**Identificação de regiões codificantes
de proteína através da
transformada modificada de Morlet**

Jesús Pascual Mena-Chalco

Dissertação apresentada ao
Instituto de Matemática e Estatística da
Universidade de São Paulo
para obtenção do título de
Mestre em Ciências

Área de Concentração: Ciência da Computação
Orientador: Prof. Dr. Roberto Marcondes Cesar Junior

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, setembro de 2005

**Identificação de regiões codificantes
de proteína através da
transformada modificada de Morlet**

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Jesús Pascual Mena-Chalco
e aprovada pela Comissão Julgadora.

São Paulo, 26 de novembro de 2005

Banca Examinadora:

- Prof. Dr. Roberto Marcondes Cesar Junior (orientador) - IME-USP.
- Prof. Dr. Roberto Hirata Júnior - IME-USP.
- Profa. Dra. Helaine Carrer - ESALQ-USP.

Agradecimentos

Com muita saudade, e temor de esquecer de alguém, posso escrever que este trabalho é o resultado dos quase vinte e seis meses na minha vida de pós-graduação no IME-USP. Meses sempre cheios do que realmente gosto: “da vida”, vida que não é a minha.

Gostaria de expressar minha gratidão profunda à minha família: Emílio, Júlia, Rafael, Senovia e Madelyn, que deram-me muito amor, imenso sustento e incentivo. Não importaria o lugar, sempre sentiria saudades de vocês. À minha família em São Paulo: Christian, Cristian, Gustavo e Karina, pelo constante apoio e por tudo o que compartilhamos.

Agradeço profundamente ao meu orientador, e amigo, professor Roberto Marcondes Cesar Junior por dar-me a oportunidade de pesquisar nesta área e fundamentalmente por acreditar em mim; por explicar-me, com todos seus comentários, a concentrar-me as vezes no *como* e as vezes no *porque*. É ele quem tem-me conduzido na pesquisa, sempre me suportou e me orientou tanto como eu podia aprender sobre minha ignorância, sendo sempre um exemplo para mim.

Sou grato aos professores Helaine Carrer e Roberto Hirata Júnior por todas as recomendações e correções dadas na versão prévia deste trabalho e no exame de qualificação. As idéias sugeridas foram de muita utilidade na continuação do trabalho. Também reitero meus agradecimentos por terem aceitado participar como a presente banca de defesa de minha dissertação de mestrado.

Sou grato ao Yossi Zana e ao David da Silva Pires pelas correções, grande ajuda crítica nos meus trabalhos, pela sempre boa disposição para tudo, muitas ensinanças e grande amizade. Agradeço, também, ao professor Paulo José da Silva e Silva pelo grande apoio inicial, exemplo e amizade. Gostaria de agradecer à professora Helaine Carrer, Henrique, Adriano, Valesca e todo o pessoal do laboratório de Biotecnologia Agrícola do Departamento de Ciências Biológicas da ESALQ-USP, por dar-me a oportunidade de conhecer mais sobre biologia molecular e bioinformática. Aos professores

Wilber Ramos, Juan Carlos Juarez, Iván Montes e Ernesto Cuadros pelo apoio quando eu o precisava antes de começar o mestrado.

Agradeço ao André Yoshiaki Kashiwabara, Henrique Alves, Said Sadique Adi e Vladimir Moreira pela amizade, sugestões, correções, e frutíferas conversas que animaram-me muito no estudo deste trabalho. Aos meus amigos Ana Beatriz, Anderson, Ariane, Celina, Claudia, Clodis, Cristian Noriega, Daniel Dantas, Daniel Vaquero, Daniel Vieira, David Junior, Eduardo Guerra, Eduardo Ueda, Elier, Fábio, Gustavo, Jishu, João, Jorge, Kelly, Lourdes, Ricardo, Silvia, Thiago Santos, Thiago Sousa e Walter obrigado pela amizade, compreensão e muitas ensinanças. Não esqueço de todos os meus demais amigos que têm feito cheia de experiências a vida toda. Agradeço especialmente à Elizabeth pela muita paciência, carinho e todo o mundo implícito.

No final mas não por último, esta dissertação nunca teria sido concebida sem o apoio financeiro da CAPES, professores, estudantes, admins e funcionários do IME-USP. Que mais linhas para escrever e dar um “muito obrigado!”?.

Resumo

Um tópico importante na análise de seqüências biológicas é a busca de genes, ou seja, a identificação de regiões codificantes de proteínas. Esta identificação permite a posterior procura de significado, descrição ou categorização biológica do organismo analisado. Atualmente, vários métodos combinam reconhecimento de padrões com conhecimento coletado de conjuntos de treinamento ou de comparações com banco de dados genômicos. Entretanto, a acurácia desses métodos está ainda longe do satisfatório. Novos métodos de processamento de seqüências de DNA e de identificação de genes podem ser criados através da busca por conteúdo (*search-by-content*). O padrão periódico de DNA em regiões codificantes de proteína, denominada periodicidade de três bases, vem sendo considerado uma propriedade dessas regiões. As técnicas de processamento digital de sinais fornecem uma base robusta para a identificação de regiões com periodicidade de três bases.

Nesta dissertação, são apresentados um *pipeline* bioinformático, os conceitos básicos da identificação genômica, e métodos de processamento digital de sinais utilizados para a identificação de regiões codificantes de proteínas. Introduzimos um novo método para a identificação dessas regiões, baseado na transformada proposta, denominada Transformada Modificada de Morlet. Apresentamos vários resultados experimentais obtidos a partir de seqüências de DNA sintéticas e reais. As principais contribuições do trabalho consistem no desenvolvimento de um *pipeline* bioinformático para projetos genoma e na criação de um método de identificação de regiões codificantes onde a periodicidade de três bases seja latente. O método apresenta desempenho superior e vantagens importantes em comparação ao método tradicional baseado na transformada de Fourier de tempo reduzido.

Palavras-chave: identificação de genes, periodicidade nos éxons, transformada modificada de Morlet, processamento digital de sinais, *pipeline*, bioinformática.

Abstract

An important topic in biological sequences analysis is gene finding, i.e. the identification of protein coding regions. This identification allows the posterior research for meaning, description or biological categorization of the analyzed organism. Currently, several methods combine pattern recognition with knowledge collected from training datasets or from comparison with genomic databases. Nonetheless, the accuracy of these methods is still far from satisfactory. New methods of DNA sequences processing and genes identification can be created through search-by-content such sequences. The periodic pattern of DNA in protein coding regions, called three-base periodicity, has been considered proper of coding regions. Digital signal processing techniques supply a strong basis for regions identification with three-base periodicity.

In this work, we present a bioinformatics pipeline, basic concepts of the genomic identification and digital signal processing methods used for protein coding regions identification. We introduce a new method for identification of these regions, based on a newly proposed transform, called Modified Morlet Transform. We present some obtained experimental results from synthetic and real DNA sequences. The main contributions consist of the bioinformatics pipeline development for genoma projects and the creation of a method for protein coding regions identification where the three-base periodicity is latent. The method presents superior performance and important advantages in comparison to traditional method based on the short time Fourier transform.

Keywords: genes identification, periodicity in exons, modified Morlet transform, digital signal processing, pipeline, bioinformatics.

Sumário

Lista de Abreviaturas	xv
Lista de Símbolos	xvii
Lista de Figuras	xix
Lista de Tabelas	xxiii
1 Introdução	1
1.1 Considerações Preliminares	1
1.2 Objetivos	2
1.3 Contribuições	3
1.4 Organização do Trabalho	3
2 Conceitos Biológicos	5
2.1 Fundamentos Biológicos	5
2.1.1 Ácidos Nucléicos	5
2.1.2 Aminoácidos	6
2.1.3 Genes	7
2.1.4 Códigos Genéticos e Proteínas	9

2.2	Sequenciamento Genômico	10
2.3	Periodicidade nas CDSs	11
2.4	Identificação de Genes	12
3	<i>Pipeline</i> Bioinformático	15
3.1	Unidades Funcionais	15
3.1.1	Identificação de Bases	16
3.1.2	Montagem do Genoma	17
3.1.3	Edição das Leituras	18
3.1.4	Anotação do Genoma	18
3.2	<i>Pipeline</i> Bioinformático Desenvolvido	19
3.2.1	Arquitetura e Implementação	19
3.2.2	Relatórios de Análise	22
4	Métodos de DSP para a Identificação de CDSs	23
4.1	Análise de Fourier	23
4.1.1	Transformada de Fourier	23
4.1.2	Espectro de Frequência	24
4.1.3	Convolução	25
4.1.4	Transformada de Fourier de Tempo Reduzido	25
4.2	Mapeamento Numérico de Nucleotídeos	26
4.3	Espectro de Frequência de DNA	28
4.4	Análise de Periodicidade de DNA	32
4.4.1	Periodicidade em Sequências Sintéticas	33
4.4.2	Periodicidades em Sequências Reais	34

4.5	Identificação de CDSs	34
4.6	Medidas de Acurácia de Identificação de Genes	41
5	Identificação de CDSs através da MMT	43
5.1	Análise em <i>Wavelets</i>	43
5.2	Transformada Modificada de Morlet	44
5.3	Método para a Identificação de CDSs através da MMT	49
5.3.1	Mapeamento	49
5.3.2	Aplicação da MMT	50
5.3.3	Projeção dos Coeficientes da MMT	51
5.3.4	Limiarização dos Coeficientes de Projeção	52
6	Resultados Experimentais	53
6.1	Testes em Sequências Sintéticas de DNA	54
6.2	Testes em Sequências Reais de DNA	55
6.3	Testes em Conjuntos de Sequências de DNA	57
7	Conclusões	71
7.1	Considerações Finais	71
7.2	Sugestões para Pesquisas Futuras	72
A	Sequências de DNA Testadas no Trabalho	75
A.1	Sequências Sintéticas	75
A.2	Sequências Reais	76
B	Periodicidades em Sequências Sintéticas com Ruído	81

C Tabelas de Acurácia estimadas para Seqüências de DNA	85
D Painéis Apresentados	93
D.1 International Conference on Bioinformatics and Computational Biology 2004	93
D.2 X-Meeting 2005	94
Referências Bibliográficas	97
Índice Remissivo	102

Lista de Abreviaturas

bp	Par de base (<i>Base pair</i>).
BAC	Cromossomo artificial bacteriano (<i>Bacterial artificial chromosome</i>).
CDS	Região codificante de proteína (<i>Coding sequence, protein coding sequence</i>).
DNA	Ácido desoxirribonucleico (<i>Desoxyribonucleic acid</i>).
DSP	Processamento Digital de Sinais (<i>Digital signal processing</i>).
EIIP	Potencial de interação elétron-íon (<i>electron-ion interaction potentials</i>).
EST	Seqüência expressa (<i>Expressed sequence tag</i>).
mRNA	RNA mensageiro (<i>Messenger RNA</i>).
MMT	Transformada modificada de Morlet (<i>Modified Morlet transform</i>).
RNA	Ácido ribonucleico (<i>Ribonucleic acid</i>).
STFT	Tranformada de Fourier de tempo reduzido (<i>Short-time Fourier transform</i>).
TBP	Periodicidade de três bases (<i>Three-base periodicity</i>).
YAC	Cromossomo artificial de levedura (<i>Yeast artificial chromosome</i>).

Lista de Símbolos

$*$	Complexo conjugado.
a	Escala.
j	Número imaginário, $\sqrt{-1}$.
s	Sinal (discreto ou contínuo).
t	Tempo ou posição.
ω	Frequência angular.
ψ	Função de análise <i>wavelet</i> .
Ψ	Transformada de Fourier de ψ .
$ \cdot $	Valor absoluto de um número.
A	Base adenina.
C	Base citocina.
G	Base guanina.
T	Base Timina.
U	Base Uracila.
N	Qualquer base (A, C, G, ou T).
$\mathcal{C}$	Conjunto de números complexos.
$\mathcal{R}$	Conjunto de números reais.
S	Transformada de Fourier de s .
E	Esperança.
u_X	Seqüência binária associada às posições da base X.
U_X	Transformada de Fourier da seqüência u_X .
arg	Argumento de um número complexo.
std	Desvio padrão.

Lista de Figuras

2.1	Representação do DNA como uma fita dupla complementar e antiparalela.	6
2.2	Representação do Dogma Central da Biologia Molecular.	9
3.1	Organização típica de um <i>pipeline</i> bioinformático para projetos genoma. Cada bloco representa uma unidade funcional e cada seta o fluxo da informação.	16
3.2	Representação da identificação de bases de leituras de três fragmentos arbitrários. As bases com valor de baixa qualidade são denotadas com letras minúsculas.	17
3.3	Interface web inicial do <i>pipeline</i> bioinformático desenvolvido.	20
3.4	Esquema da arquitetura do <i>pipeline</i> bioinformático desenvolvido.	21
4.1	Representação gráfica da sequência de DNA: ATGCTTGACTAGGGCTCAGT. Em (a) são consideradas quatro seqüências binárias correspondentes às regras 4 a 7 do mapeamento binário. Em (b) considera-se uma seqüência numérica com valores de (A, C, G, T)=(1,2,3,4).	29
4.2	Espectros de frequência de seqüências sintéticas de 600bp que contêm regiões com periodicidade três: GCT GCT . . . (direita). Representações gráficas das seqüências de DNA (esquerda).	31
4.3	Espectros de frequência de seqüências sintéticas de 600bp que contêm regiões com periodicidade três: GCT GCT . . . (direita). Representações gráficas das seqüências de DNA (esquerda). (<i>continuação</i>)	32

4.4	Espectro de frequências de seqüências. Em (a) e (b) mostram-se os coeficientes da transformada de Fourier para um conjunto de seqüências sintéticas de 600bp em que foram percentualmente inseridos dois tipos de ruído. Em (c) e (d) mostram-se os coeficientes da transformada de Fourier para seqüências reais.	33
4.5	Periodicidades estimadas para seqüências de DNA. Foram calculadas as periodicidades 2 a 24 das regiões codificantes (que simulam TBP) e regiões não-codificantes (regiões inter-gênicas e, no caso dos organismos eucariotos, íntrons).	35
4.6	Coeficientes W estimados, para seqüências de DNA, com um conjunto de tamanhos de janela (esquerda), e com um tamanho de janela determinado (direita).	37
4.7	Coeficientes W , segundo [4], estimados para o gene F56F11.4, com um conjunto de tamanhos de janela (esquerda), e com um tamanho de janela determinado (direita).	39
4.8	Os quatro possíveis resultados de regiões identificadas contra CDSs conhecidas.	41
5.1	Representações da transformada modificada de Morlet (esquerda) e suas correspondentes transformadas de Fourier (direita). Foram considerados $\omega_0 = 6.6$, t no intervalo $(-10, 10)$ e escalas a de 0.5, 1 e 4.	47
5.2	Representações das funções de análise. Em (a) varia-se a frequência da exponencial complexa, mantendo constante o desvio padrão da Gaussiana. Em (b) varia-se o desvio padrão da Gaussiana juntamente com a frequência da exponencial complexa. Em (c) varia-se o desvio padrão da Gaussiana, mantendo constante a frequência da exponencial complexa.	48
5.3	Diagrama de blocos do novo método utilizando a MMT para a identificação de CDSs onde a periodicidade de três bases seja latente. Cada bloco representa uma operação e cada seta o fluxo da informação.	49
6.1	Desempenho da MMT e da STFT para a seqüência <code>sintética1_0</code>	59
6.2	Desempenho da MMT e da STFT para a seqüência <code>sintética2_0</code>	60
6.3	Desempenho da MMT e da STFT para o gene BTU02285.	61
6.4	Desempenho da MMT e da STFT para o gene F56F11.4.	62

6.5	Desempenho da MMT e da STFT para o gene GGVITIIG.	63
6.6	Desempenho da MMT e da STFT para o gene HSDAO.	64
6.7	Desempenho da MMT e da STFT para o gene HUMCS3.	65
6.8	Desempenho da MMT e da STFT para o gene MMACLGNA.	66
6.9	Desempenho da MMT e da STFT em termos de Sn e Sp (esquerda), e de CC e limiar (direita) para os conjuntos A, Am30 e Am100.	67
6.10	Desempenho da MMT e da STFT em termos de Sn e Sp (esquerda), e de CC e limiar (direita) para os conjuntos B, Bm30 e Bm100.	68
6.11	Histogramas dos comprimentos dos éxons (esquerda), e dos íntrons (direita) para os conjuntos A, Am30 e Am100 identificados através da MMT.	69
6.12	Histogramas dos comprimentos dos éxons (esquerda), e dos íntrons (direita) para os conjuntos B, Bm30 e Bm100 identificados através da MMT.	70
A.1	Histogramas dos comprimentos dos éxons (esquerda), e dos íntrons (direita) para os conjuntos A, Am30 e Am100.	79
A.2	Histogramas dos comprimentos dos éxons (esquerda), e dos íntrons (direita) para os conjuntos B, Bm30 e Bm100.	80
B.1	Periodicidades estimadas para seqüências sintéticas com ruído. Foram calculadas as periodicidades 2 a 24 das regiões periódicas (que simulam TBP) e regiões não-periódicas (regiões que simulam às regiões inter-gênicas e, no caso dos organismos eucariotos, íntrons).	82
B.2	Periodicidades estimadas para as seqüências sintéticas com ruído. Foram calculadas as periodicidades 2 a 24 das regiões periódicas (que simulam TBP) e regiões não-periódicas (regiões que simulam às regiões inter-gênicas e, no caso dos organismos eucariotos, íntrons). (<i>continuação</i>)	83

Lista de Tabelas

2.1	Códigos, abreviaturas e nomes dos aminoácidos.	7
2.2	Código genético padrão e de bactérias/plastídeos de plantas. Os códons de iniciação estão representados pela letra M e os códons de parada pelo símbolo *.	10
4.1	Regras para o mapeamento binário de seqüências de DNA. No mapeamento é associada uma seqüência binária para cada regra que indica uma característica ou posição de cada base.	27
4.2	Seqüências binárias que representam as posições de A, C, G e T, em uma seqüência de DNA, correspondentes as regras 4 a 7 do mapeamento binário.	28
4.3	Éxons do gene F56F11.4.4 de <i>C. elegans</i> . NCBI, número de acesso AF099922 e região 7021-15020.	38
A.1	Posições das regiões periódicas na seqüência <code>sintética1_0</code>	76
A.2	Posições das regiões periódicas da seqüência <code>sintética2_0</code>	76
A.3	Estatísticas extraídas do conjunto A (570 seqüências).	77
A.4	Estatísticas extraídas do conjunto Am30 (469 seqüências).	77
A.5	Estatísticas extraídas do conjunto Am100 (103 seqüências).	78
A.6	Estatísticas extraídas do conjunto B (195 seqüências).	78
A.7	Estatísticas extraídas do conjunto Bm30 (166 seqüências).	78
A.8	Estatísticas extraídas do conjunto Bm100 (83 seqüências).	78

C.1	Medidas de acurácia estimadas para a sequência <i>sintética1_0</i>	86
C.2	Medidas de acurácia estimadas para a sequência <i>sintética2_0</i>	86
C.3	Medidas de acurácia estimadas para o gene BTU02285.	87
C.4	Medidas de acurácia estimadas para o gene F56F11.4.	87
C.5	Medidas de acurácia estimadas para o gene GGVITIIG.	88
C.6	Medidas de acurácia estimadas para o gene HSDAO.	88
C.7	Medidas de acurácia estimadas para o gene HUMCS3.	89
C.8	Medidas de acurácia estimadas para o gene MMACLGNA.	89
C.9	Medidas de acurácia estimadas para o conjunto A.	90
C.10	Medidas de acurácia estimadas para o conjunto Am30.	90
C.11	Medidas de acurácia estimadas para o conjunto Am100.	91
C.12	Medidas de acurácia estimadas para o conjunto B.	91
C.13	Medidas de acurácia estimadas para o conjunto Bm30.	92
C.14	Medidas de acurácia estimadas para o conjunto Bm100.	92

Capítulo 1

Introdução

1.1 Considerações Preliminares

A bioinformática ou biologia computacional, vem ganhando importância pelas suas descobertas na busca de um maior entendimento dos organismos [7,39]. Importantes avanços nessa área devem-se ao desenvolvimento dos métodos e técnicas de seqüenciamento automático de grandes volumes de fragmentos de DNA, predição de estruturas secundárias do RNA e construção de árvores filogenéticas.

Quando um novo organismo é seqüenciado, tipicamente deseja-se obter toda a informação útil do organismo. Um passo fundamental nesse processo é a identificação de todos os genes presentes na estrutura do DNA. Um passo posterior e igualmente importante para atingir sua anotação genômica é procurar seqüências similares às identificadas. Pode-se, assim, conhecer diferentes aspectos sobre uma seqüência molecular mediante as classificações e comparações de organismos similares já bem estudados.

Com os projetos genoma¹ em fase de seqüenciamento e em grande escala, métodos computacionais tornam-se necessários para identificar genes sobre as seqüências de DNA e saber, com eficiência e confiabilidade, a estrutura dos genes; e, como, quando e quanto é expresso [24].

O problema da identificação de genes corresponde, na verdade, à tarefa de determinar as regiões codificantes de proteínas (CDSs, *coding sequences*) das duas fitas de um fragmento de DNA, que representem as possíveis estruturas biológicas, parciais ou completas, dos genes [61]. Com tal identificação é possível determinar as funções de cada uma das proteínas produzidas pelos genes, e pode-se

---

¹Material genético de um organismo, contido em um conjunto haplóide de cromossomos.

compreender mais sobre o organismos biológico. Em termos computacionais, podemos ver esse problema como sendo a tarefa de determinar a sequência de caracteres componentes (sobre o alfabeto A, C, G, T) de cada um das CDSs em um fragmento de DNA, onde A, C, G e T representam os nucleotídeos correspondentes à sequência de DNA.

Muitos problemas práticos foram encontrados na utilização de abordagens propostas anteriormente, ou combinações delas, para a identificação de genes [9], visto que ainda não se conhecem completamente, ou não existem formulações robustas que poderiam ser utilizadas para modelar a expressão e estrutura dos genes sobre sequências de DNA. Entretanto, dispõe-se de métodos, como os descritos em [26, 61], que auxiliam os pesquisadores e dão, com certo nível, resultados próximos aos reais. Além dos métodos desenvolvidos nos últimos 20 anos [26], foram criadas metodologias computacionais que combinam busca de padrões com conhecimento obtido das comparações dos bancos de dados de sequências ou conjunto de treinamento de genes conhecidos.

Novos métodos de processamento digital de sinais (DSP, *digital signal processing*) aplicados em bioinformática ou biologia computacional [13, 34, 62] estão sendo usados nessas áreas por serem rápidos, e consistirem em uma outra abordagem alternativa com um arcabouço matemático robusto. Os métodos de DSP foram usados para identificar éxons em algumas células eucariotas [4, 51, 57] apresentando resultados promissores.

Todos esses métodos tratam a abordagem de busca por conteúdo por meio da procura de periodicidade de três nucleotídeos (TBP, *three-base periodicity*) presentes somente nas CDSs. Esta abordagem não se limita a conjuntos de genes bem caracterizados necessários na etapa de treinamento de métodos que utilizam outras abordagens. A TBP é considerada um indicador da localização das CDSs em organismos eucariotos e procariotos. Em combinação com outras abordagens esperamos obter um método mais robusto e com alto nível de acurácia.

1.2 Objetivos

O objetivo principal deste trabalho consiste no estudo e desenvolvimento de um método novo que utilize uma transformada de sinais, chamada aqui de Transformada Modificada de Morlet (MMT, *Modified Morlet transform*) para a detecção das CDSs com TBP em sequências de DNA. Esta transformada permitirá analisar uma sequência de DNA somente na frequência correspondente à TBP.

Além desse objetivo, temos o desenvolvimento de um *pipeline* bioinformático em colaboração

com o Laboratório de Biotecnologia Agrícola para o apoio no processamento e montagem do genoma completo de cloroplasto de *Eucalyptus grandis* e análise de ESTs (*expressed sequence tags*)² da bactéria *Pantoea agglomerans*. Para o cumprimento desse objetivo, contamos com a orientação da Profa. Dra. Helaine Carrer do Departamento de Ciências Biológicas da ESALQ-USP.

1.3 Contribuições

As principais contribuições deste trabalho estão discriminadas abaixo:

- Levantamento bibliográfico, com o estudo dos principais métodos de DSP que compõem o estado da arte no que se refere à identificação de CDSs baseada em técnicas de processamento digital de sinais.
- Proposição e avaliação de um novo método automático para a detecção da localização das CDSs (sendo essa a principal contribuição deste trabalho). O desenvolvimento desse método levou à definição de uma nova transformada de sinais, denominada MMT.
- Experimentos comparativos de identificação de CDSs, entre os métodos através da MMT e a transformada de Fourier de tempo reduzido (STFT).
- Estudo e desenvolvimento de um *pipeline* bioinformático para projetos genoma.

1.4 Organização do Trabalho

No Capítulo 2, apresentamos os conceitos biológicos fundamentais da biologia molecular necessários para a compreensão do problema da identificação de CDSs. Discutimos também uma representação das regiões constituintes dos genes, bem como o Dogma Central da biologia molecular. Descrevemos o processo adotado para seqüenciar um genoma completo, e explicamos as evidências de existência de periodicidade nas CDSs. No final do capítulo, realizamos uma breve descrição das categorias de estratégias para a identificação de genes.

No Capítulo 3, apresentamos a definição de um *pipeline* bioinformático e mostramos uma configuração intuitiva das ferramentas computacionais existentes para o processo de identificação de bases, montagem, edição das leituras e anotação para projetos genoma. No final desse capítulo descrevemos

---

²Seqüências pequenas de mRNA (cDNA) que representam o que é expresso pelo gene.

ainda o *pipeline* bioinformático desenvolvido para o recebimento e administração de placas de seqüenciamento, análises de qualidade, filtragem, mascaramento, montagem de seqüências, e comparações dos fragmentos obtidos com seqüências de DNA de organismos conhecidos.

No Capítulo 4 são revisadas as definições básicas da análise de Fourier e mapeamento de nucleotídeos para o tratamento das seqüências simbólicas de DNA como seqüências numéricas. Para avaliar os métodos de identificação de CDSs, realizamos uma análise de espectros de frequência e de periodicidade para seqüências sintéticas e reais de DNA amplamente utilizadas na literatura. Discutimos vários métodos de DSP baseados em uma análise tempo-frequência referentes à identificação de CDSs. Apresentamos alguns resultados preliminares do uso de tais métodos. Finalmente, tratamos as medidas de acurácia utilizadas na avaliação dos métodos de identificação de genes.

No Capítulo 5 são revisadas as definições básicas da análise em *wavelets* e é apresentada a definição da transformada modificada de Morlet usada neste trabalho. Descrevemos o novo método desenvolvido para a identificação de CDSs em seqüências de DNA onde a TBP seja latente. Este método divide o processo no mapeamento de seqüências de DNA, aplicação da MMT, projeção dos coeficientes da MMT, e limiarização dos coeficientes de projeção para a detecção das possíveis fronteiras entre as CDSs.

No Capítulo 6 apresentam-se resultados experimentais do método proposto para a identificação de CDSs. Discutimos experimentos comparativos entre os métodos através da MMT e da STFT em seqüências sintéticas e reais. Para avaliar o desempenho dos métodos na identificação de CDSs utilizamos medidas de acurácia. Um aspecto explorado nas comparações dos experimentos é a questão da limiarização como uma forma lógica de obtenção das fronteiras entre as CDSs.

No Capítulo 7 discutimos algumas conclusões obtidas neste trabalho. Analisamos as vantagens e desvantagens do método proposto para a identificação de CDSs. Apresentamos um esquema que ilustra o posicionamento de nossa contribuição em relação as principais referências da área. Finalmente, sugerimos algumas pesquisas futuras como continuação do trabalho.

Em anexos constam uma descrição das seqüências de DNA testados no trabalho (Apêndice A), estimativas de periodicidades em seqüências sintéticas com ruído (Apêndice B), tabelas de acurácia obtidas para a avaliação dos métodos utilizados para a identificação de CDSs (Apêndice C), assim como as publicações realizadas deste trabalho (Apêndice D).

Capítulo 2

Conceitos Biológicos

Consideramos, ao longo deste trabalho, de maneira indistinta, o uso dos termos nucleotídeos e bases de DNA. Consideramos também que o comprimento de uma sequência de DNA ou RNA é medido pela quantidade de nucleotídeos existentes somente em uma fita. A unidade utilizada é par de base (bp, *base pair*).

2.1 Fundamentos Biológicos

2.1.1 Ácidos Nucléicos

Todo organismo vivo, sem nenhuma exceção conhecida, armazena sua informação biológica na forma de moléculas de ácidos nucleicos, formadas por nucleotídeos, para se construir e manter. Cada nucleotídeo, por sua vez, consiste de: (1) uma molécula de açúcar (desoxirribose ou ribose), (2) um grupo fosfato e, (3) uma segunda molécula chamada base nitrogenada [2].

Existem cinco tipos de bases nitrogenadas. As encontradas no ácido desoxirribonucleico (ADN ou DNA, *Desoxyribonucleic acid*) são adenina (A), citosina (C), guanina (G) e timina (T), de modo que uma sequência de bases nitrogenadas do DNA será formada apenas por A, C, G e T. No ácido ribonucleico (RNA) há uma substituição da timina pela uracila (U) e a sequência do RNA é formada pelas bases A, C, G e U.

O ácido desoxirribonucleico, é representado como uma fita dupla complementar e antiparalela¹, sendo por meio dos nucleotídeos que as duas fitas componentes de molécula de DNA permanecem

---

¹As fitas de DNA tem polaridade oposta, isto é, correm em sentido oposto (estão orientadas em antiparalelo).

Figura 2.1: Representação do DNA como uma fita dupla complementar e antiparalela.

ligadas. O DNA possui o esquema de emparelhamento onde o nucleotídeo A sempre liga-se ao nucleotídeo T e o nucleotídeo C sempre liga-se ao nucleotídeo G por pontes de hidrogênio. Por convenção, uma molécula de DNA começa no lado 5' e termina no lado 3', onde 5' e 3' correspondem aos átomos de carbono livres da molécula de açúcar presentes em cada nucleotídeo [2]. Cada uma das fitas componentes da molécula possui uma orientação própria, onde o final de uma corresponde ao início da outra (Fig. 2.1).

O ácido ribonucleico (ARN ou RNA, *Ribonucleic acid*), é representado como uma fita única de ácidos nucleicos, que diferente do DNA, têm o açúcar ribose e a uracila em vez da timina. Essa fita é obtida através da transcrição de uma das fitas do DNA (Fig. 2.2(c)). São três os tipos de RNA: (1) o RNA mensageiro (mRNA), que contém a informação para a codificação das proteínas, (2) o RNA transportador (tRNA), que é o responsável pelo transporte de aminoácidos, e (3) o RNA ribossomal (rRNA), que possui um papel estrutural.

As diferenças entre o DNA e o RNA não se restringem aos tipos de nucleotídeos constituintes. Na maioria das vezes, o DNA é apresentado como uma longa hélice dupla e com uma estrutura secundária regular e simples. Já o RNA apresenta-se como uma fita única e de tamanho menor que o DNA, com uma grande diversidade de estruturas secundárias relacionadas às funções do DNA e do RNA na célula [2].

2.1.2 Aminoácidos

Os aminoácidos são moléculas orgânicas que contém átomos de carbono, hidrogênio, oxigênio e nitrogênio em sua composição, sendo o resultado da tradução de agrupamentos funcionais de moléculas de DNA. Os aminoácidos formam pequenas cadeias de polímeros chamados polipeptídeos, peptídeos ou monômeros de proteínas.

Dos mais de 500 tipos de aminoácidos encontrados na natureza, são 20 os mais comuns ou

Código	Abreviatura	Nome completo
A	Ala	Alanina
C	Cys	Cisteína
D	Asp	Ácido aspártico
E	Glu	Ácido glutâmico
F	Phe	Fenilalanina
G	Gly	Glicina
H	His	Histidina
I	Ile	Isoleucina
K	Lys	Lisina
L	Leu	Leucina
M	Met	Metionina
N	Asn	Asparagina
P	Pro	Prolina
Q	Gln	Glutamina
R	Arg	Arginina
S	Ser	Serina
T	Thr	Treonina
V	Val	Valina
W	Trp	Triptofano
Y	Tyr	Tirosina

Tabela 2.1: Códigos, abreviaturas e nomes dos aminoácidos.

universalmente aceitos para a síntese, isto é, codificação de proteínas. Os vegetais têm a capacidade de produzir esses 20 aminoácidos necessários para a produção de suas proteínas. Entretanto, as células de animais não produzem todos eles, sendo que alguns devem ser ingeridos. Assim, os aminoácidos são classificados em essenciais, que não podem ser produzidos pelos animais e, os não-essenciais que podem ser produzidos por animais.

Adotaremos a representação clássica [2] de uma e três letras para cada um dos 20 aminoácidos (Tab. 2.1).

2.1.3 Genes

Os genes são os responsáveis pela expressão de uma proteína, ou pelo controle da expressão de outros genes [2]. Todos os genes têm uma estrutura complexa, mas para fins práticos consideramos os genes como constituídos pelas seguintes regiões características:

1. Região de reconhecimento (região promotora);

2. Início de transcrição;
3. Região não-traduzida 5', que regula a transcrição gênica;
4. Início de tradução, ou códon de início (*start codon*);
5. Região para a codificação de proteínas (CDS). Para procariotos considera-se uma única região. Entretanto, nos genes dos organismos eucariotos é considerada uma sequência alternada de éxon/íntron separadas por regiões de corte doadora (*donor splice site* ou 5' *splice site*) e corte aceitadora (*acceptor splice site* ou 3' *splice site*). Consideramos um éxon como uma região necessária para a codificação de proteínas, e um íntron como uma região não presente na codificação com funções ainda desconhecidas;
6. Parada de tradução, ou códon de parada (*stop codon*);
7. Região não-traduzida 3';
8. Região de poliadenilação rica em Adenina (*polyA*) unicamente presente nos eucariotos;
9. Parada de transcrição.

As funções das regiões de DNA existentes entre os genes (regiões inter-gênicas) ainda não são conhecidas por completo [2]. O processo de duplicação² do DNA é chamado de replicação de DNA (Fig. 2.2(a)). A transcrição de uma sequência de DNA em uma outra é chamada de pré-mRNA (Fig. 2.2(b)) e o processo que permite a retirada dos íntrons para a obtenção do mRNA é chamada de corte (*splicing*) (Fig. 2.2(c)). Certos genes podem fazer esta retirada de forma variante, chamada de corte alternativo (*alternative splicing*). O processo biológico que faz a codificação do mRNA, mediante o tRNA e ribossomo, é chamado de tradução (Fig. 2.2(d)).

Existe um caso não usual de genes que são transcritos mas não traduzidos, chamados de genes de RNA não-codificantes (*noncoding RNA genes*). Veja [61] para uma descrição detalhada da estrutura de um gene e classificação dos tipos de éxons.

Neste trabalho, consideramos as CDSs como sendo uma única região dentro dos genes procariotos. Nos eucariotos, como várias regiões chamadas de éxons e separados por íntrons. Portanto, todos os íntrons e as regiões inter-gênicas são consideradas como não-codificantes.

---

²Processo semiconservativo que permite a auto-duplicação de todo o material genético mantendo assim o padrão de herança ao longo das gerações.

Figura 2.2: Representação do Dogma Central da Biologia Molecular: (a) fita de DNA contendo regiões gênicas, (b) fita de pré-mRNA onde estão representados os éxons e íntrons, (c) fita de mRNA, com íntrons removidos, (d) uma sequência de aminoácidos que, traduzidos, formam uma proteína.

2.1.4 Códigos Genéticos e Proteínas

O código genético, ou código de aminoácidos, é um conjunto de regras que usam as células vivas para decodificar seu genoma necessário para a síntese de proteínas. Esta codificação é feita de cada CDS. Usualmente, o código genético é escrito como um conjunto de 64 códons³ dos quais três são indicadores de parada da síntese de proteínas.

O código genético é talvez o maior descobrimento depois do DNA, pois ajudou a compreender melhor a estrutura genética dos organismos. Todos os organismos vivos usam o mesmo ou pequenas variações do código genético padrão [30] (Tab. 2.2) onde os códons de início estão representados pela letra M e os códons de parada pelo símbolo *. Em [30,42], S. Osawa e T. H. Jukes atribuíram para cada taxonomia dos organismos variações do código genético padrão com base nas evidências de sequências homólogas e/ou relações filogenéticas. Uma das variações do padrão utilizadas na comparação de sequências de proteínas é a de bactérias e plastídeos de plantas, onde foram encontrados sete códons

³Um códon é considerado como uma trinca de bases que especificam um aminoácido ou indicador de parada de síntese.

tadas em fragmentos pequenos por enzimas de restrição ou por processos físicos. Esse método é chamado *shotgun*, e depois do primeiro passo os fragmentos são clonados em vetores plasmidiais que são seqüenciados em suas extremidades. Após a primeira etapa de montagem desse genoma, fragmentos maiores são clonados em cosmídeos e seqüenciados. Essa etapa é importante para a montagem do genoma completo do organismo porque a primeira normalmente produz uma seqüência incompleta apresentando alguns “buracos” de seqüência. Os fragmentos sobrepostos criados são a parte fundamental para a montagem.

Em projetos genoma de eucariotos, que tratam uma enorme quantidade de DNA, comumente é usado o método *shotgun* hierárquico onde o DNA inteiro do organismo é primeiramente inserido em grandes vetores de clonagem, como cromossomos artificiais de bactérias (BAC, *bacterial artificial chromosome*) ou leveduras (YAC, *yeast artificial chromosome*). Depois, é realizado o *shotgun* desses grandes fragmentos dos vetores, gerando fragmentos menores que são clonados em vetores plasmidiais para o seqüenciamento.

Estes projetos, geralmente, consistem de duas etapas: (1) a montagem de cada um dos fragmentos clonados nos BACs e YACs, e (2) a montagem final que reunirá as seqüências completas dos BACs e YACs montados para a reconstituição da informação genômica inicial. Um fragmento pode ser único ou uma das muitas seqüências repetidas. Diferentes partes do genoma são bastante similares na seqüência, e certamente essas regiões são as mais difíceis de montar. O genoma deve ser seqüenciado nessa grande escala para garantir estatisticamente a existência de redundância entre os fragmentos e reduzir as chances que se perca regiões do genoma. Os resultados do seqüenciamento são conhecidos como fragmentos ou leituras (*reads*).

2.3 Periodicidade nas CDSs

As CDSs, isto é, as regiões no DNA que codificam em proteínas, tipicamente exibem uma organização periódica de três bases que não é encontrada em outras regiões como (1) as regiões intergênicas, e (2) os íntrons nos eucariotos. Nos últimos anos essa característica nas CDSs tem sido analisada para explicar sua causa [25,43,53,56,64] e quantificá-la [20,47]. Na literatura, esta periodicidade é chamada de periodicidade de três bases (TBP, *three-base periodicity*), tendo sido observada de maneira semelhante para di-nucleotídeos em cromossomos de bactérias [35].

Esta periodicidade mostra relações entre as posições dos nucleotídeos nos éxons que, acredita-se,

seja causada pela assimetria na composição das bases nas três posições codificantes [25]. Em [20] concluiu-se que as frequências não uniformes do *codon usage* são a principal causa para a determinação dessa periodicidade e que a natureza do código genético é responsável pelo comprimento do período, mas não pela própria periodicidade.

Trifonov [54,55], explica que a TBP é dada pela preferência por *G* na primeira posição, e na falta de *G* na segunda posição dos códons $[(G\text{-}n\text{ão}G\text{-}N)n]$ tanto para seqüências de organismos procariotos quanto eucariotos, sugerindo que esse padrão no mRNA possa ser responsável pela monitoração da leitura correta da tripla de nucleotídeos (*frame*) durante a tradução. Vale salientar que foram encontradas algumas exceções desta propriedade genômica nas CDSs em alguns procariotos e seqüências virais e mitocôndriais [32]. Para as regiões inter-gênicas da bactéria *E. coli* foi encontrada periodicidade próxima a 11 bases [29], e sugeriu-se que seja uma propriedade típica das regiões inter-gênicas de organismos procariotos acreditando-se que regula a transcrição.

Entretanto, essa TBP nas CDS não é uniforme, ou seja, é latente, contendo repetições imperfeitas. Entenda-se como periodicidade latente de uma seqüência de DNA a periodicidade com baixo nível de homologia entre quaisquer dois períodos na seqüência de DNA [31].

Finalmente, acreditamos que estamos em uma etapa inicial da explicação da causa e quantificação da periodicidade. O estudo da periodicidade no DNA é importante para o entendimento de uma seqüência de DNA e para o desenvolvimento de métodos de identificação gênica.

2.4 Identificação de Genes

Atualmente, a identificação de genes é a maior área de pesquisa em análise de seqüências de DNA [33]. No processo tratado sobre o Dogma Central da biologia molecular (Seção 2.1.3) várias características podem ser notadas e usadas na dedução computacional da estrutura dos genes, ou seja, a identificação de genes. Esta inclui promotores e regiões reguladoras, limites entre éxons e íntrons, e sinais de início e parada de tradução [7,37,39].

No nível de DNA, são três as categorias que agrupam as abordagens tipicamente adotadas para a identificação de regiões funcionais que codificam elementos responsáveis para a estrutura dos genes, regulação e transcrição. A seguir, descrevemos brevemente cada uma delas:

1. Métodos baseados em reconhecimento de padrões:

- (a) Busca por sítios⁴ (*search-by-sites*), baseados em modelos probabilísticos, onde procura-se a presença ou ausência de uma sequência específica, padrão ou consenso associado à expressão gênica (como a procura de promotores, ou sítios de transcrição ou tradução, ou regiões de poliadenilação).

Nesta categoria, está incluída a procura por uma sequência consenso ou uma expressão que descreva uma sequência consenso, com algumas variações. O uso de matrizes de pesos no lugar do consenso são geralmente utilizadas [26].

A grande dificuldade para estes métodos é que os sítios nem sempre estão presentes nas sequências de DNA, e quando o estão, nem sempre estão no mesmo contexto de onde foram concebidos [7].

- (b) Busca por conteúdo (*search-by-content*), onde procura-se segmentos com propriedades existentes somente nas regiões da sequência de DNA (como a procura de frequência de nucleotídeos, composição de nucleotídeos ricos em G/C ou em A/T, composição de códons e ilhas de CpG⁵).

Para discriminar CDS de não-CDS, geralmente são usados modelos estatísticos da frequência de nucleotídeos e dependências presentes na estrutura do códon [26], analisando a complexidade composicional da sequência.

Uma vantagem desta abordagem é que nenhuma sequência similar é necessária para identificar os genes codificantes de proteínas, pois a identificação é obtida da própria sequência [33].

2. Métodos baseados em comparações por homologia com proteínas, onde são inferidas funções e localização dos genes por homologia se a sequência é similar a uma outra sequência anotada existente. Estes métodos são restritivos pois, para genes recentemente descobertos, muitas vezes não tem-se nenhum casamento com outras proteínas conhecidas [7, 39].
3. Métodos baseados no uso de *expressed sequence tags* (ESTs), onde são procuradas subsequências de cDNA, que é derivado do mRNA, dentro das sequências de DNA [7].

Todos os métodos conhecidos, atualmente, são sensíveis e dependentes da natureza dos dados

---

⁴Coluna individual de resíduos em um alinhamento de aminoácidos ou nucleotídeos. Os resíduos em um sítio são presumidos como homólogos.

⁵Presença de sub-regiões ricas em citosina e guanina na região 5', que está associada ao efeito da expressão gênica.

analisados, sendo a intuição ou domínio do conhecimento dos especialistas um fator fundamental no processo da identificação de genes e posterior anotação genômica.

O problema de identificação de genes está longe de ser resolvido e a principal dificuldade é encontrada no que define realmente um gene. Biólogos moleculares encontraram alguns genes de comprimento mais curtos do que os esperados inicialmente [27]. Um desempenho melhor de uma identificação requererá melhor conhecimento biológico sobre porque os genes têm as características até agora encontradas [33].

Neste trabalho, nos concentramos na busca por conteúdo para identificação de CDSs, sendo usada a TBP (Seção 2.3) como um bom indicador de suas localizações. Pode ser encontrada em [26,61] uma descrição ampla das ferramentas utilizadas para a identificação computacional de genes de eucariotos.

Capítulo 3

Pipeline Bioinformático

Depois do seqüenciamento genômico, é particularmente necessária a determinação da longa seqüência original de DNA para sua posterior anotação. Assim, procedimentos biológicos e/ou computacionais, tal como um *pipeline* bioinformático, são fundamentais para auxiliar todo o processo de uma maneira simples, lógica e ordenada.

Pipeline bioinformático (*bioinformatics pipeline*) é um termo usado, mas vagamente definido em bioinformática. Aqui o definimos como uma seqüência de unidades ou etapas funcionais que realizam uma tarefa genômica em diferentes passos biológicos e/ou computacionais, em que cada unidade funcional recebe entradas e produz saídas que são armazenadas em arquivos ou bancos de dados especiais para sua posterior interpretação biológica. Uma saída de uma unidade funcional é uma entrada da unidade seguinte, sendo que operações dentro de uma unidade podem ser realizadas em paralelo.

3.1 Unidades Funcionais

A organização das unidades funcionais em um *pipeline* bioinformático para projetos genoma depende do tipo, características e necessidades do projeto em que deseja-se utilizar (para recentes exemplos, ver [21, 41, 46]). Entretanto, nas seguintes subseções, apresentamos uma organização típica e simplificada dessas unidades funcionais (Fig. 3.1).

Figura 3.1: Organização típica de um *pipeline* bioinformático para projetos genoma. Cada bloco representa uma unidade funcional e cada seta o fluxo da informação.

3.1.1 Identificação de Bases

As entradas para o *pipeline* bioinformático são as leituras (*reads*) dos fragmentos produzidos pelo seqüenciador automático de DNA com informações analógicas que representam os nucleotídeos lidos deste equipamento (*raw data*), chamados de arquivos de cromatograma, de um dos fragmentos próprios do método de seqüenciamento.

Para converter esses dados analógicos em fragmentos de bases, as leituras são submetidas a um programa de identificação, denominado *base-caller*, o qual as identifica como A, C, G ou T, atribuindo um valor de qualidade para cada um. O rótulo N é atribuído no caso de um nucleotídeo não ser identificado ou possuir uma qualidade muito baixa (Fig. 3.2).

A orientação de um fragmento é desconhecida e freqüentemente as máquinas de seqüenciamento automático apresentam erros nas leituras dos fragmentos de DNA. Uma boa parte dos fragmentos contém erros em uma taxa de 1 a 5%. Muito mais erros ocorrem nos extremos das leituras [39,45].

Nos trabalhos de Ewing [22,23], as atribuições dos valores confiáveis aos nucleotídeos são dadas por um algoritmo que tem como base a análise de Fourier [38]. O valor de qualidade de uma base em um fragmento é $q = -10 \log_{10} p$, onde p é a probabilidade de erro estimado para a base. Assim, os valores de qualidade estão definidos para inteiros positivos pequenos para cobrir um amplo intervalo de probabilidade de erro. Esse valor é importante para determinar se um fragmento precisa ser re-seqüenciado.

*replicons*¹ do genoma e todas as bases tenham um valor de qualidade aceitável dependendo do valor estabelecido para o projeto.

Existem diferentes maneiras para certificar que uma sequência genômica seja a correta [2]. A sequência pode ser comparada com pequenas partes de um genoma que tenha sido sequenciado e anotado previamente ou com regiões de mapas genômicos. Se uma montagem é consistente com regiões dispersas de informação conhecida, então é bastante provável que toda a sequência seja a correta [48].

3.1.3 Edição das Leituras

Esta etapa consiste nas operações de substituição, remoção ou inserção de bases. Permitindo assim a correção de possíveis erros cometidos na identificação de nucleotídeos e nos erros da própria montagem. Em geral, isso pode ser feito devido à existência de um ou mais fragmentos que tenham bases adequadas com qualidade alta para inferir a correção da base incorreta [41].

3.1.4 Anotação do Genoma

Esta etapa é crítica porque consiste na identificação de regiões onde estão localizados os possíveis genes para depois determinar quais são suas funções, descrições ou categorizações biológicas [45, 48]. As operações comuns nesta etapa são:

1. Identificação dos genes (Seção 2.4);
2. Descrição de genes, geralmente realizada com ajuda de outras montagens de outros organismos, cujas funções já são conhecidas, pois as sequências de organismos estruturalmente similares podem ter também funções similares. Esta tarefa pode ser realizada de forma semi-automática para a ajuda na classificação dos genes identificados;
3. Construção do mapa genético, mostrando a disposição, orientação e categoria geral dos genes do organismo montado.

Veja [48], para uma descrição detalhada das categorias típicas de anotação, técnicas usadas, e modelos sociológicos que têm sido adotados pelos pesquisadores.

---

¹Uma molécula de DNA capaz de se replicar e se manter.

3.2 *Pipeline* Bioinformático Desenvolvido

Nesta seção, apresentamos uma descrição do *pipeline* bioinformático que desenvolvemos no Laboratório de Bioinformática do Departamento de Ciência da Computação do IME-USP, em colaboração com o Laboratório de Biotecnologia do Departamento de Ciências Biológicas da ESALQ-USP, sob a orientação da Profa. Dra. Helaine Carrer.

O *pipeline* bioinformático desenvolvido é tratado como um conjunto de ferramentas computacionais configuradas para (1) o recebimento e administração de placas de seqüenciamento, (2) análises de qualidade por placa, (3) filtragem e mascaramento de seqüências, (4) montagem dos fragmentos de DNA e, (5) comparações dos fragmentos obtidos com seqüências de DNA de organismos conhecidos.

Atualmente, este *pipeline* bioinformático (Fig. 3.3), disponível para uso em <http://malariadb.ime.usp.br:8026/pipeline>, sob o sistema operacional SUNOS^{TM} 5.8, apóia a montagem do genoma completo de cloroplasto de *Eucaliptus grandis* e na análise de ESTs da bactéria *Pantoea agglomerans*.

3.2.1 Arquitetura e Implementação

Na Figura 3.4 ilustramos um esquema correspondente à arquitetura do *pipeline* bioinformático desenvolvido. A seguir descrevemos a interface web, o sistema de gerenciamento e as cinco unidades funcionais deste *pipeline*.

- A *interface web* fornece um ambiente interativo para a especificação de parâmetros usados nas diferentes ferramentas, bem como para o recebimento e administração de placas de seqüenciamento associadas a cada projeto. Esta interface também permite ao usuário monitorar o processo da montagem e visualizar os resultados de cada projeto. Para a implementação desta interface foram criados vários *scripts* em JSP^{TM} sobre o servidor APACHE TOMCAT^{TM} 4.1.30.
- O *sistema de gerenciamento* utiliza os parâmetros especificados pelo usuário para iniciar e dirigir todo o processo, podendo ser selecionadas diferentes maneiras de execução para cada projeto. Uma vez determinada a maneira de execução de todo o processo, são utilizadas ferramentas individuais para cada unidade funcional no *pipeline* bioinformático. Foram criados *scripts* em PERL^{TM} 5.8.4 para a implementação deste sistema.
- As unidades funcionais *identificação de bases*, *filtragem de seqüências* e *montagem de seqüências* são realizadas em dois níveis: por agrupamento de todas as placas e por placas individuais

Figura 3.3: Interface web inicial do *pipeline* bioinformático desenvolvido.

de seqüenciamento. Elas permitem a validação de qualidade das seqüências, mascarando as seqüências correspondentes aos vetores de clonagem e *primers* se existirem, nas leituras. Estas unidades funcionais são realizadas mediante o EGENE [19] que é um *pipeline* bioinformático modular flexível e de fácil integração.

- O *alinhamento de seqüências* realiza as comparações por pares de bases, comparando cada *contig* e *singlet* obtido da montagem, com outras seqüências de organismos registrados no *National center for biotechnology information* (NCBI, <http://www.ncbi.nlm.nih.gov/>). A ferramenta utilizada nesta unidade é o BLASTCL3 2.2.9 que é um cliente BLAST^{TM}2 [3].

²Blast, do inglês *Basic Local Alignment Search Tool*, é um algoritmo de comparação de seqüências usado para a

Figura 3.4: Esquema da arquitetura do *pipeline* bioinformático desenvolvido.

- Os *alinhamentos com genomas completos* dos *contigs* e *singlets* obtidos da montagem do genoma e de comprimento suficientemente grande (e.g. 2000 bp). O genoma a ser usado é definido no sistema de gerenciamento. Este alinhamento ajuda na análise de regiões cromossômicas sistêmicas, comparações evolucionárias e duplicações genômicas. A ferramenta utilizada para esta unidade funcional é o MUMMER 3.0 [17].

Um processo típico para um projeto arbitrário poderia ser o seguinte. Quando uma placa de sequenciamento é submetida ao *pipeline* bioinformático, será realizada a identificação de bases, filtragem, mascaramento, e análise de qualidade de seqüências. Este primeiro passo permite descartar seqüências com qualidade baixa. Um processo seguinte permite a montagem genômica somente com as seqüências obtidas no processo anterior para a placa submetida, como também para o processo de montagem das seqüências de todas as placas. Toda a informação sobre a montagem é coletada (*coleção de informação*), isto é, são armazenados o número de seqüências válidas, de *singlets*, e de *contigs*.

---

busca de um alinhamento local ótimo.

Para o processo de montagem de todas as placas, são realizadas as comparações dos *singlets* e *contigs* obtidos com (1) seqüências conhecidas e registradas no NCBI, e (2) um genoma completo conhecido, previamente indicadas no sistema de gerenciamento. Os resultados finais da montagem e das comparações são agrupados para a geração de relatórios da análise genômica (*fusão de informação*). Todos os resultados são apresentados através da interface web.

3.2.2 Relatórios de Análise

O *pipeline* bioinformático apresenta vários relatórios correspondentes a cada projeto registrado, sendo os mais importantes:

- Placas submetidas e processadas.
- Estado atual da montagem.
- *Singlets*, e *contigs* obtidos pela montagem das seqüências.
- Alinhamentos, dos *singlets* e *contigs* obtidos, com uma seqüência completa de um genoma conhecido.

No tempo da escrita deste trabalho conta-se com 87 placas submetidas e processadas, fazendo um total de 7399 leituras, 11 *singlets*, e 212 *contigs* na montagem do genoma de cloroplasto de *Eucaliptus grandis*. Conta-se com 26 placas, fazendo um total de 1899 leituras, 693 *singlets*, e 194 *contigs* na análise de ESTs da bactéria *Pantoea agglomerans*.

Capítulo 4

Métodos de DSP para a Identificação de CDSs

As regiões codificantes de proteínas (CDSs) em seqüências de DNA apresentam tipicamente uma organização periódica de três bases não uniforme (TBP) e que ainda não é compreendida completamente (Seção 2.3). Nota-se que vários métodos foram criados para a identificação de CDSs (Seção 2.4) baseados na busca dessa característica.

Neste capítulo, revisamos as definições básicas da análise de Fourier, mapeamento numérico de nucleotídeos, espectros de frequência de DNA, análise de periodicidade de DNA, e apresentamos os principais métodos de processamento digital de sinais (DSP) que compõem o estado da arte no que se refere à identificação de CDSs através da busca da TBP.

4.1 Análise de Fourier

Um aspecto importante da análise de Fourier deve-se ao fato dela possibilitar uma representação de uma classe ampla de funções em termos de uma combinação linear de funções base senos, cossenos ou exponenciais complexos [15]. Uma outra forma de pensar na análise de Fourier é como uma técnica matemática para transformar nossa visão de informação baseada no tempo (posição ou espaço) naquela baseada na frequência [44].

4.1.1 Transformada de Fourier

A transformada de Fourier é utilizada para decompor um sinal em componentes que representem frequências, descrevendo o sinal de forma uni-dimensional e global. Existem variações, desta transformada, diretamente relacionadas a ela, que dependem do tipo de função a transformar [38].

Esta transformada invertível representa de forma não ambígua o sinal e define a noção de frequência global no sinal.

A transformada contínua de Fourier (CFT, *Continuous Fourier transform*) é referida para funções contínuas, que representem qualquer função integrável $f(t)$ como a integral de exponenciais complexas com frequência angular ω e amplitude complexa $F(\omega) \in \mathcal{C}$ [10],

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t} dt \quad (4.1)$$

Assim, a transformada inversa de Fourier para funções contínuas, $f(t) \in \mathcal{C}$, é definida por

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)e^{j\omega t} d\omega \quad (4.2)$$

Analogamente, a transformada discreta de Fourier (DFT, *Discrete Fourier transform*) da função $f[b]$ pode ser definida como

$$F[k] = \sum_{b=0}^{n-1} f[b]e^{-2\pi jkb/n}, \quad k = 0, \dots, n-1 \quad (4.3)$$

sendo definida a transformada inversa de Fourier para funções discretas como

$$f[b] = \frac{1}{n} \sum_{k=0}^{n-1} F[k]e^{-2\pi jkb/n}, \quad b = 0, \dots, n-1 \quad (4.4)$$

A transformada de Fourier tem um inconveniente pois transformar ao domínio da frequência resulta em uma perda da informação relativa ao tempo. Portanto, é impossível indicar quando um evento particular ocorre [44].

4.1.2 Espectro de Frequência

O espectro de frequência é uma representação de um sinal em termos de frequência (Seção 4.1.1, sendo esse considerado uma projeção do sinal sobre um conjunto de funções base senoidais. Os espectros de frequência contêm informação de amplitude e de fase [38].

4.1.3 Convolução

A convolução é um operador matemático entre duas funções f e g , que produz uma terceira função, que de certa forma, represente a quantidade de sobreposição entre f e uma versão invertida e deslocada de g em diferentes instantes de tempo. A convolução de f e g , denotada $f * g$, é definida como a integral do produto de duas funções após uma ser invertida e deslocada [15, 38].

Para o caso de funções contínuas, a convolução $h(\tau)$ será dada por

$$h(\tau) = f(\tau) * g(\tau) = (f * g)(\tau) = \int_{-\infty}^{\infty} f(t)g(\tau - t)dt \quad (4.5)$$

em que a faixa de integração depende do domínio no qual as funções estejam definidas. Por outro lado, para funções discretas, a convolução é dada por

$$h[b] = f[b] * g[b] = (f * g)[b] = \sum_k f[k]g[b - k] \quad (4.6)$$

4.1.4 Transformada de Fourier de Tempo Reduzido

Uma alternativa para a análise bi-dimensional e local de dados é dada pela transformada de Fourier de tempo reduzido (STFT, do inglês *Short-time Fourier transform*), que introduz uma janela de extensão limitada que permite selecionar uma parte do sinal. A STFT é usada para determinar como a amplitude senoidal e fase de um sinal muda no tempo, dando alguma informação sobre quando e em que frequências um evento em um sinal ocorre, com uma precisão limitada e determinada pelo tamanho da janela. Esta transformada mapeia o sinal em um plano tempo¹-frequência [38, 44].

Para um sinal contínuo $f(t)$ define-se a STFT como

$$\text{STFT}(b, \omega) = \int_{-\infty}^{\infty} g^*(t - b)f(t)e^{-j\omega t}dt \quad (4.7)$$

em que $g(t)$ representa a função janela centrada em b [38]. Analogamente, a STFT para um sinal

---

¹O tempo também poderia ser considerada como a posição de um elemento em um sinal.

discreto $f[k]$ é definida como

$$\text{STFT}[b, \omega] = \sum_{k=0}^{n-1} g^*[k-b]f[k]e^{-2\pi j\omega k/n} \quad (4.8)$$

Imagine-se $g(t-b)$ como um deslizamento de uma janela através do sinal, e para cada deslocamento de g é calculada a transformada de Fourier (Seção 4.1.1). Esta transformada, que permite uma análise local, depende criticamente da escolha da janela $g(t)$. Quando a janela é determinada, a resolução tempo-frequência é fixada sobre o plano tempo-frequência pois a mesma janela é usada para analisar todo o sinal [44].

Vale salientar que o módulo da STFT define o espectrograma da função [10], e pode ser construída em duas dimensões com o tempo no eixo das ordenadas e a frequência no eixo das abscissas. A amplitude pode ser dada por uma escala de cores na visualização. O espectrograma é definido por

$$E(b, \omega) = |\text{STFT}(b, \omega)|^2 \quad (4.9)$$

4.2 Mapeamento Numérico de Nucleotídeos

Vários métodos foram utilizados para converter dados categóricos ou simbólicos de seqüências de DNA para serem tratados como seqüências numéricas [1, 59]. Esta conversão, ou mapeamento, deve-se à dificuldade de uso direto de dados simbólicos nas técnicas de DSP pois elas têm uma base algébrica totalmente ausente no caso dos dados simbólicos. Um mapeamento de símbolos a números não deve impor uma estrutura matemática que não existe em seqüências de DNA [1]. Existem duas abordagens que podem ser utilizadas para representar numericamente uma seqüência de DNA:

1. *Mapeamento fixo*, consiste na atribuição de números inteiros, reais ou complexos, aos quatro caracteres A, C, G e T correspondentes aos tipos de bases presentes na seqüência de DNA (Seção 2.1.1).

No mapeamento binário [8] são criadas sete seqüências numéricas. Cada seqüência numérica corresponde a uma das sete formas de mapear uma seqüências de DNA baseadas nas características próprias das bases, considerando, por exemplo, as ligações de hidrogênio e se as bases nitrogenadas são purinas ou pirimidinas.

Regra	Atribuição			
	A	C	G	T
1 Ligações de hidrogênio	0	1	1	0
2 Purina/pirimidina	1	0	1	0
3 Híbrida	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

Tabela 4.1: Regras para o mapeamento binário de seqüências de DNA. No mapeamento é associada uma seqüência binária para cada regra que indica uma característica ou posição de cada base.

Na Tabela 4.1 estão representadas as sete regras utilizadas em [8] para as atribuições de valores binários para cada base. Vale salientar que as seqüências obtidas por cada uma dessas regras de mapeamento são independentes, pois estas referem-se a aspectos diferentes da seqüência de DNA, mantendo assim invariância à atribuição numérica.

Para o mapeamento por números inteiros [56], reais [40] ou complexos [4] são utilizados números arbitrários ou valores que representam alguma propriedade das bases nitrogenadas como, por exemplo, os potenciais de interação elétron-íon (EIIP, *electron-ion interaction potentials*) encontrada nas bases. Os autores de [40] analisaram seqüências de DNA em que foram atribuídos valores para cada base como sendo $(A, C, G, T) = (0.1260, 0.1340, 0.0806, 0.1335)$. A dificuldade com este último mapeamento é, na verdade, a dependência na atribuição de números (inteiros, reais ou complexos) adotada, permitindo que alguma estrutura harmônica relevante seja oculta ou artificialmente criada pela atribuição [1, 59]. Considerando, por exemplo, a seqüência periódica (GTGCGA GTGCGA . . .) com a atribuição dos valores $(A, C, G, T) = (1, 0, 0, 0)$, forma-se uma seqüência numérica de período seis, e para a atribuição de $(A, C, G, T) = (0, 0, 1, 0)$, forma-se outra de período dois.

2. *Mapeamentos baseados em critérios de otimização* [12, 49, 59], de maneira não arbitrária, e que enfatiza qualquer característica periódica que poderia existir sobre séries estacionárias ou não-estacionárias.

Neste trabalho, utilizamos somente as regras 4 a 7 do mapeamento binário para a atribuição de valores para quatro seqüências, u_A , u_C , u_G e u_T , correspondentes às bases nitrogenadas, em que

Seqüência	A	T	G	C	T	T	G	A	C	T	A	G	G	G	C	T	C	A	G	T
u_A	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
u_C	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0
u_G	0	0	1	0	0	0	1	0	0	0	0	1	1	1	0	0	0	0	1	0
u_T	0	1	0	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1

Tabela 4.2: Seqüências binárias que representam as posições de A, C, G e T, em uma seqüência de DNA, correspondentes as regras 4 a 7 do mapeamento binário.

cada seqüências binária² especifica a presença ou não da base A, C, G ou T em uma dada posição da seqüência de DNA. Tal abordagem foi considerada também nos trabalhos [1, 4, 51]. Por exemplo, a seqüência ATGCTTGACTAGGCTCAGT e suas correspondentes seqüências binárias são mostradas na Tabela 4.2.

Observa-se que um mapeamento fixo inteiro, real ou complexo pode ser obtido baseado no mapeamento binário. Se a , c , g e t são valores numéricos arbitrários correspondentes às bases A, C, G e T, uma seqüência s de comprimento N pode ser representada como uma combinação linear de suas seqüências binárias u_A , u_C , u_G e u_T [4], tal que

$$s[b] = a.u_A[b] + c.u_C[b] + g.u_G[b] + t.u_T[b], \quad (4.10)$$

em que b é uma determinada posição da seqüência. Note que essa representação é tal que $u_A[b] + u_C[b] + u_G[b] + u_T[b] = 1$, para qualquer posição b . Na Figura 4.1, mostram-se as representações gráficas do mapeamento binário e do mapeamento inteiro de uma seqüência de DNA (Tab. 4.2). Para este último mapeamento consideramos valores atribuídos para (A, C, G, T)=(1,2,3,4).

4.3 Espectro de Frequência de DNA

Segundo [4], a transformada discreta de Fourier (4.3) da seqüência (4.10) de comprimento N é definida como

$$S[k] = \sum_{b=0}^{N-1} s[b]e^{-2\pi jkb/N}, \quad k = 0, \dots, N-1 \quad (4.11)$$

---

²Este tipo de seqüências também são conhecidas como seqüências indicadoras binárias, por serem indicadoras de posições.

Figura 4.1: Representação gráfica da sequência de DNA: ATGCTTGACTAGGGCTCAGT. Em (a) são consideradas quatro sequências binárias correspondentes às regras 4 a 7 do mapeamento binário. Em (b) considera-se uma sequência numérica com valores de (A, C, G, T)=(1,2,3,4).

Assim, a sequência de coeficientes S pode ser expressa como uma combinação linear das transformadas discretas de Fourier das sequências binárias, dando uma medida do conteúdo da frequência k definida como

$$S[k] = a.U_A[k] + c.U_C[k] + g.U_G[k] + t.U_T[k], \quad k = 0, \dots, N-1 \quad (4.12)$$

em que U_A , U_C , U_G e U_T são as transformadas discretas de Fourier das sequências binárias pertencentes à sequência s . Essas sequências representam a frequência para cada nucleotídeo que, combinadas, dão uma representação quatro-dimensional do espectro de frequência (Seção 4.1.2) da sequência de DNA. Portanto, o espectro de frequência total é definido como [4]

$$E[k] = |U_A[k]|^2 + |U_C[k]|^2 + |U_G[k]|^2 + |U_T[k]|^2, \quad k = 0, \dots, N-1 \quad (4.13)$$

Observe que, com (4.10) e o resultado de (4.11), temos que

$$U_A[k] + U_C[k] + U_G[k] + U_T[k] = \begin{cases} N, & k = 0 \\ 0, & k \neq 0 \end{cases}$$

Como exemplo, nas Figuras 4.2 e 4.3, apresentamos os espectros de frequência correspondentes a seqüências sintéticas de 600bp: (1) seqüência com periodicidade três de 600bp (Fig. 4.2(a)), (2) seqüência com uma região periódica de 200bp, correspondente às posições 201 a 400 (Fig. 4.2(b)), (3) seqüência com uma região periódica de 100bp, correspondente às posições 251 a 350 (Fig. 4.2(c)), (4) seqüência com uma região periódica de 50bp, correspondente às posições 276 a 325 (Fig. 4.3(d)),

Pela constituição dessas seqüências sintéticas, podemos observar que o valor da energia³ associada à frequência três é, geralmente, maior em comparação às outras frequências. Note, também, que a transformada de Fourier permite realizar uma análise global da seqüência. Note que quanto menor a CDS, menos nítido é o pico de frequência em relação ao espectro total. Portanto, os valores das energias associadas à frequência três, correspondentes às regiões com periodicidade três, serão indistintos na medida que a região periódica seja menor.

Na Figura 4.4 apresentamos os espectros de frequências para uma seqüência periódica de 600bp onde foram percentualmente inseridos dois tipos de ruídos: (1) ruído *inclusivo*, que corresponde à escolha aleatória de uma determinada porcentagem de bases para depois serem trocados com a mesma probabilidade por A, C, G e T; (2) ruído *privativo*, que corresponde também à escolha aleatória de uma porcentagem de bases, mas estas podem ser trocadas exclusivamente por outras. Por exemplo, se a base escolhida for A, esta pode ser trocada com igual probabilidade por C, G ou T. A energia correspondente à frequência três nos espectros de frequências diminui quando a porcentagem do ruído é maior (Fig. 4.4(a)), mas no caso do ruído privativo próximo a 100%, a energia nessa frequência é maior do que as outras frequências, pois a probabilidade de troca de uma base selecionada, por outra que não seja a mesma, é de 1/3 (Fig. 4.4(b)).

Mostramos alguns exemplos de espectros de frequências de seqüências de DNA correspondentes a uma CDS de 6885bp (região 86474-93358) (Fig. 4.4(c)), e uma região não-codificante de 1389bp (Fig. 4.4(d)) do genoma de *Arabidopsis thaliana* com número de acesso GenBank NC_000932. Em todos os exemplos mostrados nesta seção, com exceção das seqüências correspondente à região não-codificante (Fig. 4.4(d)), a energia presente na frequência três nos espectros de frequências demonstram a existência da TBP presente nas CDSs (Seção 2.3). Entretanto, a energia associada à periodicidade três depende do tamanho da seqüência analisada [11]. É de esperar-se uma energia alta quando a seqüência analisada possua uma alta periodicidade latente de três bases. Na seção seguinte realizamos um estudo aprofundado da periodicidade nas seqüências de DNA e em particular da TBP.

---

³Entenda-se a energia como o módulo do coeficiente da transformação.

(a) Região periódica (1 a 600)

(b) Região periódica (201 a 400)

(c) Região periódica (251 a 350)

Figura 4.2: Espectros de frequência de seqüências sintéticas de 600bp que contêm regiões com periodicidade três: GCT GCT ... (direita). Representações gráficas das seqüências de DNA (esquerda).

(d) Região periódica (276 a 325)

Figura 4.3: Espectros de frequência de seqüências sintéticas de 600bp que contêm regiões com periodicidade três: GCT GCT . . . (direita). Representações gráficas das seqüências de DNA (esquerda). (*continuação*)

4.4 Análise de Periodicidade de DNA

Neste trabalho usamos dois tipos de conjuntos de seqüências de DNA (Apêndice A) para avaliar os métodos estudados. O primeiro tipo trata de seqüências criadas sinteticamente, enquanto o segundo corresponde a conjuntos de seqüências reais de eucariotos usados para avaliar programas computacionais de identificação de genes.

Nesta seção, apresentamos para cada tipo de conjunto, as primeiras 23 estimativas inteiras de periodicidades, isto é, estimativas da periodicidade 2 a 24 das regiões codificantes (que simulam TBP) e regiões não-codificantes (regiões inter-gênicas e, no caso dos organismos eucariotos, íntrons). Observamos, para faixas maiores de periodicidade, um comportamento similar nas estimativas destas periodicidades, isto é, a energia associada à frequência três nas regiões periódicas (CDSs) é sempre maior que as das outras frequências. Esse fato motiva o desenvolvimento de métodos de identificação de CDSs com base na periodicidade.

Para o cálculo dos valores de periodicidade, nesta seção, utilizamos somente os coeficientes da transformada de Fourier (Seção 4.1.1) de cada região nas seqüências de DNA. Em [20,47] encontra-se uma descrição detalhada dos métodos comuns para o cálculo de periodicidade em seqüências de DNA.

Figura 4.4: Espectro de frequências de seqüências. Em (a) e (b) mostram-se os coeficientes da transformada de Fourier para um conjunto de seqüências sintéticas de 600bp em que foram percentualmente inseridos dois tipos de ruído. Em (c) e (d) mostram-se os coeficientes da transformada de Fourier para seqüências reais.

4.4.1 Periodicidade em Seqüências Sintéticas

As estimativas de periodicidade das seqüências sintéticas `sintética1_0` e `sintética2_0` (Apêndice A.1) são representadas nas Figuras 4.5(a) e 4.5(b). A seqüência `sintética1_0` contém regiões

periódicas de 2 até 6, sendo exatamente nessas periodicidades que as energias são maiores às dos outros não considerados na sequência sintética. Já para a sequência `sintética2_0`, que contém nove regiões periódicas de três bases (que simulam TBP), a energia associada à periodicidade três é maior as das outras regiões. Observemos que, para as regiões que não tem alguma periodicidade (que simulam regiões não-codificantes), nenhum comportamento similar é encontrado nas medidas de periodicidade.

No Apêndice B, estão representadas as periodicidades de algumas das sequências sintéticas `sintética2_XA` e `sintética2_XB` em que as bases pertencentes às regiões periódicas foram alteradas com ruído.

4.4.2 Periodicidades em Sequências Reais

Na Figuras 4.5(c) e 4.5(d) mostramos a representação das periodicidades estimadas para os éxons, íntrons e regiões inter-gênicas para conjuntos de sequências reais (Apêndice A.2). Observemos que os éxons apresentam uma energia associada às CDSs maior em comparação às outras energias das periodicidades estimadas. Esta estimativa demonstra a existência da TBP presente somente nas CDSs (Seção 2.3).

4.5 Identificação de CDSs

As pesquisas sobre DSP demonstraram que uma análise de Fourier (e também *Wavelet*) pode ser de grande ajuda na descrição da estrutura das sequências de DNA [13, 34, 62, 63]. A seguir, apresentamos uma visão geral dos métodos de DSP utilizadas para a identificação de CDSs.

S. Tiwari, em [51], para uma sequência de comprimento N , definiu quatro coeficientes normalizados na frequência três⁴, isto é, o coeficiente da STFT na posição $N/3$, como

$$A = \frac{1}{N}U_A\left[\frac{N}{3}\right], C = \frac{1}{N}U_C\left[\frac{N}{3}\right], G = \frac{1}{N}U_G\left[\frac{N}{3}\right], T = \frac{1}{N}U_T\left[\frac{N}{3}\right] \quad (4.14)$$

tal que,

$$W = |A|^2 + |C|^2 + |G|^2 + |T|^2, \quad (4.15)$$

---

⁴As demais frequências não são consideradas no cálculo dos coeficientes, pois pretende-se concentrar-se nas regiões onde exista TBP (Seção 2.3).

Figura 4.5: Periodicidades estimadas para seqüências de DNA. Foram calculadas as periodicidades 2 a 24 das regiões codificantes (que simulam TBP) e regiões não-codificantes (regiões inter-gênicas e, no caso dos organismos eucariotos, íntrons).

é definido como um identificador de CDSs⁵, com $A + C + G + T = 0$. U_A , U_C , U_G e U_T representam a STFT para sinais discretos (Seção 4.1.4) das seqüências binárias u_A , u_C , u_G e u_T associadas a seqüência de DNA (Seção 4.2).

⁵Veja sua semelhança com a equação do espectro de freqüência (4.13)

Como exemplo, apresentamos os valores dos coeficientes W para três seqüências de DNA:

1. Seqüência **sintética1.0** (Apêndice A.1) de 6000bp utilizando para o calculado das STFTs variações do tamanho de janela de 2bp a 400bp (Fig. 4.6(a)). Observemos que para as regiões com periodicidade três os coeficientes são maiores em comparação às das outras regiões (com periodicidade dois, quatro, cinco e seis). Observemos, também, que em grande parte as energias para a região de periodicidade seis são consideráveis pois esta periodicidade é múltiplo de três. Entretanto, todos estes coeficientes variam de acordo com o tamanho da janela utilizada. Uma boa aproximação para a identificação de regiões com periodicidade três para esta seqüência foi obtida usando uma janela de 21bp (Fig. 4.6(b)), pois a periodicidade nessas regiões é latente (Seção 2.3), isto é, a periodicidade é uniforme. Resultados similares foram obtidos para tamanhos de janela de 9 a 200.
2. Seqüência **sintética2.0** (Apêndice A.1) de 10000bp utilizando para o cálculo das STFTs variações do tamanho de janela de 2bp a 400bp (Fig. 4.6(c)). Esta seqüência sintética contém nove regiões com periodicidade três de diferentes comprimentos. Observemos que, para essas nove regiões, os coeficientes são maiores em comparação às regiões que contém bases aleatórias. Igualmente, como no caso da seqüência **sintética1.0**, estes coeficientes são dependentes do tamanho de janela utilizada no cálculo, pois são obtidas energias diferentes para tamanhos de janela diferentes. Na Figura 4.6(d) apresentamos uma aproximação para a identificação dessas regiões usando um tamanho de janela de 21bp.
3. Gene F56F11.4 de *C. elegans* de 8000bp cujos éxons são mostrados na Tabela 4.3, incluindo sua posição relativa no gráfico. Para o cálculo dos coeficientes utilizamos arbitrariamente variações de tamanho de janela de 2bp a 500bp (Fig 4.6(e)). Esta análise local permite identificar regiões que contenham TBP, mas os resultados dessa identificação são sempre dependentes do tamanho utilizado de janela na análise. Em [4], a identificação dos éxons usando uma janela de 351bp foi apresentada. Observemos que os coeficientes correspondentes às regiões dos quatro últimos éxons são maiores comparando-se com as das outras regiões. Podemos presumir que o primeiro éxon apresenta pouca TBP ou o tamanho de janela utilizada nessa análise não é adequada. Resultados similares foram obtidos mediante tamanhos de janela de 300bp a 400bp.

(a) *sintética1.0* com janelas de 2 a 400bp (b) *sintética1.0* com janela de 21bp (c) *sintética2.0* com janelas de 2 a 400bp (d) *sintética2.0* com janela de 21bp

(e) Gene F56F11.4 com janelas de 2 a 500bp

(f) Gene F56F11.4 com janela de 351bp

Figura 4.6: Coeficientes W estimados, para seqüências de DNA, com um conjunto de tamanhos de janela (esquerda), e com um tamanho de janela determinado (direita).

Posição relativa	Bases	Reading frame
928-1039	112	2
2528-2857	330	2
4114-4377	264	1
5465-5644	180	2
7255-7605	351	1

Tabela 4.3: Éxons do gene F56F11.4.4 de *C. elegans*. NCBI, número de acesso AF099922 e região 7021-15020.

Em um trabalho subsequente de D. Anastassiou [4], define-se

$$W = |a.A + c.C + g.G + t.T|^2 \quad (4.16)$$

como um identificador de CDSs, tal que a , c , g e t são números complexos arbitrários com $A + C + G + T = 0$. Para maximizar a capacidade discriminatória, estes valores estatísticos devem ser obtidos de outras seqüências similares ou homólogas⁶ cuja codificação seja comprovada, isto é, de éxons anotados. Usando uma seqüência sintética aleatória do mesmo tamanho da seqüência cujas CDSs são conhecidas, pode-se obter os coeficientes A_R , C_R , G_R e T_R ; e pelas definições da Seção 4.2, as seqüências A, C, G e T representam um conjunto redundante, um dos coeficientes podendo ter o valor zero. Em [4] atribuiu-se o coeficiente $c = 0$. Assim $W = |a.A + g.G + t.T|^2$. O problema se reduz a achar números complexos a , g , e t que maximizem a quantidade

$$p(a, g, t) = \frac{E\{|aA + gG + tT|\} - E\{|aA_R + gG_R + tT_R|\}}{std(|aA + gG + tT|) + std(|aA_R + gG_R + tT_R|)} \quad (4.17)$$

com as condições $E\{\arg\{a.A + t.T + g.G\}\} = 0$ e $|a| + |t| + |g| = 1$.

Para os genes do cromossomo XVI de *S. cerevisiae* Anastassiou estimou [4]: $a = 0.10 + 0.12j$, $c = 0$, $g = 0.45 - 0.19j$, $t = -0.30 - 0.20j$. Na Figura 4.7 apresentamos os valores dos coeficientes W (Eq. 4.16) para o gene anteriormente tratado. Vemos que o resultado é significativamente melhor que o método usado em (Eq. 4.15). Pois os coeficientes correspondentes à região do primeiro éxon do gene F56F11.4 (Tab. 4.3) são comparáveis aos coeficientes dos outros éxons. No entanto, a desvantagem deste método é que precisa-se de uma outra seqüência com CDSs conhecidas para a obtenção das constantes a , c , g e t .

---

⁶Seqüência similar com relacionamento evolucionário.

Figura 4.7: Coeficientes W , segundo [4], estimados para o gene F56F11.4, com um conjunto de tamanhos de janela (esquerda), e com um tamanho de janela determinado (direita).

Vaidyanathan [57] propôs a utilização de filtros digitais⁷ para a identificação de CDSs utilizando como base a TBP presente nas CDSs. Basicamente, esta técnica consiste nos seguintes passos:

- Desenhar um filtro digital passa-banda, centrado em $\omega_0 = 2\pi/3$ (filtro que isola as frequências de periodicidade três das outras frequências);
- Processar, com o filtro digital, as quatro seqüências binárias (u_A , u_C , u_G e u_T) associadas à seqüência de DNA. Suas saídas correspondentes são seqüências y_A , y_C , y_G e y_T ;
- Assim, as CDSs serão indicadas por $Y = |y_A|^2 + |y_C|^2 + |y_G|^2 + |y_T|^2$.

Este método em grande parte elimina o ruído presente nos métodos anteriores, mas a desvantagem é a perda de resolução. Em [57] apresentam-se diferentes filtros e mostram-se resultados similares à identificação de CDSs usando a Equação (4.15).

Os métodos de DSP baseados na transformada de Fourier [4, 51] e filtros digitais [57] não apresentam, atualmente, formulações bastante elaboradas devido à dependência do tamanho de janela

⁷Os filtros digitais permitem separar ou dividir um sinal nas suas componentes de frequências [38].

que deve ser definida previamente para analisar localmente uma dada região de DNA. Quanto maior for o tamanho da janela, maior será o número de coeficientes relevantes, e maior o custo computacional gasto (Figs 4.6 e 4.7). Vale salientar que estes métodos consideram somente a energia associada à periodicidade três para a identificação de CDSs.

A definição do tamanho da janela a ser utilizada na análise das seqüências de DNA é crítica para a identificação das CDSs [33], pois compromete fortemente a resolução e o resultado. Com o objetivo de reduzir a dependência ao tamanho, surgiu uma outra classe de métodos para a identificação de CDSs. Estes métodos exploram fortemente o mecanismo de mudança de tamanho da janela, utilizando transformadas em *wavelets* (tratados na Seção 5.1). Na atualidade, valores ótimos para o tamanho de janela e/ou escala são encontrados através de experimentação [33].

Assim, existem estudos preliminares na utilização de transformadas em *wavelets* para a identificação de CDSs. Em [40], utilizaram o mapeamento numérico fixo baseado no EIIP (Seção 4.2), atribuindo valores correspondentes aos potenciais das bases (A, C, G, T)=(0.1260, 0.1340, 0.0806, 0.1335). As *wavelets* utilizadas foram as Coiflets e Daubechies para decompor o sinal.

Todos os métodos de análise tempo-frequência, revisados nesta dissertação, são independentes das propriedades estruturais biológicas existentes nas seqüências de DNA. Em [12], os autores utilizaram uma transformada em *wavelet* adaptativa⁸, chamada *bionic* para a identificação de CDSs. A *wavelet bionic* baseia-se no modelo auditivo humano, onde a resolução tempo-frequência pode ser ajustada adaptativamente e a módulo dos coeficientes associada às frequências pode ter uma melhor distribuição.

Este método permite tanto a utilização do mapeamento binário fixo, como do mapeamento baseado em critérios de otimização. Nos dois tipos de mapeamento, é utilizado um vetor de pesos associado à seqüência de DNA. O mapeamento baseado no critério de otimização, refere-se ao processamento iterativo das seqüências binárias conjuntamente com o vetor de pesos. Este vetor é modificado pelos módulos dos coeficientes da transformada em *wavelets* de uma iteração prévia. A maximização da distribuição da energia (os módulos dos coeficientes) é realizada mediante o processamento das seqüências numéricas pela transformada em *wavelet bionic*. Os maiores módulos dos coeficientes associados às seqüências analisadas foram as correspondentes à frequência próxima a 0.33Hz [12].

---

⁸Refere-se ao fato de mudança iterativa dos valores nas seqüências utilizadas.

Vários outros estudos foram realizados, como o uso de técnicas de análise espectral, para identificar periodicidades latentes e características de interesse biológico [50]. Em seqüências de DNA foram utilizados transformadas em *wavelets* para explorar a complexidade da composição de bases [6], e visualização de padrões regulares [18]. Comparações de seqüências de proteínas, baseadas na transformada em *wavelets*, estão descritas em [52].

4.6 Medidas de Acurácia de Identificação de Genes

As medidas de acurácia no nível dos nucleotídeos, descritas por M. Burset e R. Guigó [9], propõem uma forma de comparação de regiões identificadas, mediante procedimentos computacionais, com CDSs conhecidas biologicamente, isto é, regiões cujas posições dos éxons, nos genes, são conhecidas. Portanto, estas medidas servem para comparar o desempenho dos programas computacionais de identificação de genes (Seção 2.4).

A medição de regiões identificadas contra CDSs conhecidas é feita mediante contagem de nucleotídeos correspondentes aos (1) verdadeiros positivos (TP, *true positive*), que são regiões corretamente identificadas, (2) falsos positivos (FP, *false positive*), que são regiões identificadas erroneamente, (3) verdadeiros negativos (TN, *true negative*), que são regiões corretamente não identificadas, e (4) falsos negativos (FN, *false negative*), que são regiões não identificadas (Fig. 4.8). Foram propostas duas medidas de acurácia baseadas nas taxas de ocorrências de sensibilidade e especificidade.

Figura 4.8: Os quatro possíveis resultados de regiões identificadas contra CDSs conhecidas.

- *Sensibilidade* (Sn), é a proporção de nucleotídeos codificantes corretamente identificados como codificantes.

$$Sn = \frac{TP}{TP + FN} \quad (4.18)$$

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

$$Sp = \frac{TP}{TP + FP} \quad (4.19)$$

Um programa computacional é considerado exato se sua Sn e Sp são simultaneamente altas. Assim, a medida de exatidão que combina a Sp e Sp é chamada de *coeficiente de correlação* (CC),

$$CC = \frac{TP.TN - FN.FP}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \quad (4.20)$$

Os valores dos CC vão de -1, quando a identificação é incorreta, a +1, quando a identificação é correta [9, 39].

Uma recente avaliação de alguns programas de identificação de genes foi realizada em [37, 45]. Outras medidas de acurácia, no nível de éxons e de proteínas, podem ser revisadas em [9]. Vale salientar que todas estas medidas não estão baseadas em componentes de desempenho como o tempo de execução ou requerimentos de memória dos programas computacionais avaliados.

Um *script* em PERL^{TM} para estimação das medidas de acurácia, de conjuntos de seqüências identificadas, está disponível no endereço <http://www.vision.ime.usp.br/~jmena/DSPgenomics/>.

Capítulo 5

Identificação de CDSs através da MMT

Como visto no capítulo anterior, a principal desvantagem dos métodos de DSP existentes para identificação de CDSs é a dependência da definição do tamanho da janela de análise. Este problema é resolvido no método apresentado nesta dissertação.

Neste capítulo, revisamos brevemente as definições básicas da análise em *wavelets* e apresentamos a definição da MMT proposta. Em seguida, descrevemos o novo método desenvolvido para a identificação de CDSs em seqüências de DNA onde a periodicidade de três bases seja latente, de maneira a contornar esse problema. Consideramos este método a principal contribuição da dissertação.

5.1 Análise em *Wavelets*

A transformada em *wavelets* permite uma análise tempo-escala de um sinal, que consiste na sua representação em termos de sinais simples construídas por translações e dilatações de uma *wavelet* de análise ψ ¹. A transformada em *wavelets* contínua de um sinal u é calculada a partir do produto interno do sinal com o conjugado complexo da *wavelet* de análise $\psi : \mathcal{R} \rightarrow \mathcal{C}$ [10]:

$$U(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u(t) \psi^* \left(\frac{t-b}{a} \right) dt, \quad (5.1)$$

em que $a, b \in \mathcal{R}$. $a > 0$, é o coeficiente de escala que permite a compressão ou expansão da função *wavelet*, sendo inversamente relacionada à frequência, b é o coeficiente que permite a sua translação através do eixo do tempo (ou posição), e $\psi(t)$ é a função base *wavelet*. O fator $\frac{1}{\sqrt{a}}$ é usado para

---

¹Também conhecida como *wavelet* mãe, protótipo ou básica.

normalização da energia.

Para sinais discretos, a transformada em *wavelets* do sinal u , é definida como

$$U(b, a) = 2^{-\frac{a}{2}} \sum_k u(k) \psi(2^{-a}k - b), \quad (5.2)$$

em que a e b são valores discretos, e a função discreta ψ pode ser tomada como uma versão amostrada da contraparte contínua. Veja [10, 14], para uma descrição maior dos tipos de transformada em *wavelets*.

A *wavelet* de análise ψ é geralmente escolhida para ser bem localizada em tempo e frequência. Esta função pode ser real ou complexa, resultando também em uma transformada real ou complexa. Na análise do sinal, nenhuma escala é privilegiada pois a mesma função ψ é utilizada, em diversas escalas. Portanto a transformada mantém uma resposta fortemente dependente da função ψ [10].

A transformada em *wavelets* permite calcular os coeficientes dos espectros de frequência para cada posição de um sinal específico, isto é, a energia em função de posição e frequência. Os espectros de frequência derivados da transformada de Fourier (Seção 4.1.1), mantêm energia como função de frequência, onde toda a informação espacial é oculta ou perdida. Já no caso da STFT (Seção 4.1.1) esta informação é obtida, mas dependente do tamanho da janela. Esta vantagem sobre a transformada de Fourier e STFT é alcançada através do uso das funções *wavelets* [44]. Analogamente ao espectrograma (Seção 4.1.4), as magnitudes da transformada em *wavelets* definem o escalograma da função [44]. O escalograma será explorado pelo método de identificação de CDSs introduzido neste trabalho e tratado na Seção 5.3.

5.2 Transformada Modificada de Morlet

A primeira *wavelet* de análise estudada, e comumente usada, no contexto da teoria das *wavelets* foi a *wavelet* de Morlet ou Gaussiana modulada. A função de análise de Morlet, como definida em [10], permite variar o desvio padrão da Gaussiana juntamente com a frequência da exponencial complexa,

$$\psi_M(t) = e^{j\omega_0 t} e^{-\frac{t^2}{2}} \quad (5.3)$$

em que w_0 é a frequência de ψ (Fig. 5.2(a)). Por outro lado, na função de análise de Gabor,

$$g(t, a) = e^{jat} e^{-\frac{t^2}{2}} \quad (5.4)$$

varia-se a frequência da exponencial complexa, mantendo constante o desvio padrão da Gaussiana [15] (Fig. 5.2(b)). A função de análise de Morlet e a de Gabor são particularmente apropriadas para a análise de padrões periódicos locais, pois são bem localizadas no domínio do tempo e da frequência [15]. Essas duas abordagens permitem analisar sinais mediante a variação da frequência da exponencial complexa. Portanto, toda região periódica existente no sinal terá coeficientes de transformação $U(b, a)$ proporcionais ao comprimento do período da região analisada. Isso sugere que regiões com diversas frequências podem ser analisadas.

No presente estudo, onde somente estamos interessados na análise de uma frequência específica², as funções de análise anteriores não são completamente adequadas pois variam as frequências da exponencial complexa na sua transformação. Aqui definimos uma modificação da função de análise de Morlet para analisar um sinal em uma dada frequência fixa (frequência da exponencial complexa constante), e de escala variável.

Segundo (5.3), temos que

$$\psi\left(\frac{t-b}{a}\right) = e^{j\omega_0\left(\frac{t-b}{a}\right)} e^{-\frac{\left(\frac{t-b}{a}\right)^2}{2}} \quad (5.5)$$

Assim, em (5.1), a transformada de um sinal u em *wavelets* de Morlet será dada por

$$U(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u(t) e^{j\omega_0\left(\frac{t-b}{a}\right)} e^{-\frac{\left(\frac{t-b}{a}\right)^2}{2}} dt \quad (5.6)$$

em que w_0 é a frequência de ψ , a o coeficiente da escala, e b o coeficiente de translação através do eixo do tempo. A MMT também utiliza uma função Gaussiana multiplicada por uma exponencial complexa. Esta função, através do parâmetro de escala a , permite variar o desvio padrão da Gaussiana, mantendo constante a frequência da exponencial complexa, ou seja,

$$U(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u(t) e^{j\omega_0(t-b)} e^{-\frac{(t-b)^2}{2a^2}} dt \quad (5.7)$$

---

²A busca de regiões com periodicidade de três bases, na sequência de DNA, associadas às regiões codificantes de proteínas.

Portanto, a função de análise ψ da MMT é definida como

$$\psi_{MM}(t, a) = e^{j\omega_0 t} e^{-\frac{t^2}{2a^2}} \quad (5.8)$$

cujas transformadas de Fourier é

$$\Psi_{MM}(\omega, a) = \sqrt{2\pi} e^{-a^2 \frac{(\omega - \omega_0)^2}{2}} \quad (5.9)$$

Note que a frequência de ψ_{MM} varia de acordo com o valor definido por ω_0 . No caso discreto, se consideramos um tamanho arbitrário de N pontos para ψ_{MM} , e $\omega_0 = N/k$, para um real positivo k , ψ_{MM} terá uma periodicidade que oscila k vezes.

Na Figura 5.1 apresentamos três representações de ψ_{MM} , e suas transformadas de Fourier, considerando arbitrariamente $\omega_0 = 6.66$ e t no intervalo $(-10, 10)$. Os valores das escalas são os correspondentes a 0.5, 1 e 4. Como $\omega_0 = 6.66 = 20/3$, espera-se que a ψ_{MM} oscile a cada três pontos. A Figura 5.2 ilustra graficamente as diferenças entre as funções de análise de Gabor, Morlet, e Morlet modificado, com valores arbitrários de ω_0 , a e b . Observe-se que, dessas três funções, na função de análise da modificação de Morlet, a frequência é mantida constante.

Finalmente, é importante destacar que, atualmente, desconhecemos se verdadeiramente a MMT é uma transformada em *wavelets*.

(a) $a = 0.5$ (b) $a = 1$ (c) $a = 4$

Figura 5.1: Representações da transformada modificada de Morlet (esquerda) e suas correspondentes transformadas de Fourier (direita). Foram considerados $\omega_0 = 6.6$, t no intervalo $(-10, 10)$ e escalas a de 0.5, 1 e 4.

Figura 5.2: Representações das funções de análise. Em (a) varia-se a frequência da exponencial complexa, mantendo constante o desvio padrão da Gaussiana. Em (b) varia-se o desvio padrão da Gaussiana juntamente com a frequência da exponencial complexa. Em (c) varia-se o desvio padrão da Gaussiana, mantendo constante a frequência da exponencial complexa.

5.3 Método para a Identificação de CDSs através da MMT

Na Seção 4.5, revisamos métodos de DSP utilizados na identificação de CDSs que precisam da definição prévia de algum critério para adotar um tamanho de janela que seja apropriado na análise de toda a seqüência de DNA. Assim, os resultados obtidos dessa análise são dependentes do tamanho de janela utilizado. Por outro lado, os métodos que reduzem a dependência ao tamanho de janela, exploram a transformação em *wavelets*, mas todos esses estudos ainda estão em uma etapa preliminar. Tais métodos apresentam uma verificação visual dos resultados obtidos, pois a grande dificuldade na identificação de CDSs, mediante a procura de periodicidade, é a determinação das fronteiras entre CDSs e regiões não-codificantes de proteínas. Esta dificuldade deve-se à transformada utilizada e em grande parte à periodicidade de três bases de baixa latência existentes em algumas CDSs.

Nesta seção, descrevemos um novo método utilizando a MMT para a identificação de CDSs, onde a periodicidade de três bases seja latente. O método divide o processo automático de identificação em quatro blocos importantes: (1) mapeamento de seqüências de DNA em quatro seqüências binárias, (2) aplicação da MMT sobre cada seqüência binária, (3) projeção dos módulos dos coeficientes normalizados da MMT, e (4) limiarização dos coeficientes de projeção, para a detecção das possíveis fronteiras entre as CDSs. Na Figura 5.3 mostramos um diagrama de blocos do método proposto.

Figura 5.3: Diagrama de blocos do novo método utilizando a MMT para a identificação de CDSs onde a periodicidade de três bases seja latente. Cada bloco representa uma operação e cada seta o fluxo da informação.

5.3.1 Mapeamento

Utilizamos as primeiras quatro regras do mapeamento fixo binário (Seção 4.2), para o mapeamento de seqüências de DNA em seqüências numericamente tratáveis.

Cada seqüência binária representa as posições de cada base nitrogenada na seqüência de DNA. Com este mapeamento, não mantém-se dependência sobre alguma atribuição numérica arbitrária. Na Tabela 4.2, apresentamos um exemplo do mapeamento binário para uma seqüência de DNA. Denotamos as seqüências binárias associadas a seqüências de DNA s como u_A , u_C , u_G , e u_T correspondentes às bases A, C, G, e T, como descrito na Equação (4.10).

5.3.2 Aplicação da MMT

No passo seguinte, para um tamanho arbitrário de N pontos, a MMT com $\omega_0 = N/3^3$ e diferentes escalas, é calculada para as seqüências binárias associadas à seqüência de DNA. A definição de $\omega_0 = N/3$ sugere uma periodicidade três em (5.8), como descrita na Seção 5.2.

A MMT para cada seqüência binária u_A , u_C , u_G e u_T , é dada por:

$$U_A(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u_A(t) \psi_{MM}^*(t - b, a) dt \quad (5.10)$$

$$U_C(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u_C(t) \psi_{MM}^*(t - b, a) dt \quad (5.11)$$

$$U_G(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u_G(t) \psi_{MM}^*(t - b, a) dt \quad (5.12)$$

$$U_T(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} u_T(t) \psi_{MM}^*(t - b, a) dt \quad (5.13)$$

em que, os sinais resultantes $U_A(b, a)$, $U_C(b, a)$, $U_G(b, a)$ e $U_T(b, a)$ representam os coeficientes da transformação utilizando ψ_{MM} .

A fim de manter uma medida comparável em todas as escalas, os módulos ao quadrado dos coeficientes da transformação são multiplicados por um fator de normalização proporcional à escala utilizada. Esta normalização deve-se ao fato que para escalas maiores os coeficientes correspondentes às regiões com TBP latente são menores. Estes coeficientes são denotados como os espectros da MMT, tal que

$$m_A(b, a) = a |U_A(b, a)|^2 \quad (5.14)$$

$$m_C(b, a) = a |U_C(b, a)|^2 \quad (5.15)$$

---

³Um comportamento similar de ψ_{MM} pode ser obtido com a utilização de $\omega_0 = 2N/3$.

$$m_G(b, a) = a |U_G(b, a)|^2 \quad (5.16)$$

$$m_T(b, a) = a |U_T(b, a)|^2 \quad (5.17)$$

em que b representa a posição dos sinais, e a a escala utilizada.

As Equações (5.14)-(5.17) representam medidas parciais da energia associada à transformação. Conseqüentemente, a medida espectral total da sequência de DNA é dada por:

$$M(b, a) = m_A(b, a) + m_C(b, a) + m_G(b, a) + m_T(b, a) \quad (5.18)$$

Uma vez que a sequência de DNA seja representada pela transformada multi-escala, diferentes aproximações podem ser consideradas afim de extrair informações dela [15]. Assim, esta transformação pode ser aplicada para um intervalo grande de escalas a , exponencialmente espaçadas e definidas *a priori*, para analisar as sequências binárias. No Capítulo 6, apresentamos os melhores resultados para valores de escala no intervalo 0.05 a 0.5. Estes valores ótimos foram encontrados através de experimentação.

No método, quando o mapeamento fixo é utilizado para converter uma sequência de DNA em quatro sequências binárias, é intrinsecamente assumido que todas as bases são independentes pois cada sequência binária é tratada de forma independente. Entretanto, sequências genômicas, incluídas as sequências de DNA, são criadas e mantidas por complicados processos biológicos [2], existindo uma interação entre bases adjacentes. Esta interação é geralmente estimada mediante correlações de sequências de DNA [5, 28]. Os coeficientes normalizados totais $M(b, a)$, para uma escala a , agrupa os coeficientes associados às transformações das sequências binárias mantendo, em parte, uma interação entre elas.

5.3.3 Projeção dos Coeficientes da MMT

Os coeficientes do espectrograma da MMT (5.18) são projetados no eixo das posições, a fim de representar as possíveis CDSs com TBP. Para uma sequência de comprimento N , os coeficientes de projeção que definem o identificador proposto de CDSs, serão representados por uma somatória dos coeficientes em todas as escalas, como

$$M_p(b) = \sum_a M(b, a) \quad (5.19)$$

para todo b entre 1 e N . Existe uma forte relação da projeção M_p com a transformada de Fourier, pois a informação da escala é perdida durante a projeção. Entretanto, na transformada de Fourier, a informação de tempo é perdida. As projeções no eixo das escalas revelam qual delas mantém maior energia através das posições,

$$M_s(a) = \sum_{b=1}^N M(b, a) \quad (5.20)$$

para todas as escalas a .

Estas duas projeções são uma maneira simples de extrair informação da transformada multiescala MMT nos eixos [15]. Tomando um exemplo dos coeficientes de projeção, é razoável concluir que se duas seqüências têm CDSs com TBP similar, então estes coeficientes devem também ser similares.

5.3.4 Limiarização dos Coeficientes de Projeção

Uma parte importante na identificação de CDSs é a determinação das fronteiras entre CDSs e regiões não codificantes de proteínas. O processo da limiarização, neste método, corresponde a uma tentativa de estabelecimento dessas fronteiras, em que são substituídos por zero qualquer coeficiente menor que um valor limite, mantendo constantes os coeficientes restantes.

A limiarização sobre M_p , na Equação (5.19), é uma maneira de eliminar detalhes “não importantes” considerados como ruído, ou como regiões com TBP com baixa latência. Consideramos um limiar percentual [58] em que são selecionados uma porcentagem das magnitudes dos coeficientes de projeção. O conjunto dos menores valores, em termos absolutos, são substituídos por zero. Portanto, uma aproximação das CDSs são as regiões isoladas cuja magnitudes M_p são maiores que zero.

Capítulo 6

Resultados Experimentais

Neste capítulo, discutimos alguns resultados experimentais obtidos com o método proposto para a identificação de CDSs onde a periodicidade de três bases seja latente (Seção 5.3). Apresentam-se experimentos comparativos de identificação de CDSs através da MMT e da STFT, em seqüências sintéticas e reais. Conjuntos de seqüências de DNA (Apêndice A) também foram utilizados na experimentação.

Mostramos nas figuras associadas à análise de cada seqüência, os coeficientes da MMT normalizados (Eq. 5.18), os coeficientes da projeção no eixo das posições da MMT (Eq. 5.19), e os coeficientes normalizados da STFT (Eq. 4.15) com comprimentos de janela de 200bp e 400bp. Mostramos a análise para cada seqüência binária associada às bases A, C, G e T. A união, que indica uma medida total da identificação, corresponde à somatória de todos os coeficientes das transformadas. As linhas pontilhadas verticalmente indicam par a par as posições das CDSs biologicamente conhecidas.

Geralmente, dos coeficientes de projeção no eixo das posições, as magnitudes associadas às CDSs são comparavelmente maiores que das outras regiões na seqüência de DNA por terem TBP latente, salvo exceções as quais serão discutidas nas seguintes seções. Observaremos que essa periodicidade não está presente em todos os coeficientes parciais de projeção.

No método de identificação de CDSs através da MMT utilizamos 40 escalas, exponencialmente espaçadas, no intervalo 0.05 e 0.5. Entretanto, no método de identificação de CDSs através da STFT consideramos arbitrariamente comprimentos de janelas de 200bp e 400bp, correspondentes aos valores aproximados da média e do desvio padrão dos comprimentos nos éxons dos conjuntos A e B (Tabelas A.3 e A.6).

Para comparar os métodos tratados utilizamos a mesma limiarização sobre os coeficientes obtidos depois da projeção, na MMT, ou da somatória dos coeficientes normalizados, na STFT. Os limiares percentuais considerados na limiarização estão no intervalo de 5% a 95%. Assim, as regiões isoladas com magnitudes maiores que zero são consideradas possíveis aproximações das localizações das CDSs.

Utilizamos medidas de acurácia para avaliar o desempenho dos métodos em termos de sensibilidade (S_n , Eq. 4.18), especificidade (S_p , Eq. 4.19) e coeficientes de correlação (CC , Eq. 4.20) nas identificações realizadas para um dado limiar. Estas medidas indicam a qualidade nas identificações de forma objetiva. Portanto algumas delas podem não ser consistentes com percepções biológicas. Medidas subjetivas não foram consideradas na análise dos resultados deste trabalho.

Para toda a análise dos resultados, no caso dos organismos eucariotos, consideramos que em cada sequência de DNA existe somente um gene, em que as primeiras e as últimas CDSs correspondem às regiões codificantes dos primeiros e dos últimos éxons, respectivamente. Já no caso dos procariotos, as regiões identificadas corresponderão aos possíveis genes, pois os genes destes organismos mantêm somente uma CDS.

No momento todos os procedimentos, dentro do método automático de identificação, foram implementados usando MATLAB®, estando disponíveis como um *toolbox* no endereço <http://www.vision.ime.usp.br/~jmena/DSPgenomics/>.

6.1 Testes em Sequências Sintéticas de DNA

Foram utilizadas as sequências sintéticas `sintética1_0` e `sintética2_0` descritas no Apêndice A.1 para avaliar a identificação de regiões onde a periodicidade é uniforme.

- Sequência `sintética1_0`, de 6000bp com duas regiões com periodicidade três (Tabela A.1). Na Figura 6.1 mostramos os coeficientes e medidas de desempenho da aplicação das transformadas para a identificação de CDSs. As medidas de acurácia estão indicadas na Tabela C.1.

Observemos que as duas regiões periódicas apresentam altas magnitudes nos coeficientes da transformada. A região que contém periodicidade seis é identificada por ser, em parte, também de periodicidade três. O desempenho da STFT com comprimento de janela de 200bp apresenta uma significativa melhora em relação à de 400bp, pois a periodicidade contida nessas regiões é uniforme.

A definição do comprimento de janela afeta fundamentalmente o desempenho na identificação de CDSs. Utilizando a MMT nessa sequência obteve-se com um limiar de 85% uma acurácia de 0.87. Entretanto, usando a STFT, obteve-se com um limiar de 85% uma acurácia de 0.95 correspondente ao comprimento de janela de 200bp, e com um limiar de 90% uma acurácia de 0.84 correspondente ao comprimento de janela de 400bp. Isso sugere que para regiões com periodicidade uniforme, e com as mesmas configurações nas transformadas, dificilmente será alcançada a identificação exata das CDSs, mas uma identificação aproximadamente razoável.

- Sequência `sintética2_0`, de 1000bp com nove regiões com periodicidade três (Tabela A.2). Na Figura 6.2 mostramos os coeficientes e medidas de desempenho da aplicação das transformadas para a identificação de CDSs. As medidas de acurácia estão indicadas na Tabela C.2.

Todas as regiões periódicas maiores que 30bp apresentam altas magnitudes nos coeficientes da transformada. O desempenho da MMT e da STFT com comprimento de janela de 200bp apresentam desempenho comparável na identificação dessas regiões. Para comprimentos de janela maiores, a acurácia obtida é menor. Com um limiar de 65% obteve-se uma acurácia de 94% utilizando a MMT e a STFT com comprimento de janela de 200bp. Entretanto, usando a STFT com comprimento de janela de 400bp obteve-se com um limiar de 70% uma acurácia de 0.87.

Isso reforça a discussão anterior, no sentido que, o nível de acurácia máximo que poderia eventualmente alcançar-se está próximo ao 90%. É evidente que para sequências reais, cuja periodicidade não é uniforme, este valor é o máximo esperado que poderia obter-se utilizando a busca de regiões com TBP.

6.2 Testes em Sequências Reais de DNA

Utilizamos oito sequências reais de DNA para avaliar a identificação de CDSs onde a TBP seja latente. Todas as sequências, com exceção do gene F56F11.4, pertencem ao conjunto de sequências A (Apêndice A.2).

- Gene BTU02285, do organismo *Bos taurus*, de 6396bp com 6 CDSs (NCBI, número de acesso U02285). Na Figura 6.3 mostramos os coeficientes e medidas de desempenho da aplicação das transformadas para a identificação de CDSs. As medidas de acurácia estão indicadas na Tabela C.3.

Observemos que as seis CDSs apresentam altas magnitudes nos coeficientes das transformadas. A utilização da MMT para a identificação de CDSs tem desempenho superior à STFT. Utilizando a MMT nessa sequência obteve-se com um limiar de 85% uma acurácia de 0.86. Entretanto, usando a STFT obteve-se com um limiar de 85% uma acurácia de 0.79, e com um limiar de 80% uma acurácia de 0.61 correspondentes aos comprimentos de janela de 200bp e 400bp.

- Gene F56F11.4, do organismo *Caenorhabditis elegans*, de 8000bp com 5 CDSs (NCBI, número de acesso AF099922). Na Figura 6.4 mostramos os coeficientes e medidas de desempenho da aplicação das transformadas para a identificação de CDSs. As medidas de acurácia estão indicadas na Tabela C.4.

Observemos que as cinco CDSs apresentam altas magnitudes nos coeficientes das transformadas, mas algumas regiões não codificantes aparentemente tem TBP com baixa latência (primeiro intron). Vale salientar que esse fenômeno ainda não compreendido completamente.

- Gene GGVITIIG, do organismo *Gallus gallus*, de 20343bp com 35 CDSs (NCBI, número de acesso X13607). Na Figura 6.5 mostramos os coeficientes e medidas de desempenho da aplicação das transformadas para a identificação de CDSs. As medidas de acurácia estão indicadas na Tabela C.5.

Observemos que somente uma CDS apresenta alta magnitude nos coeficientes das transformadas. As outras 34 têm comprimento pequeno ou TBP com baixa latência.

- Gene HSDAO, do organismo *Homo sapiens*, de 9903bp com 4 CDSs (NCBI, número de acesso X78212). Na Figura 6.6 mostramos os coeficientes e medidas de desempenho da aplicação das transformadas para a identificação de CDSs. As medidas de acurácia estão indicadas na Tabela C.6.

Observemos que as quatro CDSs apresentam altas magnitudes nos coeficientes das transformadas, mas algumas regiões não codificantes aparentemente têm TBP com baixa latência. É importante notar que a TBP nas CDSs não é uniforme, isto é, não mantém a mesma magnitude no seus coeficientes (veja as magnitudes das transformadas associadas ao primeiro éxon).

- Sequência HUMCS3, do organismo *Homo sapiens*, de 2740bp com 5 CDSs (NCBI, número de acesso M15894). Na Figura 6.7 mostramos os coeficientes e medidas de desempenho da aplicação

das transformadas para a identificação de CDSs. As medidas de acurácia estão indicadas na Tabela C.7.

Observemos que somente as três últimas CDSs apresentam altas magnitudes nos coeficientes das transformadas. A primeira CDS não foi identificada pois o comprimento dela é de 10bp. A segunda CDS, de 161bp, claramente não tem TBP latente. Isso sugere que em uma sequência de DNA, não toda CDS tem TBP latente e diferenciável das outras regiões.

Por último, na primeira região não codificante, apresenta-se uma sub-região com TBP comparável à da terceira CDS. Possivelmente esta observação deve-se a que a primeira sub-região pertença à uma CDS não anotada.

- Sequência MMACLGNA, do organismo *Mus musculus*, de 3881bp com 3 CDSs (NCBI, número de acesso Z24722). Na Figura 6.8 mostramos os coeficientes e medidas de desempenho da aplicação das transformadas para a identificação de CDSs. As medidas de acurácia estão indicadas na Tabela C.8.

Observemos que muitas sub-regiões nas regiões não-codificantes apresentam altas magnitudes nos coeficientes das transformadas. Isso reforça a discussão anterior no sentido que possivelmente estas sub-regiões pertençam a CDSs não anotadas ou sejam sub-regiões pertencentes a pseudo-genes¹ [61]. Este é um fenômeno ainda pouco estudado e ainda não compreendido.

6.3 Testes em Conjuntos de Sequências de DNA

Para seis conjuntos de sequências de DNA (Apêndice A.2) estimamos as medidas de acurácia (Tabela C.9 a C.14). Mostramos, nas Figuras 6.9 e 6.10, as medidas de desempenho da MMT e da STFT em termos de Sn e Sp , e de CC e limiar para todos os conjuntos tratados.

Para os conjuntos de sequências A e B, a utilização da MMT, no método, apresenta desempenho superior à STFT com comprimentos de janela de 200bp e 400bp. Para o conjunto A, obteve-se uma acurácia máxima de 0.47, com um limiar de 80%. Por outro lado, para o conjunto B, obteve-se uma acurácia máxima de 0.42, com um limiar de 80%. Níveis de acurácia similares foram obtidos para os conjuntos Am30, e Bm30.

---

¹Sequência de DNA derivada originalmente de genes codificantes de proteínas, que foram perdendo sua função.

Vale salientar que o método proposto para a identificação de CDSs, através da MMT, apresenta melhor desempenho em seqüências de DNA que possuam CDSs cujos comprimentos sejam maiores que 100bp. Para o conjunto Am100, obteve-se uma acurácia máxima de 0.55, com um limiar de 75%. Por outro lado, para o conjunto Bm100, obteve-se uma acurácia máxima de 0.51, com um limiar de 70%

Finalmente, mostramos os histogramas dos comprimentos dos éxons e dos íntrons para os conjuntos através da MMT (Figs. 6.11 e 6.12) em que obtiveram-se os melhores níveis de acurácia. Notemos que as distribuições nas frequências dos comprimentos dos éxons e íntrons mantém a mesma forma que as distribuições calculadas para as seqüências reais (Figs. A.1 e A.2). Essa forma nas distribuições é um bom sinal do desempenho do método desenvolvido. Acreditamos que uma análise mais aprofundada do método deve seguir essa abordagem.

Figura 6.1: Desempenho da MMT e da STFT para a sequência `sintética1_0`.

(a) Coeficientes MMT normalizados

(b) Coeficientes da projeção da MMT

(c) Coeficientes STFT com janela de 200bp

(d) Coeficientes STFT com janela de 400bp

(e) Desempenho em termos de S_n e S_p (f) Desempenho em termos de CC e limiarFigura 6.2: Desempenho da MMT e da STFT para a sequência `syntica2_0`.

(a) Coeficientes MMT normalizados

(b) Coeficientes da projeção da MMT

(c) Coeficientes STFT com janela de 200bp

(d) Coeficientes STFT com janela de 400bp

(e) Desempenho em termos de S_n e S_p (f) Desempenho em termos de CC e limiar

Figura 6.3: Desempenho da MMT e da STFT para o gene BTU02285.

(a) Coeficientes MMT normalizados

(b) Coeficientes da projeção da MMT

(c) Coeficientes STFT com janela de 200bp

(d) Coeficientes STFT com janela de 400bp

(e) Desempenho em termos de S_n e S_p (f) Desempenho em termos de CC e limiar

Figura 6.4: Desempenho da MMT e da STFT para o gene F56F11.4.

(a) Coeficientes MMT normalizados

(b) Coeficientes da projeção da MMT

(c) Coeficientes STFT com janela de 200bp

(d) Coeficientes STFT com janela de 400bp

(e) Desempenho em termos de S_n e S_p (f) Desempenho em termos de CC e limiar

Figura 6.5: Desempenho da MMT e da STFT para o gene GGVITIIG.

(a) Coeficientes MMT normalizados

(b) Coeficientes da projeção da MMT

(c) Coeficientes STFT com janela de 200bp

(d) Coeficientes STFT com janela de 400bp

(e) Desempenho em termos de S_n e S_p (f) Desempenho em termos de CC e limiar

Figura 6.6: Desempenho da MMT e da STFT para o gene HSDAO.

(a) Coeficientes MMT normalizados

(b) Coeficientes da projeção da MMT

(c) Coeficientes STFT com janela de 200bp

(d) Coeficientes STFT com janela de 400bp

(e) Desempenho em termos de S_n e S_p (f) Desempenho em termos de CC e limiar

Figura 6.7: Desempenho da MMT e da STFT para o gene HUMCS3.

(a) Coeficientes MMT normalizados

(b) Coeficientes da projeção da MMT

(c) Coeficientes STFT com janela de 200bp

(d) Coeficientes STFT com janela de 400bp

(e) Desempenho em termos de S_n e S_p (f) Desempenho em termos de CC e limiar

Figura 6.8: Desempenho da MMT e da STFT para o gene MMACLGNA.

(a) Conjunto A

(b) Conjunto Am30

(c) Conjunto Am100

Figura 6.9: Desempenho da MMT e da STFT em termos de S_n e S_p (esquerda), e de CC e limiar (direita) para os conjuntos A, Am30 e Am100.

(a) Conjunto B

(b) Conjunto Bm30

(c) Conjunto Bm100

Figura 6.10: Desempenho da MMT e da STFT em termos de S_n e S_p (esquerda), e de CC e limiar (direita) para os conjuntos B, Bm30 e Bm100.

(a) Conjunto A, identificado com limiar de 75%

(b) Conjunto Am30, identificado com limiar de 80%

(c) Conjunto Am100, identificado com limiar de 75%

Figura 6.11: Histogramas dos comprimentos dos éxons (esquerda), e dos íntrons (direita) para os conjuntos A, Am30 e Am100 identificados através da MMT.

(a) Conjunto B, identificado com limiar de 70%

(b) Conjunto Bm30, identificado com limiar de 70%

(c) Conjunto Bm100, identificado com limiar de 70%

Figura 6.12: Histogramas dos comprimentos dos éxons (esquerda), e dos íntrons (direita) para os conjuntos B, Bm30 e Bm100 identificados através da MMT.

Capítulo 7

Conclusões

7.1 Considerações Finais

Atualmente a identificação de genes é um problema complexo, e a identificação de regiões com TBP é unicamente mais um passo na frente para a identificação de CDSs [57]. Assim, neste estudo, um novo método foi introduzido para a identificação computacional de CDSs, de organismos procariotos ou eucariotos, em que a TBP seja latente. Este método, que emprega uma nova transformada, ou seja, a MMT, pode ajudar aos especialistas anotadores de genoma, fornecendo uma estrutura geral para encontrar regiões codificantes em uma dada sequência de DNA desconhecida.

A principal vantagem do método proposto é a de basear-se unicamente na TBP existente, geralmente, nas CDSs. Não é usada nenhuma outra informação adicional, como conjuntos de treinamento ou sequências homólogas de outras espécies cujas sequências e genes já sejam conhecidas [4]. Nesse sentido, o método em parte é flexível, aceitando sequências de DNA que não sejam bem caracterizadas [39]¹. Acredita-se que, com o uso dessa informação, pode-se obter um identificador com um nível de acurácia maior. Também é importante observar que este novo método é mais robusto à variação de escala. Tal dependência é um dos maiores problemas dos outros métodos similares propostos até hoje.

Mediante experimentos comparativos dos métodos de identificação de CDSs podemos concluir que a MMT tem um desempenho superior à STFT. Em ambos métodos obtiveram-se melhor acurácia quando o valor de limiarização é próximo de 85%. Isso indica que aproximadamente o 15% dos

---

¹Este tipo de sequências, geralmente, é usado pelos métodos que utilizam a abordagem de identificação de genes mediante a busca por sítios (Seção 2.4).

coeficientes de projeção serão indicadores de TBP. Esta porcentagem está diretamente relacionada com as estatísticas extraídas dos conjuntos de seqüências de organismos eucariotos utilizadas nas experimentações (Veja o apêndice A.2). Deverá ser considerado um valor de limiarização menor para análise de seqüências de organismos procariotos (e.g. 40%, correspondente à porcentagem das bases não pertencentes às CDSs nos organismos procariotos [27]). Finalmente, um nível de acurácia maior é alcançado quando os comprimentos das CDSs, nas seqüências de DNA analisadas, são maiores que 100bp.

7.2 Sugestões para Pesquisas Futuras

Na análise de seqüências de DNA e, em particular, a identificação de CDSs, muitas questões estão em aberto [16, 61]. Apesar do grande progresso da identificação de genes e do reconhecimento de padrões baseados na análise de DNA, ainda estamos em uma etapa inicial da bioinformática [48]. A seguir, descrevemos algumas questões importantes, na nossa pesquisa, às quais por enquanto não podemos responder. Estas questões precisam ser pesquisadas a futuro.

- A MMT, derivada da *wavelet* de Morlet, suporta particularmente detecção de uma determinada periodicidade latente no sinal analisado. Entretanto, outras transformadas em *wavelets*, como *wavelets packets* [60], baseadas em critérios matemáticos ou de entropia estatística podem ser pesquisadas. *Wavelets packets* e transformadas em *wavelets* adaptativa já provaram ter melhor desempenho que a transformada em *wavelets* padrão em determinadas situações particulares [36].
- Acreditamos que, levando em consideração um mapeamento de nucleotídeos dependente da seqüência, ou adaptativo, baseado em otimização, podemos representar, de alguma forma, a estrutura específica da seqüência de DNA analisada. Usando vetores de peso sobre as seqüências binárias associadas à seqüência de DNA, em um processo iterativo, pode-se obter maior informação de regiões periódicas?
- Uma questão importante na projeção dos coeficientes da MMT, no eixo das posições, associadas à seqüência de DNA analisada, é a determinação das escalas utilizadas. Resultados diferentes foram obtidos utilizando diferentes escalas. Será suficiente a utilização de uma determinada escala? Qual seria um valor dessa escala? Além da experimentação, como poderíamos encontrar um valor adequado de escala e dependente da estrutura genômica da seqüência analisada?

- As magnitudes da projeção, no eixo das posições, são consideradas como indicadoras de TBP. Estas medidas poderiam ser utilizadas como possíveis medidas de probabilidade de CDSs? Acreditamos que as magnitudes de tais coeficientes serviriam para a construção de um método híbrido de identificação de CDSs mais robusto, utilizando alguma outra abordagem, tal como a busca por sítios (Seção 2.4).
- As medidas de acurácia [9,45] utilizadas na comparação dos programas computacionais de identificação de genes são restringidas, no nível de nucleotídeos, à determinação de especificidade e sensibilidade, mas não consideram características como a distribuição do número de CDSs por gene, distribuição dos tamanhos dos CDSs identificados, dentre outros. Assim, outras medidas devem ser pesquisadas.
- A interação entre bases existente no DNA não é ainda conhecida completamente [2,24]. Como poderia ser representada esta interação com uma forma diferente de mapeamento ou transformação?

Apêndice A

Seqüências de DNA Testadas no Trabalho

A.1 Seqüências Sintéticas

Descrevemos as composições das regiões periódicas das seqüências sintéticas usadas nos experimentos apresentados neste trabalho. As posições não indicadas nessas seqüências correspondem às regiões aleatórias com probabilidade de atribuição de 25% para cada uma das bases. Todas as seqüências criadas sinteticamente estão disponíveis no endereço <http://www.vision.ime.usp.br/~jmena/DSPgenomics/>.

- **sintética1_0**: seqüência de 6000bp contendo 6 regiões periódicas (Tab. [A.1](#)).
- **sintética2_0**: seqüência de 10000bp contendo 9 regiões que representam uma TBP uniforme (Tab. [A.2](#)).
- **sintética2_XA**, seqüência que mantém a mesma composição da seqüência **sintética2_0**, mas com ruído em X% das bases nas regiões periódicas. O ruído inserido nestas regiões corresponde à escolha aleatória de X% das bases para depois serem trocadas com a mesma probabilidade por A, C, G ou T. Aqui chamamos a este tipo de ruído como *inclusivo*.
- **sintética2_XB**, seqüência que mantém a mesma composição da seqüência **sintética2_0**, mas com ruído em X% das bases nas regiões periódicas. O ruído inserido nestas regiões corresponde à escolha aleatória de X% das bases para serem trocadas exclusivamente por outras. Por exemplo, se a base escolhida for A, esta pode ser trocada com igual probabilidade por C, G ou T. Aqui chamamos a este tipo de ruído como *privativo*.

Posições	Comprimento	Periodicidade	Seqüência periódica
301 - 800	500	2	AT
1401 - 1520	120	3	CGT
1571 - 1770	200	4	TGCA
2571 - 2970	400	5	ACTNN
3371 - 4170	800	6	TGCNNN
5171 - 5870	700	3	TCA

Tabela A.1: Posições das regiões periódicas na seqüência **sintética1.0**.

Posições	Comprimento	Periodicidade	Seqüência periódica
501 - 900	400	3	ATC
1501 - 1620	120	3	TAC
1721 - 1870	150	3	CAG
3791 - 3820	30	3	CAG
4421 - 4670	250	3	GAT
5071 - 7070	2000	3	ACG
8231 - 8310	80	3	TCG
9111 - 9460	350	3	ACT
9661 - 9760	100	3	GAC

Tabela A.2: Posições das regiões periódicas da seqüência **sintética2.0**.

A.2 Seqüências Reais

As seqüências reais utilizadas são as pertencentes a organismos eucariotos cujos limites éxon-íntron foram anotados, com base nas interpretações de evidências biológicas.

- Conjunto A: Composto por 570 seqüências de DNA de vertebrados e apresentado por M. Burset e R. Guigó no artigo [9], no ano 1996. Cada seqüência contém exatamente um gene com pelo menos um íntron, e todas as seqüências consideradas são aquelas que codificam em produtos de proteína completos, iniciam a sua codificação com o códon ATG, não contém um códon de parada em um segmento das regiões codificantes, o segmento codificante tem um comprimento múltiplo de três, a seqüência tem regiões de corte doadora e aceitadora bem-caracterizados (sítios GT e AG) correspondentemente; dentre outras. Para a descrição completa das seqüências consulte a página disponível em <http://www1.imim.es/databases/genomics96/index.html>.
- Conjunto B: Composto por 195 seqüências de DNA apresentados no artigo [45] por S. Rodic, A. Mackworth e F. Ouellette, no ano 2001. Todas as seqüências desse conjunto foram extraídas dos organismos *Homo sapiens*, *Mus musculus*, e *Rattus norvegicus*, cujos comprimentos sejam

menor que 200 000 bases, e tendo características similares às descritas no conjunto A. Para a descrição completa das seqüências consulte a página disponível em <http://www.cs.ubc.ca/~rogic/evaluation/dataset.html>.

Para a análise e avaliação do desempenho do método desenvolvido neste trabalho, criamos sub-conjuntos correspondentes às seqüências com comprimentos mínimos, nos éxons, de 30bp e 100bp. Assim, para o conjunto A, criamos dois sub-conjuntos denominados setAm30 e setAm100 correspondentes aos comprimentos mínimos de 30bp e 100bp nos éxons. Analogamente, para o conjunto B, criamos dois sub-conjuntos denominados setBm30 e setBm100 correspondentes aos comprimentos mínimos de 30bp e 100bp nos éxons.

Algumas estatísticas extraídas desses conjuntos de seqüências são mostradas nas Tabelas A.3 a A.8. Apresentamos a quantidade de regiões, número de bases nas regiões, e a média e desvio padrão para cada tipo de região nas seqüências de DNA. Nas Figuras A.1 e A.2 são mostradas os histogramas dos comprimentos dos éxons e dos íntrons para todos os conjuntos de seqüências reais aqui tratados.

Região	Quantidade	Bases	Comprimento		Conteúdo				
			Média	Desvio	A	C	G	T	N
Éxon	2649	444498 (15.4%)	168	222	108013	120970	124055	91459	1
Íntron	2079	1310452 (45.3%)	630	909	332478	298396	310586	368597	395
Inter-gênica	1132	1137199 (39.3%)	1004	1464	296109	270067	267702	302844	477
Total	5860	2892149	-	-	736600	689433	702343	762900	873

Tabela A.3: Estatísticas extraídas do conjunto A (570 seqüências).

Região	Quantidade	Bases	Comprimento		Conteúdo				
			Média	Desvio	A	C	G	T	N
Éxon	1993	355968 (15.4%)	178	240	86200	96924	99335	73508	1
Íntron	1524	999122 (43.3%)	655	927	256824	225685	234286	281956	371
Inter-gênica	930	951584 (41.3%)	1023	1552	249184	224757	223252	253961	430
Total	4447	2306674	-	-	592208	547366	556873	609425	802

Tabela A.4: Estatísticas extraídas do conjunto Am30 (469 seqüências).

Região	Quantidade	Bases	Comprimento		Conteúdo				
			Média	Desvio	A	C	G	T	N
Éxon	367	114219 (22.1%)	311	477	27027	32426	33492	21274	0
Intron	264	182947 (35.3%)	693	824	45305	44493	44229	48890	30
Inter-gênica	206	220991 (42.6%)	1072	1193	55229	53386	55502	56867	7
Total	837	518157	-	-	127561	130305	133223	127031	37

Tabela A.5: Estatísticas extraídas do conjunto Am100 (103 seqüências).

Região	Quantidade	Bases	Comprimento		Conteúdo				
			Média	Desvio	A	C	G	T	N
Éxon	948	199176 (14.4%)	210	271	45313	56391	54972	42497	3
Intron	753	642788 (46.4%)	854	130	160740	148259	156493	177184	112
Inter-gênica	390	544044 (39.2%)	1395	2261	137851	130025	131901	144012	255
Total	2091	1386008	-	-	343904	334675	343366	363693	370

Tabela A.6: Estatísticas extraídas do conjunto B (195 seqüências).

Região	Quantidade	Bases	Comprimento		Conteúdo				
			Média	Desvio	A	C	G	T	N
Éxon	745	168126 (13.9%)	226	275	38194	47758	46288	35884	2
Intron	579	550301 (45.8%)	950	1403	139721	125292	132271	152913	104
Inter-gênica	332	484087 (40.3%)	1458	2397	123649	115034	116326	128977	101
Total	1656	1202514	-	-	301564	288084	294885	317774	207

Tabela A.7: Estatísticas extraídas do conjunto Bm30 (166 seqüências).

Região	Quantidade	Bases	Comprimento		Conteúdo				
			Média	Desvio	A	C	G	T	N
Éxon	185	88971 (18.5%)	481	429	20331	25394	23938	19306	2
Intron	102	123658 (25.8%)	1212	1122	33775	27053	28138	34666	26
Inter-gênica	166	267236 (55.7%)	1609	3001	70259	61416	62699	72786	76
Total	453	479865	-	-	124365	113863	114775	126758	104

Tabela A.8: Estatísticas extraídas do conjunto Bm100 (83 seqüências).

Figura A.1: Histogramas dos comprimentos dos éxons (esquerda), e dos íntrons (direita) para os conjuntos A, Am30 e Am100.

Figura A.2: Histogramas dos comprimentos dos éxons (esquerda), e dos íntrons (direita) para os conjuntos B, Bm30 e Bm100.

Apêndice B

Periodicidades em Sequências Sintéticas com Ruído

As estimativas de periodicidade das sequências sintéticas `sintética2_XA` e `sintética2_XB` (Apêndice A.1) são representadas nas Figuras B.1 e B.2.

Observemos que a magnitude da periodicidade três nas regiões periódicas (que simulam TBP existente nas CDSs) é claramente maior que as magnitudes das não periódicas (íntron e inter-gênica) e, na medida que a porcentagem do ruído aumenta, as magnitudes das periodicidades vão tornando-se indistintas para essas regiões. Entretanto, para a sequência `sintética2_100B`, que tem ruído em 100% das bases pertencentes a suas regiões periódicas, existe uma magnitude maior para a periodicidade três em comparação às demais regiões, pois o ruído inserido é privativo. Veja na Figura 4.4(b) um comportamento similar no cálculo dos espectros de frequência.

Figura B.1: Periodicidades estimadas para seqüências sintéticas com ruído. Foram calculadas as periodicidades 2 a 24 das regiões periódicas (que simulam TBP) e regiões não-periódicas (regiões que simulam às regiões inter-gênicas e, no caso dos organismos eucariotos, íntrons).

Figura B.2: Periodicidades estimadas para as seqüências sintéticas com ruído. Foram calculadas as periodicidades 2 a 24 das regiões periódicas (que simulam TBP) e regiões não-periódicas (regiões que simulam às regiões inter-gênicas e, no caso dos organismos eucariotos, íntrons). (*continuação*)

Apêndice C

Tabelas de Acurácia estimadas para Seqüências de DNA

Apresentamos medidas de acurácia (Seção 4.6) estimadas para seqüências de DNA (Apêndice A). Foram utilizadas a MMT com 40 escalas exponencialmente espaçadas no intervalo 0.05 a 0.5, e a STFT com comprimentos de janela de 200bp e 400bp, para a identificação (1) das regiões com periodicidade três, nas seqüências sintéticas; e (2) das regiões com TBP, nas seqüências reais.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	1.00	0.14	0.09	1.00	0.14	0.09	1.00	0.14	0.09
10%	1.00	0.15	0.13	1.00	0.15	0.13	1.00	0.15	0.13
15%	1.00	0.16	0.17	1.00	0.16	0.17	1.00	0.16	0.17
20%	1.00	0.17	0.20	1.00	0.17	0.20	1.00	0.17	0.20
25%	1.00	0.18	0.23	1.00	0.18	0.23	1.00	0.18	0.23
30%	1.00	0.20	0.26	1.00	0.20	0.26	1.00	0.20	0.26
35%	1.00	0.21	0.29	1.00	0.21	0.29	1.00	0.21	0.29
40%	1.00	0.23	0.32	1.00	0.23	0.32	1.00	0.23	0.32
45%	1.00	0.25	0.36	1.00	0.25	0.36	1.00	0.25	0.36
50%	1.00	0.27	0.40	1.00	0.27	0.40	1.00	0.27	0.40
55%	1.00	0.30	0.44	1.00	0.30	0.44	1.00	0.30	0.44
60%	1.00	0.34	0.49	1.00	0.34	0.49	1.00	0.34	0.49
65%	1.00	0.39	0.54	1.00	0.39	0.54	1.00	0.39	0.54
70%	1.00	0.46	0.61	1.00	0.46	0.61	0.97	0.44	0.58
75%	1.00	0.55	0.69	1.00	0.55	0.69	0.85	0.47	0.55
80%	0.96	0.66	0.76	1.00	0.68	0.80	0.85	0.58	0.65
85%	0.93	0.85	0.87	1.00	0.91	0.95	0.85	0.78	0.78
90%	0.73	1.00	0.84	0.73	1.00	0.84	0.73	1.00	0.84
95%	0.36	1.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00

Tabela C.1: Medidas de acurácia estimadas para a sequência `sintética1_0`.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	1.00	0.37	0.17	1.00	0.37	0.17	1.00	0.37	0.17
10%	1.00	0.39	0.24	1.00	0.39	0.24	1.00	0.39	0.24
15%	1.00	0.41	0.31	1.00	0.41	0.31	1.00	0.41	0.31
20%	1.00	0.44	0.37	1.00	0.44	0.37	1.00	0.44	0.37
25%	1.00	0.46	0.42	1.00	0.46	0.42	1.00	0.46	0.42
30%	1.00	0.50	0.48	1.00	0.50	0.48	1.00	0.50	0.48
35%	1.00	0.54	0.54	1.00	0.54	0.54	1.00	0.53	0.53
40%	1.00	0.58	0.60	1.00	0.58	0.60	0.99	0.58	0.58
45%	1.00	0.63	0.66	1.00	0.63	0.66	0.99	0.63	0.65
50%	1.00	0.70	0.73	1.00	0.70	0.73	0.97	0.68	0.69
55%	0.99	0.77	0.80	1.00	0.77	0.80	0.97	0.75	0.76
60%	0.99	0.86	0.88	0.99	0.86	0.88	0.92	0.80	0.77
65%	0.97	0.96	0.94	0.96	0.96	0.94	0.89	0.89	0.83
70%	0.86	1.00	0.90	0.86	1.00	0.90	0.85	0.98	0.87
75%	0.72	1.00	0.79	0.72	1.00	0.79	0.72	1.00	0.79
80%	0.57	1.00	0.68	0.00	1.00	0.02	0.57	1.00	0.68
85%	0.43	1.00	0.57	0.00	1.00	0.02	0.00	0.00	0.00
90%	0.29	1.00	0.46	0.00	1.00	0.02	0.00	0.00	0.00
95%	0.00	0.00	0.00	0.00	1.00	0.02	0.00	0.00	0.00

Tabela C.2: Medidas de acurácia estimadas para a sequência `sintética2_0`.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	1.00	0.19	0.11	1.00	0.19	0.11	1.00	0.19	0.11
10%	1.00	0.20	0.15	1.00	0.20	0.15	1.00	0.20	0.15
15%	1.00	0.21	0.20	1.00	0.21	0.20	1.00	0.21	0.20
20%	1.00	0.22	0.23	1.00	0.22	0.23	1.00	0.22	0.23
25%	1.00	0.24	0.27	1.00	0.24	0.27	0.99	0.23	0.26
30%	1.00	0.25	0.30	1.00	0.25	0.30	0.98	0.25	0.29
35%	0.99	0.27	0.33	1.00	0.27	0.34	0.95	0.26	0.29
40%	0.99	0.29	0.37	1.00	0.29	0.37	0.92	0.27	0.30
45%	0.99	0.32	0.41	0.99	0.32	0.41	0.90	0.29	0.33
50%	0.99	0.35	0.45	0.99	0.35	0.45	0.89	0.32	0.37
55%	0.98	0.39	0.50	0.98	0.39	0.50	0.88	0.35	0.40
60%	0.98	0.43	0.55	0.98	0.43	0.55	0.86	0.38	0.44
65%	0.98	0.49	0.61	0.97	0.49	0.60	0.83	0.42	0.47
70%	0.97	0.57	0.67	0.94	0.56	0.65	0.81	0.48	0.51
75%	0.95	0.67	0.75	0.90	0.64	0.69	0.77	0.55	0.56
80%	0.93	0.82	0.84	0.83	0.74	0.73	0.73	0.65	0.61
85%	0.81	0.96	0.86	0.76	0.89	0.79	0.58	0.69	0.56
90%	0.56	1.00	0.72	0.56	1.00	0.72	0.44	0.78	0.53
95%	0.28	1.00	0.49	0.28	1.00	0.49	0.20	0.71	0.32

Tabela C.3: Medidas de acurácia estimadas para o gene BTU02285.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	1.00	0.16	0.10	1.00	0.16	0.10	1.00	0.16	0.10
10%	1.00	0.17	0.14	1.00	0.17	0.14	1.00	0.17	0.14
15%	1.00	0.18	0.18	1.00	0.18	0.18	1.00	0.18	0.18
20%	1.00	0.19	0.21	1.00	0.19	0.21	1.00	0.19	0.21
25%	1.00	0.21	0.25	1.00	0.21	0.25	1.00	0.21	0.25
30%	1.00	0.22	0.28	1.00	0.22	0.28	1.00	0.22	0.28
35%	1.00	0.24	0.31	1.00	0.24	0.31	1.00	0.24	0.31
40%	1.00	0.26	0.35	1.00	0.26	0.35	1.00	0.26	0.35
45%	1.00	0.28	0.38	0.99	0.28	0.38	1.00	0.28	0.39
50%	0.99	0.31	0.42	0.98	0.30	0.41	0.98	0.30	0.41
55%	0.99	0.34	0.47	0.96	0.33	0.44	0.96	0.33	0.44
60%	0.99	0.38	0.51	0.94	0.36	0.47	0.93	0.36	0.47
65%	0.98	0.43	0.57	0.91	0.40	0.51	0.91	0.40	0.50
70%	0.97	0.50	0.63	0.89	0.46	0.55	0.88	0.45	0.54
75%	0.96	0.60	0.71	0.86	0.53	0.60	0.85	0.53	0.59
80%	0.95	0.74	0.80	0.81	0.63	0.65	0.80	0.62	0.64
85%	0.87	0.90	0.86	0.71	0.73	0.67	0.71	0.73	0.67
90%	0.59	0.92	0.71	0.54	0.84	0.63	0.55	0.85	0.64
95%	0.32	1.00	0.54	0.28	0.88	0.46	0.31	0.96	0.51

Tabela C.4: Medidas de acurácia estimadas para o gene F56F11.4.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	0.98	0.28	0.08	0.98	0.28	0.09	0.98	0.28	0.07
10%	0.97	0.29	0.14	0.96	0.29	0.11	0.93	0.28	0.06
15%	0.94	0.30	0.16	0.93	0.30	0.13	0.89	0.28	0.06
20%	0.91	0.31	0.17	0.90	0.31	0.15	0.85	0.29	0.07
25%	0.88	0.32	0.18	0.87	0.32	0.17	0.81	0.29	0.08
30%	0.85	0.33	0.20	0.82	0.32	0.17	0.76	0.30	0.08
35%	0.82	0.34	0.21	0.78	0.33	0.17	0.73	0.31	0.10
40%	0.78	0.36	0.23	0.74	0.34	0.18	0.69	0.31	0.11
45%	0.74	0.37	0.23	0.70	0.35	0.18	0.66	0.33	0.13
50%	0.71	0.39	0.25	0.65	0.35	0.18	0.62	0.34	0.15
55%	0.66	0.40	0.26	0.60	0.37	0.19	0.58	0.35	0.16
60%	0.61	0.41	0.26	0.55	0.38	0.19	0.54	0.37	0.18
65%	0.56	0.44	0.28	0.51	0.40	0.20	0.50	0.39	0.19
70%	0.52	0.47	0.30	0.45	0.41	0.21	0.44	0.40	0.19
75%	0.45	0.49	0.29	0.40	0.44	0.22	0.39	0.42	0.20
80%	0.40	0.55	0.31	0.34	0.47	0.22	0.33	0.45	0.20
85%	0.34	0.61	0.32	0.28	0.52	0.23	0.27	0.49	0.21
90%	0.26	0.70	0.32	0.22	0.61	0.25	0.21	0.58	0.23
95%	0.15	0.84	0.29	0.15	0.80	0.27	0.15	0.84	0.29

Tabela C.5: Medidas de acurácia estimadas para o gene GGVITHG.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	1.00	0.24	0.12	1.00	0.24	0.12	1.00	0.24	0.12
10%	1.00	0.25	0.18	1.00	0.25	0.17	1.00	0.25	0.18
15%	1.00	0.27	0.23	0.99	0.27	0.21	1.00	0.27	0.23
20%	1.00	0.28	0.27	0.99	0.28	0.26	1.00	0.28	0.27
25%	1.00	0.30	0.31	0.99	0.30	0.30	1.00	0.30	0.31
30%	1.00	0.33	0.36	0.98	0.32	0.34	1.00	0.32	0.35
35%	1.00	0.35	0.40	0.98	0.34	0.38	0.99	0.35	0.39
40%	1.00	0.38	0.44	0.98	0.37	0.42	0.99	0.38	0.43
45%	1.00	0.41	0.49	0.97	0.40	0.46	0.99	0.41	0.48
50%	0.99	0.45	0.54	0.96	0.44	0.50	0.98	0.45	0.53
55%	0.99	0.50	0.59	0.94	0.47	0.53	0.98	0.50	0.58
60%	0.98	0.56	0.64	0.91	0.52	0.57	0.98	0.56	0.64
65%	0.97	0.63	0.71	0.88	0.57	0.60	0.96	0.62	0.69
70%	0.94	0.72	0.76	0.83	0.63	0.63	0.92	0.70	0.74
75%	0.91	0.83	0.83	0.76	0.70	0.64	0.87	0.79	0.77
80%	0.83	0.95	0.86	0.66	0.75	0.63	0.80	0.91	0.81
85%	0.65	0.98	0.76	0.54	0.81	0.59	0.63	0.96	0.73
90%	0.44	1.00	0.61	0.40	0.92	0.55	0.44	1.00	0.61
95%	0.22	1.00	0.42	0.22	1.00	0.42	0.22	1.00	0.42

Tabela C.6: Medidas de acurácia estimadas para o gene HSDAO.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	1.00	0.25	0.13	1.00	0.25	0.13	0.99	0.25	0.09
10%	1.00	0.26	0.19	1.00	0.26	0.19	0.96	0.25	0.12
15%	1.00	0.28	0.23	1.00	0.28	0.23	0.94	0.26	0.14
20%	0.95	0.28	0.21	0.99	0.29	0.27	0.93	0.28	0.18
25%	0.93	0.30	0.24	0.99	0.31	0.31	0.92	0.29	0.21
30%	0.92	0.31	0.26	0.98	0.33	0.35	0.90	0.31	0.25
35%	0.90	0.33	0.29	0.98	0.36	0.39	0.89	0.33	0.28
40%	0.89	0.35	0.33	0.97	0.38	0.42	0.86	0.34	0.30
45%	0.88	0.38	0.38	0.95	0.41	0.45	0.81	0.35	0.29
50%	0.88	0.42	0.42	0.91	0.43	0.46	0.79	0.38	0.33
55%	0.87	0.46	0.47	0.87	0.46	0.47	0.78	0.41	0.37
60%	0.84	0.50	0.51	0.82	0.49	0.48	0.77	0.46	0.42
65%	0.79	0.54	0.52	0.75	0.51	0.47	0.74	0.50	0.45
70%	0.71	0.56	0.50	0.68	0.54	0.46	0.71	0.56	0.50
75%	0.67	0.64	0.54	0.60	0.57	0.45	0.68	0.65	0.55
80%	0.63	0.75	0.61	0.52	0.62	0.45	0.60	0.72	0.56
85%	0.54	0.85	0.61	0.40	0.64	0.40	0.50	0.80	0.55
90%	0.39	0.93	0.54	0.34	0.81	0.45	0.33	0.79	0.44
95%	0.21	1.00	0.41	0.17	0.83	0.32	0.15	0.74	0.27

Tabela C.7: Medidas de acurácia estimadas para o gene HUMCS3.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	1.00	0.10	0.08	1.00	0.10	0.08	1.00	0.10	0.08
10%	1.00	0.11	0.11	1.00	0.11	0.11	1.00	0.11	0.11
15%	1.00	0.11	0.14	1.00	0.11	0.14	1.00	0.11	0.14
20%	1.00	0.12	0.16	1.00	0.12	0.16	1.00	0.12	0.16
25%	1.00	0.13	0.19	1.00	0.13	0.19	0.96	0.12	0.16
30%	1.00	0.14	0.21	1.00	0.14	0.21	0.90	0.12	0.14
35%	1.00	0.15	0.24	0.99	0.15	0.24	0.83	0.12	0.12
40%	1.00	0.16	0.27	0.99	0.16	0.26	0.77	0.12	0.11
45%	0.99	0.17	0.29	0.97	0.17	0.28	0.73	0.13	0.12
50%	0.96	0.19	0.30	0.95	0.18	0.30	0.71	0.14	0.14
55%	0.93	0.20	0.31	0.93	0.20	0.31	0.65	0.14	0.13
60%	0.89	0.22	0.33	0.87	0.21	0.32	0.59	0.14	0.13
65%	0.87	0.24	0.36	0.82	0.23	0.32	0.49	0.14	0.10
70%	0.85	0.27	0.39	0.77	0.25	0.33	0.42	0.13	0.08
75%	0.83	0.32	0.44	0.69	0.27	0.33	0.30	0.12	0.04
80%	0.78	0.38	0.48	0.58	0.28	0.31	0.17	0.08	-0.03
85%	0.55	0.36	0.37	0.44	0.29	0.27	0.06	0.04	-0.08
90%	0.25	0.24	0.16	0.31	0.30	0.23	0.00	0.00	-0.11
95%	0.00	0.00	-0.08	0.13	0.25	0.12	0.00	0.00	-0.07

Tabela C.8: Medidas de acurácia estimadas para o gene MMA2L2A.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	0.99	0.21	0.10	0.99	0.21	0.10	0.99	0.21	0.09
10%	0.99	0.23	0.15	0.98	0.22	0.14	0.97	0.22	0.13
15%	0.98	0.24	0.18	0.97	0.23	0.17	0.96	0.23	0.15
20%	0.97	0.25	0.21	0.95	0.24	0.19	0.94	0.24	0.18
25%	0.96	0.26	0.24	0.93	0.26	0.22	0.92	0.25	0.20
30%	0.94	0.28	0.27	0.92	0.27	0.24	0.89	0.26	0.22
35%	0.93	0.29	0.29	0.90	0.28	0.26	0.87	0.28	0.23
40%	0.91	0.31	0.31	0.87	0.30	0.28	0.84	0.29	0.25
45%	0.89	0.33	0.34	0.84	0.31	0.30	0.81	0.30	0.26
50%	0.86	0.35	0.36	0.82	0.33	0.31	0.78	0.32	0.28
55%	0.84	0.38	0.39	0.78	0.35	0.33	0.74	0.34	0.29
60%	0.81	0.41	0.41	0.75	0.38	0.35	0.70	0.35	0.30
65%	0.77	0.44	0.44	0.70	0.40	0.36	0.65	0.37	0.31
70%	0.73	0.48	0.46	0.65	0.43	0.37	0.60	0.40	0.32
75%	0.68	0.53	0.47	0.59	0.47	0.38	0.54	0.43	0.32
80%	0.61	0.58	0.47	0.52	0.50	0.38	0.47	0.46	0.32
85%	0.51	0.63	0.46	0.44	0.55	0.37	0.38	0.49	0.30
90%	0.39	0.69	0.43	0.33	0.60	0.35	0.28	0.52	0.28
95%	0.22	0.75	0.34	0.19	0.66	0.28	0.15	0.55	0.21

Tabela C.9: Medidas de acurácia estimadas para o conjunto A.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	0.99	0.22	0.11	0.99	0.22	0.10	0.99	0.22	0.09
10%	0.99	0.23	0.15	0.98	0.23	0.14	0.97	0.23	0.13
15%	0.98	0.24	0.19	0.97	0.24	0.17	0.96	0.24	0.16
20%	0.97	0.25	0.22	0.95	0.25	0.20	0.94	0.25	0.18
25%	0.96	0.27	0.25	0.94	0.26	0.22	0.92	0.26	0.20
30%	0.95	0.28	0.28	0.92	0.27	0.24	0.89	0.27	0.22
35%	0.93	0.30	0.30	0.90	0.29	0.26	0.87	0.28	0.24
40%	0.91	0.32	0.32	0.87	0.30	0.28	0.84	0.29	0.25
45%	0.89	0.34	0.35	0.85	0.32	0.30	0.81	0.31	0.27
50%	0.87	0.36	0.37	0.82	0.34	0.32	0.78	0.33	0.29
55%	0.85	0.39	0.40	0.78	0.36	0.34	0.74	0.34	0.30
60%	0.81	0.42	0.43	0.75	0.39	0.35	0.70	0.36	0.31
65%	0.78	0.46	0.45	0.70	0.41	0.37	0.66	0.39	0.32
70%	0.74	0.50	0.47	0.65	0.44	0.38	0.60	0.41	0.33
75%	0.68	0.54	0.48	0.59	0.48	0.39	0.54	0.44	0.33
80%	0.61	0.59	0.49	0.52	0.52	0.39	0.48	0.47	0.33
85%	0.52	0.65	0.47	0.44	0.56	0.37	0.39	0.50	0.32
90%	0.39	0.70	0.43	0.33	0.61	0.35	0.29	0.54	0.29
95%	0.22	0.77	0.34	0.19	0.68	0.28	0.16	0.57	0.22

Tabela C.10: Medidas de acurácia estimadas para o conjunto Am30.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	1.00	0.27	0.12	0.99	0.27	0.12	0.99	0.27	0.12
10%	0.99	0.28	0.18	0.99	0.28	0.17	0.98	0.28	0.17
15%	0.98	0.30	0.22	0.98	0.29	0.21	0.97	0.29	0.21
20%	0.98	0.31	0.26	0.97	0.31	0.24	0.96	0.31	0.24
25%	0.97	0.33	0.30	0.96	0.33	0.28	0.94	0.32	0.27
30%	0.96	0.35	0.33	0.94	0.34	0.31	0.93	0.34	0.29
35%	0.95	0.37	0.37	0.93	0.36	0.34	0.91	0.36	0.32
40%	0.94	0.40	0.40	0.91	0.38	0.36	0.88	0.38	0.34
45%	0.93	0.43	0.44	0.88	0.41	0.39	0.86	0.40	0.37
50%	0.91	0.46	0.47	0.86	0.43	0.41	0.83	0.42	0.39
55%	0.89	0.49	0.50	0.82	0.46	0.43	0.80	0.45	0.40
60%	0.85	0.53	0.52	0.78	0.48	0.44	0.76	0.47	0.41
65%	0.81	0.57	0.54	0.74	0.51	0.45	0.71	0.50	0.42
70%	0.76	0.61	0.55	0.68	0.55	0.45	0.66	0.53	0.42
75%	0.70	0.65	0.55	0.62	0.58	0.45	0.60	0.56	0.42
80%	0.62	0.70	0.54	0.54	0.62	0.44	0.52	0.60	0.42
85%	0.52	0.75	0.51	0.45	0.66	0.42	0.43	0.64	0.40
90%	0.40	0.81	0.47	0.34	0.72	0.39	0.32	0.69	0.36
95%	0.22	0.86	0.36	0.19	0.78	0.31	0.18	0.75	0.29

Tabela C.11: Medidas de acurácia estimadas para o conjunto Am100.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	0.99	0.25	0.12	0.99	0.25	0.10	0.99	0.25	0.11
10%	0.98	0.26	0.16	0.97	0.26	0.14	0.98	0.26	0.16
15%	0.97	0.27	0.20	0.95	0.27	0.17	0.96	0.27	0.19
20%	0.96	0.29	0.23	0.93	0.28	0.19	0.94	0.28	0.21
25%	0.95	0.30	0.26	0.91	0.29	0.21	0.92	0.30	0.23
30%	0.93	0.32	0.28	0.89	0.30	0.23	0.90	0.31	0.25
35%	0.92	0.33	0.30	0.87	0.32	0.25	0.87	0.32	0.26
40%	0.90	0.35	0.32	0.84	0.33	0.27	0.84	0.33	0.28
45%	0.87	0.37	0.34	0.81	0.35	0.28	0.81	0.35	0.29
50%	0.85	0.39	0.36	0.78	0.36	0.30	0.78	0.36	0.30
55%	0.82	0.41	0.38	0.75	0.38	0.31	0.75	0.38	0.31
60%	0.78	0.43	0.39	0.71	0.40	0.33	0.71	0.40	0.33
65%	0.75	0.46	0.41	0.67	0.43	0.34	0.66	0.43	0.34
70%	0.70	0.49	0.42	0.62	0.45	0.35	0.62	0.45	0.34
75%	0.64	0.53	0.42	0.56	0.48	0.35	0.56	0.48	0.35
80%	0.57	0.57	0.42	0.50	0.51	0.35	0.49	0.51	0.34
85%	0.49	0.62	0.42	0.42	0.56	0.34	0.41	0.54	0.33
90%	0.37	0.66	0.38	0.32	0.60	0.32	0.31	0.57	0.30
95%	0.22	0.72	0.31	0.19	0.65	0.26	0.17	0.61	0.23

Tabela C.12: Medidas de acurácia estimadas para o conjunto B.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	0.99	0.26	0.12	0.99	0.26	0.10	0.99	0.26	0.12
10%	0.99	0.27	0.17	0.97	0.27	0.14	0.98	0.27	0.16
15%	0.98	0.28	0.20	0.95	0.28	0.17	0.97	0.28	0.20
20%	0.96	0.30	0.23	0.93	0.29	0.20	0.95	0.29	0.22
25%	0.95	0.31	0.26	0.91	0.30	0.22	0.93	0.30	0.24
30%	0.93	0.32	0.28	0.89	0.31	0.24	0.91	0.32	0.26
35%	0.92	0.34	0.31	0.87	0.33	0.26	0.88	0.33	0.27
40%	0.90	0.36	0.33	0.84	0.34	0.27	0.85	0.34	0.29
45%	0.87	0.38	0.35	0.81	0.35	0.29	0.82	0.36	0.30
50%	0.85	0.40	0.37	0.78	0.37	0.30	0.79	0.37	0.31
55%	0.82	0.42	0.38	0.75	0.39	0.32	0.75	0.39	0.32
60%	0.79	0.44	0.39	0.71	0.41	0.33	0.71	0.41	0.33
65%	0.75	0.47	0.41	0.67	0.43	0.34	0.67	0.44	0.34
70%	0.70	0.50	0.42	0.62	0.46	0.35	0.62	0.46	0.35
75%	0.64	0.53	0.42	0.56	0.49	0.35	0.56	0.49	0.35
80%	0.57	0.58	0.42	0.50	0.52	0.35	0.50	0.52	0.35
85%	0.49	0.62	0.42	0.42	0.56	0.34	0.42	0.56	0.34
90%	0.37	0.66	0.38	0.32	0.60	0.32	0.31	0.59	0.30
95%	0.22	0.72	0.30	0.19	0.65	0.26	0.17	0.63	0.24

Tabela C.13: Medidas de acurácia estimadas para o conjunto Bm30.

Limiar	MMT			STFT (200)			STFT (400)		
	S_n	S_p	CC	S_n	S_p	CC	S_n	S_p	CC
05%	1.00	0.36	0.16	0.99	0.35	0.15	0.99	0.36	0.16
10%	0.99	0.37	0.22	0.98	0.37	0.20	0.99	0.37	0.23
15%	0.98	0.39	0.27	0.97	0.38	0.23	0.98	0.39	0.27
20%	0.97	0.40	0.30	0.95	0.39	0.26	0.97	0.40	0.30
25%	0.96	0.42	0.34	0.93	0.41	0.29	0.95	0.42	0.33
30%	0.95	0.44	0.37	0.91	0.43	0.31	0.93	0.44	0.36
35%	0.93	0.46	0.40	0.89	0.44	0.33	0.92	0.46	0.38
40%	0.91	0.49	0.42	0.87	0.46	0.35	0.90	0.48	0.40
45%	0.89	0.51	0.44	0.84	0.48	0.37	0.87	0.50	0.42
50%	0.87	0.54	0.47	0.81	0.51	0.39	0.84	0.53	0.44
55%	0.84	0.57	0.48	0.77	0.53	0.40	0.81	0.55	0.45
60%	0.81	0.60	0.50	0.73	0.56	0.42	0.78	0.58	0.47
65%	0.77	0.63	0.51	0.69	0.58	0.43	0.74	0.61	0.48
70%	0.71	0.66	0.51	0.64	0.61	0.43	0.69	0.65	0.48
75%	0.65	0.70	0.50	0.58	0.64	0.42	0.63	0.68	0.48
80%	0.58	0.74	0.49	0.51	0.68	0.42	0.56	0.72	0.47
85%	0.49	0.79	0.47	0.43	0.72	0.40	0.47	0.77	0.45
90%	0.36	0.82	0.41	0.32	0.77	0.36	0.35	0.81	0.40
95%	0.20	0.85	0.31	0.18	0.81	0.28	0.19	0.85	0.30

Tabela C.14: Medidas de acurácia estimadas para o conjunto Bm100.

Apêndice D

Painéis Apresentados

D.1 International Conference on Bioinformatics and Computational Biology 2004

Bioinformatics Tools for Assembling and Analysis of Chloroplast Genomes

Jesus P. Mena-Chalco¹, Henrique S. Alves², Helaine Carrer², Roberto M. Cesar-Jr¹

1 Depto. Ciência da Computação, IME-USP. Rua do Matão, 1010. São Paulo-SP 05508-090.

2 Depto. Ciências Biológicas, ESALQ-USP. Av. Pádua Dias, 11. Piracicaba-SP 13418-900.

Chloroplasts are organelles found only in plant and algae cells. They are responsible for photosynthesis and for the synthesis of key molecules required for the basic architecture and functioning of plant cells. These organelles have their own genetic machinery and together with the nucleus and mitochondrial genomes are responsible for cellular coordination activity. At the moment 29 higher plant plastid genomes (plastomes) have been sequenced (<http://ncbi.nlm.nih.gov/>). The plastome sequences are conserved among species but the genes arrangements are different for divergent plant groups. The knowledge of the nucleotide sequence of chloroplast genomes is important for evolution studies and for biotechnology applications. The chloroplast organelle being used as a model in this study was isolated from *Eucalyptus grandis*, an important economical tree for the production of paper and cellulose and in Brazil is located the main germoplasm collection of Eucalyptus outside Australia.

We have sequenced 3500 sequences from an Eucalyptus DNA library. These sequences represent so far, 50% of the total plastome sequence of *Eucalyptus grandis*. These sequences are stored through a special pipeline at the bioinformatics servers at URL <http://malariadb.ime.usp.br:8026/pipeline/>. Once this phase is accomplished, the next step is the search for similar sequences in other related organisms. Some tentative results towards this direction have been already obtained.

In this study, we apply digital signal processing (DSP) techniques [1, 2, 3] on the genomic data sequences in order to identify and compare DNA and protein sequences of *Eucalyptus grandis* to the other available higher plant plastomes. We have chosen different approaches to identify protein coding DNA regions and to compare protein sequences. In particular, traditional Fourier analysis and the wavelet transform will be evaluated [4, 5].

References

- [1] J. Chen, H. Li, K. Sun and B. Kim, "How will bioinformatics impact signal processing research", IEEE Signal Processing Magazine, November 2003.
- [2] X. Zhang et al. "Signal processing techniques in genomic engineering", Proceedings of the IEEE, Vol 90. Nro 12, December 2002.
- [3] D. Anastassion, "Genomic signal processing", IEEE Signal Proc. Mag., pp. 8 20, July 2001.
- [4] P. Lio, "Wavelets in bioinformatics and computational biology: state of art and perspectives". Bioinformatics Review, 19(1), 2003.
- [5] C. H. Trad, Q. Fang and I. Cosic. "Protein sequence comparison based on the wavelet transform approach", Protein engineering, 15(3):193-203, 2002.

D.2 X-Meeting 2005

Protein Coding Regions Identification through the Modified Morlet Transform

Jesús P. Mena-Chalco¹, Roberto M. Cesar-Jr¹

¹ Depto. Ciência da Computação, IME-USP. Rua do Matão, 1010. São Paulo-SP 05508-090.

An important topic in biological sequences analysis area is the protein coding regions identification. This identification allows the posterior research for meaning, description or biological categorization of the analyzed organism [1]. Currently, several methods combine pattern recognition with knowledge collected from training datasets of known genes or from comparison with genomic databases. Nonetheless, the accuracy of these methods is still far from satisfactory. New methods of DNA sequences processing and genes identification can be created through *search-by-content* such sequences [2]. The periodic pattern of DNA in protein coding regions, called three-base periodicity (TBP), has been considered proper of coding regions. This phenomenon was not observed for nonprotein coding. The digital signal processing techniques supply a strong basis for regions identification with TBP [2,3].

In this work we introduce a new method for protein coding regions identification with TBP, based on a wavelet transform, called Modified Morlet Transform (MMT), which does not need to be trained on sequences

databases. We use a fixed binary mapping rules to create four binary sequences. Where each one represents the positions of each nitrogenate base in DNA sequence. Next the MMT, with different scales is applied to all binary sequences. The module of each normalized coefficient is projected onto the position axis. Projection onto the scale axis reveal which scale carry more signal energy throughout the positions. The result of the projection position axis represents the protein coding region identifier. These projection coefficients correspond to regions with TBP. Thus, we use thresholding coefficients, based on both shrinking values and inflection points, to exclude positions where the associated energy is lower. At the moment, we consider arbitrary length region criterions for discarding possible very short protein coding regions identification. The performance of the proposed transform was examined by analyzing synthetic and real DNA sequences (RGRC2 and F56F11.4 genes of *O. sativa* and *C. elegans* organism, respectively). Preliminary results show that MMT is better than traditional methods by presenting greater sensitivity to TBP and discriminatory capability between protein coding regions.

References

- [1] A. W.-C. Liew, H. Yan, and M. Yang, Pattern recognition techniques for the emerging field of bioinformatics: A review, *Pattern Recognition* 38 (2005), no. 11, 2055-2073.
- [2] X. Zhang, F. Chen, Y. Zhang, S. C. Agner, M. Akay, Z. Lu, M. M. Y. Waye, and S. K. Tsui, Signal processing techniques in genomic engineering, *Proceedings of the IEEE* 90 (2002), no. 12, 1822-1833.
- [3] J. P. Mena-Chalco, H. S. Alves, H. Carrer and R. M. Cesar-Jr, Bioinformatics tools for assembling and analysis of chloroplast genomes, *International Conference on Bioinformatics and Computational Biology* (2004), Rio de Janeiro.

Referências Bibliográficas

- [1] V. Afreixo, P. J. S. G. Ferreira, and D. Santos, *Fourier analysis of symbolic data: A brief review*, Digital Signal Process. **14** (2004), no. 6, 523–530. [26](#), [27](#), [28](#)
- [2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular biology of the cell*, fourth ed., Garland Science, New York, 2002. [5](#), [6](#), [7](#), [8](#), [10](#), [17](#), [18](#), [51](#), [73](#)
- [3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *Basic local alignment search tool*, Journal of Molecular Biology **215** (1990), 403–410. [20](#)
- [4] D. Anastassiou, *Genomic signal processing*, IEEE Signal Processing Magazine **8** (2001), no. 4, 8–20. [xx](#), [2](#), [27](#), [28](#), [29](#), [36](#), [38](#), [39](#), [71](#)
- [5] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, *Characterizing long-range correlations in DNA sequences from wavelet analysis*, Physical Review Letters **74** (1995), no. 16-17, 3293–3296. [51](#)
- [6] A. Arneodo, Y. D’Aubenton-Carafa, B. Audit, E. Bacry, J.-F. Muzy, and C. Thermes, *What can we learn with wavelets about DNA sequences?*, Physica A **249** (1998), no. 1-4, 439–448. [41](#)
- [7] A. D. Baxevanis and B. F. Ouellette, *Bioinformatics: A practical guide to the analysis of genes and proteins*, second ed., John Wiley and Sons, New York, 2001. [1](#), [12](#), [13](#)
- [8] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K Peng, M. Simons, and H. E. Stanley, *Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis*, Physical Review E **51** (1995), no. 5, 5084–5091. [26](#), [27](#)
- [9] M. Burset and R. Guigó, *Evaluation of gene structure prediction programs*, Genomics **34** (1996), no. 3, 353–367. [2](#), [41](#), [42](#), [73](#), [76](#)
- [10] Y. T. Chan, *Wavelet basics*, Kluwer Academic Publishers, Boston, 1995. [24](#), [26](#), [43](#), [44](#)
- [11] V. R. Chechetkin and A. Y. Turygin, *Size-dependence of three-periodicity and long-range correlations in DNA sequences*, Physics Letters A **199** (1995), no. 1-2, 75–80. [30](#)

- [12] F. Chen and Y.-T. Zhang, *A DNA structure-based bionic wavelet transform and its applications to DNA sequence analysis*, Applied Bionics and Biomechanics **1** (2003), no. 1, 3–9. [27](#), [40](#)
- [13] J. Chen, H. Li, K. Sun, and B. Kim, *How will bioinformatics impact signal processing research?*, IEEE Signal Processing Magazine **20** (2003), no. 6, 16–26. [2](#), [34](#)
- [14] C. K. Chui, *An introduction to wavelets*, Wavelet Analysis and its Applications, vol. 1, Academic Press, Inc., 1992. [44](#)
- [15] L. F. Costa and R. M. Cesar Jr., *Shape analysis and classification: Theory and practice*, CRC Press, Inc., Boca Raton, FL, USA, 2000. [23](#), [25](#), [45](#), [51](#), [52](#)
- [16] R.V. Davuluri, I. Grosse, and M.Q. Zhang, *Computational identification of promoters and first exons in the human genome*, Nature Genetics **29** (2001), no. 4, 412–7. [72](#)
- [17] A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg, *Alignment of whole genomes*, Nucleic Acids Research **27** (1999), no. 11, 2369–2376. [21](#)
- [18] G. Dodin, P. Vandergheynst, P. Levoir, C. Cordier, and L. Marcourt, *Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences*, Journal of Theoretical Biology **206** (2000), no. 3, 323–326. [41](#)
- [19] A. M. Durham, A. Y. Kashiwabara, F. T. G. Matsunaga, P. H. Ahagon, F. Rainone, L. Varuzza, and A. Gruber, *EGene: a configurable pipeline generation system for automated sequence analysis*, Bioinformatics **21** (2005), no. 12, 2812–2813. [20](#)
- [20] S. T. Eskesen, F. N. Eskesen, B. Kinghorn, and A. Ruvinsky, *Periodicity of DNA in exons*, Journal Molecular Biology **5** (2004), no. 12, 1–11. [11](#), [12](#), [32](#)
- [21] E. W. Myers et al., *A whole-genome assembly of Drosophila*, Science **287** (2000), no. 5461, 2196–2204. [10](#), [15](#)
- [22] B. Ewing and P. Green, *Base-calling of automated sequencer traces using Phred. II. Error probabilities*, Genome Research **8** (1998), no. 3, 186–194. [16](#)
- [23] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, *Base-calling of automated sequencer traces using Phred. I. Accuracy assessment*, Genome Research **8** (1998), no. 3, 175–185. [16](#)
- [24] R. Guigó, *DNA composition, codon usage and exon prediction*, pp. 53–80, Academic Press, 1999. [1](#), [73](#)
- [25] G. Gutierrez, J. L. Oliver, and A. Marin, *On the origin of the periodicity of three in protein coding DNA sequences*, Journal of Theoretical Biology **167** (1994), no. 4, 413–414. [11](#), [12](#)
- [26] D. Haussler, *Computational genefinding*, Trends in Biochemical Sciences **16** (1998), no. 1, 12–15. [2](#), [13](#), [14](#)

- [27] J. D. Hawkins, *A survey on intron and exon lengths*, Nucleic Acids Research **16** (1988), no. 21, 9893–9908. [14](#), [72](#)
- [28] H. Herzel, E. N. Trifonov, O. Weiss, and I. Große, *Interpreting correlations in biosequences*, Physica A **249** (1998), 449–459. [51](#)
- [29] S. Hosid, E. N. Trifonov, and A. Bolshoy, *Sequence periodicity of Escherichia coli is concentrated in intergenic regions*, BMC Molecular Biology **5** (2004), no. 14, 1–7. [12](#)
- [30] T. H. Jukes and S. Osawa, *Evolutionary changes in the genetic code*, Comp Biochem Physiol B **106** (1993), no. 3, 489–494. [9](#)
- [31] E. V. Korotkov and D. A. Phoenix, *Latent periodicity of DNA sequences of many genes*, Pacific Symposium on Biocomputing **2** (1997), 222–231. [12](#)
- [32] W. Li, *The study of correlation structures of DNA sequences: A critical review*, Computers & Chemistry **21** (1997), no. 4, 257–271. [12](#)
- [33] A. W.-C. Liew, H. Yan, and M. Yang, *Pattern recognition techniques for the emerging field of bioinformatics: A review*, Pattern Recognition **38** (2005), no. 11, 2055–2073. [12](#), [13](#), [14](#), [40](#)
- [34] P. Liò, *Wavelets in bioinformatics and computational biology: State of art and perspectives*, Bioinformatics **19** (2003), no. 1, 2–9. [2](#), [34](#)
- [35] I. Lopez-Villasenor, M. V. Jose, and J. Sanchez, *Three-base periodicity patterns and self-similarity in whole bacterial chromosomes*, Biochemical and Biophysical Research Communications **325** (2004), no. 2, 467–478. [11](#)
- [36] S. Mallat, *A wavelet tour of signal processing*, Academic Press, San Diego, 1998. [72](#)
- [37] C. Mathe, M. F. Sagot, T. Schiex, and P. Rouze, *Current methods of gene prediction, their strengths and weakness*, Nucleic Acids Research **30** (2002), no. 19, 4103–4117. [12](#), [42](#)
- [38] N. Morrison, *Introduction to Fourier analysis*, John Wiley & Sons, Brisbane, 1994. [16](#), [23](#), [24](#), [25](#), [39](#)
- [39] D. W. Mount, *Bioinformatics: Sequence and genome analysis*, second ed., Cold Spring Harbor Press, Cold Spring Harbor, NY, 2004. [1](#), [12](#), [13](#), [16](#), [17](#), [42](#), [71](#)
- [40] J. Ning, C. N. Moore, and J. C. Nelson, *Preliminary wavelet analysis of genomic sequences*, IEEE Computer Society Bioinformatics Conference, 2003, pp. 509–510. [27](#), [40](#)
- [41] V. K. Okura, *Bioinformática de projetos genoma de bactérias*, Master’s thesis, Universidade Estadual de Campinas, Departamento de Teoria da Computação, 2002. [15](#), [18](#)
- [42] S. Osawa, T. H. Jukes, K. Watanabe, and A. Muto, *Recent evidence for evolution of the genetic code*, Microbiol Reviews **56** (1992), no. 1, 229–264. [9](#)

- [43] B. Pierre, S. Brunak, Y. Chauvin, J. Engelbrecht, and A. Krogh, *Periodic sequence patterns in human exons*, Proc Int Conf Intell Syst Mol Biol **3** (1995), no. 3, 30–38. [11](#)
- [44] O. Rioul and M. Vetterli, *Wavelet and signal processing*, IEEE Signal Processing Magazine **8** (1991), no. 4, 14–38. [23](#), [24](#), [25](#), [26](#), [44](#)
- [45] S. Rogic, A. K. Mackworth, and B. F. Ouellette, *Evaluation of gene-finding programs on mammalian sequences*, Genome Research **11** (2001), no. 5, 817–832. [16](#), [17](#), [18](#), [42](#), [73](#), [76](#)
- [46] M. Shah, S. Passovets, D. Kim, K. Ellrott, L. Wang, I. Vokler, P. LoCascio, D. Xu, and Y. Xu, *A computational pipeline for protein structure prediction and analysis at genome scale*, Bioinformatics **19** (2003), no. 15, 1985–1996. [15](#)
- [47] B. D. Silverman and R. Linsker, *A measure of DNA periodicity*, Journal of Theoretical Biology **118** (1986), no. 3, 295–300. [11](#), [32](#)
- [48] L. Stein, *Genome annotation: From sequence to biology*, Nature Reviews Genetics **2** (2001), no. 7, 493–503. [18](#), [72](#)
- [49] D. S. Stoffer, D. E. Tyler, and A. J. McDougall, *Spectral analysis for categorical time series: Scaling and the spectral envelope*, Biometrika **80** (1993), no. 3, 611–622. [27](#)
- [50] D. Sussillo, A. Kundaje, and D. Anastassiou, *Spectrogram analysis of genomes*, EURASIP Journal of Applied Signal Processing, 2004, pp. 29–42. [41](#)
- [51] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, *Prediction of probable genes by Fourier analysis of genomic sequences*, Bioinformatics **13** (1997), no. 3, 263–270. [2](#), [28](#), [34](#), [39](#)
- [52] C. H. Trad, Q. Fang, and I. Cosic, *Protein sequence comparison based on the wavelet transform approach*, Protein Engineering **15** (2002), no. 3, 193–203. [41](#)
- [53] E. N. Trifonov and J. L. Sussman, *The pitch of chromatin DNA is reflected in its nucleotide sequence*, Proc Natl Acad Sci USA **77** (1980), no. 7, 3816–3820. [11](#)
- [54] E.N. Trifonov, *Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences*, Journal of Molecular Biology **194** (1987), no. 4, 643–652. [12](#)
- [55] ———, *3-, 10.5-, 200- and 400-base periodicities in genome sequences*, Physica A: Statistical And Theoretical Physics **249** (1998), no. 1-4, 511–516. [12](#)
- [56] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, *Periodicity in DNA coding sequences: Implications in gene evolution*, Journal of Theoretical Biology **151** (1991), no. 3, 323–331. [11](#), [27](#)

- [57] P. P. Vaidyanathan and B. Yoon, *The role of signal-processing concepts in genomics and proteomics*, Journal of the Franklin Institute **341** (2004), no. 1-2, 111–135. [2](#), [39](#), [71](#)
- [58] B. Vidaković and P. Müller, *Wavelets for kids: A tutorial introduction*, Discussion Paper 94-13, Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina, U.S.A., 1994. [52](#)
- [59] W. Wang and D. H. Johnson, *Computing linear transforms of symbolic signals*, IEEE Transaction on Signal Processing **50** (2002), no. 3, 628–634. [26](#), [27](#)
- [60] M. V. Wickerhauser, *Adapted wavelet analysis from theory to software*, A. K. Peters, Wellesley, Massachusetts, 1994. [72](#)
- [61] M. Q. Zhang, *Computational prediction of eukaryotic protein-coding genes*, Nature Reviews Genetics **3** (2002), no. 9, 698–709. [1](#), [2](#), [8](#), [14](#), [57](#), [72](#)
- [62] X. Zhang, F. Chen, Y. Zhang, S. C. Agner, M. Akay, Z. Lu, M. M. Y. Waye, and S. K. Tsui, *Signal processing techniques in genomic engineering*, Proceedings of the IEEE **90** (2002), no. 12, 1822–1833. [2](#), [34](#)
- [63] J. Zhao, X. W. Yang, J. P. Li, and Y. Y. Tang, *DNA sequences classification based on wavelet packet analysis*, Proc. Wavelet Analysis and Its Applications, 2nd Int. Conf. (2001), 424–429. [34](#)
- [64] V. B. Zhurkin, *Periodicity in DNA primary structure is defined by secondary structure of the coded protein*, Nucleic Acid Research **9** (1981), no. 8, 1963–1971. [11](#)

Índice Remissivo

- ácido
  - amino, [6–7](#)
  - desoxirribonucleico, [5](#)
  - nucleico, [5](#)
  - ribonucleico, [6](#)
- éxons, [11](#)
- ADN, *veja* ácido desoxirribonucleico
- ARN, *veja* ácido ribonucleico
- bases
  - identificação, [16](#)
  - nitrogenadas, [5](#), [26](#)
- biologia molecular
  - dogma, [8](#), [10](#)
  - fundamentos, [5](#)
- busca no DNA
  - por conteúdo, [2](#), [13](#)
  - por sítios, [13](#)
- código genético, [9](#)
- códon, [9](#)
- CDS, *veja* região codificante
- coeficiente de correlação, [42](#)
- convolução, [25](#)
- DFT, *veja* transformada discreta de Fourier
- DNA, *veja* ácido desoxirribonucleico
- DSP, *veja* processamento digital de sinais
- escalograma, [44](#)
- especificidade, [42](#)
- espectro
  - frequência, [24](#), [28](#)
- espectrograma, [26](#)
- filtros digitais, [38](#)
- Fourier
  - análise, [23](#)
  - métodos para a identificação de CDSs, [34](#)
  - transformada, *veja* transformada de Fourier
- genes
  - constituição, [7](#)
  - descrição, [18](#)
  - expressão, [2](#), [10](#)
  - identificação, [1](#), [12](#), [18](#), [23](#)
- genoma
  - anotação, [15](#), [18](#)
  - diferença, [17](#)
  - fragmentação, [10](#)
  - mapa, [18](#)
  - montagem, [11](#), [17](#)
  - projetos, [1](#), [10](#)
  - seqüenciamento, [1](#), [10](#), [15](#), [17](#)
    - shotgun, [11](#)
    - shotgun hirárquico, [11](#)
- histograma, [58](#), [77](#)

- íntrons, 11
- leituras
  - edição, 18
  - identificação, 16
- limiarização, 52
- mapeamento de nucleotídeos, 26
  - baseado em critérios de otimização, 27
  - fixo, 26, 49
- medidas de acurácia, 41, 54, 85
- MMT, *veja* transformada modificada de Morlet
- nucleotídeos, 5
- periodicidade
  - análise, 32
  - latente, 12
  - região codificante, 2, 11, 30, 32, 53
  - região inter-gênica, 12
  - seqüências reais, 34
  - seqüências sintéticas, 33
- pipeline* bioinformático
  - definição, 15
  - desenvolvido, 2, 19
  - unidades funcionais, 15
- processamento digital de sinais, 2, 26
- projeção, 51
- proteínas, 9
- região
  - codificante, 9, 11, 30
  - inter-gênica, 8, 11
  - não-codificante, 8
- resolução
  - tempo-freqüência, 26
- RNA, *veja* ácido ribonucléico
- ruído sintético, 30
- sensibilidade, 42
- seqüência
  - binária, 26, 28
  - comprimento, 5, 28
  - numérica, 26
  - real, 34
  - representação, 28
  - sintética, 30, 32, 33
- STFT, *veja* transformada de Fourier de tempo reduzido
- TBP, *veja* periodicidade região codificante
- transformada
  - Fourier, 23, 30
    - contínua, 24
    - discreta, 24
    - tempo reduzido, 25, 53, 71
  - modificada de Morlet, 2, 45, 53, 71
  - aplicação, 50
- Wavelet
  - análise, 43
  - Morlet, 44