CADEIAS DE MARKOV COM ESTADOS LATENTES COM APLICAÇÕES EM ANÁLISES DE SEQÜÊNCIAS DE DNA

Deive Ciro de OLIVEIRA[1]
Cibele Queiroz da SILVA[2]
Lucas Monteiro CHAVES[3]

- RESUMO: Neste trabalho a teoria das cadeias de Markov com estados latentes, HMM's (*Hidden Markov Models*), é aplicada ao problema de discriminação de regiões homogêneas em seqüências de DNA. Utiliza-se o algoritmo EM na obtenção de estimativas de máxima verossimilhança dos parâmetros associados ao modelo. Foram analisados trechos do genoma das bactérias *Xylella fastidiosa*, *Xanthomonas axonopodis* pv. *citri*, *Streptococcus pneumoniae* e *Escherichia coli*. Os resultados são discutidos e comparados com as organizações reais dos genes, de acordo com a classificação COG (Cluster of Orthologous Groups).
- PALAVRAS-CHAVE: Processos estocásticos; cadeias de Markov com estados latentes; análise de DNA.

1 Introdução

Os ácidos nucléicos têm papel fundamental no metabolismo dos seres vivos. Entre os ácidos nucléicos mais importantes estão o *DNA* (ácido desoxi ribonucléico) e *RNA* (ácido ribonucléico). A estrutura fundamental dos ácidos nucléicos é composta de monômeros (moléculas similares), denominados nucleotídeos. Os nucleotídeos são distintos pela presença de diferentes bases nitrogenadas. As bases são Adenina (A), Citosina (C), Guanina (G) e Timina (T – Ocorrência no *DNA*) ou Uracila (U – Ocorrência no RNA). Vamos tratar, neste trabalho, especificamente da análise de seqüências de *DNA*.

O *DNA* apresenta forma helicoidal (dupla hélice) e a característica de complementaridade das bases, sendo importantíssimo no processo de síntese de proteínas (Processos de *Transcrição* e *Tradução*). A função vital do *DNA* nesta síntese se deve ao fato dele codificar as proteínas que serão sintetizadas. As proteínas são moléculas que atuam na regulação das vias metabólicas dos organismos e algumas organelas. Sendo assim, falhas na seqüência de *DNA* podem causar distúrbios graves na regulação do metabolismo. O esquema básico da síntese protéica é ilustrado na Figura 1.

1Fundação Educacional de Oliveira, CEP 35540-000, Oliveira, MG, Brasil. E-mail: deiveufla@ig.com.br
[2]Departamento de Estatística, Universidade de Brasília - UNb, CEP 70910-900, Brasília, DF, Brasil. E-mail: *cibeleqs@unb.br*
[3]Departamento de Ciências Exatas, Universidade Federal de Lavras - UFLA, CEP 37200-000, Lavras, MG, Brasil. E-mail: *lucas@ufla.br*

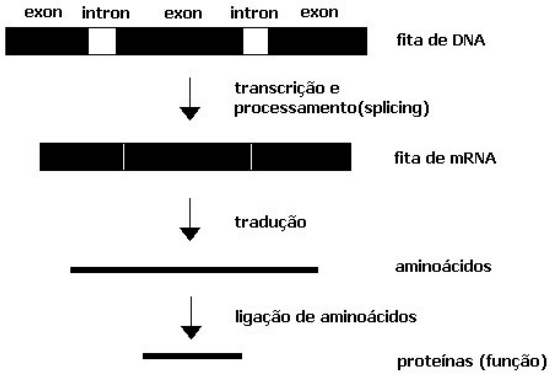

Figura 1 - Ilustração resumida do processo de Síntese Protéica.

Pode-se observar, na Figura 1, que nem toda a seqüência de *DNA* codifica uma proteína, ocorrendo a presença de trechos que não atuam na codificação. Essas regiões são denominadas íntrons. As regiões codificadoras são denominadas éxons. Observe que a seqüência de *DNA* presente no organismo é linear (desconsidera-se, para efeito de análise, a seqüência complementar), na qual trechos específicos codificam proteínas específicas. Uma questão natural é como localizar os trechos que codificam diferentes proteínas, ou de maneira mais genérica, localizar trechos que codificam proteínas com o mesmo tipo de funcionalidade. Um fator complicador é a presença dos íntrons, que não codificam proteínas.

Uma particularidade importante de regiões ou trechos de *DNA* que codificam proteínas com funções similares, é que essas regiões podem apresentar expressões similares de *C+G* ("*isochores*") (Churchill, 1989), isto é, segmentos onde as proporções da ocorrência de uma base ou de um determinado grupo de bases são similares. Assim, o problema da discriminação de trechos que codificam proteínas similares reduz-se a encontrar regiões com proporções homogêneas (similares) de *C+G*. Este problema, muitas vezes, é denominado *Segmentação de DNA* (Boys e Henderson, 2004) (Boys et al., 2000).

A estratégia para a resolução do problema é considerar a seqüência de *DNA* como um processo aleatório. Pode-se abordar o problema utilizando a inferência sobre o *Ponto de mudança em seqüências de Variáveis Aleatórias* (Hinkley e Hinkley, 1970), (Smith, 1975). A dificuldade, em tal abordagem, ocorre em razão da existência da dependência entre as bases na seqüência de *DNA*. Existem ainda algumas técnicas baseadas em testes de hipóteses para a detecção de diferenças entre proporções (Mood; Graybii e Boes, 1963) e inspeção seqüencial de proporções (Staden, 1984) que podem ser utilizadas. Essas técnicas apresentam escolhas inicias arbitrárias ou imprecisões na distinção das regiões com comportamento homogêneo. Um modelo que permite capturar as dependências entre as bases de uma seqüência de *DNA*, e discriminar claramente as distinções entre as regiões funcionalmente similares é denominado cadeias de Markov com estados latentes (Rabiner, 1989), (da-Silva, 2002), (Oliveira, 2005).

2 Cadeias de Markov com estados latentes

As cadeias de Markov com estados latentes (*HMM – Hidden Markov Models*) são modelos compostos por dois processos estocásticos. Um *processo estocástico* pode ser definido como um conjunto de variáveis aleatórias indexadas que possuem o mesmo domínio (espaço amostral), cuja distribuição de probabilidade conjunta é conhecida. Normalmente, o índice das variáveis é associado a unidades de tempo ou espaço. Este índice pode assumir valores discretos ou contínuos. No caso da aplicação em questão, os índices irão assumir valores discretos.

Uma característica dos Modelos de Markov com Estados Latentes (*HMM*) é o pressuposto da existência de dois processos estocásticos interdependentes. Um dos processos é *latente*, isto é, não observável $\{S\} = \{S_1, S_2, S_3..., S_t, ...\}$, e o outro é *observável*, $\{O\} = \{O_1, O_2, O_3..., O_t, ...\}$. O processo não observável é modelado segundo uma Cadeia de Markov. Uma Cadeia de Markov de ordem 1 é um tipo específico de processo estocástico onde a probabilidade de ocorrência da variável aleatória S_t, dada a ocorrência dos S_k´s com $k < t$, é dependente apenas da variável aleatória S_{t-1}. Caso a cadeia fosse de ordem 2, S_t dependeria apenas de S_{t-1} e S_{t-2}. Dependências mais longas podem ser adotadas, porém será utilizada a cadeia de ordem 1 para proceder as análises. Pode-se modelar o processo estocástico observável, com ou sem a presença de dependência, entre as variáveis aleatórias $\{O\} = \{O_1, O_2, O_3..., O_t, ...\}$. De acordo com a especificação do *HMM* e com a aplicação é possível assumir relações complexas de dependência, bem como independência entre as variáveis aleatórias do processo estocástico $\{O\} = \{O_1, O_2, O_3..., O_t, ...\}$. As variáveis aleatórias dos processos observável e latente, com mesmo índice, respectivamente O_t e S_t, são dependentes. Neste trabalho considera-se o caso onde as variáveis condicionais $O_t \mid S_t$ são independentes $\forall\, t$, sendo S_t e O_t variáveis aleatórias discretas (espaço amostral discreto). Os elementos do contradomínio de S_t são chamados de *estados do processo*, enquanto os elementos do contradomínio de O_t são denominados *observações do processo*. Os parâmetros associados a um *HMM* denotado por $\lambda = (\pi, A, B)$ são:

$$\pi_{s_1} = P(S_1 = s_1); \tag{2.1}$$

$$a_{s_{t-1}s_t} = P(S_t = s_t \mid S_{t-1} = s_{t-1}); \tag{2.2}$$

$$b_{s_t o_t} = P(O_t = o_t \mid S_t = s_t). \tag{2.3}$$

O parâmetro π_{s_1} representa a probabilidade de, no início do processo, a variável latente assumir o estado s_1. O parâmetro $a_{s_{t-1}s_t}$ caracteriza a dependência de ordem 1 da Cadeia de Markov no processo estocástico latente. Por fim, a interdependência entre os processos latente e observável é estabelecida pelo parâmetro $b_{s_t o_t}$.

A modelagem do problema de discriminação de regiões homogêneas em seqüências de *DNA*, via utilização de *HMM*´s, consiste em se considerar a seqüência de *DNA* como a realização de um processo estocástico observável $\{O_t\} = \{O_1, O_2, O_3..., O_t, ...\}$. Isto implica,

necessariamente, que o espaço amostral das variáveis O_t esteja restrito ao conjunto $\{A,G,T,C\}$. No entanto, não estamos interessados nas freqüências individuais de cada uma das bases, e sim nas freqüências das ocorrências de *C*+*G* e *A*+*T*. O espaço amostral a ser modelado terá dois possíveis resultados, que são, 1 se ocorrer *C* ou *G* e 0 se ocorrer *A* ou *T*. Assim, o espaço amostral do processo estocástico observável é $\Omega_O = \{0,1\}$. As variáveis S_t, do processo estocástico latente auxiliarão na descrição dos segmentos com diferentes proporções de *C*+*G*. A descrição do domínio Ω_S (conjunto dos estados do processo estocástico latente) de S_t requer a especificação do número de regiões distintas, em freqüência de *C*+*G*, que estão presentes ao longo da seqüência analisada. Tal informação é desconhecida, visto que o processo S_t é latente. Por isso, especificaremos o conjunto Ω_S com vários possíveis números de estados distintos e, finalmente, por um método de adequação de modelos, adotaremos o mais adequado.

Existem três problemas básicos associados aos *HMM´*s. O primeiro deles (*problema 1*) é a partir de uma seqüência de realizações do processo estocástico observável $O = \{O_1 = o_1, O_2 = o_2, O_3 = o_3, ..., O_T = o_T\}$ e da especificação de um *HMM* λ – calcular a verossimilhança $P(O \mid \lambda)$. O segundo (*problema 2*) é caracterizado pela obtenção da seqüência mais provável de estados $s_1, s_2, s_3, ...s_T$, referente ao processo estocástico latente, dada a realização do processo estocástico observável $O_1 = o_1, O_2 = o_2, O_3 = o_3, ..., O_T = o_T$, onde T representa o tamanho da seqüência observada. O terceiro problema (*problema 3*) é caracterizado pela obtenção de estimativas dos parâmetros do *HMM* λ, dada uma realização do processo observável O.

Para a resolução do *problema 1* existem dois métodos, baseados em programação dinâmica, equivalentes, denominados *forward* e *backward*. Vamos apresentar os dois, visto que no *problema 3* utilizaremos algumas quantidades calculadas na solução do *problema 1*. Para se obter $P(O \mid \lambda)$ pelo método *forward*, inicialmente é necessário obter uma matriz α cuja construção se dá através das relações de recorrência explicitadas pela expressão (2.4).

$$\alpha_{(t,s_t)} = \begin{cases} b_{s_t o_t} \times \pi_{s_t}, & \text{se } t = 1; \\ b_{s_t o_t} \times \sum_{s_{t-1} \in \Omega_s} \left[a_{s_{t-1} s_t} \times \alpha_{(t-1, s_{t-1})} \right], & \text{se } 1 < t \leq T. \end{cases} \tag{2.4}$$

Observe em (2.4) que T é o tamanho da seqüência observada, enquanto o_t representa a realização da *t*-ésima variável O_t do processo observável. Para se obter $P(O \mid \lambda)$, após a construção da matriz α, basta calcular o somatório da última linha de α dada pela expressão (2.4). Portanto, considerando o modelo λ conhecido, a verossimilhança de interesse é dada por:

$$P(O \mid \lambda) = \sum_{s_T \in \Omega_s} \alpha_{(T, s_T)}. \tag{2.5}$$

Pode-se obter $P(O \mid \lambda)$ utilizando-se o método *backward*. Isto é feito com a construção da matriz β que é calculada a partir das relações de recorrência expressas em (2.6):

$$\beta_{(t,s_t)} = \begin{cases} 1, \text{ se } t = T; \\ \sum_{s_{t+1} \in \Omega_s} b_{o_{t+1} s_{t+1}} \times \beta_{(t+1, s_{t+1})} \times a_{s_t s_{t+1}}, \text{ se } 1 \leq t < \text{T}. \end{cases} \tag{2.6}$$

A verossimilhança $P(O \mid \lambda)$ é obtida pelo somatório da primeira linha da matriz β, isto é:

$$P(O \mid \lambda) = \sum_{s_1 \in \Omega_s} \left[b_{o_1 s_1} \times \beta_{(1,s_1)} \times \pi_{s_1} \right]. \tag{2.7}$$

Nos casos em que as seqüências de realizações do processo estocástico observável apresentam grandes dimensões, e dependendo do recurso computacional empregado na implementação do método, é necessário utilizar uma técnica adicional de normalização das matrizes α e β para que seja possível o cálculo de $P(O \mid \lambda)$. Tal técnica esta descrita em (Rabiner, 1989) e (Oliveira, 2005).

O *problema 2* é solucionado com a aplicação do algoritmo de Viterbi (da-Silva, 2002). A obtenção da seqüência mais provável de estados latentes não é um passo essencial para a discriminação das regiões com distintas freqüências de *C+G*. Porém, tal conhecimento pode ser utilizado para a visualização destas regiões. No algoritmo de Viterbi, tal como nos métodos *forward* e *backward*, também se trabalha com a manipulação de matrizes. Ao final do processo, obtém-se uma seqüência de estados $S = \{S_1 = s_1, S_2 = s_2, S_3 = s_3, ..., S_T = s_T\}$ cuja verossimilhança conjunta, $P(O, S \mid \lambda)$, é máxima (Rabiner, 1989), (da-Silva, 2003), (Oliveira, 2005).

O *problema 3* trata da inferência sobre o modelo λ. Existem diversos métodos para a obtenção de estimativas dos parâmetros presentes em λ. Como exemplo, podemos citar o método de mínimos quadrados, métodos dos momentos e o método bayesiano. Neste trabalho adotamos o Método da Máxima Verossimilhança. Este consiste em se obter a estimativa $\hat{\lambda}$ de λ tal que $l(\hat{\lambda}) = P(O \mid \hat{\lambda})$ seja máxima. Como na estimação de λ o estimador de máxima verossimilhança não pode ser obtido de forma fechada, faz-se necessário a utilização de métodos numéricos. Dessa forma, na obtenção de $\hat{\lambda}$, utiliza-se, neste trabalho, o método iterativo *EM – Expectation-Maximization Algorithm* (Dempster; Laird e Rubin, 1977). Este método é baseado na obtenção de estimativas $\hat{\lambda}^{(l)}$ de λ, sendo $\hat{\lambda}^{(l)}$ a estimativa no l-ésimo passo iterativo. Por construção do algoritmo, temos que $P(O \mid \hat{\lambda}^{(l+1)}) \geq P(O \mid \hat{\lambda}^{(l)})$. O método deve ser iniciado a partir de uma estimativa inicial $\hat{\lambda}^{(0)}$. Considerando a função $I(O_t, k) = 1$ quando $O_t = k$ e $I(O_t, k) = 0$ quando $O_t \neq k$, as estimativas $\hat{\lambda} = (\hat{\pi}, \hat{A}, \hat{B})$, a cada passo iterativo do algoritmo *EM*, devem ser calculadas de acordo com as equações explicitadas em (2.8) e (2.9):

$$\hat{\gamma}_{t(i)}^{(l)} = \frac{\hat{\beta}_{(t,i)}^{(l)} \text{ x } \hat{\alpha}_{(t,i)}^{(l)}}{P(O \mid \hat{\lambda}^{(l)})}, \quad \hat{\xi}_{t(i,j)}^{(l)} = \frac{\hat{\beta}_{(t,j)}^{(l)} \text{ x } \hat{b}_{j0_t}^{(l)} \text{ x } \hat{a}_{ij}^{(l)} \text{ x } \hat{\alpha}_{(t-1,i)}^{(l)}}{P(O \mid \hat{\lambda}^{(l)})} \tag{2.8}$$

$$\hat{\pi}_i^{(l+1)} = \hat{\gamma}_{1(i)}^{(l)}, \quad \hat{a}_{ij}^{(l+1)} = \frac{\sum_{t=2}^{T} \hat{\xi}_{t(i,j)}^{(l)}}{\sum_{t=2}^{T} \hat{\gamma}_{t(i)}^{(l)}}, \quad \hat{b}_{ik}^{(l+1)} = \frac{\sum_{t=2}^{T} \hat{\gamma}_{t(i)}^{(l)} \; x \, I(O_t, k)}{\sum_{t=2}^{T} \hat{\gamma}_{t(i)}^{(l)}} \tag{2.9}$$

Como o algoritmo *EM* é um método iterativo, um critério de convergência deve ser estabelecido. Em geral para $\varepsilon \geq 0$ fixo, o procedimento deve ser iterado enquanto $P(O \mid \hat{\lambda}^{(l+1)}) - P(O \mid \hat{\lambda}^{(l)}) \geq \varepsilon$. Como nos métodos *forward* e *backward*, é necessário utilizar técnicas de normalização em γ e ξ. Tais técnicas são apresentadas em (Rabiner, 1989) e (Oliveira, 2005). Após a obtenção da estimativa $\hat{\lambda}$ de λ, adotando-se modelos com diferentes números de estados, ou equivalentemente, com diferentes domínios para Ω_s, devemos por meio de um critério, escolher o modelo mais adequado. Existem critérios de escolha de modelos, que de modo a se obter um equilíbrio entre vício e variabilidade na estimação dos parâmetros, consideram o "custo" e o "benefício" proporcionados pelo modelo. Dentre estes, podem ser citados o *AIC – Akaike´s Information Criterion* (Sakamoto, 1978), o *BIC – Bayesian Information Criterion* (Schwarz, 1978) e o ΔBIC (Churchill, 1992). Sendo $\hat{\lambda}$ o estimador de máxima verossimilhança, K os graus de liberdade do modelo em questão, T o tamanho da seqüência de observações e n o número de elementos de Ω_o (contradomínio de O_t´s), os valores dos respectivos critérios são dados pelas equações expressas em (2.10):

$$\begin{aligned} AIC &= -2\ln(P(O \mid \hat{\lambda})) + 2K, \\ BIC &= -2\ln(P(O \mid \hat{\lambda})) + K\ln(T), \\ \Delta BIC &= \left[-2\ln(P(O \mid \hat{\lambda})) + K\ln(T)\right] - \left[2 \times T \times \ln(n)\right]. \end{aligned} \tag{2.10}$$

Calculando-se o valores associados a cada modelo, aquele que minimizar o *BIC*, ΔBIC e *AIC* é o mais adequado, segundo o respectivo critério.

Após a obtenção das estimativas e seleção do modelo mais adequado, devemos utilizar ferramentas gráficas, que auxiliarão na discriminação das regiões com diferentes características em termos de freqüência de *C+G*. Uma dessas ferramentas é o gráfico da *composição local*, construído a partir das estimativas $\hat{\lambda}$. Este gráfico é construído a partir de valores que representam a esperança da probabilidade estimada de ocorrência de uma base *C* ou *G* em uma posição t da seqüência. O gráfico terá, portanto, o eixo ordenado

variando entre 0 e 1, enquanto o eixo das abscissas terá variação discreta com o tamanho T. A partir do modelo estimado, os valores para a construção do gráfico estão ilustrados em (2.11):

$$E_{S_t|O,\hat{\lambda}}\left[P(O_t = k \mid S_t, \hat{\lambda})\right] = \sum_{i \in \Omega_S} \hat{b}_{ik} \times \hat{\gamma}_{t(i)}. \qquad (2.11)$$

O gráfico da *composição local*, cujos pontos são constituídos dos pares $\left(t, E_{S_t|O,\lambda}\left[P(O_t = k \mid S_t, \hat{\lambda})\right]\right)$, explicitará as regiões com características homogêneas em proporção de *k* (*k* equivale a C ou G). Para identificarmos regiões homogêneas, basta observar trechos onde ocorre baixa variabilidade da *composição local*. Quando houver mudanças bruscas de padrão de emissão de *C+G*, é possível que isto seja devido à presença de distintas regiões do *DNA* responsáveis por diferentes características funcionais.

Uma outra forma de discriminar as diferentes regiões em proporção de *C+G* e, portanto, capturar diferenças funcionais entre os segmentos, é analisar a seqüência de estados do processo latente mais provável (obtenção pelo algoritmo de Viterbi).

3 Resultados

Nesta seção analisamos fragmentos de seqüências de *DNA* de quatro organismos (*Xylella fastidiosa, Xanthomonas axonopodis* pv. citri, *Streptococcus Pneumoniae, Escherichia coli*) na tentativa de evidenciar a utilidade dos *HMM´s* na discriminação de regiões que indiquem possíveis diferenças em funcionalidade protéica. Para tanto, comparamos os resultados obtidos com a classificação *COG* (*Cluster of Orthologous Group*) dos genes, disponível no Genbank (www.ncbi.nih.gov). Tal classificação discrimina genes que codificam proteínas com atividades distintas. Todas as implementações dos métodos foram realizadas no ambiente Delphi. Passamos, a seguir, a relatar as análises para cada um dos organismos em estudo.

3.1 *Xylella fastidiosa*

A *Xylella fastidiosa* é uma bactéria patógena que ataca culturas cítricas, sendo de grande importância econômica (Simpson et al., 2000). O trabalho de da-Silva (2003) apresenta a análise de regiões homogêneas sobre os éxons da seqüência de *DNA* associado ao segmento que se inicia no gene *XF1141* e finaliza no gene *XF1196*. Procuramos, aqui, analisar a mesma seqüência, considerando íntrons e éxons, presentes entre os genes *XF1141* e *XF1196*. No total, este segmento possui 50.716 bases. A seqüência foi obtida a partir do *Genbank,* sob código AE003849. Foram obtidos os modelos ajustados, pelo método da máxima verossimilhança, considerando-se 2, 3, 4 e 5 estados. Na Tabela 1 apresenta-se os valores das funções *AIC*, *BIC,* ΔBIC e os graus de liberdade associados a cada modelo.

Tabela 1 - Valores de graus de liberdade, BIC, ΔBIC e AIC associados á seqüência *Xylella fastidiosa*

Nº de estados	Grau de liberdade	BIC	ΔBIC	AIC
2	5	68.640,542	-1.667	68.596,442
3	12	68.614,233	-1.693	68.517,215
4	21	68.694,287	-1.613	68.526,711
5	32	68.788,308	-1.519	68.532,534

Com base em todos os critérios, observa-se que o modelo com 3 estados é o que melhor descreve os dados. Pelo gráfico da *composição local* de *C+G* (Figura 2), associado ao modelo com 3 estados, é possível observar segmentos homogêneos na expressão de *C+G*.

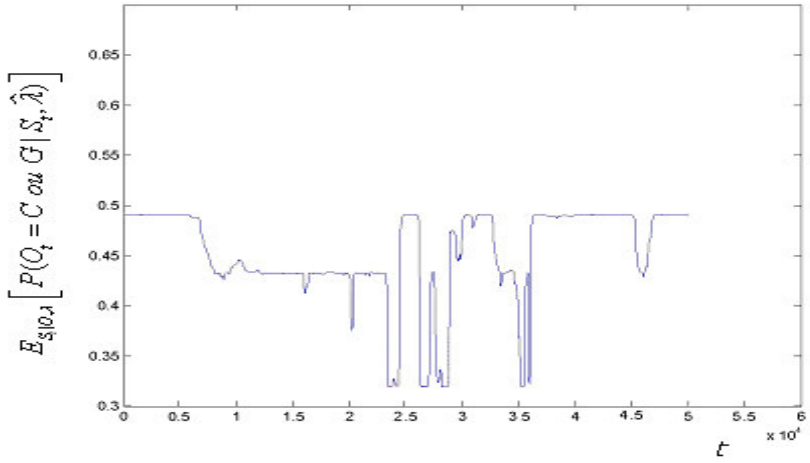

Figura 2 - Gráfico da composição local de C+G respectivo à seqüência da *Xylella fastidiosa*, sob o modelo com 3 estados.

Analisando-se o perfil descrito pelo gráfico da *composição local*, é possível discriminar 4 trechos caracterizados pelo comportamento homogêneo ou heterogêneo na ocorrência de *C+G*. Um trecho inicial que vai da primeira base até, aproximadamente, a base 7.000, apresenta uma característica bem homogênea (pouca variação) em relação ao conteúdo de *C+G*, sendo esta em torno de 0,48. É possível observar ainda um segundo trecho homogêneo que se inicia, aproximadamente, na base 9.000 e vai até a base 22.000, e os valores da *composição local* variam em torno de 0,42. Embora exista a presença de homogeneidade neste trecho, ela não é tão evidente quanto no primeiro, devido à presença de algumas poucas variações muito bruscas ao longo da seqüência. O terceiro trecho apresenta uma característica heterogênea (variação alta nos valores da *composição local* de *C+G*) em relação ao conteúdo de *C+G*. Este trecho varia, aproximadamente, da base na posição 24.000 até a base na posição 36.000. No quarto e último trecho é possível verificar uma característica homogênea, embora exista uma variação considerável por volta da base 46.000. Os valores da *composição local* neste trecho são próximos de 0,48.

Todas estas observações sugerem que o segmento analisado é descrito como 4 trechos (e 3 estados) com características distintas. O primeiro dos trechos está localizado entre as bases 1 e 7.000 (permanência no estado 3) e o segundo está localizado da base 7.000 até 23.000 (permanência no estado 2). Um terceiro trecho possui uma característica altamente heterogênea (alternância entre estados 1, 2 e 3) em termos de ocorrência de *C+G*, agregando pequenas regiões de funcionalidades diversas (relativo ao tamanho da seqüência). Este terceiro trecho está localizado, aproximadamente, a partir da base 23.000 até a base 36.000. A análise sugere a presença de outra região funcional a partir da base 36.000 (permanência no estado 3).

Se observarmos a estrutura deste segmento (genes *XF1141* até *XF1196*), segundo os mapas funcionais e classificação de proteínas similares (COG´s) funcional ou estruturalmente (Figura 3), podemos notar que o *HMM* proporcionou boa discriminação das distintas regiões funcionais. Na Figura 3, adaptada do *Genbank*, apresenta-se uma ilustração gráfica das regiões funcionais no segmento analisado.

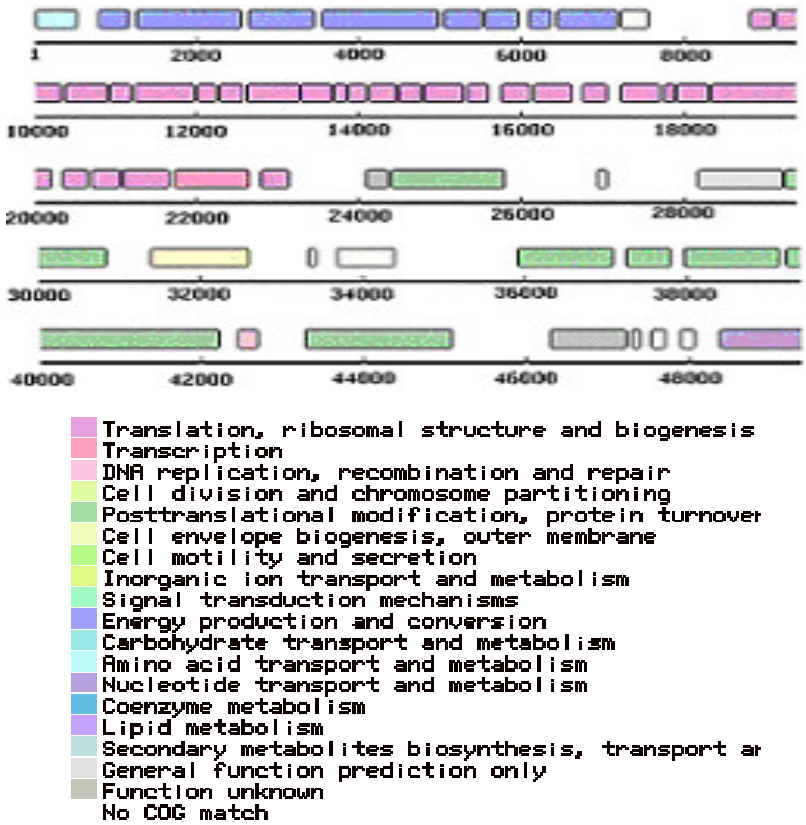

Figura 3 - Regiões de expressão do gene *XF1141* até o gene *XF1196* da *Xylella fastidiosa* e legenda de classificação funcional.

Os resultados obtidos foram coerentes com os resultados encontrados em da-Silva (2003), onde 4 trechos foram discriminados. Os dois primeiros trechos (1 a 7.000 e 7.000 a 21.000) são largamente homogêneos. Por possuir poucas regiões não codificadoras, os dois primeiros trechos são caracterizados de maneira similar ao presente trabalho.

3.2 *Xanthomonas axonopodis* pv. *citri*

O gênero *Xanthomonas* é um importante grupo de bactérias fitopatogênicas. Sua importância se deve ao fato de causar doenças em culturas de relevância econômica em todo o mundo. O *Xanthomonas axonopodis* pv.*citri* causa o "Cancro Cítrico", doença que ataca culturas cítricas causando perdas significativas de produção e, conseqüentemente, perdas de ordem financeira (Silva et al., 2002). Um segmento do genoma desta bactéria foi obtido a partir de consulta no Genbank sob registro NC003919. Essa seqüência inclui bases do gene XAC0965 até o gene XAC1014, e tem tamanho total de 49.611 bases. Os modelos competidores considerados variam de 2 a 5 estados latentes. Os resultados dos critérios de adequação *BIC*, ΔBIC e *AIC* associados aos modelos ajustados são apresentados na Tabela 2.

Tabela 2 - Valores de graus de liberdade, *BIC*, ΔBIC e *AIC* associados a seqüência do *Xanthomonas axonopodis* pv. *citri*

Nº de estados	graus de liberdade	*BIC*	ΔBIC	*AIC*
2	5	65.571,414	-3.204	65.527,354
3	12	65.701,336	-3.074	65.523,592
4	21	65.620,524	-3.155	65.533,908
5	32	65.807,290	-2.968	65.551,742

O modelo com 2 estados foi escolhido como o mais adequado pelo critério *BIC*. O critério *AIC* indica o modelo com 3 estados como o mais adequado. Embora os *HMM´s* sejam distintos (2 e 3 estados), os gráficos associados às esperanças condicionais da freqüência de *C+G* são similares, indicando um mesmo perfil.

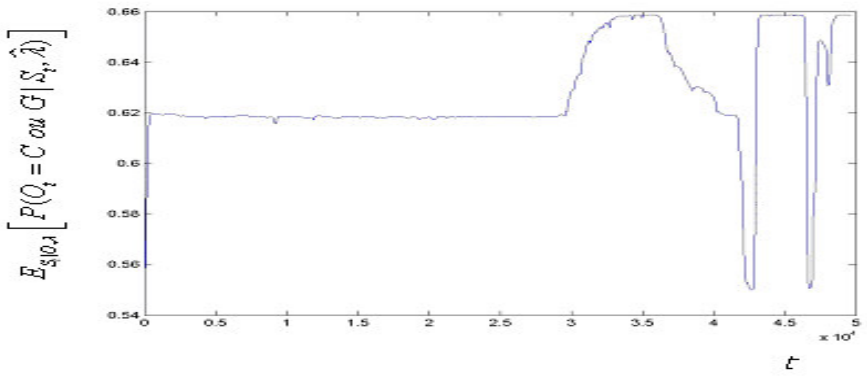

Figura 4 - Gráfico da composição local de C+G respectivo à seqüência do *Xanthomonas Axonopodis* pv. *citri* associado ao modelo com 3 estados.

Observando o gráfico da composição local associado ao modelo com 3 estados (Figura 4), podemos notar um comportamento altamente homogêneo nas 30..000 primeiras bases da seqüência. Este comportamento sugere a presença de uma única região funcional presente neste trecho. No segundo trecho da seqüência, após a posição 30.000, existe uma variação maior que no primeiro trecho (maior heterogeneidade) com relação aos valores da composição local. Isto sugere a possível presença de regiões que codificam proteínas com várias funcionalidades.

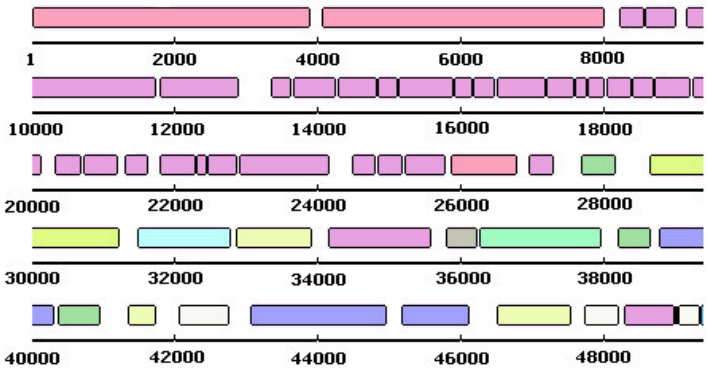

Figura 5 - Regiões de expressão do gene *XAC0965* até o gene *XAC1014* da bactéria *Xanthomonas axonopodis* pv. *citri*.

Observando a Figura 5, notamos que, aproximadamente, nas primeiras 30.000 posições, existe a codificação de proteínas que atuam na tradução e transcrição, processos ligados à síntese de proteínas. Nas 20.000 posições seguintes, existem várias funcionalidades associadas ao transporte de íons, metabolismo, além de regiões sem funcionalidade conhecida. Desta forma, o modelo ajustado foi capaz de identificar tais aspectos.

3.3 *Streptococcus pneumoniae*

O *Streptococcus Pneumoniae* é um importante patógeno humano. Esta bactéria causa a maioria das infecções respiratórias agudas e otites, evidenciando a importância de seu estudo (Tettelin et al., 2001). O segmento obtido a partir do *Genbank*, com código NC003098, é correspondente a seqüência de bases a partir do gene *SPR0584* até o gene *SPR0641*. Este segmento possui um total de 49.105 pares de bases. O segmento escolhido é marcadamente heterogêneo em termos de funcionalidade. Assim, o resultado da análise, utilizando o *HMM*, deve refletir este comportamento. Foram considerados 4 modelos competidores (modelos com 2, 3, 4 e 5 estados). Os valores dos critérios, *BIC*, ΔBIC , *AIC* e os graus de liberdade são apresentados na Tabela 3.

Tabela 3 - Valores de graus de liberdade, *BIC*, ΔBIC e *AIC* associados à seqüência da bactéria *Streptococcus pneumoniae*

Nº de estados	Graus de liberdade	*BIC*	ΔBIC	*AIC*
2	5	65.401,579	-2.672	65.357,571
3	12	65.435,606	-2.638	65.338,787
4	21	65.509,551	-2.564	65.342,318
5	32	65.623,531	-2.450	65.368,281

Observe que, segundo o critério *BIC*, o modelo mais adequado é o *HMM* com 2 estados. Porém, segundo o critério *AIC*, o modelo considerado mais adequado é o *HMM* com 3 estados. Embora os dois modelos sejam diferentes, os respectivos gráficos da composição local de *C+G*, são bem similares e representam, de maneira satisfatória, a conhecida heterogeneidade funcional do segmento.

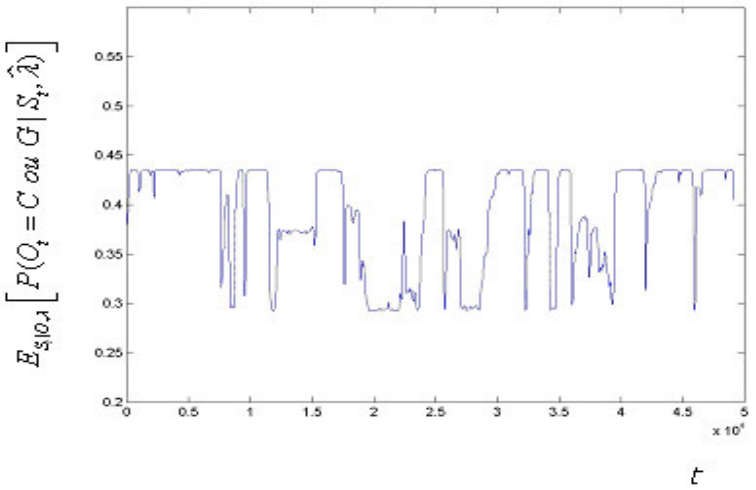

Figura 6 - Gráfico da *composição local* de *C+G* respectivo à seqüência do *Streptococcus pneumoniae*, sob o modelo com 3 estados.

Podemos observar, pela Figura 6, a característica altamente heterogênea quanto à *composição local* de *C+G* ao longo da seqüência. Isto sugere que o fragmento em estudo possui vários segmentos pequenos (proporcionalmente ao tamanho da seqüência) relacionados a características funcionais distintas. A divisão funcional no segmento analisado é apresentada na Figura 7.

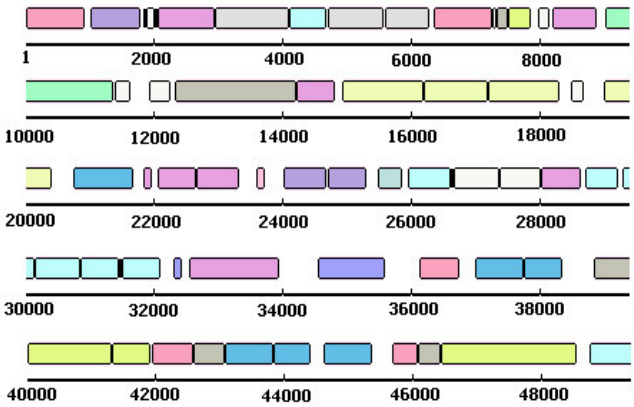

Figura 7 - Regiões de expressão do gene *SPR0584* até o gene *SPR0681* do *Streptococcus pneumoniae*.

Na Figura 7 podemos notar que não existem segmentos de grande dimensão (maiores que 5.000) que codificam um determinado tipo de proteína. O que se observa no segmento analisado é a presença de várias regiões com funções distintas, sendo que o *HMM* conseguiu descrever a alta heterogeneidade funcional do trecho.

3.4 *Escherichia coli*

Existem vários tipos de infecções causadas por esta bactéria. Entre elas, podemos citar colites, infecções urinárias e gastrintestinais entre outras (Santos, 2003). A seqüência de bases da bactéria *Escherichia coli* analisada foi obtida a partir do *Genbank* sob código NC004431. Esta subseqüência inclui bases do gene *C4208* até o gene *C4253*, totalizando 44.600 bases. Os valores dos critérios *BIC*, ΔBIC e *AIC*, e graus de liberdade associados aos modelos competidores são apresentados na Tabela 4.

Tabela 4 - Valores de Graus de Liberdade, *BIC*, ΔBIC e *AIC* associados à seqüência da bactéria *Escherichia coli*

Nº de estados	Graus de liberdade	*BIC*	ΔBIC	*AIC*
2	5	61.200,477	-8.114	61.156,950
3	12	61.202,015	-8.113	61.106,255
4	21	61.253,340	-8.061	61.087,935
5	32	61.357,238	-7.957	61.104,779

Segundo o critério *BIC*, o *HMM* com dois estados é o que melhor descreve os dados. Já o critério *AIC* aponta o modelo com 4 estados, como o mais adequado. O gráfico das esperanças condicionais do conteúdo de *C+G*, associado ao modelo com 4 estados, é apresentado na Figura 8.

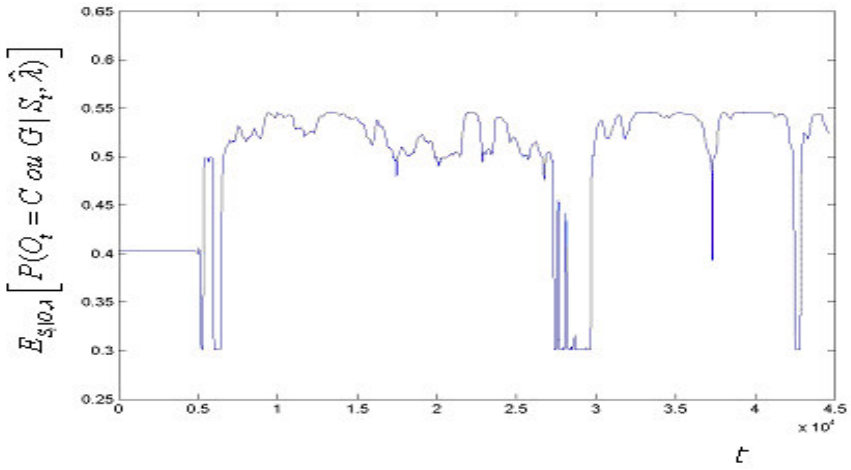

Figura 8 - Gráfico da composição local de *C+G* respectivo a seqüência do *Escherichia coli*.

Podemos notar, pelo comportamento do gráfico (Figura 8), um trecho inicial com alta homogeneidade com relação à *composição local* de *C+G*. Este trecho vai até a posição 5.000, sugerindo a presença de alguma região funcional. Todo o restante da seqüência (posição 5.000 até posição 44.600) não apresenta, em nenhum trecho, um comportamento homogêneo claro (as posições 30.000 até 40.000 apresentam um indício de homogeneidade). Isto indica a possível presença de um número expressivo de distintas funcionalidades associadas a este segmento. A estrutura funcional real da seqüência analisada é apresentada na Figura 9.

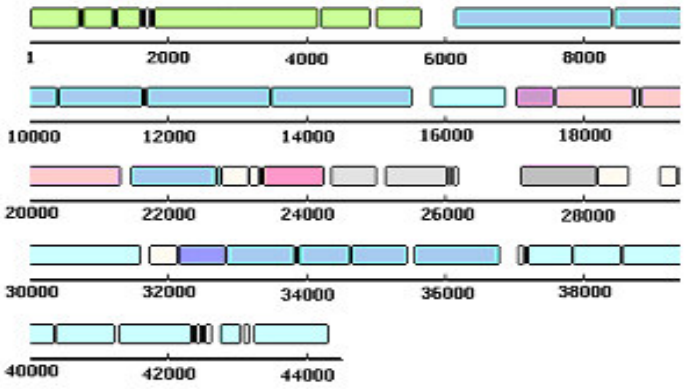

Figura 9 - Regiões de expressão do gene *C4208* até o gene *C4253* da *Escherichia coli.*

De acordo com a Figura 9, o primeiro trecho da seqüência, composto pelas 5.000 primeiras bases, está responsável pela codificação de proteínas com funções ligadas à secreção celular. Nesta região conseguiu-se boa discriminação com o *HMM*. No entanto,

na região seguinte, partindo da base 6.000 até a base 15.000, existe uma região funcional associada ao transporte de carboidratos e ao metabolismo, que não foi evidenciada pelo gráfico da *composição local* (trecho bastante heterogêneo). O trecho do segmento de *DNA* que inicia, aproximadamente, na base 16.000 e vai até a base 32.000, é heterogêneo com relação às características funcionais existentes, o que é evidenciado pela análise utilizando o *HMM*. Da base 33.000 da seqüência até 37.000, existe a presença de uma região funcional ligada ao transporte de aminoácidos que fica moderadamente evidenciada no gráfico das esperanças condicionais. Situação análoga ocorre no segmento que vai da posição 37.000 até a posição 44.000, onde ocorre uma região funcional ligada ao transporte e metabolismo de carboidratos. Vale ressaltar que o modelo com maior número de estados (5 estados), embora não escolhido pelos critérios adotados, conseguiu capturar, mais claramente, as regiões funcionais do segmento analisado.

Conclusões

Visto que o número de organismos seqüenciados cresceu nos últimos anos, são necessários métodos para análises destas seqüências. Extrair as informações presentes nestas seqüências constitui um desafio. Neste sentido, uma das análises úteis é a discriminação de regiões com distintas funcionalidades biológicas em dados relativos a genomas de organismos. Os *HMM´s* se mostraram eficientes nesta tarefa, uma vez que na quase totalidade dos dados analisados, os resultados fornecidos pelo *HMM´s* se mostraram coerentes com as estruturas funcionais conhecidas dos trechos analisados.

OLIVEIRA, D. C.; da-SILVA, C. Q.; CHAVES, L. M. Hidden Markov models applied to DNA sequence analysis. *Rev. Mat. Est*., São Paulo, v.24, n.2, p.51-66, 2006.

- *ABSTRACT: Hidden Markov models are applied to the analysis of homogeneous segments of DNA sequences. A software was developed and applied to the analysis of the bacteriophage lambda, Xanthomonas axonopodis pv. citri, Escherichia coil, Strepctococcus pneumonia and Xylella fastidiosa DNA sequences. The results are compared to the ones obtained by Churchill (1989) and da-Silva (2003).*

- *KEYWORDS: Sthocastic processes; hidden Markov models; DNA analysis.*

Referências

BOYS, R. J.; HENDERSON, D. A. A Bayesian approach to DNA sequence segmentation. *Biometrics*, Washington, v.6 , n.3, p.573-581, 2004.

BOYS, R. J.; HENDERSON, D. A.; WILKINSON, D. J. Detecting homogeneous segments in DNA sequences by using hidden Markov models. *J. R. Stat. Soc. Ser. C*, London, v.49, n.2, p.269-285, 2000.

CHURCHILL, G. A. Stochastic Models for Heterogeneous DNA sequence. *Bull. Math. Biol.*, New York ,v.51, p.79-94, 1989.

CHURCHILL, G. Hidden Markov Chain and the analysis of genome structure. *Comp. Chem.*, Amsterdam, v.16, n.2, p.107-115, 1992.

da-SILVA, C. Q. Hidden Markov models applied to a subsequence of the Xylella fastidiosa genome. *Genet. Mol. Biol.*, Ribeirão Preto, v.2, p.529-535, 2003.

da-SILVA, C. Q. Tutorial sobre modelos markovianos com estados latentes. In: CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL, 25, 2002. Nova Friburgo. *Anais...* Nova Friburgo: SBMAC, 2002.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum Likehood from incomplete data via the EM algorithm. *J. R. Stat. Soc.*, London, v.39, p.1-38, 1977.

HINKLEY, D. V.; HINKLEY, E. A. Inference about the change-point in a sequence of binomial variables. *Biometrika*, London, v.57, p.477-488, 1970.

MOOD, A.; GRAYBILL, F. A.; BOES, D. C. *Introduction to the Theory of statistics*. 3. ed. Tokio: Mcgraw-hill Kogakusha, 1963. 564 p.

OLIVEIRA, D. C. de. *Cadeias de Markov com Estados Latentes com aplicações em análises de seqüências de DNA*. 2005. 190f. Dissertação (Mestrado em Agronomia – Estatística e Experimentação Agropecuária) – Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, 2005.

RABINER, L. R. A Tutorial on hidden Markov models and selected applications on speech recognition. *Proc. IEEE*, New York, v.77, n.2, 1989.

SAKAMOTO, Y.; ISHIGURO, M.; KITAGAWA, G. *Akaike information criterion statistics*. Tokio: D. Reidel, 1986. 240p.

SANTOS, E. *Estudo dos fatores de virulência de Escherichia coli, isolada de infecção urinária em humanos*. 2003. 135f. Dissertação (Mestrado em Microbiologia Agropecuária) – Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista, Jaboticabal, 2003.

SCHWARZ, G. Estimating the dimension of a model. *Ann. Stat.*, Washington, v.6, p.461-464, 1978.

SILVA, A. C. et al. Comparison of genomes of two Xanthomonas pathogens with differing host specificities. *Nature*, London, v.417, p.459-463, 2002.

SIMPSON, A. J. G. et al. The genome sequence of the plant pathogen Xylella fastidiosa. The Xylella fastidiosa consortium of the organization for nucleotide sequencing and analysis. *Nature*, London, v.406, n.6792, p.151-157, 2000.

SMITH, A. F. M. A baysean Approach to Inference about a Change Point in a Sequence of Random Variables. *Biometrika*, London, v.62, p.407-416, 1975.

STADEN, R. Graphic methods to determine the function of nucleic acid sequence. *Nucleic Acids Res.*, Oxford, v.12, p.521-538, 1984.

TETTELIN, H. et al. Complete genome sequence of a virulent isolate of streptococcus pneumoniae. *Science*, Washington, v.293, p.498-505, 2001.

Recebido em 20.09.2005.

Aprovado após revisão em 26.06.2006.