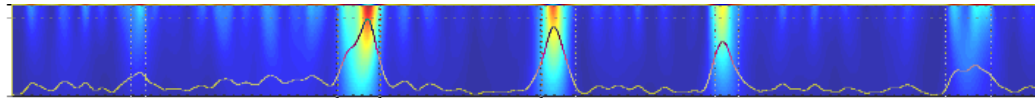


Identificação de Regiões Codificantes de Proteína através da Transformada Modificada de Morlet



Candidato

Jesús P. Mena-Chalco

Orientador

Prof. Dr. Roberto Marcondes Cesar Jr.

Departamento de Ciência da Computação - IME - USP

19/10/2005

Roteiro

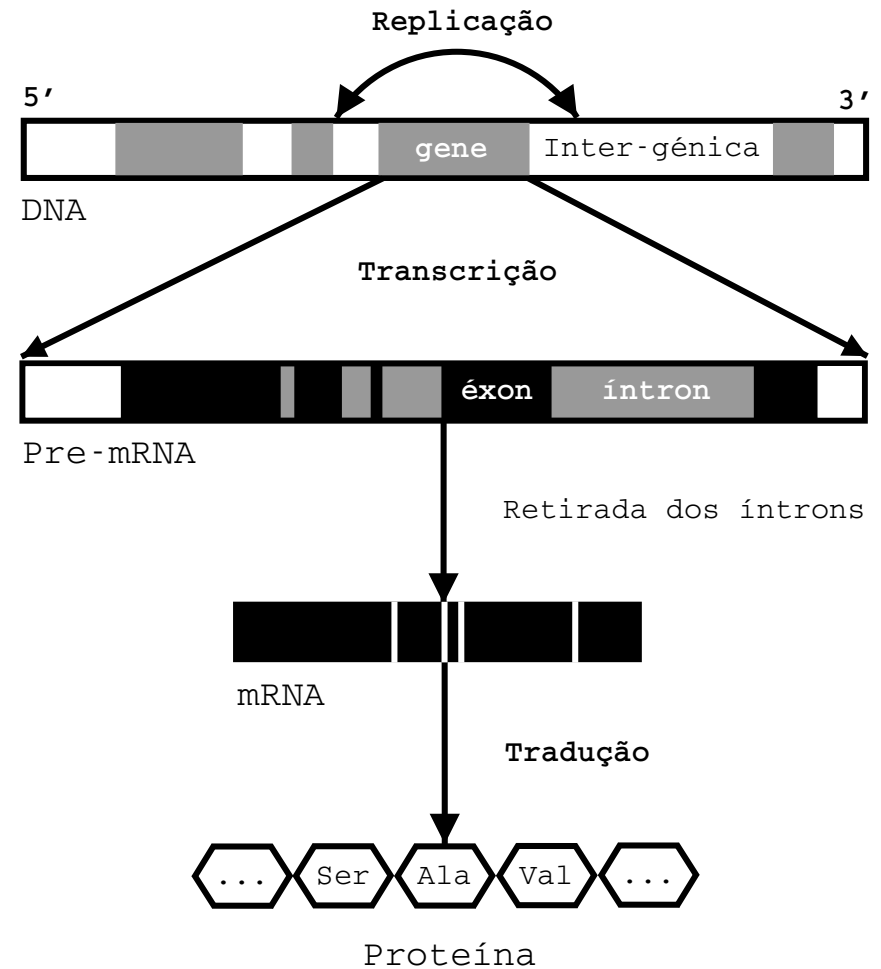
- *Pipeline* Bioinformático.
- O Problema da Identificação de Genes.
- Métodos de DSP para a Identificação de CDSs.
- Transformada Modificada de Morlet.
- Método Proposto.
- Resultados.
- Conclusões.
- Pesquisas futuras.

Introdução

Um **gene** é uma região que expressa ou controla uma proteína.

Sub-regiões:

1. De reconhecimento (promotora);
2. De transcrição;
3. Região não-traduzida 5';
4. De início de tradução (*start codon*);
5. **Região para a codificação de proteína (CDS);**
6. De tradução (*stop codon*);
7. Região não-traduzida 3';
8. De poliadenilação (*polyA*, eucariotos);
9. De transcrição.



Introdução

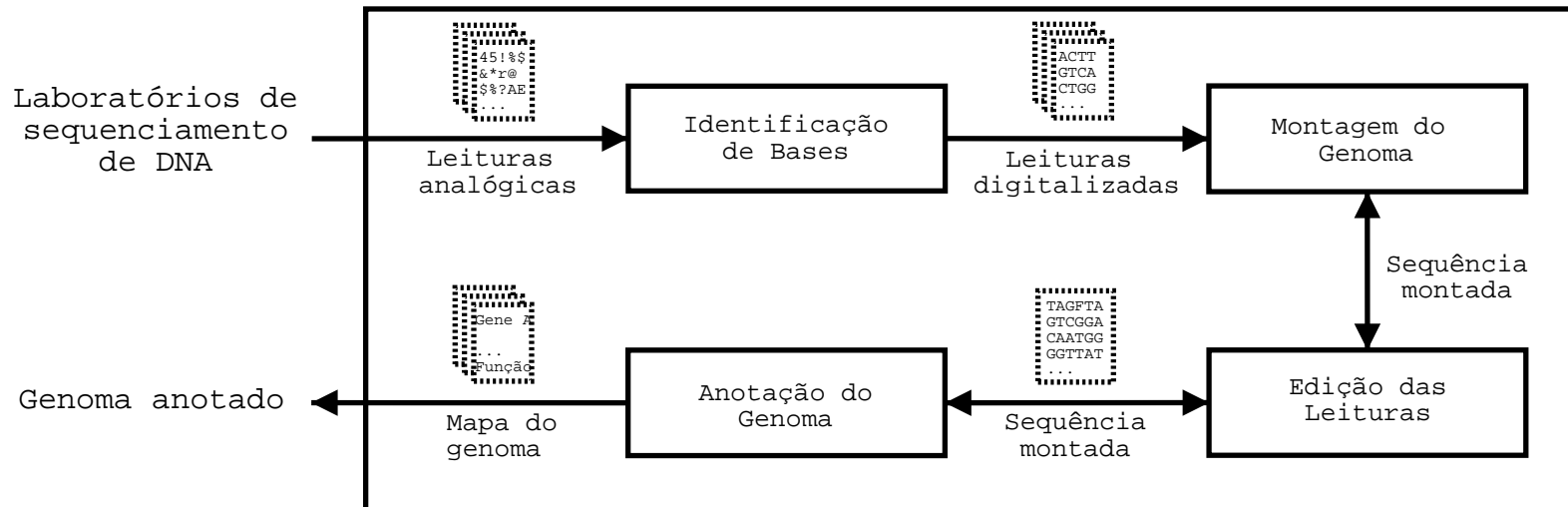
- Um tópico importante na análise de seqüências biológicas é a **busca de genes** (identificação de regiões codificantes de proteína).
- Metodologias computacionais para **identificar genes** e outras regiões funcionais foram desenvolvidas nos últimos 20 anos.
- Os métodos de **processamento digital de sinais** (DSP) têm um papel importante nesse contexto.
- Os métodos de DSP fornecem uma **base robusta** para a identificação de regiões codificantes de proteína (CDSs).

Objetivos

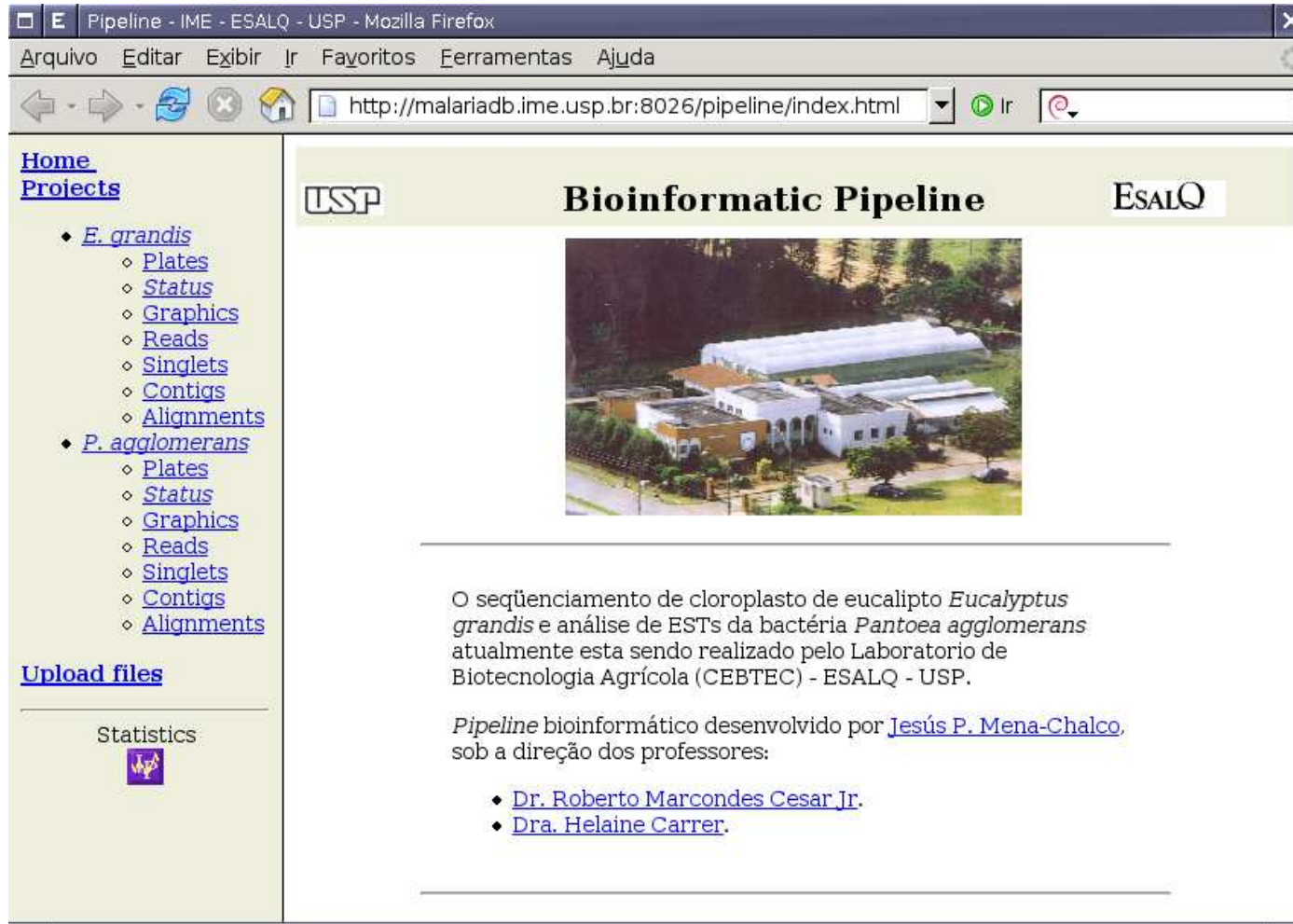
- Desenvolvimento de um *pipeline* bioinformático para o Laboratório de Biotecnologia Agrícola, ESALQ-USP.
- Estudo dos métodos de DSP para a identificação de regiões codificantes de proteínas.

Pipeline Bioinformático

Considerado como uma **seqüência de unidades** funcionais que realizam uma tarefa genômica em passos biológicos e/ou computacionais.



Pipeline Bioinformático Desenvolvido




Arquivo Editar Exibir Ir Favoritos Ferramentas Ajuda

http://malariadb.ime.usp.br:8026/pipeline/index.html


[Home](#)
[Projects](#)

- ♦ [E. grandis](#)
 - ◊ [Plates](#)
 - ◊ [Status](#)
 - ◊ [Graphics](#)
 - ◊ [Reads](#)
 - ◊ [Singlets](#)
 - ◊ [Contigs](#)
 - ◊ [Alignments](#)
- ♦ [P. agglomerans](#)
 - ◊ [Plates](#)
 - ◊ [Status](#)
 - ◊ [Graphics](#)
 - ◊ [Reads](#)
 - ◊ [Singlets](#)
 - ◊ [Contigs](#)
 - ◊ [Alignments](#)

[Upload files](#)

Statistics


USP **Bioinformatic Pipeline** **ESALQ**

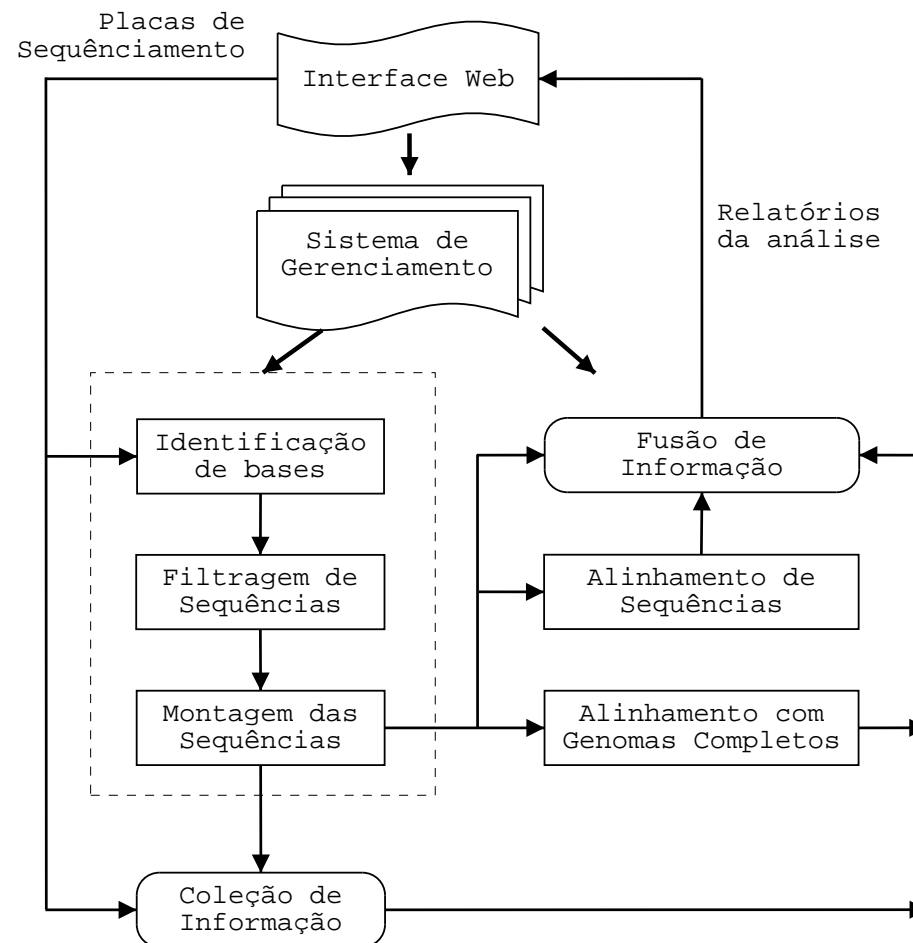


O seqüenciamento de cloroplasto de eucalipto *Eucalyptus grandis* e análise de ESTs da bactéria *Pantoea agglomerans* atualmente esta sendo realizado pelo Laboratório de Biotecnologia Agrícola (CEBTEC) - ESALQ - USP.

Pipeline bioinformático desenvolvido por [Jesús P. Mena-Chalco](#), sob a direção dos professores:

- ♦ [Dr. Roberto Marcondes Cesar Jr.](#)
- ♦ [Dra. Helaine Carrer.](#)

Arquitectura do *Pipeline* Bioinformático



O Problema: Identificação de Genes

Categorias que agrupam as abordagens para sua solução:

1. Métodos baseados em reconhecimento de padrões:
 - Busca por sítios: procura-se a presença ou ausência de uma seqüência específica, padrão ou consenso associado à expressão gênica;
 - **Busca por conteúdo:** procura-se por segmentos com **propriedades** específicas.
2. Métodos baseados em comparações por homologia com proteínas.
3. Métodos baseados no uso de *expressed sequence tags* (ESTs).

Periodicidade nas Regiões Codificantes

- As CDSs, tipicamente exibem uma **organização periódica** de três bases (TBP, *three-base periodicity*) não uniforme (latente) que não é encontrada em outras regiões [EEKR04].
- Essa **propriedade** tem sido analisada para explicar sua causa e quantificá-la [SL86].
- As frequências não uniformes do **codon usage** determinam a periodicidade. O **código genético** é responsável pelo comprimento do período [EEKR04].

Mapeamento Numérico de Nucleotídeos

Análise dos dados simbólicos de seqüências de DNA para serem tratados como seqüências numéricas.

- **Mapeamento fixo.**

	Regra	Atribuição			
		A	C	G	T
1	Ligações de hidrogênio	0	1	1	0
2	Purina/pirimidina	1	0	1	0
3	Hibrida	1	1	0	0
4	Base A	1	0	0	0
5	Base C	0	1	0	0
6	Base G	0	0	1	0
7	Base T	0	0	0	1

- Mapeamento baseados em critérios de otimização.

Mapeamento Fixo Binário

Seja a , c , g e t valores numéricos arbitrários correspondentes às bases de uma seqüência de DNA A , C , G e T .

Uma seqüência s de DNA de tamanho N pode ser representada como

$$s[b] = a.u_A[b] + c.u_C[b] + g.u_G[b] + t.u_T[b], \quad b = 0, 1, \dots, N - 1$$

em que $u_X[n]$ representa a seqüência binária associada à seqüência s .

Esta representação permite que $u_A[b] + u_C[b] + u_G[b] + u_T[b] = 1$

Análise de Fourier

A transformada de Fourier possibilita **decompor** um sinal em componentes que representem frequências.

Dado um sinal $f(t)$ a transformada de Fourier é definida como

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t} dt$$

A transformada de Fourier de tempo reduzido é definida como

$$STFT(b, \omega) = \int_{-\infty}^{\infty} g^*(t - b)f(t)e^{-j\omega t} dt$$

Espectro de Frequência de DNA

A Transformada Discreta de Fourier da seqüência s é definida como [Ana01]

$$S[k] = \sum_{b=0}^{N-1} s[b]e^{-2\pi jkb/N}, \quad k = 0, 1, \dots, N - 1$$

$$s[b] = a.u_A[b] + c.u_C[b] + g.u_G[b] + t.u_T[b]$$

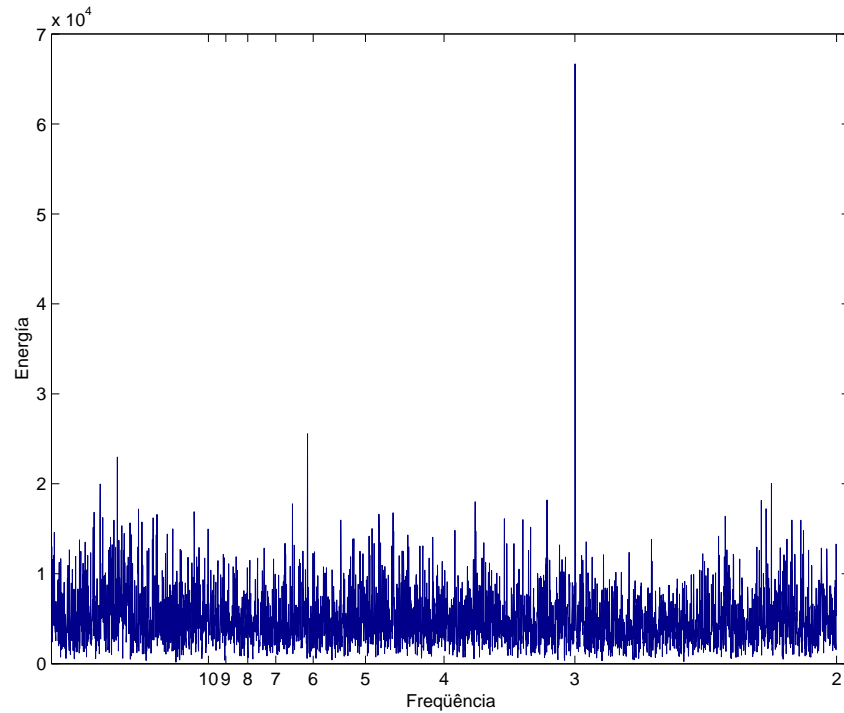
$$S[k] = a.U_A[k] + c.U_C[k] + g.U_G[k] + t.U_T[k]$$

O espectro de frequência total é representado por

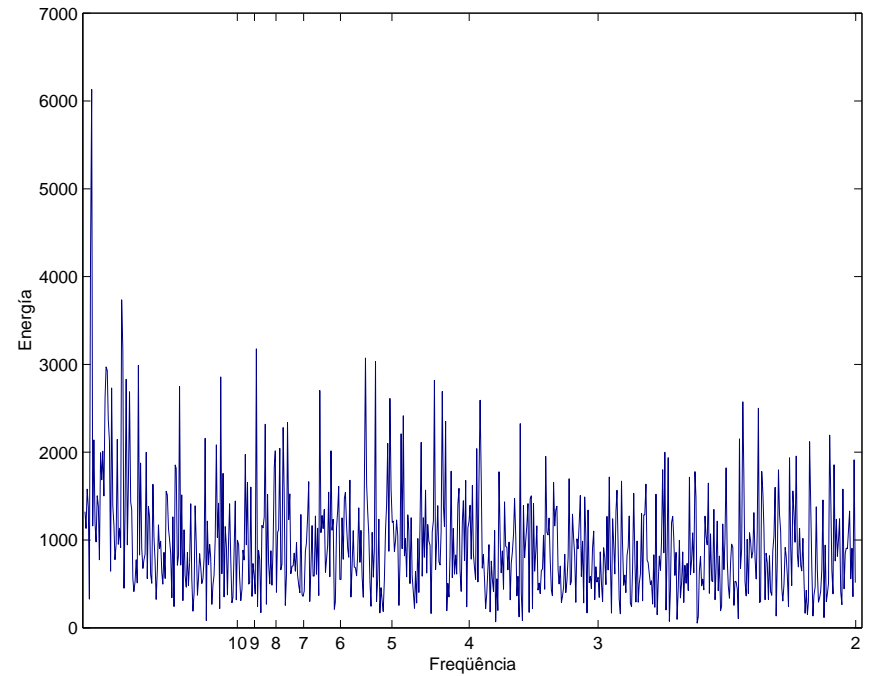
$$E[k] = |U_A[k]|^2 + |U_C[k]|^2 + |U_G[k]|^2 + |U_T[k]|^2$$

Espectro de Frequência

Arabidopsis thailana



Região codificante (CDS)



Região inter-gênica

Métodos de Fourier e DNA

Em [TRB⁺97] define-se 4 coeficientes normalizados na frequência três ($\frac{N}{3}$), como

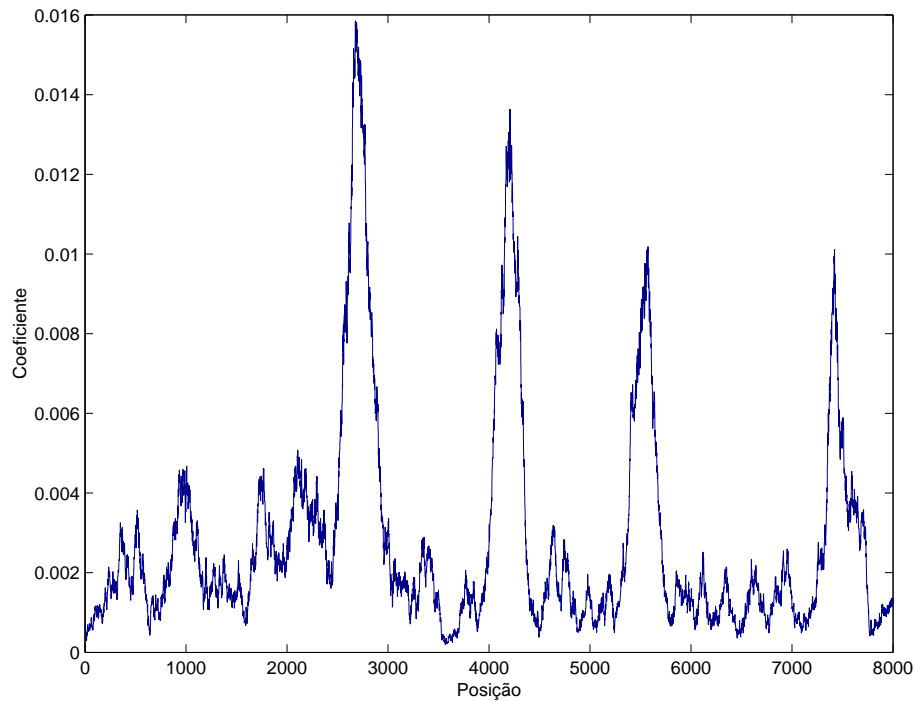
$$A = \frac{1}{N}U_A \left[\frac{N}{3} \right], \quad C = \frac{1}{N}U_C \left[\frac{N}{3} \right],$$

$$G = \frac{1}{N}U_G \left[\frac{N}{3} \right], \quad T = \frac{1}{N}U_T \left[\frac{N}{3} \right],$$

e o identificador de CDSs como

$$W = |A|^2 + |C|^2 + |G|^2 + |T|^2,$$

Métodos de Fourier e DNA



Baseado no trabalho [TRB⁺97]

Gene F56F11 de *C. elegans*

Posição relativa	Tamanho
928-1039	112
2528-2857	330
4114-4377	264
5465-5644	180
7255-7605	351

Métodos de Fourier e DNA

Em [Ana01] define-se

$$W = |a.A + c.C + g.G + t.T|^2$$

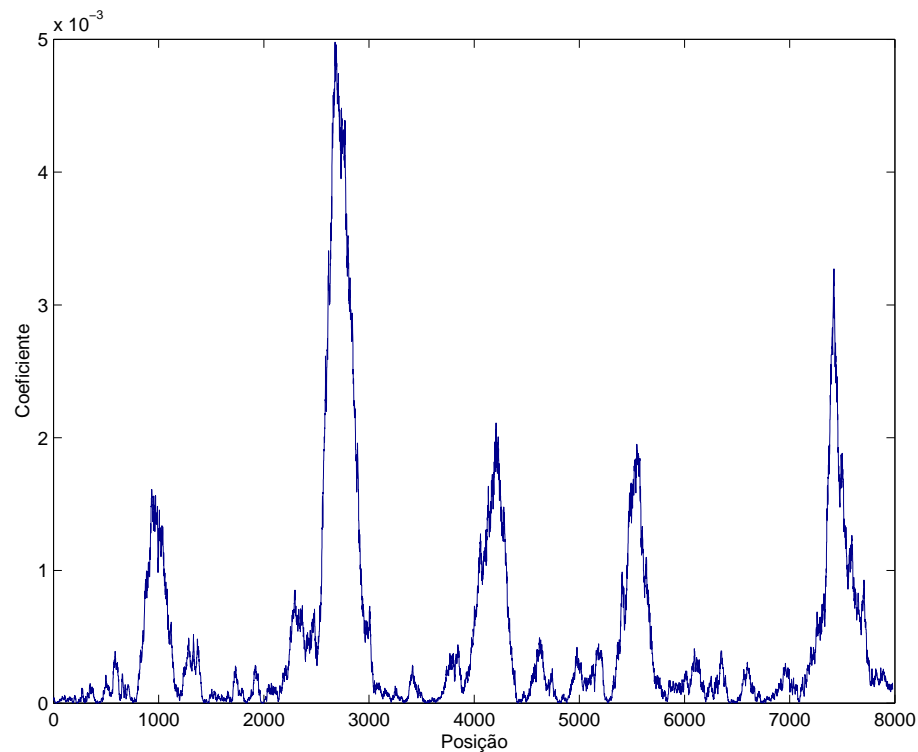
onde a , t , c e g são números complexos arbitrários tal que $A + T + C + G = 0$.

$$p(a, g, t) = \frac{E\{|a.A + t.T + g.G|\} - E\{|a.A_R + t.T_R + g.G_R|\}}{std(|a.A + t.T + g.G|) + std(|a.A_R + t.T_R + g.G_R|)}$$

Para os genes do cromossomo XVI de *S. cerevisiae*

$$a = 0.10 + 0.12j \quad c = 0 \quad g = 0.45 - 0.19j \quad t = -0.30 - 0.20j$$

Métodos de Fourier e DNA



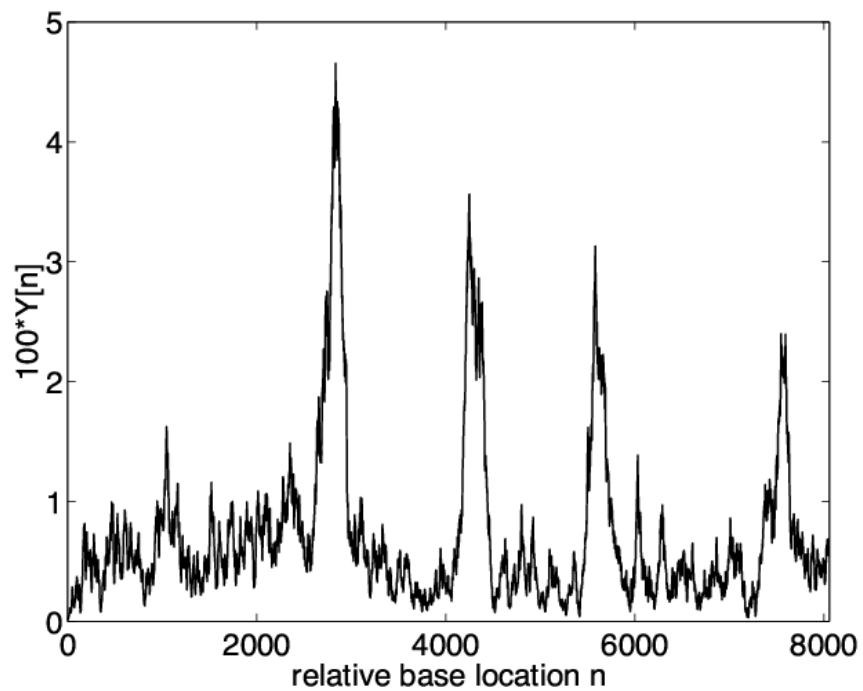
Baseado no trabalho [Ana01]

Gene F56F11 de *C. elegans*

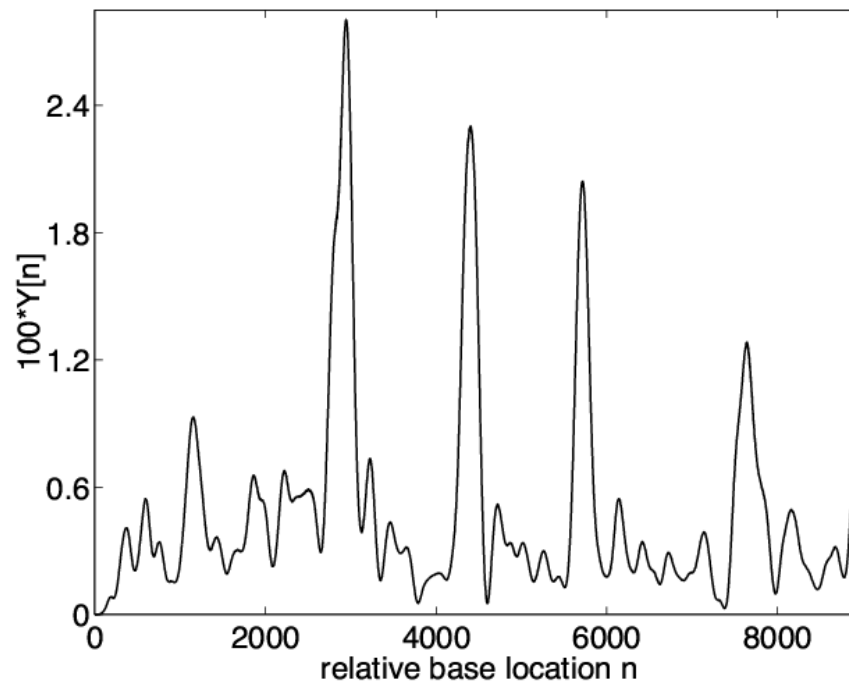
Posição relativa	Tamanho
928-1039	112
2528-2857	330
4114-4377	264
5465-5644	180
7255-7605	351

Filtros Digitais no DNA

C. Elegans



Filtro passa-banda



Filtro *multistage*

[VY04]

Análise em *Wavelets*

A transformada em *wavelets* permite uma análise tempo-escala de um sinal em termos de sinais simples (*wavelet*).

Para um sinal u a transformada em *wavelets* contínua é dada por:

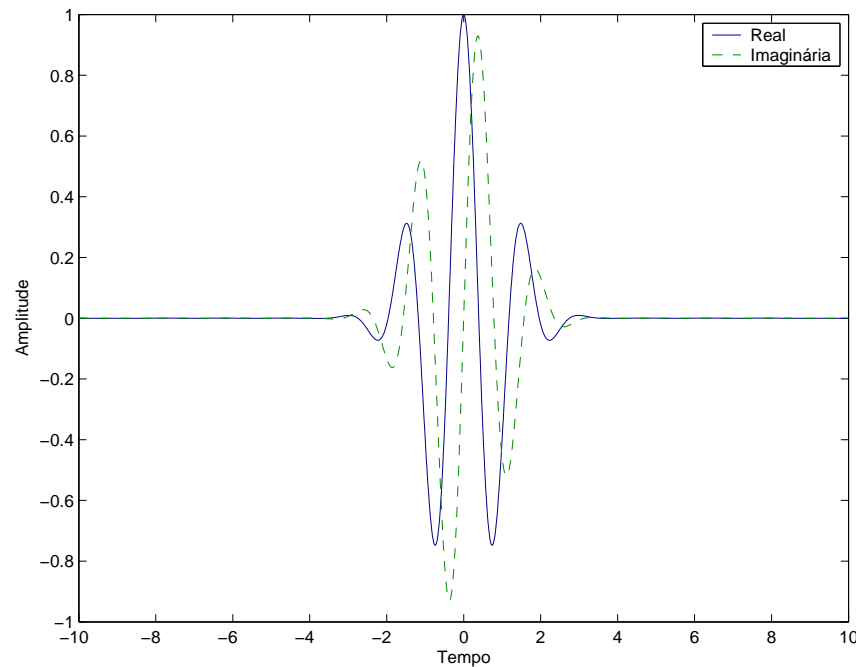
$$U(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u(t) \psi^* \left(\frac{t - b}{a} \right) dt$$

- $a > 0$ coeficiente de escala.
- b coeficiente de translação através do eixo do tempo (ou posição).
- $\psi(t)$ função de análise *wavelet*.
- $\frac{1}{\sqrt{a}}$ fator de normalização da energia.

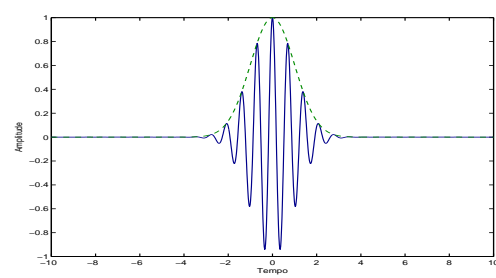
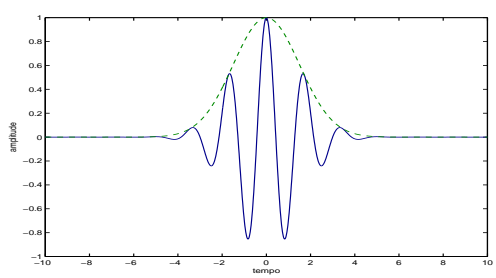
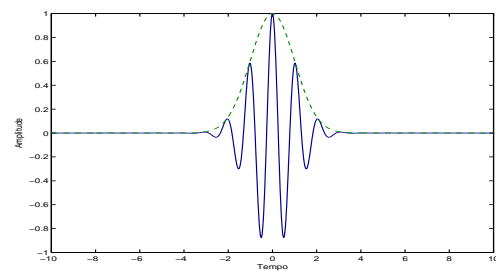
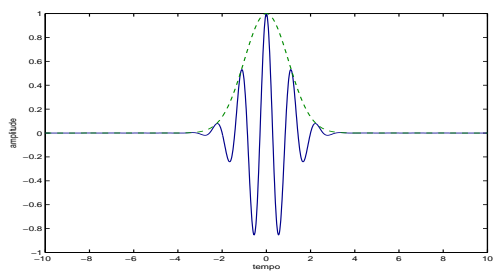
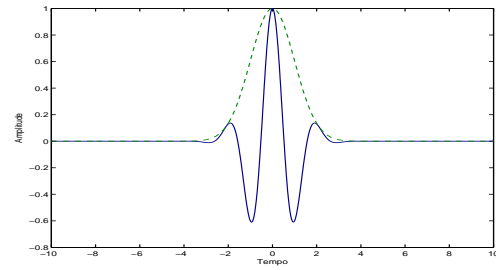
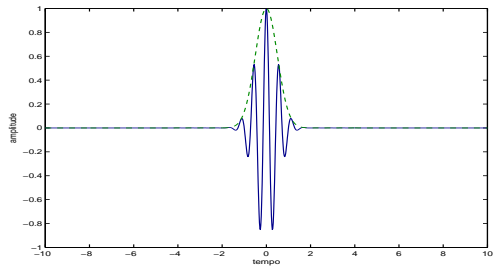
Função de Análise de Morlet

Apropriada para a análise de padrões periódicos locais, pois é bem localizada no domínio do tempo e da frequência.

$$\psi_M(t) = e^{j\omega_0 t} e^{-\frac{t^2}{2}}$$



Diferenças entre as Funções de Análise



Morlet

“Gaborettes”

$$\psi_M(t) = e^{j\omega_0 t} e^{-\frac{t^2}{2}}$$

$$G(t, a) = e^{jat} e^{-\frac{t^2}{2}}$$

Transformada Modificada de Morlet (MMT)

Da função de análise de Morlet temos que

$$\psi\left(\frac{t-b}{a}\right) = e^{j\omega_0\left(\frac{t-b}{a}\right)} e^{-\frac{\left(\frac{t-b}{a}\right)^2}{2}}$$

$$U(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u(t) e^{j\omega_0\left(\frac{t-b}{a}\right)} e^{-\frac{\left(\frac{t-b}{a}\right)^2}{2}} dt$$

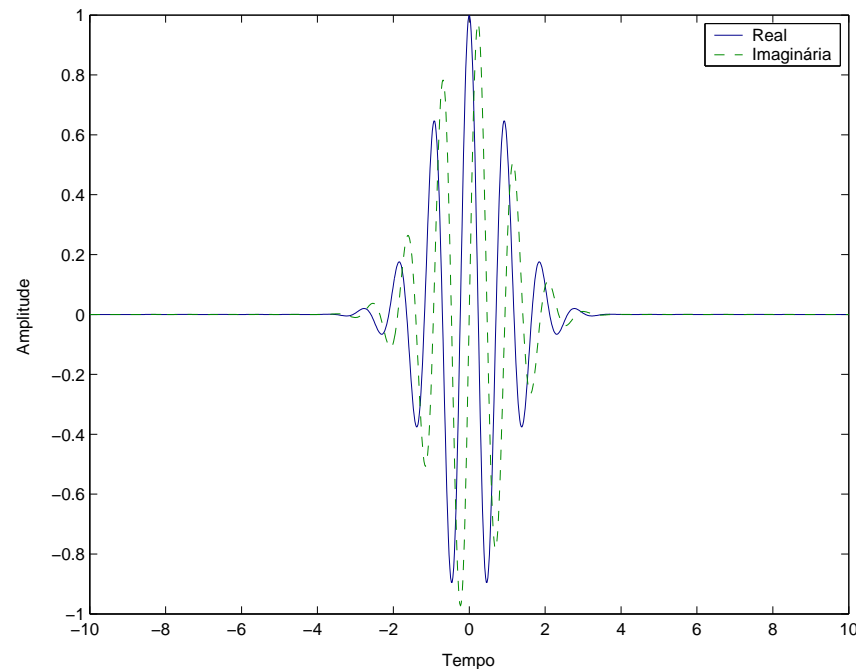
Usamos o parâmetro de escala a para manter **constante** a frequência

$$U(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u(t) e^{j\omega_0(t-b)} e^{-\frac{(t-b)^2}{2a^2}} dt$$

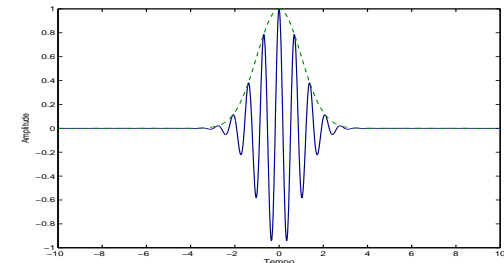
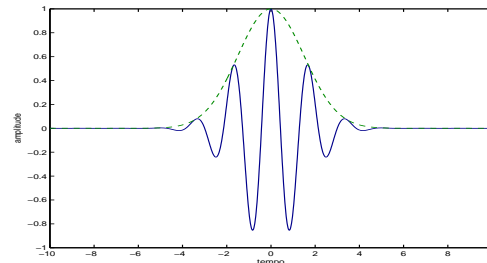
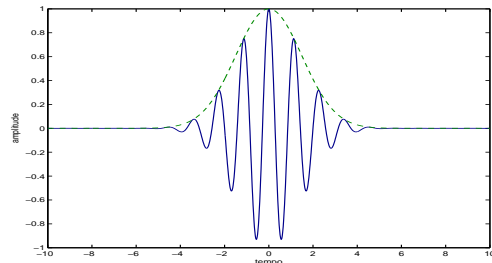
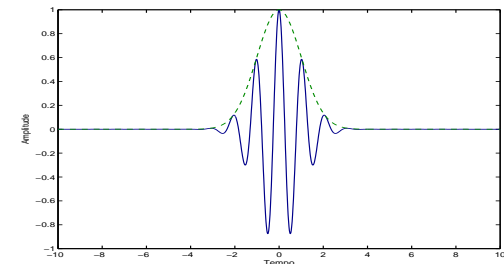
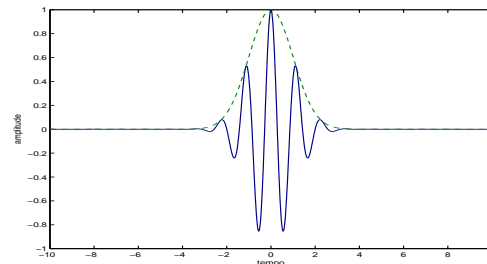
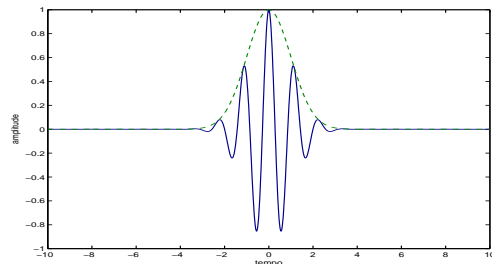
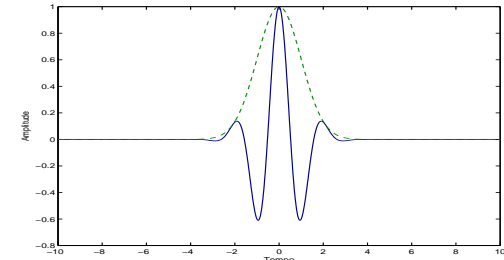
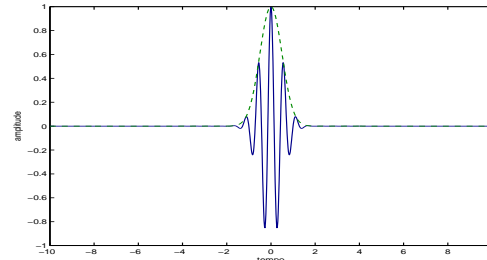
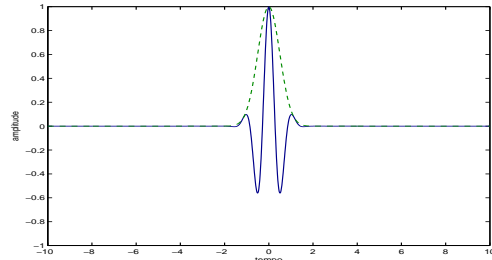
Função de Análise da MMT

Apropriada para a análise de padrões periódicos locais de frequência fixa, e de escala variável.

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



Diferenças entre as Funções de Análise



Morlet modificado

Morlet

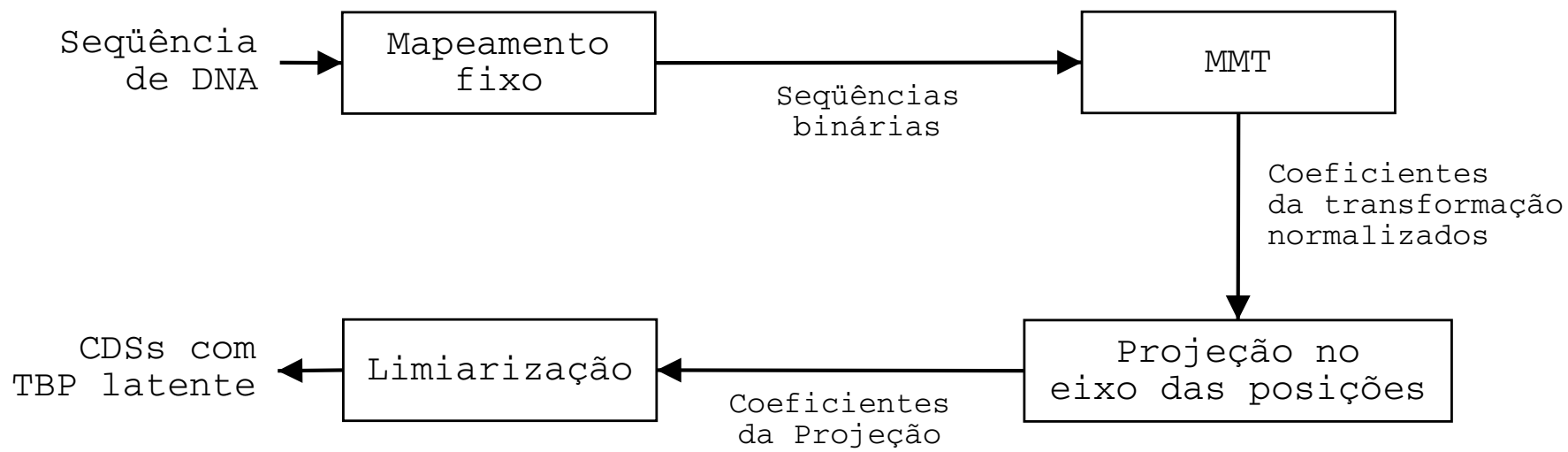
“Gaborettes”

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

$$\psi_M(t) = e^{j\omega_0 t} e^{-\frac{t^2}{2}}$$

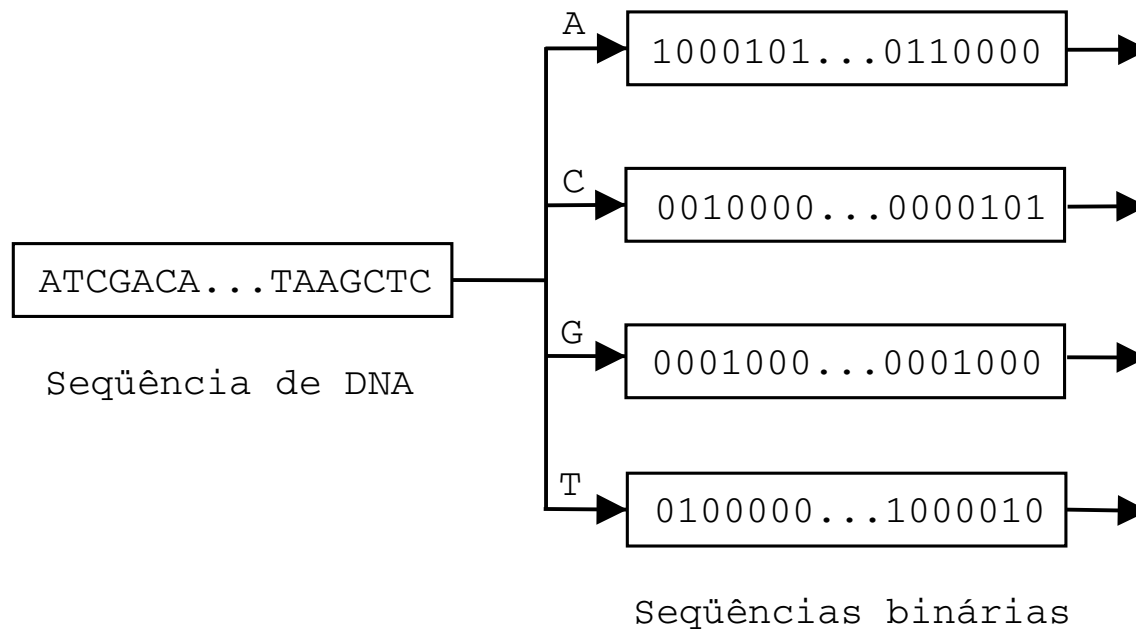
$$G(t, a) = e^{jat} e^{-\frac{t^2}{2}}$$

Novo Método para Identificação de CDSs



Mapeamento de Bases

Utiliza-se 4 regras do mapeamento fixo binário (u_A , u_C , u_G , e u_T).



Aplicação da MMT

Calcula-se a MMT das seqüências binárias, para um tamanho N arbitrário de ψ_{MM} com $\omega_0 = \frac{N}{3}$ e diferentes escalas.

$$U_A(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u_A(t) \psi_{MM}^*(t - b, a) dt$$

$$U_C(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u_C(t) \psi_{MM}^*(t - b, a) dt$$

$$U_G(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u_G(t) \psi_{MM}^*(t - b, a) dt$$

$$U_T(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u_T(t) \psi_{MM}^*(t - b, a) dt$$

Normalização dos Coeficientes

Os coeficientes são normalizados a fim de manter uma **medida comparável** em todas as escalas.

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

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

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

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

A medida normalizada total da seqüência de DNA é dada por:

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

Projeção dos Coeficientes

Os coeficientes da MMT são projetados no eixo das posições, a fim de representar as **possíveis CDSs**.

$$M_p(b) = \sum_a M(b, a), \quad 1 \leq b \leq N$$

As projeções no eixo das escalas revelam qual delas mantém maior energia através das posições

$$M_s(a) = \sum_{b=1}^N M(b, a), \quad \forall a$$

Limiarização dos Coeficientes de Projeção

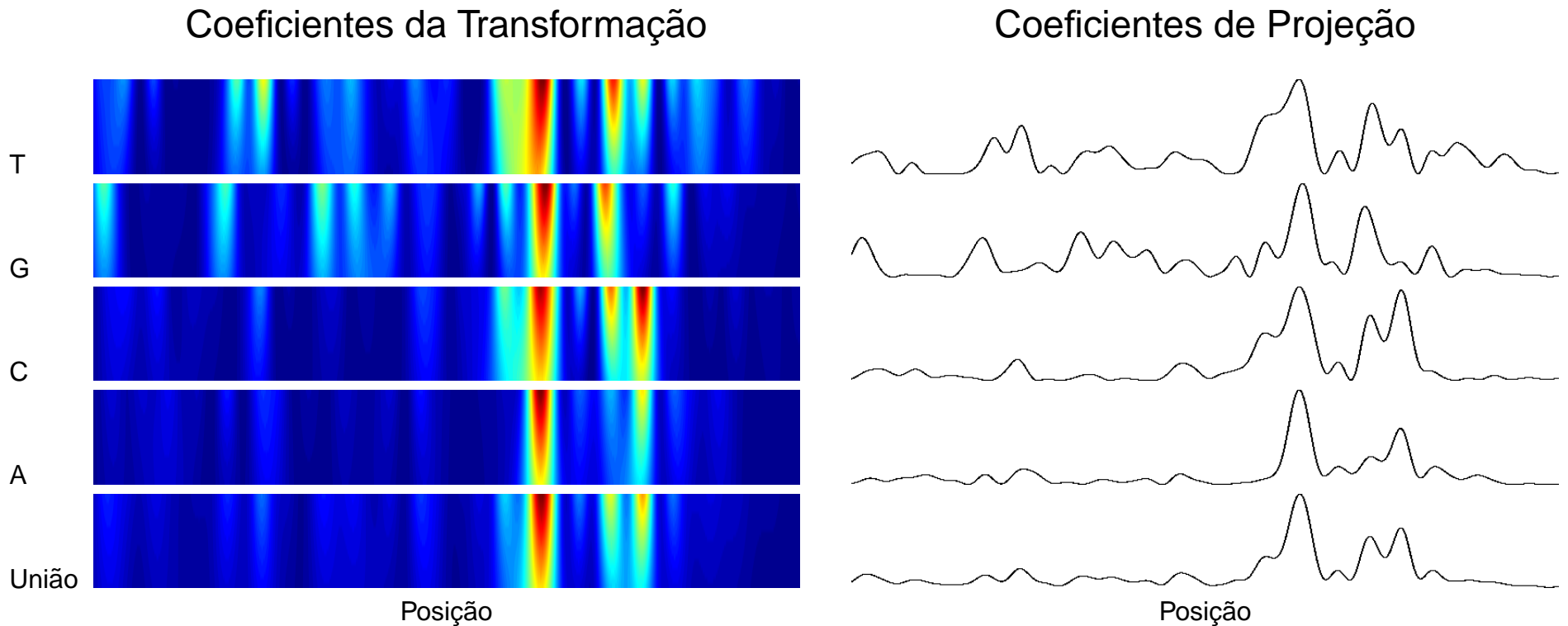
O processo da limiarização corresponde a uma tentativa de estabelecer as **fronteiras** entre as CDSs.

Uma das formas é mediante *Wavelet shrinkage*, no qual coeficientes abaixo de um limite (“erro máximo”) são **substituídos por zero**.

É considerado um **limiar percentual** nas magnitudes dos coeficientes de projeção.

A Importância da Escala

Com 20 escalas a exponencialmente espaçadas no intervalo $[0.25, 0.5]$

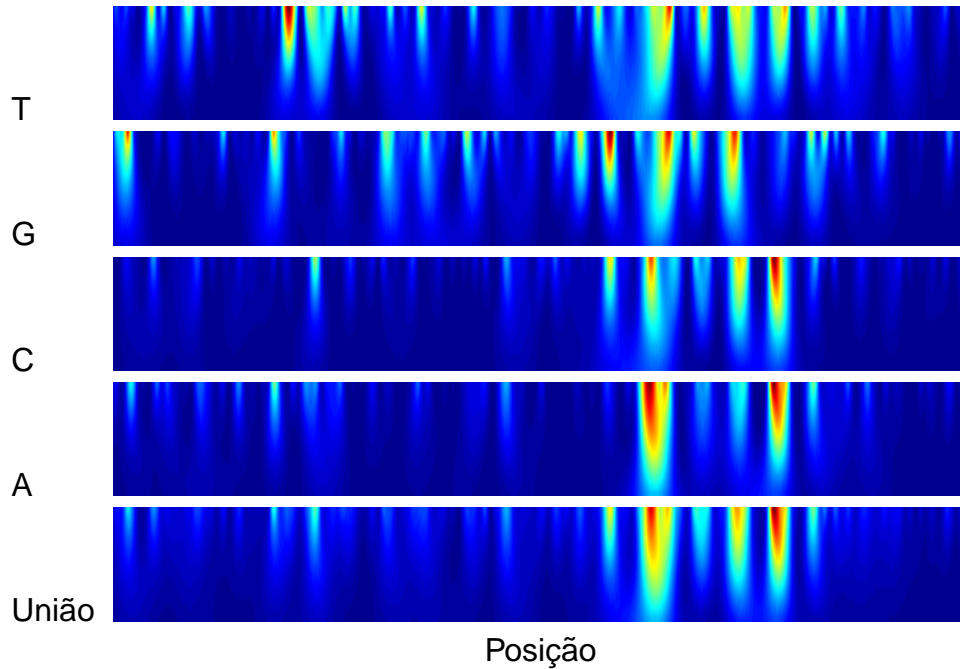


Gene BTU02285 (*Bos taurus*) de 6396bp com 6 CDSs.

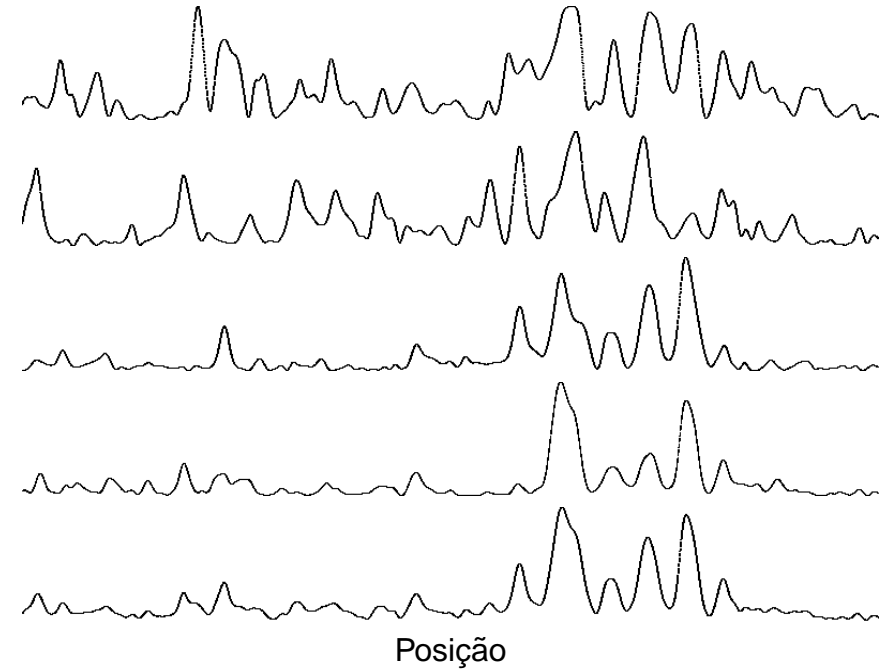
A Importância da Escala

Com 20 escalas a exponencialmente espaçadas no intervalo $[0.025, 0.5]$

Coeficientes da Transformação



Coeficientes de Projeção

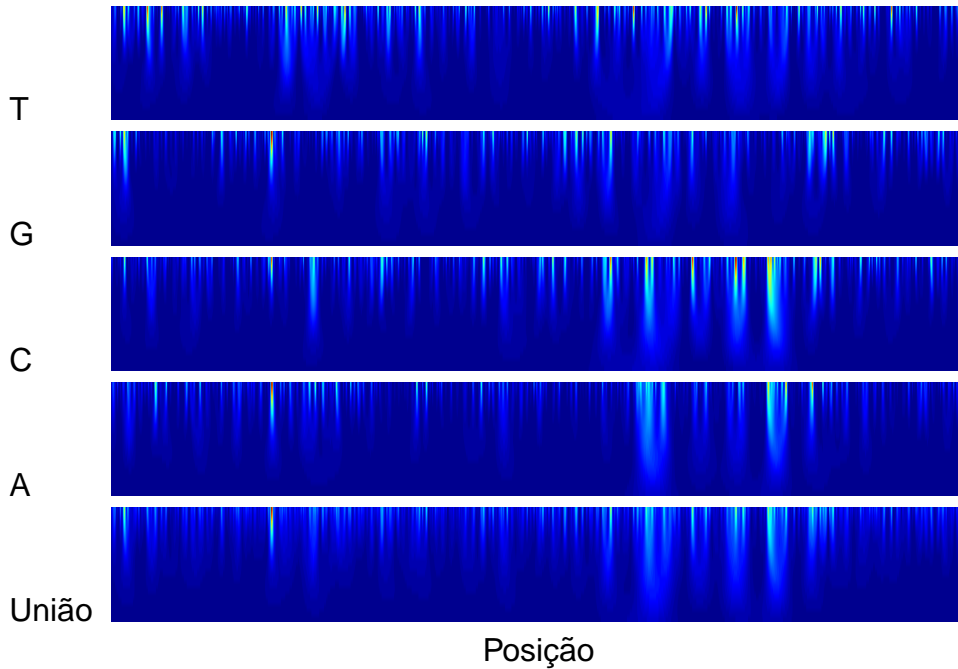


Gene BTU02285 (*Bos taurus*) de 6396bp com 6 CDSs.

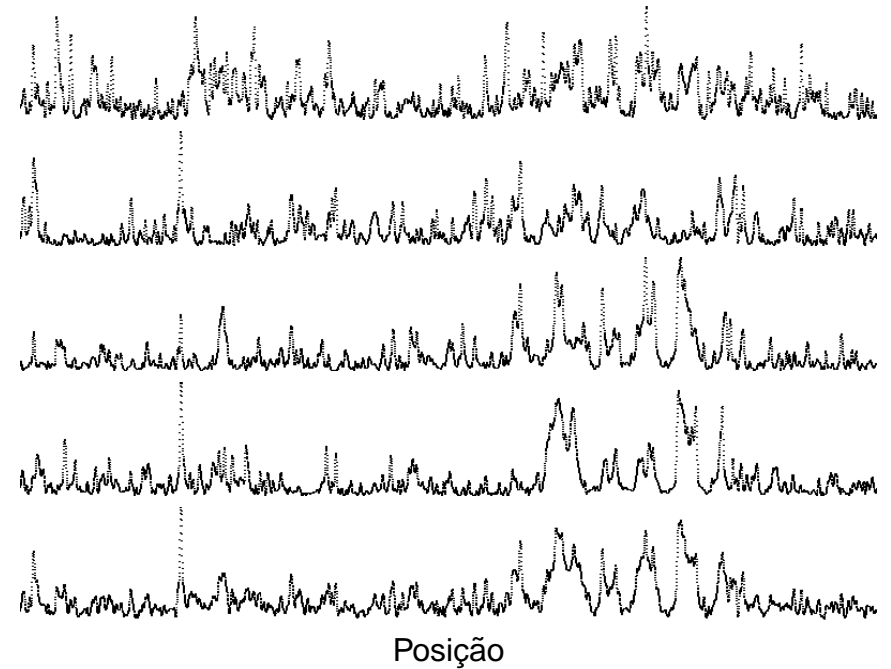
A Importância da Escala

Com 20 escalas a exponencialmente espaçadas no intervalo $[0.001, 0.5]$

Coeficientes da Transformação



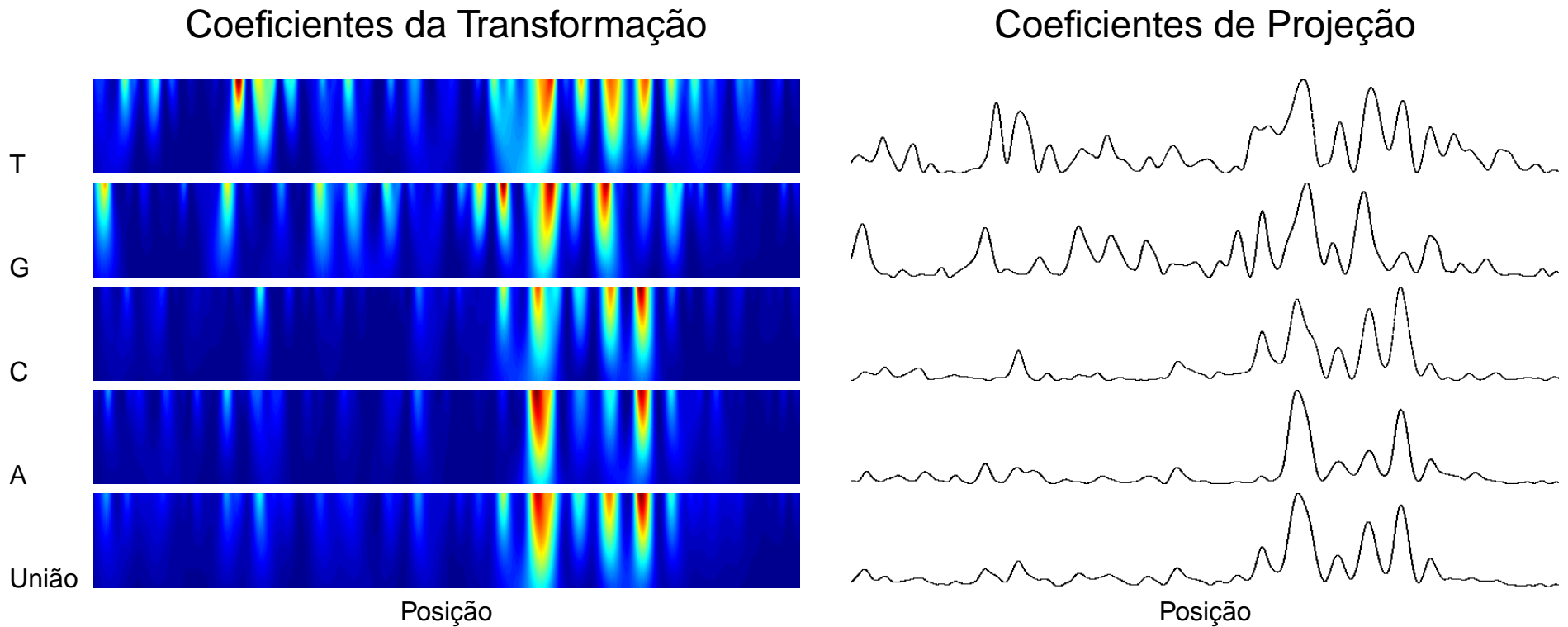
Coeficientes de Projeção



Gene BTU02285 (*Bos taurus*) de 6396bp com 6 CDSs.

A Importância da Escala

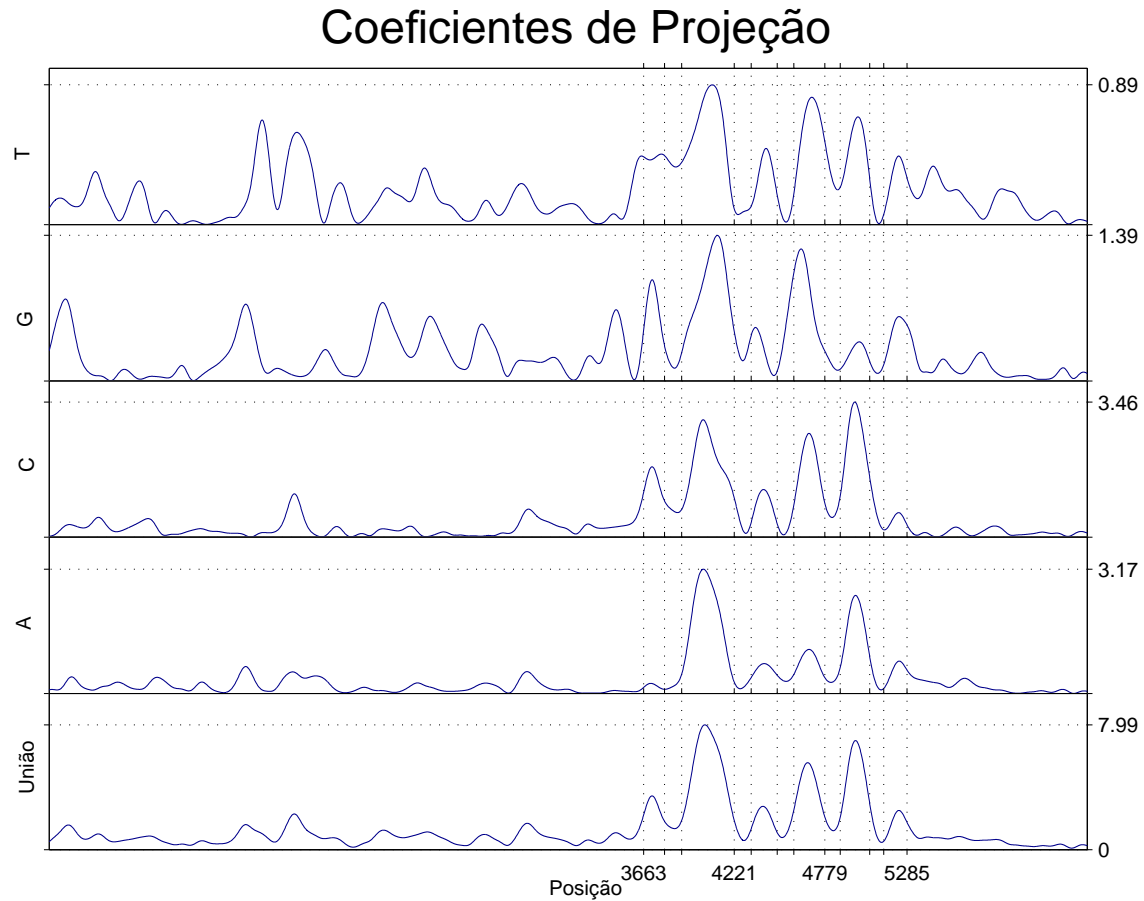
Com 20 escalas a exponencialmente espaçadas no intervalo $[0.05, 0.5]$



Gene BTU02285 (*Bos taurus*) de 6396bp com 6 CDSs.

A Importância da Projeção

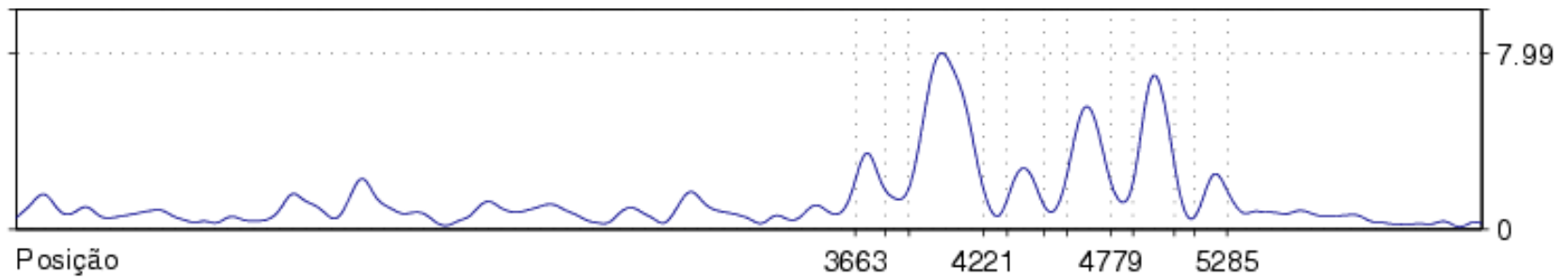
Para 20 escalas a exponencialmente espaçadas no intervalo $[0.05, 0.5]$



A Importância da Limiarização

Para 20 escalas a exponencialmente espaçadas no intervalo $[0.05, 0.5]$

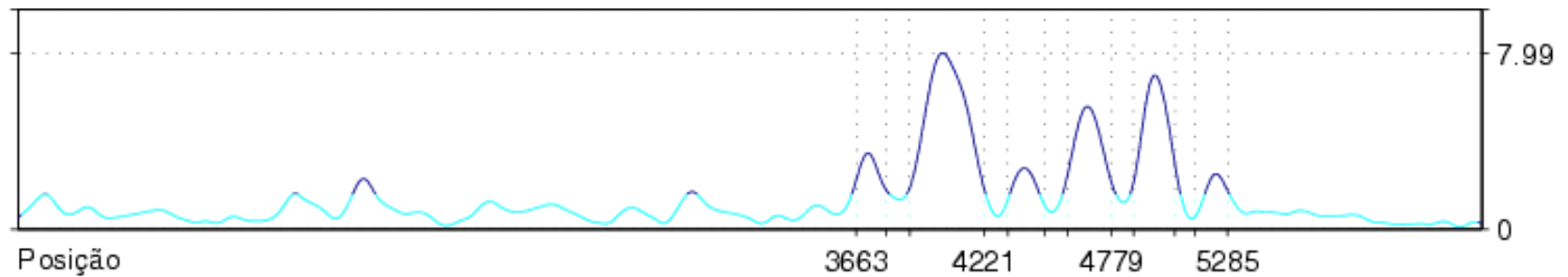
Coeficientes de Projeção da União



A Importância da Limiarização

Para 20 escalas a exponencialmente espaçadas no intervalo $[0.05, 0.5]$

Coeficientes de Projeção da União

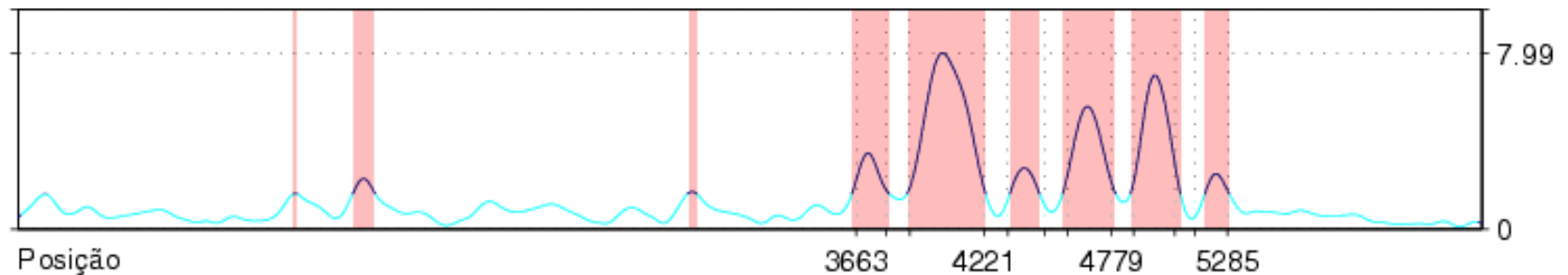


Limiarização arbitrária

A Importância da Limiarização

Para 20 escalas a exponencialmente espaçadas no intervalo $[0.05, 0.5]$

Coeficientes de Projeção da União



Limiarização arbitrária
Possíveis CDSs

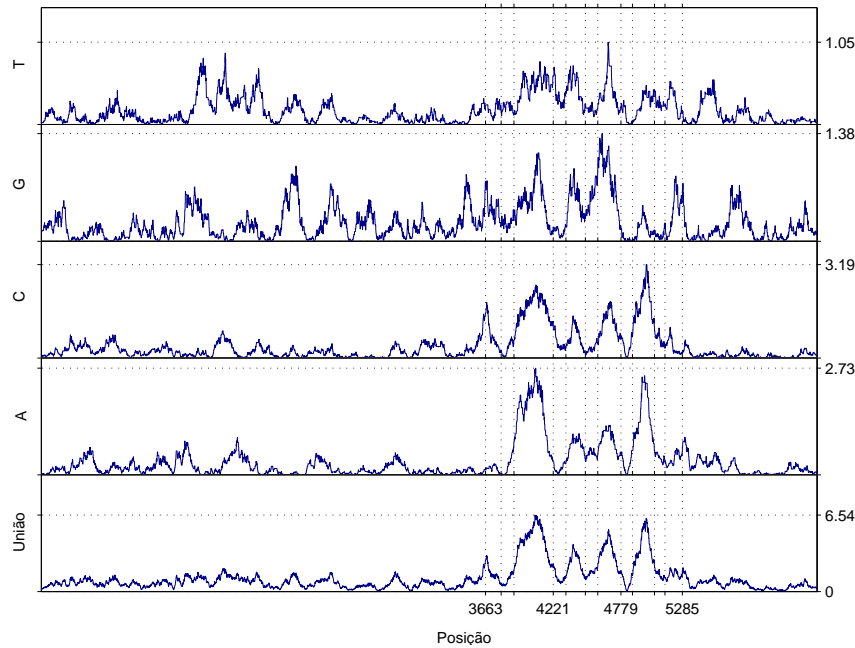
Seqüências de Teste

- Seqüências sintéticas.
- Seqüências reais.

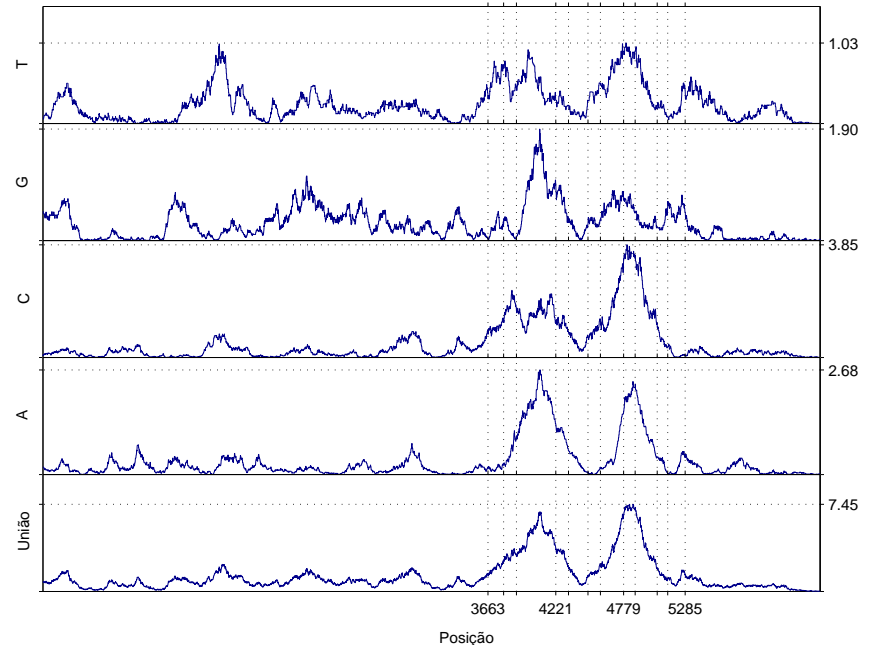
Conjunto	Região	Quantidade	Bases	Comprimento	
				Média	Desvio
A (570 seqs.)	Éxons	2649	444498 (15.4%)	168	222
	Íntrons	2079	1310452 (45.3%)	630	909
	Inter-gênicas	1132	1137199 (39.3%)	1004	1464
B (195 seqs.)	Éxons	948	199176 (14.4%)	210	271
	Íntrons	753	642788 (46.4%)	854	130
	Inter-gênicas	390	544044 (39.2%)	1395	2261

Testes Preliminares

Coeficientes da STFT gene BTU02285



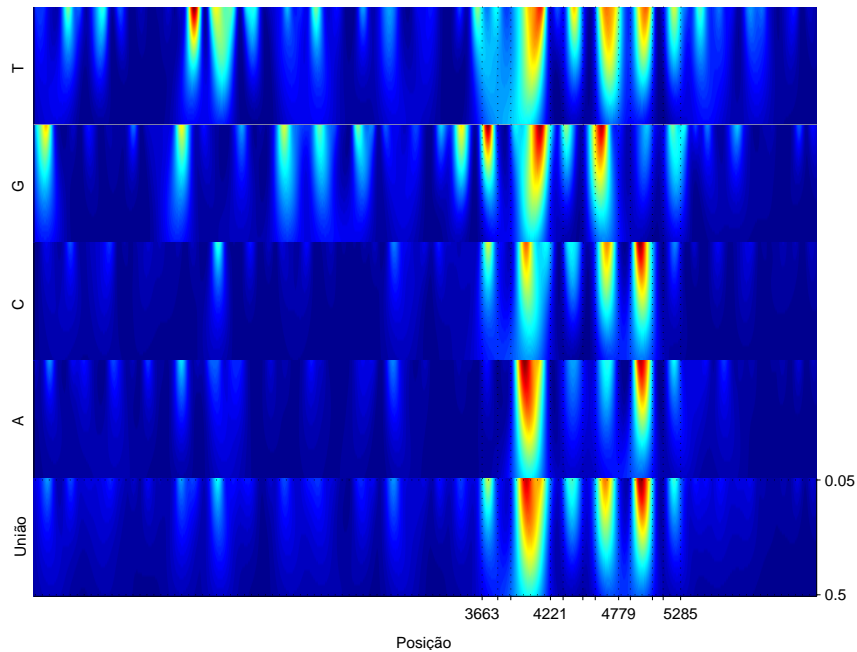
Janela de 200bp



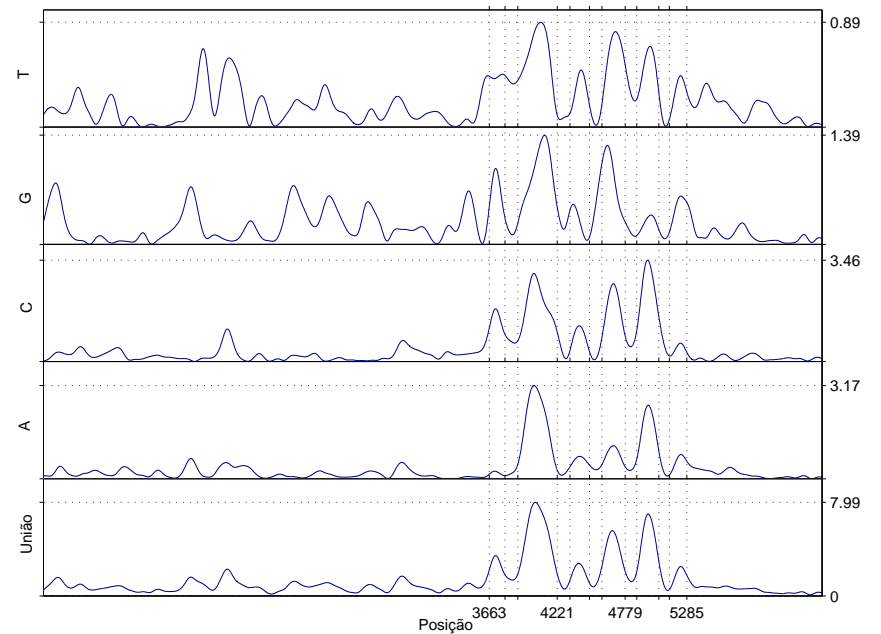
Janela de 400bp

Testes Preliminares

Gene BTU02285



Coeficientes MMT normalizados

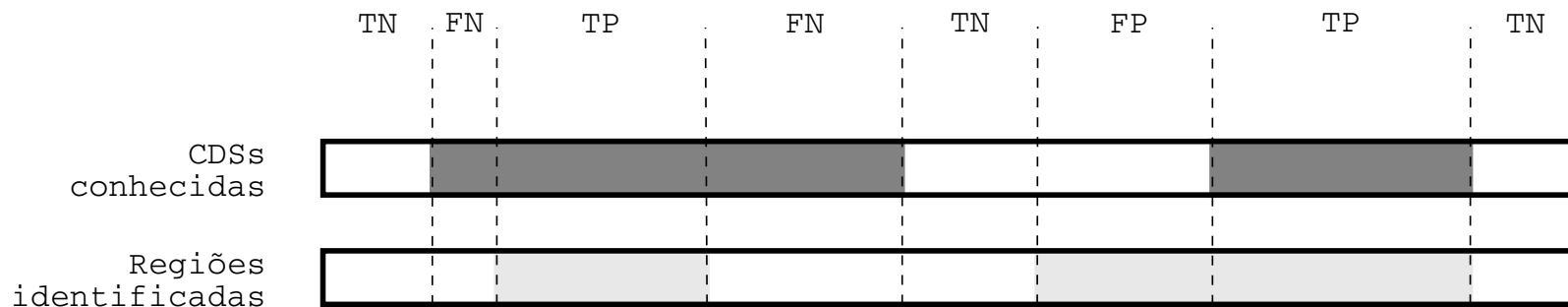


Coeficientes de projeção

Medidas de Acurácia

As medidas de acurácia no nível dos nucleotídeos [BG96], propõem uma **forma de comparação** de regiões identificadas com CDSs conhecidas biologicamente.

A medição de regiões identificadas contra CDSs conhecidas é feita mediante contagem de nucleotídeos.



Medidas de Acurácia

- **Sensibilidade** (S_n), proporção de nucleotídeos codificantes corretamente identificados como codificantes.

$$S_n = \frac{TP}{TP+FN}$$

- **Especificidade** (S_p), proporção de nucleotídeos identificados como codificantes que são realmente codificantes.

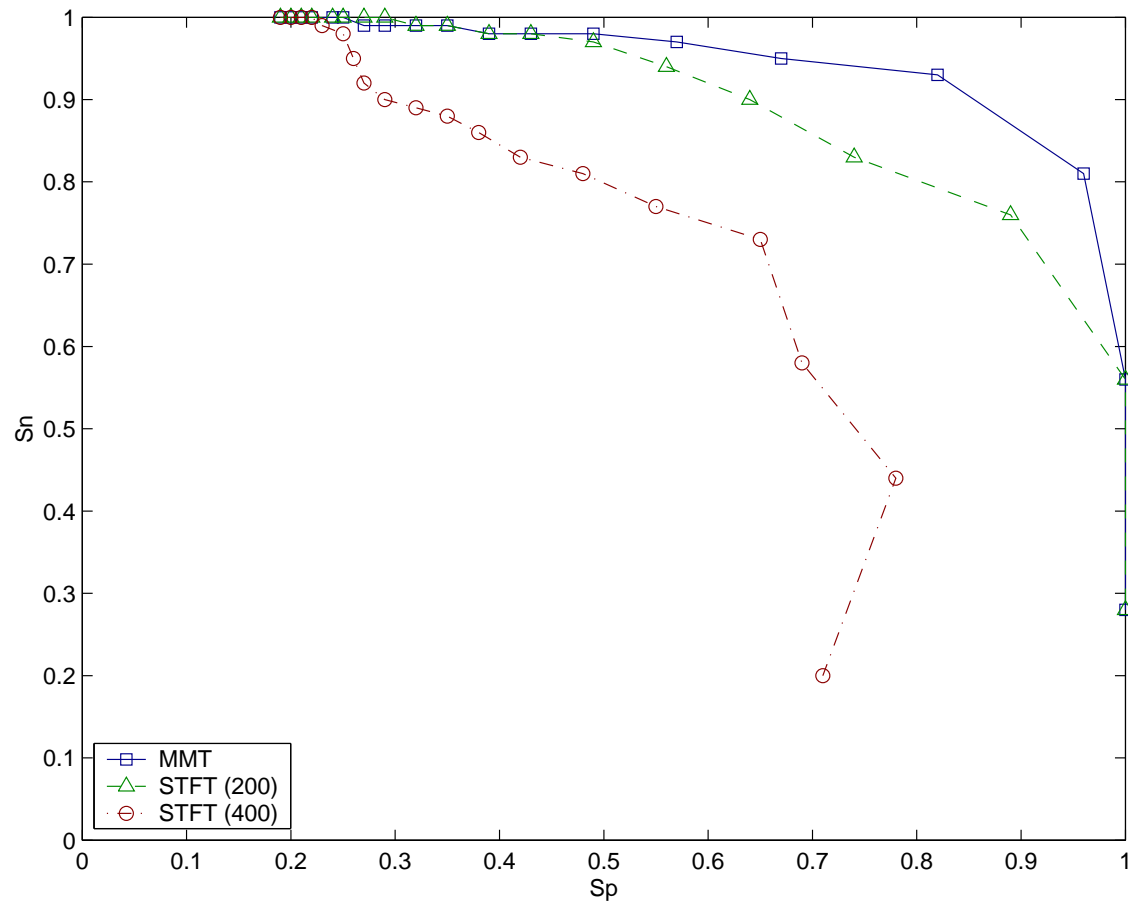
$$S_p = \frac{TP}{TP+FP}$$

- **Coeficiente de correlação** (CC), medida que combina a S_n e S_p .

$$CC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}$$

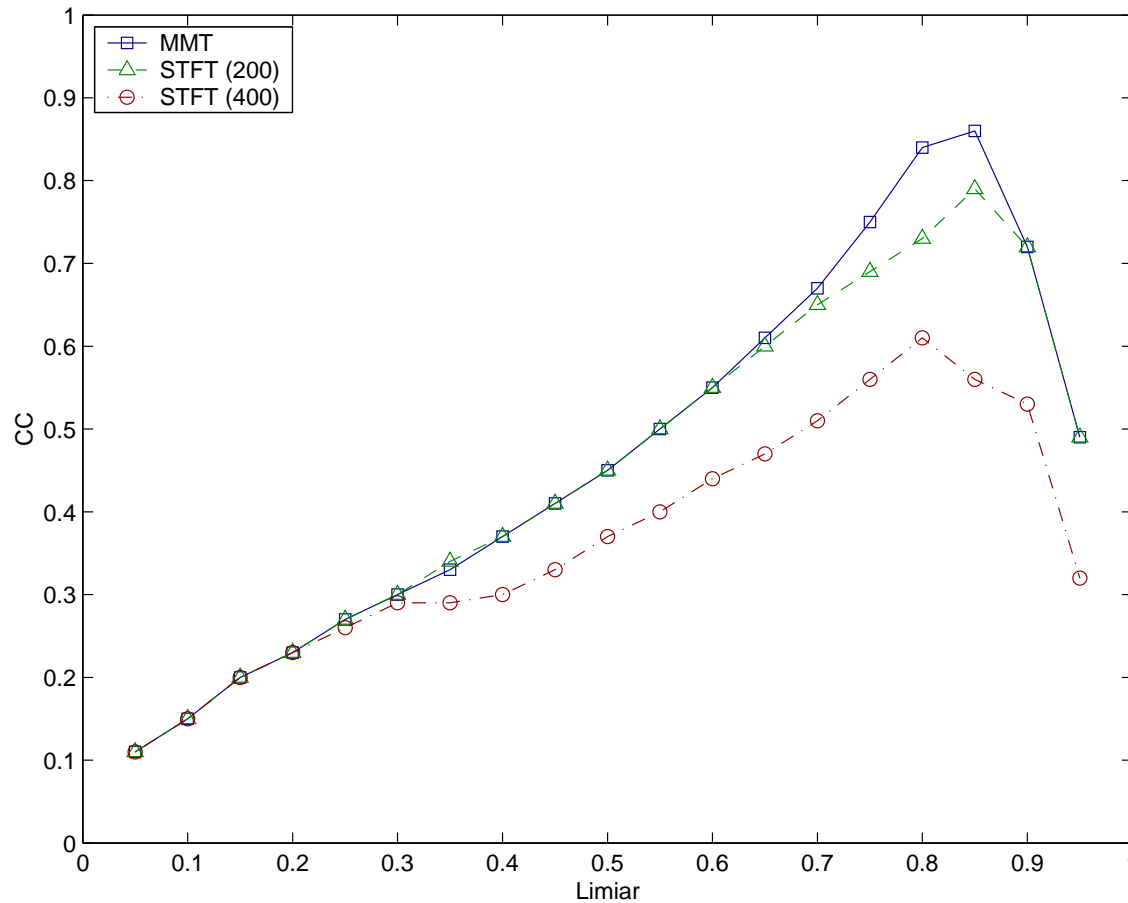
Resultados Preliminares

Desempenho da MMT e da STFT para o gene BTU02285



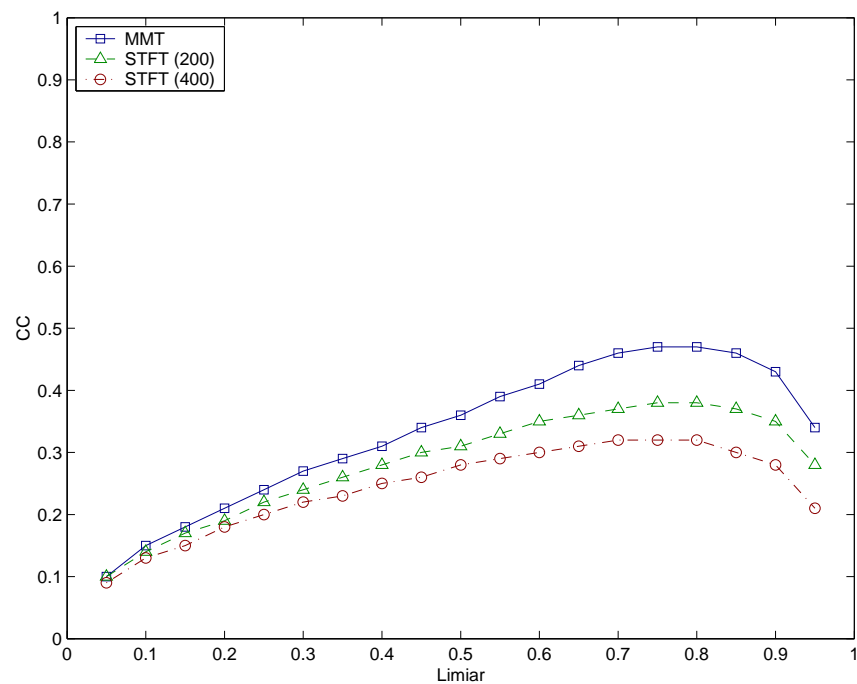
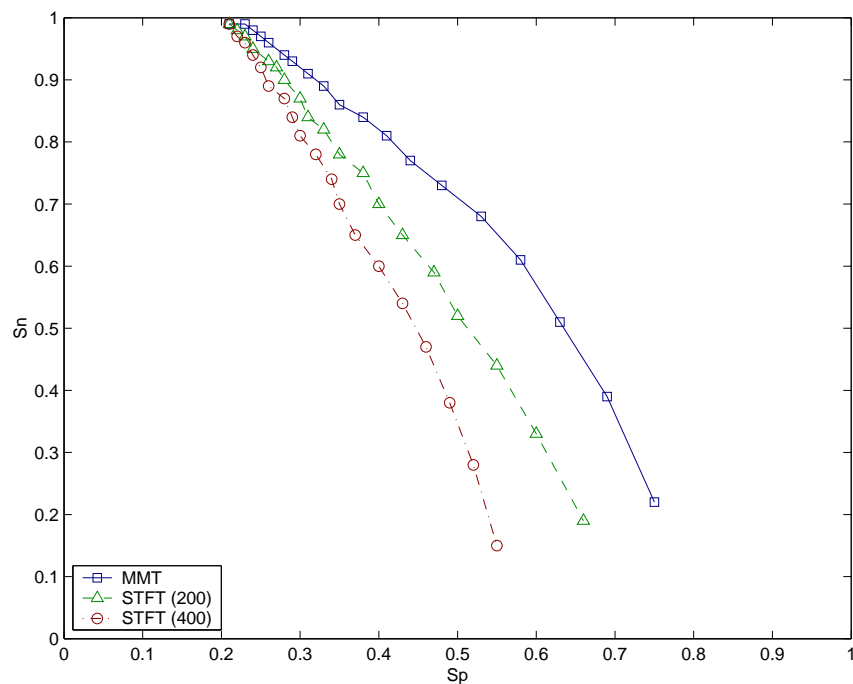
Resultados Preliminares

Desempenho da MMT e da STFT para o gene BTU02285



Resultados Preliminares

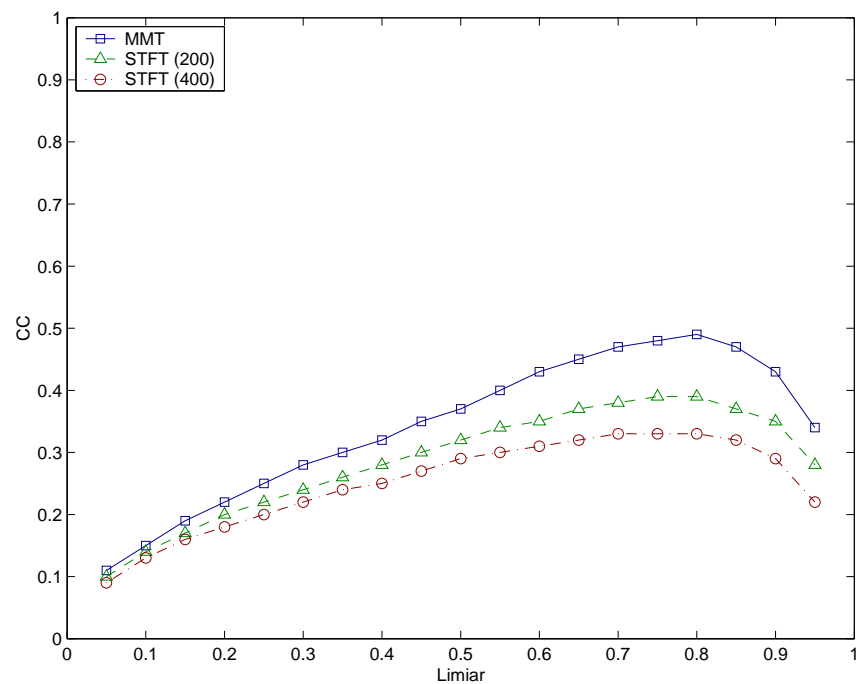
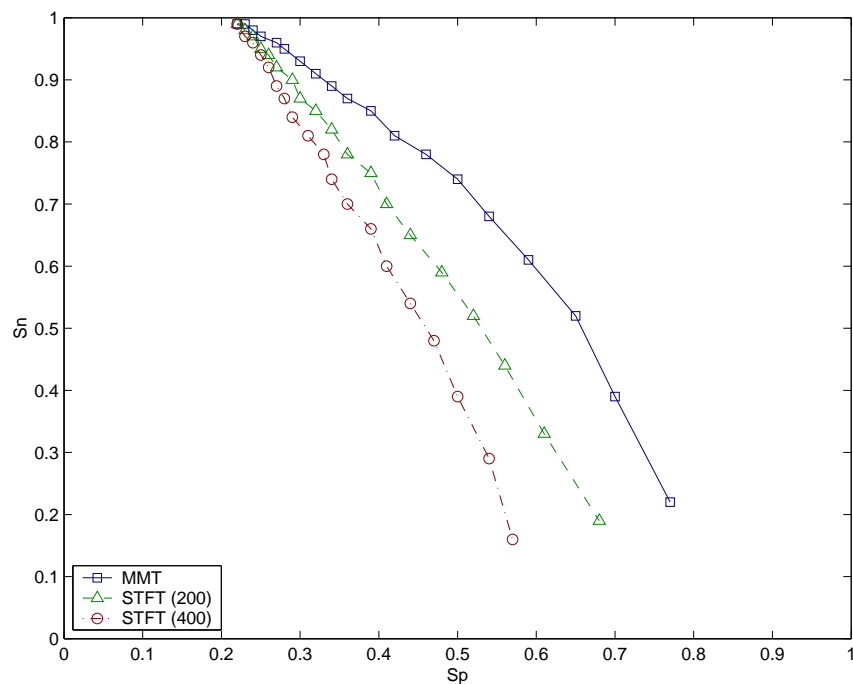
Desempenho da MMT e da STFT para o conjunto A



570 seqüências

Resultados Preliminares

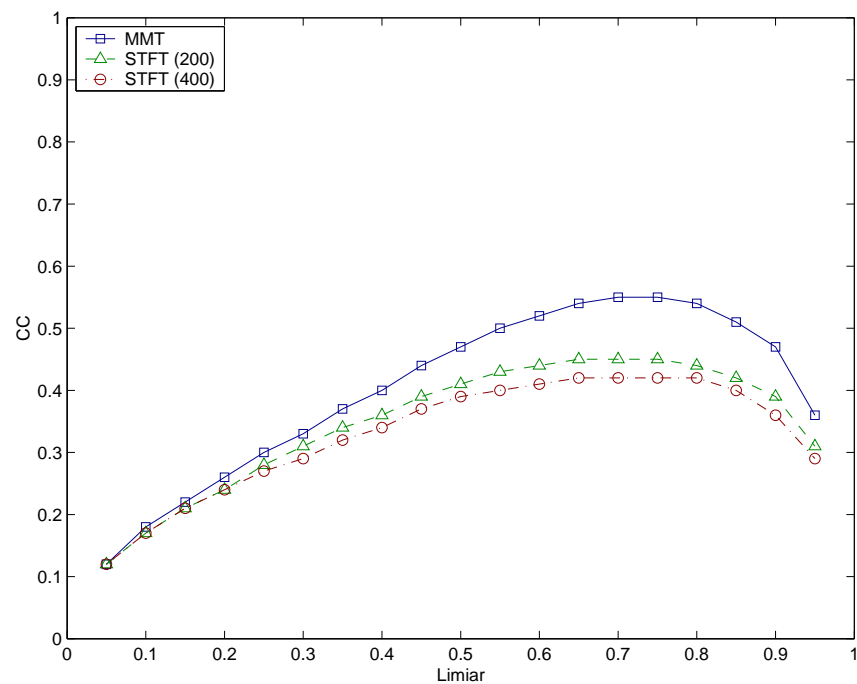
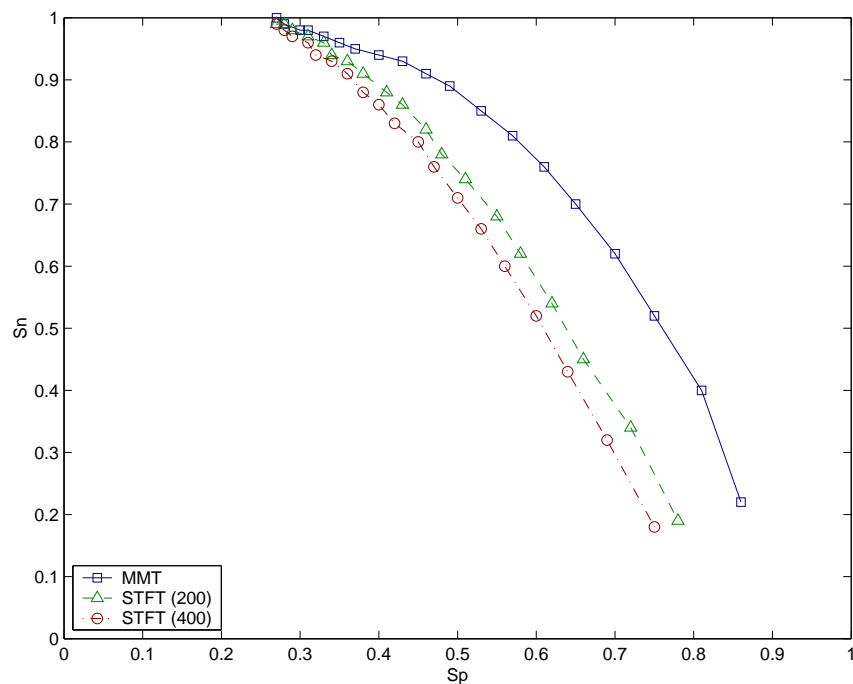
Desempenho da MMT e da STFT para conjunto Am30



469 seqüências. Comprimento dos éxons maiores a 30bp.

Resultados Preliminares

Desempenho da MMT e da STFT para conjunto Am100



103 seqüências. Comprimento dos éxons maiores a 100bp.

Conclusões

- A MMT tem um desempenho superior à STFT. Um nível de acurácia maior é alcançado quando os comprimentos das CDSs são maiores a 100bp.
- O método basea-se unicamente na TBP existente nas CDSs. Não é usada nenhuma outra informação adicional.
- Este novo método é mais robusto à variação de escalas.

Contribuições

- Desenvolvimento de um ***pipeline* bioinformático** para projetos genoma.
- Definição de uma **nova transformada** de análise de padrões periódicos locais de frequência fixa.
- Introdução de um **novo método** para a identificação de regiões codificantes de proteína.

Pesquisas Futuras

- Outras transformadas em *wavelets*, como a *wavelets packets*, baseadas em critérios matemáticos ou de entropia estatística podem ser pesquisadas.
- Os coeficientes de projeção podem ser considerados indicadores da probabilidade de região ser codificante.
- Estudo da representação da interação existente entre as bases de DNA.
- Estudo de medidas de acurácia não restritas à determinação de S_n e S_p , considerando a distribuição do número de CDSs por gene, comprimento dos CDSs identificados.

Referências

- [Ana01] D. Anastassiou. Genomic signal processing. *IEEE Signal Processing Magazine*, 8(4):8–20, 2001.
- [BG96] M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–367, 1996.
- [EEKR04] S. T. Eskesen, F. N. Eskesen, B. Kinghorn, and A. Ruvinsky. Periodicity of DNA in exons. *Journal Molecular Biology*, 5(12):1–11, 2004.
- [SL86] B. D. Silverman and R. Linsker. A measure of DNA periodicity. *Journal of Theoretical Biology*, 118(3):295–300, 1986.
- [TRB⁺97] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy. Prediction of probable genes by Fourier analysis of genomic sequences. *Bioinformatics*, 13(3):263–270, 1997.
- [VY04] P. P. Vaidyanathan and B. Yoon. The role of signal-processing concepts in genomics and proteomics. *Journal of the Franklin Institute*, 341(1-2):111–135, 2004.