MÉTODOS ESTATÍSTICOS PARA A AVALIAÇÃO DA REDUNDÂNCIA EM ESTUDOS DE ETIQUETAS DE SEQUÊNCIA EXPRESSA

Fernanda Vital de PAULA[1]
Fabyano Fonseca e SILVA[2]
Carlos Souza do NASCIMENTO[3]
Simone Eliza Facioni GUIMARÃES[3]
Sebastião MARTINS FILHO[2]
Moysés NASCIMENTO[2]
Alencariano FALCÃO[1]

- RESUMO: Pesquisas de Etiquetas de Sequência Expressa (ESTs) são importantes para identificação de genes em estudos de seqüenciamento de organismos, porém na presença de elevada redundância de transcritos tal técnica torna-se inviável, pois produz poucas seqüencias que ainda não foram previamente amostradas. Dentre as medidas de redundância destaca-se o número de genes, $\Delta(t)$, que podem ser descobertos em uma amostra futura de EST t vezes maior que a amostra original. Tal estatística é útil para direcionar protocolos de sequenciamento de uma biblioteca de cDNA, uma vez que indica quando este deve ser encerrado, evitando seqüenciar material já seqüenciado e os custos inerentes a tal fato. O presente trabalho teve como objetivo apresentar os aspectos teóricos da estatística $\Delta(t)$, e propor abordagens clássica e bayesiana para a estimação da mesma. Foram utilizados dados de ESTs obtidos de duas bibliotecas de cDNA referentes ao organismo Mastigamoeba Balamuthi, e os resultados mostraram que as estimativas por intervalo obtidas para $\Delta(t)$, foram consideravelmente mais precisas quando se utilizou a abordagem bayesiana.

- PALAVRAS-CHAVE: EST; bibliotecas de cDNA, inferência bayesiana.

1. Introdução

Desde a sua introdução em Adams et al. (1991), as pesquisas de Etiqueta de Sequência Expressa (EST) têm sido uma ferramenta poderosa que permite identificar, detectar e caracterizar rapidamente genes expressos de um determinado organismo, o que faz de tal técnica um meio eficiente para descoberta de genes em projetos de sequenciamento do genoma funcional.

[1] Universidade Federal do Tocantins – UFT, Escola de Medicina Veterinária e Zootecnia, BR 153, Km 112, CEP 77800-000, Araguaína, TO. E-mail: *fernandavmat@yahoo.com.br* / *ajs.falcao@gmail.com*
[2] Universidade Federal de Viçosa - UFV, Departamento de Estatística, CEP 36571-000 Viçosa, MG. E-mail: *fabyanofonseca@ufv.br* / *martinsfilho@ufv.br* / *moysesnascim@ufv.br*
[3] Universidade Federal de Viçosa - UFV, Departamento de Zootecnia, CEP 36571-000 Viçosa, MG. E-mail: *carsouzarj@hotmail.com* / *sfacioniguima@gmail.com*

O processo de sequenciamento com a utilização de ESTs envolve a produção de bibliotecas de cDNA e em muitos casos, quando há redundância de transcritos altamente expressos nessas bibliotecas, os protocolos experimentais demandam gastos excessivos relacionados com a normalização, a qual pode ser superficialmente caracterizada como procedimentos laboratoriais planejados para uniformizar as frequências gênicas em bibliotecas de cDNA. De forma geral, espera-se que após a normalização o sequenciamento continue 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 provenientes de diferentes bibliotecas de cDNA.

Estes métodos são fundamentados em técnicas de estimação de números de espécies e, permitem acessar, por exemplo, o número esperado de novos genes, $\Delta(t)$, que podem ser descobertos em uma amostra futura de EST t vezes maior que a amostra original. A estatística $\Delta(t)$ é mais importante apresentada por Susko e Roger (2004) sob o ponto de vista prático, sendo útil para direcionar protocolos de sequenciamento de uma biblioteca de cDNA, uma vez que indica quando este deve ser encerrado, evitando sequenciar material já seqüenciado (redundância) e os custos inerentes a tal fato.

Devido à importância de $\Delta(t)$, , uma proposta interessante seria obter sua estimativa sob o ponto de vista bayesiano, uma vez que a abordagem clássica já foi executada por Susko e Roger (2004). Em termos gerais, a inferência Bayesiana configura-se como um dos principais assuntos da comunidade científica envolvida com o desenvolvimento e aplicação de procedimentos estatísticos. Segundo Gamerman e Lopes (2006) tal sucesso é decorrente grande versatilidade de tal abordagem na resolução de problemas nunca antes solucionados pelos métodos clássicos, além de possibilitar, na maioria das vezes, estimação por intervalo (intervalos de credibilidade) mais precisa que aquela proveniente da abordagem clássica (intervalos de confiança).

Diante do exposto, o principal objetivo do presente trabalho foi propor um método bayesiano de estimação do número de genes, $\Delta(t)$, esperado em uma nova amostra t vezes maior que a original. Além disso, objetivou-se também avaliá-lo em relação ao método clássico apresentado por Susko e Roger (2004). Tais metodologias foram aplicadas a dados de EST referentes a duas bibliotecas de cDNA, uma não-normalizada e outra normalizada, do protista Mastigamoeba balamuthi.

2. Material e métodos

Os dados de EST referentes descritos na Tabela 1 foram obtidos do sequenciamento de bibliotecas de cDNA não-normalizadas e uma normalizadas 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 e 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 v

	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	0,000
9	1	0,013	1	0,025
10	1	0,014	0	0,000
13	1	0,018	0	0,000
14	0	0,000	1	0,039
15	5	0,105	0	0,000
Total	715	1,000	363	1,000

De acordo com Susko e Roger (2004), tem-se a seguinte expressão para o número esperado de novos genes, $\Delta(t)$, , em uma nova amostragem:

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

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 da distribuição binomial negativa 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 apresentados na Tabela 1, 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 da Tabela 1 por meio do método dos quadrados mínimos generalizados para modelos de regressão não linear via função nls do software R (Development Core Team, 2010).

Em resumo, as estimativas dos parâmetros da distribuição binomial negativa, $\hat{\alpha}$ e $\hat{\gamma}$, juntamente como o valor de η_1, irão compor as fórmulas do estimador do número esperado de novos genes e se sua variância (Susko e Roger , 2004). Tais fórmulas são:

$$\hat{\Delta}(t)=\eta_1\hat{\alpha}^{-1}\hat{\gamma}^{-1}\left[1-(1+\hat{\gamma}t)^{-\hat{\alpha}}\right],$$

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

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 $\Delta(t)$.

Uma proposta bayesiana diz respeito à obtenção de uma distribuição a posteriori para $\Delta(t)$, sendo α e γ estimados via aplicação do teorema de Bayes:

$$P(\alpha,\gamma|x)\mu P(x|\alpha,\gamma)P(\alpha)p(\gamma),$$

em que: $P(\alpha)$ é a distribuição a priori de α e $P(\gamma)$ é a distribuição a priori de γ independentemente dos dados x, $P(x|\ \alpha,\ \gamma)$ é a distribuição dos dados amostrais dado os parâmetros α e γ (Função de Verossimilhança) e $P(\alpha,\ \gamma|\ x)$ é a distribuição dos parâmetros considerando os dados x (Distribuição a Posteriori).

Susko e Roger (2004) utilizaram o modelo binomial negativo truncado, obtido da função de probabilidade da distribuição Binomial Negativa para descrever os dados dispostos na Tabela 1, como já citado. Assim, considerando uma amostra aleatória $x_1, x_2, ..., x_n$, e independência entre tais observações, e assumindo que $x_i \sim \text{BinNeg}(\alpha,\gamma)$, a Função de Verossimilhança é dada por:

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

Sob o enfoque Bayesiano, assumiu-se que o parâmetro γ tem distribuição de probabilidade a priori Beta com parâmetros a e b e α tem distribuição de probabilidade Beta Reparametrizada no intervalo [-1,1] com parâmetros c e d. Os parâmetros a, b, c e d são denominados de hiperparâmetros, e definem a forma das distribuições de probabilidade a priori assumidas para os parâmetros de interesse. Tais distribuições são apresentadas a seguir, respectivamente para os parâmetros γ e α:

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

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

De acordo com o Teorema de Bayes, obteve-se a seguinte distribuição conjunta a posteriori:

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

da qual obtém-se as seguintes distribuições condicionais completas a posteriori para γ e para α, respectivamente:

$$P(\gamma|\alpha,x)=\gamma^{\sum_{i=1}^{N}x_i+a-1}(1-\gamma)^{Na+b-1},$$

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

Nota-se que a distribuição condicional completa de γ diz respeito à função de distribuição de probabilidade (f.d.p) de uma distribuição Beta, cujos parâmetros são designados por: $a^* = \sum_{i=1}^{N} x_i + a$ e $b^* = Na+b$. $\gamma|\alpha,x \sim B(a^*,b^*)$

Dessa forma, tem-se uma distribuição condicional completa a posteriori representada por uma distribuição de probabilidade conhecida: $\gamma \mid \alpha, x \sim B(a^*, b^*)$. Porém, o mesmo não ocorre com a distribuição condicional completa de α, a qual não se caracteriza como uma distribuição de probabilidade conhecida. Tal fato implicou na utilização do algoritmo Metropolis-Hastings (MH) para amostrar valores da distribuição marginal a posteriori de α, e na utilização do algoritmo Gibbs Sampler para amostrar valores da distribuição marginal a posteriori de γ.

Na implementação do algoritmo Metropolis-Hastings, os valores gerados para α foram valores provenientes de uma distribuição candidata, dada por 2B(c,d) -1. Tal distribuição foi usada porque proporcionou uma taxa de aceitação de valores candidatos 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 uma cadeia de 20.000 iterações, das quais 5.000 foram eliminadas (*burn-in*), restando 15.000 valores, dos quais amostrou-se seqüencialmente apenas um de cada cinco valores (*thin*). Este procedimento refletiu em uma cadeia final com 3.000 iterações, sendo a constatação da convergência (verificação do tamanho ideal da cadeia) realizada por meio dos critérios de Geweke (1992) e de Raftery e Lewis (1992), ambos disponíveis no pacote *BOA – Bayesian Output Analysis Program* (SMITH, 2007) do software R.

Uma vez obtidos os 3.00 valores para α e γ, ou seja, amostras de suas respectivas distribuições marginais a posteriori, obteve-se indiretamente as amostras da distribuição marginal a posteriori para a estatística Δ(t) da seguinte maneira:

$$\Delta(t)^{(k)} = \eta_1 \alpha^{-1^{(k)}} \gamma^{-1^{(k)}} \left[1-(1+\gamma^{(k)}t)^{-\alpha^{(k)}}\right],$$

sendo k a indicação de cada um dos 3.000 valores de α e γ usados para obter Δ(t).

3 Resultados e discussão

Os parâmetros estimados pela abordagem frequentista de Susko e Roger (2004), foram $\hat{\alpha} = -0,778$ e $\hat{\gamma} = 0,944$ para a biblioteca não-normalizada e $\hat{\alpha} = -0,715$ e $\hat{\gamma} = 0,889$ para a biblioteca normalizada.

Tais valores foram, então, substituídos na fórmula de Δ(t) a fim de obter as estimativas do número esperado de genes considerando t=0; 0,1; 0,2;...; 1. Assim, foi possível construir o gráfico da Figura 1 cujo eixo x representa o tamanho da nova amostra (valores de nt) e o eixo y os valores estimados de Δ(t). Vale ressaltar que os intervalos de confiança (assumindo Normalidade assintótica) mostrados neste mesmo gráfico foram obtidos por meio da especificação da estimativa da variância de Δ(t) apresentada anteriormente.

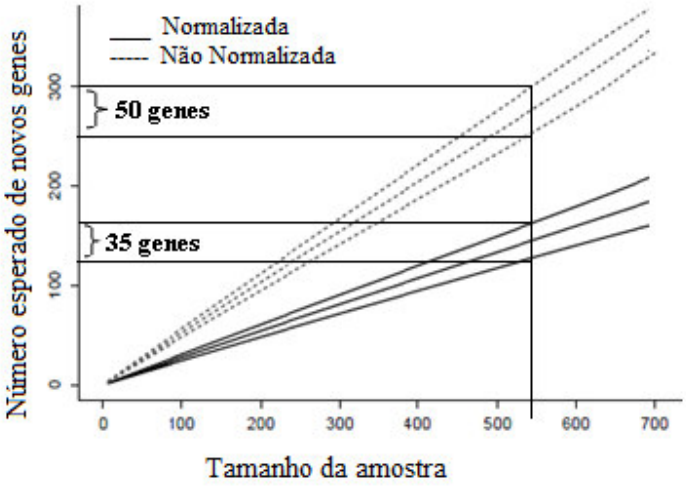

Figura 1 - Estimativa do número esperado de genes como uma função do tamanho da amostra considerando a abordagem freqüentista proposta por Susko e Roger (2004). A linha central representa a estimativa pontual à e as linhas em torno da mesma representa o intervalo de confiança assintótico de 95%.

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, como se pode observar na Figura 1, 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. 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.

Para a obtenção de α e γ pelo procedimento bayesiano, as estimativas de tais parâmetros foram obtidas a partir das médias das distribuições marginais a posteriori que por sua vez foram obtidas pelos métodos MCMC (Markov Chain Monte Carlo).

Na Tabela 2 são apresentados os resultados referentes a avaliação da convergência das cadeias geradas pelos algoritmos MCMC. Nota-se pelo teste de Geweke ($P_G > 0,05$, sendo H_0:convergência da cadeia) que as cadeias convergiram, e nota-se pelo critério de Raftery e Lewis que o tamanho efetivo das cadeias utilizadas foi suficiente para assegurar a convergência, pois este indicou cadeias de tamanho de 11.931 e 18.480 iterações para as bibliotecas não-normalizada e normalizada, respectivamente, sendo esses valores menores que o valor usado de 20.000 iterações. Os valores de *burn-in* indicados por este critério também foram inferiores que o valor de 5.000 iterações usado.

Tabela 2 - Diagnósticos de convergência de Geweke e Raftery e Lewis para as cadeias de valores gerados via MCMC para α e γ considerando 20.000 iterações, com burn-in de 5.000 e thin de 5 iterações

	Biblioteca não-normalizada				Biblioteca normalizada			
Par.	Geweke (P_G)	Niter	Burn-in	Thin	Geweke (P_G)	Niter	Burn-in	Thin
α	0, 8041 (0, 4213)	11.931	9	3	0, 3998 (0, 6893)	18.480	15	5
γ	0, 1412 (0, 8877)	3.795	2	1	−1, 0577 (0, 2902)	3.811	2	1

Par.: Parâmetro.

Na Tabela 3 são apresentadas as estimativas dos parâmetros α e γ obtidas via inferência bayesiana.

Tabela 3 - Média, desvio-Padrão, limites inferiores e superiores do intervalo de credibilidade de 95% para α e γ considerando as cadeias geradas (com tamanha efetivo de 3.000 iterações) para as bibliotecas não-normalizada e normalizada

	Biblioteca não-normalizada				Biblioteca normalizada			
Parâmetro	Média	DP	Linf	Lsup	Média	DP	Linf	Lsup
α	-0, 743	0, 039	-0, 813	-0, 6665	-0, 753	0, 037	-0, 826	-0, 681
γ	0, 997	0, 001	0, 993	0, 999	0, 992	0, 004	0, 984	0, 999

Por meio da utilização dos 3.000 valores considerados de α e γ foi possível obter o mesmo número de valores (amostras da distribuição marginal a posteriori) para o parâmetro $\hat{\Delta}(t)$ (número esperado de novos genes). Os resultados referentes a este parâmetro podem ser visualizados na Figura 2.

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

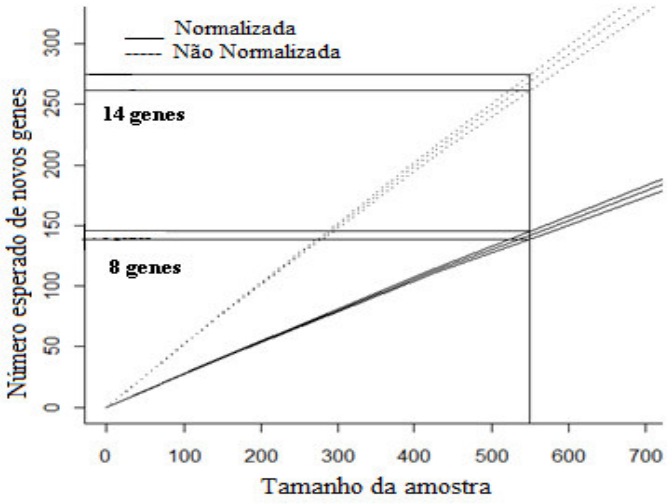

Figura 2 - Estimativa do número esperado de genes como uma função do tamanho da amostra considerando a abordagem bayesiana. A linha central representa a estimativa pontual à e as linhas em torno da mesma representa o intervalo de credibilidade de 95%.

De forma geral, as estimativas pontuais de α e γ (Tabela 3) foram substancialmente semelhantes aquelas obtidas pelo método frequentista, sendo a maior diferença obtida para o parâmetro γ, cujas estimativas para a biblioteca normalizada foram, respectivamente, 0,8890 e 0,992 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 clássica. De forma geral, esta melhora pode ser observada nas Figuras 3 e 4.

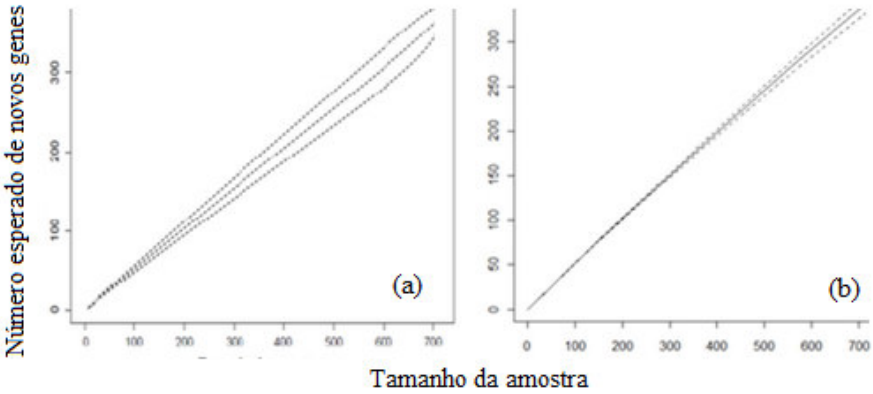

Figura 3 - Estimativas do número esperado de novos genes em função do tamanho da amostra obtidas pelas metodologias frequentista (a) e bayesiana (b) considerando a biblioteca não-normalizada.

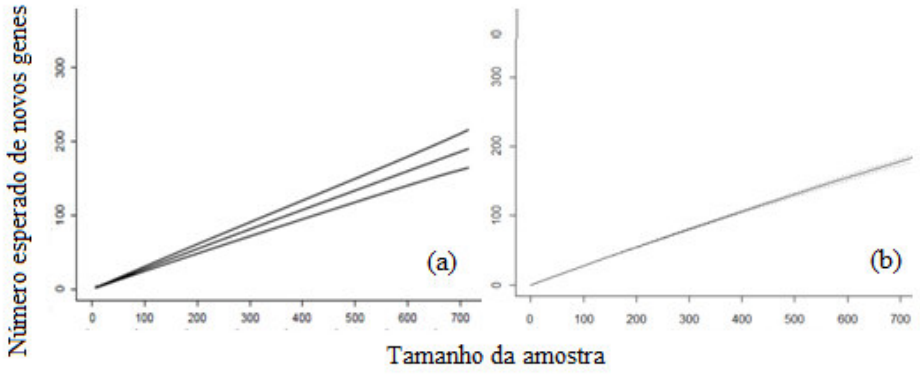

Figura 4 - Estimativas do número esperado de novos genes em função do tamanho da amostra obtidas pelas metodologias frequentista (a) e bayesiana (b) considerando a biblioteca normalizada.

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 estes autores destacam-se Silva et al. (2008) 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 em tempos futuros.

Conclusões

1. A abordagem bayesiana para a estimação de $\Delta(t)$, mostra-se eficiente sendo útil para direcionar protocolos de sequenciamento de uma biblioteca de cDNA.
2. Os resultados apresentados mostram que as estimativas por intervalo obtidas para $\Delta(t)$ foram consideravelmente mais precisas quando se utilizou a abordagem bayesiana.

PAULA; F. V.; SILVA, F. F.; NASCIMENTO, C. S. do; GUIMARÃES. S. E. F.; MARTINS FILHO, S.; NASCIMENTO, M. Statistical methods to evaluate redundancy in Expressed sequence tags studies. *Rev. Bas. Biom.*, São Paulo, v.29, n3, p.462-471, 2011.

- *ABSTRACT: Expressed sequence tags (ESTs) surveys are very important to identify genes in sequencing studies of various organisms, but in the presence of high transcript redundancy rate this technique is unviable, since it produces few sequences that were not previously sampled. One of the most relevant redundancy measures is the number of genes, Δ (t), which may be discovered in a future EST sample t times larger than the original sample. This statistics is useful to direct sequencing protocols of cDNA libraries, since it indicates when this process must be closed, avoiding to sequence stuff already sequenced and the costs related with this one. The present work had as objective to present the theoretical aspects of the statistic Δ(t) and to propose a classical and bayesian approach for its estimation. Were used data from two cDNA libraries from Mastigamoeba Balamuthi organism, and the results showed that the interval estimates obtained for Δ (t) were significantly more accurate when the Bayesian inference was used.*

- *KEYWORDS: EST; cDNA libraries; bayesian inference.*

Referências

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.; MCCOMBIE, W. R.; VENTER, V. C. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, v.252, p.1651–1656, 1991.

BLASCO, A.; PILES, M.; VARONA, L. Bayesian analysis of the effect of selection for growth rate on growth curves in rabbits. *Genetics Selection Evolution*, v.35, n.1, p.21-41, 2003.

GAMERMAN, D.; LOPES, H.F. *Markov chain Monte Carlo*: stochastic simulation for Bayesian inference. 2.ed. New York: Chapman e Hall, 2006. 324p.

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: BERNARDO, J. M. et al. *Bayesian statistics 4***.** Oxford: Oxford University, 1992, p.625-631.

HUANG, X.; MADAN, A. CAP3: A DNA sequence assembly program. *Genome Research*, v.9, p. 868-877, 1999.

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

RAFTERY, A. E.; LEWIS, S. How many iterations in the Gibbs sampler? In: BERNARDO, J. M. et al. *Bayesian statistics 4***.** Oxford: Oxford University, 1992. p.763-773.

SILVA, F. F. , SÁFADI, T. , MUNIZ, J.A. , ROSA, G. J. M. , AQUINO, L.H. , MOURÃO, G.B. , SILVA, C. H. O. 2011. Bayesian analysis of autoregressive panel data model: application in genetic evaluation of beef cattle. *Scientia Agrícola*. v.68, p.237-245, 2011.

SILVA, F. F.; SÁFADI, T.; MUNIZ, J.A.; AQUINO, L.H.; MOURÃO, G.B. Comparação bayesiana de modelos de previsão de diferenças esperadas nas progênies no melhoramento genético de gado Nelore. *Pesquisa Agropecuária Brasileira*, v.43, p.37-45, 2008.

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.

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

Recebido em 21.06.2010.

Aprovado após revisão 21.09.2010.