

# HMMs e reconhecimento de promotores procarióticos: investigação de alternativas para reduzir a taxa de falsos positivos

**Adriana Neves dos Reis, Ney Lemke**

Programa de Pós-Graduação em Computação Aplicada, PIPCA, UNISINOS,  
93022-000, São Leopoldo, RS  
E-mail: adriana@exatas.unisinos.br, lemke@exatas.unisinos.br

A expressão do código de um gene em informação útil para a célula é desencadeada pela interação da enzima RNA-polimerase (RNAP) com a seqüência de nucleotídeos que antecede à região gênica propriamente dita. Nessa região, denominada de região promotora ou somente promotor [6], estão localizados os elementos que orquestram a fase inicial da expressão dos genes. Esses sinais são reconhecidos por uma subunidade da RNAP, conhecida como fator  $\sigma$ .

Diante de sua importância, essas regiões são foco de interesse de diferentes estudos desde a primeira compilação com 298 promotores identificados por Lisser e Margalit em 1993 [5]. Contudo, apesar dos avanços da tecnologia *in vitro*, identificá-los é caro e demorado. Mesmo em organismo procarióticos (que não possuem núcleo celular), os quais apresentam genomas menores e mais compactos.

Abordagens *in silico*, mesmo com taxa de acerto média entre 13-54% [8] no reconhecimento de promotores, são uma alternativa competitiva para o problema. Do ponto de vista de análise de seqüências de DNA, o fator  $\sigma$  reconhece pequenas regiões em média de 6 nucleotídeos (hexâmeros) que são conservadas em todos os promotores em um dado organismo.

A partir desta abordagem, um promotor é caracterizado por 3 regiões (Figura 1): um hexâmero centrado em -35 do ponto inicial de transcrição (+1), outro centrado em -10, e a que separa os hexâmeros, em média de 17 bases e sem conservação relevante em sua composição.

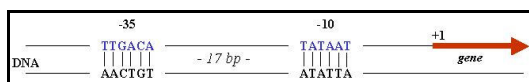


Figura 1 – Representação esquemática dos elementos que caracterizam um promotor em procariotos.

Para a *Escherichia coli* (*E. coli*), modelo biológico padrão entre os procariotos, o promotor ideal é reconhecido por conter a seqüência  $TTGACAN_{17}TATAAT$ , onde  $N$  corresponde a qualquer um dos 4 nucleotídeos (A, T, C e G). Entretanto, esse padrão não ocorre em nenhuma região promotora real [7].

Dentre os métodos *in silico* para analisar essas

seqüências, os *hidden Markov models* (HMMs) são amplamente utilizados, devido à sua natureza estocástica. Porém eles geram um grande número de Falsos Positivos (*FP*), ou seja, seqüências não-promotoras reconhecidas como promotor. Este trabalho investiga possíveis alternativas para a redução da taxa de *FP*, quando HMMs são aplicados para o reconhecimento de promotores.

HMM é um tipo de modelo de Markov, podendo ser definido, essencialmente, como um autômato estocástico de estados finitos [3]. Esse é dado por:

$$M = (Q, \Sigma, \pi, a, b),$$

onde

- $Q = \{q_0, \dots, q_n\}$  é um conjunto de  $n$  estados;
- $\Sigma = \{\sigma_1, \dots, \sigma_m\}$  é o alfabeto finito de saída;
- $(\pi_i)$  são as probabilidades iniciais dos estados;
- $(a_{i,j})$  são as probabilidades de transição do estado  $i$  para cada estado  $j$  possível, dado que o modelo está em  $i$ ;

- e  $(b_{i,k})$  as probabilidades de emissão dos elementos  $k$  pertencentes ao alfabeto no estado  $i$ .

Nosso trabalho parte da criação de um cenário padrão de investigação, sobre o qual os experimentos para redução da taxa de *FP* foram executados. Segue sua descrição.

A arquitetura dos HMMs é do tipo linear com transições possíveis para todos os tipos de estados, ou seja, de pareamento (*match*), de não-pareamento (*insert*) e de deleção (*delete*). Isso está em conformidade com a modelagem usada para alinhamento de seqüências biológicas (Figura 2).

Além disso, os HMMs possuem 81 estados principais, mesmo número de nucleotídeos que compõem os promotores da base de dados do RegulonDB ([http://www.cifn.unam.mx/Computational\\_Genomics/regulondb/](http://www.cifn.unam.mx/Computational_Genomics/regulondb/)); o qual fornece informações confiáveis sobre a rede regulatória de *Escherichia coli* com base experimental, incluindo dados dos promotores e sua respectiva classe de fator  $\sigma$  associado [10]. As regiões promotoras propriamente ditas foram extraídas do banco,

sendo consideradas as versões 3.1 [4] e 4.0 [10] disponibilizadas, respectivamente, em dezembro de 2003 e em julho de 2004.

O alfabeto de emissão é constituído dos 4 nucleotídeos e a probabilidade inicial de emissão de cada um é igual a 0,25.

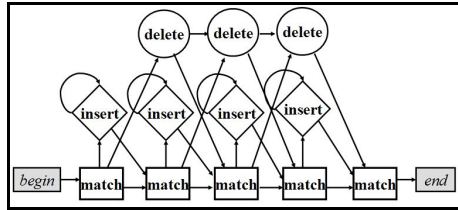


Figura 2 – Arquitetura dos HMMs utilizados em alinhamento de seqüências biológicas.

Com a estrutura e os parâmetros iniciais definidos, o HMM é treinado por 30 épocas no modo *on-line*, período considerado adequado para este tipo de problema [2]. Após o treinamento, tem-se o melhor modelo que descreve as seqüências do conjunto de treinamento submetidas ao HMM nesta etapa.

A primeira diferença deste trabalho para os anteriores, é o uso da metodologia de *10-fold cross validation* [11], para avaliar os resultados estatisticamente. Isso se faz necessário, principalmente em razão do grande número de parâmetros a serem computados em relação ao número de promotores disponíveis para treinamento dos modelos.

O conjunto de teste é constituído dos 10% separados do conjunto de treinamento mais o mesmo número de seqüências da classe gênica.

A partir desse cenário, a primeira hipótese para reduzir FP foi a determinação criteriosa do limiar de reconhecimento. Normalmente, o valor máximo para uma seqüência ser considerada promotora é definido através da comparação do valor médio de verossimilhança para duas classes de seqüências distintas. Isso torna esse limiar fortemente dependente da classe considerada como não-promotora.

Para evitar isso, foi considerado um método mais rigoroso para determinar o valor de escore limite ou crítico de reconhecimento ( $S_c$ ), tal que as seqüências com pontuação menor que ele sejam classificadas como promotoras.

Para especificar o  $S_c$ , utilizaram-se dois conjuntos de seqüências: o de promotores submetido para treinar o HMM e o de fragmentos gênicos de mesmo tamanho. Esses dois conjuntos são avaliados pelo HMM através do algoritmo de *Viterbi*, que calcula a verossimilhança de uma seqüência ter sido gerada por um dado modelo.

Como resultado, obtivemos os escores para o conjunto de VP ( $S_{promotores}$ ) e para o conjunto de VN ( $S_{genes}$ ). Essas duas variáveis podem ser modeladas

como variáveis aleatórias normalmente distribuídas, sendo computados para cada conjunto seus histogramas. Ajustando esses dados a Gaussianas, conforme mostra a Figura 2, têm-se duas distribuições  $D_{promotores}$  e  $D_{genes}$ , respectivamente para verdadeiros positivos e verdadeiros negativos.

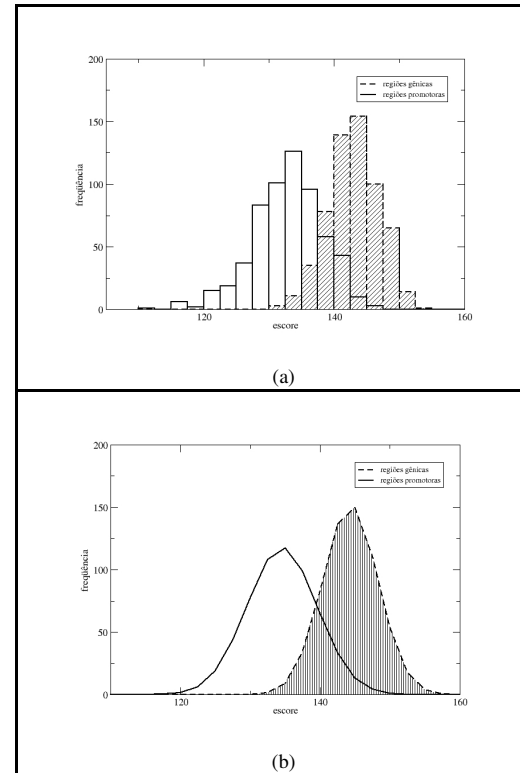


Figura 3 – Distribuições de probabilidade dos escores do algoritmo de *Viterbi* para regiões promotoras e gênicas. (a) Histogramas. (b) Curvas de ajuste a Gaussianas.

O cálculo de  $S_c$  baseou-se na Regra de Decisão de Bayes, na qual, usando a melhor estimativa das probabilidades de duas classes, calcula-se o valor esperado de decisão para cada uma delas, escolhendo a alternativa com máximo valor.

Considerando que as distribuições  $D$  representam fielmente os dados, podem-se determinar as funções  $D_{promotores}(S)$  e  $D_{genes}(S)$ , as quais denotam a fração de verdadeiros negativos e de verdadeiros positivos em função do escore limiar  $S$ .  $S_c$  é o valor esperado que maximiza a soma dessas duas funções.

Com base nesta idéia, escolheu-se o critério de exatidão ( $A$ ) para escolher o  $S_c$ . Assim, o  $S_c$  é o valor que maximiza a exatidão do modelo, a partir do número de Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN). O cálculo de exatidão é dado por:

$$A = \frac{VN + VP}{VN + VP + FN + FP}$$

O cálculo de exatidão também é empregado como medida de desempenho sobre os conjuntos de teste. Assim, são calculados dois valores de exatidão: a esperada ( $A_e$ ), resultante da soma das distribuições, e a observada ( $A_o$ ), resultante do cálculo sobre o conjunto de teste.

Determinada essa metodologia, verificou-se a influência do tamanho do conjunto de treinamento no desempenho do HMM. Uma vez que existem 2 versões do RegulonDB, no primeiro experimento foram consideradas 600 regiões promotoras da versão 3.1 e 929 da versão 4.0. Os resultados são apresentados na Tabela 1.

RegulonDB	Medida	Média	Desvio-padrão
Versão 3.1	$S_c$	139,177	0,957
	$A_e$	0,843	0,03
	$A_o$	0,805	0,047
Versão 4.0	$S_c$	136,587	0,615
	$A_e$	0,851	0,011
	$A_o$	0,817	0,07

Tabela 1: Medições de escore limite e exatidão para o reconhecimento de promotores de *E. coli*, utilizando as versões 3.1 e 4.0 do RegulonDB.

Neste experimento, observou-se que mesmo com as extensões no cenário, o erro tem apenas uma redução de 1,05% de acordo com as medidas de exatidão.

Isso pode ser explicado pelo fato de que quando um HMM é treinado com as regiões promotoras, o modelo captura não apenas padrões específicos dessas seqüências, mas também outras características intrínsecas à constituição e organização do genoma do organismo. Um caso conhecido é a influência do conteúdo A+T do genoma [9].

Assim, além da determinação do limiar de decisão, verificou-se a hipótese do uso de um par de HMMs no reconhecimento de promotores, um para regiões promotoras e o outro para gênicas, com o objetivo de eliminar o ruído causado por outros padrões genômicos no modelo promotor. Essa alteração está fundamentada na Teoria de Estimação de Limite de Decisão de Verossimilhança para Análise de Discriminação [1].

Para isso, 2 modelos foram criados:

- um para representar seqüências promotoras de *E. coli*;
- um *modelo nulo* para representar demais regiões.

As seqüências gênicas continuam sendo consideradas

como Verdadeiros Negativos. Para computar o escore ( $S_p$ ) de uma seqüência  $x$  ser promotora, consideram-se 2 modelos: o de promotor ( $hmm_p$ ) e o de genes ( $hmm_g$ ), sendo dado por:

$$S_p = \log \frac{P(x|hmm_p)}{P(x|hmm_g)}$$

Esse cálculo tem a vantagem de não requerer que sejam conhecidas previamente as probabilidades das duas classes de seqüências mencionadas. Além disso, com ele reduzimos o ruído das características intrínsecas do genoma no modelo promotor, principalmente o conteúdo A+T. Em resumo, com o  $S_p$  calculamos o quanto mais provável uma seqüência é promotora do que gênica. A Tabela 2 compara os resultados dos experimentos para as duas hipóteses investigadas.

Acrescentando a Análise de Discriminação à metodologia estabelecida para determinação do limiar de decisão, a exatidão, tanto observada quanto a esperada, ultrapassa 0,9. Além disso, comparando-se os mesmos resultados com os obtidos apenas com a primeira hipótese, adquiriu-se um aumento de performance em 12,48% e uma redução na taxa de erro em 44,26%.

Extensão	Medida	Média	Desvio-padrão
Determinação do Limiar de Decisão	$S_c$	136,587	0,615
	$A_e$	0,851	0,011
	$A_o$	0,817	0,07
Determinação do Limiar de Decisão e Análise de Discriminação	$S_c$	1,343	0,604
	$A_e$	0,921	0,005
	$A_o$	0,919	0,03

Tabela 2: Medições de escore limite e exatidão para o reconhecimento de promotores de *E. coli*, utilizando ambas extensões no RegulonDB versão 4.0.

Este trabalho reafirma a necessidade de considerar aspectos funcionais e estruturais dos promotores, para que os modelos matemáticos de seqüências ou as metodologias de reconhecimento consigam representar diferentes influências sob o mecanismo de expressão.

## Agradecimentos

Ao CNPq pelo apoio financeiro, e à HP Brazil R&D pela colaboração no desenvolvimento do trabalho.

## Referências

- [1] L. Arslan, J.H.L. Hansen, Likelihood Decision Boundary Estimation between HMM pairs in Speech Recognition, *IEEE Transactions on Speech & Audio Processing*, 6, 4 (1998) 410-414.
- [2] P. Baldi, S. Brunak, “Bioinformatics: the machine learning approach”, MIT Press, 2<sup>a</sup> ed., 2001.
- [3] P. Clote, R. Backofen, “Computational Molecular Biology: an introduction”, Wiley, 2000.
- [4] A.M. Huerta, H. Salgado, D. Thieffry, J. Collado-Vides, RegulonDB: a database on transcriptional regulation in *Escherichia coli*, *Nucleic Acids Research*, 26, 1 (1998) 55-59.
- [5] S. Lissner, H. Margalit, Compilation of *E. coli* mRNA promoter sequences, *Nucleic Acids Research*, 21, 7 (1993) 1507-1516.
- [6] A.G. Pedersen, P. Baldi, S. Brunak, Y. Chauvin, Characterization of prokaryotic and eukaryotic promoters using hidden Markov models, em Proceedings for Fourth International Conference on Intelligent Systems for Molecular Biology, 182-191, 1996.
- [7] A.P. Pevzner, “Computational Molecular Biology: an algorithmic approach”, MIT Press, 2000.
- [8] P. Qiu, Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochemical and Biophysical Research Communications*, 309 (2003) 495-501.
- [9] A.N. Reis, N. Lemke, Análise de um Modelo HMM para Predição de Regiões Promotoras Procarióticas com Base em Dados de *Escherichia coli*, em XXVII Congresso Nacional de Matemática Aplicada e Computacional - CNMAC, 2004.
- [10] H. Salgado, S. Gama-Castro, A. Martínez-Antonio, E. Díaz-Peredo, F. Sánchez-Solano, M. Peralta-Gil, D. Garcia-Alonso, V. Jiménez-Jacinto, A. Santos-Zavaleta, C. Bonavides-Martínez, J. Collado-Vides, RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12, *Nucleic Acids Research*, 32 (2004) D303-D306.
- [11] C.H. Wu, J.W. McLarty, “Neural Networks and Genome Informatics”, Elsevier Science, 2000.