tp} + \log \pi_t^p) \right] \right\} \tag{2}$$

The goal of Bayesian inference is to estimate the posterior distribution of the parameters in the model, given the observed data. For our purposes, the parameters of interest are the stick-breaking proportions u_t, which we can use to estimate the relative abundances of targets in the sample, together with their confidence intervals. The other parameters that may be of use later are the base frequencies $\Phi_{x,b}$ used in the position bias model. However, we do not need to infer posterior distributions for the $\Phi_{x,b}$, since point estimates are sufficient for our purposes. Maximum likelihood estimates for the $\Phi_{x,b}$ can be obtained easily by examining the subset of reads that map to unique (t, p) locations, and computing the base frequencies at each offset relative to the read position p over this read subset. Therefore, we will concern ourselves only with estimating the distributions of u_t.

In variational Bayes inference, we approximate the posterior distribution of the hidden variables and parameters (which is intractable) as a product of independent distributions for each variable. We find a set of such distributions that minimizes the KL divergence between the true and approximated posteriors. In our case, the hidden variables are t_n and p_n for $n = 1 \ldots N$. We write the variational approximation to the posterior as:

$$q(\mathbf{t}, \mathbf{p}, \mathbf{u}) = \prod_{n=1}^{N} q_t(t_n) q_p(p_n|t_n) \prod_{t=1}^{t_{\max}} q_u(u_t)$$

The notation q_t, q_p or q_u for the respective probability mass or density functions is overloaded here; these are actually different functions for the individual t_n, p_n and u_t variables. Each component distribution belongs to the same family as its counterpart in the likelihood, but has different parameters. Thus, t_n and p_n follow categorical distributions, while u_t has a Beta distribution:

$$\begin{aligned} t_n &\sim \mathrm{Cat}(\theta_{\mathbf{n}}) \quad \text{where} \quad \theta_{\mathbf{n}} \equiv \{\theta_n^t, \;\; 1 \le t \le t_{\max}\} \\ p_n|t_n &\sim \mathrm{Cat}(\phi_{\mathbf{nt_n}}) \quad \text{where} \quad \phi_{\mathbf{nt_n}} \equiv \{\phi_{nt_n}^p, \;\; 1 \le p \le \tilde{l}(t_n)\} \\ u_t &\sim \mathrm{Beta}(\alpha_t, \beta_t) \end{aligned}$$

We write the distribution of p_n as conditional on t_n, since its range of values depends on t_n.

The variational Bayes algorithm provides a method for inferring the set of parameters $\{\theta_{\mathbf{n}}, \phi_{\mathbf{nt_n}}, \alpha_t, \beta_t\}$ that minimizes the KL divergence between $q(\mathbf{t}, \mathbf{p}, \mathbf{u})$ and the true posterior. It is an iterative procedure, similar to the EM algorithm, which alternates between the following two steps:

VBE step: For each $n \in \{1 \ldots N\}$, set

$$q_{tp}^{(\text{new})}(t_n, p_n) = \frac{1}{Z_{t_n p_n}} \exp\left\{\mathbb{E}_{q_u}\left[\log \mathbb{P}\left(r_n, t_n, p_n | \mathbf{u}\right)\right]\right\} \tag{3}$$

where $\mathbb{E}_{q_u}\left[f(\mathbf{u})\right]$ represents the expectation of $f(\mathbf{u})$ under the distribution $q_u(\mathbf{u}) = \prod_t q_u(u_t)$ from the previous VBM step, and $Z_{t_n p_n}$ is a normalization factor.

VBM step: For each $t \in \{1 \ldots t_{\max}\}$, set

$$q_u^{(\text{new})}(u_t) = \frac{1}{Z_{u_t}} \mathbb{P}\left(u_t | \alpha_0, \beta_0\right) \exp\left\{\sum_{n=1}^{N} \mathbb{E}_{q_{tp}}\left[\log \mathbb{P}\left(r_n, t_n, p_n | \mathbf{u}\right)\right]\right\} \tag{4}$$

where $\mathbb{E}_{q_{tp}}\left[f(\mathbf{t}, \mathbf{p})\right]$ represents the expectation of $f(\mathbf{t}, \mathbf{p})$ under the distribution $q_{tp}(\mathbf{t}, \mathbf{p}) = \prod_n q_t(t_n) q_p(p_n | t_n)$ from the previous VBE step, $\mathbb{P}\left(u_t | \alpha_0, \beta_0\right)$ is a prior probability density for u_t, and Z_{u_t} is another normalization factor.

It can be shown [1] that, by iterating the VBE and VBM steps, the product distribution $q(\mathbf{t}, \mathbf{p}, \mathbf{u})$ converges to a local optimum with minimal KL divergence from the true posterior distribution. Each step corresponds to an update rule for the parameters of the variational posterior, which we will now derive.

For the VBE step, we rewrite the LHS of equation 3 as a product of categorical distributions:

$$q_{tp}^{(\text{new})}(t_n, p_n) \propto \exp\left\{\sum_{t=1}^{t_{\max}} t_n^t \log \theta_n^{t(\text{new})} + \sum_{t=1}^{t_{\max}} \sum_{p=1}^{\tilde{l}(t)} t_n^t p_n^p \log \phi_{nt}^{p(\text{new})}\right\}$$

Plugging in the relevant terms from equation 2 gives us the following for the RHS:

$$\begin{aligned}
q_{tp}^{(\text{new})}(t_n, p_n) &\propto \exp\left\{\mathbb{E}_{q_u}\left[\sum_{t=1}^{t_{\max}} t_n^t \left(\log \tilde{l}(t) + \log u_t + \sum_{j<t} \log(1 - u_j) + \sum_{p=1}^{\tilde{l}(t)} p_n^p (\log y_n^{tp} + \log \pi_t^p)\right)\right]\right\} \\
&= \exp\left\{\sum_{t=1}^{t_{\max}} t_n^t \left(\log \tilde{l}(t) + \mathbb{E}_{q_u}\left[\log u_t\right] + \sum_{j<t} \mathbb{E}_{q_u}\left[\log(1 - u_j)\right] + \sum_{p=1}^{\tilde{l}(t)} p_n^p (\log y_n^{tp} + \log \pi_t^p)\right)\right\} \\
&= \exp\left\{\sum_{t=1}^{t_{\max}} t_n^t \left(\log \tilde{l}(t) + \psi(\alpha_t) - \psi(\alpha_t + \beta_t) + \sum_{j<t} [\psi(\beta_j) - \psi(\alpha_j + \beta_j)]\right)\right. \\
&\qquad \left. + \sum_{t=1}^{t_{\max}} \sum_{p=1}^{\tilde{l}(t)} t_n^t p_n^p (\log y_n^{tp} + \log \pi_t^p)\right\}
\end{aligned} \tag{5}$$

In equation 5, we use formulas for the expectations of u_t and $(1 - u_t)$ under a Beta(α_t, β_t) distribution; here $\psi()$ is the digamma function. By comparing multipliers of the sufficient statistics t_n^t and $t_n^t p_n^p$, we obtain the update rules for the variational parameters $\theta_n^{t(\text{new})}$ and $\phi_{nt}^{p(\text{new})}$:

$$\begin{aligned}
\theta_n^{t(\text{new})} &\propto \exp\left[\log \tilde{l}(t) + \psi(\alpha_t) - \psi(\alpha_t + \beta_t) + \sum_{j<t} [\psi(\beta_j) - \psi(\alpha_j + \beta_j)]\right] \\
\phi_{nt}^{p(\text{new})} &\propto \exp\left[\log y_n^{tp} + \log \pi_t^p\right] = y_n^{tp} \pi_t^p
\end{aligned} \tag{6}$$

The categorical parameters $\theta_n^{t(\text{new})}$ are constrained by the requirement that $\sum_{t=1}^{t_{\max}} \theta_n^{t(\text{new})} = 1$; therefore we normalize them by dividing the terms in equation 6 by their sum. Note that $q_{tp}^{(\text{new})}(t_n, p_n) = 0$ for any t such that $y_n^{tp} = 0$ for all p, i.e. any target that does not have an alignment from read n. This greatly simplifies calculation of the updates, since we only need to consider targets with alignments from each read. Note also that $\phi_{nt}^{p(\text{new})}$ does not change, once π_t^p and the observed data are given; this is because we did not include π_t^p in the variational Bayes framework.

For the VBM step, we write $q_u^{(\text{new})}(u_t)$ as a Beta distribution:

$$q_u^{(\text{new})}(u_t) \propto \exp[(\alpha_t - 1) \log u_t + (\beta_t - 1) \log(1 - u_t)] \tag{7}$$

To compute the RHS of equation 4, we first collect terms from equation 2 that multiply u_t and $1 - u_t$:

$$\begin{aligned}
\mathbb{P}(t_n | \mathbf{u}) &\propto \sum_{t=1}^{t_{\max}} \left\{ t_n^t \left[\log \tilde{l}(t) + \log u_t + \sum_{j<t} (1 - u_j) \right] \right\} \\
&= \sum_{t=1}^{t_{\max}} \left[t_n^t (\log \tilde{l}(t) + \log u_t) + \sum_{j=t+1}^{t_{\max}} t_n^j \log(1 - u_t) \right]
\end{aligned}$$

Substituting this expression into equation 4 yields:

$$\begin{aligned}
q_u^{(\text{new})}(u_t) \propto\ \mathbb{P}(u_t | \alpha_0, \beta_0) \exp \Bigg\{ \sum_{n=1}^{N} \mathbb{E}_{q_{tp}} &\left[t_n^t (\log \tilde{l}(t) + \log u_t) + \sum_{j=t+1}^{t_{\max}} t_n^j \log(1 - u_t) \right] \\
&+ \mathbb{E}_{q_{tp}} \left[\sum_{p=1}^{\tilde{l}(t)} t_n^t p_n^p (\log y_n^{tp} + \log \pi_t^p) \right] \Bigg\}
\end{aligned}$$

Entering the Beta prior for u_t and rearranging expectations gives us:

$$
\begin{aligned}
q_u^{(\text{new})}(u_t) \propto \exp\{(\alpha_0 - 1)\log u_t + (\beta_0 - 1)\log(1 - u_t)\} \\
\times \exp \sum_{n=1}^{N} \left\{ \mathbb{E}_{q_{tp}}\left[t_n^t\right](\log u_t + \log \tilde{l}(t)) + \sum_{j=t+1}^{t_{\max}} \mathbb{E}_{q_{tp}}\left[t_n^j\right]\log(1 - u_t) \right\} \\
\times \exp \sum_{n=1}^{N} \left\{ \sum_{p=1}^{\tilde{l}(t)} \mathbb{E}_{q_{tp}}\left[t_n^t p_n^p\right](\log y_n^{tp} + \log \pi_t^p) \right\} \\
= \exp\{(\alpha_0 - 1)\log u_t + (\beta_0 - 1)\log(1 - u_t)\} \\
\times \exp \left\{ \sum_{n=1}^{N} \theta_n^{t(\text{new})}(\log u_t + \log \tilde{l}(t)) + \sum_{j=t+1}^{t_{\max}} \sum_{n=1}^{N} \theta_n^{j(\text{new})} \log(1 - u_t) \right\} \\
\times \exp \sum_{p=1}^{\tilde{l}(t)} \sum_{n=1}^{N} \left\{ \theta_n^{t(\text{new})} \phi_{nt}^{p(\text{new})} (\log y_n^{tp} + \log \pi_t^p) \right\} \\
= \frac{1}{Z_{u_t}} \exp \left\{ (\alpha_0 + \sum_{n=1}^{N} \theta_n^{t(\text{new})} - 1)\log u_t + (\beta_0 + \sum_{j=t+1}^{t_{\max}} \sum_{n=1}^{N} \theta_n^{j(\text{new})} - 1)\log(1 - u_t) \right\}
\end{aligned}
\tag{8}
$$

where the normalization factor $\frac{1}{Z_{u_t}}$ captures all the factors in the posterior that *don't* depend on u_t. Here $\theta_n^{t(\text{new})}$ is the estimate of the categorical probability that read n comes from target t computed in the previous VBE step.

Comparing coefficients of $\log u_t$ and $\log(1 - u_t)$ between equations 7 and 8 gives us the following update rules for the VBM step:

$$
\begin{aligned}
\alpha_t^{(\text{new})} &= \alpha_0 + \sum_{n=1}^{N} \theta_n^{t(\text{new})} \\
\beta_t^{(\text{new})} &= \beta_0 + \sum_{j=t+1}^{t_{\max}} \sum_{n=1}^{N} \theta_n^{j(\text{new})}
\end{aligned}
\tag{9}
$$

The update rules defined in equations 6 and 9 are the core of the variational Bayes algorithm. To analyze a metagenomic read data set, we perform the following steps:

1. Align the reads to a database of reference genomes using a tool such as Bowtie 2 [2].
2. Initialize the parameters θ_n^t to assign equal probability to each target t with alignments from read n.

3. Use the VBM update rules (equation 9) to assign values to α_t and β_t for each target.
4. Use the VBE update rules (equation 6 to compute new values for the θ^t_n. Normalize the values so that $\sum_{t=1}^{t_{\max}} \theta^t_n = 1$ for each read n.
5. Continue iterating the VBE and VBM updates until the α_t and β_t values converge, within some specified tolerance.

4 Abundance and confidence limit estimation

The result of the variational Bayes iterations is a set of estimates for the (α_t, β_t) parameters of the Beta distributions describing the breakpoints u_t. To turn these into estimates and confidence intervals for the relative abundances ρ_t, we compute logarithms of both sides of equation 1:

$$\log \rho_t = \log u_t + \sum_{j=1}^{t-1} \log(1 - u_j)$$

Since the u_t are independent draws from their respective distributions, the expectations and variances of their logs are additive. For a variable X distributed as Beta(α, β), the expectations and variances are as follows:

$$\begin{aligned}
\mathbb{E}\left[\log X\right] =&\ \psi(\alpha) - \psi(\alpha + \beta) \\
\mathbb{E}\left[\log(1 - X)\right] =&\ \psi(\beta) - \psi(\alpha + \beta) \\
\text{Var}\left(\log X\right) =&\ \psi'(\alpha) - \psi'(\alpha + \beta) \\
\text{Var}\left(\log(1 - X)\right) =&\ \psi'(\beta) - \psi'(\alpha + \beta)
\end{aligned}$$

where $\psi()$ and $\psi'()$ are the digamma and trigamma functions, respectively. Therefore:

$$\begin{aligned}
\mathbb{E}\left[\log \rho_t\right] =&\ \psi(\alpha_t) - \psi(\alpha_t + \beta_t) + \sum_{j=1}^{t-1} \left[\psi(\beta_j) - \psi(\alpha_j + \beta_j)\right] \\
\text{Var}\left(\log \rho_t\right) =&\ \psi'(\alpha_t) - \psi'(\alpha_t + \beta_t) + \sum_{j=1}^{t-1} \left[\psi'(\beta_j) - \psi'(\alpha_j + \beta_j)\right]
\end{aligned} \tag{10}$$

We can then compute the 95% confidence interval for $\log \rho_t$ by a normal approximation:

$$\begin{aligned}
\log \rho_t^{(\text{lower})} &= \mathbb{E}\left[\log \rho_t\right] - 1.96\sqrt{\text{Var}\left(\log \rho_t\right)} \\
\log \rho_t^{(\text{upper})} &= \mathbb{E}\left[\log \rho_t\right] + 1.96\sqrt{\text{Var}\left(\log \rho_t\right)}
\end{aligned}$$

Note that, when these are exponentiated to form CI bounds for ρ_t, the bounds will be asymmetric about the estimate $\exp(\mathbb{E}\left[\log \rho_t\right])$.

5 Implementation and testing

We implemented the variational Bayes algorithm in Python and tested it with simulated read data. Reads were simulated by selecting genomes from the RefSeq viral database (containing 5,301 targets) according to the Chinese restaurant process described in section 2, using Beta parameters $\alpha_0 = 1, \beta_0 = 5$; these parameters were chosen to create a steep dropoff of abundances with rank, and thus a wide range of relative abundances with a small set of targets. We computed position bias parameters by analyzing base frequencies surrounding read starts in one of the example sequence datasets provided with the eXpress software [3], and used these to compute read position probabilities for each of the targets selected by the CRP. We then sampled 50-mer reads from each selected target, distributed according to the computed probabilities. We generated data sets for testing ranging in size from 10,000 to 1,000,000 reads. We aligned reads to the RefSeq viral DB using Bowtie 2 and estimated abundances, using the algorithm described above.

Figure 2 shows natural logarithms of the abundances fit by the variational Bayes algorithm to a 100,000 read simulated data set, with error bars representing the 95% confidence intervals, plotted against the input abundances used to generate the reads. The confidence intervals are wider for targets with lower abundance, since these are represented by fewer reads. For the more abundant targets, the estimates are very close to the true input values in almost all cases. The one exception is the fourth most abundant target shown in Figure 2. This is a human adenovirus sequence which is 100% identical to another adenovirus genome in RefSeq. Because there was no reason to prefer one sequence over the other, the variational Bayes algorithm assigned equal abundances to both targets, each being half of the true abundance. This example shows that curation of target databases to remove duplicate sequences is essential for obtaining accurate abundance estimates.

Figure 3 shows the convergence trajectories for the targets matched by at least 1,000 reads from the same 100,000 read data set. Trajectories drawn in red correspond to genomes not belonging to the actual target set used to generate reads; rather, these genomes have sequences similar enough to members of the actual target set that many reads map ambiguously to both the actual and the near-neighbor targets. The convergence plot shows that the fitted abundances for the near-neighbor targets decrease with successive iterations; for this particular data set, about 15 iterations are needed to reach convergence. This is exactly the effect we hoped to achieve by using the Chinese restaurant process model; it shows that the variational algorithm converges toward a sparse solution for the relative abundance profile, assigning most of the reads to the best matching targets rather than splitting them between similar targets (as happens with other metagenomic analysis tools). The time to complete each iteration is roughly linear in the number of alignments, requiring about 1 second per 60,000 alignments.

6 Conclusions and next steps

Here we have described our development of a simplified generative model for single-end metagenomic sequence read data, together with the design and implementation of a variational Bayes algorithm to fit relative abundance parameters to data based on this model. Our initial testing shows that the algorithm produces sparse solutions for the relative abundances of the microbial constituents of a metagenomic sample. The next phase of this research will involve further testing of the algorithm, using different settings for the tuning parameters, to see if the accuracy of the fitted abundances can be improved. Once this work is completed, we will integrate the variational Bayes algorithm with existing code in the eXpress software [3] that accounts for positional bias and errors in the read sequence. This will result in a more accurate tool for abundance estimation that works even when reads don't perfectly match a genome in the reference database. The integrated tool will also deal correctly with paired-end read data, which will further reduce the effect of ambiguous read mapping. Finally, we will investigate using an online version of the variational Bayes algorithm, to see if we can improve performance without compromising the accuracy of the fitted abundances.

References

[1] Matthew J Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College London, 2003.

[2] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, April 2012.

[3] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71–73, January 2013.

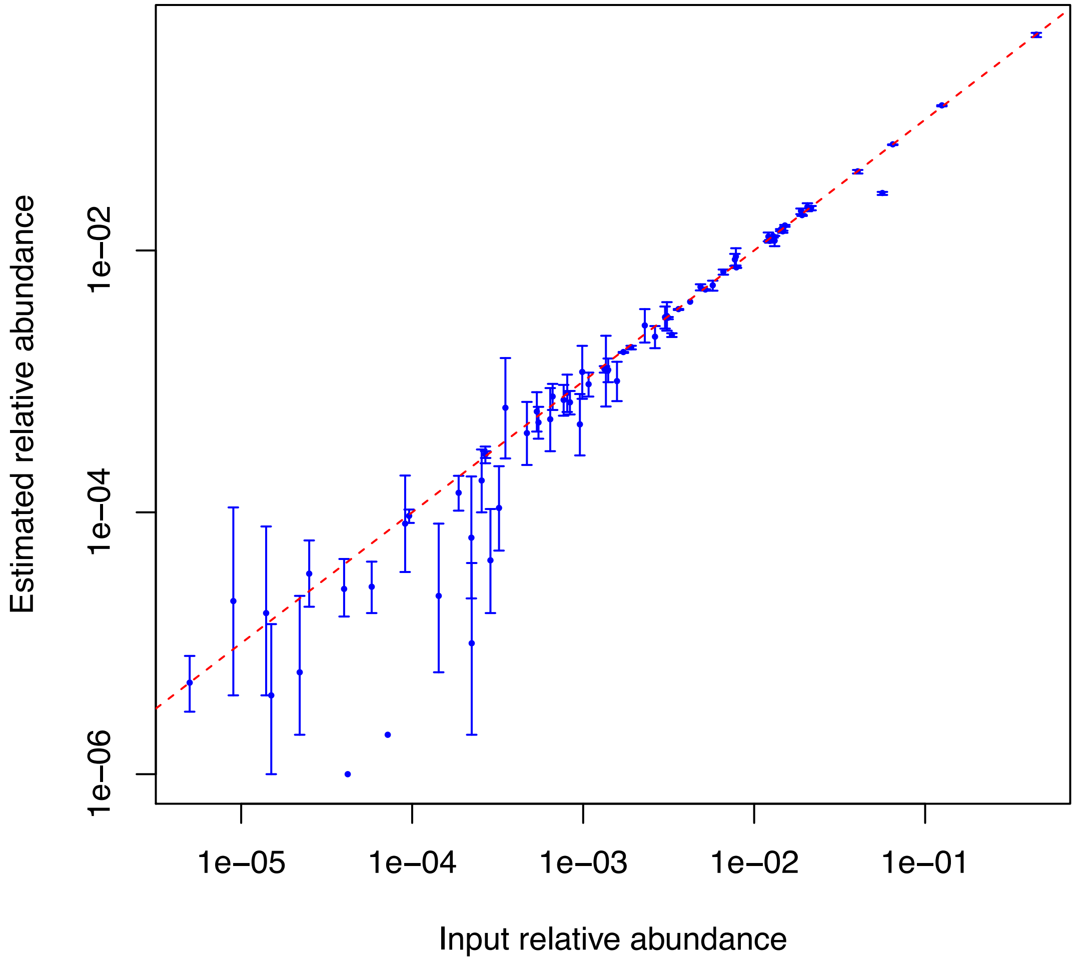

Figure 2: Comparison of relative abundances fit by variational Bayes algorithm to input abundances used to generate simulated data

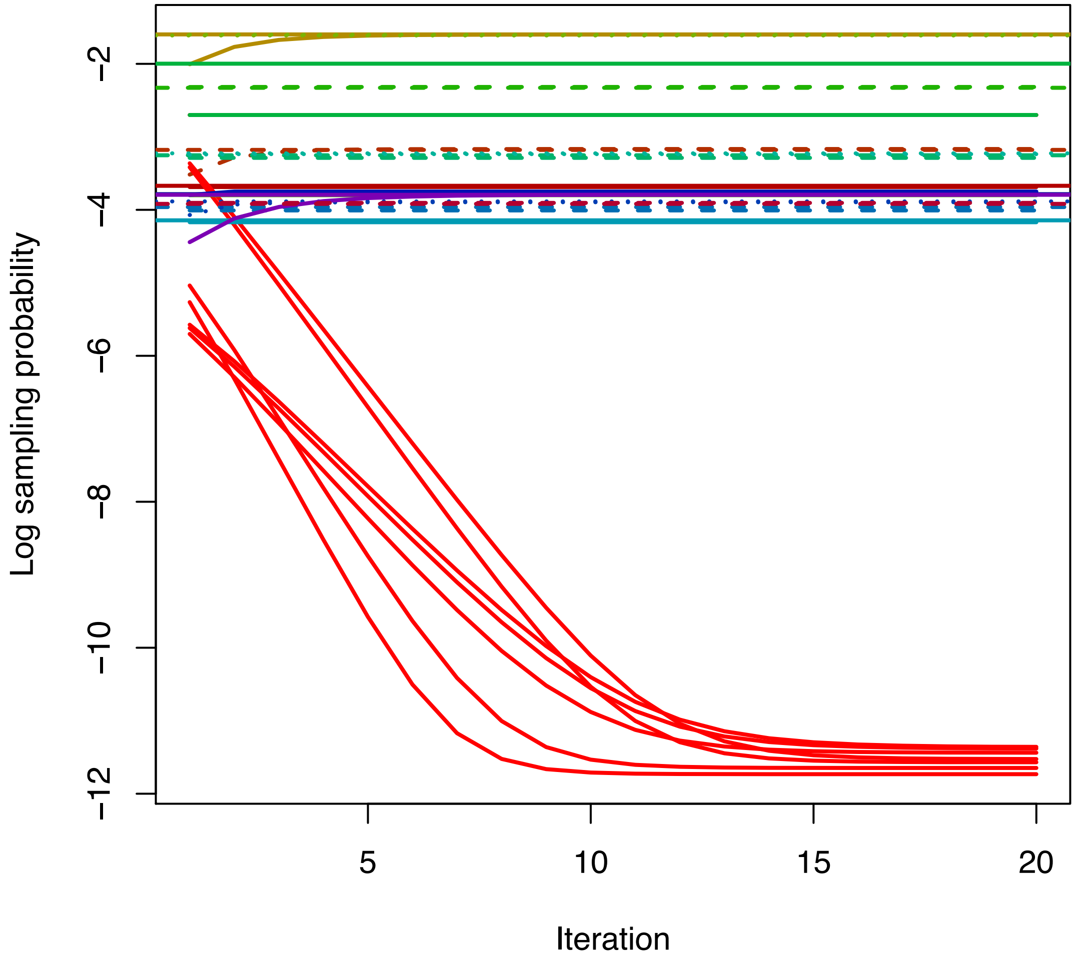

Figure 3: Convergence trajectories of relative abundance estimates over 20 iterations of the variational Bayes algorithm for the targets aligned by at least 1,000 of the 100,000 simulated reads