Universidade Federal Fluminense

Análise de quantidade de informação de seqüências de DNA utilizando alinhamento múltiplo

Alan do Amaral Ribeiro
Victor Hugo Simões Pinheiro Filho

Monografia apresentada ao Departamento de Ciência da Computação da Universidade Federal Fluminense como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação

Banca Examinadora

Helena Cristina da Gama Leitão (Orientador)

Teresa Cristina de Aguiar

José Raphael Bokehi

Departamento de Ciência da Computação

Niterói - Rio de Janeiro - Brasil

Julho de 2006

Monografia apresentada ao departamento de Ciência da Computação da Universidade Federal Fluminense como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação pelos alunos: Victor Hugo Simões Pinheiro Filho e Alan do Amaral Ribeiro.

Prof. Orientador: Helena Cristina da Gama Leitão

Resumo

Neste trabalho, apresentamos uma forma de calcular a quantidade de informação presente em trechos de DNA, a partir do método de alinhamento de múltiplas seqüências. Analisaremos a quantidade de informação que as seqüências homólogas e não-homólogas fornecem sobre o seu ancestral biológico. Para este objetivo, utilizamos técnicas de análise espectral e teoria da informação.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

O DNA (Ácido Desoxirribonucleico) tem sido grande objeto de estudo de pesquisadores. Atualmente, há vários projetos em andamento visando não só sequenciá-lo, mas também tentando entender como a informação genética é armazenada. Nesse sentido, é necessário o uso de computadores que nos auxiliem nestas tarefas biológicas. A Biologia Computacional é uma área interdisciplinar que consiste no desenvolvimento de modelos quantitativos para explicar fenômenos biológicos[1].

1.1 Motivação

A busca por similaridade entre trechos de DNA é essencial para entendermos o mecanismo de evolução dos genes nas espécies. Quanto mais similares são as seqüências, mais próximas estão em um dado ramo da evolução.

Comparar seqüências de DNA é um processo custoso, portanto muitas heurísticas foram desenvolvidas com intuito de localizar trechos similares em bancos de dados biológicos. Dentre as heurísticas mais utilizadas podemos citar o BLAST e o FAST[2].

1.2 Trabalhos relacionados

O cálculo da quantidade de informação é prejudicado pela correlação existente entre os nucleotídeos que compõem a cadeia de DNA. Devido a este fato, não podemos simplesmente somar a contribuição individual de cada nucleotídeo, pois não aparecem com mesma probabilidade.

Luo[3] demonstra que a correlação dos nucleotídeos é normalmente de curta distância, e o nível de correlação depende da evolução do organismo. Landini[4] mostrou que existe correlação entre bases mesmo bastante separadas entre si.

Anastassiou[5]e muitos outros abordaram a questão do reconhecimento de regiões codificadoras e não-codificadoras de proteínas, e verificaram que em uma relação $k = N/3$, sendo k a freqüência e N o tamanho da amostra, o espectro de potência apresenta um pico nesta região, dada pela distribuição não-aleatória dos códons.

Em seu trabalho, Pessoa[6] através de técnicas de processamento de sinais e teoria da informação, calcula a informação mútua que uma seqüência A fornece a respeito de uma seqüência B. Nosso trabalho é uma adaptação deste método, mas acreditamos que a abordagem de alinhamento de múltiplas seqüências tenha uma melhor aproximação na reconstrução de famílias biológicas.

Nosso objetivo com este trabalho, é contribuir com uma nova abordagem para o cálculo da quantidade de informação, que trechos de DNA fornecem a respeito de um ancestral comum, através de técnicas de processamento de sinais, estatística e teoria da informação.

1.3 Estrutura da Monografia

O trabalho foi estruturado em 9 capítulos. No capítulo 2 revisamos alguns conceitos fundamentais de biologia molecular. No capítulo 3 são apresentados algoritmos de alinhamento, parte da biologia computacional. Nos capítulos 4 e 5 são apresentadas respectivamente a codificação numérica utilizada, e conceitos sobre a análise de Fourier. Alguns fundamentos concernentes à teoria da informação são

revisados no capítulo 6. O processamento dos coeficientes de Fourier, assim como o metodologia utilizada para o cálculo da quantidade de informação está presente no capítulo 7. No capítulo 8 apresentamos e discutimos os resultados e no capítulo 9 apresentamos considerações finais e trabalhos futuros.

Capítulo 2

Fundamentos da Biologia Molecular

Neste capítulo serão apresentados conceitos fundamentais de biologia molecular, necessários ao entendimento da análise, como: genoma, DNA e sua composição, RNA, genes, proteínas e o processo de tradução e transcrição.

2.1 Células

A célula é a unidade fundamental dos seres vivos, ou a menor unidade capaz de manifestar as propriedades de um ser vivo; ela é capaz de sintetizar seus componentes, de crescer e de multiplicar-se; tais propriedades são descritas pelo *genoma*. O genoma é composto por moléculas de ácido nucléico(DNA e RNA) e funciona como um *roteiro* para regular as atividades celulares, entre elas, a fabricação de proteínas. Cada trecho do genoma que especifica uma proteína é denomidado *gene*.

As estruturas subcelulares (organelas) são comuns a muitos tipos de células. Essas organelas desenvolvem funções distintas, que no total, produzem as características de vida associada com a célula.

2.2 Eucariotos e procariotos

A organização interna de uma célula divide o mundo dos seres vivos em dois domínios de organismos: eucariotos e procariotos. Nas células dos eucariotos são encontrados compartimentos internos bem definidos como o núcleo e organelas. Como organismos eucariotos, podemos citar todos os animais e as plantas. Já as células de procariotos não possuem compartimentos internos e em particular não possuem núcleo. Como organismos procariotos, podemos citar as bactérias e as algas azuis. Nas células de organismos eucariotos o material genético reside no núcleo, e nas células dos procariotos o material genético permanece livre no citoplasma.

Ainda existe outro grupo, o dos *vírus*. O termo vírus geralmente refere-se às partículas que infectam eucariotos, enquanto o termo *bacteriófago* ou *fago* é utilizado para descrever aqueles que infectam procariotos. Tipicamente, estas partículas carregam uma pequena quantidade de ácido nucléico (seja DNA ou RNA) cercada por alguma forma de cápsula protetora consistente de proteína. Como os vírus não possuem células, ainda existe discussão se podem ser considerados seres vivos.

2.3 DNA e RNA

Todo ser vivo armazena sua informação genética na forma de moléculas de ácidos nucléicos, que por sua vez têm essa informação armazenada em um conjunto de nucleotídeos; esta informação é necessária para a célula se manter e replicar. Os nucleotídeos têm em sua composição: um fosfato, uma pentose e uma base nitrogenada que os diferencia.

Os ácidos nucléicos mais importantes são o Ácido Desoxirribonucléico (DNA) e o Ácido Ribonucléico (RNA). Nos nucleotídeos de ambos, o grupo fosfato permanece o mesmo, mas a pentose e as bases nitrogenadas são diferentes. A pentose no RNA chama-se ribose, e desoxirribose no DNA. Existem ainda, quatro tipos de bases nitrogenadas no DNA: adenina, citosina, timina e guanina (A, C, T, G), e no RNA, adenina, uracila, citosina, guanina (A, U, C, G) respectivamente. A adenina e a guanina são chamadas de bases púricas e a timina, uracila e citosina de bases

pirimídicas. Por convenção, as bases nitrogenadas é que dão nomes aos nucleotídeos.

Tabela 2.1: Tabela de Nucleotideos

Grupo	**DNA**	**RNA**
Bases Púricas	C	C
	T	U
Bases Pirimidicas	A	A
	G	G

2.3.1 Estrutura do DNA

Na estrutura dos nucleotídeos, o grupo fosfato é ligado à pentose pelo carbono 5' da mesma e a base nitrogenada pelo carbono 1'. Os nucletídeos ligam-se uns aos outros na mesma fita pelo carbono 3'. É isso que permite que uma longa fita seja construída.

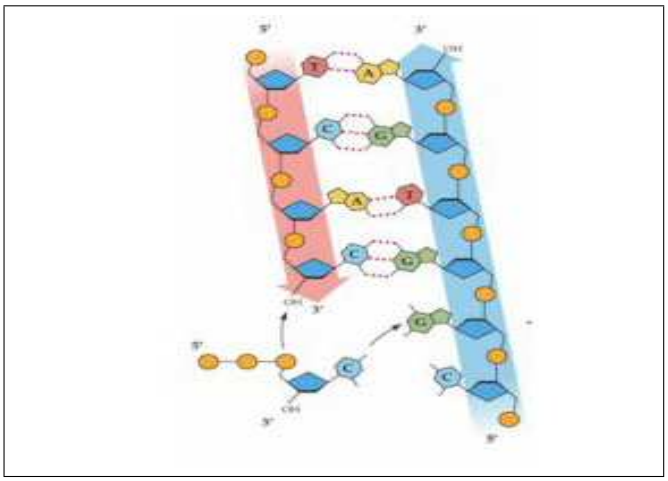

Figura 2.1: Cadeia de Nucleotídeos
Fonte: http://www.cientic.com/

Ao invés de sempre ver um diagrama molecular enorme de uma fita de DNA, o que vemos freqüentemente é uma sequência de letras, tais como $ATCTTAG$. Esta seqüência representa que bases estão em um determinado lado de uma fita de DNA. A sequência acima $ATCTTAG$ representa a fita: adenina-timina-citosina-timina-timina-adenina-guanina.

O DNA é formado por duas fitas complementares e antiparalelas. Os nucleotídeos que estão em uma fita se alinham com nucleotídeos da outra fita da seguinte forma(A com T e G com C). Isso significa que correm em sentidos opostos. Uma fita começa com 5' e termina com 3' enquanto a outra começa com 3' e termina com 5'. Por convenção a fita de sentido 5'→ 3' é colocada na parte de cima de um desenho bidimensional[7]. A figura abaixo dá um exemplo visual deste conceito e também mostra como as fitas são complementares.

```
5'  ...  A  T  A  C  T  G  G  ...  3'
         |  |  |  |  |  |  |
3'  ...  T  A  T  G  A  C  C  ...  5'
```

Figura 2.2: Estrutura do DNA

2.4 Proteínas

As proteínas são moléculas essenciais para manter a estrutura e funcionamento de todos os organismos vivos e podem ter diferentes propriedades e funções. A hemoglobina, por exemplo, é a proteína responsável pelo transporte de oxigênio no corpo.

A decodificação do genoma está baseada em trincas de nucleotídeos, chamadas *códons*, que são usados para especificar aminoácidos. Os aminoácidos são as unidades estruturais básicas das proteínas. Um aminoácido é constituído de um grupamento amina, uma carboxila, um átomo de hidrogênio e um grupamento R diferenciado, todos eles ligados à um carbono.

Combinando-se os 4 nucleotídeos em triplas obtém-se 64 combinações. Embora esse número seja superior aos 20 aminoácidos existentes, mais do que um códon pode representar um mesmo aminoácido.

Dentre os códons possíveis, três não especificam aminoácidos, e referem-se à sinais de terminação da síntese de um cadeia de aminoácidos. Esses códons são chamados de códons de parada.

Tabela 2.2: Tabela de Aminoácidos

Símbolo	**Aminoácido**	**códon**
Ala	Alanina	GCT,GCC,GCA,GCG
Cis	Cisteína	TGT,TGC
Asp	Ácido Aspártico	GAT,GAC
Glu	Ácido Glutâmico	GAA,GAG
Fen	Fenilalanina	TTT,TTC
Gli	Glicina	GGT,GGC,GGA,GGG
His	Histidina	CAT,CAC
Ile	Isoleucina	ATT,ATC,ATA
Lis	Lisina	AAA,AAG
Leu	Leucina	TTA,TTG,CTT,CTC,CTA,CTG
Met	Metionina	ATG
Asn	Asparagina	AAT,AAC
Pro	Prolina	CCT,CCC,CCA,CCG
Gln	Glutamina	CAG,CAA
Arg	Arginina	CGT,CGC,CGA,CGG,AGA,AGG
Ser	Serina	TCT,TCC,TCA,TCG,AGT,AGC
Tre	Treonina	ACT,ACC,ACA,ACG
Val	Valina	GTT,GTC,GTA,GTG
Trp	Triptofano	TGG
Tir	Tirosina	TAT,TAC
X	PARADA	UAA,UAG,UGA

2.4.1 Transcrição e Tradução

Para que a síntese da proteína ocorra, o DNA deve passar por um processo de transcrição, onde será criada por uma enzima chamada transcriptase, uma cópia dele em RNA; este RNA é chamado de RNA mensageiro (mRNA). De posse do mRNA, entra em processo na célula a etapa de tradução. Assim como o DNA, as proteínas são polímeros lineares formadas pela variação do alfabeto químico dos aminoácidos. Os aminoácidos têm além do conteúdo informativo, uma natureza físico-química, que também determina o propósito da proteína [7].

A tradução é o processo de síntese ou fabricação de proteínas (construção da cadeia de aminoácidos). Para a fabricação das proteínas é necessário que estruturas

celulares chamadas ribossomos decodifiquem a mensagem contida na molécula de mRNA para uma cadeia de aminoácidos. O processo de tradução é realizado da seguinte maneira: ao combinar-se com os ribossomos, o mRNA tem sua seqüência de códons lida, e para cada códon o respectivo tRNA é atraído até os ribossomos, e pela complementariedade de bases é feita a ligação entre o códon (do mRNA) e o anticódon (do tRNA), liberando o aminoácido carregado pelo tRNA que é então ligado à cadeia crescente do polipeptídeo. A síntese da proteína é encerrada quando os ribossomos encontram um códon de parada no mRNA.

2.4.2 Éxons e Íntrons

Os genes são formados por regiões codificadoras e não-codificadoras. As regiões codificadoras são denominadas *éxons* e não-codificadoras, *íntrons*. Nos procariotos, quase toda a seqüência de DNA é formada por éxons, o que não ocorre com os eucariotos, que possuem grandes regiões de DNA que aparentemente não apresentam funcionalidade ou possuem funcionalidade desconhecida. Nos eucariotos, quando é feita a transcrição do DNA em mRNA, apenas os éxons são copiados. Estima-se que cerca de 90% do DNA dos procariotos especifique proteínas, enquanto que nos eucariotos, a estimativa é que apenas 5% do DNA especique proteínas. Acredita-se que os íntrons nos eucariotos sejam regiões sem função, resultado da evolução das espécies[8].

2.5 Evolução

A *evolução* é o processo através no qual ocorrem as mudanças ou transformações nos seres vivos ao longo do tempo, dando origem a espécies novas, isto porque, o DNA de um organismo não é uma molécula estática, mas está freqüentemente exposto a agentes, naturais ou artificiais, que provocam modificações na sua estrutura ou composição química.

2.5.1 Mutações

As *mutações* são modificações súbitas e hereditárias no material genético. Geralmente, os organismos portadores de uma mutação em um determinado gene apresentam problemas na sua sobrevivência e diversas alterações. Todos os seres vivos sofrem um certo número de mutações, como resultado de funções celulares normais ou interações aleatórias com o ambiente. Tais mutações são denominadas *espontâneas*. A ocorrência de mutações pode ser aumentada pelo tratamento com determinados compostos. Tais compostos são denominados *agentes mutagênicos* e as modificações que eles causam *mutações induzidas*.

Adotou-se que quando a mutação envolve grandes porções do DNA, como um gene ou vários genes, ou ainda regiões de repetição, denomina-se apenas *mutação*. Quando a alteração é de uma base (que pode sofrer substituição, adição ou deleção) e o resultado é o mau funcionamento do sistema celular que replica ou repara o DNA, denomina-se *mutação de ponto* ou *mutação pontual*.

2.5.2 Homologia

A *homologia* é o estudo biológico das semelhanças entre estruturas de diferentes organismos que possuem a mesma origem embriológica. Tais estruturas podem ou não ter a mesma função. Devido às mutações, um ser vivo pode dar origem a outros, mas as novas espécies conservarão alta similiridade em trechos dos seus genomas. Por exemplo, a cadeia de nucleotídeos que codifica a hemoglobina no homem, deve ser extremamente similar à dos outros mamíferos.

Os conceitos de homologia, ortologia e paralogia são de extrema importância quando estamos interessados em comparar organismos. Considere os genes g e g' pertencentes ao genoma G e o gene h pertencente ao genoma H:

- g e h são homólogos se descendem de um mesmo ancestral comum. Neste caso dizemos que g e h são **ortólogos**;
- g e g' são homólogos, embora com funções diferentes no organismo. Neste caso dizemos que g e g' são **parálogos**;

Embora trechos homólogos não sejam *idênticos* entre si, são altamente *similares* por derivarem de um ancestral comum.

Capítulo 3

Comparação de seqüências

A comparação de seqüências é a operação primitiva mais importante em biologia computacional, servindo de base para muitas outras mais complexas [9]. Como procuramos homologia entre cadeias, determinar o quão parecidas são as seqüências nos permitirá dizer se provém de um ancestral comum.

Para isso, será utilizada uma técnica conhecida como *alinhamento*, que pode ser definido como a inserção de espaços arbitrários de forma que as seqüências possuam o mesmo tamanho e fiquem o mais similares possíveis [9].

Serão apresentadas duas formas de alinhamento: *comparação de duas seqüências* e *comparação de várias seqüências*.

3.1 Comparação de duas seqüências

Nesta forma de comparação, teremos apenas duas seqüências envolvidas. Por exemplo, as cadeias S_1 = ACTG e S_2 = ATG, podem ser alinhadas da seguinte forma

S_1	A	C	T	G
S_2	A	-	T	G

Figura 3.1: Alinhamento de cadeias

Portanto, para acharmos o melhor alinhamento precisamos calcular a melhor pontuação, chamada similaridade, através de um critério que penaliza erros e espaços, mas que premia acertos. Por padrão, os acertos em cada coluna são pontuados com +1, os erros com -1 e os espaços com -2. Logo, a similaridade das seqüências S_1 e S_2 é 1.

S_1	A	C	T	G		
S_2	A	-	T	G		
	+1	-2	+1	+1	=	+1

Figura 3.2: Pontuação de cadeias

O primeiro passo é construir uma matriz de similaridade que armazenará em uma entrada (i, j) a pontuação entre S_1[1..i] e S_2[1..j]. As primeiras linha e coluna são preenchidas com múltiplos de -2, este é o único alinhamento possível caso uma das seqüências seja vazia. O cálculo de uma entrada (i, j) é feito utilizando-se três entradas anteriores. O algoritmo possui o parâmetro g que é a penalidade para espaços, e $p(i, j)$ que é igual a 1, caso $S_1[i] = S_2[j]$ e $p(i, j) = -1$, se $S_1[i] \neq S_2[j]$. Ao fim desta etapa, a similaridade entre as duas seqüências estará na posição $a[m, n]$. Este algoritmo baseia-se no método da programação dinâmica[10] e está descrito na figura 3.3.

De posse da matriz, o alinhamento pode ser realizado partindo da posição a[m,n] e recursivamente percorrer cada coluna, sempre tomando um novo ponto de partida, até chegar na posição a[0,0]. Na volta da recursão, o algoritmo retorna pelas suas posições anteriores gerando o alinhamento ótimo. Cada posição percorrida é uma coluna do alinhamento. Dependendo se o deslocamento será na horizontal, vertical ou diagonal, isto equivale à inserção de espaços em uma das seqüências ou à correspondência das seqüências em uma dada posição.

Ao final da execução do algoritmo da figura 3.4 teremos as seqüências s' e t' completamente alinhadas. Com isso, podemos saber se em uma dada posição i, t' difere de s'.

Este algoritmo produz o alinhamento ótimo entre duas seqüências e seu custo computacional é $O(mn)$, onde m e n são os tamanhos das seqüências sendo com-

```
procedure SIMILARIDADE(S1, S2, g)                    ▷ As entradas do algoritmo
    m ← |S1|
    n ← |S2|
    for i ← 0, m do
        a[i, 0] ← i * g
    end for
    for j ← 0, n do
        a[0, j] ← j * g
    end for
    for i ← 0, m do
        for j ← 0, n do
            a[i, j] ← max(a[i − 1, j] + g, a[i − 1, j − 1] + p(i, j), a[i, j − 1] + g)
        end for
    end for
    return ← a[m, n]
end procedure
```

Figura 3.3: Algoritmo de Similaridade

paradas. Se m e n tiverem mesmo tamanho teremos $O(n^2)$. Como veremos, este algoritmo torna-se inviável para múltiplas seqüências.

3.2 Comparação de várias seqüências

A definição de alinhamento múltiplo entre seqüências é uma generalização natural do alinhamento entre duas seqüências [11]. O algoritmo exato para a solução do problema, ou seja, aquele que encontra o alinhamento de maior valor, consiste em uma extensão do algoritmo exato de programação dinâmica para o alinhamento entre duas seqüências, porém este algoritmo se aplica apenas para seqüências pequenas, devido ao tempo e espaço gastos em sua computação pois, para um conjunto de k seqüências a serem alinhadas, onde a maior delas possui tamanho n, o custo do algoritmo seria $O(n^k)$.

Neste caso, será necessária a utilização de uma heurística, que embora não garanta o resultado ótimo, pode nos dar um bom resultado com custo menor. A heurística utilizada será a do *Alinhamento Estrela*.

```
procedure ALINHAMENTO(i, j, a)                    ▷ As entradas do algoritmo
    if i = 0 and j = 0 then
        tam ← 0
    else if i > 0 and a[i, j] = a[i − 1, j] + g then
        Alinhamento(i − 1, j, tam)
        tam ← tam + 1
        s'[tam] ← s[i]
        t'[tam] ← −
    else if i > 0 and j > 0 and a[i, j] = a[i − 1, j − 1] + p(i, j) then
        Alinhamento(i − 1, j − 1, tam)
        tam ← tam + 1
        s'[tam] ← s[i]
        t'[tam] ← t[j]
    else
        Alinhamento(i, j − 1, tam)
        tam ← tam + 1
        s'[tam] ← −
        t'[tam] ← t[j]
    end if
end procedure
```

Figura 3.4: Algoritmo que constrói o alinhamento ótimo

No alinhamento estrela, o primeiro passo é escolher a seqüência *centro*, ou seja, será feito alinhamento ótimo dois-a-dois entre todos os pares possíveis, o centro será a que apresentar maior somatória de similaridade quando comparada com as demais. Considerando s_c como o centro e s_i e s_j como as seqüências a serem alinhadas, a escolha do centro é dada pela fórmula

$$s_c = max(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} similaridade(s_i, s_j)) \tag{3.1}$$

Uma vez obtida a seqüência centro, o alinhamento múltiplo é construído progressivamente. O processo se inicia com o alinhamento ótimo entre s_c e digamos por exemplo, s_1. Progressivamente o centro(já com buracos) é alinhado com s_2, os novos buracos precisam ser propagados para as seqüências já alinhadas, neste caso s_1 e assim sucessivamente.

O custo para obtenção da seqüência centro é de $O(k^2n^2)$ sendo n o tamanho

da maior seqüência, enquanto que o custo para a inserção de todas as seqüência no alinhamento é de $O(k^2l)$, onde l é o número máximo de colunas do alinhamento. Assim, o custo total do algoritmo é de $O(k^2n^2 + k^2l)$.

Por exemplo, consideremos as seqüências s_1, s_2, s_3 e s_4.

s_1	A	C	G	T		
s_2	A	T	C	G	T	
s_3	A	G	C	G	C	
s_4	A	A	C	G	A	T

Figura 3.5: Cadeias para alinhamento múltiplo

Para escolhermos o centro, comparamos as seqüências duas a duas e guardamos a similaridade em uma matriz dada pela figura 3.6.

	s1	s2	s3	s4	soma
s1	0	2	0	0	2
s2	2	0	1	1	4
s3	0	1	0	-1	0
s4	0	1	-1	0	0

Figura 3.6: Matriz que guarda as similaridades entre as seqüências

É realizado o somatório das similaridades da matriz linha a linha e armazenado na coluna *soma*. A linha que obtiver a maior similaridade, indica a seqüência central, neste caso s_2.

De posse da seqüência central s_2, deve ser realizado o alinhamento ótimo de s_2 com todas as outras.

```
        0  1  2  3  4
s2  =   A  T  C  G  T
        |  |  |  |  |
s1  =   A  -  C  G  T
```

```
        0  1  2  3  4
s2  =   A  T  C  G  T
        |  |  |  |  |
s3  =   A  G  C  G  C
```

```
        0  1  2  3  4  5
s2  =   A  T  C  G  -  T
        |  |  |  |  |  |
s4  =   A  A  C  G  A  T
```

Figura 3.7: Alinhamento entre as seqüências

Ao ser inserido um buraco, por exemplo, na posição 4 da seqüência central, este buraco deve ser inserido na mesma posição nas seqüências já alinhadas para manter o alinhamento de todas as seqüências. O buraco inserido em s_2 permanecerá nesta seqüência para os próximos alinhamentos. A figura 3.8 mostra o alinhamento múltiplo das seqüências.

```
s1   A  T  C  G  -  T
s2   A  -  C  G  -  T
s3   A  G  C  G  -  C
s4   A  A  C  G  A  T
```

Figura 3.8: Alinhamento de várias cadeias

Capítulo 4

Codificação numérica das seqüências de DNA

As seqüências de DNA representadas como cadeia de caracteres precisam ser codificadas em representação numérica para trabalharmos com processamento de sinais.

4.1 Técnicas de codificação de DNA

Umas das técnicas mais utilizadas é a codificação com quatro vetores[12], em que uma seqüência de DNA com comprimento N pode ser escrita como

$$x[n] = au_A[n] + tu_T[n] + cu_C[n] + gu_G[n] \qquad (4.1)$$
$$n = 0, 1, 2, ..., N-1$$

no qual $u_A[n], u_T[n], u_C[n]$ e $u_G[n]$ são indicadores binários de seqüência, que tomam o valor 1 ou 0 na posição n, dependendo se caracter correspondente existe ou não naquela posição. Por exemplo, a seqüência $ACCTG$ tem $N = 5$, $u_A[0] = 1, u_T[0] = 0, u_C[2] = 1$ e $u_C[3] = 0$.

A codificação com quatro vetores embora muito utilizada, possui um inconveniente constatado por Anastassiou[5]. Para cada n, três dos quatro vetores assumem valor 0 e um assume valor 1. Isso significa que são redundantes, pois

$$u_A[n] + u_T[n] + u_C[n] + u_G[n] = 1, para\ todo\ n \tag{4.2}$$

Se os quatro vetores são linearmente dependentes, uma solução seria utilizar apenas três vetores, mas isto poderia acarretar perda de informação. Diante deste fato, a codificação complexa nos pareceu mais apropriada.

4.2 Codificação complexa

Esta codificação foi proposta por Cheever[13] e também utilizada por Pessoa[6] em seu trabalho.

Nesta codificação os nucleotídeos A, T, C e G são convertidos para $+1, -1, +i, -i$, onde $i = \sqrt{-1}$ é a unidade imaginária. Cada base púrica é mapeada com um valor complementar a sua correspondente pirimídica.

Temos que as bases A, T, C e G são representadas respectivamente por $(+1, 0)$, $(-1, 0)$, $(0, +1)$ e $(0, -1)$. Para buracos introduzidos pelo alinhamento, devemos representá-los por $(0, 0)$. A seqüência codificada é chamada de *sinal genômico*.

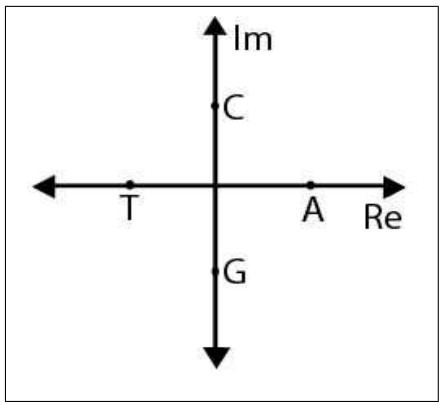

Figura 4.1: Mapeamento numérico no plano complexo

Capítulo 5

Análise de Fourier

A análise de Fourier possibilita uma representação de uma classe ampla de funções em termos de uma combinação linear de funções base senos, cossenos ou exponenciais complexos. Uma outra forma de pensar na análise de Fourier é como uma técnica matemática para transformar nossa visão de informação baseada no tempo (posição ou espaço) naquela baseada na freqüência.

5.1 Representação de Fourier

Um sinal contínuo com período T pode ser aproximado com qualquer grau de precisão desejado por uma série de Fourier finita, que é uma soma finita de senóides e cossenóides, cujos períodos dividem T:

$$f(t) = a_0 + \sum_{k=1}^{n} \left[a_k \cos\left(\frac{2\pi}{T} kt\right) + b_k \sin\left(\frac{2\pi}{T} kt\right) \right] \tag{5.1}$$

Os coeficientes a_k e b_k são números reais, os coeficientes de Fourier do sinal. Cada termo $a_k \cos(2\pi kt/T) + b_k \sin(2\pi kt/T)$ é chamado componente do sinal e o índice k é a freqüência da componente.

5.1.1 Transformada Discreta de Fourier

A Transformada Discreta de Fourier(DFT) pode ser definida pela equação

$$X(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n)\ exp\Big(-\mathbf{i}\frac{2\pi}{N} nk \Big) \tag{5.2}$$

onde k é o parâmetro que corresponde à freqüência do coeficiente $X(k)$.

A transformação pode ser vista como uma mudança de base no domínio complexo, ou seja, temos uma conversão de um sinal de N pontos no domínio do tempo em um sinal de N pontos no domínio da frequência. O sinal de entrada *tempo* contém as amostras do sinal a ser decomposto, e o sinal de saída *frequência* contêm as amplitudes dos cossenos e senos, representados respectivamente por $Re(k)$ e $Im(k)$.

A transformada discreta de fourier é utilizada sobre sinais periódicos e discretos, por isso consideramos as amostras $x(n)$ e os coeficientes $X(k)$ da equação 5.2 como periódicas, com período N.

5.2 Espectro de potência

O espectro de potência, ou potência média do sinal, é dado pela soma dos quadrados das amplitudes

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

A potência também pode ser obtida a partir da (TDF)

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

No plano complexo da transformada, temos um conjunto de vetores ψ_k; quando $k \in \{0 < k < N/2\}$, o gráfico do sinal ψ_k, forma uma espiral que dá k voltas completas em um período N. Temos ainda um conjunto de vetores ψ_k que correspondem à freqüência $-k$, que se diferencia apenas pelo sentido de rotação na espiral. Portanto na equação 5.4, os termos $A(k)$ e $A(-k)$, medem ondas de mesma freqüência k. Portanto, para fins de análise no sinal, é necessário somar essas duas componentes.

O espectro de potência do sinal complexo de período N é dado pela equação

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

Capítulo 6

Teoria da informação

Neste capítulo serão apresentados alguns conceitos básicos sobre teoria da informação, necessários para o cálculo de informação mútua entre seqüências de DNA.

6.1 Probabilidade

A probabilidade pode ser definida como o número de ocorrências de cada símbolo X, dada por $t(X)$, dividido pelo número total de símbolos, dado por M.

$$p(X) = \frac{t(X)}{M}$$

6.1.1 Probabilidade condicional

Dadas duas variáveis aleatórias quaisquer X e Y, sendo $P(Y) > 0$, definimos a *probabilidade condicional* de X, dado Y, como sendo

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \tag{6.1}$$

6.1.2 Valor Esperado

O valor esperado de uma variável aleatória X que assuma os valores $x_1, x_2, ..., x_n$ é dado por

$$E(X) = \sum_{i=1}^{n} x_i p_i \tag{6.2}$$

o valor esperado também pode ser estimado pela média aritmética

$$E(X) = \frac{1}{N} \sum_{i=1}^{N} X[i] \tag{6.3}$$

a variância de uma variável aleatória é definida por

$$V(X) = E(X^2) - E(X)^2 \tag{6.4}$$

e o desvio padrão dado pelo símbolo σ é definido como

$$\sigma(X) = \sqrt{V(X)} \tag{6.5}$$

6.2 Quantidade de informação

A *quantidade de informação* está relacionada com a probabilidade de ocorrência de um dado símbolo na transmissão de mensagens[14], ou seja, quanto menor a probabilidade de que um símbolo X apareça, maior será a quantidade de informação que teremos quando este símbolo aparecer.

Sendo a informação dada por $I(p)$ e p, como a probabilidade de ocorrência de um símbolo ou mensagem, teremos $\lim_{p\to 1} I(p) = 0$ e $\lim_{p\to 0} I(p) = \infty$. Logo, se eventos muito prováveis carregam pouca informação e improváveis carregam muita informação, a quantidade de informação pode ser dada pela fórmula

$$I(p) \simeq \log \frac{1}{p} \tag{6.6}$$

Pode-se usar o logaritmo em qualquer base para se calcular a quantidade de informação, desde que se use a mesma base para todos os cálculos. A base usada

determina a unidade em que se mede a quantidade de informação. Se for usada a base 2, a unidade é o bit.

6.3 Entropia Informacional

Uma outra grandeza importante na teoria da informação é a *entropia informacional*, ou *entropia de Shannon*. Ela é definida como a média da quantidade de informação contida em um conjunto de símbolos $x_1, ..., x_n$ com probabilidade $p_1, ..., p_n$. A entropia de uma variável aleatória X é

$$H(X) = \sum_{i=1}^{n} p_i \left(\log_2 \frac{1}{p_i} \right) \tag{6.7}$$

6.3.1 Entropia de Variável gaussiana

Os processos aleatórios independentes igualmente prováveis costumam se agrupar em uma distribuição chamada de *normal* (curva na forma de sino). A distribuição de Gauss originalmente serve para mostrar como se distribuem os erros em uma medida experimental. Porém, pode também mostrar como se distribuem os dados em várias situações originadas de eventos mutuamente independentes. Essa distribuição é dada como

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}}\, exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{6.8}$$

sendo μ o valor esperado da distribuição. A entropia para distribuição normal de uma variável aleatória X é dada por

$$H(X) = \frac{1}{2} \log_2(2\pi e V(X)) \tag{6.9}$$

onde V(X) é a variância de X.

6.3.2 Entropia condicional e informação mútua

Podemos definir como entropia condicional, o valor da entropia de uma variável aleatória X, dado que conhecemos o valor de Y. A partir da entropia condicional, podemos calcular a quantidade de informação que X fornece a respeito de Y.

$$H(X|Y) = \sum_{i=1}^{n} H(X|Y = y_i)p(Y = y_i) \tag{6.10}$$

A partir da equação 6.10 e das propriedades de Shannon [14], temos

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y|X) \\ I(X,Y) &= H(X) - H(X|Y) \end{aligned} \tag{6.11}$$

De posse das equações 6.10 e 6.11, concluímos que a informação que X fornece a respeito de Y é a mesma que Y fornece a respeito de X. Por isso, o termo *informação mútua*.

6.4 Informação mútua de variáveis normais

Em nossa análise podemos considerar as seqüências homólogas A e B como sinais e descendentes de um mesmo ancestral, ou uma mesma fonte transmissora S. Como A e B sofreram mutações, podemos dizer que foram corrompidas por ruídos R e Q. Desta forma temos que $A = S + Q$ e $B = S + R$. Segundo Pessoa[6], podemos considerar S, Q e R como variáveis independentes e com variâncias $\hat{S}$, $\hat{Q}$ e $\hat{R}$, o que nos daria $\hat{A} = \hat{S} + \hat{Q}$ e $\hat{B} = \hat{S} + \hat{R}$.

A equação 6.9 é reescrita como

$$H(B) = \frac{1}{2}\log_2(2\pi e\hat{B}) = \frac{1}{2}\log_2[2\pi e(\hat{S} + \hat{R})] \tag{6.12}$$

Ainda segundo o mesmo raciocínio[6], a entropia condicional pode ser reduzida a

$$H(B|A) = \frac{1}{2}\log_2\left[2\pi e\left(\frac{\hat{S}\hat{Q}}{\hat{A}} + \hat{R}\right)\right] \tag{6.13}$$

Realizando as substituições das equações 6.12 e 6.13 na equação 6.11 teremos a informação que A fornece a respeito de B.

$$I(A,B) = \frac{1}{2}\log_2\left[\frac{(\hat{S}+\hat{Q})(\hat{S}+\hat{R})}{\hat{S}(\hat{Q}+\hat{R})+\hat{Q}\hat{R}}\right] \tag{6.14}$$

Para o caso $R = 0$, ou seja $B = S$, a equação 6.14 se reduz a

$$I(A,S) = \frac{1}{2}\log_2\left(\frac{\hat{A}}{\hat{Q}}\right) = \frac{1}{2}\log_2\left(\frac{\hat{S}+\hat{Q}}{\hat{Q}}\right) \tag{6.15}$$

Capítulo 7

Metodologia

Neste capítulo apresentamos o método utilizado para o cálculo da quantidade de informação. Este método é uma adaptação da técnica utilizada por Pessoa[6].

7.1 Estimativa sobre sinal genômico

A transformada de Fourier aplicada ao sinal genômico nos permite superar o obstáculo da correlação entre os nucleotídeos, pois os coeficientes de Fourier real e imaginário são variáveis aleatórias e independentes com distribuição normal, conforme observado experimentalmente[6].

A variância sobre os coefientes de Fourier também é verificada experimentalmente. Quando $k = 0$, a variância é dada pela fórmula

$$\begin{aligned} V[ReA(k)] &= \frac{1}{M}\sum_{i=1}^{M-1}(ReA(k) - E(ReA(k)))^2 \\ V[ImA(k)] &= \frac{1}{M}\sum_{i=1}^{M-1}(ImA(k) - E(ImA(k)))^2 \end{aligned} \tag{7.1}$$

para $0 < k < N/2$, temos que o valor esperado para os coeficientes é zero, e que a variância $V[ReA(k)] = V[ImA(k)]$. Logo a variância é dada pela fórmula

$$V[A(k)] = \frac{1}{4M}\sum_{i=1}^{M-1}(ReA(k))^2 + (ImA(k))^2 + (ReA(-k))^2 + (ImA(-k))^2 \tag{7.2}$$

se N é par, a variância para $k = N/2$, que não possui par $A(-k)$ é dada por

$$V[A(k)] = \frac{1}{2M} \sum_{i=1}^{M-1} (ReA(k))^2 + (ImA(k))^2 \tag{7.3}$$

Notamos pelas equações acima, que a variância é dada pelo espectro de potência médio do sinal.

7.2 Relações entre variáveis

O primeiro passo é extrair de um banco de dados biológico[15] as seqüências a serem utilizadas. As seqüências são divididas em grupos e então alinhadas conforme explicado na seção 3.2. Posteriormente, devemos codificá-las numericamente como proposto na seção 4.2.

7.2.1 Média e Ruído

Conforme explicado no capítulo 6, as cadeias de DNA podem ser consideradas sinais corrompidos por ruídos. A equação 6.15 nos diz a informação que uma mensagem corrompida A fornece sobre o ancestral, ou a mensagem original S.

Desta forma, precisamos estimar a variância do ancestral e do ruído procedendo da seguinte forma: para cada grupo de seqüências, calculamos a média entre elas e a diferença de cada seqüência em relação à média.

$$m = \frac{1}{L} \sum_{i=1}^{L} s_i \tag{7.4}$$

$$d = s_i - m \tag{7.5}$$

A partir das equações 7.4 e 7.5 deduzimos a variância dos coeficientes de Fourier da seqüência ancestral $\hat{S}_k$ e do ruído $\hat{Q}_k$. Consideramos a diferença entre a média e a seqüência ancestral como o somatório de todos os ruídos divididos pelo número de seqüências. Desta forma, estimamos $\hat{S}_k$.

$$M_k - S_k = \frac{1}{L}\sum_{i=1}^{L} Q_k \tag{7.6}$$

$$\hat{M}_k - \hat{S}_k = \frac{1}{L}\hat{Q}_k \tag{7.7}$$

$$\hat{S}_k = \hat{M}_k - \frac{1}{L}\hat{Q}_k \tag{7.8}$$

tendo $\hat{M}_k$ como a variância da média. A variância do ruído é estimada a partir da equação 7.5. Consideramos tanto a média m como a seqüência s_i corrompida por ruídos.

$$D_k = S_k + Q_k - \left[\frac{1}{L}\sum_{i=1}^{L}(S_k + Q_k)\right] \tag{7.9}$$

$$\hat{D}_k = \left(1 - \frac{1}{L}\right)^2 \hat{Q}_k - \left(\frac{1}{L}\right)^2 \hat{Q}_k \tag{7.10}$$

$$\hat{Q}_k = \hat{D}_k\left(\frac{L}{L-1}\right) \tag{7.11}$$

Logo, as variâncias $\hat{S}_k$ e $\hat{Q}_k$ são dadas por

$$\hat{S}_k = \hat{M}_k - \frac{\hat{D}_k}{L-1} \tag{7.12}$$

$$\hat{Q}_k = \hat{D}_k\left(\frac{L}{L-1}\right) \tag{7.13}$$

7.2.2 Cálculo da informação

Para calcular a informação, consideramos a mensagem A da equação 6.15 como uma seqüência de consenso C estimada pelas variâncias $\hat{D}_k$ e $\hat{M}_k$. Realizando as substituições das equações 7.13 e 7.12 na equação 6.15, temos

$$I(C,S) = \frac{1}{2}\log_2\left(\frac{L(\hat{D}_k + \hat{M}_k) - (\hat{D}_k + \hat{M}_k)}{\hat{D}_k L}\right) \tag{7.14}$$

Como a freqüência k que está no intervalo $0 < k < N/2$ possui duas componentes, precisamos multiplicar a quantidade de informação obtida em k por 2.

$$I_k = \begin{cases} \log_2 \left(\frac{L(\hat{D}_k + \hat{M}_k) - (\hat{D}_k + \hat{M}_k)}{\hat{D}_k L} \right) \text{se } k = 0 \text{ ou } k = N/2 \\ \\ 2 \log_2 \left(\frac{L(\hat{D}_k + \hat{M}_k) - (\hat{D}_k + \hat{M}_k)}{\hat{D}_k L} \right) \text{se } 0 < k < N/2 \end{cases} \tag{7.15}$$

A quantidade de informação total é dada pela soma da contribuição de cada freqüência k

$$I_{total} = \sum_{k=0}^{N/2} I_k \tag{7.16}$$

7.3 Procedimento da análise

Os programas foram desenvolvidos em linguagem C, utilizando o compilador GNU(gcc 3.2.2), a biblioteca FFTW[16], sobre a plataforma Linux.*

Inicialmente, retiramos de um banco de dados biológico[15] aproximadamente $L = 50$ trechos de DNA. Procedemos agrupando estes trechos em $G = 10$ grupos com $P = 5$ seqüências. Conforme explicado na seção 3.2, alinhamos as seqüências de cada grupo e então, extraímos um trecho de $N = 512$ bases de cada seqüência.

Extraídas as N bases e codificadas numericamente, calculamos a média m entre as seqüências de um grupo e também a diferença de cada seqüência do grupo em relação à média. Obtivemos então m e d_i, para $i \in \{1, ..., P\}$. Aplicamos a transformada de Fourier sobre os sinais m e d_i, e de posse de M_k e D_{ik} calculamos o espectro de potência. Este procedimento é aplicado para todos os grupos.

*código-fonte e documentação encontram-se disponíveis em http://www.ic.uff.br/∼hcgl

Capítulo 8

Resultados

Neste capítulo apresentamos os resultados obtidos com a aplicação do método proposto no capítulo 7. Em uma primeira análise, utilizamos apenas trechos de DNA de procariotos que fossem homólogos. Esperando obter uma quantidade de informação menor, agrupamos trechos de eucariotos e procariotos em uma segunda etapa.

8.1 Seqüências homólogas

Inicialmente, construímos o gráfico do sinal média e de sinais diferença de um grupo. Cada grupo possui 5 sinais diferença, mas para fins de análise, utilizamos apenas 2 sinais diferença. A figura 8.1 mostra os espectros de potência M_k e D_{ik} de um grupo de seqüências homólogas.

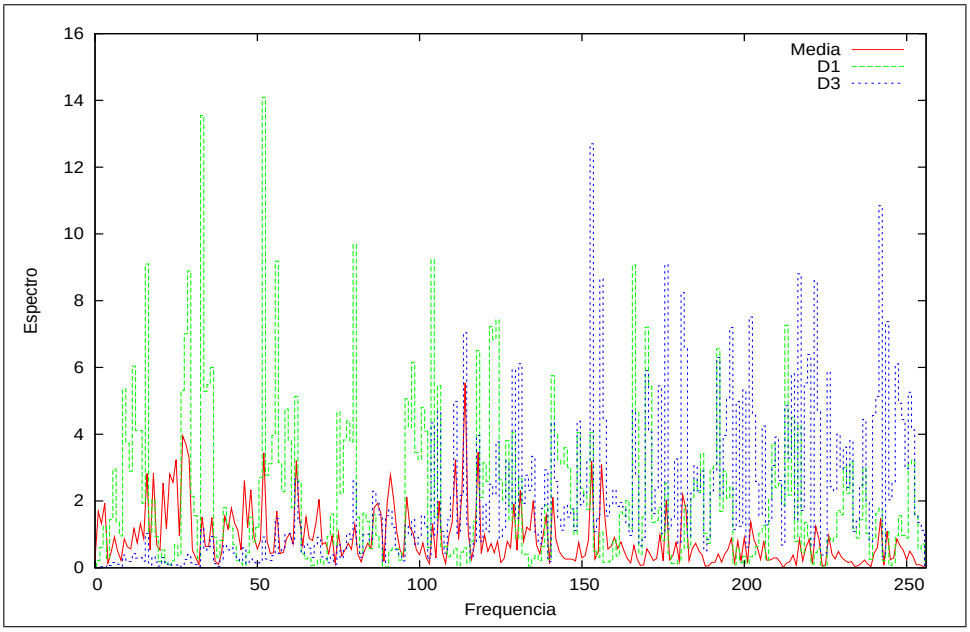

Figura 8.1: Espectros de potência de um grupo homólogo

8.2 Informação entre trechos homólogos

Conforme discutido na seção 7.1, a variância $\hat{M}_k$ e $\hat{D}_k$ pode ser estimada a partir do espectro de potência médio. A variância $\hat{M}_k$ é calculada sobre os G grupos e a variância $\hat{D}_k$ sobre as L seqüências. A figura 8.2 mostra os espectros de potência médio de M_k e D_k.

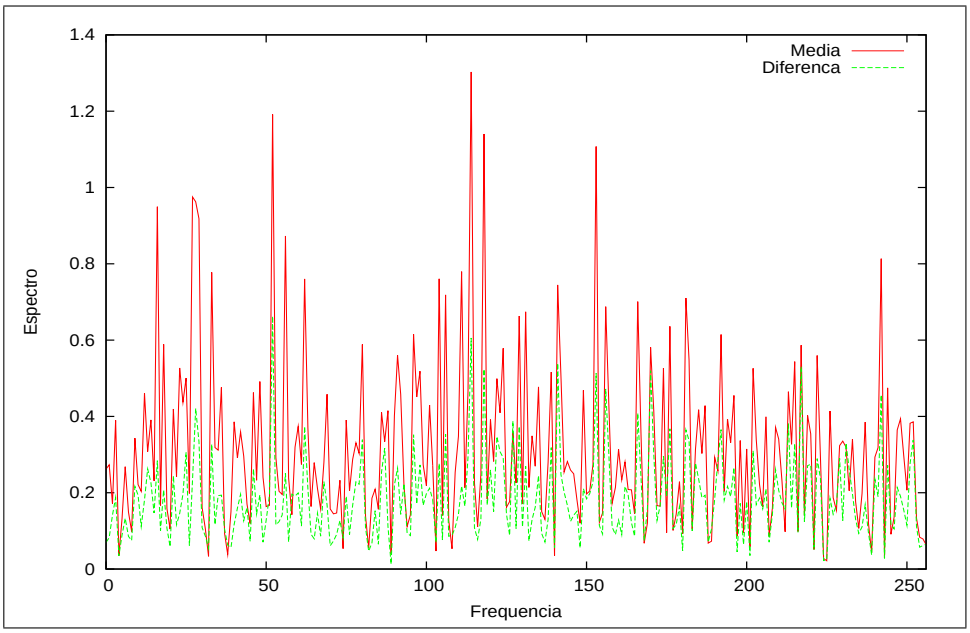

Figura 8.2: Espectros de potência médio

Com as variâncias $\hat{M}_k$ e $\hat{D}_k$ podemos aplicar a fórmula 7.15 para encontrar a quantidade de informação I presente em cada freqüência k, dada por I_k. A figura 8.3 mostra a quantidade de informação presente em cada freqüência.

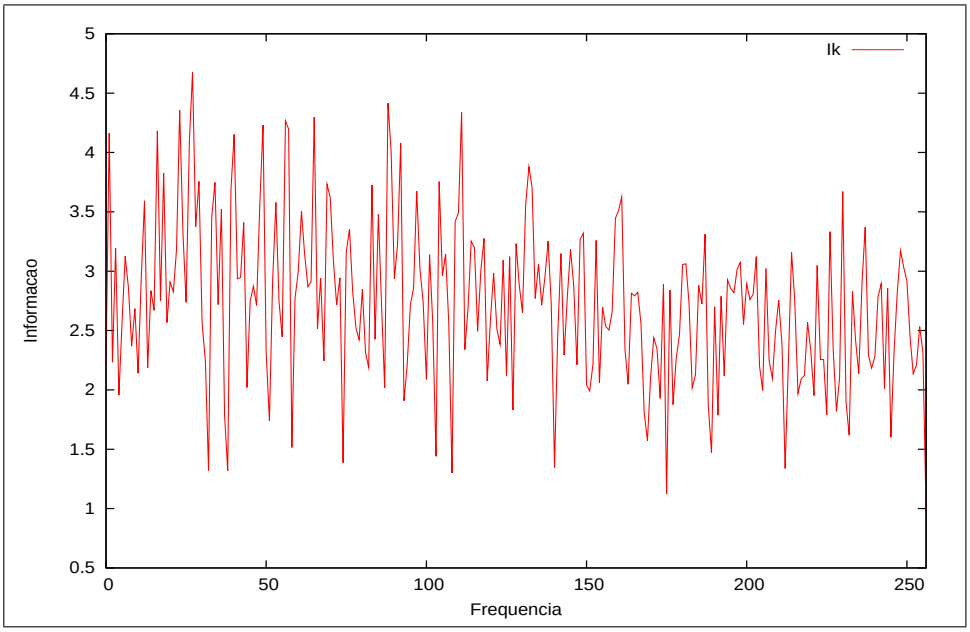

Figura 8.3: Quantidade de informação

A informação total é a soma da contribuição de cada freqüência k, dada pela fórmula 7.16. A quantidade de informação total para trechos homólogos encontrada foi aproximadamente 701 bits.

8.3 Informação entre trechos não-homólogos

Para seqüências não-homólogas, procedemos exatamente da mesma forma, mas utilizando trechos de vários organismos eucariotos e procariotos. A quantidade de informação total encontrada foi 294 bits, ou seja, um valor menor que o encontrado em trechos homólogos. A figura 8.4 mostra os gráficos da quantidade de informação presente em trechos homólogos e não-homólogos.

8.4 Discussão

Como constatado na figura 8.1, algumas seqüências tiveram um nível de ruído bastante alto em relação à média, o que pode explicar a redução da informação em

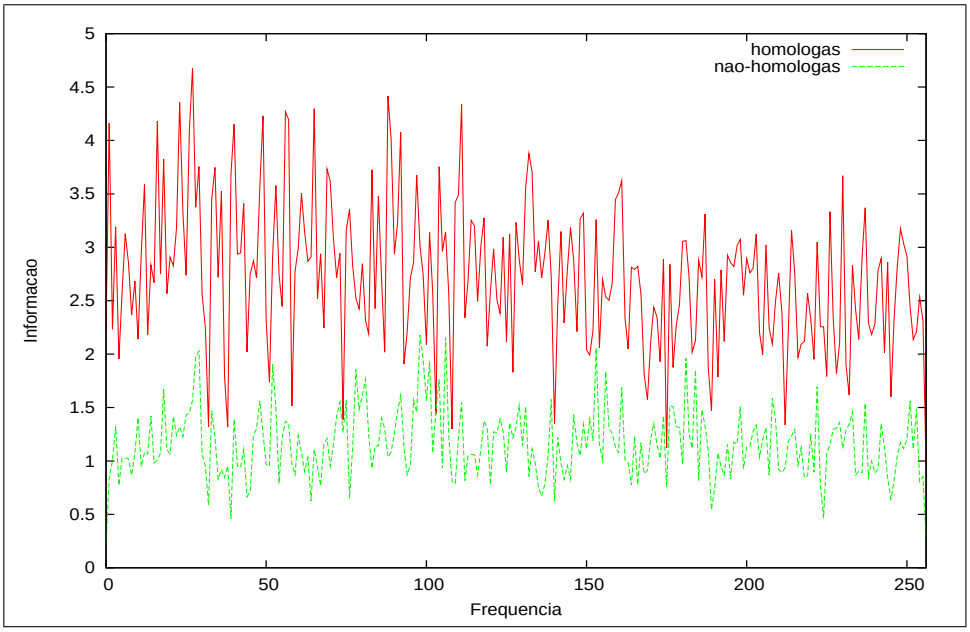

Figura 8.4: Comparação da informação

relação à obtida por Pessoa[6] em seu trabalho. A utilização de trechos homólogos que não codificam a mesma proteína também pode ter reduzido a informação.

A perda de informação em seqüências não-homólogas deve-se à grande quantidade de buracos inseridos durante a etapa de alinhamento. Os resultados podem ter sofrido redução da informação devido as restrições encontradas na obtenção de amostras, pois existe dificuldade de se coletar uma grande quantidade de trechos homólogos em bancos de dados públicos.

Capítulo 9

Considerações Finais e Trabalhos Futuros

Neste trabalho, propomos um modelo de cálculo de informação baseado no método de alinhamento múltiplo, utilizando trechos homólogos de procariotos. Acreditamos que esta abordagem ofereça uma melhor aproximação na identificação de famílias biológicas.

Este processo permitiu identificar que trechos homólogos fornecem uma quantidade de informação maior do que sequências que não possuem muita homologia entre si.

Cabe ressaltar que mesmo com uma quantidade limitada de amostras, temos uma boa representação da informação contida em grupos biológicos, o que nos faz acreditar que este método demonstra confiabilidade nos resultados.

Para confirmar a viabilidade do método proposto neste trabalho, algumas sugestões para trabalhos futuros seriam utilizar grupos de trechos homólogos de eucariotos, ou então utilizar grupos homólogos de diversas espécies. Uma alternativa para o aumento da quantidade de informação seria propor uma codificação numérica que resultasse em uma menor perda de informação.

Referências Bibliográficas

[1] Gibas C. e Jambeck P. *Desenvolvendo Bioinformática*. Campus, 2001.

[2] Miller E.W. Lipman D.J. Altschul S.F., Gish W. http://www.ncbi.nih.gov. NCBI.

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

[4] G. Landini. Comunicação pessoal, Setembro 2000.

[5] Dimitris Anastassiou. Digital signal processing of biomolecular sequences. *Columbia University*.

[6] Luciana Pessoa. Análise da informação mútua em seqüências homólogas. Tese de mestrado, UFF, Outubro 2004.

[7] UFSC. Proteína. Internet, Setembro 2005. http://www.enq.ufsc.br/labs.

[8] João Piazza. Uma metodologia para determinação do organismo de origem de seqüências de dna com aplicação em projetos est. tese, Unicamp, junho 2004.

[9] João Setubal, João / Meidanes. *Introduction to Computational Molecular Biology*. 1997.

[10] Ronald L. Rivest Clifford Stein Thomas H. Cormen, Charles E. Leiserson. *Introduction to Algorithms*. Second edition.

[11] D. Gusfield. *Algorithms on Strings, Trees, and Sequences. Computer Science and Computational Biology*. Cambridge University Press, 1997.

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

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

[14] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948.

[15] NCBI. Genbank. Disponível em http://www.ncbi.nlm.nhi.gov/entrez. Banco de Dados Biológico Público.

[16] FFTW. Fastest fourier transform in the west. Disponível em http://www.fftw.org.