FERNANDA VITAL DE PAULA

MÉTODOS ESTATÍSTICOS APLICADOS À ANÁLISE DE DADOS DE ETIQUETA DE SEQUÊNCIA EXPRESSA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2011

FERNANDA VITAL DE PAULA

MÉTODOS ESTATÍSTICOS APLICADOS À ANÁLISE DE DADOS DE ETIQUETA DE SEQUÊNCIA EXPRESSA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 11 de fevereiro de 2011

Carlos Souza do Nascimento
(Co-orientador)

Paulo Roberto Cecon

Gérson Rodrigues dos Santos

Sidney Martins Caetano

Fabyano Fonseca e Silva
(Orientador)

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

PAULA, Fernanda Vital de, M.Sc., Universidade Federal de Viçosa, fevereiro de 2011. **Métodos estatísticos aplicados à análise de dados de etiqueta de sequência expressa.** Orientador: Fabyano Fonseca e Silva. Co-orientador: Carlos Souza Nascimento.

Pesquisas de Expressed Sequence Tags (ESTs) são uma ferramenta fundamental para identificação de genes em estudos de seqüenciamento de vários organismos. Dado uma amostra preliminar de EST de uma certa biblioteca de cDNA, vários problemas estatísticos de predição podem surgir. Em particular, é de interesse calcular o número de genes, $\Delta(t)$, que podem ser descobertos em uma amostra futura de EST t vezes maior que a amostra original. Esta e outras estatísticas, apresentadas por Susko e Roger (2004), tais como cobertura e o número de leituras necessárias para se descobrir um novo gene são úteis para direcionar protocolos de sequenciamento por meio do cálculo do grau de redundância de uma biblioteca de cDNA. Este cálculo visa maximizar a obtenção de genes durante um sequenciamento de ESTs, porém, este ainda é visto como um procedimento de custo elevado e adequações de técnicas para redução de tal custo é de fundamental importância. O presente trabalho tem como objetivo apresentar os aspectos teóricos da metodologia proposta por Susko e Roger (2004), implementá-la computacionalmente no software livre R e principalmente propor uma abordagem bayesiana para a estimação de $\Delta(t)$. Toda a metodologia foi aplicada a dois conjuntos de dados: o primeiro diz respeito a duas bibliotecas de cDNA referentes ao organismo *Mastigamoeba Balamuthi* e o segundo a duas bibliotecas de cDNA referentes à pele de bovinos F_2 (Holandês × Gir) infestados pelo carrapato R*iphicephalus (Boophilus) microplus.* Para os dois conjuntos de dados as estimativas por intervalo obtidas para $\Delta(t)$ foram consideravelmente mais precisas quando se utilizou a inferência bayesiana, indicando que a mesma apresenta-se como uma alternativa viável para estudos relacionados ao cálculo da redundância em análises de ESTs.

ABSTRACT

PAULA, Fernanda Vital de, M.Sc., Universidade Federal de Viçosa, February, 2011. **Statistical methods applied to expressed sequence tag data analisys.** Adviser: Fabyano Fonseca e Silva. Co-adviser: Carlos Souza Nascimento.

Expressed sequence tags (ESTs) surveys are a fundamental tools to identify genes in sequencing studies of various organisms. Given a EST preliminary sample from a certain cDNA library, several prediction statistical problems can arise. Particularly, to calculate the number of genes, $\Delta(t)$, which may be discovered in a future EST sample t times larger than the original sample is interesting. This and other ststistics, presented by Susko and Roger (2004), such as coverage and number of necessary readings to discover a new gene are useful for direct sequencing protocols by calculating the degree of redundancy of a cDNA library. This calculation seeks to maximize the obtaining of genes during a EST sequencing, however this is still seen as a costly procedure and adequacy techniques for reducing such costs is of fundamental importance. The present work has as objective to present the theoretical aspects of the methodology proposed by Susko and Roger (2004), to implement computationally the methodology in the free software R and mainly to propose a bayesian approach for estimating $\Delta(t)$. All the methodology was applied to two data sets: the first concerns two cDNA libraries from *Mastigamoeba balamuthi* organism and the second concerns two cDNA libraries from skin of F_2 (Holstein × Gyr) bovine infested with the *ticks Riphicephalus (Boophilus) microplus*. For both data sets the interval estimates obtained for $\Delta(t)$ were significantly more accurate when the Bayesian inference was used, indicating that it is an aviable alternative for studies related to the calculation of the redundancy in analysis of ESTs.

1. INTRODUÇÃO

As pesquisas de Etiqueta de Sequência Expressa (EST) são uma ferramenta poderosa que permite caracterizar rapidamente genes expressos de um determinado organismo, sendo meios eficientes para descoberta de gene em projetos de seqüenciamento do genoma funcional. Em muitos casos, quando há redundância de transcritos altamente expressos, os protocolos experimentais demandam gastos excessivos relacionados com a normalização, a qual pode ser superficialmente caracterizada como procedimentos laboratoriais que envolvem gastos adicionais.

A normalização é planejada para uniformizar a análise de frequências gênicas em bibliotecas de cDNA, de forma que a seleção de um clone aleatório e seu posterior sequenciamento continuem a render sequências expressas que ainda não foram amostradas previamente. Porém, atualmente, há poucos métodos rigorosos disponíveis para avaliar a redundância relativa de várias bibliotecas preparadas do mesmo organismo e/ou para avaliar se protocolos de normalização mostraram-se eficientes. Alguns métodos estatísticos apresentados por Susko e Roger (2004), podem ser usados para estimar e comparar a taxa de descoberta de novos genes em amostras de ESTs agrupadas de diferentes bibliotecas de cDNA. De forma geral, tais métodos são úteis para medir o grau de redundância de uma biblioteca e conduzir a seleção dos números de clones a serem amostrados futuramente de várias bibliotecas de cDNA com o intuito de maximizar a taxa de descoberta de novos genes.

Uma proposta interessante seria uma abordagem bayesiana para tais métodos, uma vez que esta configura-se como um dos principais assuntos da comunidade científica envolvida com o desenvolvimento e aplicação de procedimentos estatísticos. Tal sucesso é decorrente da grande versatilidade de tal abordagem na resolução de problemas nunca antes solucionados por outros métodos. Segundo O'Hagan (1994) outro fato que merece destaque é que na metodologia bayesiana a obtenção de intervalos de credibilidade é imediata, levando em consideração a incerteza existente sobre todos os parâmetros simultaneamente, sendo, portanto, a estimação por intervalo geralmente mais precisa em relação à outras metodologias, que em geral utilizam variâncias assintóticas e aproximadas.

Diante do exposto, o presente trabalho objetiva apresentar detalhadamente os métodos estatísticos propostos por Susko e Roger (2004) relacionados à estimação da redundância em análises de EST, e principalmente propor uma abordagem bayesiana para o número esperado de genes em uma nova amostragem. Objetiva-se também aplicar a referida metodologia em

dois conjuntos de dados de EST, provenientes respectivamente de bibliotecas de cDNA do protista *Mastigamoeba balamuthi* e de bovinos F_2 (1/2 Holandês:1/2 Gir) resistentes e suscetíveis ao carrapato *Riphicephalus (Boophilus) microplus*.

2. REVISÃO DE LITERATURA

2.1.ESTs (Expressed Sequence Tags)

Um dos principais objetivos nos estudos em genômica é a obtenção de uma boa aproximação da relação de genes existentes no genoma do organismo estudado. Um recurso necessário para atender esse objetivo é obter sequências dos clones que representam a expressão do mRNA presente em uma célula ou tecido submetido a determinada circunstância de estudo. Vale ressaltar que o mRNA é a molécula de RNA produzida pela célula, a partir da transcrição do gene contido no DNA, e que será utilizada para produção de proteínas na fase de tradução. Cada uma das sequências obtidas dos clones é chamada Etiqueta de Sequência Expressa (EST – Expressed Sequence Tag) (BAUDET, 2006; NASCIMENTO, 2009).

Desde a sua introdução em Adams et al. (1991), as ESTs têm desempenhado um papel importante na identificação, detecção e caracterização genômica de organismos, o que proporciona uma alternativa atraente e eficiente de sequenciamento. Susko e Roger (2004) relatam que pesquisas de EST não são usadas somente para descoberta de genes, mas são conduzidas frequentemente para avaliar diferenças em expressão de gene em tecidos diferentes ou células expostas a condições diferentes.

O processo de sequenciamento com a utilização de ESTs envolve a produção de bibliotecas de cDNA, a clonagem dos cDNAs com a utilização de vetores, tais como plasmídeos e bacteriófago, e o sequenciamento dos clones através de uma única leitura em uma máquina de sequenciamento, que torna esta técnica de baixo custo, em relação às outras técnicas existentes.

Na Figura 1 observa-se que o primeiro passo na construção da biblioteca de cDNA consiste em isolar a população de mRNAs do tecido ou tipo celular de interesse (I). As moléculas de mRNA (II) são então utilizadas como moldes para a síntese de moléculas de DNA por uma enzima denominada transcriptase reversa. Essas moléculas de DNA transcritas a partir de mRNAs são denominadas cDNAs (III), as quais são ligadas em vetores de clonagem e inseridas em bactérias para multiplicação. Os clones da biblioteca de cDNA são então seqüenciados para produzir as ESTs (IV) (FIGUEIRA, 2006). Os passos II e III, que demandam um maior entendimento de conceitos bioquímicos, são representados de forma mais detalhada na Figura 2.

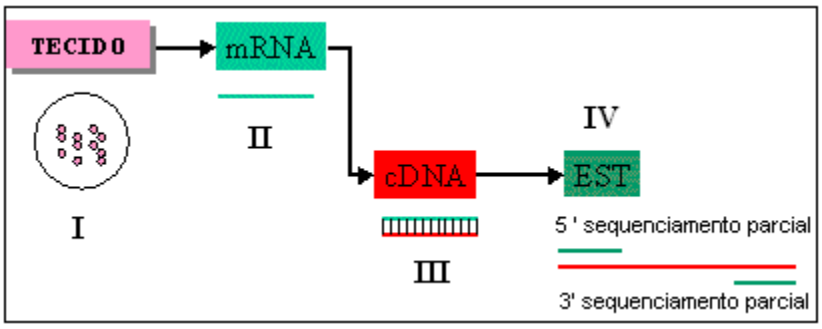

Figura 1. Esquema da construção de ESTs. Fonte: http://binfo.ym.edu.tw/yang/talks/genann/sld007.htm

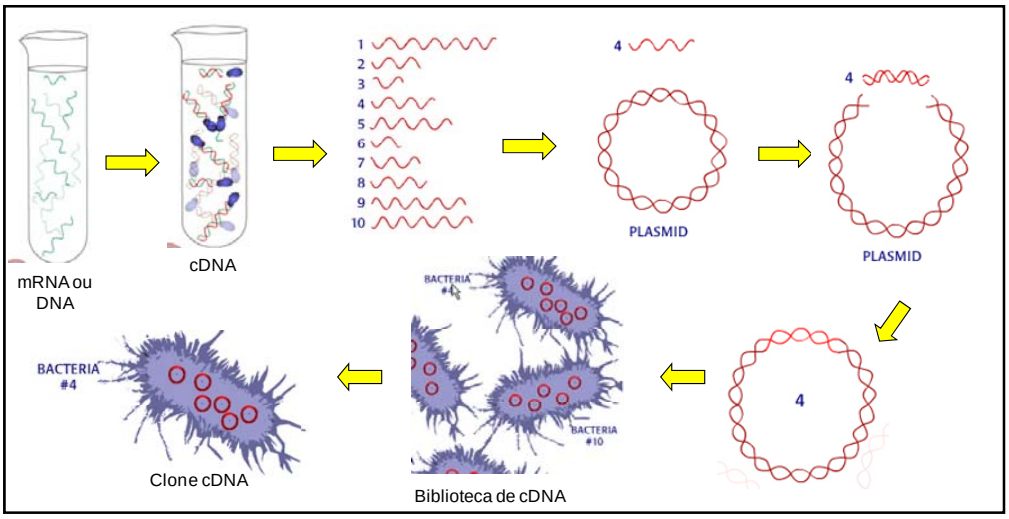

Figura 2. Ilustração da construção de uma biblioteca genômica ou de cDNA: Adaptado de Figueira (2006)

Ao realizar o sequenciamento de cDNAs, as ESTs geram informações sobre os genes que estão sendo expressos no organismo de maneira imediata, porém os genes não são expressos com igual frequência. Existem genes que são expressos continuamente por produzirem proteínas pertencentes a vias metabólicas que regem as reações químicas que ocorrem no organismo, e existem genes que são expressos apenas quando o organismo é submetido à condições especiais, e, além disso, tecidos diferentes expressam genes diferentes. Em função destas características, este tipo de sequenciamento necessita que sejam produzidas bibliotecas de cDNA com origem em diversos tecidos, extraídos sob diferentes condições, tais como, idade, ambiente, presença de doenças, entre outras (BAUDET, 2006).

A alta redundância de determinados transcritos entre as ESTs significa que estes possuem similaridade suficientemente elevada a ponto de se assumir que são derivados de um mesmo gene (COWARD et al., 2002). Essa redundância, segundo Patanjali et al. (1991), pode ser reduzida pelo uso de bibliotecas de cDNA normalizadas, nas quais a frequência de

genes altamente expressos é reduzida por técnicas específicas que envolvem procedimentos laboratoriais de custo elevado. De forma geral, a normalização é uma técnica utilizada para tentar equilibrar o sequenciamento de forma que os genes mais expressos não sejam sequenciados muitas vezes e que os raros possam ser sequenciados um maior número de vezes.

2.2.Redundância em Bibliotecas de cDNA

Em uma biblioteca de cDNA, um gene pode ser representado por várias ESTs, o que frequentemente dificulta o uso efetivo das mesmas. Em função da redundância inerente se faz necessária uma análise especial, dada por agrupamento de ESTs, onde cada grupo representaria um gene, de forma a identificar a real representatividade das sequências. Dessa forma, o objetivo é separar seqüências similares em grupos homogêneos e distintos (FIGUEIRA, 2006).

Uma vez organizadas em grupos, as ESTs podem ser efetivamente exploradas de acordo com os objetivos do estudo, como descoberta de novos genes e mapeamento de genomas (FIGUEIRA, 2006). Assim, informações decorrentes deste processo permitem identificar perfis de sequências únicas existentes e o número de EST dentro de cada grupo, os quais representam a abundância de transcritos em cada biblioteca avaliada.

A referida abundância pode ser analisada de forma alternativa mediante a avaliação da redundância relativa das bibliotecas preparadas a partir do mesmo organismo ou dos mesmos tecidos. Assim, bibliotecas que produzem novos genes a uma taxa mais elevada, são menos redundantes, e ao mesmo tempo, mais abundantes. Tais conceitos estão relacionados com a decisão de se avançar com o sequenciamento de uma biblioteca não normalizada ou de recorrer a procedimentos de normalização, os quais envolvem custos laboratoriais adicionais.

Essa decisão é baseada em estimativas de parâmetros tais como a cobertura dos transcritos (C) e o número de leituras esperadas (E) para se descobrir um novo gene, embora atualmente o parâmetro mais importante tem sido o número esperado de genes ($\Delta(t)$) em uma nova amostragem t vezes maior que a original (SUSKO E ROGER, 2004).

2.3.Metodologia proposta por Susko e Roger (2004)

Dada uma amostra preliminar de ESTs de uma determinada biblioteca, algumas estatísticas podem prever os problemas que podem surgir no que diz respeito ao custo e

eficiência do sequenciamento desta. Em particular, é de grande interesse estimar quantos genes novos podem ser detectados em uma amostra futura de EST de determinado tamanho, o que representa a base para a decisão de se prosseguir com sequenciamento da biblioteca e, em caso de uma decisão positiva, uma diretriz para a seleção do tamanho de uma nova amostra.

Os métodos propostos por Susko e Roger (2004) foram fundamentados em técnicas de estimação de números de espécies (GOOD, 1953), e de forma geral, permitem acessar, por exemplo, a cobertura (C) de uma biblioteca, definida como a proporção de genes únicos na biblioteca representados na amostra de leituras em questão. Com base na estimativa da cobertura, outra medida de biblioteca pode ser utilizada, sendo esta o número de leituras esperadas (E) para se descobrir um novo gene.

Outra estatística é dada pelo número esperado de novos genes ($\Delta(t)$) em uma nova amostragem, sendo talvez a estatística mais importante apresentada por Susko e Roger (2004) no ponto de vista prático, já que fornece ao pesquisador uma estimativa do número de genes que serão descobertos em uma amostra t vezes maior que a amostra original de tamanho n. Para acessar $\Delta(t)$ é preciso ajustar um modelo não-linear, segundo a metodologia proposta pelos autores em questão.

2.4.Modelo Não-Linear

Um modelo é dito não-linear quando ele não é linear em relação aos parâmetros e nem pode ser linearizado por meio de transformações, uma vez que admite uma estrutura de erros aditiva.

Os modelos não lineares podem ser escritos como:

$$y_i = f(x_i, \boldsymbol{\theta}^o) + \varepsilon_i, \quad i = 1, \ldots, n;$$

em que: y_i representa a observação da variável dependente, $f(x_i, \boldsymbol{\theta}^o)$ é a função esperança ou função resposta conhecida, x_i representa a observação da variável independente, $\boldsymbol{\theta}^o = \left[\theta_1^o, \theta_2^o, \ldots, \theta_p^o\right]'$ é um vetor de parâmetros p dimensional desconhecido e ε_i representa o efeito do erro aleatório não observável suposto com distribuição Normal Independentemente e Identicamente Distribuída (NIID) com média zero e variância desconhecida σ^2.

Um exemplo de modelo não linear é dado pelo modelo exponencial, com erros assumidos aditivos, cuja forma é dada por:

$$y_i = \gamma_0 \exp(\gamma_1 x_i) + \varepsilon_i,$$

em que: γ_0 e γ_1 são os parâmetros do modelo; x_i são os valores da variável preditora e ε_i são os termos do erro, independentes, com distribuição normal com média zero e variância σ^2.

Derivando *f* com respeito à γ_0 e γ_1 obtem-se:

$$\frac{\partial f}{\partial \gamma_0} = \exp(\gamma_1 x)$$

$$\frac{\partial f}{\partial \gamma_1} = \gamma_0 x \exp(\gamma_1 x)$$

Como estas derivadas envolvem pelo menos um dos parâmetros, o modelo é reconhecido como não-linear e tal fato implica na utilização de métodos iterativos de estimação. Sob o ponto de vista frequentista, estes procedimentos de estimação muitas vezes apresentam problemas de não-convergência, além de proporcionarem estimativas de erros-padrão aproximadas e/ou fundamentadas em teorias assintóticas, sendo, portanto, questionáveis na presença de um pequeno tamanho de amostra. Metodologias mais robustas, como técnicas de reamostragem (Bootstrap) e principalmente a inferência bayesiana podem ser úteis para evitar tais problemas (Silva, 2006).

2.5.Inferência Bayesiana

A inferência bayesiana consiste em uma abordagem estatística que trata os parâmetros populacionais como variáveis aleatórias e que, a menos de limitações computacionais, permite obter intervalos de confiança (aqui chamados intervalos de credibilidade), fazendo uso de métodos de simulação de Monte Carlo. Ou seja, trata-se de uma alternativa quando, na inferência clássica, a distribuição de amostragem de um estimador for por demais intratável, inviabilizando a construção de intervalos de confiança. Além disso, a inferência bayesiana tem a vantagem adicional de permitir a incorporação de informações passadas *(a priori*), caso existam, enriquecendo o processo de inferência.

Métodos Bayesianos são atualmente um dos principais assuntos da comunidade científica envolvida com o desenvolvimento e aplicação de procedimentos estatísticos, pois estes têm apresentado grande versatilidade na resolução de problemas nunca antes solucionados, principalmente em relação à análise de modelos hierárquicos. A metodologia

bayesiana ficou resguardada durante um grande período por necessitar de resoluções matemáticas, mais precisamente de integrações, inviáveis de serem feitas algebricamente.

Por volta da década de 60 ela ressurgiu em alguns trabalhos teóricos como o de Jeffreys (1961), mas somente em 1990 com o trabalho de Gelfan e Smith (1990), os quais utilizaram algoritmos da classe MCMC (Markov Chain Monte Carlo) é que o problema da resolução de integrais foi solucionado de uma maneira alternativa, uma vez que os resultados foram excelentes, o que atraiu a atenção de muitos outros pesquisadores.

Em recentes estudos envolvendo análise Séries Temporais e de modelos de regressão não linear (Barreto e Andrade, 2004; Silva et al., 2005 e Silva et al. 2006) a metodologia bayesiana foi utilizada com sucesso, pois sua característica de considerar todos os parâmetros como variáveis aleatórias, segundo esses autores, reduziu substancialmente o número de estimativas viesadas. Além disso, relatam também que a utilização desta metodologia requer um número menor de observações, pois os conceitos probabilísticos envolvidos diminuem a dependência do ajuste do modelo em relação ao número de dados utilizados, uma vez que o conceito de graus de liberdade não é considerado.

Segundo O'Hagan (1994) outro fato que merece destaque é que na metodologia bayesiana a obtenção de intervalos de credibilidade é imediata, levando em consideração a incerteza existente sobre todos os parâmetros simultaneamente, sendo, portanto, a estimação por intervalo geralmente mais precisa em relação àquela apresentada pela metodologia freqüentista, que em geral utiliza variâncias assintóticas e aproximadas.

A inferência bayesiana consiste de uma informação a priori, dos dados amostrais e do cálculo da densidade a posteriori dos parâmetros. A informação a priori é dada pela densidade de probabilidade $P(\theta)$, a qual expressa o conhecimento do pesquisador sobre os parâmetros a serem estimados. Quando em determinado estudo o pesquisador tem pouca ou nenhuma informação para incorporar à priori considera-se uma não-informativa, por exemplo, a priori de Jeffreys (JEFFREYS, 1961). Os dados $Y = \{y_1, y_2, \dots, y_n\}$, representados por uma amostra aleatória de uma população com densidade f, são utilizados na análise bayesiana através da função de verossimilhança $L(y_1, \dots, y_n|\theta\}$, que é a densidade conjunta destes dados.

Portanto, a partir do momento que se opta por uma distribuição a priori, seja ela informativa ou não, e obtém-se a função de verossimilhança, é possível, por meio do Teorema de Bayes representado pela expressão seguinte, obter a distribuição densidade a posteriori de θ, de forma que qualquer conclusão a seu respeito é realizada a partir desta distribuição,

$$P(\theta|Y_n) = \frac{P(Y_n|\theta)P(\theta)}{\int L(Y_n|\theta)P(\theta)d\theta}$$

sendo $Y_n = \{y_1, y_2, \ldots, y_n\}$,. O denominador, chamado de constante de integração, não depende de θ, portanto temos:

$$P(\theta|Y_n) \propto L(Y_n|\theta)P(\theta)$$

ou seja, a expressão acima pode ser entendida como: Posteriori ∝ Verossimilhança x Priori, em que ∝ representa proporcionalidade.

Segundo Broemiling (1989), pode-se pensar no Teorema de Bayes como um mecanismo de atualização da opinião do estatístico sobre θ, portanto este teorema constitui a base da inferência Bayesiana, pois toda prática inferencial é realizada a partir da distribuição a posteriori obtida.

A distribuição a posteriori de um parâmetro contém toda a informação probabilística a respeito do mesmo. Assim, toda a inferência sobre o parâmetro é realizada por meio desta distribuição. Segundo Rosa (1998) para se inferir em relação a qualquer elemento de θ, a distribuição a posteriori conjunta dos parâmetros, P(θ|Y), deve ser integrada em relação a todos os outros elementos que a constituem. Assim, se o interesse do pesquisador se concentra sobre determinado conjunto de θ, por exemplo, θ_1, tem-se a necessidade da obtenção da distribuição P(θ_1|Y), denominada de marginal, a qual é dada por:

$$P(\theta_1|Y) = \int_{\theta \neq \theta_1} P(\theta|Y)d\theta_{\theta \neq \theta_1}$$

A integração da distribuição conjunta a posteriori para a obtenção das marginais geralmente não é analítica, necessitando de algoritmos iterativos especializados como o Gibbs Sampler e o Metropolis-Hastings, os quais são denominados de algoritmos MCMC. Portanto, para a utilização desses algoritmos, é necessário que se obtenha a partir da distribuição a posteriori um conjunto de distribuições chamadas de distribuições condicionais completas.

Recentemente, nota-se um interesse crescente em métodos estatísticos Bayesianos aplicados a diversas áreas da ciência, como em Epidemiologia, Bioestatística, Engenharia, e outras. De acordo com Asai (2005), desde a segunda metade da década de noventa o maior desenvolvimento desta metodologia, especialmente no que se diz respeito aos algoritmos relacionados com o método de Monte Carlo via cadeias de Markov (MCMC), tem sido observado na área de Séries Temporais aplicadas principalmente a área de Econometria e Finanças. Isto se deve ao fato desses algoritmos fornecerem resultados mais rápidos e

confiáveis direcionados com solução numérica de expressões complexas resultantes do tratamento estatístico de modelos utilizados nessas áreas.

2.6.Algoritmos MCMC

Os métodos MCMC são uma alternativa aos métodos não iterativos tendo em vista a resolução de integrais complexas. A idéia é obter uma amostra das distribuições marginais a posteriori dos parâmetros de interesse por meio de um processo iterativo. Segundo Gamerman (1997) uma cadeia de Markov é um processo estocástico em que a probabilidade de estar em um certo estado em um tempo futuro pode depender do estado atual do sistema, mas não dos estados em tempos passados, ou seja, os valores gerados no processo apresentam uma dependência. Por sua vez, estes valores gerados são considerados amostras aleatórias de uma determinada distribuição de probabilidade, caracterizando assim o método de simulação Monte Carlo. Dessa forma tem-se uma ação conjunta desses dois métodos que resulta no método MCMC, cujos principais algoritmos são o Metropolis-Hastings e o Gibbs Sampler.

O algoritmo de Metropolis-Hastings é utilizado para a obtenção da distribuição marginal a posteriori quando o amostrador de Gibbs não se mostra eficiente, ou seja, para parâmetros cuja distribuição condicional não se caracteriza como uma distribuição de probabilidade conhecida. Neste caso, geram-se valores do parâmetro a partir de uma distribuição proposta e esse é aceito ou não com uma certa probabilidade de aceitação (CHIB E GREENBERG, 1995).

Segundo Berg (2004) para descrever o algoritmo Metropolis-Hastings, suponha que a distribuição de interesse é a distribuição a posteriori, $P(\boldsymbol{\theta}|Y)$, sendo $\boldsymbol{\theta}$ um vetor de parâmetros a ser estimado, $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_K)$, k=1, 2,..., K. Considere também que todas as condicionais completas a posteriori, $P(\theta_k|\theta_{-k}, Y)$, estejam disponíveis, mas não se sabe gerar amostras diretamente de cada uma, e que amostras de um novo valor de θ_k, serão geradas a partir de uma distribuição proposta, ou candidata, $C(\theta_k)$. Este valor gerado, denominado inicialmente de θ_k^c, constituirá, ou não, uma amostra da distribuição marginal a posteriori de θ_k conforme determinação de um critério probabilístico imposto pelo pesquisador.

Os valores dos parâmetros gerados por esses algoritmos, após verificação de q iterações, q=1,2,...,Q, são utilizados para formar uma amostra aleatória, $\left\{\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_K^{(1)}, \theta_1^{(2)}, \theta_2^{(2)}, ..., \theta_K^{(2)}, ..., \theta_1^{(q)}, \theta_2^{(q)}, ..., \theta_K^{(q)}\right\}$. À medida que o número de iterações

aumenta, o conjunto de valores gerados aproxima de sua condição de equilíbrio. Assim assume-se que a convergência é atingida em uma iteração cuja distribuição esteja arbitrariamente próxima da distribuição de equilíbrio, ou seja, da distribuição marginal desejada (Sorensen e Gianola, 2004).

O algoritmo de Metropolis-Hastings é bastante geral, e pode, pelo menos em princípio, ser implementado com qualquer distribuição condicional completa a posteriori e para qualquer proposta. Entretanto sob o ponto de vista prático, a escolha da proposta é crucial para o bom desenvolvimento do algoritmo, ou seja, para sua convergência para a distribuição marginal a posteriori (Chib e Greenberg, 1995). No Quadro abaixo é apresentado um esquema ilustrativo do algoritmo Metropolis-Hastings.

I - Inicialize $\boldsymbol{\theta}^{(0)} = \theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_K^{(0)}$;

II - Obtenha um novo valor para $\theta_k^{(q)}$ por meio do seguinte critério:

Gere um valor proposto, θ_k^c, para θ_k : $\theta_k^c \sim C(\theta_k \mid \theta_{-k}^{(q-1)})$

Aceite θ_k^c com probabilidade dada por:

$$\alpha = \min\left\{1, \frac{P\left(\theta_k^{(q)} \middle| \theta_{-k}, Y\right) \Rightarrow valor\ gerado: condicional\ completa\ de\ \theta_k\ via\ \theta_k^c}{P\left(\theta_k^{(q-1)} \middle| \theta_{-k}, Y\right) \Rightarrow valor\ aceito\ para\ \theta_k\ na\ iteração\ anterior\ (q-1)}\right\}$$

onde $\theta_{-i}^{(a)} = (\theta_1^{(k)}, ..., \theta_{i-1}^{(k)}, \theta_{i+1}^{(k-1)}, ..., \theta_S^{(k-1)})$.

Se α > u, u~U[0,1], aceita-se o valor, caso contrário permanece o valor obtido na iteração anterior.

III - Faça $q = q + 1$ até atingir o número estipulado (Q) de iterações.

Um problema no algoritmo Matropolis-Hastings é que a taxa de aceitação pode ser muito baixa, isto é, teremos que gerar muitos valores da distribuição candidata até conseguir um número suficiente de valores da posteriori. Isto ocorrerá se as informações da priori e da verossimilhança forem conflitantes já que neste caso os valores gerados terão baixa probabilidade de serem aceitos (Ehlers, 2007).

O algoritmo Gibbs Sampler é, essencialmente, um esquema iterativo de amostragem de uma cadeia de Markov, cujo núcleo de transição é formado pelas distribuições

condicionais completas. É uma técnica para gerar variáveis aleatórias de uma distribuição marginal quando se conhece a sua densidade (Gamerman, 1997). À medida que o número q de iterações aumenta, a seqüência de valores gerados se aproxima da distribuição de equilíbrio, ou seja, da densidade marginal desejada para cada parâmetro, quando se assume que a convergência foi atingida.

Para descrever o algoritmo, suponha que a distribuição de interesse seja uma distribuição a posteriori $P(\boldsymbol{\theta}|Y)$, com $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_K)$ e considere também que todas as condicionais completas a posteriori $P(\theta_k \mid \theta_{-k}, Y)$, k=1,2,...,K, estejam disponíveis e que sabe-se gerar amostras de cada uma delas (Casella e George, 1992). No Quadro a seguir é apresentado o esquema ilustrativo do algoritmo Gibbs Sampler.

I - Inicialize $\boldsymbol{\theta}^{(0)} = \theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_K^{(0)}$;

II - Obtenha um novo valor para $\theta_k^{(q)}$ por meio do seguinte critério:

i) Gere um valor para $\theta_k^{(q)}$ diretamente de sua condicional completa:

$$\theta_k^{(q)} \sim P\left(\theta_k \left| \theta_1^{(q-1)}, ..., \theta_{k+1}^{(q-1)}, ..., \theta_K^{(q-1)}, Y \right.\right)$$

III - Faça $q = q + 1$ até atingir o número estipulado (Q) de iterações.

2.7.Critérios de Convergência

Os métodos de MCMC são uma ótima ferramenta para resolução de muitos problemas práticos na análise Bayesiana. Porém, algumas questões relacionadas à convergência nestes métodos ainda merecem bastante pesquisa, como por exemplo, o número de iterações que deve ter o processo de simulação para garantir que a cadeia convirja para o estado de equilíbrio.

Os algoritmos Metropolis-Hastings e o amostrador de Gibbs são processos iterativos, e as cadeias resultantes desses métodos necessitam ter sua convergência constatada. Para avaliação da convergência desses algoritmos opta-se pelos testes de diagnóstico de GEWEKE (1992), GELMAN E RUBIN (1992), RAFTERY-LEWIS (1992B) e o de HEIDELBERG E

WELCH (1983). Todos esses testes encontram-se disponíveis no pacote BOA (Bayesian Output Analisys) do software livre R (R Development Core Team, 2010).

2.7.1. Critério de Raftery e Lewis

Ao se analisar a convergência de uma seqüência gerada por meio do amostrador de Gibbs, é comum descartar as primeiras iterações, em geral, de 40% a 50% do total, considerando-se que essa primeira parte esteja sendo influenciada pelos valores iniciais. Este início da cadeia é chamado de período de "aquecimento" ou burn-in.

Outro aspecto importante refere-se á dependência entre as observações subseqüentes da cadeia. Para se obter uma amostra independente, as observações devem ser espaçadas por um determinado número de iterações, ou seja, considerar saltos (thin) de tamanho k, usando, para compor a amostra, os valores a cada k iterações.

O critério de Raftery e Lewis fornece estimativas do número de iterações necessárias para se obter a convergência, do número de iterações iniciais que devem ser descartadas (burn-in) e da distância mínima (k) de uma iteração à outra para se obter uma amostra independente. Esses valores são calculados mediante especificações para garantir que um quantil u de uma determinada função seja estimado com uma precisão pré-definida.

2.7.2. Critério de Geweke

Usando técnicas de análise espectral, o critério de Geweke fornece um diagnóstico para a ausência de convergência.

Este propõe o diagnóstico de convergência para Cadeias de Markov baseados no teste de igualdade de médias da primeira e última parte da cadeia de Markov (geralmente dos primeiros 10% e dos últimos 50%).

De acordo com toda a teoria apresentada no item 2.5, nota-se que a Inferência Bayesiana está fundamentada em distribuições de probabilidade, as quais são assumidas para os dados (função de verossimilhança) e para os parâmetros (prioris).

Devido ao fato da análise de ESTs gerar observações referentes a contagens (frequências) de sucessos (descoberta de genes) em uma sequência de avaliações (análise das bibliotecas), a distribuição Binomial Negativa apresenta-se como uma alternativa viável para descrever tais observações amostrais.

2.8. Distribuição Binomial Negativa

Essa família de distribuições de probabilidade é uma generalização da distribuição geométrica, ou seja, a geométrica é um caso particular da binomial negativa, uma vez que a soma de variáveis com distribuição geométrica, independentes e identicamente distribuídas, possui distribuição binomial negativa.

Para caracterizar tal distribuição, considere uma sequência de ensaios independentes de Bernoulli com probabilidade de sucesso constante p para cada ensaio. Seja a variável X definida como sendo igual ao número de fracassos requeridos para que ocorra o r-ésimo sucesso, então, X se distribui de acordo com a binomial negativa. Percebe-se que ocorreram x fracassos e r-1 sucessos antes do r-ésimo sucesso no último ensaio. Cada configuração de x fracassos, com probabilidade 1 – p, com r – 1 sucessos, com probabilidade p cada, tem probabilidade igual à $(1\text{-}p)^x\ p^{r-1}$. O número de combinações possíveis para organizar x fracassos em um total de x + r – 1 configurações é dado por:

$$\binom{x+r-1}{x} = \binom{x+r-1}{r-1}$$

Assim, a função de probabilidade da distribuição binomial negativa pode ser obtida multiplicando-se esse número de combinações das configurações possíveis de fracassos e sucessos nos x + r – 1 ensaios anteriores ao r-ésimo sucesso pela probabilidade de cada configuração, $(1\text{-}p)^x\ p^{r-1}$, e ainda, pela probabilidade p do r-ésimo sucesso, obtido independentemente no último ensaio. A função de probabilidade assim obtida é dada por:

$$P(X = x) = \binom{x+r-1}{x} p^r (1-p)^x$$

em que x = 0, 1, 2, 3, ... representa o número de falhas até ocorrência do r-ésimo sucesso, r = 1, 2, ... e $0 < p \leq 1$. No âmbito de estudos de EST, x representa o número de ESTs amostrados que não resultaram na obtenção de um gene.

A média e a variância da distribuição binomial negativa são:

$$\mu_X = \frac{r(1-p)}{p} \ e \ \ \sigma_X^2 = \frac{r(1-p)}{p^2}$$

Vale ressaltar que a variável aleatória binomial é uma contagem do número de sucessos em n tentativas de Bernoulli, ou seja, o número de tentativas é predeterminado e número de sucessos é aleatório. Uma variável aleatória binomial negativa é uma contagem do número de tentativas requeridas para obter r sucessos, isto é, o número de sucessos é predeterminado e o número de tentativas é aleatório. Nesse sentido, uma variável aleatória

binomial negativa pode ser considerada o oposto, ou o negativo, de uma variável aleatória binomial.

Como exemplo, considere que tomando uma amostra de ESTs de uma biblioteca de cDNA, estamos interessados em saber qual a probabilidade de que, com o vigésimo EST amostrado se obtenha o quinto gene, sabendo que a probabilidade de encontrar um gene em uma amostra de ESTs em tal biblioteca é de 0,4.

Fazendo a variável aleatória X denotar o número de genes encontrados até o vigésimo EST amostrado, tem-se então que X tem uma distribuição binomial negativa com r = 5 (5 sucessos ou 5 genes) e x = 20 – 5 = 15 (15 fracassos)

Seja p, a probabilidade de sucesso, ou seja, a probabilidade de encontrar um gene na amostra selecionada e q, a probabilidade de fracasso, ou seja, a probabilidade de não encontrar um gene na amostra. No caso p = 0,4 e q = 1 – 0,4 = 0,6. Segue-se,

$$P(X = 15) = \binom{15 + 5 - 1}{15} \times 0{,}4^5 0{,}6^{15} = 0{,}44788 = 44{,}788\%$$

Então, a probabilidade de se obter o quinto gene (sucesso) no vigésimo EST amostrado é de 0, 44788 ou 44, 788%.

3. METODOLOGIA

3.1. Dados

Para ilustrar os métodos propostos por Susko e Roger (2004) e testar a metodologia bayesiana proposta, serão utilizados dois conjuntos de dados.

3.1.1. Conjuntos de dados referentes ao protista *Mastigamoeba balamuthi*

Primeiramente será utilizado os dados descritos na Tabela 1 apresentados por Susko e Roger (2004), referentes à duas bibliotecas de cDNA.

Nesse caso, ESTs foram obtidas por seleção aleatória e sequenciamento de clones de uma biblioteca não-normalizada e uma normalizada de cDNA do protista *Mastigamoeba balamuthi*. A biblioteca normalizada foi preparada a partir da biblioteca não-normalizada e, portanto, a biblioteca não-normalizada contém todos os genes na biblioteca normalizada (mas não vice-versa). Depois que as ESTs foram obtidas, as sequências foram agrupadas em grupos de sequências que apresentavam regiões de similaridade entre si, utilizando o programa de clusterização CAP3 (Contig Assembling Program 3) (Huang & Madan, 1999).

Tabela 1. Números de sequências n_x que foram lidas x vezes nas bibliotecas normalizadas e não-normalizadas do protista Mastigamoeba e as respectivas probabilidades associadas à proporção de cada grupo de sequências lidas x vezes.

	Não-Normalizada		**Normalizada**	
x	**n_x**	**P(x)**	**n_x**	**P(x)**
1	378	0,529	200	0,551
2	33	0,092	21	0,116
3	21	0,088	14	0,116
4	9	0,050	4	0,044
5	6	0,042	3	0,041
6	1	0,008	3	0,050
7	3	0,029	1	0,019
8	1	0,011	0	0
9	1	0,013	1	0,025
10	1	0,014	0	0
13	1	0,018	0	0
14	0	0	1	0,039
15	5	0,105	0	0
Total	715	1	363	1

3.1.2. Conjunto de dados referentes a bovinos F_2 (Holândes x Gir)

O carrapato representa grande problema para a bovinocultura, seja ela de corte ou leite, pois além do aumento do custo de produção com a aplicação de carrapaticidas, os mesmos podem proporcionar resíduos químicos nos produtos de origem animal e poluição ambiental. Uma solução viável para este problema é selecionar animais geneticamente resistentes ao ataque de carrapatos. Sabe-se que raças zebuínas (*Bos indicus*) são mais resistentes que raças européias (*Bos taurus*), pois o gado zebuíno, proveniente da Índia, tem convivido há milhares de anos com o carrapato *Riphicephalus B. microplus*, ocorrendo provavelmente uma eliminação natural dos animais mais sensíveis, permitindo assim maiores oportunidades reprodutivas para os animais geneticamente resistentes (Lemos et al., 1986). Atualmente, com o advento de modernas técnicas biotecnológicas, um dos principais objetivos dos estudos em genômica bovina é identificar genes envolvidos na resistência/susceptibilidade dos bovinos ao carrapato e caracterizar o padrão de expressão dos mesmos que corresponde a eventos fisiológicos importantes relacionados com a produção e a saúde dos animais. Tais estudos é um dos fatores mais promissores para reduzir as perdas de produção e diminuir o custo de controle desse parasita na pecuária bovina.

Devido à importância de estudos de resistência ao carrapato para a bovinocultura brasileira e à necessidade de se aumentar o número de ESTs relacionado à resistência/susceptibilidade dos bovinos, a metodologia também será aplicada a conjuntos de dados provenientes de amostras de ESTs obtidas de bibliotecas de cDNA não-normalizada geradas a partir de 2 pools de tecidos de pele proveniente de animais F_2 (1/2Holandês:1/2 Gir) avaliados como resistentes (RES) e susceptíveis (SUS) infestados com carrapato *Riphicephalus (Boophilus) microplus*. As ESTs foram agrupadas e montadas em contigs (grupos de similaridade) pelo programa CAP3 (Contig Assembling Program 3) (Huang e Madan, 1999). Os dados em questão são apresentados na Tabela 2.

Tabela 2. Números de sequências n_x que foram lidas *x* vezes nas bibliotecas suscetíveis e resistentes não-normalizadas referentes aos bovinos F_2 (Holandês x Gir) e as respectivas probabilidades associadas à proporção de cada grupo de sequências lidas *x* vezes.

	SUS		**RES**	
x	**n_x**	**P(x)**	**n_x**	**P(x)**
1	819	0,375	619	0,340
2	108	0,099	135	0,148
3	21	0,029	29	0,048
4	19	0,035	17	0,037
5	8	0,018	5	0,014
6	7	0,019	6	0,020
7	3	0,010	5	0,019
8	3	0,011	0	0
9	2	0,008	0	0
10	2	0,009	1	0,005
11	1	0,005	2	0,012
13	1	0,006	0	0
31	0	0	1	0,017
Total	2182	1	1822	1

3.2. Estatísticas propostas por Susko e Roger (2004)

3.2.1. Cobertura (C)

Para se obter a estimativa aproximadamente imparcial de cobertura que é frequentemente usada, inicialmente obtém-se o número n_1 de agrupamentos de EST que consistem de uma única sequência, sendo n o número total de sequências. A estimativa da cobertura é calculada como um menos a proporção de genes que apareceram em um agrupamento consistindo de uma única sequência. Esta é dada por:

$$\hat{C} = \frac{n - n_1}{n} = 1 - \frac{n_1}{n}$$

Em outras palavras, a cobertura é a proporção de agrupamentos compostos por duas ou mais sequências. Assim, como $\hat{C}$ e n são diretamente proporcionais, se $\hat{C}$ é um valor alto, n também o é, sendo menos improvável a descoberta de um gene em um novo sequenciamento. Além disso, para que $\hat{C}$ seja igual a zero o quociente n_1/n deve ser igual a um, o que significa, como se trata de um quociente, que o numerador deve ser igual ao numerador, ou seja, n_1 deve ser igual a n. Uma interpretação prática para tal igualdade é que o número de sequências lidas apenas uma vez deve ser igual ao número total de sequências da amostra. Se tal fato ocorrer em uma determinada amostra de ESTs, qualquer amostragem futura de sequências resultaria em uma taxa de descoberta de genes de 100% ou não haveria indícios de

redundância na biblioteca da qual foi retirada a amostra em questão. Assim conclui-se que para o pesquisador é interessante que a cobertura dê um valor próximo de zero, o que representa menor redundância na biblioteca e consequentemente maior número de genes encontrados na amostra de EST.

Como a cobertura atual de uma biblioteca não é diretamente observável, a mesma está sujeita a erros de amostragens, portanto a comparação entre duas ou mais bibliotecas, requer que os erros padrão sejam levados em conta.

Se o número de genes, N, na biblioteca e o número de genes amostrados são altos, então geralmente a diferença $C - \hat{C}$ tende a se distribuir normalmente com média zero e erro-padrão:

$$se(\hat{C}) = n^{-1/2}[(n_1/n) + (2n_2/n) - (n_1/n)^2]^{1/2}$$

em que n é o número total de sequências, n_1 é o número de agrupamentos que consistem em uma única sequência e n_2 é o número de agrupamentos que consistem em duas seqüências. Dessa forma, um intervalo de confiança assintótico $(1-\propto) \times 100\%$ é assim determinado por $\hat{C} \pm z_{\propto/2} se\ (\hat{C})$.

A cobertura também pode ser usada no cálculo de outras estatísticas como o intuito de facilitar a interpretação e possibilitar um maior entendimento a respeito da dinâmica do sequenciamento.

3.2.2. Número de leituras esperadas para descobrir um novo gene (E)

O número esperado de leituras para descobrir um gene novo é estimado como:

$$\hat{E} = 1/(1 - \hat{C}),$$

sendo $\hat{C}$ a cobertura estimada conforme 3.2.1.

O erro padrão para cobertura pode ser convertido a um erro padrão para o número esperado de leituras necessário na descoberta de um novo gene pela transformação:

$$se(\hat{E}) = se(\hat{C})/(1 - \hat{C})$$

Embora, as estatística C e E sejam importantes para estudos de redundância em bibliotecas de cDNA, os mesmos não permitem acessar informações a respeito da viabilidade do sequenciamento em amostrar futuras. Para tanto, Susko e Roger (2004), propuseram uma nova estatística denominada número esperado de novos genes $\Delta(t)$.

3.2.3. Número esperado de genes ($\Delta(t)$)

Para acessar $\Delta(t)$ em uma nova amostra t vezes maior que a amostra original de tamanho n, a seguinte equação é recomendada:

$$\Delta(t) = \eta_1 \alpha^{-1} \gamma^{-1} \{1 - (1 + \gamma t)^{-\alpha}\},$$

em que t é o valor que define o tamanho de uma nova amostra, α e γ são os parâmetros da distribuição binomial negativa e o termo η_1 é dado por $P(x = 1)nt$, ou seja, η_1 é a proporção de genes que foram lidos apenas uma vez.

O uso de tal distribuição diz respeito ao fato da mesma possibilitar o cálculo da probabilidade de um gene selecionado aleatoriamente aparecer x vezes em uma nova amostra de tamanho nt, sendo n o tamanho da amostra original. Dessa forma, a distribuição em questão pode ser ajustada aos dados amostrais como aqueles apresentados nas Tabelas 1 e 2, considerando-se a seguinte expressão:

$$P(x) \propto \frac{\Gamma(x+\alpha)}{x!\Gamma(1+\alpha)} \gamma^{x-1}, x = 1, \ldots$$

O modelo probabilístico descrito acima, devido a sua não linearidade em relação aos parâmetros, foi ajustado aos dados das Tabela 1 e 2 por meio do método dos quadrados mínimos generalizados para modelos de regressão não linear via função nls do software R (Development Core Team, 2009).

Em resumo, as estimativas dos parâmetros da distribuição binomial negativa, $\hat{\alpha}$ e $\hat{\gamma}$, juntamente como o valor de η_1 irão compor a fórmula do número esperado de novos genes ($\Delta(t)$), a fim de proporcionar o seu estimador: $\hat{\Delta}(t) = \eta_1 \hat{\alpha}^{-1} \hat{\gamma}^{-1} \{1 - (1 + \hat{\gamma} t)^{-\hat{\alpha}}\}$.

De acordo com Susko e Roger (2004), o estimador da variância de $\hat{\Delta}(t)$ pode ser derivada assintoticamente, o qual é dado por:

$$\hat{V}[\hat{\Delta}(t)] = n^{-1} \sum_{x \geq 1} t^{2x} \eta_x - n^{-1} \sum_{x \geq 1} \eta_x (-1)^x [1 - 2(1 + t)^x + (1 + 2t)^x]$$

em que: x é o número de vezes que as sequências apareceram na amostra e N o número total de sequências na biblioteca. Por meio deste estimador da variância é possível construir intervalos de confiança assintoticamente normais para $\Delta(t)$.

3.3. Proposta Bayesiana

Uma proposta bayesiana diz respeito à obtenção de uma distribuição a posteriori para Δ(t), sendo α e γ estimados via aplicação do teorema de Bayes, através da relação de proporcionalidade:

$$P(\alpha, \gamma|x) \propto P(x|\alpha, \gamma)P(\alpha)P(\gamma),$$

onde P(α) é a distribuição a priori de α e P(γ) é a distribuição a priori de γ independentemente dos dados *x*, P(x|α, γ) é a distribuição dos dados amostrais dado os parâmetros α e γ (Função de Verossimilhança) e P(α, γ|x) é a distribuição dos parâmetros considerando os dados *x* (Distribuição a Posteriori).

3.3.1. Distribuição Binomial Negativa

Susko e Roger (2004) utilizaram o modelo binomial negativo truncado de Fisher et al. (1943) para descrever os dados dispostos na Tabela 1.

$$P(x) \propto \frac{\Gamma(x+\alpha)}{x!\,\Gamma(\alpha+1)}\gamma^{x-1}, x = 1,2, \ldots$$

Tal modelo é obtido da função de probabilidade da distribuição Binomial Negativa, e essa obtenção é descrita a seguir. Vale lembrar que a função de probabilidade da distribuição Binomial Negativa com parâmetros α e p, onde $0 \leq p \leq 1$ é dada (Casella e Berger, 1990) por:

$$P(x) = \binom{\alpha + x - 1}{x} p^{\alpha}(1-p)^{x} \; ou \; P(x) = \frac{(\alpha + x - 1)!}{x!\,(\alpha - 1)!} p^{\alpha}(1-p)^{x}, x = 0,1, \ldots$$

Aplicando a propriedade $\Gamma(n+1) = n!\,, n \geq 0$ nos fatoriais $(\alpha + x - 1)!$ E $(\alpha - 1)!$ Em P(x), obtém-se:

$$P(x) = \frac{\Gamma(\alpha + x)}{\mathrm{x}!\,\Gamma(\alpha)} p^{\alpha}(1-p)^{x}.$$

Pela propriedade $\Gamma(n+1) = n\,\Gamma(n), \quad n \geq 0$ da função gama, decorre que $\Gamma(n) = \frac{\Gamma(n+1)}{n}$. Tal propriedade utilizada em $\Gamma(\alpha)$ implica em:

$$P(x) = \frac{\Gamma(\alpha + x)}{x! \frac{\Gamma(\alpha + 1)}{\alpha}} p^{\alpha}(1-p)^{x}.$$

Multiplicando o quociente do denominador invertido pelo numerador obtém-se:

$$P(x) = \frac{\Gamma(\alpha + x)}{x!\,\Gamma(\alpha + 1)} \alpha p^{\alpha}(1-p)^{x}.$$

Através da propriedade $b^{x} = b^{x-1}.b$, para um valor b qualquer, podemos substituir $(1-p)^{x}$ por $(1-p)^{x-1}\,(1-p)$, obtendo:

$$P(x) = \frac{\Gamma(\mathrm{x} + \alpha)}{\mathrm{x}!\,\Gamma(\alpha + 1)} \alpha p^{\alpha}(1-p)^{x-1}(1-p).$$

Agora, considerando $1 - \mathrm{p} = \gamma$ e consequentemente $\mathrm{p} = 1 - \gamma$, tem-se:

$$P(x) = \frac{\Gamma(x + \alpha)}{x!\,\Gamma(\alpha + 1)} \alpha(1-\gamma)^{\alpha}\gamma^{x-1}\,\gamma,$$

portanto:

$$P(x) = \frac{\Gamma(x + \alpha)}{x!\,\Gamma(\alpha + 1)} (\alpha(1-\gamma)^{\alpha}\gamma)\gamma^{x-1} \propto \frac{\Gamma(x + \alpha)}{x!\,\Gamma(\alpha + 1)} \gamma^{x-1}, \alpha > -1 \text{ e } 0 \leq \gamma \leq 1.$$

A proporcionalidade inserida deve-se ao fato do termo $(\alpha(1-\gamma)^{\alpha}\gamma)$ ser uma constante.

3.3.2. Função de Verossimilhança

A distribuição conjunta dos dados amostrais dado os parâmetros da distribuição assumida para os dados é denominada Função de Verossimilhança. Foi visto no item anterior que a Distribuição Binomial Negativa foi adotada para descrever os dados. Assim, considerando uma amostra aleatória $x_1, x_2, \ldots, x_n$, independência entre tais observações e assumindo que $x_i \sim BinNeg(\alpha, \gamma)$, a Função de Verossimilhança é dada por:

$$P(x|\alpha,\gamma) = \prod_{i=1}^{N} P(x_i) = \prod_{i=1}^{N} \frac{\Gamma(x_i + \alpha)}{x_i!\,\Gamma(\alpha + 1)} (\alpha(1-\gamma)^{\alpha}\gamma)\gamma^{x_i - 1}.$$

Pela utilização da propriedade $\prod_{i=1}^{N} x_i y_i = \prod_{i=1}^{N} x_i \prod_{i=1}^{N} y_i$, tem-se que:

$$P(x|\alpha,\gamma) = \prod_{i=1}^{N}\left[\frac{\Gamma(x_i+\alpha)}{x_i!\,\Gamma(\alpha+1)}\right]\prod_{i=1}^{N}(\alpha(1-\gamma)^{\alpha}\gamma)\gamma^{x_i-1},$$

portanto o termo $\prod_{i=1}^{N}(\alpha(1-\gamma)^{\alpha}\gamma)\gamma^{x_i-1}$ pode ser reescrito da seguinte maneira:

$$\prod_{i=1}^{N}(\alpha(1-\gamma)^{\alpha}\gamma)\gamma^{x_i-1} = \prod_{i=1}^{N}(\alpha(1-\gamma)^{\alpha})\gamma^{x_i} = \alpha(1-\gamma)^{\alpha}\gamma^{x_1}\ldots\alpha(1-\gamma)^{\alpha}\gamma^{x_N}.$$

Utilizando as propriedades da multiplicação de N fatores, $(1-\gamma)^{\alpha}(1-\gamma)^{\alpha}\ldots(1-\gamma)\alpha = (1-\gamma)N\alpha$ e $\gamma x1\gamma x2\ldots\gamma xn = \gamma x1+x2+\ldots+xn = \gamma xn$. Dessa forma, obtém-se:

$$\prod_{i=1}^{N}(\alpha(1-\gamma)^{\alpha})\gamma^{x_i} = \alpha(1-\gamma)^{\alpha}\gamma^{x_1}\ldots\alpha(1-\gamma)^{\alpha}\gamma^{x_N} = \alpha^{N}(1-\gamma)^{N\alpha}\gamma^{\sum x_i}.$$

Portanto,

$$P(x|\alpha,\gamma) = \prod_{i=1}^{N}\left[\frac{\Gamma(x_i+\alpha)}{x_i!\,\Gamma(\alpha+1)}\right]\alpha^{N}(1-\gamma)^{N\alpha}\gamma^{\sum x_i},$$

é a Função de Verossimilhança assumida no estudo em questão.

3.3.3. Distribuições a Priori para γ e α

Sob o enfoque Bayesiano, assumiu-se que os parâmetros γ e α têm distribuições de probabilidade a priori Beta com parâmetros a e b $\big(B(a,b)\big)$ e Beta Reparametrizada $(B(c,d)I(-1,1))$ no intervalo [-1,1] com parâmetros c e d, respectivamente, isto é,

$$\gamma \sim B(a,b) \text{ e } \alpha \sim B(c,d)I(-1,1)$$

com as constantes a, b, c e d conhecidas. Tais constantes são denominadas hiperparâmetros.

As funções de distribuição de probabilidade correspondentes às prioris utilizadas seguem abaixo com suas respectivas médias e variâncias. As Figuras 3 e 4 ilustram tais distribuições e mostram os respectivos valores dos hiperparâmetros utilizados para cada biblioteca (não-normalizada e normalizada).

Os hiperparâmetros em questão foram obtidos de acordo com análises prévias realizadas por Susko e Roger (2004) para os dados referentes ao protista *Mastigamoeba*

balamuthi e por Nascimento (2009) para os dados referentes aos bovinos F_2 (1/2 Holandês:1/2 Gir).

$$P(\gamma) = \frac{1}{B(a,b)}\gamma^{a-1}(1-\gamma)^{b-1} \propto \gamma^{a-1}(1-\gamma)^{b-1}, 0 \leq \gamma \leq 1, a > 0, b > 0,$$

com $E(\gamma) = \frac{a}{a+b}$ e $V(\gamma) = \frac{ab}{(a+b)^2(a+b+1)}$.

$$P(\alpha) = \frac{1}{B(c,d)}(\alpha-1)^{c-1}(1-\alpha)^{d-1} \propto (\alpha-1)^{c-1}(1-\alpha)^{d-1},$$

com $E(\alpha) = \frac{c-d}{c+d}$ e $V(\alpha) = \frac{4cd}{(c+d)^2(c+d+1)}$.

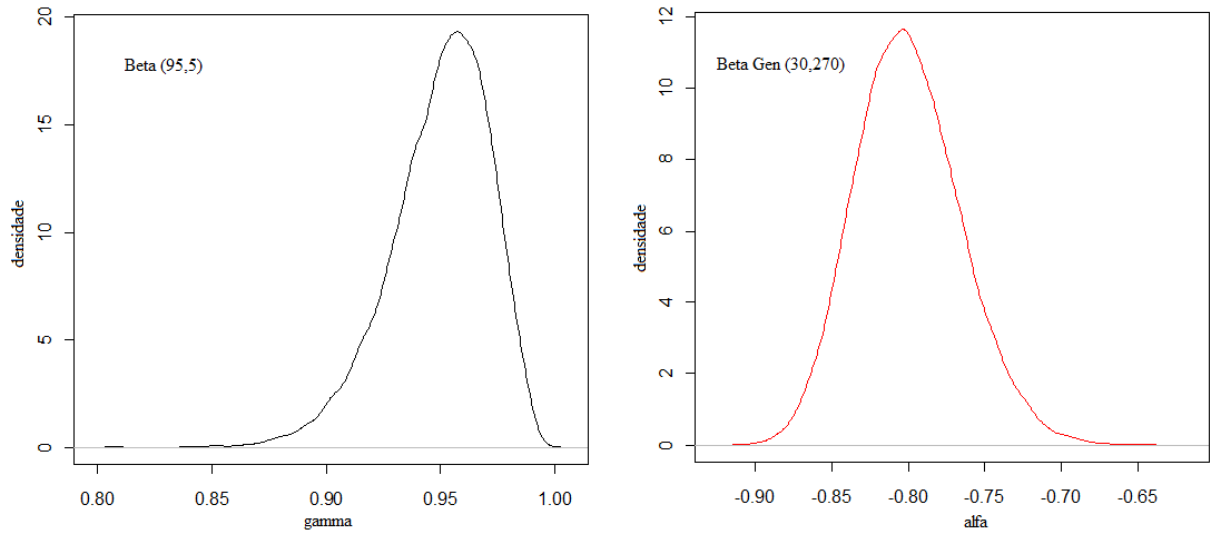

Figura 3. Prioris utilizadas para gamma e alfa, respectivamente considerando a biblioteca não-normalizada.

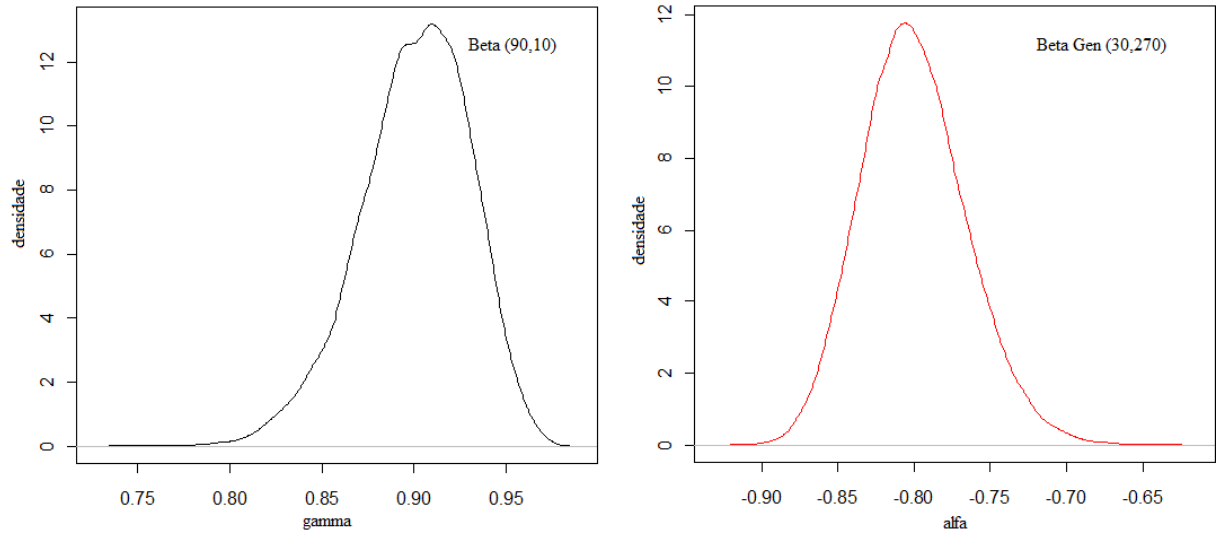

Figura 4. Prioris utilizadas para gamma e alfa, respectivamente, considerando a biblioteca normalizada.

3.3.4. Distribuição conjunta à posteriori e distribuições condicionais completas à posteriori (DCCP)

Para obter as distribuições condicionais completas a posteriori, simplesmente assumimos que ao condicionar a posteriori conjunta em relação a um parâmetro, o mesmo passa a ser considerado constante, e como já está explicito no termo de proporcionalidade, este pode ser desconsiderado. Dessa forma, a expressão resultante pode ser denominada de condicional completa a posteriori do outro parâmetro, ou seja, aquele para a qual a condicional completa não foi condicionada.

Pelo Teorema de Bayes, a distribuição conjunta à posteriori ($P(\gamma, \alpha|x)$) será obtida pelo produto da função de verossimilhança com as prioris, isto é,

$$P(\gamma, \alpha|x) \propto P(x|\alpha, \gamma)P(\gamma)P(\alpha)$$

Assim,

$$P(\gamma, \alpha|x) \propto \prod_{i=1}^{N} \left[\frac{\Gamma(x_i + \alpha)}{x_i!\,\Gamma(\alpha + 1)}\right] \alpha^N (1-\gamma)^{N\alpha} \gamma^{\Sigma x_i} \gamma^{a-1} (1-\gamma)^{b-1} (\alpha - 1)^{c-1} (1-\alpha)^{d-1}.$$

Agrupando as potências com bases iguais $(1-\gamma)^{N\alpha}$ e $(1-\gamma)^{b-1}$, $\gamma^{\Sigma x_i}$ e γ^{a-1}, por meio da propriedade $a^x a^y = a^{x+y}$, sendo a um valor qualquer, obém-se:

$$P(\gamma,\alpha|x) \propto \prod_{i=1}^{N}\left[\frac{\Gamma(x_i+\alpha)}{x_i!\,\Gamma(\alpha+1)}\right]\alpha^N(\alpha-1)^{c-1}(1-\alpha)^{d-1}\,\gamma^{\sum x_i+a-1}(1-\gamma)^{N\alpha+b-1}.$$

3.3.4.1. DCCP para γ

Para obter a distribuição condicional completa a posteriori para γ, condicionamos a posteriori conjunta em relação ao parâmetro α, que passa a ser considerado constante, podendo eliminar, por uma questão de proporcionalidade, os termos da posteriori conjunta que dependem apenas do parâmetro α, os quais são enfatizados abaixo:

$$P(\gamma,\alpha|x) \propto \prod_{i=1}^{N}\underbrace{\left[\frac{\Gamma(x_i+\alpha)}{x_i!\,\Gamma(\alpha+1)}\right]\alpha^N(\alpha-1)^{c-1}(1-\alpha)^{d-1}}\,\gamma^{\sum x_i+a-1}(1-\gamma)^{N\alpha+b-1}$$

Dessa forma, a condicional completa a posteriori de γ é denominada por:

$$P(\gamma|x,\alpha) \propto \gamma^{\sum x_i+a-1}(1-\gamma)^{N\alpha+b-1}\,,$$

cujo aspecto permite notar que tal expressão apresenta função de distribuição de probabilidade (f.d.p) de uma distribuição Beta, cujos parâmetros são designados por:

$$a^* = \sum x_i + a \;e\; b^* = N\alpha + b$$

Dessa forma, temos uma condicional completa representada por uma distribuição de probabilidade conhecida, isto é, $\gamma|x,\alpha \sim B(a^*, b^*)$. Este fato implica na utilização do algoritmo Gibbs Sampler para gerar amostras da distribuição marginal a posteriori para γ.

3.3.4.2. DCCP para α

Seguindo a mesma idéia para obtenção da condicional completa para α, pode-se eliminar os termos da posteriori conjunta que dependem apenas do parâmetro γ, indicados abaixo:

$$P(\gamma,\alpha|x) \propto \prod_{i=1}^{N}\left[\frac{\Gamma(x_i+\alpha)}{x_i!\,\Gamma(\alpha+1)}\right]\alpha^N(\alpha-1)^{c-1}(1-\alpha)^{d-1}\,\underbrace{\gamma^{\sum x_i+a-1}(1-\gamma)^{N\alpha+b-1}}$$

Assim, a condicional completa a posteriori de α é denominada por:

$$P(\alpha|x,\gamma) \propto \prod_{i=1}^{N}\left[\frac{\Gamma(x_i+\alpha)}{x_i!\,\Gamma(\alpha+1)}\right]\alpha^N(\alpha-1)^{c-1}(1-\alpha)^{d-1},$$

podendo ser reescrita, devido ao termo $(\alpha-1)^{c-1}(1-\alpha)^{d-1}$ representar o núcleo da f.d.p de uma distribuição Beta Reparametrizada no intervalo [-1,1] com parâmetros c e d, da seguinte forma:

$$P(\alpha|x,\gamma) \propto \prod_{i=1}^{N}\left(\frac{\Gamma(x_i+\alpha)}{x_i!\,\Gamma(\alpha+1)}\right)\alpha^N B(c,d)\mathrm{I}(-1,1)$$

Nota-se que tal expressão não se caracteriza como uma distribuição de probabilidade conhecida, fato este que exige a aplicação do algoritmo Metropolis- Hastings na geração de valores de α.

3.3.5. Implementação dos algoritmos MCMC

Ao se considerar algoritmos MCMC, a cada iteração i, tem-se as estimativas $\alpha^{(i)}$ e $\gamma^{(i)}$, portanto as amostras que irão compor a distribuição a posteriori do número esperado de genes ($\Delta(t)$) serão obtidas indiretamente por: $\Delta(t)^{(i)} = \eta_1 {\alpha^{(i)}}^{-1} {\gamma^{(i)}}^{-1}\{1-\left(1+\gamma^{(i)}t\right)^{-\alpha^{(i)}}\}$.

Como visto, a distribuição condicional completa para o parâmetro γ é dada por uma distribuição Beta, apresentando uma forma conhecida e sendo, portanto, passível ao uso do algoritmo Gibbs Sampler.

O mesmo não acontece para a distribuição condicional do parâmetro α a qual não apresenta uma forma definida, devendo-se então utilizar, nesta situação, o algoritmo Metropolis-Hastings. Assim, os valores gerados para alfa são os valores candidatos a serem aceitos ou não pelo algoritmo Metropolis-Hastings e a variância de alfa interfere diretamente na porcentagem de aceitação dos valores propostos para tal parâmetro. Portanto, a escolha da distribuição candidata foi dada por $2 * B(c,d) - 1$, a fim de que a taxa de aceitação dos valores candidatos permaneça entre 17 e 45%, conforme recomendação de Blasco et al. (2003).

Os algoritmos Gibbs Sampler e Metropolis-Hastings foram implementados matricialmente no software estatístico R (R Development Core Team, 2010). Considerou-se, na aplicação ao conjunto de dados referentes ao protista *Mastigamoeba balamuthi,* uma cadeia de 20.000 iterações, das quais 5.000 foram eliminadas (burn-in) e foram considerados

os primeiros de cada cinco elementos dessa cadeia (thin). Na aplicação ao conjunto de dados referentes aos bovinos F_2 (holandês x gir), foi considerada uma cadeia de 20.000 iterações, das quais 5.000 foram eliminadas (burn-in) e foram considerados os primeiros de cada cinco elementos dessa cadeia (thin), para a biblioteca SUS, enquanto para a biblioteca RES foi considerada uma cadeia de 72.000 iterações, das quais 18.000 foram eliminadas (burn-in) e foram considerados os primeiros de cada quinze elementos dessa cadeia (thin). Tais procedimentos de eliminação foram realizados com o intuito de minimizar os efeitos dos valores iniciais adotados no processo iterativo.

A constatação final da convergência foi realizada por meio dos critérios de Geweke (1992) e de Raftery e Lewis (1992), isto porque Nogueira (2004) recomenda a utilização de diferentes critérios para constatar a convergência. Ambos os critérios estão disponíveis no pacote BOA ("Bayesian Output Analysis") do software R.

Os códigos utilizados para a implementação do programa no software R são apresentados no Apêndice A.

4. RESULTADOS E DISCUSSÕES

4.1. Conjunto de dados referentes ao protista *Mastigamoeba balamuthi*

Quanto às estimativas dos parâmetros da metodologia de Susko e Roger (2004), para a biblioteca não-normalizada obteve-se uma cobertura de 0,47 com erro-padrão 0,02 baseado em 715 leituras, enquanto para a biblioteca normalizada, surpreendentemente, já que feita a normalização é esperado a obtenção de bibliotecas mais uniformes, obteve-se uma cobertura quase idêntica à biblioteca não-normalizada, que foi 0,45, com erro-padrão 0,03 baseado em 363 leituras. O número esperado de EST para descobrir um novo gene foi 1,89, com erro-padrão 0,08, para a biblioteca não-normalizada e 1,82, com erro-padrão 0,11, para a biblioteca normalizada, o que indica que para se descobrir um novo gene é necessário amostrar aproximadamente 2 ESTs de ambas bibliotecas. Um resumo destas estimativas se encontra na Tabela 3.

Tabela 3. Estimativas paramétricas para a cobertura e para o número de sequências/reads necessárias para se descobrir um novo gene e os respectivos intervalos de confiança.

Conjunto de ESTs	Cobertura (C)	Intervalo de confiança 95%	Número de leituras (E)	Intervalo de confiança 95%
NãoNormalizada	0, 47	[0, 45; 0, 49]	1, 89	[1, 81; 1, 97]
Normalizada	0, 45	[0, 42; 0, 48]	1, 82	[1, 70; 1, 93]

Conforme já comentado, o parâmetro mais importante do método de Susko e Roger (2004) é o número esperado de genes, $\Delta(t)$, em uma nova amostra t vezes maior que a amostra original. Para a obtenção do mesmo, já foi visto que é necessário ajustar um modelo não-linear a fim de estimar α e γ, que por sua vez irão compor a fórmula de Δ (t). Na Figura 5 é mostrado o ajuste do modelo não-linear Binomial Negativo de Fisher (item 3.2.3) aos dados, mais especificamente, às frequências dos genes que apareceram x vezes na amostra. Dessa forma nota-se o bom ajuste do modelo binomial negativo aos dados referentes à biblioteca não-normalizada e à biblioteca normalizada pelos valores obtidos para o coeficiente de determinação (R^2), os quais foram muito próximos de um. Os parâmetros estimados por tal método foram $\alpha = -0, 778$ e $\gamma = 0, 944$ para a biblioteca não-normalizada e $\alpha = -0, 715$ e $\gamma = 0, 889$ para a biblioteca normalizada.

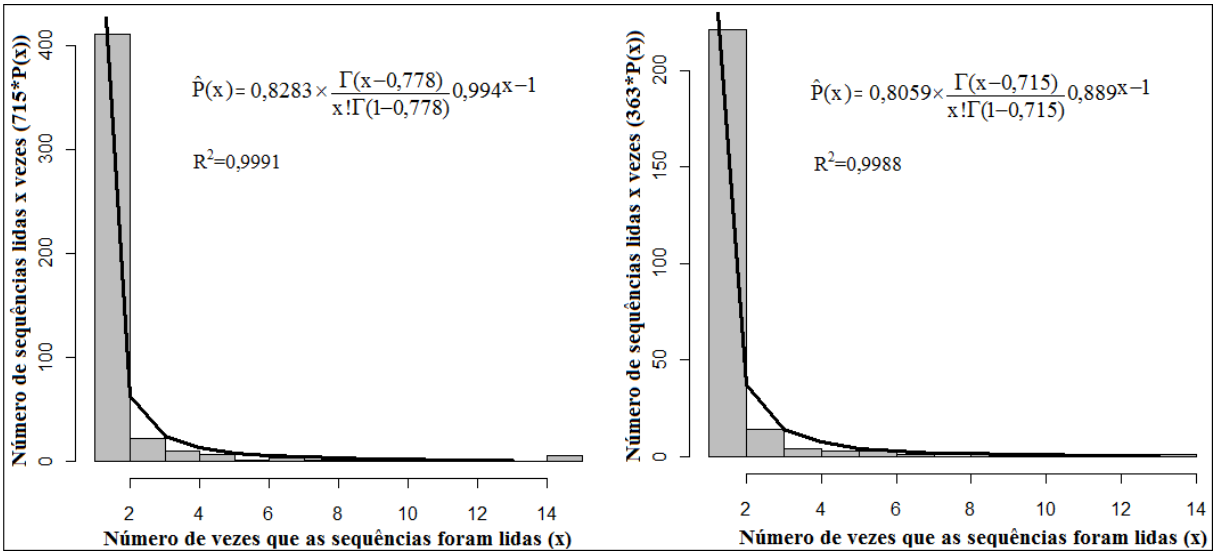

Figura 5. Histogramas das frequências do número de sequências que foram lidas k vezes e o ajuste da distribuição Binomial Negativa aos dados, referentes à biblioteca não-normalizada e normalizada, respectivamente.

Os valores em questão ($\hat{\alpha}$ e $\hat{\gamma}$) foram então substituídos na fórmula de $\Delta(t)$ a fim de obter as estimativas do número esperado de genes considerando t=0, 0,1,..., 1. Assim foi possível construir o gráfico da Figura 6 cujo eixo x representa o tamanho da nova amostra (valores de nt), sendo n o tamanho original da amostra, e o eixo y, $\hat{\Delta}(t)$, cujos valores são dados por: $\hat{\Delta}(t) = \eta_1 \hat{\alpha}^{-1} \hat{\gamma}^{-1}\{1 - (1 + \hat{\gamma}t)^{-\hat{\alpha}}\}$. Vale ressaltar que os intervalos de confiança mostrados neste mesmo gráfico foram obtidos por meio da especificação da variância $\hat{V}\,[\hat{\Delta}\,(t)]$ apresentada no item 3.2.3, a qual foi utilizada sob o ponto de vista da teoria assintótica normal a fim de obter os limites inferiores e superiores de tal intervalo.

Observa-se na Figura 6 que o número esperado de novos genes é praticamente linear com inclinação menos acentuada para a biblioteca normalizada, o que indica menor probabilidade de descoberta gênica em amostragem futura. Como exemplo, ao observar a Figura 6, pode perceber que quando o tamanho de uma nova amostra é de 400 ESTs, na biblioteca normalizada espera-se 100 novos genes, enquanto na biblioteca não-normalizada espera-se encontrar nas ESTs amostradas 200 novos genes. Nesse caso, uma taxa máxima de descoberta de genes seria alcançada amostrando ESTs da biblioteca não-normalizada, não sendo necessário nesse caso o procedimento de normalização. Este tipo de informação é extremamente inestimável ao pesquisador para avaliar a qualidade de bibliotecas de cDNA e evitar o desperdício com tempo e recursos financeiros para produzir ESTs de bibliotecas com alta redundância e relativamente "pobres" em informações gênicas.

De forma geral, pode-se questionar que a amplitude do intervalo de confiança para certos tamanhos de amostras realmente inviabilizam a eficiência das práticas utilizadas, uma vez que ao se utilizar, por exemplo, uma amostra futura de tamanho 550, as amplitudes dos intervalos de confiança para $\hat{\Delta}(t)$ são respectivamente, de 50 e 35 genes para as bibliotecas não-normalizadas e normalizadas. Tomando por exemplo a primeira biblioteca, observa-se que a estimativa pontual é $\hat{\Delta}(t)=275$, portanto, a redução de 25 genes influencia significativamente os resultados esperados para o prosseguimento de pesquisas de ordem de seqüenciamento genômico ou de expressão gênica. Portanto, assegurar que tais processos de estimação sejam mais precisos apresenta-se como uma importante fonte de pesquisa nas áreas de estatística genética e biometria.

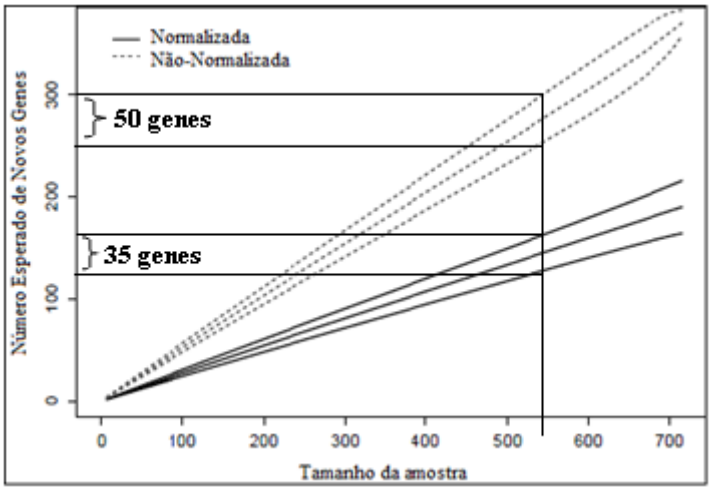

Figura 6. Estimativa do número esperado de genes como uma função do tamanho da amostra.

4.1.1. Proposta Bayesiana

As estimativas dos parâmetros α e γ foram obtidas a partir das médias das distribuições marginais a posteriori que por sua vez foram obtidas pelos métodos MCMC (Markov Chain Monte Carlo), mais especificamente, os algoritmos Metropolis-Hastings e Gibbs sampling.

Para verificar a convergência das cadeias geradas pelos algoritmos MCMC foram utilizados os diagnósticos de convergência de Raftery e Lewis e Geweke, dispostos no pacote Bayesian Output Analysis Program (BOA) no R-project. Foi considerada inicialmente uma cadeia piloto com 5.000 iterações, a fim de obter estimativas dos tamanhos *burn-in* e *thin* via critério de Raftery e Lewis. Nas tabelas 4 e 5 são apresentados os resultados provenientes destas cadeias piloto.

Tabela 4. Média, Desvio-Padrão, Limites inferiores e superiores e Diagnósticos de convergência de Geweke e Raftery e Lewis para as estimativas de α e γ considerando 5.000 iterações (biblioteca não-normalizada).

Parâmetro	Média	DP	L_{inf}	L_{sup}	Geweke (P_G)	Niter	Burn-in	Thin
α	−0, 7469	0, 0375	−0, 8154	-0, 6693	0, 9803 (0, 3270)	10.102	6	2
γ	0, 9975	0, 0023	0, 9930	0, 9998	−1, 2216 (0, 2219)	3.900	2	1

Tabela 5. Média, Desvio-Padrão, Limites inferiores e superiores e Diagnósticos de convergência de Geweke e Raftery e Lewis para as estimativas de α e γ considerando 5.000 iterações (biblioteca normalizada).

Parâmetro	Média	DP	L_{inf}	L_{sup}	Geweke (P_G)	Niter	Burn-in	Thin
α	−0, 7543	0, 0363	−0, 8199	−0, 6693	0, 7357 (0, 4619)	7.240	4	2
γ	0, 9930	0, 0042	0, 9836	0, 9984	−1, 7344 (0, 0828)	3.900	2	1

Pelo teste de Geweke, a cadeias teriam convergido ($P_G > 0,05$), porém, pelo diagnóstico de convergência de Raftery e Lewis, que indicou 10.102 e 7.240 iterações para as bibliotecas não-normalizada e normalizada, respectivamente, necessitando, portanto, de cadeias maiores. Assim, reconsiderou-se uma cadeia com 20.000 iterações para as duas bibliotecas. Nas Tabelas 6 e 7 são apresentados as estimativas obtidas de tais cadeias.

Tabela 6. Média, Desvio-Padrão, Limites inferiores e superiores e Diagnósticos de convergência de Geweke e Raftery e Lewis para as estimativas de α e γ considerando 20.000 iterações (biblioteca não-normalizada).

Parâmetro	Média	DP	L_{inf}	L_{sup}	Geweke(P_G)	Niter	Burn-in	Thin
α	−0, 7539	0,0374	−0, 8207	−0, 6693	0, 8041 (0, 4213)	11.931	9	3
γ	0, 9929	0,0041	0, 9832	0, 9985	0, 1412 (0, 8877)	3.795	2	1

Tabela 7. Média, Desvio-Padrão, Limites inferiores e superiores e Diagnósticos de convergência de Geweke e Raftery e Lewis para as estimativas de α e γ considerando 20.000 iterações (biblioteca normalizada).

Parâmetro	Média	DP	L_{inf}	L_{sup}	Geweke(P_G)	Niter	Burn-in	Thin
α	0, 3207	0,0849	-0, 5001	-0, 1712	0, 3998 (0, 6893)	18.480	15	5
γ	0, 9845	0,0060	0, 9710	0, 9940	-1, 0577 (0, 2902)	3.811	2	1

Pelo teste de Geweke, a cadeias convergiram ($P_G > 0,05$) e também pelo diagnóstico de convergência de Raftery e Lewis, que indicou 11.931 e 18.480 iterações para as bibliotecas não-normalizada e normalizada, respectivamente, sendo esses valores menores que 20.000. Assim não foi necessário reconsiderarmos uma cadeia com um número maior de iterações.

Os 5.000 primeiros elementos destas cadeias foram descartados respeitando o burn-in indicado de 9 e 2 para alfa e beta, respectivamente, na biblioteca não-normalizada e de 15 e 2 para alfa e beta, respectivamente, na biblioteca normalizada. Dos restantes foram selecionados os primeiros de cada cinco elementos respeitando o thin de 3 e 1 para alfa e beta, respectivamente na biblioteca não-normalizada e de 5 e 1 para alfa e beta, respectivamente na biblioteca normalizada. Desse modo, foram obtidas amostras das distribuições a posteriori com 3.000 elementos em cada cadeia. As cadeias geradas pelas 20.000 iterações e as cadeias resultantes após o burn-in e thin, são ilustradas pelas Figuras 7, 8, 9 e 10.

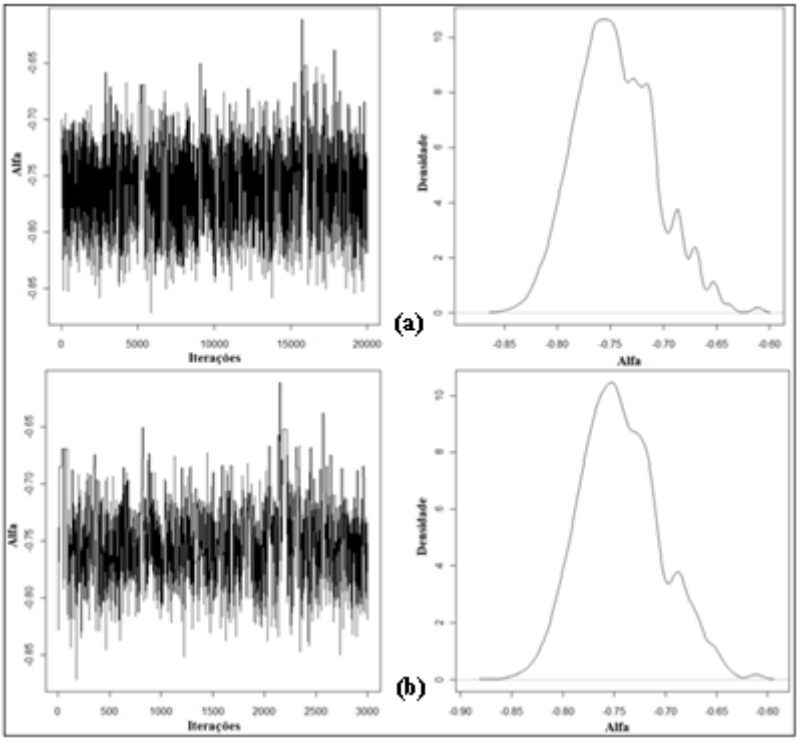

Figura 7. Valores e densidade das estimativas do parâmetro α obtidas através das 20.000 iterações não considerando (a) e considerando (b) o burn-in e o thin da cadeia gerada referente à biblioteca não-normalizada do protista *Mastigamoeba balamuthi.*

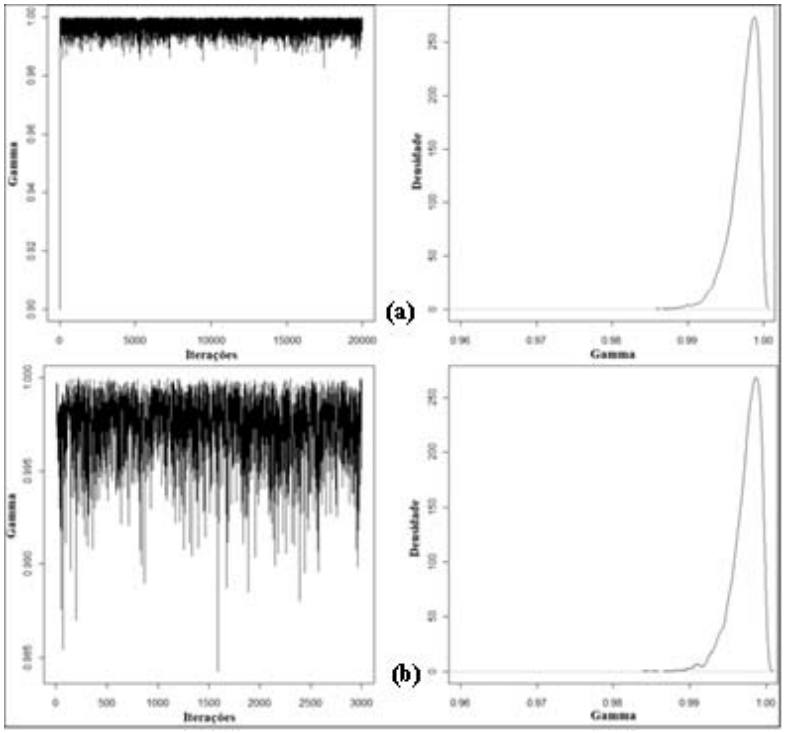

Figura 8 - Valores e densidade das estimativas do parâmetro γ obtidas através das 20.000 iterações não considerando (a) e considerando (b) o burn-in e o thin da cadeia gerada referente à biblioteca não-normalizada do protista *Mastigamoeba balamuthi.*

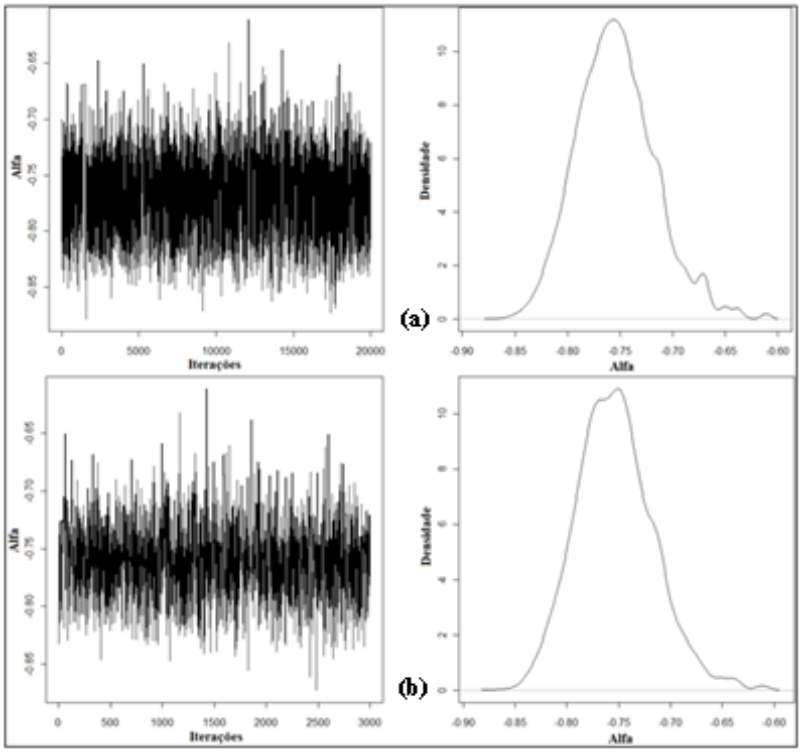

Figura 9. Valores e densidade das estimativas do parâmetro α obtidas através das 20.000 iterações não considerando (a) e considerando (b) o burn-in e o thin da cadeia gerada referente à biblioteca normalizada do protista *Mastigamoeba balamuthi.*

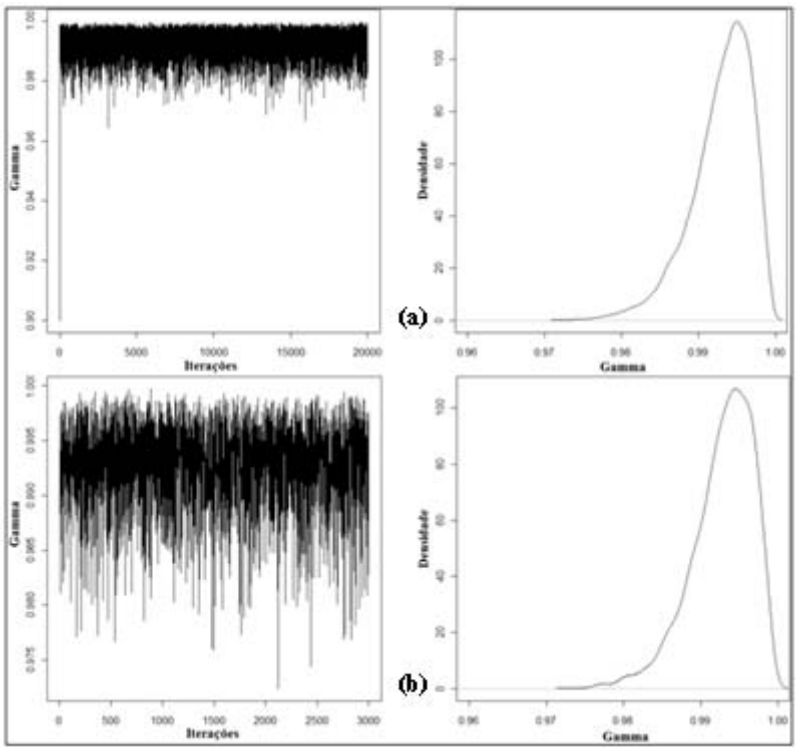

Figura 10. Valores e densidade das estimativas do parâmetro γ obtidas através das 20.000 iterações não considerando (a) e considerando (b) o burn-in e o thin da cadeia gerada referente à biblioteca normalizada do protista *Mastigamoeba balamuthi.*

Algumas descrições das cadeias resultantes, referente aos parâmetros alfa e beta, para as duas bibliotecas consideradas se encontram na Tabela 8.

Tabela 8. Média, Desvio-Padrão, Limites inferiores e superiores para α e γ considerando as cadeias geradas (3.000 iterações) para as bibliotecas não-normalizada e normalizada, considerando seus respectivos burn-in e thin.

	Não-Normalizada				Normalizada			
Parâmetro	Média	DP	L_{inf}	L_{sup}	Média	DP	L_{inf}	L_{sup}
α	-0, 7434	0, 0392	-0, 8134	-0, 6666	-0, 7534	0, 0375	-0, 8262	-0, 6812
γ	0, 9975	0, 0018	0, 9939	0, 9999	0, 9928	0, 0040	0, 9849	0, 9990

Implementando os 3.000 valores considerados de alfa e beta para a obtenção do número esperado de novos genes, obtém-se o gráfico da Figura 11. Nesta pode-se verificar que houve uma melhora significativa na estimação por intervalo para o número esperado de novos genes ao se utilizar a inferência bayesiana, uma vez que as amplitudes dos intervalos de credibilidade para um tamanho de amostra de 550 foram, respectivamente, de 14 e 8 genes para as bibliotecas não-normalizada e normalizada. Na primeira biblioteca, observa-se que a estimativa pontual é $\hat{\Delta}(t)=270$, portanto, a redução de sete genes não influencia

significamente os resultados obtidos pelo seqüenciamento no que diz respeito ao nível de expressão da biblioteca.

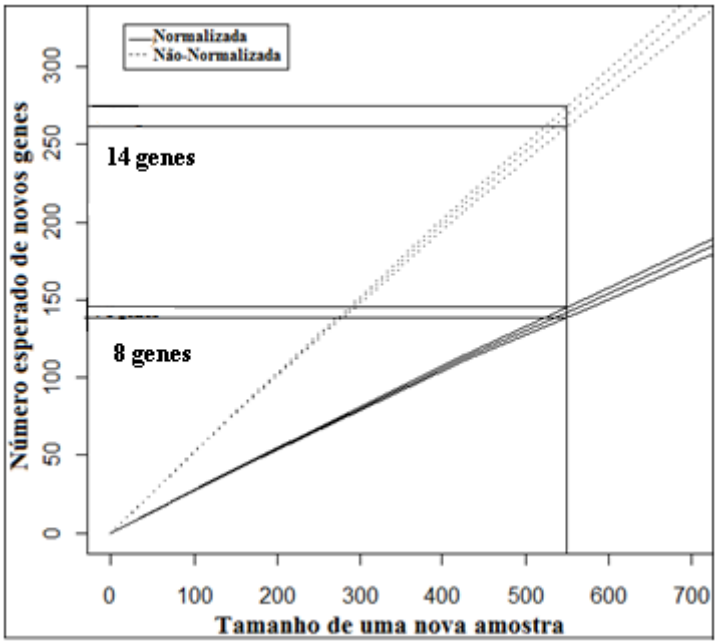

Figura 11. Estimativa bayesiana para o número esperado de novos genes em função de uma nova amostragem de transcritos para as bibliotecas não-normalizada e normalizada. A linha central fornece a estimativa e as linhas em torno, o intervalo de credibilidade a 95%.

Pode ser verificado na Figuras 5 e na Tabela 8 que de forma geral as estimativas pontuais de α e γ foram substancialmente semelhantes, sendo a maior diferença obtida para o parâmetro γ, cujas estimativas para a biblioteca normalizada foram respectivamente 0, 8890 e 0, 9928 para os métodos frequentista e bayesiano. Porém, o fator mais relevante é que houve uma melhora significativa na estimação por intervalo para o número esperado de novos genes ao se utilizar a inferência bayesiana em relação à metodologia frequentista. De forma geral, esta melhoria pode ser observada nas Figuras 12 e 13. Em relação à superioridade da estimação por intervalo bayesiana em relação à frequentista, vários autores relatam que o intervalo de credibilidade realmente se sobressai ao intervalo de confiança principalmente quando este último lança mão de propriedades assintóticas. Dentre estas autores destacam-se Silva (2006) e Silva et. al. (2011), os quais compararam a estimação por intervalo entre os métodos frequentista e bayesiano na previsão de valores genéticos de touros Nelores para tempos futuros.

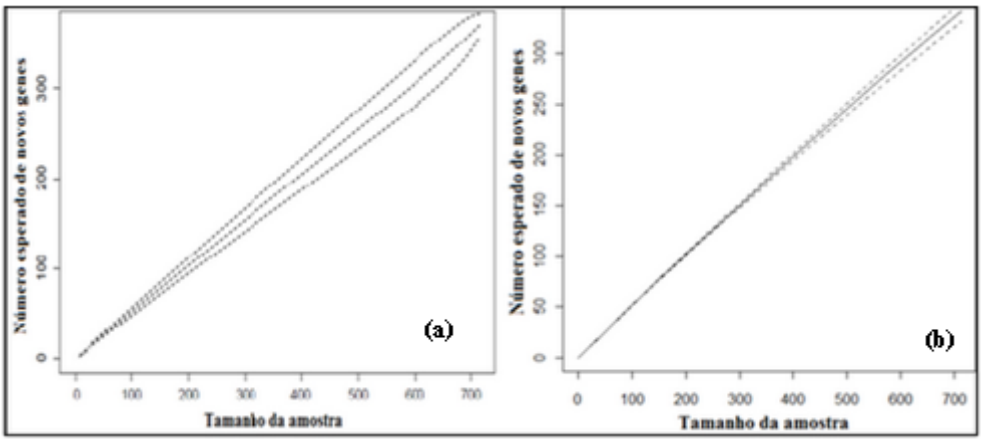

Figura 12. Número esperado de genes em função do tamanho da amostra pela metodologia proposta por Susko e Roger (2004) (a) e pela abordagem bayesiana (b) proposta no presente trabalho, referente à biblioteca não-normalizada.

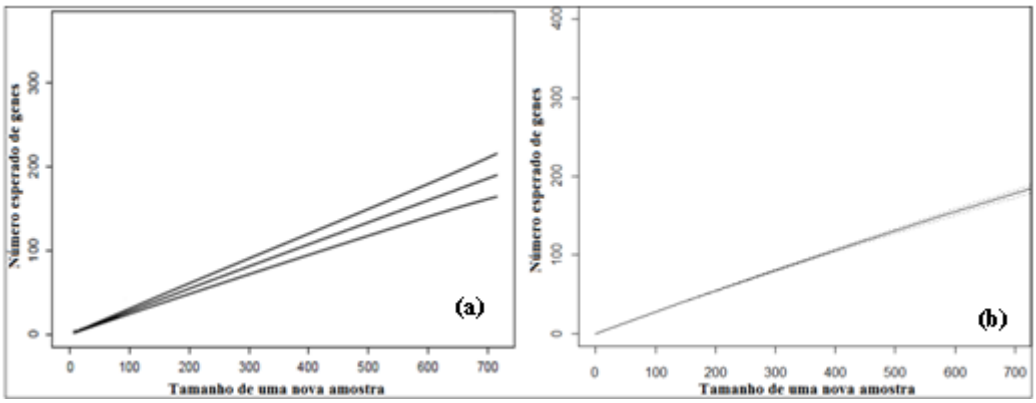

Figura 13. Número esperado de genes em função do tamanho da amostra pela metodologia proposta por Susko e Roger (2004) (a) e pela abordagem bayesiana (b) proposta no presente trabalho, referente à biblioteca normalizada.

4.2. Conjunto de dados referentes aos bovinos F_2 (holandês x gir)

Duas bibliotecas de cDNA não-normalizadas foram geradas a partir de biópsias de pele obtidas de animais F_2 infestados com *Riphicephalus B. microplus* e foram descritas na Tabela 2. As ESTs foram agrupadas usando CAP3 com os parâmetros *default*. Para os dados de EST, o perfil agrupamento de n genes foi diretamente sumarizado dos resultados do CAP3. A Figura 14 sumariza os dois conjuntos de *EST* quanto à distribuição em possíveis grupos de genes. Para a biblioteca do grupo dos animais resistentes (RES), foram gerados 1.207 transcritos com 820 genes únicos, os quais foram distribuídos em dez grupos com níveis de expressão 619, 135, 29, 17, 5, 6, 5, 1, 2, 1. No primeiro nível, 619 genes apareceram uma única vez e, no nível mais baixo, um gene foi representado 31 vezes no total de 820 genes.

Para a biblioteca SUS, foram gerados 1.350 transcritos com 981 genes únicos, agrupados em 12 clusters com níveis de expressão 806, 108, 21, 19, 8, 7, 3, 3, 2, 2, 1, 1. Nesse conjunto de dados, 819 genes apareceram apenas uma vez e 1 gene com 13 possíveis transcritos representando o mesmo gene.

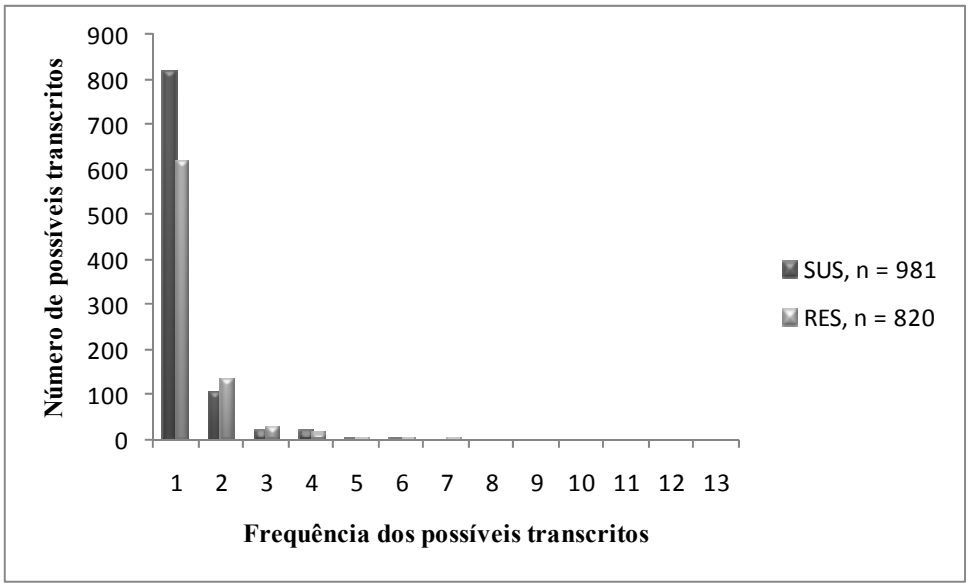

Figura 14. Distribuição das ESTs em relação ao número possível de transcritos. Fonte: Nascimento, 2009.

As estimativas para cobertura foram 0,49 (±0,02) e 0,40 (±0,02) para as bibliotecas RES (n=820) e SUS (n=981) e indicam, respectivamente, que as ESTs amostradas representam redundância de 49 e 40%, ou seja, para cada 100 transcritos gerados, 51 (RES) e 60 (SUS) deles representam genes expressos apenas uma vez. As estimativas para RES foram 1,94 (RES) e 1,66 (SUS) e os intervalos de confiança (α=5%), foram 1,94 (±0,04) e 1,66 (±0,03), para RES e SUS, respectivamente (Tabela 9). Esses valores indicam que, em média, serão necessários aproximadamente duas ESTs para descobrir um novo gene. Ambas estimativas indicam maior número de genes redundantes na biblioteca RES, embora esta tenha menor número de ESTs (n=820), o que indica possivelmente, erros de agrupamento neste conjunto de dados e estes podem estar inflacionando essas estimativas. Segundo Nascimento (2009), provavelmente, uma nova amostragem de transcritos na biblioteca RES conduzirá a menor descoberta de novos genes.

As elevadas taxas de redundância obtidas indicam que protocolos mais eficientes que favoreçam a identificação de genes raros devem ser usados. Sugerem ainda que a construção de bibliotecas normalizadas pode ser mais efetiva em detectar genes raros, já que SUS e RES dizem respeito à bibliotecas não-normalizadas (Nascimento, 2009).

Tabela 9. Estimativas para a cobertura e para o número de leituras necessárias para se descobrir um novo gene e os respectivos intervalos de confiança.

Conjuntos de ESTs	Cobertura	Intervalo de confiança de 95%	Número de leituras	Intervalo de confiança de 95%
SUS	0,40	[0,38; 0,42]	1,66	[1,63; 1,69]
RES	0,49	[0,47; 0,51]	1,94	[1,90; 1,98]

Os valores estimados para α e γ a fim de calcular o número esperado de novos genes em uma amostra futura através do método frequentista foram: $\hat{\alpha}$ = -0, 7115 e $\hat{\gamma}$ =1, 0586 para SUS e $\hat{\alpha}$ = -0, 4944 e $\hat{\gamma}$= 0, 8549 para RES.

Da mesma forma que apresentado no item 4.1, os valores em questão ($\hat{\alpha}$ e $\hat{\gamma}$) foram utilizados a fim de obter as estimativas do número esperado de genes considerando t=0, 0,1,..., 1. Nas Figuras 15 e 16 são mostrados, respectivamente para as bibliotecas SUS e RES, os gráficos provenientes dos valores de $\hat{\Delta}(t)$ em função do tamanho das novas amostras (valores de nt). Nota-se também nestes gráficos que os intervalos de confiança assintóticos de 95% foram plotados tendo como base as estimativas das variâncias apresentadas no item 3.2.3.

O número esperado de novos genes é praticamente linear e com inclinação menos acentuada para RES, o que indica menor probabilidade de descoberta de novos genes em uma amostragem futura. A título de ilustração, tomando, por exemplo, uma nova amostra de tamanho 400, observa-se que a biblioteca SUS providenciará um número esperado de 305 genes enquanto que para a biblioteca RES tal número é de 230 genes. De forma geral, estas informações são importantes para a condução de novas pesquisas nas quais tais valores auxiliarão na tomada de decisão quanto o prosseguimento do seqüenciamento com a utilização de técnicas de normalização para ambas as bibliotecas ou não. A princípio estas técnicas são mais recomendadas para a biblioteca RES, uma vez que esta gerando um número menor de genes, embora esta apresente uma quantidade maior de seqüências.

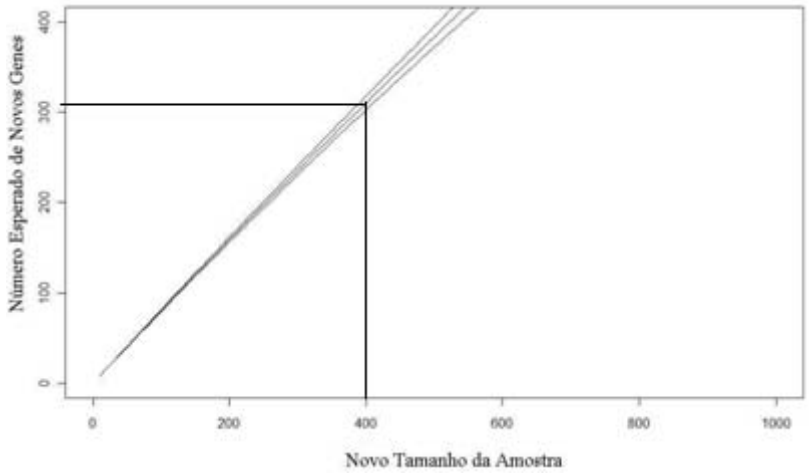

Figura 15. Estimativas para número esperado de novos genes em função de uma nova amostragem de transcritos para o grupo suscetível (SUS).

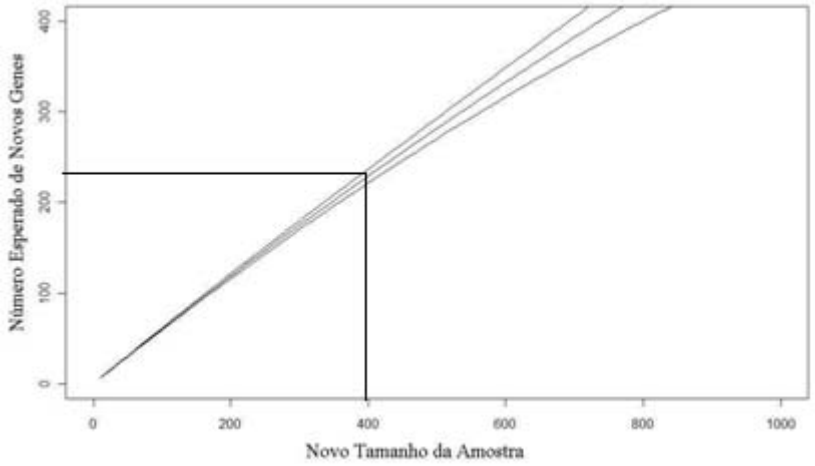

Figura 16. Estimativas para número esperado de novos genes em função de uma nova amostragem de transcritos para o grupo resistente (RES).

4.2.1. Proposta bayesiana

As estimativas dos parâmetros α e γ foram obtidas a partir das médias das distribuições marginais a posteriori, que por sua vez foram obtidas pelos métodos MCMC (Markov Chain Monte Carlo), mais especificamente, os algoritmos Metropolis-Hastings e Gibbs sampling, respectivamente. As rotinas relativas à implementação de tais algoritmos se encontram nos Apêndice.

Para verificar a convergência das cadeias geradas pelos algoritmos MCMC utilizamos os diagnósticos de convergência de Raftery e Lewis e Geweke, dispostos no pacote Bayesian Output Analysis Program (BOA) do software R (R Development Core Team, 2010). Foi considerada inicialmente uma cadeia piloto de 5.000 iterações, cujos resumos estão apresentados nas Tabelas 10 e 11.

Nota-se que o critério de Geweke (1992) não rejeitou a hipótese de convergência (PG > 0,05) em nenhuma situação, enquanto que o critério de Raftery e Lewis (1992) mostrou que é necessário um número maior de iterações para que as cadeias convirjam.

Tabela 10. Média, desvio-padrão, limites inferiores e superiores e diagnósticos de convergência de Geweke e Raftery e Lewis para as estimativas de α e γ considerando 5.000 iterações (biblioteca referente à pele dos animais suscetíveis).

Parâmetro	Média	DP	L_{inf}	L_{sup}	Geweke (P_G)	Niter	Burn-in	Thin
α	−0, 7304	0, 0465	−0, 8160	−0, 6375	1, 1112 (0, 2665)	9.614	6	2
γ	0, 9991	0, 0008	0, 9969	0, 9999	−1, 6353 (0, 1020)	4.030	3	1

Tabela 11. Média, desvio-padrão, limites inferiores e superiores e diagnósticos de convergência de Geweke e Raftery e Lewis para as estimativas de α e γ considerando 5.000 iterações (biblioteca referente à pele dos animais resistentes).

Parâmetro	Média	DP	L_{inf}	L_{sup}	Geweke (P_G)	Niter	Burn-in	Thin
α	−0, 5718	0, 0505	−0, 6718	-0, 4746	0, 9440 (0, 3452)	71.786	55	11
γ	0, 9829	0, 0040	0, 9753	0, 9892	−0, 4548 (0, 6492)	3.840	2	1

Devido a indicação de 9.614 e 71.786 iterações, respectivamente para as bibliotecas SUS e RES, pelo diagnóstico de convergência de Raftery e Lewis, foram efetuadas novas análises considerando 20.000 iterações para a biblioteca SUS e 72.000 iterações para a biblioteca RES (Tabelas 12 e 13).

Para as novas cadeias geradas, como se pode perceber nas Tabelas 12 e 13, nota-se que o critério de Geweke não rejeitou a hipótese de convergência ($P_G > 0,05$) em nenhuma situação, e que o critério de Raftery e Lewis (1992), fundamentado no fator de dependência, também mostrou que todas as cadeias convergiram, já que os números de iterações apontados, de 12.615 e 65.268, são respeitados pelos números de iterações considerados, os quais foram de 20.000 e 72.000, respectivamente para as bibliotecas SUS e RES.

Tabela 12. Média, desvio-padrão, limites inferiores e superiores e diagnósticos de convergência de Geweke e Raftery e Lewis para as estimativas de α e γ considerando 20.000 iterações (biblioteca referente à pele dos animais suscetíveis).

Parâmetro	Média	DP	L_{inf}	L_{sup}	Geweke(P_G)	Niter	Burn-in	Thin
α	−0, 7307	0, 0459	−0, 8152	−0, 6371	0, 7108 (0, 4772)	12.615	9	3
γ	0, 9992	0, 0008	0, 9969	0, 9999	−0, 8436 (0, 3989)	3.938	2	1

Tabela 13. Média, desvio-Padrão, limites inferiores e superiores e diagnósticos de convergência de Geweke e Raftery e Lewis para as estimativas de α e γ considerando 72.000 iterações (biblioteca referente à pele dos animais resistentes).

Parâmetro	Média	DP	L_{inf}	L_{sup}	Geweke (P_G)	Niter	Burn-in	Thin
α	0,5719	0, 0497	−0, 6663	−0, 4713	0,8204 (0,4120)	65.268	42	14
γ	0, 9829	0, 0036	0, 9751	0, 9892	−0,4159 (0, 6775)	3.772	2	1

Para ilustrar a análise das cadeias geradas pelos algoritmos Gibbs Sampler e Metropolis-Hastings na análise dos dados conjunto de dados referentes aos bovinos F_2 (holandês x gir), os gráficos com os valores gerados a cada iteração e a densidade da distribuição a posteriori para o parâmetro alfa e gama, para as bibliotecas referentes à pele dos animais suscetíveis e resistentes são apresentados nas Figuras 17, 18, 19 e 20.

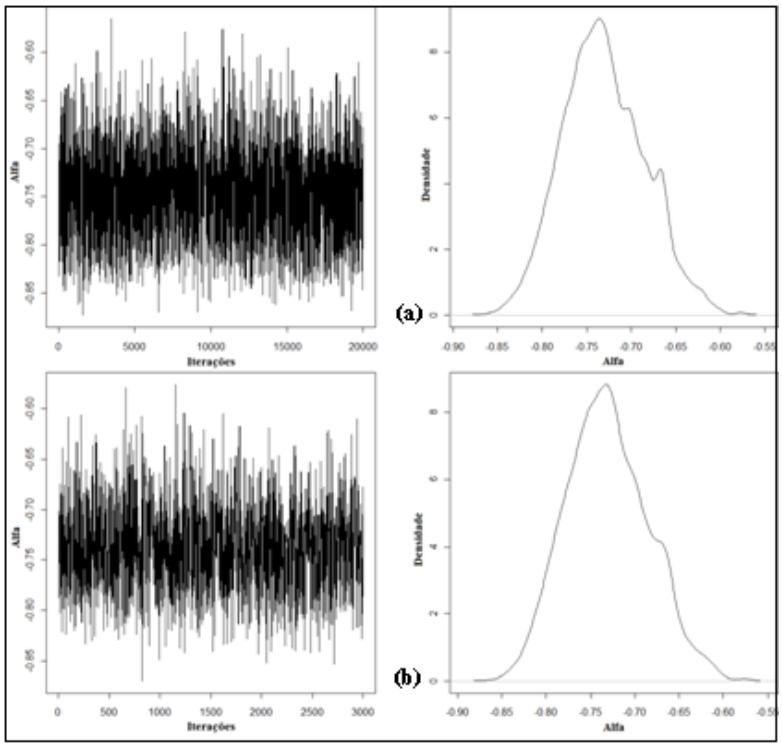

Figura 17. Valores e densidade de α obtidos através das 20.000 iterações não considerando (a) e considerando (b) o burn-in e o thin da cadeia gerada referente à pele dos animais suscetíveis.

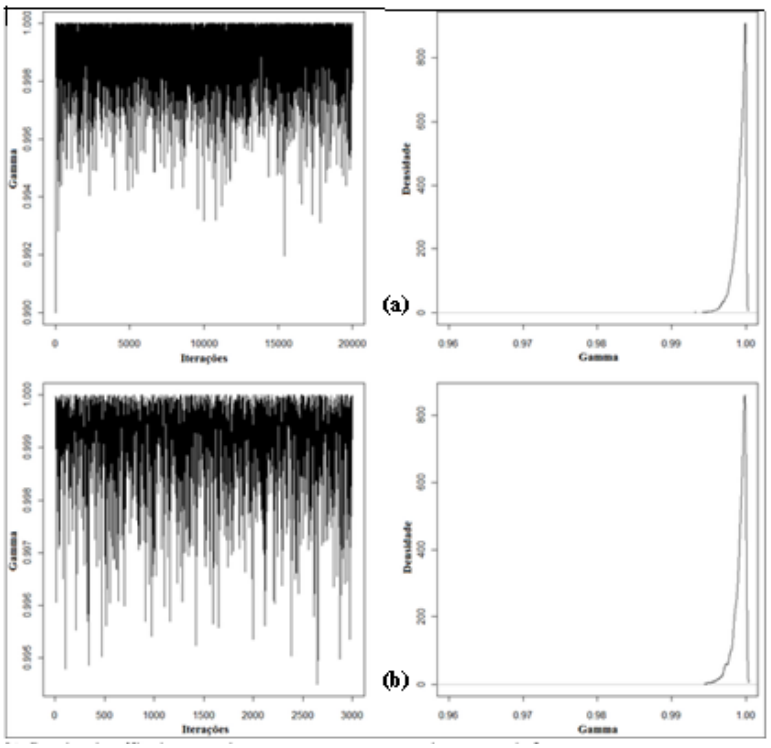

Figura 18. Valores e densidade de γ obtidos através das 20.000 iterações não considerando (a) e considerando (b) o burn-in e o thin da cadeia gerada referente à pele dos animais suscetíveis.

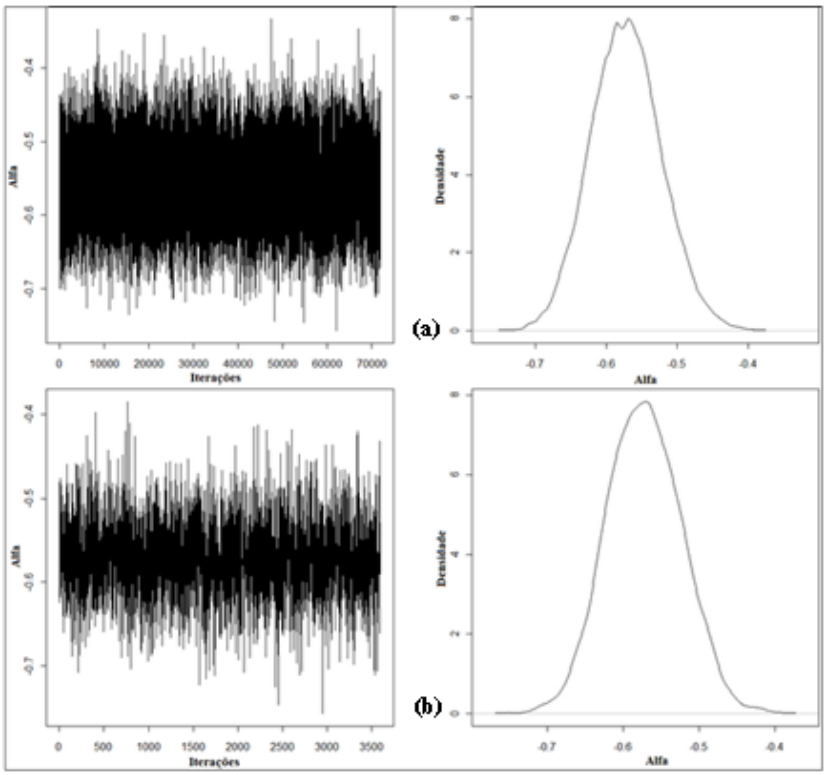

Figura 19. Valores e densidade de α obtidos através das 72.000 iterações não considerando (a) e considerando (b) o burn-in e o thin da cadeia gerada referente à pele dos animais resistentes.

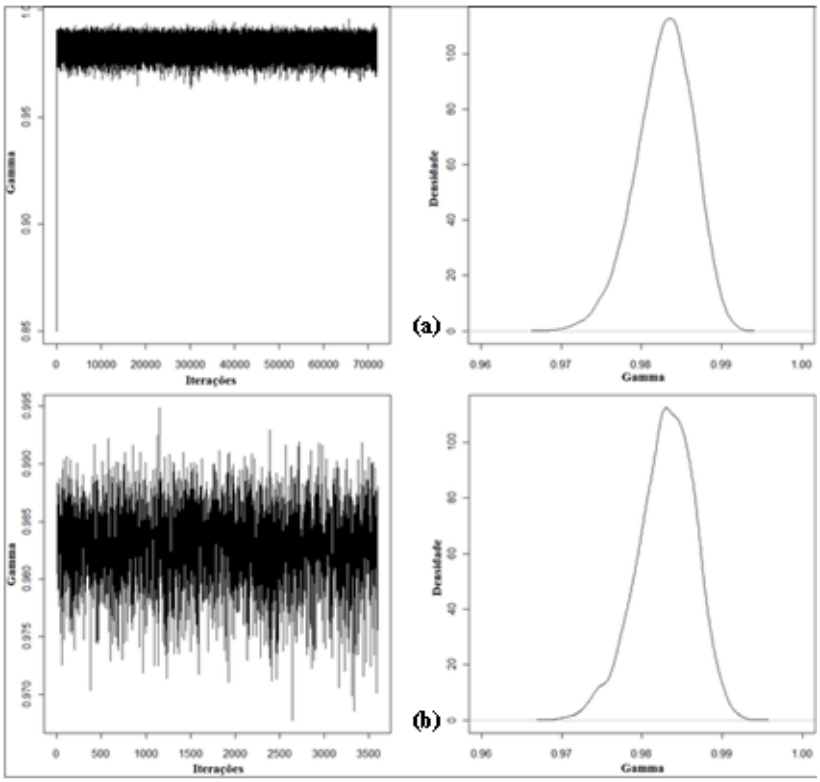

Figura 20. Valores e densidade de γ obtidos através das 72.000 iterações não considerando (a) e considerando (b) o burn-in e o thin da cadeia gerada referente à pele dos animais resistentes.

Os valores indicados para o burn-in e para o thin por meio do critério de Raftery e Lewis (1992), como pode ser visto nas Tabelas 12 e 13, foram todos inferiores aos utilizados. Para o grupo SUS o burn-in utilizado foi de 5.000 e o thin de 5, restando 3.000 valores na cadeia gerada, enquanto para o grupo RES o burn-in utilizado foi de 18.000 e o thin de 15, restando 3.600 valores na cadeia gerada. Tal procedimento certifica que as inferências a posteriori foram realizadas desconsiderando independência entre os valores gerados para os parâmetros alfa e gama e possíveis influências dos valores iniciais. Para enfatizar tal discussão vale lembrar que as cadeias resultantes após o burn-in e o thin, foram ilustradas nas Figuras 17, 18, 19 e 20, juntamente com suas respectivas cadeias originais. Os resultados de tais análises encontram-se na Tabela 14.

Tabela 14. Média, desvio-padrão, limites inferiores e superiores para α e γ considerando as cadeias geradas para as bibliotecas referentes à pele dos animais suscetíveis (SUS) e resistentes (RES), considerando seus respectivos burn-in e thin.

	SUS				RES			
Parâmetro	Média	DP	L_{inf}	L_{sup}	Média	DP	L_{inf}	L_{sup}
α	-0, 7309	0, 0460	-0, 8185	-0, 6442	-0, 5721	0, 0492	-0, 6635	-0, 4764
γ	0, 9992	0, 0008	0, 9974	1, 0000	0, 9830	0, 0036	0, 9759	0, 9900

As Figuras 21 e 22 mostram os gráficos do número esperado de genes em função do tamanho de uma nova amostra, obtidos através das estimativas de alfa e gama geradas pelas cadeias consideradas, com relação aos dois conjuntos de dados SUS e RES, respectivamente.

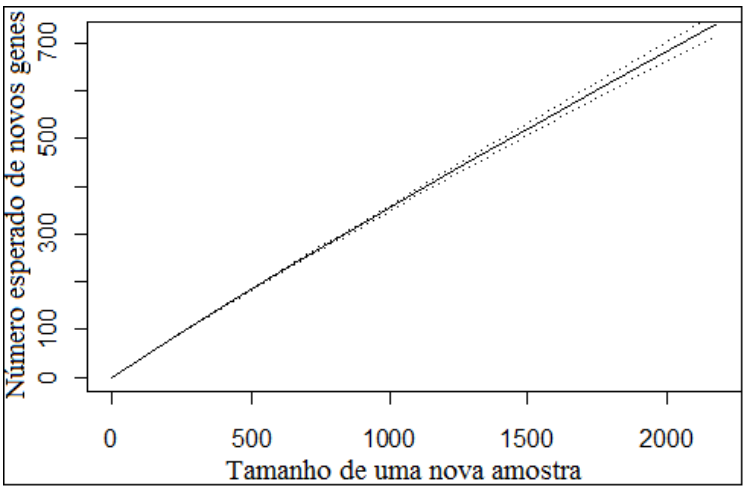

Figura 21**.** Estimativa bayesiana para o número esperado de novos genes em função de uma nova amostragem de transcritos para o grupo suscetível (SUS). A linha central fornece a estimativa e as linhas em torno do intervalo de credibilidade a 95%.

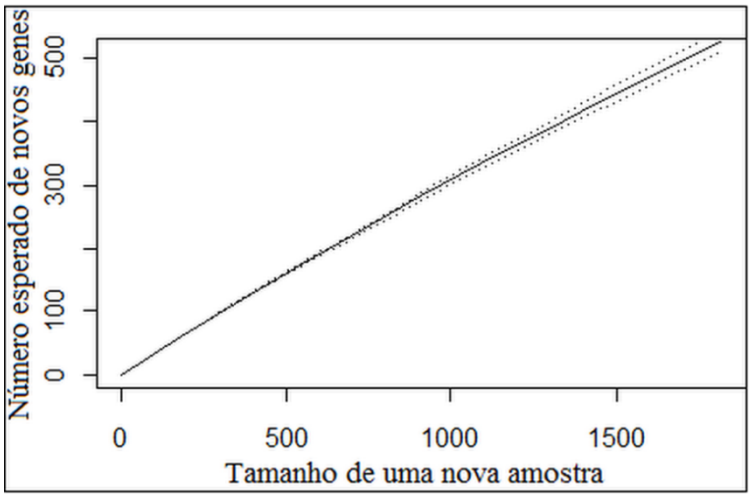

Figura 22. Estimativa bayesiana para o número esperado de novos genes em função de uma nova amostragem de transcritos para o grupo suscetível (RES). A linha central fornece a estimativa e as linhas em torno do intervalo de credibilidade a 95%.

É possível notar nas Figuras 23 e 24 que realmente a metodologia Bayesiana apresentada foi eficiente, já que as mesmas mostram que a estimação por intervalo para o número esperado de genes, quando se considera a inferência bayesiana foi consideravelmente menor que estimada segundo Susko e Roger (2004).

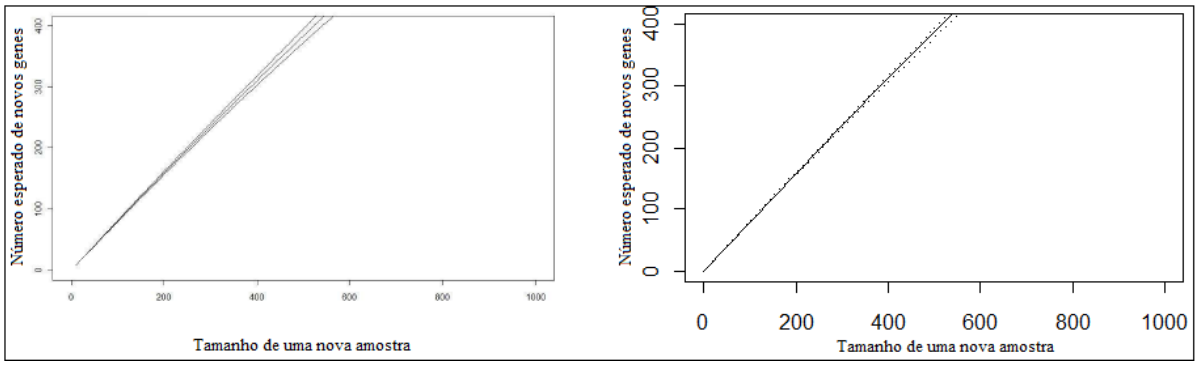

Figura 23. Número esperado de genes em função do tamanho da amostra pela metodologia proposta por Susko e Roger (2004) e pela abordagem bayesiana proposta no presente trabalho, referente ao grupo suscetível (SUS), respectivamente.

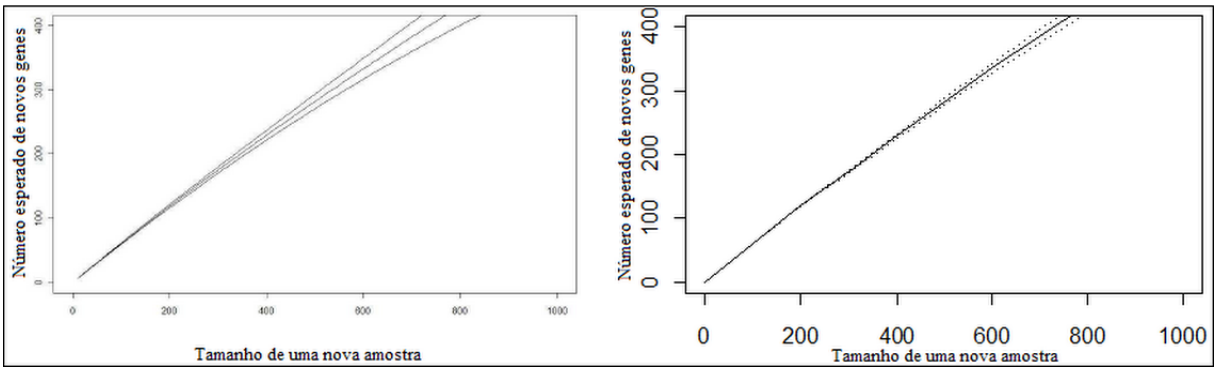

Figura 24. Número esperado de genes em função do tamanho da amostra dada pela metodologia proposta por Susko e Roger (2004) e pela abordagem bayesiana proposta no presente trabalho, referente ao grupo resistente (RES), respectivamente.

5. CONCLUSÕES

A descrição detalhada dos métodos apresentados por Susko e Roger (2004) permitiu compreender os aspectos teóricos relacionados ao conceito de redundância em estudos de bibliotecas de cDNA e realçar sua importância a ponto de decidir quais bibliotecas estão mais propensas a produção de novos genes.

A abordagem bayesiana para o parâmetro número esperado de genes em uma nova amostra, $\Delta(t)$, mostrou-se adequada e proporcionou estimação por intervalo mais precisa em relação ao método frequentista proposto por Susko e Roger (2004).

A metodologia proposta por Susko e Roger e a proposta bayesiana, foram implementadas de forma prática e eficiente no software R para todos os conjuntos de dados utilizados, caracterizando assim como uma ferramenta útil para estudos desta natureza.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ADAMS, M. D.; KELLEY, J. M.; GOCAYNE, J. D.; DUBNICK, M.; POLYMEROPOULOS, M. H.; XIAO, H.; MERRIL, C. R.; WU, A.; OLDE, B.; MORENO, R. F.; KERLAVAGE, A. R.; MC-COMBIE, W. R.; VENTER, J. C**. Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project.** Science, 252:1651–1656, June 1991.

ASAI, M. 2005. Comparison of MCMC Methods for Estimating Stochastic Volatility Models. **Computational Economics**, New York, v. 25, n.4, p.281-301, June 2005.

BARRETO, G.; ANDRADE, M.G. Robust Bayesian Approach for AR(p) Models Applied to Streamflow Forecasting. **Journal Applied Statistical Science**, New York, v.12, n.3, p.269-292, Mar. 2004.

BAUDET, C. **Uma abordagem para detecção e remoção de artefatos em sequências ESTs.** Dezembro, 2006. 228 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Estadual de Campinas, Campinas, SP, 2006.

BERG, B.A, **Markov Chain Monte Carlo Simulations and Their Statistical Analysis: With Web-based Fortran Code.** World Scientific, Singapore, 2004. 35p.

BOX, G. E. P., TIAO, G. C. Bayesian inference in statistical analysis. Addison-Wesley, 588p. 1973.

BLASCO, A.; PILES, M.; VARONA, L.A. Bayesian analysis of the effect of selection for growth rate on growth curves in rabbits. **Gen. Sel. Evol.**, v.35, p.21-41, 2002.

BROEMILING, L.D. **Bayesian Analysis of linear models**. New York, USA: John Wiley and Sons, 1989, 412p.

CASELLA, G.; GEORGE, E.I. Explaining the Gibbs sampler. **The American Statistician**, Salt Lake, v.46, n.3, p.167-174, July 1992.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hastings algorithm. **The American Statistician**, Salt Lake, v.49, n.4, p.327-345, Nov. 1995.

COWARD, E.; HAAS, S.A.; VINGRON, M. SpliceNest: visualization of gene structure and alternative splicing based on EST clusters. **Trends in Genetics,** Oxford, v. 18. 2002.
EHLERS, R.S, 2007. **Introdução à Inferência Bayesiana**. Disponível em http://leg.ufpr.br/~ ehlers/bayes. Acesso em: 26/03/10.

FIGUEIRA, A. P**. Análise Comparativa de Algoritmos de Agrupamento De ESTs.** Agosto, 2006. 96 f. Dissertação (Ciências Genômicas e Biotecnologia) - Universidade Católica de Brasília, Brasília, 2006.

GAMERMAN, D. **Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference**. Londres, UK: Chapman & Hall, 1997. 245 p.

GELFAND, A.E.; SMITH, A.F..M. Sampling based approaches to calculating marginal densities. **Journal of the American Statistical Association**, Alexandria, v.85, n.4, p.398-409, June 1990.

GELMAN, A.; RUBIN, D.B. Inference from iterative simulation using multiple sequence. **Statistical Science**, Hayward, v.7, n.4, p.457-511, May 1992.

GEWEKE, J. Evaluating the accurary of sampling-based approaches to the calculation of posterior moments. **Bayesian Statistics** 4 (eds. Bernardo, J.M.; Berger, J.O.; Dawid, A.P.; Smith, A.F.M.), New York, USA: Oxford University Press, 1992, p.625-631.

GOOD, I. J. The population frequencies of species and the estimation of population parameters. **Biometrika**, v. 40, p. 237-264, 1953.

GOOD, I. J.; TOULMIN, G. H. The number of new species, and the increase in population coverage, when a sample is increased**. Biometrika**, v. 43, p. 45-63, 1956.

HASTINGS, W. K. Monte Carlo Sampling Methods Using Markov Chains and their Applications. **Biometrika**, v. 57, p. 97-109, 1970.

HEIDELBERGER, P.; WELCH, P. Simulation run length control in the presence of an initial transient. **Operations Research**, Maryland, v.31, n.6, p.1109-44, Nov.-Dec. 1983.

HUANG, S.; WEIR, B. S. Estimating the Total Number of Alleles Using a Sample Coverage Method. **Genetics** 159:1365-1373, 2001.

HUANG, X.; MADAN, A. Cap 3: a DNA sequence assembly program. **Genome Research**, v.9, p.868-877, 1999.

JEFFREYS, H. **Theory of Probability**. Oxford, UK: Claredon Press, 1961.

LIJOI, A.; MENA, R. H.; PRÜNSTER, I. A Bayesian nonparametric method for prediction in EST analysis**. BMC Bioinformatics** 2007, 8:339.

NASCIMENTO, C. S. **Expressão gênica em bibliotecas de cdna de pele de bovinos F_2 (holandês × gir) infestados com o carrapato *riphicephalus (boophilus) microplusm*,** 2009.136 f. Tese (Doutorado em Zootecnia) - Universidade Federal de Viçosa, Viçosa, MG, 2009.

O'HAGAN, A. **Kendall's advanced theory of statistics**. Volume 2B. Bayesian Inference. New York, USA: Edward Arnold Press, 1994, 402p

PATANJALI, S.R.; PARIMOO, S.; WEISSMAN, S.M. (1991). **Construction of a uniform-abundance (normalized) cDNA library.** Proc. Natl. Acad. Sci. (88): 1943-1947.

PAULA, F. V.; SILVA, F. F.; NASCIMENTO. **Descrição e Implementação de Métodos Estatísticos aplicados à análise de dados de EST (*Expressed Sequence Tags*) em bibliotecas de cDNA.** In: 55ª RBRAS (Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria) e 15ª RARG (Reunião Anual da Região Argentina da Sociedade

Internacional de Biometria), 2010, Florianópolis-SC. Anais da 55ª RBRAS e 15ª RARG. Florianópolis-SC: UFSC, 2010.

R DEVELOPMENT CORE TEAM. **R: a language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <http://www.R-project.org>. Online. Acesso em: 2010.

RAFTERY, A. E.; LEWIS, S. **How many iterations in the Gibbs sampler?** In Bayesian Statistics (Eds.: J. M. Bernardo et al.), Oxford, USA: University Press, p.763-773. 1992.

ROSA, G.J.M. **Análise Bayesiana de Modelos Mistos Robustos via Amostrador de Gibbs.** 1998. 57 p. Tese (Doutorado em Estatística e Experimentação Agronômica) – Universidade de São Paulo, Piracicaba, SP.

SILVA, F.F. SILVA, F.F. **Análise bayesiana do modelo auto-regressivo para dados em painel**: **aplicação na avaliação genética de touros da raça Nelore**. 2006. 100 f. Tese (Doutorado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras, Lavras, MG, 2006.

SILVA, N.M.A.; LIMA, R.R.; SILVA, F.F.; MUNIZ, J.A; AQUINO, L.H. Modelo hierárquico bayesiano aplicado na avaliação genética de curvas de crescimento de bovinos de corte. Ciência Animal Brasileira. **Arq. Bras. Med. Vet. Zootec.,** v.62, n.2, p.409-418, 2010.

SMITH, B.J. Boa: an R package for MCMC output convergence assessment and posterior inference. **Journal of Statistical Software**, v.21, n.11, p.1-37, 2007.

SNOEIJER, C, Q. **Geração e análise de etiquetas de sequências transcritas - ESTs – de *trypanosoma rangeli***. Fevereiro, 2004. 64 f. Dissertação (Mestrado em Biotecnologia) – Universidade Federal de Santa Catarina, Florianópolis, SC, 2004.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian and MCMC Methods in Quantitative Genetics**, New York, USA: Springer, 2002, 770p.

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

VASCONCELOS, S. S. **Uma investigação: ESTs (Expressed Sequence Tags) podem ser usados no desenvolvimento de marcadores moleculares baseados em introns?** Outubro, 2003. 98 f. Dissertação (Mestrado em Ciências Genômicas e Biotecnologia) – Universidade Católica de Brasília, Brasília, 2003.

WANG, J. Z.; LINDSAY, B. G.; CUI, L.; WALL, P. K.; MARION, J.; ZHANG, J.; PAMPHILIS, C. W. Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries. **BMC Bioinformatics** 2005. 6:300.

APÊNDICE

Códigos de Programação no Software R

Aqui são descritas as rotinas utilizadas no software R-project considerando as bibliotecas não-normalizada e normalizada referentes ao protista *Mastigamoeba balamuthi*.

```
##########Cobertura##########

a=(read.table("C: /dadossusko.txt")) #entrando com os dados dispostos na
#Tabela 1 (considerando as colunas 1, 2 e 3) no R.

L1=a$V1*a$V2

L2=a$V1*a$V3

C1=(sum(L1)-a$V2[1])/sum(L1) #cobertura da biblioteca não-normalizada

C2=(sum(L2)-a$V3[1])/sum(L2) #cobertura da biblioteca normalizada

se_C1=(sqrt(sum(L1))^-1)*sqrt((a$V2[1]/sum(L1))+(2*a$V2[2]/sum(L1))
((a$V2[1]/sum(L1))^2)) #erro padrão referente à cobertura da biblioteca
não-normalizada

se_C2=(sqrt(sum(L2))^-1)*sqrt((a$V3[1]/sum(L2))+(2*a$V3[2]/sum(L2))-
((a$V3[1]/sum(L2))^2)) #erro padrão referente à cobertura da biblioteca
#normalizada

####número esperado de leituras exigidas para descobrir um novo gene####

eta1=1/(1-C1) #número esperado de leituras exigidas para se descobrir um
#novo gene na biblioteca não-normalizada

eta2=1/(1-C2) #número esperado de leituras exigidas para se descobrir um
#novo gene na biblioteca normalizada

se_eta1=se_C1/(1-C1) #erro padrão referente ao número esperado de leituras
#exigidas para descobrir um novo gene na biblioteca não-normalizada

se_eta2=se_C2/(1-C2) #erro padrão referente ao número esperado de leituras
#exigidas para descobrir um novo gene na biblioteca normalizada

##########Número esperado de novos genes##########

##########Biblioteca Normalizada##########

k <- c(1,2, 3, 4, 5, 6, 7, 9, 14) #número de vezes que as sequências foram
lidas
Nk <- c(200, 21, 14, 4, 3, 3, 1, 1, 1) #número de sequências lidas k vezes
p1= (k*Nk)/sum(k*Nk) #proporção das sequências lidas k vezes
d1=data.frame(cbind(p1,k))
fit1=nls(p1 ~ ((gamma(k + b1)/(factorial(k)*gamma(1 + b1))))*(b2^(k-1)),
start = list(b1=-0.87,b2=1.5),data=d1) # modelo probabilístico descrito
#ajustado aos dados via função nls
par1=coef(fit1) # coeficientes estimados do modelo probabilístico ajustado
aos dados
alfa1= par1[1]
gama1= par1[2]
```

```
eta1=200 # número de sequências que apareceram apenas uma vez na biblioteca
#normalizada
t=seq(0,2,0.01) # número de vezes que uma nova amostra é maior que a
original
deltat1=eta1*(1-((1+gama1*t)^-alfa1))/(alfa1*gama1) # número esperado de
#novos genes em uma nova amostra t vezes maior que a amostra original de
#tamanho n
plot(715*t,deltat1,type="l") # plotando o número esperado de novos genes em
#função de uma amostra de tamanho nt

##########Biblioteca Não-Normalizada##########
k <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13, 15) #número de vezes que as
#sequências foram lidas
Nk <- c(378, 33, 21, 9, 6, 1, 3, 1, 1, 1, 1, 5) #número de sequências lidas
#k vezes
eta1=378 # número de sequências que apareceram apenas uma vez na biblioteca
```

OBS 1.: Prossegue-se com os mesmos passos para a biblioteca normalizada a fim de plotar o número esperado de novos genes em função de uma amostra de tamanho nt.

```
##########Inferência Bayesiana ##########
set.seed(1234)
##########Não-Normalizada##########
 k=c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13, 15)
Nk=c(378, 33, 21, 9, 6, 1, 3, 1, 1, 1, 1, 5)
x=k*Nk
x=matrix(x,1,length(x))
N=length(x)
sum_x=sum(x)
#priori de gamma Beta(a,b)
a=99 # a e b devem ser especificado de acordo com os gráficos da dist. Beta
#(Fig. 4)
b=11
plot(density(rbeta(10000,a,b)))
#priori de alfa Beta_gen(c,d)I(-1,1)
c=30 # c e d devem ser especificado de acordo com os gráficos da dist. Beta
#(Fig. 4)
d=270
plot(density((2*rbeta(100000,c,d)-1)),col="red")
#Gibbs e Metropolis
Niter=5000 #número de iterações
a_star=matrix(0,Niter,1)
b_star=matrix(0,Niter,1)
alfa=matrix(0,Niter,1)
gama=matrix(0,Niter,1)
cand=matrix(0,Niter,1)
alfa_denom=matrix(0,Niter,1)
alfa_denom1=matrix(0,Niter,1)
alfa_num=matrix(0,Niter,1)
alfa_num1=matrix(0,Niter,1)
fator=matrix(0,Niter,1)
prob=matrix(0,Niter,1)
u=matrix(0,Niter,1)
z=matrix(0,Niter,length(x))
z1=matrix(0,Niter,length(x))
prod_z1=matrix(0,Niter,1)
z2=matrix(0,Niter,length(x))
z3=matrix(0,Niter,length(x))
```

```
prod_z3=matrix(0,Niter,1)
for(i in 1:1)
 {
  alfa[i]=-0.7
  gama[i]=0.9
 }
 for(i in 2:Niter)
  {
   a_star[i]=sum_x+a
   b_star[i]=(N*alfa[i-1]) + b
   gama[i]=rbeta(1,a_star[i],b_star[i]) #condicional completa de gamma
   cand[i]=(2*rbeta(1,c,d)-1)
for (j in 1:length(x))
{
 z[i,j]=gamma(x[j]+alfa[i-1]-1)
}
for(j in 1:length(x))
{
 if (z[i,j]>3.345253e+49) z1[i,j]=1e50 else z1[i,j]=z[i,j]
}
 prod_z1[i]=prod(z1[i,])
alfa_denom[i]=prod_z1[i]*(gamma(alfa[i-1]+1)^(-N))*(alfa[i-1]^N)*((alfa[i-
1]-1)^(c-1))*(1-alfa[i-1]^(d-1))  #condicional completa de alfa
for(l in 1:Niter)
{
 if          (alfa_denom[l]==0)          alfa_denom1[l]=1e-50          else
alfa_denom1[l]=alfa_denom[l]
}
for (j in 1:length(x))
{
z2[i,j]=gamma(x[j]+cand[i]-1)
}
for(j in 1:length(x))
{
 if (z2[i,j]>3.345253e+49) z3[i,j]=1e50 else z3[i,j]=z2[i,j]
}
 prod_z3[i]=prod(z3[i,])
alfa_num[i]=prod_z3[i]*(gamma(cand[i]+1)^(-N))*(cand[i]^N)*((cand[i]-1)^(c-
1))*(1-cand[i]^(d-1))
for(m in 1:Niter)
{
 if (alfa_num[m]==0) alfa_num1[m]=1e-50 else alfa_num1[m]=alfa_num[m]
}
fator[i]=(alfa_num1[i]/alfa_denom1[i])
    prob[i]=min(1,fator[i]) # cálculo probabilidade MH
   u[i]=runif(1,0,1) # geração do valor uniforme que será comparado com
#prob MH
   if (prob[i]> u[i]) alfa[i]=cand[i] else alfa[i]=alfa[i-1]
}
N_acept=length(unique(alfa)) #número de valores não repetidos para alfa
(aceitos)
taxa_acpet=(N_acept*100)/Niter # taxa de aceitação (ideal entre 20-40%)
taxa_acpet # taxa de aceitação do algoritmo Metropolis-Hastings
dados=cbind(alfa,gama) # valores gerados para alfa e beta por meio das
#iterações

##########burn-in e thin ##########
naonorm=dados
naonorm_burn=naonorm[5001:20000,] #burn-in de 5.000
aux1=rep(c(1,0,0,0,0),3000) #thin de 5
naonorm_burn_th=cbind(naonorm_burn,aux1)
```

```
naonorm_burn_th_fim=naonorm_burn_th[naonorm_burn_th$aux1==1,] #cadeia final
#para alfa e beta com 3.000 valores

##########Número esperado de novos genes##########
t=seq(0,2,0.2)
nt=length(t)
delta_t_nonnorm=matrix(0,nrow(naonorm_burn_th_fim),nt)
 for (i in 1:nrow(naonorm_burn_th_fim))
  {
   for (j in 1:nt)
    {
     delta_t_nonnorm[i,j]=378*(naonorm_burn_th_fim[i,1]^-
1)*(naonorm_burn_th_fim[i,2]^-1)*(1-(1+naonorm_burn_th_fim[i,2]*t[j])^-
naonorm_burn_th_fim[i,1])
    }
   }
delta_t_nonnorm_med=matrix(0,1,nt)
delta_t_nonnorm_quant=matrix(0,2,nt)
 for (j in 1:nt)
 {
  delta_t_nonnorm_med[,j]=mean(delta_t_nonnorm[,j])
  delta_t_nonnorm_quant[,j]=quantile(delta_t_nonnorm[,j],  probs  =  c(2.5,
97.5)/100)
 }
final_nonnorm=cbind(delta_t_nonnorm_quant[1,],t(delta_t_nonnorm_med),delta_
t_nonnorm_quant[2,])
plot(715*t,final_nonnorm[,1],type="l", lty=3)
lines(715*t,final_nonnorm[,2],type="l")
lines(715*t,final_nonnorm[,3],type="l", lty=3) # Gráfico do número esperado
#de novos genes disposto na Figura 22.
```

OBS 2.: Os códigos no item anterior dizem respeito à implementação das estatísticas apresentadas levando em consideração a biblioteca não-normalizada referente ao organismo *Mastigamoeba Balamuth*, para as demais bibliotecas consideradas, normalizada, SUS e RES, os códigos foram os mesmo, porém foram alterados os valores de a, b, c e d (parâmetros das prioris consideradas) e o número de iterações para a geração da cadeia para obtenção de estimativas para os parâmetros alfa e beta.