Universidade Federal Fluminense

Análise da informação mútua em seqüências de DNA homólogas

Dissertação submetida para a obtenção do título de

Mestre em Computação

ao Programa de Pós-Graduação

em Computação da UFF

por

Luciana de Souza Pessôa

Outubro 2004

Análise da informação mútua em seqüências de DNA homólogas

Luciana de Souza Pessôa

Dissertação de Mestrado submetida ao Curso de Pós-Graduação em Ciência da Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre em Ciência da Computação

Aprovada em outubro de 2004 por:

Prof(a). Helena Cristina da Gama Leitão / UFF.

Prof. Jorge Stolfi / UNICAMP

Prof. Marcos Craizer / PUC RIO

Prof(a). Aura Conci / UFF

Prof. José Ricardo de Almeida Torreão / UFF

NITERÓI, RJ - BRASIL

Em memória do meu pai Manoel Alves Pessôa

Agradecimentos

O primeiro agradecimento é a Deus, por mais esta oportunidade de evolução intelectual.

Agradeço à Prof. Helena e ao Prof. Stolfi pela orientação e por me ensinarem a tornar reais os planos mais complexos.

Em especial, agradeço à Prof. Helena, pela amizade e companheirismo, principalmente nos momentos mais difíceis.

À minha família, que sempre me incentivou, apoiou, patrocinou e soube compreender as minhas ausências.

Ao meu marido, Rafael, por todo amor que alimentava a paciência e a confiança de que em breve teríamos dias mais tranquilos.

Aos amigos Ivairton Santos, Renata Cápua, Adriana Bechara por compartilharem os grandes momentos do mestrado, desde os mais alegres aos mais desesperadores.

Ao Stênio Soares, por toda a solidariedade destas últimas semanas que antecederam a defesa. Obrigada também pelas sugestões sobre a dissertação.

Ao Super Jacques, nosso amigo que sempre nos salva dos perigos. Especialmente, os do laboratório.

Obrigada a todos os professores, em especial ao Prof. Alexandre Plastino e à Prof. Anna Dolejsi que muito me incentivaram sobre a decisão de ingressar no

mestrado.

À Geiza, ao Eduardo, à Aline, ao Alexandre... A todos os amigos do laboratório e das salas de estudo pelas quais passei.

Às funcionárias Isabela e Ângela, pela atenção que dedicam a todos os alunos.

À CAPES por ter financiado este trabalho.

Resumo da Dissertação apresentada à UFF como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação (M.Sc.)

Análise da informação mútua em seqüências de DNA homólogas

Luciana de Souza Pessôa

Outubro/2004

Orientador: Helena Cristina da Gama Leitão
Co-orientador: Jorge Stolfi
Programa de Pós-Graduação em Computação

Neste trabalho, descrevemos um método para estimar a informação mútua de sequências de DNA homólogas, ou seja, a quantidade de informação contida em uma seqüência de DNA que pode ser utilizada para identificar trechos de DNA que originaram de um mesmo ancestral. Para isso utilizamos técnicas de processamento de sinais, especialmente análise espectral, filtragem de sinais e teoria da informação. Analisando a quantidade de informação mútua podemos estimar a probabilidade de falsos positivos — trechos que não são homólogos à seqüência, mas que por coincidência são tão similares à mesma quanto às homólogas.

Em 1999, Leitão e Stolfi desenvolveram um algoritmo eficiente para reconstrução de objetos cerâmicos fragmentados, usando a técnica de comparação multi-escala de seqüências. Esta técnica poderia ser aplicada também ao problema de localização de trechos similares em um banco de bio-seqüências, que é um problema fundamental na identificação de genes homólogos e na montagem de genomas a partir de fragmentos seqüenciados. A viabilidade da comparação multi-escala para este problema depende da hipótese de que, mesmo nas escalas mais grosseiras utilizadas, uma cadeia de DNA ainda contém informação mútua suficiente para eliminar uma fração significativa de falsos positivos. Verificamos esta hipótese utilizando o método aqui descrito.

Abstract of Dissertation presented to UFF as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

Mutual information analysis of homologous DNA sequences

Luciana de Souza Pessôa

October/2004

Advisor: Helena Cristina da Gama Leitão
Co-advisor: Jorge Stolfi
Department: Computer Science

In this dissertation, we describe a method for estimating the amount of mutual information of homologous DNA sequences, i.e., the information contained in a DNA sequence that can be used to identify homologous blocks that have a same ancestor. For that purpose, we use signal processing techniques, especially spectral analysis, signal filtering, and information theory. The analysis of the mutual information content allows us to estimate the probability of false positives — strings that are not homologous to the given sequence, but are just as similar to it as the homologous ones.

In 1999, Leitão and Stolfi developed an efficient algorithm for the reconstruction of fragmented seramic objects, using the technique of multiscale sequence comparison. This technique may be applicable to the problem of finding similar strings in a bio-sequence database, which is a fundamental problem for the identification of homologous genes and for the assembly of genomes from sequenced fragments. The viability of multiscale comparison for this problem relies on the hypothesis that, even in the coarsest scales used, a DNA sequence still contains enough mutual information to eliminate a significative fraction of false positives. We verify this hipothesis by the method described here.

Palavras-chave

1. Análise de seqüências de DNA
2. Teoria da informação
3. Análise de Fourier
4. Seqüências homólogas

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A informação necessária para reproduzir e manter um organismo está codificada em moléculas de ácidos nucléicos (DNA ou RNA), e pode ser representada abstratamente por uma seqüência definida sobre um alfabeto de quatro letras. Aprofundar o entendimento sobre como a informação é armazenada nestas *bio-seqüências* pode levar a métodos mais eficientes de comparação de seqüências genômicas.

Neste trabalho, analisamos a quantidade de informação contida em uma seqüência de DNA que pode ser utilizada para identificar seqüências homólogas à mesma. Para isso utilizamos técnicas de processamento de sinais, especialmente análise espectral [2], filtragem de sinais[14] e teoria da informação[17]. Os programas utilizados nesta análise foram desenvolvidos em linguagem C utilizando o compilador GNU (gcc 3.2.2) e o pacote FFTW (Fastest Fourier Transform in the West) [9].

1.1 Motivação

1.1.1 Comparação de bio-seqüências

Segundo Setubal e Meidanis[24], a comparação de bio-seqüências é a operação primitiva mais importante em biologia computacional, servindo como base para muitas outras manipulações mais complexas. Através da comparação é possível

buscar trechos de seqüências semelhantes entre várias seqüências contidas numa base de dados de DNA ou proteínas, obedecendo a um dado critério de similaridade.

A busca por similaridade é uma etapa necessária na reconstrução de genomas a partir de dados brutos gerados pelo equipamento de sequenciamento de DNA. A busca por similaridade é também essencial para a identificação de trechos homólogos de DNA — ou seja, trechos que descendem de uma mesma cadeia ancestral; e a principal ferramenta para identificação da função e da evolução dos genes [3].

A simplicidade do conceito de comparação oculta a complexidade dos algoritmos e estruturas de dados envolvidos em tal operação, principalmente se estamos tratando da comparação de seqüências em grandes volumes de dados. O custo de comparar, por exemplo, duas seqüências de comprimento N por programação dinâmica é $O(N^2)$ [6].

Para contornar esta situação, novos métodos baseados em heurísticas foram criados. Os mais conhecidos são o BLAST(Basic Local Alignment Search Tool)[1, 13, 24] e o FAST[13, 22, 24]. Entretanto, embora tenham resultados satisfatórios e sejam muito utilizados, não esgotam os esforços para a busca de soluções ainda mais eficientes.

1.1.2 Comparação multi-escala de bio-seqüências

Uma idéia inovadora para um método de comparação de seqüências seria a utilização da técnica de múltiplas escalas que se basearia na utilização de versões mais simplificadas das seqüências genômicas a serem comparadas.

Em [20], Leitão e Stolfi utilizaram esta técnica com sucesso no problema de reconstrução de objetos fragmentados, onde o objetivo era encontrar os pares de fragmentos que eram adjacentes nos objetos originais. Deste modo, foi possível reduzir o custo assintótico da comparação de um grande número de contornos de fragmentos. Apesar de aplicada a contornos de fragmentos de cerâmica, a técnica pode ser aplicada a todo tipo de seqüências, inclusive os sinais genômicos considerados neste trabalho.

Um método de comparação de seqüências que utilizasse múltiplas escalas na busca por seqüências similares deveria, inicialmente, reduzir a quantidade de detalhes das seqüências em vários níveis. Em seguida, iniciaria a comparação utilizando as seqüências representadas com o menor nível de detalhes, assim, apesar de todas as seqüências da base de dados serem comparadas com a seqüência alvo, isto seria feito a um custo mais reduzido do que se estivéssemos utilizando todas as seqüências representadas com todos os detalhes. Naturalmente, nesta etapa seriam encontradas muitas seqüências supostamente semelhantes que, por falta de detalhes, pareceriam ser tão similares quanto as corretas. Em seguida, a busca seria feita sobre uma versão mais detalhada das seqüências, porém avaliando apenas as seqüências que resultaram do processamento anterior. O procedimento se repetiria para versões cada vez mais detalhadas e sobre um conjunto com número de seqüências cada vez mais reduzido. O fato de compararmos seqüências com mais detalhes implica no aumento do custo computacional desta operação, mas por outro lado, se torna possível discriminar melhor os pares correspondentes dos não-correspondentes, e eliminar estes últimos. Desta forma, por ocasião da última iteração estaríamos usando todos os detalhes disponíveis nas seqüências originais; mas teríamos um conjunto pequeno de seqüências a comparar, boa parte dos quais apresentariam uma similaridade muito alta.

1.1.3 Informação mútua em trechos homólogos

A viabilidade da comparação multi-escala depende da hipótese de que uma cadeia de DNA de certo comprimento, apesar das mutações sofridas pela mesma, ainda contém informação suficiente para eliminar quase todos os trechos do banco de dados que não são homólogos à mesma; e que isto ainda continua valendo, em parte, para as versões simplificadas das cadeias. O objetivo deste trabalho é justamente analisar a quantidade de *informação mútua* em trechos homólogos de DNA, ou seja, a informação contida num trecho que pode ser usada para identificar outros trechos homólogos ao mesmo.

1.2 Trabalhos relacionados

O cálculo da quantidade de informação de seqüências de DNA é dificultado pela correlação entre os nucleotídeos que as compõem, ou seja, pelo fato de que a ocorrência de determinado nucleotídeo em determinada posição da seqüência depende dos nucleotídeos presentes em outras posições. Devido a essas correlações, que são difíceis de medir, a quantidade de informação contida numa seqüência de nucleotídeos não é simplesmente a soma das contribuições individuais de cada nucleotídeo.

A correlação entre os nucleotídeos foi objeto de várias pesquisas visando medir a quantidade de informação contida numa seqüência de DNA. Entretanto, nenhuma dessas pesquisas procurou medir a informação mútua entre seqüências homólogas, que é o tema deste trabalho. Muitas delas tinham por objetivo derradeiro a identificação de regiões codificadoras de proteínas, ou a reconstrução de bio-seqüências.

Por meio de análises estatísticas, Luo[21] demonstra duas propriedades importantes da correlação entre os nucleotídeos: (1) a correlação é predominantemente de curta distância, isto é, a probabilidade da ocorrência de determinado nucleotídeo em uma dada posição depende principalmente dos nucleotídeos próximos, distantes em uma ou duas posições; e (2) um aumento da correlação entre nucleotídeos adjacentes proporcional à evolução do organismo.

Landini [16] e Voss [30] estudaram o espectro de potência de seqüências de DNA, e observaram que o mesmo possui a forma característica de processos fractais [23], implicando a existência de correlações significativas mesmo entre bases bastante separadas entre si.

Vieira [8] também verificou uma correlação de curta distância nas seqüências de DNA através de análise espectral e estatística, e mostrou que o comportamento fractal nem sempre prevalece na seqüência genômica inteira.

Grosse e outros [12] motivaram-se no problema de reconhecimento de regiões codificadoras de proteínas para investigar a existência de padrões estatísticos que fossem independentes da espécie dos organismos estudados e diferentes nas regiões

codificadoras e não-codificadoras do DNA. Para isso, os autores estudaram a função *informação mútua* I(k) entre bases próximas, que quantifica a informação que se obtém de um *nucleotídeo* Y a partir da identificação de um *nucleotídeo* X que está a uma distância k de Y. Assim, descobriram que a informação mútua de regiões não-codificadoras é muito baixa, enquanto que nas regiões codificadoras $I(k)$ varia de acordo com a distância k. Se k é múltiplo de 3 então $I(k)$ é maior que o obtido em caso contrário. A partir desta observação Grosse e os demais autores definiram *informação mútua média* e observaram que a distribuição desta informação é bastante diferente para regiões codificadoras e não-codificadoras do DNA.

Otu e Sayood [25] utilizaram o conceito de informação mútua média definido por Grosse e outros [12] para resolver o problema de reconstrução de bio-seqüências [24]. Considerando que fragmentos originados nas mesmas regiões da seqüência alvo têm perfis similares de informação mútua média, o método proposto reduziu o custo computacional da comparação dos fragmentos utilizando este critério para agrupá-los. Deste modo, em cada iteração, todos os agrupamentos resultam em trechos reconstruídos da seqüência alvo. Estes trechos formam uma nova coleção de fragmentos que terão seus perfis de informação mútua média novamente avaliados, gerando novos agrupamentos. O processo encerra retornando a seqüência alvo reconstruída.

1.3 Visão geral do método proposto

O método que propomos para análise da informação mútua entre cadeias de DNA homólogas utiliza técnicas de processamento de sinais, como análise espectral e filtragem de sinais. Para isso, precisamos codificar as seqüências genômicas (simbólicas) em seqüências numéricas.

Em seguida, aplicamos a transformada de Fourier para representar o sinal numérico no domínio das freqüências e analisá-lo através de espectros de potência. Assim, o cálculo da quantidade de informação mútua foi simplificado superando os obstáculos impostos pela correlação entre os nucleotídeos.

Para confirmar a viabilidade do método de comparação multi-escala, aplicamos

a mesma análise às cadeias de DNA resultantes de um processo de filtragem e simplificação. O objetivo é verificar se as seqüências simplificadas ainda contêm uma quantidade de informação suficiente para identificar seqüências homólogas a ela.

1.4 Estrutura da dissertação

O restante deste trabalho está assim dividido: no capítulos 2,3 e 4 revisamos os principais conceitos em biologia molecular, biologia computacional e teoria da informação. O capítulo 5 é dedicado à revisão dos aspectos fundamentais da análise de Fourier. No capítulo 6, apresentamos o esquema de codificação que utilizamos para transformar as seqüências de DNA (simbólicas) em seqüências numéricas. A aplicação de técnicas de processamento de sinais sobre o sinal genômico é vista no capítulo 7, onde também avaliamos algumas características dos coeficientes de Fourier que simplificam o cálculo da quantidade de informação mútua. O método utilizado para determinação da quantidade da informação mútua das seqüências genômicas é apresentado no capítulo 8. A aplicação do método proposto é vista no capítulo 9, onde também apresentamos e analisamos os resultados obtidos. No capítulo 10, aplicamos o mesmo método às versões simplificadas das cadeias que seriam utilizadas no método de comparação multi-escala. Finalmente, no capítulo 11, apresentamos nossas conclusões e fazemos sugestões de trabalhos futuros.

Capítulo 2

Elementos de biologia molecular

Neste capítulo, fazemos uma revisão de alguns conceitos de biologia molecular essenciais para o entendimento dos capítulos subseqüentes.

2.1 Ácidos nucléicos

A informação necessária para a manutenção da vida dos organismos está armazenada nos *ácidos nucléicos*, macromoléculas que controlam a síntese de proteínas e, portanto, os principais processos químicos da vida. É também através da duplicação dos ácidos nucléicos que este "conhecimento" é transmitido para as gerações seguintes. Ácidos nucléicos estão presentes em todos organismos vivos, incluindo bactérias e vírus. (Alguns agentes infecciosos recentemente descobertos, os *príons*, são constituídos apenas de proteína, sem ácidos nucléicos; mas é discutível se eles podem ser considerados "organismos vivos".)

Cada molécula de ácido nucléico é uma longa cadeia (um *polímero*) de pequenas moléculas, os *nucleotídeos*. Os nucleotídeos são formados por três resíduos ou radicais (moléculas incompletas), respectivamente de ácido fosfórico (ou fosfato), de uma pentose (um açúcar simples com cinco carbonos), e de uma base orgânica nitrogenada.

Distinguem-se dois tipos de ácidos nucléicos: o *ácido ribonucléico* (ARN, ou RNA em inglês), e o *ácido desoxiribonucléico* (ADN, ou DNA). O radical fosfato é identico no DNA e no RNA, mas as pentoses são diversas: no RNA, encontramos a ribose e no DNA, encontramos a desoxirribose. As bases nitrogenadas do DNA podem ser a citosina (C), timina (T), adenina (A) ou guanina (G). As duas primeiras possuem o anel carbono-nitrogênio da purina, e as duas últimas possuem os dois anéis fundidos da pirimidina. As bases adenosina, guanina e citosina também ocorrem no RNA, mas a timina é substituída pela uracila (U). Veja a tabela 2.1.

Grupo	Base no DNA	Símbolo	Base correspondente no RNA	Símbolo
Bases púricas ou purinas	Citosina	C	Citosina	C
	Timina	T	Uracila	U
Bases pirimídicas ou pirimidinas	Adenina	A	Adenina	A
	Guanina	G	Guanina	G

Tabela 2.1: Bases nitrogenadas: grupos e símbolos

2.1.1 Estrutura dos ácidos nucléicos

Em 1953, Watson e Crick propuseram um modelo para a estrutura do DNA que explicava as regularidades de composição das bases, suas propriedades biológicas e, especialmente, sua duplicação na célula [27]. A "espinha dorsal" da macromolécula de DNA é formada pelos resíduos de fosfato e pentose, encadeados alternadamente; as bases nitrogenadas ligam-se lateralmente às pentoses. Em seu estado mais comum, as macromoléculas de DNA ocorrem aos pares. Em cada par, as espinhas dorsais das duas cadeias estão torcidas, formando uma hélice dupla em torno de um eixo central. As duas cadeias se prendem uma à outra por meio de pontes de hidrogênio (ligações fracas devidas à atração elétrica) , estabelecidas entre os pares de bases. Cada base púrica de uma cadeia se liga a uma base pirimídica da outra cadeia. Mais precisamente, cada adenina (A) de uma cadeia corresponde a uma timina (T) da outra, e cada guanina (G) com uma citosina (C). As bases A e T são ditas *complementares* uma da outra, como C e G [24].

As duas cadeias do DNA podem ocasionalmente ocorrer parcialmente ou total-

mente separadas, por exemplo durante a divisão celular ou outros processos vitais. As cadeias de RNA, por outro lado, geralmente ocorrem isoladas.

Segundo o modelo de Watson e Crick (confirmado por milhares de experimentos e estudos), toda informação genética de um organismo está codificada na seqüência com que as quatro bases (A, T ou U, C e G) ocorrem ao longo de certas moléculas-mestre de DNA ou RNA, seu *genoma*. Na reprodução do organismo, ou na divisão de suas células, esta seqüência é duplicada por processos químicos e as cópias são passadas aos organismos ou células descendentes, o que explica o fenômeno da hereditariedade. Muitos outros fenômenos genéticos, como a combinação de características na reprodução sexual e a identidade imunológica, são explicados por outros processos que alteram as cadeias de DNA ou RNA do genoma.

2.2 Proteínas

O genoma controla os processos vitais do organismo principalmente controlando a síntese de *proteínas*. Esta é outra grande classe de macromoléculas orgânicas, que desempenham uma enorme variedade de papéis no ser vivo — suporte físico, geração de movimento, separação e transporte de substâncias, sinalização e comunicação, etc. As proteínas mais importantes são as *enzimas*, que catalisam e controlam a maioria das reações químicas que ocorrem no organismo, incluindo a síntese e degradação (desmantelamento) de outras proteínas.

Cada parte de um genoma que especifica a síntese de uma espécie de proteína é denominada o *gene* da mesma. Assim, por exemplo, no genoma de todo vertebrado há pelo menos um gene que descreve a hemoglobina, a proteína que transporta oxigênio pela corrente sanguínea. Outro gene descreve a enzima peptose, que é liberada no estômago para digerir as proteínas dos alimentos. Outro codifica a substância que, na retina, detecta a luz vermelha. E assim por diante.

Cada proteína é também um polímero linear, cujas unidades são resíduos de *aminoácidos* — compostos orgânicos simples que possuem um grupo amina ($-NH_2$, básico) e um grupo carboxílico (–COOH, ácido) presos ao mesmo átomo de car-

bono. Na síntese da proteína, os aminoácidos são unidos entre si por esses radicais, com eliminação de moléculas de água, formando uma espinha dorsal com estrutura $\cdots$–NH–C–CO–NH–C–CO–$\cdots$. Em geral, cada proteína ocorre dobrada sobre si mesma de uma forma característica, complexa e irregular. Assim como no DNA, o tipo e seqüência dos radicais ligados aos átomos –C– desta cadeia determinam as propriedades da proteína, incluindo seu dobramento e sua função no organismo.

2.2.1 O código genético

A seqüência dos aminoácidos de uma proteína está codificada no seu gene segundo um esquema relativamente simples, comum a todos os seres vivos, o *código genético*. Cada aminoácido é especificado por um grupo de três nucleotídeos consecutivos denominado *códon*. Cada combinação de três bases num códon especifica um único aminoácido, mas alguns aminoácidos podem ser especificados de várias formas; portanto, apesar de haver $4^3 = 64$ códons possíveis, a maioria das proteínas é formada por apenas 20 aminoácidos diferentes. Uma das 64 combinações é o "ponto final" que indica o fim do gene. A tabela 2.2 apresenta os vinte aminoácidos mais comumente encontrados nas proteínas, e seus respectivos códons. Os símbolos são convencionalmente usados em biologia computacional e em bancos de dados para descrever as seqüências de aminoácidos de uma proteína.

2.2.2 RNA mensageiro

A síntese de uma proteína é normalmente feita em duas etapas. Na etapa de *transcrição*, uma enzima especial (transcritase) percorre o trecho do DNA do organismo que contém o gene correspondente, construindo uma cópia em RNA, denominada *RNA mensageiro* (mRNA). Na etapa de *síntese* propriamente dita, esta cópia é processada por uma segunda enzima, o *ribossomo*; este "lê" um códon de cada vez, e acrescenta o aminoácido correspondente à proteína sendo formada.

O ribossomo na verdade é um aglomerado de diversas proteínas separadas e de uma cadeia de RNA especial, o *RNA ribossomal* (rRNA). Este RNA não contém

Símbolos		Aminoácido	Códon(s)
A	Ala	Alanina	GCT, GCC, GCA, GCG
C	Cis	Cisteína	TGT, TGC
D	Asp	Ácido Aspártico	GAT, GAC
E	Glu	Ácido Glutâmico	GAA, GAG
F	Fen	Fenilalanina	TTT, TTC
G	Gli	Glicina	GGT, GGC, GGA, GGG
H	His	Histidina	CAT, CAC
I	Ile	Isoleucina	ATT, ATC, ATA
K	Lis	Lisina	AAA, AAG
L	Leu	Leucina	TTA, TTG, CTT, CTC, CTA, CTG
M	Met	Metionina	ATG
N	Asn	Asparagina	AAT, AAC
P	Pro	Prolina	CCT, CCC, CCA, CCG
Q	Gln	Glutamina	CAG, CAA
R	Arg	Arginina	CGT, CGC, CGA, CGG, AGA, AGG
S	Ser	Serina	TCT, TCC, TCA, TCG, AGT, AGC
T	Tre	Treonina	ACT, ACC, ACA, ACG
V	Val	Valina	GTT, GTC, GTA, GTG
W	Trp	Triptofano	TGG
Y	Tir	Tirosina	TAT, TAC

Tabela 2.2: Os 20 aminoácidos mais comumente encontrado nas proteínas, e os respectivos códons no DNA.

informação genética, mas serve principalmente como armação para as proteínas.

Para ser usado como matéria-prima pelo ribossomo, cada aminoácido deve estar quimicamente "rotulado" por um pequeno trecho de RNA, o *RNA de transferência* (tRNA). Cada um dos 20 aminoácidos é rotulado com um tRNA diferente, que possui em certo ponto a seqüência de três bases complementar ao códon do aminoácido. É pelo casamento entre este trecho do tRNA e o códon corrente do mRNA que o ribossomo consegue identificar o aminoácido correto a ser acrescentado. Cada um dos 20 tipos de complexo tRNA-aminoácido é fabricado por uma enzima especializada.

2.3 Eucariotos, procariotos, e vírus

Há três grandes divisões de organismos vivos, os *eucariotos*, os *procariotos*, e os *vírus*.

Um eucarioto é geralmente constituído de várias células relativamente estanques; seu genoma é constituído de DNA, e normalmente está agrupado num núcleo bem definido, separado do restante da célula por uma membrana. Este grupo inclui todos os animais e plantas.

Um organismo procarioto é formado por apenas uma célula isolada que não possui núcleo. O genoma (na forma de um ou mais anéis de DNA) fica espalhado no citoplasma. Exemplos de procariotos são as bactérias e as algas azuis.

Os vírus geralmente consistem de um envelope de proteína contendo uma cadeia de RNA ou (raramente) DNA; sem metabolismo próprio, eles precisam invadir organismos procariotos ou eucariotos, e usar as enzimas e processos metabólicos destes para se reproduzir. Os vírus são responsáveis por muitas doenças de animais e plantas, como o resfriado, a varíola e a AIDS. Vírus que atacam bactérias são chamados de *bacteriófagos* ou *fagos*. Os *retrovírus* conseguem incorporar seu RNA (depois de oportunamente "traduzido" para DNA) ao genoma do organismo infectado.

2.3.1 Éxons e íntrons

Nos organismos procariotos e nos vírus, cada gene normalmente está codificado como um único trecho de DNA, e quase todo o genoma consiste de genes. Já nos eucariotos, cada gene normalmente consiste de vários trechos separados do DNA genômico, os *éxons*, intercalados com trechos de DNA que não codificam proteínas, os *íntrons*.

Sabe-se que alguns dos íntrons exercem funções de controle, por exemplo liberar ou bloquear a transcrição de genes vizinhos dependendo da presença de certa substância na célula. Outros íntrons podem ser identificados como cópias do RNA de retrovírus. Entretanto, a maior parte dos íntrons (que podem tomar uma fração considerável do genoma) ainda tem função desconhecida, e talvez sejam apenas "lixo" — cadeias sem função, acumuladas por acaso durante bilhões de anos de evolução.

2.4 Homologia

Dizemos que dois trechos de genoma (quer obtidas de dois organismos diferentes, quer de um mesmo organismo) são *homólogos* quando eles descendem de uma mesma cadeia de DNA ou RNA ancestral. É importante notar que dois trechos homólogos não têm, usualmente, seqüências de nucleotídeos *idênticas*, pois eles sofreram diferentes alterações devidas às suas mutações (troca, remoção ou inserção de bases); mas têm seqüências *similares* porque estas alterações ocorreram sobre a mesma seqüência original, o gene do ancestral comum.

Capítulo 3

Elementos de biologia computacional

A *biologia molecular computacional* é a área da ciência da computação que objetiva prover métodos computacionais para resolver problemas envolvendo bio-seqüências(DNA, RNA e proteínas).

3.1 Comparação de bio-seqüências

O problema mais importante da biologia computacional é a comparação de bio-seqüências. Ela envolve os conceitos de *alinhamento* e *similaridade*. Nesta seção apresentamos estes conceitos e abordamos também a relação entre similaridade e homologia.

3.1.1 Alinhamento

Na comparação de duas bio-seqüências a, b, comparar as bases na mesma ordem — isto é a_i com b_i, para todo i — não é uma boa idéia, porque a evolução e/ou a aquisição de uma bio-seqüência pode causar a inserção ou eliminação de bases.

Para contornar esse problema, é necessário primeiro *alinhar* as duas seqüências,

isto é, encontrar um emparelhamento das bases de modo que cada base seja comparada com a base que, com maior probabilidade, descende da mesma base ancestral.

O *alinhamento* pode ser definido como a inserção de espaços em locais arbitrários ao longo das seqüências de modo que elas terminem com o mesmo tamanho [24]. Assim, por exemplo, as seqüências $s_1 = GACGGTCAC$ e $s_2 = GATCGGTTAG$, podem ser alinhadas da seguinte forma

	1	2	3	4	5	6	7	8	9	10
s_1	G	A	–	C	G	G	T	C	A	C
s_2	G	A	T	C	G	G	T	T	A	G
			↑					↑		↑

Figura 3.1: Alinhamento de seqüências de DNA

Ao alinharmos as seqüências desta forma, estamos supondo que o T na posição 3 de s_2 foi inserido, ou a base correspondente de s_1 foi eliminada; e que houve alterações nas bases das posições 8 e 10, em s_1 ou em s_2.

Para cada alinhamento de duas seqüências podemos comparar as bases correspondentes e calcular a sua *similaridade*, segundo um critério de pontuação que recompensa acertos e penaliza erros e espaços. Quando o alinhamento correto entre as bases não é conhecido, supõe-se geralmente que o alinhamento que fornece a maior similaridade (o *alinhamento ótimo*) é o mais provável.

Algoritmo para alinhamento ótimo

O algoritmo padrão para encontrar o alinhamento ótimo de duas seqüências s e t, e calcular sua similaridade baseia-se na técnica de programação dinâmica [6]. O custo deste algoritmo é $O(nm)$, onde n e m são os tamanhos das seqüências.

O algoritmo utiliza uma *matriz de similaridade* a, com $m+1$ linhas e $n+1$ colunas. Numa primeira etapa, o algoritmo preenche cada elemento $a[i, j]$ com a similaridade entre os prefixos $s[1..i]$ e $t[1..j]$. Ao final do preenchimento, a similaridade entre as duas seqüências estará na posição $a[m, n]$.

A figura 3.2 apresenta o algoritmo para preenchimento da matriz de similaridade. O parâmetro g é a penalidade para espaços. A função $p(i, j)$ assume o valor 1 se $s_i = t_j$, e -1 se $s_i \neq t_j$.

```
1.  Algoritmo Similaridade
2.    entrada:  seqüências s e t; penalidade g
3.    saída:  similaridade entre s e t
4.       m ← |s|
5.       n ← |t|
6.       para i ←0 até m faça
8.          a[i,0] ← i * g
9.       para j ←0 até n faça
10.          a[0,j] ← j * g
11.       para i ←0 até m faça
12.          para j ←0 até n faça
13.             a[i,j] ← max(a[i-1,j] + g, a[i-1,j-1]+p(i,j), a[i,j-1]+g)
14.   retorna a[m,n]
```

Figura 3.2: Algoritmo de Similaridade

A figura 3.3 mostra a matriz de alinhamento para as seqüências $s = AAAC$ e $t = AGC$, calculada usando penalidade $g = -2$.

			A	G	C
		0	1	2	3
	0	0	-2	-4	-6
A	1	-2	-1	-1	3
A	2	-4	-1	0	2
A	3	-6	-3	-2	-1
C	4	-8	-5	-4	-1

Figura 3.3: Matriz de alinhamento

Numa segunda etapa, o algoritmo obtém o alinhamento ótimo a partir da matriz de similaridade. O algoritmo parte da posição $a[m, n]$ da matriz e identifica o maior dentre os valores $a[m-1, n]+g$, $a[m, n-1]+g$ e $a[m-1, n-1]+p(i, j)$. A posição correspondente passa a ser a nova posição de partida. O algoritmo recursivamente assume novas posições de partida até chegar na posição $[0, 0]$. Neste ponto, o algoritmo retorna passo-a-passo pelas suas posições anteriores gerando, efetivamente, o alinhamento. Cada posição percorrida na volta nos dá uma coluna do alinhamento. Dependendo se o deslocamento será na horizontal, vertical ou diagonal, isto equivale à inserção de um espaço em uma das seqüências ou à correspondência das seqüências naquela posição.

A figura 3.4 apresenta o algoritmo recursivo que constrói o melhor alinhamento entre duas seqüências[24]. Como resultado para o caso das seqüências $s = AAAC$ e $t = AGC$, o algoritmo retorna a seguinte solução:

s'	A	A	A	C
t'	A	G	–	C

3.1.2 Homologia e similaridade

A homologia entre duas (ou mais) bio-seqüências não pode ser determinada diretamente, pois o processo de replicação ou leitura que as produziu a partir da seqüência ancestral comum não pode ser observado; e, no caso de genomas diferentes, o ancestral comum freqüentemente não está mais disponível. Portanto, a homologia entre genes é deduzida a partir da similaridade entre os mesmos. Ou seja, quanto maior a similaridade entre duas seqüências, menor a probabilidade de que elas tenham sido geradas independentemente uma da outra e de que a similaridade ocorra por acaso. Como exemplificado por Koonin e Galperin [15], tomando duas seqüências de 100 aminoácidos que apresentam 80% de identidade, podemos calcular a probabilidade disto ocorrer por acaso e verificar que ela é tão pequena que se torna improvável que tal evento tenha ocorrido nos últimos 5 bilhões de anos

```
1.  Algoritmo Alinhamento
2.    entrada:  índices i, j, matriz a dada pelo algoritmo Similaridade
3.    saída:  alinhamento em s', t', e tamanho em tam
4.       se i = 0 e j = 0 então
5.          tam ← 0
6.       senão se i > 0 e a[i,j] = a[i-1,j] + g então
7.          Alinhamento(i-1, j, tam)
8.          tam ← tam + 1
9.          s'[tam] ← s[i]
10.         t'[tam] ← -
11.      senão se i > 0 e j > 0 e a[i,j] = a[i-1,j-1] + p(i,j) então
12.         Alinhamento(i-1, j-1, tam)
13.         tam ← tam + 1
14.         s'[tam] ← s[i]
15.         t'[tam] ← t[j]
16.      else
17.         Alinhamento(i, j-1, tam)
18.         tam ← tam + 1
19.         s'[tam] ← -
20.         t'[tam] ← t[j]
```

Figura 3.4: Algoritmo recursivo para alinhamento ótimo

— exceto por duplicação de uma seqüência ancestral comum, e concluir daí que as seqüências em questão são homólogas.

Capítulo 4

Conceitos básicos de teoria da informação

Neste capítulo revisamos alguns conceitos de probabilidade e teoria da informação que auxiliam na compreensão da análise do conteúdo de informação presente nas seqüências de DNA.

4.1 Conceitos de probabilidade e estatística

4.1.1 Densidade de probabilidade

Denotaremos por $Pr(X \approx x)$ a densidade de probabilidade de X no ponto x, ou seja,

$$Pr(X \approx x) = \lim_{h \to 0} \frac{Pr(x \leq X \leq x + h)}{h}$$

4.1.2 Probabilidade condicional

Sejam duas variáveis aleatórias A e B que podem assumir os valores $\{a_1, \ldots a_p\}$ e $\{b_1, \ldots b_q\}$, respectivamente. Denotaremos por $P(B = b_i | A = a_j)$, a probabilidade

condicional de $B = b_i$ dado $A = a_j$, ou seja,

$$\Pr(B = b_j | A = a_i) = \frac{\Pr(B = b_j \wedge A = a_i)}{\Pr(A = a_1)} \tag{4.1}$$

4.1.3 Média, variância e desvio padrão

A *média* (ou *valor médio*, ou *valor esperado*) de uma variável aleatória real X é definida como $E[X] = \int_{-\infty}^{+\infty} x \Pr(X \approx x)\, dx$, onde $\Pr(X \approx x)$ é a densidade de probabilidade de X no ponto x. Como é sabido, o valor de $E[X]$ pode ser estimado a partir de uma amostra de valores $X[i]$, $i \in \{0..M-1\}$, pela média aritmética

$$E[X] \approx \frac{1}{M} \sum_{i=0}^{M-1} X[i] \tag{4.2}$$

A variância de uma variável aleatória Y, denotada por $V(Y)$ ou $\hat{Y}$, é definida por

$$\hat{Y} = V(Y) = E(Y^2) - [E(Y)]^2 \tag{4.3}$$

O desvio padrão da variável X é a raiz quadrada da sua variância, ou seja,

$$\sigma[X] = \sqrt{V[X]}$$

4.2 Informação

Lathi[17] nos explica que a *quantidade de informação* recebida em uma mensagem está diretamente relacionada com a incerteza ou inversamente relacionada com a probabilidade de sua ocorrência. Denotando por p a probabilidade de ocorrência de uma mensagem específica e por $I(p)$ a informação proporcionada pela mesma, intuitivamente espera-se que $\lim_{p\to 1} I(p) = 0$, e $\lim_{p\to 0} I(p) = \infty$. Uma medida de informação útil deve satisfazer também várias outras propriedades intuitivas. Por exemplo, se uma mensagem consiste de duas partes independentes, então a informação contida na mesma deve ser a soma da informação contida nessas partes —

ou seja, $I(p_1 p_2) = I(p_1) + I(p_2)$. Esta e outras considerações implicam na seguinte fórmula

$$I(p) \sim \log \frac{1}{p} \tag{4.4}$$

Podemos ainda utilizar o ponto de vista da engenharia[17] que considera que o conteúdo de informação de qualquer mensagem é igual ao número mínimo de dígitos requeridos para codificá-la.

A base do logaritmo na fórmula (4.4) é arbitrária e afeta apenas um fator de escala, isto é, a unidade de medida da informação. Adotando-se a base 2, a unidade será um *bit* — a informação contida numa mensagem que tem probabilidade 1/2.

4.3 Entropia

A entropia de uma variável aleatória X é a quantidade média de informação fornecida por cada elemento de uma seqüência de valores sucessivos de X. Se X assume os valores $x_1 \ldots x_m$ com probabilidades $p_1 \ldots p_m$ e cada valor de X é independente dos anteriores, então a entropia de X é

$$H[X] = \sum_{i=1}^{m} p_i I(p_i) = -\sum_{i=1}^{m} p_i \log_2 p_i \tag{4.5}$$

Se os valores assumidos por X não são independentes entre si então a entropia de X será menor do que o valor fornecido pela equação (4.5). Isto porque a dependência de um valor com valores prévios reduz sua incerteza[17].

4.3.1 Entropia de um par de variáveis independentes

Em particular, suponha que a variável X seja na verdade um par de variáveis $X = (Y, Z)$, de modo que cada valor X_i seja um par (Y_j, Z_k). Se as variáveis Y e Z forem independentes entre si — ou seja, se $\Pr(X = (Y_j, Z_k)) = \Pr(Y = Y_j) \Pr(Z = Z_k)$ — então as entropias de Y e Z podem ser somadas e $H(X)$ é dada por:

$$H[X] = H[Y] + H[Z] \tag{4.6}$$

4.3.2 Entropia de variável gaussiana

Se X é uma variável aleatória cuja função densidade de probabilidade é gaussiana (ou normal),

$$G_\sigma(\tau) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(\tau-\mu)^2}{2\sigma^2}) \tag{4.7}$$

onde μ e σ representam, respectivamente, a média e o desvio padrão da distribuição; neste caso, verifica-se que a sua entropia, para o caso de X ser real, é dada por

$$H(X) = \frac{1}{2}\log_2(2\pi e\hat{X}) \tag{4.8}$$

onde $\hat{X}$ é a variância de X.

4.3.3 Entropia condicional

Definimos *entropia condicional*, de B dado A, como a quantidade média de informação fornecida sobre o valor de B uma vez que conhecemos o valor de A:

$$H(B|A) = \sum_{i=1}^{p} H(B|A=a_i)\Pr(A=a_i) \tag{4.9}$$

onde $H(B|A=a_i)$ é o resultado da aplicação da equação (4.5) para a distribuição de probabilidade condicional de B dado A (equação (4.1)).

A partir da entropia condicional, podemos calcular a quantidade de informação média que A fornece a respeito de B, ou seja, a informação que obtemos a respeito de B quando conhecemos o valor assumido por A.

$$I(A \to B) = H[B] - H[B|A] \tag{4.10}$$

A informação média $I(A \to B)$ dada por A sobre B também é o valor esperado de $I(A=a_i \to B)$, para um valor genérico de a_i, ou seja,

$$I(A \to B) = \sum_{i=1}^{p} \Pr(A=a_i) I(A=a_i \to B) \tag{4.11}$$

4.3.4 Informação mútua de variáveis gaussianas

Um caso particular no cálculo da informação condicional ocorre quando $A = S + Q$ e $B = S + R$, onde S, Q e R são variáveis independentes com distribuições gaussianas simétricas e variâncias $\hat{S}$, $\hat{Q}$, e $\hat{R}$. Podemos supor que S é uma "mensagem" a partir da qual foram feitas duas cópias independentes, e que foram corrompidas por "ruídos" Q e R gerando as "mensagens" A e B. Neste caso, segue que A e B também têm distribuições gaussianas simétricas com variâncias $\hat{A} = \hat{S}+\hat{Q}$ e $\hat{B} = \hat{S} + \hat{R}$, respectivamente.

Então a equação (4.8) é simplesmente

$$H[B] = \frac{1}{2}\log_2(2\pi e\hat{B}) = \frac{1}{2}\log_2[2\pi e(\hat{S} + \hat{R})] \tag{4.12}$$

Além disso, a distribuição condicional $\Pr(S \approx x|A = y)$ passa a ser outra gaussiana com média $y\hat{S}/\hat{A}$ e variância $\hat{S}\hat{Q}/\hat{A}$. Como R é independente de Q e S, a distribuição condicional de $B = S + R$ dado que $A = y$ é também uma gaussiana, com a mesma média $y\hat{S}/\hat{A}$ e variância $\hat{S}\hat{Q}/\hat{A} + \hat{R}$. Observe que este valor não depende de y; então a equação(4.9) pode ser reduzida a

$$H[B|A] = \frac{1}{2}\log_2\left[2\pi e\left(\frac{\hat{S}\hat{Q}}{\hat{A}} + \hat{R}\right)\right] \tag{4.13}$$

De acordo com as equações 4.10, 4.12 e 4.13, a quantidade de informação dada por A sobre B é então

$$I(A \to B) = \frac{1}{2}\log_2\left[\frac{\hat{A}\hat{B}}{\hat{S}\hat{Q} + \hat{A}\hat{R}}\right] =$$
$$\frac{1}{2}\log_2\left[\frac{(\hat{S} + \hat{Q})(\hat{S} + \hat{R})}{\hat{S}\hat{Q} + (\hat{S} + \hat{Q})\hat{R}}\right] = \frac{1}{2}\log_2\left[\frac{(\hat{S} + \hat{Q})(\hat{S} + \hat{R})}{\hat{S}(\hat{Q} + \hat{R}) + \hat{Q}\hat{R}}\right] \tag{4.14}$$

Note que esta equação é simétrica em relação a $\hat{Q}$ e $\hat{R}$, portanto concluímos que $I(A \to B) = I(B \to A)$ o que justifica o nome "informação mútua".

A equação de Shannon-Harley

Em particular, no caso em que $R = 0$ (ou seja, $B = S$) a equação 4.14 se reduz à equação de Shannon-Hartley[17],

$$I(A \to S) = \frac{1}{2}\log_2\left[\frac{\hat{A}}{\hat{Q}}\right] = \frac{1}{2}\log_2\left[\frac{\hat{S}+\hat{Q}}{\hat{Q}}\right] \tag{4.15}$$

que fornece a quantidade de informação dada pela mensagem corrompida A sobre a mensagem original S.

Capítulo 5

Análise de Fourier e filtragem de sinais

Neste capítulo, estudamos a representação de sinais no domínio das freqüências através das transformadas de Fourier e revisamos as principais características de alguns tipos de filtros que são utilizados para atenuar a potência do sinal em determinadas faixas de freqüência.

5.1 Transformada discreta de um sinal

Definimos um sinal discreto como sendo uma seqüência x de n valores $x(0), \ldots x(N-1)$. A *transformada discreta de Fourier* (TDF) de x é a seqüência X de N *coeficientes* $X(0), \ldots X(N-1)$ definida pela equação

$$X(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) \exp\left[-\mathbf{i}\frac{2\pi}{N}nk\right] \tag{5.1}$$

onde o parâmetro k é a freqüência do coeficiente $X(k)$.

Esta transformação pode ser vista como uma mudança de base no espaço $\mathbf{C}^N$. Ou seja, os coeficientes $X(k)$ são as coordenadas do vetor complexo $x = (x(0), x(1), ...x(N-$

1)) na *base de Fourier*, o conjunto de vetores $\phi_0, \phi_1, ...\phi_{N-1}$ definidos por

$$\phi_k(n) = \frac{1}{\sqrt{N}} exp\left[\mathbf{i}\frac{2\pi}{N}nk\right]$$

O gráfico da seqüência $\phi_k(n)$ em função de n é uma espiral que dá k voltas em torno do eixo n. Esta base é ortonormal segundo o produto escalar do espaço $\mathbf{C}^N$, definido por

$$\langle x|y\rangle = \sum_{n=0}^{N-1} x(n)y^*(n)$$

onde z^* é o conjugado do número complexo z. Ou seja, $\langle\phi_j|\phi_j\rangle = 1$, e $\langle\phi_j|\phi_k\rangle = 0$ para todo j,k em $0, \ldots N-1$ com $j \neq k$ Portanto, a transformada inversa é dada por

$$x(n) = \sum_{k=0}^{N-1} X(k)\phi_k(n) \tag{5.2}$$

onde cada seqüência $X(k)\phi_k$ é a *componente (espectral)* do sinal x de freqüência k.

Na transformada discreta de Fourier, tanto a seqüência de amostras x quanto a seqüência de coeficientes X são implicitamente consideradas periódicas, com período N. Isso decorre do fato que, segundo a equação (5.1), $X(k) = X(k+N)$ para todo k; e, segundo (5.2) e $x(n) = x(n+N)$ para todo n. Em particular, as componentes $X(k)\phi_k$ com freqüências k entre $\lfloor N/2\rfloor+1$ e $N-1$ são idênticas às componentes com freqüências entre $1-\lceil N/2\rceil$ e -1, respectivamente. Portanto, na representação da transformada podemos usar indiferentemente k entre 0 e $N-1$ ou entre $1-\lceil N/2\rceil$ e $\lfloor N/2\rfloor$.

5.2 Espectro de potência de um sinal

Em física, verifica-se que a *potência média* $\|x\|^2$ (a energia média por unidade de tempo) de uma seqüência de impulsos elétricos a com período N é proporcional à soma dos quadrados de suas amplitudes; ou seja, numa unidade apropriada,

$$\|a\|^2 = \sum_{n=0}^{N-1} |a(n)|^2 \tag{5.3}$$

Esta fórmula pode ser diretamente estendida a sinais complexos, e fornece uma medida do "tamanho" do sinal. (Tecnicamente, $\|\cdot\|^2$ é o quadrado de uma norma para o espaço $\mathbf{C}^N$ das seqüências de N números complexos.)

Segundo o teorema de Parseval [17], a potência $\|a\|^2$ também pode ser obtida a partir da transformada discreta de Fourier A do sinal a, pela fórmula

$$\|a\|^2 = \sum_{k=0}^{N-1} |A(k)|^2 \tag{5.4}$$

Cada termo $|A(k)|^2$ representa a fração da potência do sinal a que está contida na sua componente de freqüência k, ou seja no sinal complexo $A(k)\phi_k$.

Observamos que, para cada freqüência $k \in \{1..\lfloor (N-1)/2 \rfloor\}$, o gráfico do sinal ϕ_k) é uma espiral de raio 1 que, quando t varia de 0 a N, dá k voltas completas em torno do eixo t. O gráfico do sinal ϕ_{k-1}, que corresponde à freqüência $-k$, é idêntico exceto pelo sentido de rotação da espiral. Portanto, na somatória (5.4), os termos $|A(k)|^2$ e $|A(-k)|^2$ ($= |A(N-k)|^2$) na verdade medem ondas de mesma freqüência k, apenas com diferentes relações entre as partes reais e imaginárias. Portanto, para fins de análise de fenômenos periódicos num sinal, é geralmente conveniente somar essas duas componentes.

Ou seja, é conveniente definir o *espectro de potência* (ou simplesmente *espectro*) de um sinal complexo $a(t)$ de período N como sendo a seqüência

$$\tilde{A}(k) = \begin{cases} |A(k)|^2 & \text{se } k = 0 \text{ ou } k = N/2, \\ |A(k)|^2 + |A(N-k)|^2 & \text{se } 0 < k < N/2 \end{cases} \tag{5.5}$$

Note que o espectro de potência tem $\lfloor N/2 \rfloor + 1$ elementos $\tilde{A}(k)$, para k entre 0 e $\lfloor N/2 \rfloor$; e que o último elemento é especial quando N é par.

5.3 Transformada rápida de Fourier(FFT)

A transformada discreta de Fourier de N amostras pode ser calculada em tempo $N \log N$ pelo algoritmo *Fast Fourier Transform* (FFT) criado por C. Runge e H. König e aperfeiçoado por J. W. Cooley e J. W. Tukey [5]

5.4 Filtragem de sinais

Filtragem é o processo pelo qual a parte útil e essencial de um sinal é separada de componentes indesejáveis ou *ruídos*.

O tipo mais comum de filtro é o *linear* e *invariante no tempo*. Seu efeito é obtido pela multiplicação da transformada de Fourier $X(k)$ de um sinal pelo fator de atenuação, ou função de transferência $H(k)$. Ou seja, a transformada de Fourier $Y(k)$ do sinal filtrado é dada por

$$Y(k) = X(k)H(k) \tag{5.6}$$

Em geral, a função $H(k)$ é escolhida de modo que as componentes espectrais que contêm a parte útil do sinal sejam preservadas ou pouco afetadas, enquanto que as que contém principalmente ruído sejam eliminadas ou reduzidas.

O efeito de um filtro linear e invariante no tempo pode ser também descrito como sendo a *convolução* do sinal original com um determinado sinal h, ou seja,

$$y(n) = \sum_{j=0}^{N-1} x(n-j)h(j)$$

O sinal h é chamado de *núcleo* do filtro; ele é a transformada de Fourier inversa da seqüência $H(k)$, e também a saída do filtro quando a entrada é um pulso que mede 1 para $n = 0$, e 0 para $n \in 1, \ldots N-1$.

5.4.1 Filtro retangular

Num *filtro retangular* a função $H(k)$ é 1 para freqüências dentro de certa faixa, e 0 fora dela — ou vice-versa. No caso mais comum, em que o ruído afeta mais as componentes de freqüência mais alta, utiliza-se um *filtro retangular passa-baixas*, cuja função de transferência é definida pela equação (5.7).

$$H(k) = \begin{cases} 1, & \mid k \mid < k_c \\ \\ 0, & \text{caso contrário} \end{cases} \tag{5.7}$$

onde k_c é a *freqüência de corte*.

Assim, a faixa de freqüências menores que k_c, denominada *banda passante*, é preservada, e a faixa das demais freqüências, denominada *banda de corte*, é eliminada.

O filtro retangular não é adequado para nossas finalidades pois pode introduzir oscilações espúrias no sinal filtrado quando este é convertido para o domínio do tempo, *o efeito de Gibbs*.

5.4.2 Filtro Gaussiano

Para evitar o efeito de Gibbs, a função de transferência não deve ter transições abruptas [28].

O melhor filtro, neste sentido, é o *filtro gaussiano*, cuja função de transferência tem a forma

$$H(k) = \exp\left(-\left(\frac{k\sigma\pi}{N}\right)^2\right) \tag{5.8}$$

O núcleo deste filtro é um sinal com perfil gaussiano com média zero e desvio padrão σ. O parâmetro σ, também chamado de *duração característica* ou *escala de tempo* do filtro gaussiano, determina o grau de suavização que o sinal receberá. Quanto maior o valor do desvio padrão σ, maior a suavização.

Observe na equação (5.8) que o peso $H(k)$ decresce muito rapidamente a medida que k aumenta. Podemos verificar em [20] que $H(k)$ pode ser considerado nulo quando

$$k \geq 2T/\sigma \tag{5.9}$$

Com esta relação é possível estabelecer o desvio padrão σ de modo que o sinal seja anulado a partir de uma freqüência de corte k_c.

Capítulo 6

Codificação de seqüências de DNA

Em biologia computacional, uma cadeia de DNA é geralmente representada por uma seqüência finita definida sobre um conjunto de símbolos $\Sigma = \{A, T, C, G\}$. Entretanto, esta representação não é conveniente para nosso objetivo, pois não é possível aplicar sobre a mesma técnicas de processamento de sinais. Para isso, devemos codificar as seqüências de DNA em seqüências numéricas.

6.1 Alguns métodos para codificar DNA

Uma opção consiste em atribuir a cada base um valor numérico arbitrário, digamos $A = 0$, $C = 1$, $T = 2$, $G = 3$. Entretanto, Wang [31] alerta para o fato de que este mapeamento sugeriria que um nucleotídeo seria maior que outro. Assim, por exemplo, a seqüência $AGAG \ldots AG$ teria espectro de potência bem maior que $ACAC \ldots AC$, mas esta é uma distinção que não tem justificativa biológica.

Voss [30] propôs decompor uma seqüência simbólica em Z seqüências binárias $U_z[n]$ as quais identificam se o símbolo z aparece na posição n. Para tanto, $U_z[n]$ assume o valor 1 se z aparece na posição n, e 0, caso contrário. Definindo os valores numéricos $\{a, c, t, g\}$ para os símbolos $\{A, C, T, G\}$, a seqüência numérica resultante

de uma seqüência simbólica de tamanho N seria:

$$x[n] \;=\; aU_A[n] + cU_C[n] + tU_T[n] + gU_G[n]\ ,\ n = 0, 1, 2, 3, ..., N-1 \qquad (6.1)$$

Tratando a seqüência genômica ATCCG de tamanho N=5 temos, por exemplo, que $U_A[0] = 1$, $U_A[1] = 0$, $U_C[2] = 1$ e $U_T[4] = 0$,

Apesar desta forma de codificação ser bastante popular [2, 8, 21, 31] , a utilização de 4 vetores para representar uma cadeia de DNA aumenta o custo computacional do cálculo da transformada discreta de Fourier.

6.2 Codificação do DNA no plano complexo

A codificação que nos pareceu mais adequada foi apresentada por Cheever *et al.* [4], onde cada base é representada por um número complexo. Especificamente, A, T, C, G são mapeados para $+1, -1, +i, -i$, onde $i = \sqrt{-1}$ é a unidade imaginária. Observe que cada base púrica e a correspondente base pirimídica são mapeadas com valores complementares. Esta codificação tem a vantagem de produzir um sinal complexo—que é um objeto já bem estudado na teoria de processamento de sinais—, e a entrada canônica para a transformada de Fourier. Chamamos esta seqüência de *sinal genômico*. Veja a figura 6.1.

O eixo vertical n representa as posições das bases na seqüência de DNA. A base T na posição 0 é representada pelo par complexo $(-1, 0)$; a base G na posição 1 é representada representada pelo par complexo $(0, -1)$; e assim por diante.

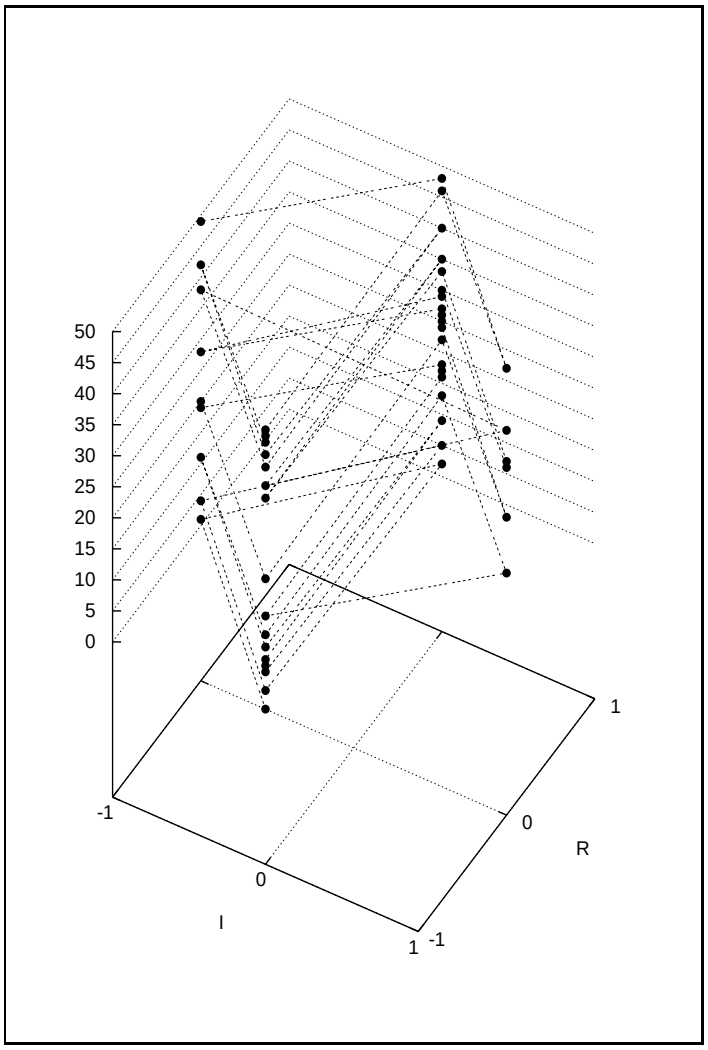

Figura 6.1: Codificação numérica da seqüência de DNA s = TGATGATTTATGTACTAA-AGGTACAAAAGAACCATATCGTATGTTTACA

Capítulo 7

Processamento do sinal genômico

Neste capítulo verificamos algumas características da transformada de Fourier de um sinal genômico que facilitam o cálculo da quantidade de informação. Também analisamos o espectro de potência do sinal genômico.

7.1 Transformada de Fourier

Conforme visto na seção 1.2, o principal obstáculo para o cálculo do conteúdo de informação útil do DNA na busca de similaridade é a correlação entre bases próximas. Como veremos a seguir, a transformada de Fourier, aplicada ao sinal complexo, nos permite superar este obstáculo porque as partes real e imaginária dos coeficientes de Fourier de uma seqüência são todas variáveis aleatórias independentes com distribuição de probabilidade gaussiana (ou normal). Unindo estas propriedades a outras da transformada de Fourier, podemos obter uma relação entre o espectro de potência médio e as variâncias dos coeficientes de Fourier.

7.1.1 Gerando uma amostra de DNA

Para confirmar experimentalmente o caráter gaussiano e a independência dos coeficientes de Fourier, utilizamos uma amostra de $M = 5000$ trechos de DNA de tamanho $N = 100$. Denotaremos os sinais complexos dos mesmos por $a[i](t)$ para $i \in \{0..M-1\}$ e $t \in \{0..N-1\}$, e as respectivas transformadas de Fourier por $A[i](k)$, para $k \in \{0..N-1\}$.

Estes trechos foram extraídos do genoma completo da bactéria *Pseudomonas aeruginosa PA01* (seqüência número NC_002516 do GenBank) que contém 6264403 bases.

Para evitar possíveis correlações entre as bases finais de um trecho $a[i]$ e as bases iniciais do trecho seguinte $a[i+1]$, que poderiam afetar a análise, o genoma acima foi dividido em M sub-genomas de tamanho $K = 1252$, e cada trecho $a[i]$ da amostra foi extraído aleatoriamente do sub-genoma correspondente, de forma a garantir um intervalo mínimo de 350 bases entre trechos consecutivos. Mais precisamente, denotando o sub-genoma i por $z[i](s)$ com $s \in \{0..K-1\}$, tomamos $a[i](t) = z[i](t+d[i])$ para $t \in \{0..N-1\}$, onde cada deslocamento $d[i]$ foi aleatoriamete escolhido entre 350 e 1151 com probabilidade uniforme.

7.1.2 Distribuição gaussiana dos coeficientes

Uma vez que cada coeficiente de Fourier é uma combinação linear de N variáveis aleatórias parcialmente independentes, pela lei dos grandes números [26] é lícito esperar que sua distribuição de probabilidade seja muito próxima de uma gaussiana.

Verificamos experimentalmente esta suposição, construindo histogramas dos valores de vários coeficientes, tanto da parte real quanto da parte imaginária, para nossa amostra de trechos de DNA. Mais precisamente, para cada gráfico escolhemos uma freqüência $k \in \{0..N-1\}$, e um operador $\mathrm{Op} \in \{\mathrm{Re}, \mathrm{Im}\}$, e construímos o histograma dos M valores $X[i] = \mathrm{Op}\, A[i](k)$, para $i \in \{0..M-1\}$, com faixas de largura $\delta = 0.1$.

A figura 7.1 mostra alguns destes histogramas. A altura de cada barra é o número de valores $X[i]$ observados na faixa correspondente, dividido por M e δ. A curva tracejada é o gráfico da distribuição gaussiana $g_\sigma(X)$, com média 0 e desvio padrão σ igual ao desvio padrão observado dos valores $X[i]$. Como podemos ver, em cada caso a distribuição observada dos valores $X[i]$ é bastante próxima de uma distribuição gaussiana, sendo que as discrepâncias podem ser atribuídas ao número limitado de amostras ($M = 5000$).

7.1.3 Independência dos coeficientes

A assertiva de que as partes real e imaginária dos coeficientes de Fourier que representam um sinal genômico são variáveis aleatórias independentes não pode ser provada matematicamente porque pode-se construir distribuições de cadeias de DNA que a violam. Entretanto, pode-se tentar verificar visualmente a existência de possíveis dependências pelo método gráfico descrito a seguir.

Sejam X e Y duas variáveis aleatórias, com distribuições gaussianas de médias $\bar{X}, \bar{Y}$ e variâncias $\hat{X}, \hat{Y}$, respectivamente. Para testar a independência destas variáveis, podemos tomar um grande número M de amostras simultâneas $X[0..M-1]$ e $Y[0..M-1]$, e plotar os M pares $(X[i], Y[i])$ como pontos isolados, obtendo uma figura conhecida como *gráfico de espalhamento*.

Se as duas variáveis forem independentes, o gráfico de espalhamento deve mostrar uma nuvem de pontos simétrica, horizontal e verticalmente, centrada no ponto $(\bar{X}, \bar{Y})$. Mais precisamente, a nuvem de pontos deve ser consistente com a distribuição gaussiana bidimensional de probabilidade

$$f(X,Y) = g_{\sqrt{\hat{X}}}(X - \bar{X})\, g_{\sqrt{\hat{Y}}}(Y - \bar{Y}) \tag{7.1}$$

O gráfico desta distribuição tem a forma de um sino; suas projeções nas direções X e Y são curvas gaussianas, e suas curvas de nível são elipses alinhadas com os eixos de coordenadas.

Portanto, se a nuvem de pontos for visivelmente inclinada ou distorcida, ou mostrar qualquer outro desvio consistente e significativo em relação à distribução (7.1),

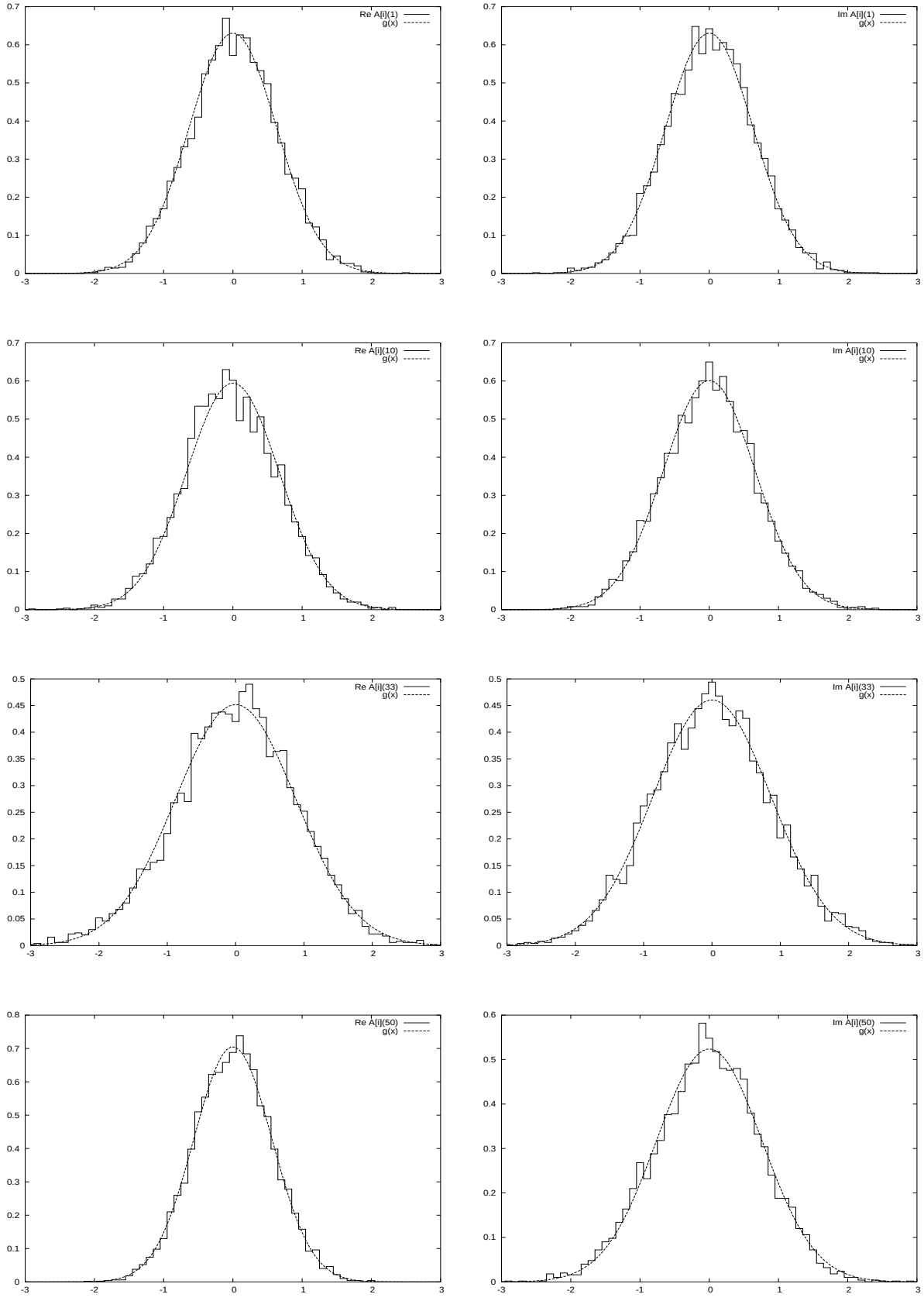

Figura 7.1: Histogramas dos coeficientes de Fourier de uma amostra de trechos de DNA, e curvas gaussianas de mesma variância

podemos concluir definitivamente que as variáveis *não* são independentes.

Por outro lado, se a nuvem de pontos for consistente com a distribuição (7.1), podemos contar isso como evidência de que elas *são* independentes. Devemos dizer *evidência* e não *prova*, pois pode-se imaginar relações sutis entre X e Y que não seriam visíveis no gráfico. Por exemplo, uma relação do tipo $X = \sin(10000Y)$ somente seria visível se plotássemos um número muito grande de pares, num gráfico com escalas muito grandes. Porém, no nosso caso, é difícil imaginar um mecanismo biológico que produzisse dependências deste tipo.

Recordando, nosso objetivo é verificar que os coeficientes de Fourier de um trecho de DNA aleatoriamente escolhido, considerados como $2N$ variáveis aleatórias reais, são independentes entre si. Construimos portanto gráficos de espalhamento para diversos pares de coeficientes de Fourier dos trechos de DNA da amostra descrita na seção 7.1.1, tanto das partes reais quanto das partes imaginárias. Mais precisamente, para cada gráfico escolhemos um par de freqüências $k', k'' \in \{0..N-1\}$, um par de operadores $\mathrm{Opa}, \mathrm{Opb} \in \{\mathrm{Re}, \mathrm{Im}\}$, e plotamos todos os pontos $(X[i], Y[i])$ para $i \in \{0, ..M-1\}$, onde $X[i] = \mathrm{Opa}\, A[i](k')$ e $Y[i] = \mathrm{Opb}\, A[i](k'')$.

Como era inviável examinar todas as $2N(2N-1)/2 = 19900$ combinações de coeficientes possíveis, concentramos o exame em alguns pares onde havia mais suspeita *a priori* de correlação, como

$$k'' = k' \qquad k'' = -k' \qquad k'' = k' + 1 \qquad k'' = 2k' \qquad k'' = 3k' \tag{7.2}$$

e para freqüências especiais como $k' = 1$, $k' = N/2$,, $k' = N/3$, etc.

A figura 7.2 mostra alguns destes gráficos. Podemos observar a distribuição simétrica dos pontos sobre o plano complexo indicando a independência entre os coeficientes.

7.1.4 Média dos coeficientes de Fourier

Recordamos que a parte real de cada coeficiente de Fourier $\mathrm{Re}\, A(k)$ de um sinal genômico a é uma combinação linear $\sum_{t=0}^{N-1} c_k(t)\, \mathrm{Re}\, a_n - s_k(t)\, \mathrm{Im}\, a_n$ das bases

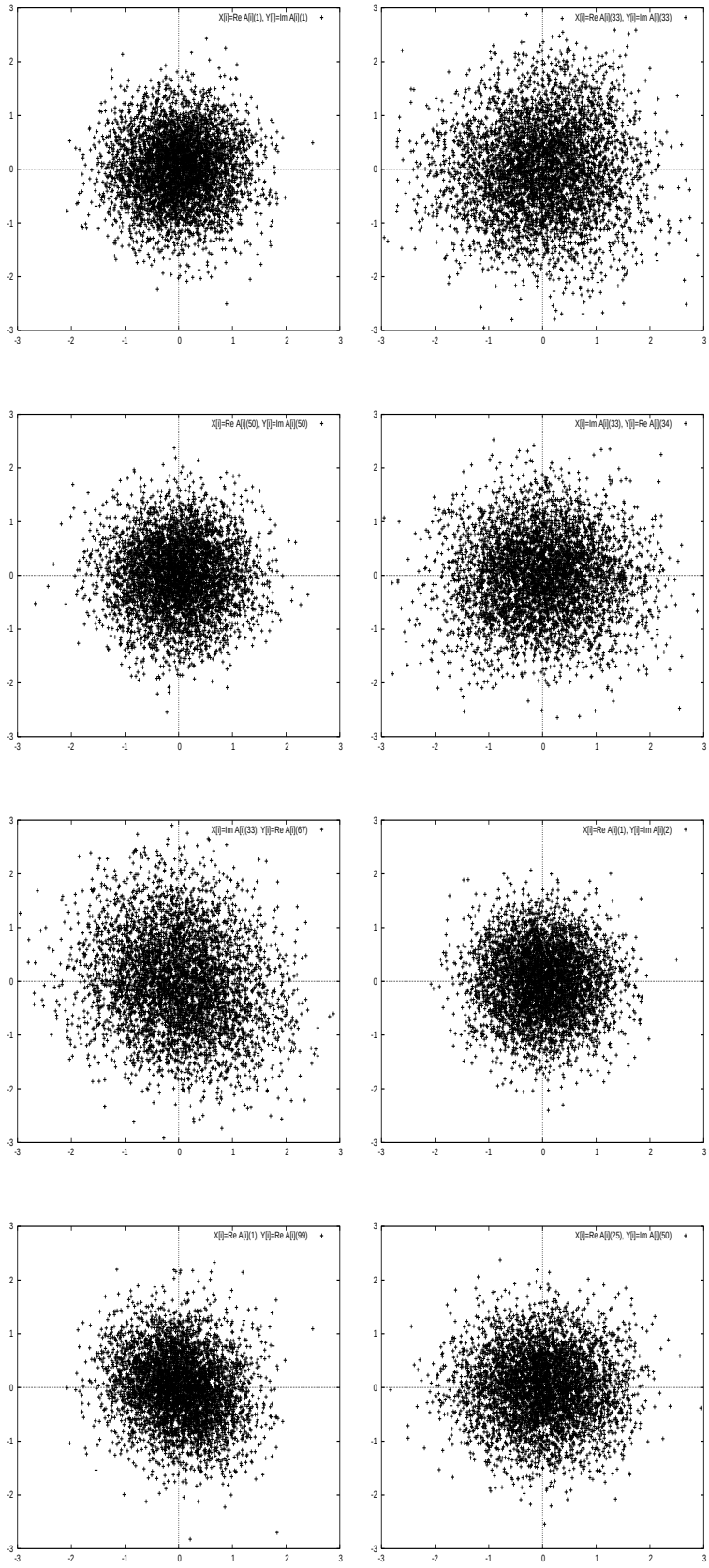

Figura 7.2: Verificação da independência dos coeficientes de Fourier de um conjunto de trechos de DNA

codificadas a_n, onde os pesos $c_k(t)$ são uma cossenóide de freqüência k e $s_k(n)$ são as senóides correspondentes. Pode-se verificar que, exceto para $c_0 = 0$, a soma desses pesos é sempre zero. Portanto, se o trecho a é extraído de um genoma arbitrário a partir de um ponto aleatório, é razoável supor que a somatória acima assume um certo valor X e seu oposto $-X$ com igual probabilidade. Isso implica que $E[\mathrm{Re}\, A(k)]$, para um trecho de DNA escolhido aleatoriamente, é zero. Um raciocínio análogo justifica a suposição de que $E[\mathrm{Im}\, A(k)]$ também é zero. Estas suposições são confirmadas pelos histogramas da seção 7.1.2.

Este raciocínio não se aplica ao coeficiente $A(0)$: nesse caso os pesos $c_0(t)$ são todos iguais e positivos, e os pesos $s_0(t)$ são todos nulos. Verifica-se que $\mathrm{Re}\, A(0)$ é o número de A's menos o número de T's no trecho, e $\mathrm{Im}\, A(0)$ é o número de C's menos o número de G's, ambos divididos por $\sqrt{N}$. Mas sabe-se que, em trechos de DNA obtidos de seres vivos, há diferenças pequenas mas significativas entre as freqüências esperadas das quatro bases. Portanto, o valor médio dos coeficientes $\mathrm{Re}\, A(0)$ e $\mathrm{Im}\, A(0)$ não é zero, mas sim

$$E[\mathrm{Re}\, A(0)] = \sqrt{N}(\Pr(A) - \Pr(T)) \tag{7.3}$$

$$E[\mathrm{Im}\, A(0)] = \sqrt{N}(\Pr(C) - \Pr(G)) \tag{7.4}$$

Na amostra de DNA que usamos para nossas análises, os valores médios observados são $E[\mathrm{Re}\, A(0)] \approx 0.010$ $E[\mathrm{Im}\, A(0)] \approx -0.210$.

7.1.5 Variâncias dos coeficientes de Fourier

Devido às dependências complicadas que sabidamente existem entre bases próximas em genomas reais, não é possível estimar teoricamente a variância dos coeficientes $\mathrm{Re}\, A(k)$ e $\mathrm{Im}\, A(k)$. (Vale lembrar, aliás, que a justificativa para estudar os coeficientes $A(k)$ em vez das amostras $a(t)$ foi justamente eliminar essas dependências.) Essas variâncias precisam ser portanto estimadas experimentalmente, a partir de um conjunto suficientemente grande de trechos de DNA $a[i](t)$, $i \in \{0..M-1\}$, como os descritos acima.

Entretanto, observamos que $c_k(t) = c_{-k}(t)$, e $s_k(t) = -s_{-k}(t)$, para todo k e todo t. Portanto, podemos supor que, para trechos aletórios de DNA, $V[\mathrm{Re}\, A(k)] = V[\mathrm{Re}\, A(-k)]$, e também $V[\mathrm{Im}\, A(k)] = V[\mathrm{Im}\, A(-k)]$. Esta observação nos autoriza a tratar os M valores $\mathrm{Re}\, A[i](k)$ e os M valores $\mathrm{Re}\, A[i](-k)$, para fins de estimativa da variância, como se fossem $2M$ amostras de uma única variável aleatória – o que melhora um pouco a precisão da estimativa.

Observamos também que um deslocamento cíclico de uma seqüência x por $N/(4k)$ bases tem o efeito de multiplicar o coeficiente $X(k)$ por $\mathbf{i}$. Como as seqüências analisadas começam em pontos aleatórios do genoma, esta observação nos autoriza a supor que $V[ImA(k)] = V[ReA(k)]$ para todo k.

Em conclusão, para $0 < k < N/2$, estimamos as variâncias dos coeficientes de Fourier pelas equações

$$V[ReA(k)] = V[ReA(-k)] = V[ImA(k)] = V[ImA(-k)] = \frac{1}{4M}\sum_{i=0}^{M-1}(\mathrm{Re}\, A(k))^2 + (\mathrm{Re}\, A(-k))^2 + (\mathrm{Im}\, A(k))^2 + (\mathrm{Im}\, A(-k))^2 = \frac{1}{4M}\sum_{i=0}^{M-1}|A(k)|^2 + |A(-k)|^2 \qquad (7.5)$$

onde $\mathrm{Op} \in \{\mathrm{Re}, \mathrm{Im}\}$. Para $k = 0$, o argumento acima não vale, portanto devemos calcular $V[ImA(k)]$ e $V[ReA(k)]$ separadamente. Como as médias $E[\mathrm{Re}\, A(k)]$ e $E[\mathrm{Im}\, A(k)]$ não são nulas, devemos usar as equações

$$V[\mathrm{Op}\, A(k)] \approx \frac{1}{M-1}\sum_{i=0}^{M-1}(\mathrm{Op}\, A(k) - E[\mathrm{Op}\, A(k)])^2 \qquad (7.6)$$

Note que o denominador é $M-1$ em vez de M, para compensar o fato de que, neste caso, as médias são estimadas a partir das mesmas amostras. Finalmente, se N é par, as variâncias para a freqüência $k = N/2$ (cujo coeficiente $A(k)$ não tem parceiro $A(-k)$) deve ser estimada por

$$V[\mathrm{Re}\, A(k)] = V[\mathrm{Im}\, A(k)] \approx \frac{1}{2}\sum_{i=0}^{M-1}(\mathrm{Re}\, A(k))^2 + (\mathrm{Im}\, A(k))^2 = \frac{1}{2}\sum_{i=0}^{M-1}|A(k)|^2 \qquad (7.7)$$

7.2 Análise espectral

As figuras 7.3 e 7.4 mostram os espectros de potência de trechos de DNA, de dois organismos, extraídos do GenBank [11]. Ambos trechos possuem 4096 bases.

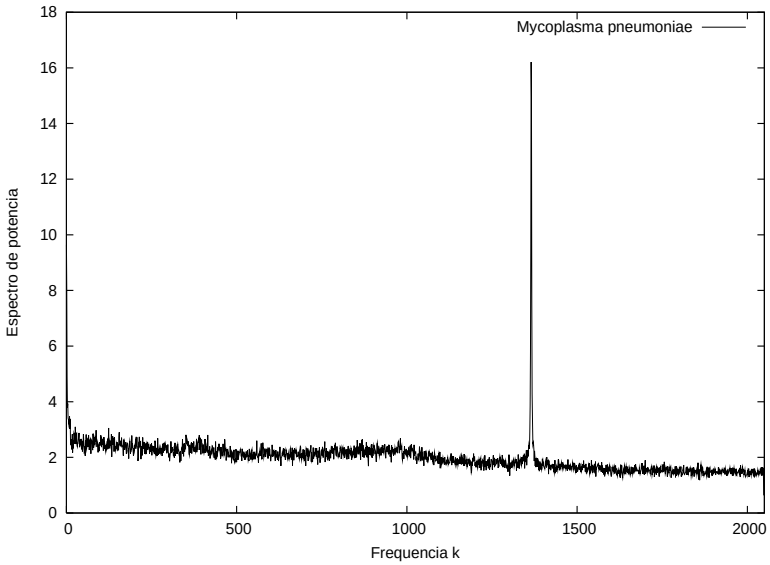

Figura 7.3: Espectro de potência de um trecho de DNA, com 4096 bases, do organismo procarioto *Mycoplasma pneumoniae*

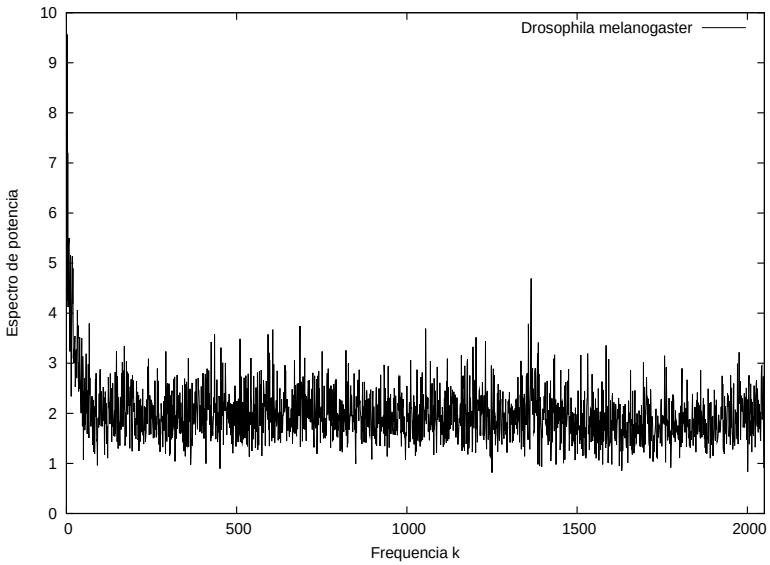

Figura 7.4: Espectro de potência de um trecho de DNA, com 4096 bases, do organismo eucarioto *Drosophila Melanogaster*

Observando o espectro de potência do DNA do procarioto, notamos um pico evidente localizado na freqüência $k = N/3$, que corresponde ao componente espec-

tral com período 3. Este pico deve-se à distribuição não-aleatória dos códons, uma característica de regiões codificadoras observada por Fickett [10]. Tiwari [29] observou este pico na análise espectral do DNA de vários organismos, examinando regiões codificadoras e não-codificadoras. Em seguida, Anasstassiou [2] utilizou este conhecimento para detectar as regiões codificadoras de proteínas nos genomas.

O pico em $k = N/3$ é muito mais fraco no espectro de potência do DNA do eucarioto, porque no DNA dos eucariotos as regiões codificadoras são interrompidas por numerosos segmentos não codificadores.

Com exceção do pico no espectro de potência dos procariotos, ambos espectros de potência são aproximadamente planos, com um ligeiro declínio (de aproximadamente 2.3 para aproximadamente 1.8) de $k = 0$ até $k = \lfloor n/2 \rfloor$. Esta queda indica que o DNA possui um caráter fractal como observado por Voss [30].

Capítulo 8

Método proposto

Quando procuramos os homólogos de uma dada seqüência num banco de genomas sempre existe a possibilidade de obtermos falsos positivos — trechos que não são homólogos mas que se assemelham tanto quanto os homólogos à seqüência dada.

Para verificarmos se um trecho de DNA é longo o bastante para que possamos localizar seus trechos homólogos sem muitos falsos positivos, precisamos avaliar quanta informação esse trecho contém sobre seus trechos homólogos.

Neste capítulo, descrevemos um método para medir a quantidade de informação mútua entre seqüências de DNA homólogas. O método é uma adaptação da técnica proposta em 1999 por Leitão e Stolfi [19]. Nesse trabalho, o método foi utilizado para determinação da quantidade média de informação, contida em um pedaço de contorno de fragmento, que podia ser usada para identificar fragmentos adjacentes ao mesmo.

8.1 Dados de entrada

Para aplicar esse método é necessário adquirir um conjunto de pares de trechos homólogos de DNA $a[i]$, $b[i]$, da população de interesse, sem repetições. As

seqüências de DNA podem ser adquiridas em bases de dados públicas que contêm genomas sequenciados de diversos organismos. Entre estas bases de dados, citamos como exemplo, o GenBank [11]. O método supõe que as duas cadeias de cada par possuem todas o mesmo tamanho N, com bases homólogas alinhadas.

Essas cadeias devem ser convertidas para sinais numéricos como explicado na seção 6.2. Espaços em branco introduzidos pelo alinhamento devem ser codificados como zeros $(0 + 0\mathbf{i})$.

8.2 Sinais genômicos como mensagens

A codificação das seqüências genômicas que as transforma em sinais complexos permite-nos tratar uma seqüência de DNA abstratamente como um sinal corrompido por um ruído (diferenças entre bases causadas por mutações e erros de leitura). Especificamente, duas seqüências homólogas, a e b, podem ser escritas como $a(n) = s(n) + q'(n)$ e $b(n) = s(n) + q''(n)$, onde s é a seqüência ancestral, e q', q'' são as funções "ruído" que representam mutações ou perdas de bases.

Seja A_k, B_k, S_k, Q'_k, e Q''_k coeficientes de Fourier de a, b, s, q', e q'', respectivamente. Verificamos no capítulo 7 que as partes real e imaginária dos coeficientes de A_k, B_k são variáveis aleatórias independentes com distribuição gaussiana. Vamos supor que o mesmo vale para S_k, Q'_k e Q''_k. Por simetria, podemos supor que $\mathrm{Re}\, Q'_k$, $\mathrm{Im}\, Q'_k$, $\mathrm{Re}\, Q''_k$ e $\mathrm{Im}\, Q''_k$ têm a mesma variância $\hat{Q}_k$. Então, pela equação (4.14), a informação dada por cada coeficiente $\mathrm{Op}\, A_k$ sobre o coeficiente correspondente $\mathrm{Op}\, B_k$, para $\mathrm{Op} \in \{\mathrm{Re}, \mathrm{Im}\}$, é

$$I(\mathrm{Op}\, A_k \to \mathrm{Op}\, B_k) \;=\; \frac{1}{2} \log \left[\frac{(\hat{S}_k + \hat{Q}_k)^2}{(2\hat{S}_k + \hat{Q}_k)\hat{Q}_k} \right] \tag{8.1}$$

8.3 Estimativa das variâncias

Na prática, nós não temos informação direta sobre a variância do sinal original $\hat{S}_k$ (a seqüência genômica do ancestral mútuo) ou do ruído $\hat{Q}_k$ (a diferença entre

as seqüências causada por mutações, inserções ou remoções de bases). Procedemos portanto da seguinte forma.

A partir das seqüências $a[i]$ e $b[i]$ calculamos as seqüências $m[i] = \frac{1}{2}(a[i] + b[i])$ (*sinal médio*) e $d[i] = a[i] - b[i]$ (*sinal diferença*). Sejam M_k e D_k os respectivos coeficientes de Fourier para cada freqüência k. Temos

$$M_k = \frac{S_k + Q'_k + S_k + Q''_k}{2} = S_k + \frac{Q'_k + Q''_k}{2} \tag{8.2}$$

$$D_k = (S_k + Q'_k) - (S_k + Q''_k) = Q'_k - Q''_k \tag{8.3}$$

Portanto,

$$\hat{M}_k = V[\mathrm{Op}\, M_k] = V\left[\mathrm{Op}\, S_k + \mathrm{Op}\, \frac{Q'_k + Q''_k}{2}\right] = \hat{S}_k + \frac{1}{2}\hat{Q}_k \tag{8.4}$$

$$\hat{D}_k = V[\mathrm{Op}\, D_k] = V[\mathrm{Op}\, Q'_k - \mathrm{Op}\, Q''_k] = 2\hat{Q}_k \tag{8.5}$$

De onde segue que as variâncias $\hat{S}_k$ e $\hat{Q}_k$ podem ser obtidas pelas equações

$$\hat{S}_k = \hat{M}_k - \frac{1}{4}\hat{D}_k \tag{8.6}$$

$$\hat{Q}_k = \frac{1}{2}\hat{D}_k \tag{8.7}$$

8.4 Cálculo da informação mútua por componente

Pela equação (8.1), a quantidade de informação contida na variável $\mathrm{Op}\, A_k$ sobre a variável correspondente $\mathrm{Op}\, B_k$ da cadeia homóloga é

$$I(\mathrm{Op}\, A_k \to \mathrm{Op}\, B_k) = \frac{1}{2}\log\left[\frac{(\hat{A}_k)^2}{\left(2(\hat{M}_k - \frac{1}{4}\hat{D}_k) + \frac{1}{2}\hat{D}_k\right)\left(\frac{1}{2}\hat{D}_k\right)}\right] = \frac{1}{2}\log\left[\frac{(\hat{A}_k)^2}{\hat{M}_k\hat{D}_k}\right] \tag{8.8}$$

Como $\hat{A}_k = \hat{S}_k + \hat{Q}'_k$, a equação (8.8) pode ser reescrita como

$$I(\mathrm{Op}\, A_k \to \mathrm{Op}\, B_k) = \frac{1}{2}\log\left[\frac{\hat{M}_k}{\hat{D}_k} + \frac{1}{2} + \frac{1}{16}\frac{\hat{D}_k}{\hat{M}_k}\right] \tag{8.9}$$

É conveniente para fins de interpretação combinar a quantidade de informação contida nas partes real e imaginária das componentes de freqüência k e $-k$. Portanto

se $0 < k < N/2$ definimos $I_k = I(\mathrm{Re}\, A_k \to \mathrm{Re}\, B_k) + I(\mathrm{Im}\, A_k \to \mathrm{Im}\, B_k) + I(\mathrm{Re}\, A_{-k} \to \mathrm{Re}\, B_{-k}) + I(\mathrm{Im}\, A_{-k} \to \mathrm{Im}\, B_{-k})$. Entretanto se $k = 0$ e $k = N/2$ como as componentes de freqüência k e $-k$ coincidem, definimos $I_k = I(\mathrm{Re}\, A_k \to \mathrm{Re}\, B_k) + I(\mathrm{Im}\, A_k \to \mathrm{Im}\, B_k)$. Temos portanto

$$I_k = \begin{cases} \log\left[\dfrac{\hat{M}_k}{\hat{D}_k} + \dfrac{1}{2} + \dfrac{1}{16}\dfrac{\hat{D}_k}{\hat{M}_k}\right] & \text{se } k = 0 \text{ ou } k = N/2 \\ 2\log\left[\dfrac{\hat{M}_k}{\hat{D}_k} + \dfrac{1}{2} + \dfrac{1}{16}\dfrac{\hat{D}_k}{\hat{M}_k}\right] & \text{se } 0 < k < N/2 \end{cases} \tag{8.10}$$

8.5 Informação mútua total

Lembrando que os coeficientes A_k, B_k são independentes, a informação total sobre b dada por a é então simplesmente

$$I_{\mathrm{tot}} = \sum_{k=0}^{\lfloor N/2 \rfloor} I_k \tag{8.11}$$

Capítulo 9

Resultados experimentais e discussão

Neste capítulo apresentamos e discutimos os resultados obtidos com o método do capítulo 8, a partir de um conjunto de trechos homólogos extraídos de vários genes de procariotos.

9.1 Descrição do experimento

9.1.1 Aquisição das cadeias homólogas

Para construir nossa amostra de cadeias homólogas, escolhemos no GenBank $H = 48$ proteínas $P[1], P[2], \ldots, P[48]$. Para cada i, escolhemos dois genes $r[i]$ e $s[i]$ de organismos procariotos diferentes que sintetizam a proteína $P[i]$. Usando o algoritmo de programação dinâmica (secão 3.1.1), contruimos o alinhamento ótimo de cada par, e extraímos um trecho aleatório do mesmo, com $N = 495$ pares de bases. Obtivemos assim dois trechos alinhados de DNA $a[i]$, $b[i]$, cada um com N bases (incluindo eventuais brancos inseridos pelo algoritmo de alinhamento).

Uma vez que genes podem conter longos trechos eliminados, inseridos ou transpostos, tomamos precauções adicionais para garantir que os trechos selecionados a e b fossem realmente homólogos: exigimos que os genes $r[i]$ e $s[i]$ tivessem compri-

mentos iguais, e que, no alinhamento ótimo das seqüências de aminoácidos houvesse pelo menos 90% de identidade.

9.1.2 Codificação numérica e cálculo das variâncias

Os H trechos homólogos alinhados $a[i]$ e $b[i]$ da amostra foram convertidos para sinais numéricos, conforme descrito na seção 8.1. Calculamos então os sinais média $m[i]$ e diferença $d[i]$, e os respectivos espectros de potência $\tilde{M}[i]$ e $\tilde{D}[i]$. A figura 9.1 mostra um desses pares, extraído dos genes que codificam a proteína *methionine aminopeptidase* nos organismos *Mycoplasma gallisepticum R* e *Mycoplasma gallisepticum*.

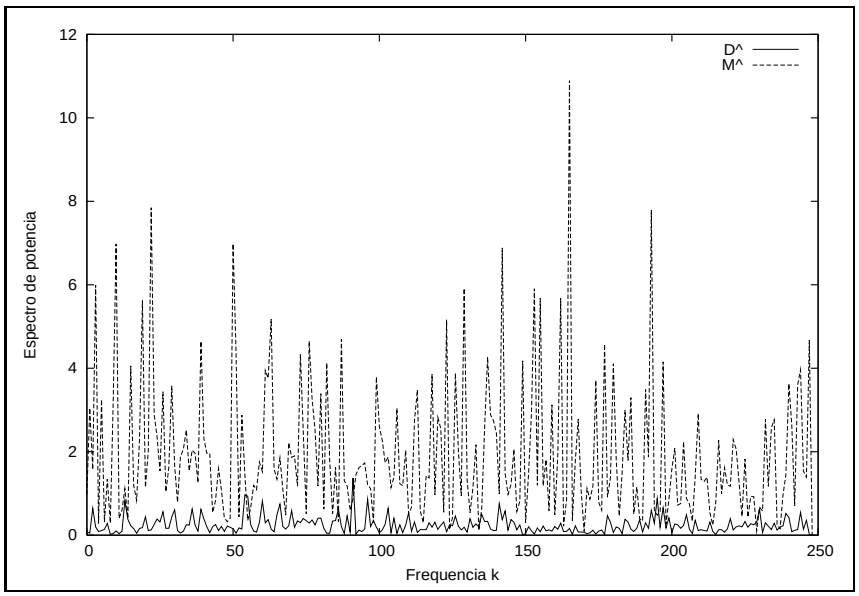

Figura 9.1: Espectros de potência $\tilde{M}[i]$ e $\tilde{D}[i]$ do sinal médio $m[i]$ e do sinal diferença $d[i]$, para um par de trechos homólogos alinhados de DNA procarioto com $N = 495$ bases.

Calculamos a seguir as médias dos espectros de potência $\tilde{M}[i]$ e $\tilde{D}[i]$, sobre todos os H pares da amostra. Os espectros médios estão ilustrados na figura 9.2. Conforme discutido na seção 8.3, extraímos destes espectros médios estimativas para as variâncias $\hat{M}_k$ e $\hat{D}_k$ de cada coeficiente de Fourier das seqüências m e d para dois trechos homólogos aleatórios.

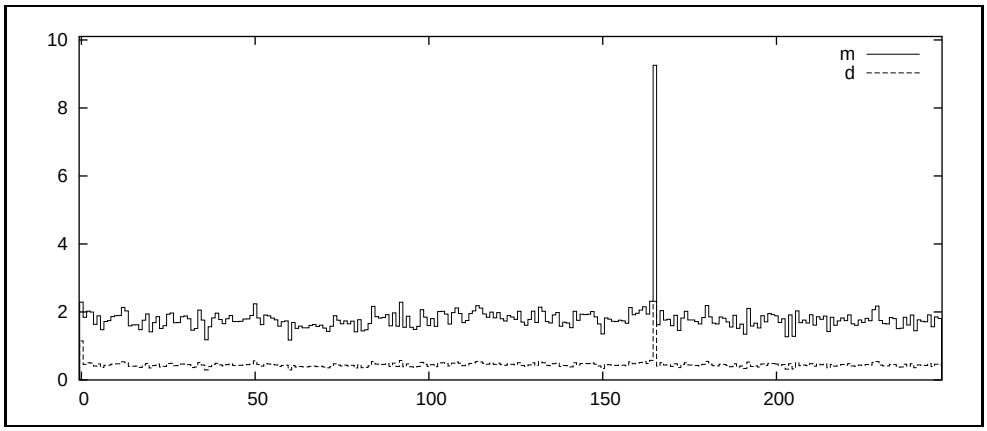

Figura 9.2: Médias dos espectros de potência $\tilde{M}[i]$ e $\tilde{D}[i]$, dos sinais $m[i]$ e $d[i]$, sobre $H = 48$ pares de trechos homólogos de DNA de procarioto com $N = 495$ bases.

9.1.3 Cálculo da informação mútua por componente

A partir das variâncias estimadas $\hat{M}_k$ e $\hat{D}_k$, calculamos a informação mútua esperada I_k para cada freqüência $k \in \{0, .. \lfloor N/2 \rfloor\}$, pela fórmula (8.8). Os valores de I_k estão ilustrados na figura 9.3 e detalhados na tabela 9.1.

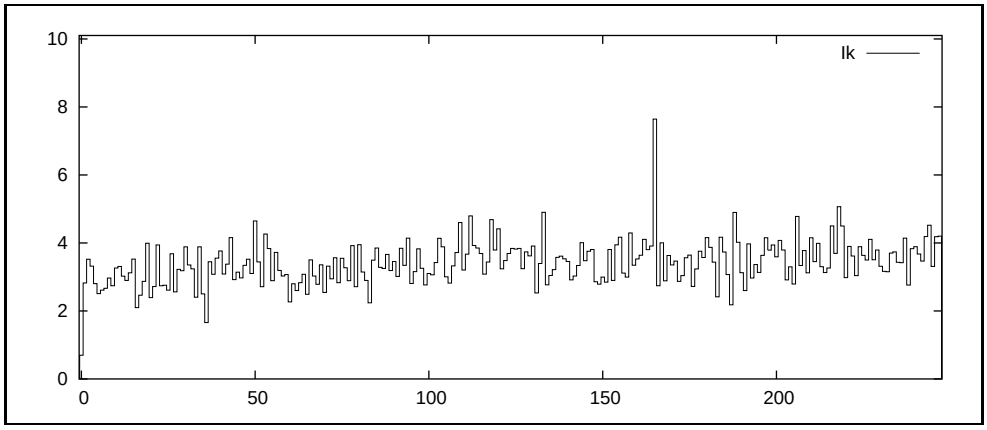

Figura 9.3: Informação mútua esperada I_k, por freqüência k, para dois trechos homólogos aleatórios de DNA de procarioto com $N = 495$ bases.

9.1.4 Cálculo da informação mútua total

Somando os valores de I_k para todas as freqüências $k \in \{0, .. \lfloor N/2 \rfloor\}$, concluímos que a informação mútua total esperada é $I_{\text{tot}} = 846$ bits — ou seja, em média $I(a \to b)/N = 1.71$ bits por base.

k	$\hat{M}_k$	$\hat{D}_k$	I_k
0	2.294	2.158	0.698
1	0.460	0.216	2.822
2	0.506	0.177	3.520
3	0.499	0.190	3.317
4	0.409	0.194	2.802
5	0.471	0.255	2.508
...	...	...	...
162	0.538	0.148	4.104
163	0.484	0.150	3.803
164	0.578	0.172	3.909
165	2.314	0.170	7.641
166	0.406	0.198	2.742
167	0.510	0.146	4.003
168	0.440	0.201	2.885
...	...	...	...
242	0.431	0.154	3.464
243	0.427	0.114	4.186
244	0.478	0.112	4.520
245	0.391	0.149	3.311
246	0.464	0.124	4.182
247	0.452	0.120	4.198
Total			845.68728

Tabela 9.1: Resultados para $\hat{M}_k$, $\hat{D}_k$ e I_k, por freqüência k.

9.2 Discussão

Uma vez que a amostra incluiu apenas DNA de procariotos, ela não se aplica a DNA de eucariotos, nem mesmo a exons homólogos, que deveriam ser objeto de análise separada.

Mesmo para procariotos, a validade do experimento foi um pouco prejudicada pelas restrições impostas na escolha dos pares de genes homólogos $r[i]$ e $s[i]$ (mesma proteína, mesmo comprimento, 90% de similaridade). Essas restricões devem ter reduzido a diferença média entre os trechos amostrados $a[i]$ e $b[i]$. Portanto o valor numérico da informação mútua (846 bits em 495 bases) é suspeito, e precisa ser confirmado com amostras mais representativas.

Vale observar que seqüências homólogas que **não** codificam a mesma proteína

(incluindo íntrons de eucariotos) devem ter informação mútua muito menor; porém, a construção de uma amostra representativa de tais pares é bastante problemática.

A figura 9.3 mostra que a informação mútua está repartida de maneira quase uniforme sobre todas as componentes do sinal genômico (aproximadamente 3.41 bits por freqüência, ou seja 1.71 bits para cada parte (Re ou Im) de cada coeficiente de Fourier (X_k ou X_{-k}). A exceção é a freqüência $k = N/3$, onde a relação entre a variância $\hat{M}_k$ do sinal médio e a variância $\hat{D}_k$ do sinal diferença é muito maior, aumentando significativamente a informação mútua entre os coeficientes A_k e B_k (para aproximadamente 7.64 bits).

Em primeira aproximação, é razoável supor que a quantidade de informação mútua I_{tot} de trechos homólogos de tamanho N é aproximadamente proporcional a N. Porém, numa análise mais precisa, esperamos observar um desvio desta regra: um trecho de tamanho $2N$ deve ter um pouco menos do que o dobro de informação de um trecho de tamanho N. A justificativa para esta hipótese é que as dependências entre bases reduzem a incerteza da segunda metade do trecho, uma vez conhecida a primeira metade. Este efeito merece ser analisado experimentalmente.

Capítulo 10

Informação mútua em múltiplas escalas

10.1 Comparação em múltiplas escalas

Na técnica de comparação em mútiplas escalas [18, 7], o banco de seqüências em questão é repetidamente "simplificado", removendo-se a cada etapa metade das amostras de cada seqüência, até que os trechos comuns a procurar sejam reduzidos a poucas amostras consecutivas. Procuram-se então trechos semelhantes nesta versão mais grosseira do banco; naturalmente serão encontrados muitos pares incorretos que, por falta de detalhes, parecem ser tão similares quanto os pares corretos. Todos estes pares são então re-comparados usando-se a versão um pouco mais fina do banco, o que permite eliminar boa parte dos pares incorretos. O processo é repetido com versões cada vez mais detalhadas, eliminando-se a cada etapa os pares menos promissores.

Para avaliar e otimizar o desempenho do método de múltiplas escalas, precisamos estimar a taxa de falsos positivos que se deve esperar na busca por similaridade de seqüências homólogas a uma seqüência dada, em cada escala.

Essa taxa depende do algoritmo de comparação, mas não pode ser menor que a taxa implicada pela quantidade de informação mútua. Seja N o tamanho do trecho dado, após redução para a escala em questão. Se, nessa escala, a informação mútua para trechos homólogos de comprimento N é q bits, isso significa que há probabilidade 2^{-q} de que uma seqüência não homóloga de tamanho N, seja indistinguível de uma seqüência homóloga, por mera coincidência.

10.2 Simplificação de seqüências

O processo de simplificação usado na comparação multi-escala é a *sub-amostragem*, que no caso consiste em reduzir um sinal a com N amostras a um sinal a' com $N/2$ amostras. (Vamos supor nesta seção que N é par.)

10.2.1 Efeito da simplificação no espectro

Segundo o teorema de Nyquist [17], se temos N amostras de um sinal, podemos apenas reconstruir no máximo N componentes de Fourier do mesmo. Verifica-se que a sub-amostragem trivial $a'(n) = a(2n)$, que toma uma amosta sim e uma não, torna idênticas as componentes de freqüencias k e $k+N/2$, para todo k. Portanto, o efeito da sub-amostragem na transformada de Fourier é fazer $A'(k) = (A(k)+A(k+N/2))/\sqrt{2}$ para todo $k \in \{0..N/2-1\}$ e eliminar as componentes de freqüência $k \in \{N/2..N-1\}$. Ou, considerando k no intervalo $-N/2 < k \leq N/2$, a sub-amostragem trivial soma as componentes de alta freqüência ($|k| > N/4$) do sinal original às de baixa freqüência ($|k| < N/4$), e elimina as de alta freqüência.

Este efeito (conhecido como *aliasing*) é indesejável para a comparação multi-escala, pois os coeficientes de Fourier de alta freqüência são muito afetados por pequenos deslocamentos do sinal. Por exemplo, se a e b forem dois trechos de N amostras de um mesmo genoma com período N, deslocados entre si por uma base, podemos verificar que $A(k) = B(k)\exp(2\pi\mathbf{i}k/N)$ – ou seja, o coeficiente de freqüência k é rodado no plano complexo por um ângulo $2\pi k/N$. Portanto, quando

k é próximo de $\pm N/2$, o deslocamento por uma única base praticamente inverte o sinal dessa componente. Como não sabemos a priori a posição inicial dos trechos homólogos, é possível que a sub-amostragem seja feita com deslocamento de 1 base num deles, modificando radicalmente o espectro e portanto tornando os sinais muito diferentes entre si.

Para ilustrar este problema, considere o sinal periódico x com $N = 8$ amostras $(1.0, 0.8, 1.0, 0.6, 1.0, 0.4, 1.0, 0.2)$, e o sinal y obtido deslocando-se x circularmente para a esquerda por 1 passo. Se aplicarmos a amostragem trivial aos dois sinais, obteremos respectivamente $x' = (1.0, 1.0, 1.0, 1.0)$ e $y' = (0.8, 0.6, 0.4, 0.2)$, que são muito diferentes entre si.

10.2.2 Filtragem do sinal genômico

Para evitar este problema, a sub-amostragem deve ser obrigatoriamente precedida de uma filtragem (ou suavização) do sinal, que elimine, ou reduza substancialmente, as freqüências altas ($|k| > N/4$). Por exemplo, se suavizarmos as seqüências x e y acima tomando a média de três amostras consecutivas com pesos $1/4$, $1/2$, e $1/4$, obteremos $x' = (0.75, 0.90, 0.85, 0.80, 0.75, 0.70, 0.65, 0.60)$ e seu deslocamento por 1 passo, que, sub-amostradas, resultam em $x' = (0.75, 0.85, 0.75, 0.65)$ e $y' = (0.90, 0.80, 0.70, 0.60)$ — muito mais semelhantes entre si.

10.2.3 Efeito da simplificação na informação mútua

Como vimos na seção 5.4.2, a suavização deve ser realizada de preferência com um filtro gaussiano. Se a duração característica do mesmo é σ, a potência da componente de freqüência k é reduzida pelo fator $|H(k)|^2 = \exp\left(-2\,(k\sigma\pi/N)^2\right)$.

Neste caso, pode-se observar que a filtragem em si não afeta a informação útil contida nas seqüências, pois ela reduz tanto o "sinal útil" $s(n)$ quanto os "ruídos" $q'(t), q''(n)$ pelo mesmo fator, e portanto não afeta a relação $\hat{S}(k)/\hat{N}(k)$. (Pode haver alguma perda de informação caso os valores filtrados sejam representados

com número finito de casas, pois o arredondamento acrescenta um ruído adicional cuja amplitude não é reduzida pela filtragem.) A perda de informação se dá na sub-amostragem, quando as componentes de freqüência alta que não foram completamente eliminadas se transformam em ruído de freqüências baixas, a amostragem introduz um deslocamento aleatório (0 ou 1) em cada trecho, e o número de componentes de Fourier é reduzido à metade. Há uma tensão portanto entre aumentar σ (o que reduz a amplitude do sinal útil) e diminuir σ (o que aumenta a amplitude do ruído criado pela sub-amostragem). Como meio-termo, adotamos $\sigma = 0.75$.

10.3 Análise experimental

10.3.1 Preparo das seqüências reduzidas

Imitando o método de comparação multi-escala, preparamos $L = 5$ versões reduzidas dos H pares de trechos homólogos $a[i]$, $b[i]$ usados no experimento do capítulo 8. Para facilitar as contas, reduzimos o comprimento N de cada trecho de 495 para $2^5 \cdot 3 \cdot 5 = 480$, o que nos permite reduzir as cadeias à metade do tamanho, exatamente, L vezes.

A primeira versão filtrada ($l = 1$) foi obtida suavizando-se o sinal original com parâmetro $\sigma_0 = 0.75$. Este valor foi escolhido de modo a otimizar a quantidade de informação mútua na escala $l = 1$.

Para cada $l \in \{2, .. L\}$, a seqüência original foi suavizada com um filtro gaussiano de desvio padrão $\sigma_l = 2^l \sigma_0$. A seqüência filtrada foi então sub-amostrada com passo 2^l, tomando-se $a'(n) = a(2^l n + t)$ onde t é uma variável aleatória inteira tal que $t/2^l \in [-1/2, +1/2]$, probabilidade uniforme.

A figura 10.1 mostra as médias $\tilde{M}_k$ e $\tilde{D}_k$ dos espectros médios de potência dos sinais m e d de todos os pares homólogos da amostra, para cada uma das escalas, e as respectivas quantidades de informação mútua I_k. Os resultados estão resumidos na tabela 10.1.

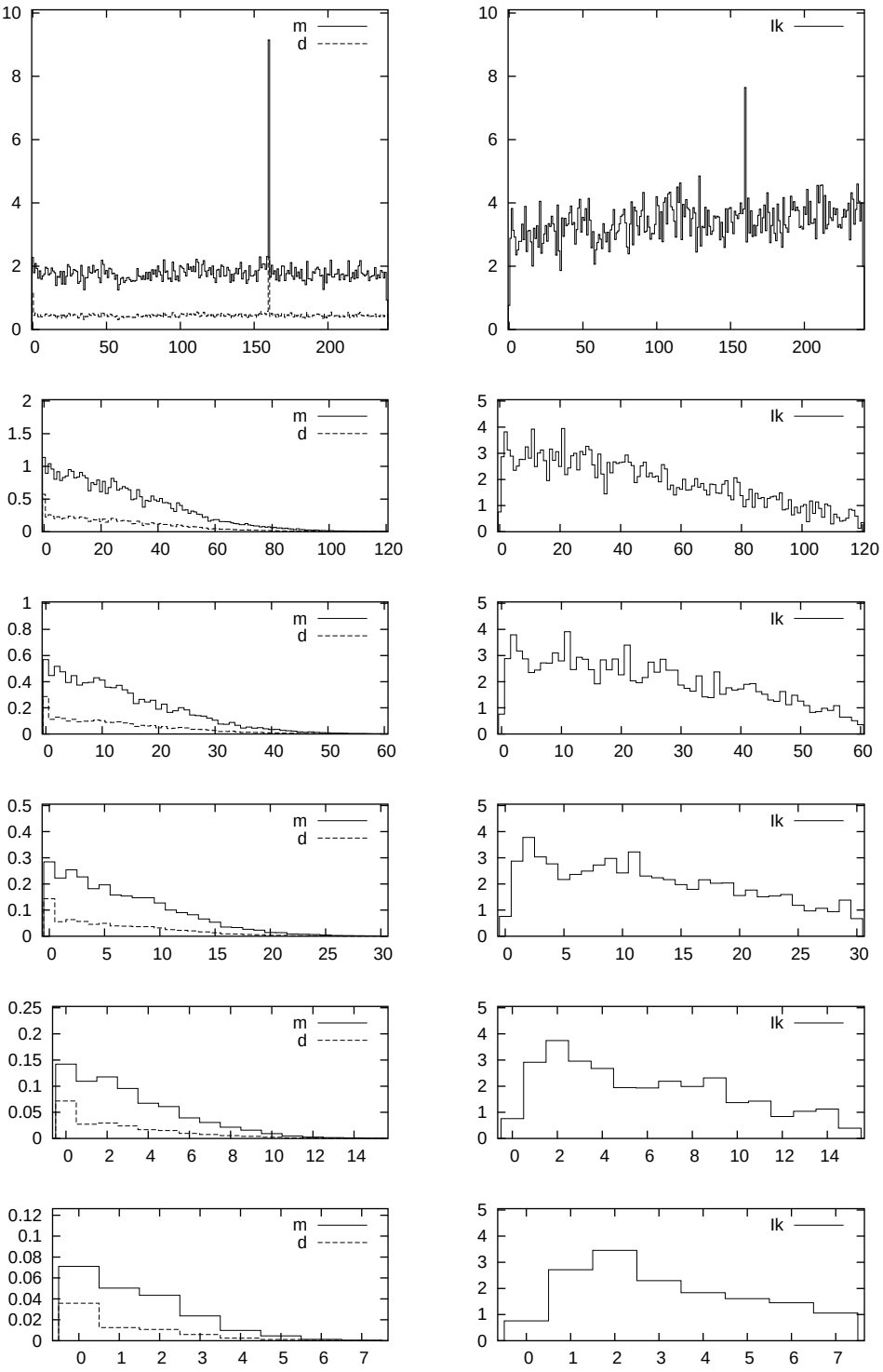

Figura 10.1: Espectros de potência médios $\tilde{M}_k$ e $\tilde{D}_k$, e quantidades de informação mútua I_k, para 48 cadeias homólogas de DNA de procarioto, filtradas e subamostradas em várias escalas.

Escala	0	1	2	3	4	5
σ	-	0,75	1,50	3,00	6,00	12,00
2^l	1	2	4	8	16	32
$N/2^l$	480	240	120	60	30	15
I_{tot}	819,4	225,4	121,2	62,4	29,6	15,2
I_{tot}/N	1,71	0,47	0,25	0,13	0,06	0,03
$I_{\mathrm{tot}}/(N/2^l)$	1,71	0,94	1,01	1,04	0,99	1,01

Tabela 10.1: Quantidade de informação mútua em seqüências homólogas de DNA de procariotos, filtradas e sub-amostradas em várias escalas.

10.4 Discussão

Para exemplificar a interpretação destes dados, vamos supor que queremos encontrar cadeias homólogas a uma cadeia dada com $N = 480$, num banco de seqüências de DNA de procariotos contendo inicialmente no total $B = 10^9$ bases. Se reduzirmos o banco como no método multi-escala, com $\sigma_0 = 0,75$, e começarmos a busca na escala $l = 5$, esperamos encontrar falsos positivos com probabilidade $p = 2^{-I_{\mathrm{tot}}} = 2^{-15,2} \approx 0,000027$. Como nessa escala temos $B/2^5 \approx 31 \times 10^6$ trechos candidatos, esperamos portanto obter $B/2^5 \times 2^{-I_{\mathrm{tot}}} \approx 830$ falsos positivos. O custo da comparação será proporcional a $(N/2^5)(B/2^5)$, ou seja 1/1024 do custo da comparação na escala original. A etapa seguinte comparará apenas as seqüências encontradas na primeira etapa — que incluem, além das seqüências homólogas, apenas os 830 falsos positivos.

Por outro lado, se começassemos a comparação diretamente na escala 4 teríamos $B/2^4 \approx 62.5 \times 10^6$ trechos candidatos, o custo seria proporcional a $(N/2^4)(B/2^4)$ (ou seja 4 vezes o da escala 5), e esperaríamos obter $(B/2^4) \times 2^{-29,6} \approx 1$ falso positivo.

Capítulo 11

Conclusão e trabalhos futuros

Neste trabalho analisamos o problema de comparação de seqüências genômicas, especificamente a quantidade de informação mútua entre seqüências homólogas de DNA, utilizando técnicas de processamento de sinais e teoria da informação. Esta análise nos permite concluir, em particular, que a técnica de múltiplas escalas é uma abordagem promissora para este problema. Mais ainda, a análise permite prever o número de falsos positivos que serão obtidos com seqüências filtradas numa determinada escala de resolução, permitindo a aplicação mais racional deste método em casos específicos.

Nossa principal contribuição está na originalidade dos resultados. Como apresentamos no capítulo 1, há diversos trabalhos tratando da informação mútua entre *nucleotídeos de uma mesma seqüência* mas até onde sabemos, a informação mútua entre *seqüências homólogas* de DNA não foi analisada em outras pesquisas.

Além disso, a mesclagem de técnicas de áreas variadas (processamento de sinais, teoria da informação e estatística) representa uma abordagem pouco explorada em biologia computacional, com grande potencial de aplicações.

Estas contribuições podem servir de base para trabalhos futuros. Citamos, em especial, o desenvolvimento de um método baseado em múltiplas escalas para a

comparação de seqüências genômicas, conforme discutimos no capítulo 1.

Num horizonte mais imediato, há muito trabalho a fazer na análise da quantidade de informação mútua entre vários tipos de bio-seqüências homólogas, incluindo genomas de eucariotos, íntrons, seqüências de aminoácidos de proteínas, etc. Para esta última tarefa seria necessário desenvolver uma codificação numérica adequada para aminoácidos.

Referências Bibliográficas

[1] Altschul, S. F., Gish, W., Millera, W., Myers, E. W., e Lipman, D. J. A basic local alignment search tool. *Journal of Molecular Biology 215* (1990), 403–410.

[2] Anasstassiou, D. Digital signal processing of biomolecular sequences. Relatório Técnico CU/EE/TR2000-20-042, Departament of Electrical Engineering. Columbia University, USA, 2002.

[3] Brown, T. A. *Genomes*, 2nd. ed. BIOS Scientific Publishers, Ltd, 2002.

[4] Cheever, E. A., Searls, D. B., Karunaratne, W., e Overton, G. C. Using signal processing techniques for dna sequence comparison. *Bioengineering Conference* (1989), 173–174.

[5] Cooley, J. W., e Tukey, J. W. *Mathematics of Computation 19* (1965), 297–301.

[6] Cormen, T. H., Lieserson, C. E., Rivest, R. L., e Stein, C. *Introduction to Algorithms.* McGraw-Hill Higher Education, 2001.

[7] da Gama Leitão, H. C. *Reconstrução Automática de Objetos Fragmentados.* PhD thesis, Instituto de Computação, Universidade de Campinas, novembro de 1999.

[8] de Souza Vieira, M. Statistics of dna sequences: A low-frequency analysis. *Physical Review E 60*, 5 (November 1999), 5932–5937.

[9] FFTW. Fastest fourier transform in the west. Biblioteca disponível em <http://www.fftw.org>.

[10] FICKETT, J. W. Recognition of protein coding regions in dna sequences. *Nucleics Acids Research 10* (1982), 5303–5318.

[11] GENBANK. Banco de dados de genomas. Disponível em <http://www.ncbi.nlm.nih.gov/entrez>.

[12] GROSSE, I., BULDYREV, S. V., STANLEY, H. E., HOLSTE, D., E HERZEL, H. Average mutual information of coding and noncoding dna. Apresentado em: Pacific Symposium on Biocomputing 5:611-620, 2000.

[13] GUSFIELD, D. *Algorithms on strings, trees, and sequences — Computer Science and Computational Biology.* Cambridge University Press, 1997.

[14] JACKSON, L. B. *Digital Filters and Signal Processing: with MATLAB exercises.* Kluwer Academic Publishers, 1995.

[15] KOONIN, E. V., E GALPERIN, M. T. *SEQUENCE - EVOLUTION - FUNCTION Computational Approaches in Comparative Genomics.* Kluwer Academic Publishers, 2003.

[16] LANDINI, G. Comunicação pessoal, setembro de 2000.

[17] LATHI, B. P. *Modern Digital and Analog Communications Systems.* Oxford University Press, 1998.

[18] LEITÃO, H. C. G., E STOLFI, J. A multiscale method for the reassembly of two-dimensional fragmented objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 9 (setembro de 2002), 1239–1251.

[19] LEITÃO, H., E STOLFI, J. Information contents of fracture lines. Relatório Técnico CU/EE/TR2000-20-042, Institute of Computing(IC), Unicamp, Brazil, 1999. 15 pages.

[20] LEITÃO, H. C. L. *Reconstrução Automática de Objetos Fragmentados.* PhD thesis, Instituto de Computação, Unicamp, 1999.

[21] Liaofu Luo, W. L., Jia, L., Ji, F., e Tsair, L. Statistical correlation of nucleotideos in a dna sequence. *Physical Review E 58*, 1 (1998), 861–871.

[22] Lipman, D. J., e Pearson, W. R. Rapid and sensitive protein similarity search. *Science 268* (1985), 1435–1441.

[23] Mandelbrot, B. B. *The Fractal Geometry of Nature*. W.H. Freeman & Company, New York, NY, USA, 1977.

[24] Meidanis, J., e Setubal, J. C. *Introduction to Computational Molecular Biology*. PWS Publishers, 1997.

[25] Otu, H. H., e Sayood, K. A divide-and-conquer approach to fragment assembly. *Bioinformatics 19*, 1 (2003), 22–29.

[26] Papoulis, A. *Probability, Random Variables, and Stochastic Process*. McGraw-Hill International Editions, 1991.

[27] Robertis, D., e Jr., D. R. *Bases da Biologia Celular e Molecular*. Guanabara Koogan, 1993.

[28] Soliman, S. S., e Srinath, M. D. *Continuous and discrete signals and systems*. Prentice-Hall International, Inc., 1990.

[29] Tiwari, S., Ramachandran, S., Bhattacharya, A., Battacharya, S., e Ramaswamy, R. Prediction of probable genes by fourier analysis of genomic sequences. *Comput Appl Biosci 13*, 3 (1997), 263–270.

[30] Voss, R. Evolution of long-range fractal correlations and $1/f$ noise in dna base sequences. *Physical Review Letters 68*, 25 (1992), 3805–3808.

[31] Wang, W., e Johnson, D. H. Computing linear transforms of symbolic signals. *IEEE Transactions on Signal processing 50*, 3 (March 2002), 628–634.