

Identificação de Regiões Codificantes de Proteína Através da Transformada Modificada de Morlet

Jesús P. Mena-Chalco¹, Roberto M. Cesar-Jr.¹

¹Departamento de Ciência da Computação – IME – USP – SP – Brasil

{jmena,cesar}@vision.ime.usp.br

Abstract. *An important topic in biological sequences analysis is gene identification, i.e., the identification of protein coding regions. In the Msc. study, we introduced a new method, which does not need training dataset for identification of protein coding regions. This method is based on a new transform, here called Modified Morlet Transform, and defined because traditional time-scale transforms are not suitable to gene identification. We present the main obtained results thus showing that the method performs better than previous approaches based on the short time Fourier transform.*

Resumo. *Um tópico importante na análise de seqüências biológicas é a identificação de genes, i.e., a identificação de regiões codificantes de proteína. No estudo de mestrado, propusemos um novo método, o qual não necessita de conjuntos de treinamento, para a identificação de regiões codificantes. Este método baseia-se em uma nova transformada, denominada Transformada Modificada de Morlet e definida porque transformadas tempo-escala tradicionais não são completamente apropriadas para a identificação de genes. Aqui apresentamos os principais resultados obtidos, os quais mostram que o método tem um melhor desempenho do que métodos prévios baseados na transformada de Fourier de tempo reduzido.*

1. Introdução

Atualmente, a área de Bioinformática vem recebendo muita importância devido ao auxílio que fornece, na descoberta genômica, a fim de levar um melhor entendimento dos organismos. Quando um novo organismo é seqüenciado, é desejável obter toda a informação possível do genoma, sendo um passo fundamental a identificação de genes. Tal identificação corresponde à determinação das regiões codificantes de proteína (CDS). As CDS tipicamente apresentam uma organização periódica imperfeita de três bases, a qual geralmente não está presente nas regiões intergênicas e nos íntrons. Nos últimos anos, essa característica independente das espécies foi analisada para que se possa explicar sua origem e quantificá-la, por exemplo [Eskesen et al. 2004]. Na literatura, é comumente chamada de periodicidade de três bases (TBP), tendo sido observada de maneira similar para di-nucleotídeos em cromossomos de bactérias.

Os métodos de processamento digital de sinais (DSP) foram usados para identificar éxons de células de eucariotos, apresentando resultados promissores. Estes métodos, enfocados somente na busca de regiões com TBP, são recomendáveis para uso em genomas nos quais não existem conjuntos de treinamento. Os métodos de DSP para a identificação de CDS baseados na transformada de Fourier [Tiwari et al. 1997, Anastassiou 2001] e filtros digitais [Vaidyanathan and Yoon 2004] não apresentam

formulações robustas devido a sua dependência em relação ao tamanho de janela para a análise local bem como devido à carência de delimitações entre CDS e não CDS¹. A definição prévia do tamanho de janela é crítica e, para reduzir tal dependência, métodos alternativos que exploram diferentes tamanhos de janela, tal como a transformada em *wavelets*, foram recentemente estudados.

Uma maneira natural para a identificação de CDS mediante técnicas de multi-resolução consiste no uso de pequenas ou grandes escalas em curtas ou longas CDS, respectivamente. Assim, as CDS curtas poderão ser analisadas por funções de suporte pequeno e as CDS longas por funções de suporte grande. Neste contexto, a transformada em *wavelets* é uma abordagem lógica para analisar as seqüências de DNA. Entretanto, não é completamente apropriada pois a freqüência das funções de análise variam com a mudança das escalas.

A principal contribuição da dissertação de mestrado [Mena-Chalco 2005] é a proposição de um método para a identificação de CDS baseado na transformada modificada de Morlet (MMT) que resolve apropriadamente o problema de análise de sinais com freqüência específica e de escala variável. A MMT foi igualmente introduzida na dissertação. Adicionalmente, avaliamos o desempenho do método comparando-o com outro baseado na transformada de Fourier de tempo reduzido (STFT).

Este artigo está organizado da seguinte maneira: na Seção 2 introduzimos a MMT; na Seção 3 detalhamos nosso método para a identificação de CDS; finalmente descrevemos nossos resultados e discutimos o desempenho do método na Seção 4.

2. Transformada Modificada de Morlet

Uma transformada multiescala de um sinal u pode ser calculada por:

$$U(b, a) = \int u(x)g(x, b, a)dx, \quad (1)$$

em que $a > 0$ é o parâmetro de escala, b o parâmetro de espaço e g a função de análise.

Na Equação (1), diversas funções podem ser adotadas para transformar u . Em particular, funções bem localizadas no domínio da freqüência, como a de Morlet, são utilizadas para analisar sinais de forma local e com diferentes freqüências. Estas funções não são completamente apropriadas para a identificação de CDS pois variam suas freqüências com a variação da escala, i.e., a variação do desvio padrão da Gaussiana envolvida.

Aqui definimos uma modificação da função de Morlet para analisar localmente sinais em uma freqüência específica e com escala variável. Na função de análise de Morlet usamos o parâmetro de escala a para manter constante a freqüência da exponencial complexa, variando o desvio padrão da Gaussiana (ou seja, a escala),

$$U(b, a) = \int u(x)e^{j\omega_0(x-b)}e^{-\frac{(x-b)^2}{2a^2}} dx, \quad (2)$$

em que ω_0 é a freqüência básica angular. Portanto, a função de análise ψ_{MM} da Transformada Modificada de Morlet está definida por:

$$\psi_{MM}(x, a) = e^{j\omega_0 x}e^{-\frac{x^2}{2a^2}} \quad (3)$$

¹Atualmente, os métodos para identificação de CDS baseados em DSP mostram somente uma verificação visual dos resultados, apresentando medidas qualitativas na identificação, as quais dificultam muito uma possível comparação com outros métodos.

3. Método para a Identificação de CDS

O método proposto para a identificação de regiões codificantes de proteína é dividido em quatro processos.

3.1. Mapeamento Numérico de Nucleotídeos

Usamos as regras 4–7 do mapeamento fixo binário [Buldyrev et al. 1995] para criar quatro seqüências binárias em que cada uma represente as posições da adenina (A), citosina (C), guanina (G) e timina (T) na seqüência de DNA. As regras 1–3 não produzem informação relevante para a identificação de CDS. Denotamos por u_A , u_C , u_G e u_T as seqüências binárias correspondentes às quatro regras associadas aos nucleotídeos A, C, G e T, respectivamente.

3.2. Aplicação da MMT

A MMT, com diferentes escalas a e frequência angular ω_0 sendo um múltiplo de três², é calculada para todas as seqüências binárias utilizando ψ_{MM} de N pontos,

$$U_A(b, a) = \int u_A(x) \psi_{MM}^*(x - b, a) dx \quad (4)$$

$$U_C(b, a) = \int u_C(x) \psi_{MM}^*(x - b, a) dx \quad (5)$$

$$U_G(b, a) = \int u_G(x) \psi_{MM}^*(x - b, a) dx \quad (6)$$

$$U_T(b, a) = \int u_T(x) \psi_{MM}^*(x - b, a) dx \quad (7)$$

Os melhores resultados obtidos foram para escalas exponencialmente espaçadas entre 0, 2 e 0, 7. Neste trabalho, o espectro de cada seqüência binária é definido como o módulo ao quadrado dos seus coeficientes depois da aplicação da transformada:

$$m_A(b, a) = |U_A(b, a)|^2 \quad (8)$$

$$m_C(b, a) = |U_C(b, a)|^2 \quad (9)$$

$$m_G(b, a) = |U_G(b, a)|^2 \quad (10)$$

$$m_T(b, a) = |U_T(b, a)|^2 \quad (11)$$

Assim, o espectro total é a soma dos espectros de cada seqüência binária:

$$M(b, a) = m_A(b, a) + m_C(b, a) + m_G(b, a) + m_T(b, a) \quad (12)$$

Com a formulação de espectro total tentamos, parcialmente, representar a interação e a dependência existente entre nucleotídeos no genoma.

3.3. Projeção dos Espectros das Seqüências

O espectro total da seqüência analisada é projetado no eixo das posições para que possamos ter uma medida local das regiões com TBP. Dada uma seqüência de tamanho N , os coeficientes de projeção do espectro total aqui são definidos como:

$$M_p(b) = \sum_a M(b, a), \quad b = 0, \dots, N - 1 \quad (13)$$

Vale salientar que estes coeficientes permitem comparar o nosso método usando a MMT e a STFT. Por outro lado, a projeção no eixo das escalas revela qual escala mantém mais energia no sinal através das posições.

²A definição de $\omega_0 = N/3$ implica que a frequência angular de ψ_{MM} é um múltiplo de três.

3.4. Limiarização dos Coeficientes de Projeção

Uma forma natural de localizar os limites entre CDS e não CDS é mediante a incorporação de um limiar nos coeficientes de projeção. A limiarização sobre M_p permite excluir posições onde os coeficientes são pequenos, i.e., todos os coeficientes abaixo de um dado valor são substituídos por zero. Em geral, regiões com pouca ou nenhuma TBP têm coeficientes de projeção baixos. Assim, os coeficientes diferentes de zero são aqueles associados a possíveis CDS.

4. Resultados e Discussão

Diversos experimentos foram realizados para avaliar a eficiência do método proposto usando a MMT e a STFT na análise de seqüências. Com a MMT usamos um sinal de 1200 pontos e 40 escalas entre 0,2 e 0,7. Por outro lado, com a STFT usamos tamanhos de janela de 200, 300 e 400 pontos correspondentes a valores aproximados da média e desvio padrão do conjunto de seqüências usado.

Para fins de comparação, usamos o limiar no intervalo de 5 e 95 nos coeficientes de projeção, quando é considerada a MMT e, a soma dos coeficientes normalizados na frequência três, quando é considerada a STFT [Tiwari et al. 1997]. Foram utilizadas a sensibilidade, a especificidade e os coeficientes de correlação [Bursat and Guigó 1996] para medir o desempenho de identificação. Medidas subjetivas e outros componentes de desempenho não foram considerados na análise dos resultados.

Na Figura 1, apresentamos as medidas de desempenho calculadas para o conjunto original [Bursat and Guigó 1996] e dois subconjuntos deste em que seqüências com tamanhos de éxons menores a 30 bp e 100 bp foram removidos. Para o conjunto original usando a MMT, obteve-se uma especificidade máxima de 0,44 em uma sensibilidade de 0,78. Com um limiar de 80% obteve-se uma exatidão máxima de 49%. Entretanto, usando a STFT com tamanho de janela de 300 pontos, obteve-se uma especificidade máxima de 0,38 em uma sensibilidade de 0,75. Com o mesmo limiar de 80% obteve-se a exatidão máxima de 40%. Nos outros dois subconjuntos, apresenta-se um comportamento similar, em que a MMT tem desempenho superior à STFT. No subconjunto cujas seqüências contêm éxons de tamanhos maiores a 100 bp obteve-se a exatidão máxima de 56%.

Os melhores desempenhos foram obtidos em conjuntos cujas seqüências contêm éxons de tamanhos superiores a 100 bp e um limiar de 80%. Isso sugere que o nosso método é adequado para analisar seqüências cujos comprimentos dos éxons sejam maiores a 100 bp. É interessante observar que o valor do limiar próximo de 85% está relacionado com o 15% dos nucleotídeos pertencentes às CDS no conjunto de seqüências usado. Acreditamos que valores de limiar adequados poderiam ser obtidos de estatísticas de organismos taxonomicamente similares ao serem analisados.

5. Conclusões

Na dissertação de mestrado [Mena-Chalco 2005] foi introduzido um novo método para a identificação de regiões codificantes de proteína, bem como uma modificação da função de análise de Morlet para a análise de regiões periódicas. Além disso, foi avaliado comparativamente o desempenho do método com outro baseado na STFT.

O método proposto na dissertação de mestrado tem um desempenho superior dos métodos prévios baseados na STFT, pois a análise multiresolução realizada permite exa-

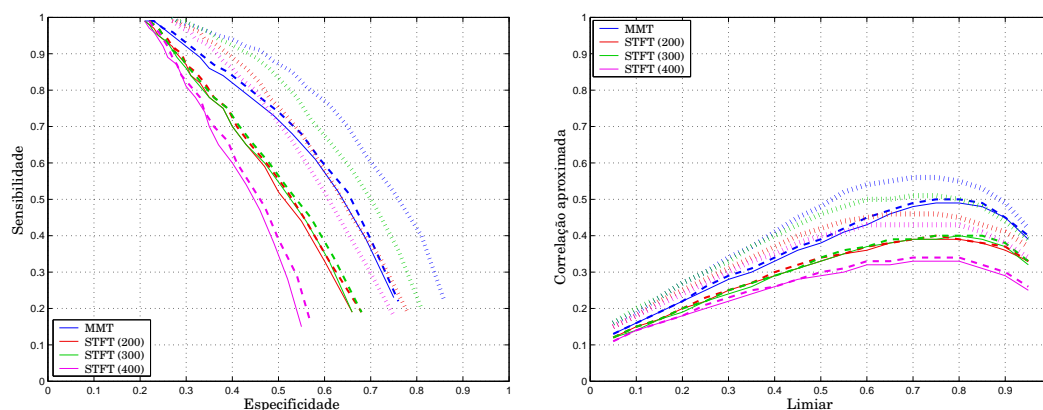


Figura 1. Desempenho em termos de especificidade e sensibilidade (esquerda) e limiar e correlação aproximada (direita) da identificação de CDS do método proposto usando a MMT. Os tamanhos de janela na STFT foram de 200, 300 e 400 pontos. O desempenho para o conjunto original é mostrado com linhas sólidas. Os subconjuntos cujas seqüências contêm éxons maiores que 30 bp e 100 bp são mostrados com linhas tracejadas e pontilhadas, respectivamente.

minar regiões curtas ou longas com funções de análise periódicas e de suporte pequeno ou grande, respectivamente. A característica significativa deste método é a robustez às variações de escala. Atualmente, esta dependência é uma grande dificuldade nos outros métodos do estado-da-arte.

Informação Complementar

O texto completo da dissertação de mestrado, os *scripts* usados e um sistema de submissão *on-line* para a análise de seqüências de DNA mediante o método descrito estão disponíveis em <http://www.vision.ime.usp.br/~jmena/DSPgenomics/>.

Referências

- Anastassiou, D. (2001). Genomic signal processing. *IEEE Signal Processing Magazine*, 8(4):8–20.
- Buldyrev, S. V., Goldberger, A. L., Havlin, S., Mantegna, R. N., Matsu, M. E., Peng, C.-K., Simons, M., and Stanley, H. E. (1995). Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Physical Review E*, 51(5):5084–5091.
- Burset, M. and Guigó, R. (1996). Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–367.
- Eskesen, S. T., Eskesen, F. N., Kinghorn, B., and Ruvinsky, A. (2004). Periodicity of DNA in exons. *Journal of Molecular Biology*, 5(12):1–11.
- Mena-Chalco, J. P. (2005). Identificação de regiões codificantes de proteína através da transformada modificada de Morlet. Master's thesis, IME – USP.
- Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., and Ramaswamy, R. (1997). Prediction of probable genes by Fourier analysis of genomic sequences. *Bioinformatics*, 13(3):263–270.
- Vaidyanathan, P. P. and Yoon, B. (2004). The role of signal-processing concepts in genomics and proteomics. *Journal of the Franklin Institute*, 341(1-2):111–135.