UNIVERSIDADE FEDERAL FLUMINENSE

ARY HENRIQUE MORAIS DE OLIVEIRA

ANÁLISE DE INFORMAÇÃO EM PROTEÍNAS HOMÓLOGAS

NITERÓI

2006

UNIVERSIDADE FEDERAL FLUMINENSE

ARY HENRIQUE MORAIS DE OLIVEIRA

ANÁLISE DE INFORMAÇÃO EM PROTEÍNAS HOMÓLOGAS

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre. Área de concentração: Otimização Combinatória e Inteligência Artificial.

Orientador:
Prof.a Helena Cristina da Gama Leitão. D.Sc.

NITERÓI

2006

ANÁLISE DE INFORMAÇÃO EM PROTEÍNAS HOMÓLOGAS

Ary Henrique Morais de Oliveira

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre. Área de concentração: Otimização Combinatória e Inteligência Artificial.

Aprovada por:

Prof.a Helena Cristina da Gama Leitão, D.Sc. / IC-UFF
(Presidente)

Prof. Jorge Stolfi, Ph.D. / IC-UNICAMP

Prof.a Maria Emília Machado Telles Walter, D.Sc. /
CIC-UNB

Niterói, 08 de Dezembro de 2006.

Aos meus pais, irmãos, à minha namorada e aos amigos.

Agradecimentos

Agradeço principalmente a Deus, maior responsável por todas as vitórias que conquistei, sempre me guiando por bons caminhos e oferecendo grandes oportunidades, tais como esta, um dos meus maiores sonhos, o mestrado em computação. Agradeço ao amor dos meus pais, Ary Oliveira e Ana Tereza, e irmãos, Bruno Tácito e Ivo Sócrates, que sempre me motivaram, estendendo suas mãos e erguendo-me a cada queda diante dos obstáculos ocorridos por essa jornada no decorrer do mestrado em Niterói-RJ.

Agradeço com amor e carinho a minha namorada, Juliana Freitas, pelo grande amor mostrado nesta empreitada, pela compreensão e paciência nos momentos em que precisei me ausentar, mostrando-se uma pessoa forte e companheira. Aos colegas da graduação CEULP/ULBRA, a todos os professores, em especial ao meu orientador Fabiano Fagundes e a coordenadora Parcilene Fernandes Brito responsáveis por me aconselhar a cursar o mestrado, mostrando sempre que esse sonho era possível de ser realizado.

Não poderia deixar de agradecer aos colegas de trabalho da Secretaria da Segurança Pública do Estado do Tocantins, dentre eles: Andrey Reis, Antônio Omar, Alex Coelho, Alessandro Brandão, Douglas, Edgar Arrais, Fernando Frota, Jacir Bombardelli, Jânio Elias, Renato, Mario Elias, Norma Sueli, Roberta Mendes, Thelciane Paranhos (Tekinha), Waldeir e Wanderson Teixeira por me substituir e cobrir minha ausência no trabalho nos momentos de afastamento necessários para escrever a dissertação. Ao diretor de Tecnologia da Informação Marco Aurélio Giralde pelo auxílio e compreensão nos afastamentos do trabalho nos momentos em que eu precisava ausentar-me para as reuniões junto com minha orientadora em Niterói.

Aos meus grandes amigos do mestrado em computação da Universidade Federal Fluminense, os quais foram minha família quando eu estava em Niterói, principalmente nos momentos finais antes da defesa, dentre eles: Idalmis, Maysinha, Viviane Thomé, Nilmax; aos companheiros responsáveis pelas atividades esportivas: Rodrigo Toso e Thiago Proença, ao grande Jonivan responsável pelas atividades culturais e aquelas fugidas para descontração. aos grandes amigos e companheiros Kennedy e Diego Brandão, que sempre

deram uma grande força reforçando a certeza dessa vitória.

Aos companheiros de apartamento e também colegas da UFF, Allysson Gustavo e Lucas Bastos (em Boa Viagem) que me abrigaram em sua moradia assim que cheguei em Niterói-RJ (com muita mala e sem lugar definido para ficar); aos colegas do segundo apartamento, no bairro de São Francisco, Stênio Sã, Bruno Novelli (grande companheiro de chopp), Luciana Brugiolo, Juliana Silva e Tatiane Martins.

Aos colegas que eventualmente cederam um colchão em seu apartamento nestes momentos finais, dentre eles: Rafael Augusto (Guto), Thiago (facada), Raphael Guerra, Glauco. Ao grande amigo e conterrâneo Warley Gramacho (T.o.c.a.) que mostrou-se ser um grande companheiro nesta reta final do mestrado. Aos conterrâneos que acabaram de iniciar essa jornada no mestrado, Yuri e Ivan (T.o.c.o.). Um agradecimento mais que especial aos amigos Cristiane Ferreira, Renatha Capua, Daniela Viana e Jacques que me auxiliaram na preparação da apresentação nos dias e momentos que antecederam a defesa.

Aos membros da banca pela importante contribuição em sugestões, o que complementaram e enriqueceram o conteúdo esta dissertação. Aos professores e funcionários do IC (Angela, Maria e Izabela). Um agradecimento muito especial a minha orientadora Helena Cristina da Gama Leitão, pelo carinho e paciência que foi mostrado no decorrer do curso, por todo conhecimento repassado, importante para o meu aprendizado e amadurecimento na minha vida acadêmica.

Resumo

Neste trabalho analisamos a quantidade de informação em proteínas homólogas, utilizando métodos de processamento de sinais e teoria da informação. Uma das principais dificuldades no desenvolvimento do trabalho foi a escolha de uma codificação numérica adequada para aplicar as técnicas acima apresentadas.

A princípio, buscamos efetuar a transformada de domínio com o objetivo de analisar as seqüências a partir de uma nova forma de representação. Entretanto, para efetuar essa mudança de domínio tornou-se necessária a representação das seqüências, inicialmente na forma de strings de caracteres, em um domínio numérico.

Em seguida, aplicamos a transformada de Fourier sob a representação numérica das seqüências e utilizamos a análise espectral para gerar os respectivos espectros de potência. Após obtidos os resultados, plotamos os gráficos no intuito de analisar visualmente o comportamento das seqüências neste novo domínio. Ao final aplicamos ferramentas de estatística, tais como média e variância junto a ferramentas de teoria da computação para quantificar a informação contida nas bio-seqüências.

Abstract

In this dissertation we analyze the amount of information in homologous proteins, using methods of processing of signals and theory of the information. One of the main difficulties was the choice of a numerical codification adjusted to apply the techniques above presented.

The principle we search to effect the domain transform with the objective of to analyse the bio-sequences to leave a new form of representation. However, to effect this change of domain one became necessary the representation of bio-sequences of one it forms in string of char to one numerical form.

After that we apply the Fourier transform under the numerical representation of bio-sequences and we use the spectral analysis to generate the respective power specter. After gotten the results we plot the graphs in intention to analyze visually the behavior of bio-sequences int this new domain. To the end we apply the joined average and variance calculation with the information theory tools to quantify the content information in the bio-sequences.

Palavras-chave

Análise da quantidade de informação em seqüências biológicas;
Processamento de dados genômicos e proteômicos;
Biologia computacional;
Processamento de sinais digitais em seqüências biológicas.

Abreviações

DNA : Ácido Desoxirribonucléico;
RNA : Ácido Ribonucléico;
DSP : *Digital Signal Processing*;
FFT : *Fast Fourier Transform*;
NCBI : *National Center of Biotechnology of Informatio*;
DDBJ : *DNA Data Bank of Japan*;
EMBL : *European Molecular Biology Laboratory*;

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Os organismos vivos possuem informações sobre sua formação armazenadas no código genético (*genoma*). O código genético é composto por uma ou mais macromoléculas de DNA (ácido desoxirribonucléico) formados a partir de 4 diferentes bases nucléicas A, C, G, T, adenina, citosina, guanina e timina, também conhecidas como ácidos nucléicos ou nucleotídeos. Os ácidos nucléicos estão organizados no DNA em uma estrutura na forma de dupla hélice, conforme mostrado na figura 1.1, a qual foi descoberta em 1953 pelos cientistas J. D. Watson e F. H. Crick[28]. O código genético é dividido em duas regiões distintas denominadas de espaço gênico, ou simplesmente genes, e a outra composta pelo espaço intergênico. Os genes contêm as informações necessárias para produção de proteínas, que são os elementos responsáveis pela estruturação e manutenção da vida nos organismos.

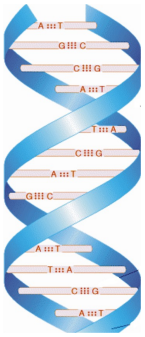


Figura 1.1: Modelo de dupla hélice do DNA proposta por J. D. Watson e F. H. Crick[28].

Desde a década de 1990, foram identificados e catalogados genomas dos diversos organismos existentes (animais, plantas, bactérias, vírus e etc), num processo conhecido como seqüênciamento e montagem de genoma. Atualmente, alguns organismos possuem seu genoma totalmente seqüênciados, dentre eles o homem (*homo sapiens*) e o rato (*mus musculus*). Porém, a identificação dos genes foi muito acelerada quando as áreas de

biologia molecular e a computação interagiram resultando na criação da *Biologia Computacional*. A computação além de apoiar a montagem do genoma, ainda auxilia na interpretação das seqüências que o compõe, concebendo novas ferramentas e aprimorando as atualmente disponíveis.

As bases de dados de seqüênicas biológicas têm sido bastante utilizadas na biologia computacional, pois contém não apenas as bases que constituem os genomas dos organismos seqüênciados, mas também as funções associadas a porções destas seqüências. Destacamos dentre as existentes atualmente: o GenBank (NCBI[5]), o DDBJ[1] e o EMBL[2]. Existem ferramentas de comparação de seqüências que utilizam estas bases de dados que são utilizadas para a análise e interpretação dos genes seqüênciados. Através da comparação podemos verificar as similaridades existentes entre seqüências, partindo do princípio de que seqüências similares implicam em estruturas similares e conseqüentemente funções similares. Assim, funções de genes podem ser inferidas por comparação de seqüências.

O conceito de comparação é associado a mais duas importantes definições na análise das seqüências, a homologia e a similaridade. A homologia entre duas ou mais seqüências distintas indica se as seqüências têm um ancestral em comum. A similaridade mede numericamente a equivalência entre duas ou mais seqüências biológicas, indicando o quão similares (parecidas) elas são[21]. A operação de comparação pode ser realizada utilizando diversos outros métodos de análise como o processamento de sinais digitais e teoria da informação, que foram utilizados neste trabalho.

O processamento de sinais digitais, responsável por tratar e extrair informações de sinais digitais, tem sido amplamente aplicado à seqüências de DNA, RNA e proteínas, motivo pelo qual criou-se uma nova subárea especializada denominada *processamento de sinais genômicos/proteômicos* [14] [27]. Suas técnicas de transformada de domínios, análise espectral e desenvolvimento de filtros digitais são alguns exemplos de abordagens aplicadas para a identificação de genes e análise de seqüências [27].

Outra técnica bastante utilizada em sistemas de comunicação digital, a qual foi estendida para a biologia computacional, é a teoria da informação. Esta área tem como principal objetivo a quantificação da informação em uma mensagem transmitida de uma origem a um destino. Autores como L. Pessoa[23], I. Grosse, S. V. Buldyrev e H. E. Stanley[15], H. Herzel, W. Ebeling e A. O. Schimitt [16] adaptaram as seqüências biológicas tratando-as como mensagens transmitidas a partir de alguma fonte, para que a partir disto fosse possível quantificar o seu conteúdo.

1.1 Trabalhos Relacionados

Encontramos diversos trabalhos na literatura que identificam características e classificações comuns entre as diversas seqüências biológicas contidas em bases de dados de genomas. Todavia, observamos que a maioria dos estudos estão concentrados em seqüências de DNA. Nesta seção vamos enumerar e descrever alguns trabalhos que utilizam ferramentas de processamento de sinais, análise estatística e teoria da informação aplicados no estudo de seqüências de DNA e de proteínas.

1.1.1 Trabalhos com DNA

Atualmente temos vários trabalhos na bibliografia propondo diversas formas para auxiliar na identificação de regiões codificadoras e não-codificadoras de DNA. Em seu trabalho, I. Grosse, S. V. Buldyrev e H. E. Stanley [15] busca a existência de padrões estatísticos com o objetivo de identificar a independência (diferença no padrão formação) entre as várias espécies de organismos. Para isto o autor utiliza o cálculo da informação mútua nos trechos de DNA em regiões codificadoras e não-codificadoras. Foram identificados alguns padrões estatísticos que variam muito de espécie para espécie, sendo que após aplicar a análise espectral foi identificado que o comportamento da região codificadora sempre apresenta-se com picos variando regularmente exceto nas freqüências múltiplas de três. Também foi observado que as regiões não-codificadoras apresentam um decréscimo na quantidade de informação à medida que a freqüência aumenta.

A teoria da informação também é utilizada em trabalhos como o de H. Herzel, W. Ebeling e A. O. Schimitt[16] onde o autor verificou que o DNA dos organismos tem milhares de seqüências dispersas e repetitivas. Para analisar estes dados, o autor utiliza o conceito de entropia para construir um modelo analítico a fim de analisar seqüências de DNA. O modelo consiste na independência de símbolos distribuídos de forma equivalente com repetições de dispersões aleatórias.

Ferramentas de processamento de sinais tem sido bastante aplicadas a área da biologia computacional, trabalhos como o de D. Anastassiou [7] mostram o desenvolvimento de ferramentas utilizando a transformada de Fourier e a análise espectral para o auxílio na identificação de regiões codificadoras de DNA. O autor apresenta uma ferramenta que utiliza-se de mapa de cores para auxiliar na identificação de tais regiões a nível de quadro de leitura, isto é, fazendo uma análise local dos espectros. Neste trabalho também são apresentadas formas de codificação numérica de seqüências biológicas utilizando números

complexos tanto para nucleotídeos quanto para aminoácidos.

Da mesma forma, P. D. Cristea [11] [10] apresenta em seu trabalho uma forma de representação numérica para nucleotídeos e códons, mas utiliza vetores de números complexos organizados de forma a representar um tetraedro. Seu trabalho tem como objetivo definir algumas propriedades as quais possibilitam representar as seqüências biológicas numericamente para a correta aplicação de técnicas de processamento de sinais.

1.1.2 Trabalhos com Proteínas

Alguns trabalhos concentram-se na análise da informação contida em proteínas, dentre eles destaca-se o de O. Weiss, M. A. J. Montaño e H. Herzel [29] que aplicaram a entropia de Shannon (Teoria da informação) e de alguns algoritmos de compressão de dados com o objetivo de verificar as redundâncias em seqüências de proteínas. Foi destacada a dificuldade em dar uma estimativa da quantidade de informação por resíduo, uma vez que, a lista de símbolos de aminoácidos é bem maior que a de nucleotídeos (são 20 símbolos para representar os aminoácidos ao contrário dos nucleotídeos que utilizam 4 símbolos).

Outra característica que dificulta a análise é o fato das seqüências apresentarem um tamanho bastante reduzido, pois em sua grande maioria as proteínas possuem centenas de aminoácidos e raramente este número chega aos milhares (os nucleotídeos geralmente apresentam seqüências contendo milhares e até mesmo milhões de nucleotídeos). Por tal motivo existem poucos esforços na busca da quantificação da informação em proteínas.

O trabalho de M. F. Macciato, V. Cuomo, A. Tramontano [20] apresenta análises de características de auto-correlação entre os aminoácidos de algumas proteínas. Os autores afirmam que as estruturas secundárias e terciárias das proteínas são totalmente determinadas por sua estrutura primária, onde foi levantada a hipótese de que o comportamento físico das proteínas é refletida em propriedades estatísticas de auto-correlação da seqüência de aminoácidos.

São descritas algumas propriedades de auto-correlação de uma sucessão de valores sugerindo a partir disto a existência de uma memória finita. Se a sucessão é feita de uma seqüência de valores independentes então sua auto-correlação é definida como sendo de ordem zero, se for definido como sendo de ordem um é porque a memória tem um valor previamente conhecido. Se nós considerarmos a probabilidade condicional na sucessão e K é o menor inteiro maior do que 0 (zero) então a seqüência de aminoácidos é de ordem K.

1.2 Motivação

Em sua dissertação, L. Pessoa[23] descreve a dificuldade de realizar o cálculo da quantidade de informação em seqüências de DNA pelo fato de existirem correlações (dependências) entre as bases próximas responsáveis pela formação dos códons. Tal característica dificulta a análise, uma vez que, a presença de um determinado nucleotídeo é produto da existência de outro previamente encontrada. Para cadeias de aminoácidos ainda não foram definidos esse tipo de correlações. Alguns trabalhos encontram algumas características de auto-correlação [20] e de redundâncias [29], o que pode sugerir repetição de blocos de aminoácidos.

Baseado no fato de que as seqüências de DNA contém informações necessárias para a geração das proteínas, podemos suspeitar que exista algum tipo de padrão de formação. Esta identificação de padrões poderia ser seguida por alguma forma de correlação refletida junto a seqüência de aminoácidos, seja em elementos próximos uns dos outros (curta distância/escala) ou em aminoácidos muitas posições atrás ou a frente (longa distância/escala). De qualquer forma, podemos inferir que as seqüências de proteínas não são formadas de forma aleatória, portanto, carregam informações que podem ser, consequentemente, quantificadas.

Ao trabalharmos com proteínas homólogas, temos como principal característica a formação das seqüências através de um ancestral comum, ou seja, ambas as proteínas comparadas conservam informações suficientes sobre o seu ancestral, apesar das transformações ocorridas por mutações. Com base nisto, podemos identificar através da informação contida em uma determinada seqüência outras homólogas a ela.

1.3 Objetivos

Este trabalho tem como objetivo desenvolver um método para analisar e quantificar a informação mútua entre proteínas homólogas, utilizando para isto técnicas de transformada de domínio com Fourier, análise espectral e teoria da informação. Para aplicar tais técnicas, foi necessário representar as seqüências de aminoácidos numericamente, motivo pelo qual foi levantada na bibliografia uma forma de codificação mais adequada ao nosso problema. Ao final, a implementação deste método recebe como entrada um conjunto de seqüências de proteínas homólogas alinhadas aos pares, e fornece como resultado a quantidade de informação mútua por componente (entre pares de seqüências), assim como a

informação mútua total entre todas as demais proteínas contidas no espaço amostral.

1.4 Estrutura da Dissertação

Esta dissertação está organizada da seguinte forma: No capítulo 2 são apresentados fundamentos básicos de biologia molecular, dentre eles a síntese protéica. Já o capítulo 3 mostra alguns conceitos de biologia computacional, abordando as definições e ferramentas importantes. O capítulo 4 mostra e discute diversas formas de codificação numérica para seqüências biológicas e propõe a codificação usada neste trabalho . Os capítulos 5 e 6 apresentam, respectivamente, as ferramentas da área do processamento de sinais, tais como a transformada de Fourier e a análise espectral; e elementos da teoria da informação, dentre eles, a forma de quantificação de mensagens, a entropia e o cálculo da informação mútua. O capítulo 7 propõe a metodologia utilizada para a análise da informação em seqüências homólogas. Ao final, nos capítulos 8 e 9, apresentamos os resultados dos experimentos e a conclusão, bem como a sugestão de trabalhos futuros.

Capítulo 2

Biologia Molecular

Neste capítulo são abordadas algumas definições importantes da área de biologia molecular, o que se faz necessário para a compreensão da abordagem proposta neste trabalho. São definidos alguns conceitos importantes tais como a classificação dos seres vivos, quanto a formação de sua célula, e algumas características de formação do código genético. São descritas também as etapas necessárias para a síntese das proteínas.

2.1 Eucariotos, procariotos e vírus

Os organismos vivos podem ser divididos em três grandes grupos: os eucariotos, os procariotos e os vírus. Com exceção dos vírus, todos os demais organismos são formados por células, sendo que dependendo do tipo estrutural da célula que compõe o corpo dos organismos, eles podem ser classificados em: **eucariotos** e **procariotos** [17].

Os eucariontes, que são organismos mais complexos, são formados por células eucarióticas, que apresentam membrana plasmática, citoplasma e o núcleo. Estes organismos apresentam em seu núcleo a maior parte do DNA celular, o qual é separado dos outros componentes celulares que ficam no citoplasma. As células eucarióticas têm formas e estruturas variadas e se diferenciam de acordo com suas funções nos diferentes tecidos. Compreendem os seguintes seres: animais, fungos, protozoários e plantas, exceto as bactérias e vírus.

Os procariontes são formados apenas por uma célula procariótica estruturalmente simples e bioquimicamente multiformes. Esses organismos apresentam membrana plasmática e citoplasma, entretanto o nucleóide - material nuclear - não é delimitado pela membrana nuclear (esta é inexistente). Os seres vivos que apresentam células procarióticas são as bactérias e as algas azuis.

Os vírus são seres muito pequenos formados basicamente de ácidos nucléicos com um envelope simples de proteínas (que abriga o material genético). Eles não apresentam estrutura celular típica. Para se manterem vivos necessitam parasitar células procarióticas ou eucarióticas, pois os vírus não possuem metabolismo próprio.

2.2 Ácidos nucléicos

Os ácidos nucléicos estão presentes em todas as células e são responsáveis pelo armazenamento e transmissão da informação genética, bem como pela tradução dessa informação e síntese protéica [17]. Todos os seres vivos possuem dois tipos de seqüências de ácidos nucléicos: DNA e RNA. Os vírus contêm apenas um dos dois tipos, ou o DNA ou o RNA. A diferença entre o DNA e o RNA está no fato de um possuir como pentose a desoxirribose enquanto o outro possui a ribose.

Os ácidos nucléicos são macromoléculas de enorme importância biológica, sendo constituídos pela polimerização de unidades mais simples chamadas nucleotídeos. Os nucleotídeos desempenham um papel central na transferência de energia e são subunidades a partir das quais as macromoléculas DNA (ácido desoxirribonucléico) e RNA (ácido ribonucléico) são formadas.

2.2.1 DNA

O DNA (ácido desoxirribonucleico) é responsável por armazenar toda a informação genética dos organismos vivos. Ele é formado pelo agrupamento dos ácidos nucléicos distribuídos em duas regiões distintas, o espaço gênico e o espaço intergênico [27]. Os genes contêm em seu interior duas sub-regiões denominadas *íntrons* e *éxons*. O éxon é um segmento de nucleotídeos responsável por codificar uma determinada proteína, enquanto que os íntrons são segmentos existentes entre os éxons que não codificam proteínas cuja função biológica ainda não foi claramente definida [8]. Os íntrons ocorrem apenas em organismos eucariotos, os demais, procariotos e vírus possuem apenas os éxons.

Cada nucleotídeo é composto de três moléculas ainda mais simples: uma base nitrogenada; uma pentose (açúcar) e ácido fosfórico (figura 2.1). As bases nitrogenadas são de dois tipos: *pirimídicas* e *púricas*. O que diferencia uma da outra é a quantidade de anéis, uma vez que as pirimídicas possuem um único anel heterocíclico e as púricas dois anéis fusionados, conforme podemos observar na figura 2.2. As bases púricas mais encontradas são a adenina (A) e a guanina (G), e as principais bases pirimídicas são a timina (T), a

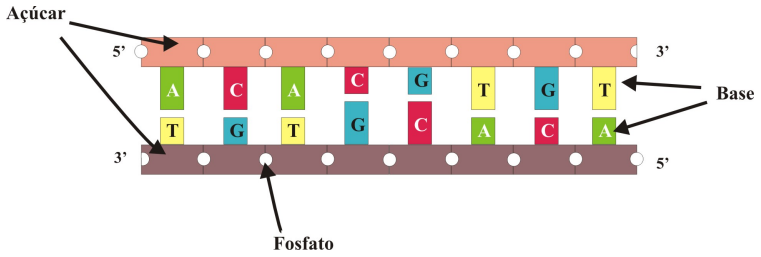


Figura 2.1: Dupla fita do DNA contendo as 04 bases nucléicas, o fosfato e o açúcar.

Classificação	Símbolo	Nucleotídeo
Purinas	A	Adenina
	G	Guanina
Pirimidinas	C	Citosina
	T	Timina
	U	Uracila

Tabela 2.1: As quatro bases nucléicas (que compõem os nucleotídeos) com respectivas classificações e símbolos

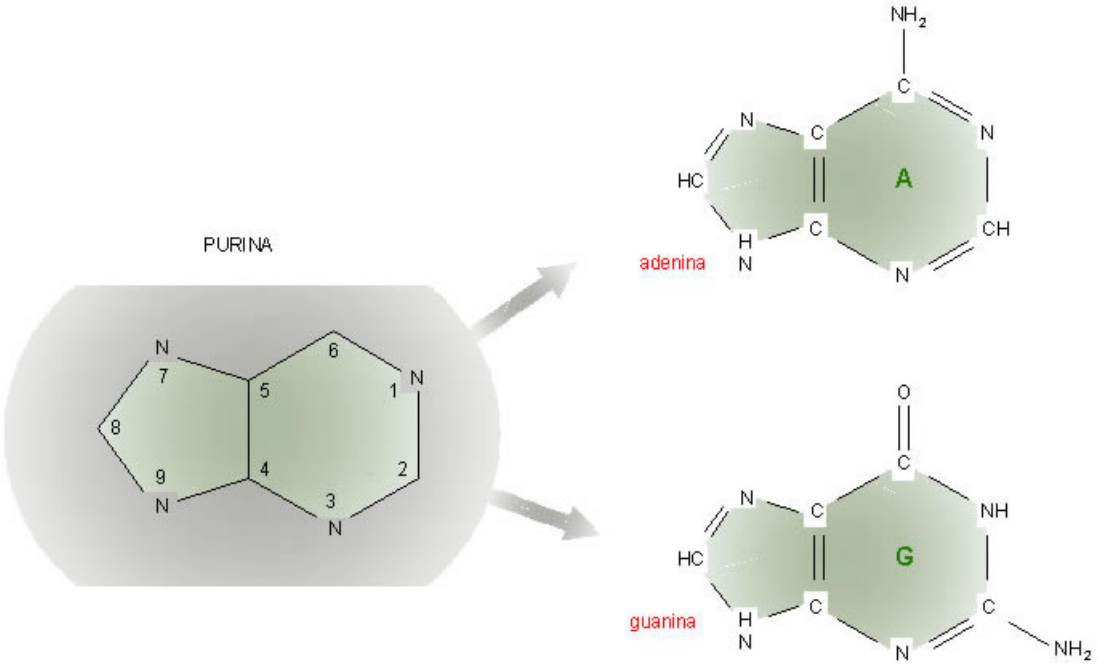
citosina (C) e a uracila (U). As bases nucléicas são apresentadas na tabela 2.1.

Conforme mencionado anteriormente, espacialmente, o DNA tem o formato de dupla hélice onde cada hélice, ou fita, é responsável pelo agrupamento da seqüência de DNA. As duas margens são ligadas através de pontes de hidrogênio, onde a primeira margem é composta pelos nucleotídeos no sentido $5'$ a $3'$, enquanto que a segunda é organizada em sentido inverso (contrário), ou seja, $3'$ a $5'$. A ligação entre os nucleotídeos de ambas as margens segue uma determinada regra, onde a adenina (A) liga-se à timina (T) e a citosina (C) à guanina (G). Por este motivo dizemos que esses pares de bases são *complementares* e que a segunda fita é complementar reversa da primeira. A figura 2.1 ilustra essas características.

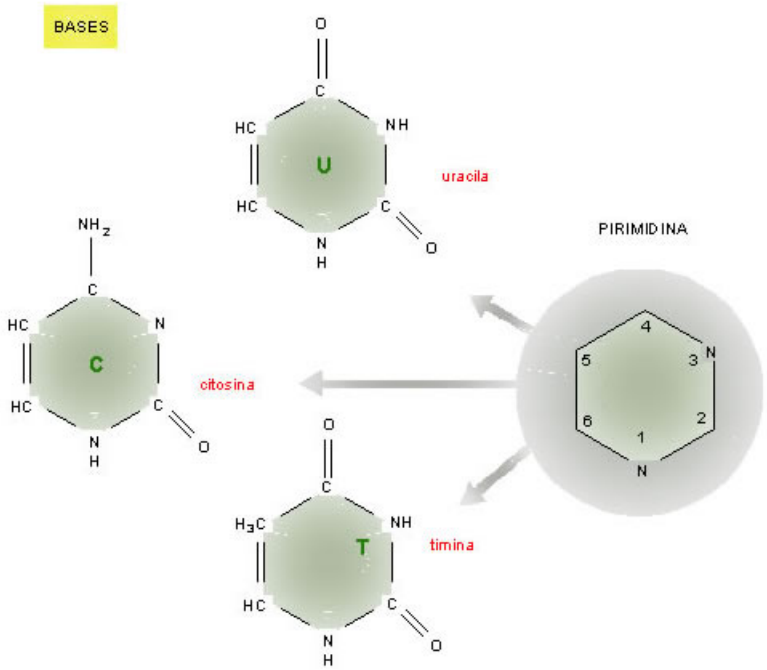
2.2.2 RNA

O RNA (ácido ribonucléico) tem sua composição bem semelhante à do DNA. Possui o mesmo grupo de moléculas, um ácido fosfórico, uma pentose e uma base nitrogenada, todavia, a pentose do RNA é a ribose enquanto no DNA é a desoxirribose. A base nitrogenada timina (T) presente no DNA é substituída pela uracila (U) no RNA. Estruturalmente, o RNA geralmente possui uma cadeia (fita) de nucleotídeos, entretanto, a exemplo do DNA pode ter a forma de dupla fita.

Existem 3 classes de RNA nas células cada uma responsável por desempenhar uma determinada função no processo de codificação de proteínas (denominado *síntese protéica*),



(a) Bases púricas



(b) Bases pirimidicas

Figura 2.2: Bases nucléicas púricas e pirimídicas [6].

são eles: o mRNA, rRNA e tRNA. O RNA mensageiro, ou mRNA tem como principal função o transporte da informação referente a estrutura de formação das proteínas. Estas informações são adquiridas a partir do DNA, sendo cada trinca de nucleotídeos denominada *códon*.

O RNA ribossômico, ou rRNA, é responsável pela formação do ribossomo participando da montagem da cadeia de aminoácidos utilizando o mRNA e o RNA transportador, ou tRNA que faz o transporte dos aminoácidos requisitados até o ribossomo.

Considerando que temos 4 diferentes bases nucléicas (A, T, C e G) e que para formar um códon precisamos agrupá-las em 3 bases, temos, portanto, 4^3 elementos distintos ou 64 códons. Cada códon é responsável pela codificação de um *aminoácido*.

2.3 Proteínas

As **proteínas** são macromoléculas compostas de aminoácidos responsáveis por realizar funções básicas da maior parte dos organismos vivos, dependendo de suas características estruturais e funcionais [12]. Elas são responsáveis pelo controle e gerenciamento dos processos químicos e biológicos nos organismos vivos. Estruturalmente as proteínas são polímeros de aminoácidos, isto é, são compostos químicos de baixa massa molecular resultantes de reações químicas de saturação, unidas por ligações peptídicas entre o grupamento aminoterminal de um aminoácido e o grupamento carboxiterminal de outro aminoácido. Os aminoácidos são constituídos essencialmente por carbono (C), hidrogênio (H), oxigênio (O), nitrogênio (N) e geralmente também por enxofre.

Os aminoácidos são pequenos componentes que agrupados formam proteínas. Um aminoácido é um composto orgânico simples que possui presos ao mesmo átomo de carbono, um grupo amina (-NH2 básico) e um grupo carboxílico (-COOH ácido). Todo aminoácido possui uma cadeia lateral ou um grupo R que é a parte variável responsável por diferenciar um aminoácido do outro. O grupo R pode variar em tamanho, polaridade e carga dos aminoácidos e tem a função de diferenciar os aminoácidos.

Existem 20 diferentes aminoácidos os quais são codificados por 61 códons. Os aminoácidos podem ser representados por grupos de $1, 2, 3, 4$ ou até mesmo 6 códons diferentes. Dentre os 64 códons existem 4 que desempenham uma função especial no código genético, o códon ATG que indica o início de uma região codificadora de proteína, além disso, ele também é responsável pela representação do aminoácido Metionina. Os códons TGA, TAG e TAA indicam o término de uma região codificadora e não codificam nenhum dos

Símbolo	Abreviação	Aminoácido	Códons
A	Ala	Alanina	GCT, GCC, GCA, GCG
C	Cys	Cisteína	TGT, TGC
D	Asp	Ácido Aspártico	GAT, GAC
E	Glu	Ácido Glutâmico	GAA, GAG
F	Phe	Fenilalanina	TTT, TTC
G	Gly	Glicina	GGT, GGC
H	His	Histidina	CAT, CAC
I	Ile	Isoleucina	ATT, ATC, ATA
K	Lys	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	Thr	Treonina	ACT, ACC, ACA, ACG
V	Val	Valina	GTT, GTC, GTA, GTG
W	Trp	Triptofano	TGG
Y	Tyr	Tirosina	TAT, TAC

Tabela 2.2: Os 20 aminoácidos padrões (primários) com respectivos símbolos e abreviaturas.

20 aminoácidos.

As proteínas possuem funções variadas e de extrema importância no organismo. Dentre estas funções temos proteínas: responsáveis pelo transporte de substâncias (ex. albumina e a hemoglobina), pelo armazenamento (ex. ferritina e ovoalbumina), a mobilidade (ex. actina e miosina), a estruturação (ex. colágeno e queratina), a defesa (ex. imunoglobulina), as regulatórias (ex. insulina) e etc. Além dessas temos uma classe especial, as denominadas enzimáticas, que agem como catalisadoras acelerando as reações químicas.

Uma característica de fundamental importância para a proteína é a sua habilidade de interagir com outras moléculas. Esta interação é possível graças a um lugar específico na proteína denominado *sítio ativo* onde outras moléculas podem se ligar a proteína. Este lugar é encontrado apenas em enzimas, as quais unem-se a outras moléculas e aceleram as reações químicas necessárias para a produção de outros elementos importantes para o organismo, tais como carboidrados, proteínas, lipídios e etc.

As proteínas possuem uma estrutura tridimensional que está diretamente relacionada com a estrutura [6]. Porém, para definir uma estrutura tão complexa como as proteínas foi necessário estabelecer quatro níveis estruturais:

- **Estrutura primária**: é formada pela seqüência de aminoácidos que constituem uma cadeia polipeptídica unida por ligações peptídicas. É o nível estrutural mais simples, porém o mais importante, uma vez que a partir dele deriva-se todo o arranjo espacial da molécula. Cada proteína tem sua própria estrutura primária a exemplo da figura 2.3.

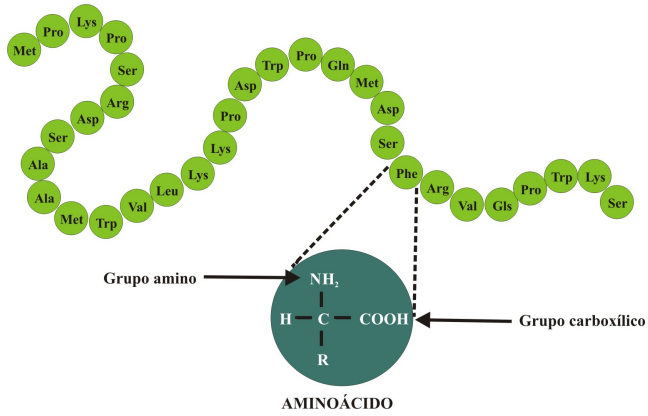


Figura 2.3: Estrutura primária da proteína.

- **Estrutura secundária**: é o arranjo (conformação) espacial da proteína e deriva da posição de certos aminoácidos na cadeia peptídica. Nesta estrutura temos dois tipos de arranjos denominados *alfa-hélice* - caracterizado por uma hélice em espiral - e *folha-beta* - envolvendo dois ou mais segmentos polipeptídicos arranjados em paralelo ou no sentido anti-paralelo.

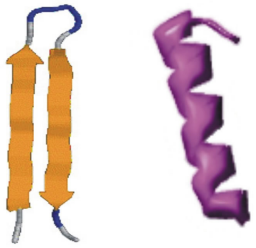


Figura 2.4: Estrutura secundária da proteína. À esquerda temos a folha-beta e à direita a alfa-hélice.

- **Estrutura terciária**: é a conseqüência da formação de novos dobramentos nas estruturas de alfa-hélice levando a uma configuração tridimensional, globosa e alongada de proteínas. Resulta do enrolamento da hélice (estrutura alfa-hélice) ou da folha pregueada (estrutura folha-beta). Esta estrutura é responsável por determinar a função biológica às proteínas. A estrutura terciária descreve o dobramento final de uma cadeia.

Figura 2.5: Estrutura terciária da proteína.

- **Estrutura quaternária**: é a combinação de duas ou mais proteínas, o que origina moléculas de grande complexidade. Essa transformação de grupos de proteínas em estruturas tridimensionais é a estrutura quaternária. Esta junção de cadeias polipeptídicas podem produzir diferentes funções para os compostos.

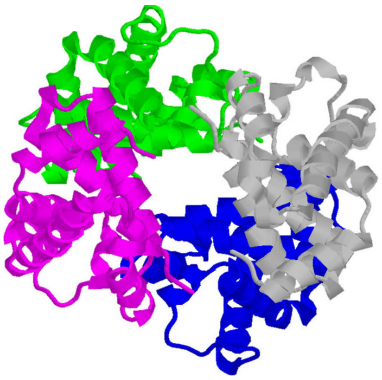


Figura 2.6: Estrutura quaternária de uma proteína.

2.4 Síntese de proteínas dos eucariotos

O DNA dos organismos contém a informação necessária para a produção de proteínas, porém deve passar por um processo denominado *síntese protéica* para a definição das proteínas [12]. A síntese protéica inicia-se com o reconhecimento de um gene ou um grupo de genes através de uma região conhecida como *promotor*. O promotor é responsável por indicar ao mecanismo celular a presença de um gene logo a sua frente (este processo de identificação também é realizado pelo códon ATG). Após identificado o início de um gene uma cópia desse gene é feita para uma molécula de mRNA (o RNA é formado por uma margem simples), momento este em que todas as bases nitrogenadas T serão substituídas pela base U. Este processo é denominado *transcrição*.

O mRNA é usado nas estruturas celulares denominadas *ribossomos* para a produção

de proteínas, portanto, é no interior dos ribossomos que ocorre a síntese protéica. Devemos observar que nos organismos eucariotos, como citado na seção 2.2.1 temos a presença de duas sub-regiões nos genes, os éxons e os íntrons. Na fase de transcrição os íntrons são retirados do mRNA o qual terá em seu interior apenas os éxons (informação necessária para a codificação de proteínas). A figura 2.7 mostra um exemplo de retirada do íntrons bem como a identificação do quadro de leitura .

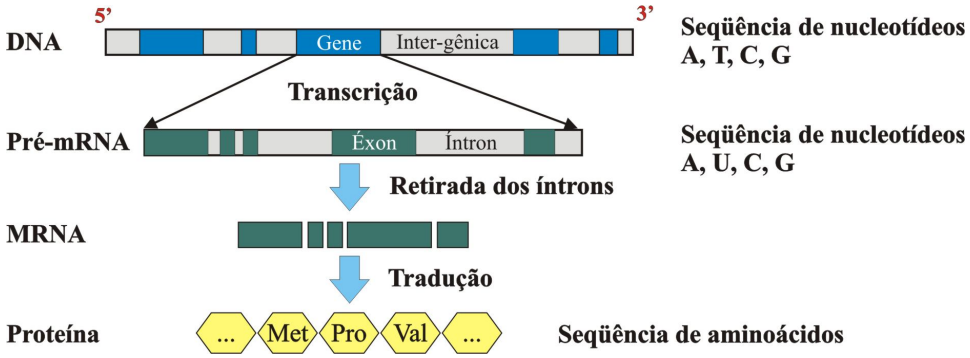


Figura 2.7: Processo de síntese protéica, quando o código genético passa por diversas fases até a formação de uma proteína [8].

Ao término da fase de transcrição, inicia-se a *tradução*. Nesta fase, os ribossomos (formados por moléculas de RNA denominadas rRNA), percorrem o mRNA e a medida que os códons são identificados o rRNA usa uma outra molécula de RNA, denominada tRNA, o aminoácido correspondente ao códon processado. Cada aminoácido é anexado à uma nova seqüência contendo todos os aminoácidos identificados pelo rRNA. Esse processo ocorre até a identificação de um dos códons de parada e tem como produto final a proteína devidamente sintetizada.

2.5 Cromossomos

Segundo [6], os cromossomos, também denominados *cromossomas* são estruturas compostas de uma longa sequência de DNA, que contém em seu interior um conjunto de genes. Os cromossomos são encontrados nos seres eucariotos e procariotos. Nos eucariotos, encontramos múltiplos cromossomos lineares dentro do núcleo celular, enquanto que nas bactérias eles podem ser circulares ou lineares, vindo o número de cromossomos variar de uma bactéria para a outra.

O conjunto de todos os cromossomos contidos em uma dada espécie é responsável por formar o seu genoma [21]. Como mencionado no capítulo 1, o genoma armazena todas as informações necessárias para a formação dos organismos vivos; ele é formado por duas

regiões distintas denominadas de região gênica (genes) e região intergênica [17]. É através da região gênica que são extraídas informações para a produção de proteínas.

Capítulo 3

Biologia Computacional

A **biologia computacional** tem por objetivo desenvolver métodos analíticos utilizando métodos teóricos, modelagem matemática e técnicas de simulação computacional para estudar sistemas biológicos. Existe uma confusão entre os termos biologia computacional e bioinformática. A **bioinformática** é a pesquisa, o desenvolvimento ou a aplicação de ferramentas e abordagens computacionais para expandir o uso de dados biológicos, médicos e de saúde, utilizando tais ferramentas para aquisição, o armazenamento, a organização e visualização dos dados [24].

Dentre as várias pesquisas e ferramentas desenvolvidas para a biologia computacional, destacam-se as de comparação entre seqüências e as de estruturação e consulta a bancos de dados de genomas [24]. Neste capítulo citamos algumas destas ferramentas trazendo definições básicas e exemplos focando as utilizadas para o desenvolvimento desta dissertação.

3.1 Comparação de duas seqüências biológicas

A comparação entre seqüências é a operação mais primitiva e importante da Biologia Computacional [21]. Ela consiste em encontrar partes semelhantes entre duas ou mais seqüências. Junto à comparação denotamos dois outros conceitos importantes para a biologia computacional. O primeiro é o *alinhamento*, que é definido como colocar uma seqüência acima da outra, adaptando-as através de deslocamento dos elementos e trechos contidos em seu interior a fim de ajustar a correspondência entre duas bases, uma de cada seqüênica. Cada alinhamento corresponde a um valor associado. O segundo conceito complementa o primeiro, é a *similaridade* entre seqüências, definida como a melhor forma de alinhar as duas seqüências.

```
gi|4557665|ref|NP_000866.1|insulin-like growth factor 1 receptor precursor [Homo sapiens]
gi|45384296|ref|NP_990363.1|insulin-like growth factor 1 receptor [Gallus gallus]
Matrix PAM30

Score = 3662 bits (8627),  Expect = 0.0
Identities = 1158/1376 (84%), Positives = 1263/1376 (91%), Gaps = 22/1376 (1%)

Query  900   NYTARIQATSLSGNGSWTDPVFFYVQAKTG-YENFIHLIIALPVAVLLIVGGLVIMLYVF  951
             NY+AR+QATSL+GNGSWT+PV FYVQ K+   Y+NF+HLII LP+A LLI+GGL+IMLYVF
Sbjct  900   NYSARVQATSLAGNGSWTEPVSFYVQPKSANYDNFLHLIIVLPIAFLLIIGGLLIMLYVF  952

Query  1259  FLEIISSIKEEMEPGFREVSFYYSEENKLPEPEELDLEPENMESVPLDPSASSSSLPLPD  1318
             FLEIISSIK+E++P F+EVSF+YSEENK P+ EELDLE ENMES+PLDP   SS+L   D
Sbjct  1260  FLEIISSIKDELDPAFKEVSFFYSEENKPPDTEELDLETENMESIPLDP---SSTLQPTD  1316
```

Figura 3.1: Exemplo de alinhamento entre trechos da proteína insulin-like growth factor 1 receptor precursor no *homo sapiens* e *Gallus gallus* utilizando a matriz de pontuação PAM30.

```
gi|4557665|ref|NP_000866.1|insulin-like growth factor 1 receptor precursor [Homo sapiens]
gi|45384296|ref|NP_990363.1|insulin-like growth factor 1 receptor [Gallus gallus]
matrix BLOSUM62

Score = 2424 bits (6283),  Expect = 0.0
Identities = 1157/1369 (84%), Positives = 1262/1369 (92%), Gaps = 8/1369 (0%)

Query  900   NYTARIQATSLSGNGSWTDPVFFYVQAKT-GYENFIHLIIALPVAVLLIVGGLVIMLYVF  958
             NY+AR+QATSL+GNGSWT+PV FYVQ K+   Y+NF+HLII LP+A LLI+GGL+IMLYVF
Sbjct  900   NYSARVQATSLAGNGSWTEPVSFYVQPKSANYDNFLHLIIVLPIAFLLIIGGLLIMLYVF  959

Query  1259  FLEIISSIKEEMEPGFREVSFYYSEENKLPEPEELDLEPENMESVPLDPSASSSSLPLPD  1318
             FLEIISSIK+E++P F+EVSF+YSEENK P+ EELDLE ENMES+PLDP   SS+L   D
Sbjct  1260  FLEIISSIKDELDPAFKEVSFFYSEENKPPDTEELDLETENMESIPLDP---SSTLQPTD  1316
```

Figura 3.2: Exemplo de alinhamento entre trechos da proteína insulin-like growth factor 1 receptor precursor no *homo sapiens* e *Gallus gallus* utilizando a matriz de pontuação BLOSUM62.

3.1.1 Alinhamentos entre pares de seqüências

O alinhamento busca ajustar duas seqüências utilizando para isso a inserção de espaços em branco, com o objetivo de torná-las do mesmo tamanho e produzir maior correspondência possível entre os seus elementos. Devemos nos ater ao fato de que não existe alinhamento entre dois espaços em branco. Portanto, baseado nesta definição, utilizamos duas seqüências contendo, respectivamente, trechos da proteína *insulin-like growth factor 1 receptor precursor* nas espécies *Homo sapiens* e *Gallus gallus*. Através das figuras 3.1 e 3.2 podemos verificar o alinhamento entre essas duas seqüências de aminoácidos.

Observe que os espaços inseridos (denotados pelo caractere '-' e denominados como *gap*) auxiliam no deslocamento dos caracteres e trechos na seqüências efetuando a adaptação e ajuste entre os elementos. Outro item que podemos observar através dos alinhamentos mostrados nas figuras 3.1 e 3.2, é que podem ocorrer alinhamentos entre elementos diferentes (por exemplo, na terceira posição - 902 - temos o alinhamento entre os aminoácidos T e S). Essas formas de alinhamento serão considerados se ajudarem na

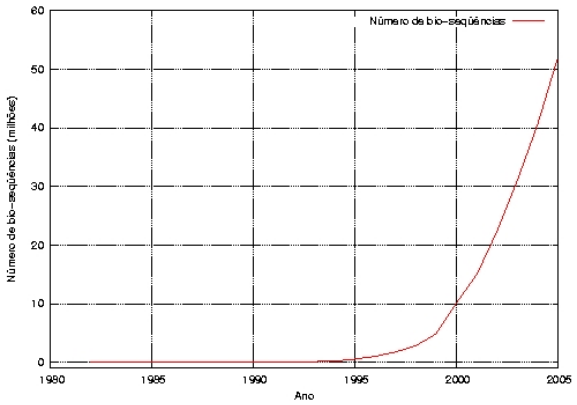


Figura 3.3: Gráfico que demonstra o crescimento do número de bases no NCBI[5].

aquisição de uma maior pontuação e conseqüentemente maior valor entre as seqüências comparadas.

A partir do alinhamento entre seqüências, obtemos uma pontuação, onde quanto mais equivalência tiverem os elementos, maior será a pontuação. Portanto, a similaridade é obtida através da pontuação do melhor alinhamento, ou seja o alinhamento ótimo.

3.2 Buscas em Bancos de Dados

Com a criação de bancos de dados específicos para o armazenamento de genomas temos verificado um grande crescimento do número de seqüências [18], principalmente nos últimos anos. Como podemos observar, o gráfico da figura 3.3 mostra que o crescimento da bases de genomas do NCBI [5] tem sido exponencial. Com essa característica, tem-se buscado criar e melhorar os métodos de consulta nestas bases de dados, resultando no desenvolvimento de ferramentas como o BLAST (responsável pelo alinhamento de seqüências) para o alinhamento de seqüências, e as matrizes de substituição responsáveis por uma melhor forma de pontuação na comparação de proteínas.

3.2.1 Matrizes de Substituição

Segundo Meidanis e Setubal [21] a comparação entre proteínas não pode seguir critérios comuns de comparação, assim como o modelo aplicado aos nucleotídeos, que utiliza penalidades de 0 (bases diferentes), 1 (bases iguais) e -1 (inserção de gaps). A comparação entre proteínas geralmente é realizada seguindo conceitos e critérios evolucionários, onde é utilizado um esquema de pontuação baseado em probabilidade de ligação entre os aminoácidos que compõem a seqüência.

As matrizes de substituição possuem valores proporcionais à probabilidade de um aminoácido X ser substituído por um Y, sendo que tal comparação é executada para todos os pares de aminoácidos possíveis. Temos dois tipos de matrizes de substituição, as matrizes PAM, *Percent Accepted Mutations* (Percentual de Mutações Aceitável) que foram as primeiras a surgirem e as matrizes BLOSUM, abreviação para *Block Substitution Matrix*, sucessoras das PAM's. Geralmente estas matrizes se apresentam com uma numeração após o nome, por exemplo, $BLOSUM62$, $PAM30$. Estes números refletem a porcentagem de divergência aceita no alinhamento entre duas seqüências.

As matrizes PAM têm como característica principal o alinhamento global de proteínas com parentesco próximo, uma vez que o procedimento padrão para a pontuação da matriz é a família das proteínas comparadas. A matriz primitiva denominada PAM1 aceita uma mutação com apenas 1% de divergência entre as seqüências comparadas, lembrado que tal comparação é a nível de aminoácidos.

As matrizes BLOSUM executam a análise utilizando blocos de alinhamentos locais tendo a vantagem de analisar uma grande diversidade de famílias de proteínas (ao contrário das PAM's que usam a família como principal característica) [18]. Estas matrizes, a exemplo das PAM's, apresentam variações identificadas por um número, no caso das BLOSUM, temos o exemplo da BLOSUM62 que indica que serão considerados alinhamento cuja similaridade apresente-se acima de 62%. Todavia temos diversas outras variações cuja numeração identifica a porcentagem mínima de identidade exigida entre duas seqüências.

As matrizes PAM têm como principal vantagem a possibilidade de se montar um modelo evolutivo gerando novas matrizes a partir da primeira. Entretanto, elas são baseadas na freqüência de substituição encontrada em proteínas de parentesco muito próximo. As matrizes BLOSUM têm como vantagem a detecção de seqüências com maior relação biológica, porém, não montam um modelo evolutivo assim como nas matrizes PAM's.

3.2.2 BLAST

O BLAST (*Basic Local Alignment Search Tool*) é uma família de ferramentas utilizadas para analisar a similaridade entre pares de seqüências catalogadas ou não em bases de dados de genomas [21]. O BLAST tem como retorno da análise uma pontuação através do alinhamento de vários pares de segmentos, isto é, entre os vários sub-trechos similares entre as duas seqüências comparadas.

Portanto, a princípio temos um grande trecho contíguo representando um conjunto de nucleotídeos ou aminoácidos, ao aplicarmos o BLAST é feito um alinhamento a fim de conseguir uma pontuação máxima de similaridade, caso não seja possível alinhar o trecho como um todo, o algoritmo efetua o alinhamento de sub-trechos e contabiliza na pontuação final.

Segundo definições apresentadas na base de dados do NCBI [5], o BLAST é programa que procura regiões de similaridade local entre seqüências e pode ser utilizado para inferir relacionamento funcionais e evolucionários entre seqüências. Temos algumas variações do BLAST aplicados para fins específicos, por exemplo, temos o BLASTN que é utilizado para o alinhamento de trechos de DNA, já o BLASTP é aplicado a trechos de proteínas (conjunto de aminoácidos).

Capítulo 4

Codificação numérica

As seqüências biológicas de DNA e proteínas são compostas por cadeias de caracteres, sendo que DNA possui um alfabeto de 4 elementos e as proteínas 20. Como enfatizado no capítulo 1, o fato das seqüências estarem representadas através de caracteres nos restringe a aplicarmos o processamento de sinais, pois precisamos de um conjunto de valores numéricos para o correto processamento matemático da seqüência.

Baseado nesta característica, buscamos na literatura várias formas de representação numéricas, tais como as apresentadas por D. Anastassiou[7]; P. D. Cristea[10]; J. P. M. Chalco[8]; L. Pessoa[23]; e P. P. Vaidyanathan[27] a fim de selecionar e propor uma mais adequada ao domínio do problema abordado neste trabalho.

4.1 Representação numérica de seqüências de DNA

A forma de associação numérica mais primitiva para seqüências biológicas é a utilização dos valores $0, 1, 2$ e 3, denotando os nucleotídeos A, C, G e T. Esse tipo de representação é abordada em alguns trabalhos, tais como o de J. P. M. Chalco[8], onde o autor busca identificar regiões codificadoras de proteínas através da transformada modificada de Morlet e o de P. D. Cristea[10], que concentra-se em propor uma forma de codificação numérica mais adequada à aplicação de processamento de sinas digitais.

Entretanto, encontramos no trabalho de L. Pessoa[23] justificativas que mostram que este tipo de codificação pode interferir na análise espectral, pois causa perda de informação quando aplicamos filtros digitais. Outro problema destacado é o fato de que esta representação sugere que um dado nucleotídeo é maior que outro, distinção esta que não segue nenhum tipo de embasamento biológico.

Nucleotídeo	Valor EIIP
A	0.1260
C	0.1340
G	0.0806
T	0.1335

Tabela 4.1: As quatro bases nucléicas associadas com seu respectivo valor EIIP (potencial de interação elétrica de íons).

J. P. M. Chalco[8] sugere a associação de uma propriedade química dos nuleotídeos denominada de valor EIIP, ou seja, medida do potencial de interação elétrica de íons (*electron-íon interaction potential*). Os valores EIIP são mostrados na tabela 4.1.

A maioria dos autores, dentre eles L. Pessoa [23] e D. Anastassiou[7], utilizam a representação numérica de seqüências de DNA por números complexos, onde os caracteres A, C, G e T são associados a números representados pelas variáveis a, c, g e t. Para cada uma dessas variáveis foram associados os valores complexos contendo uma parte real e outra imaginária. No caso de L. Pessoa, os valores foram associados da seguinte forma: $a = (1, 0), c = (-1, 0), g = (0, 1)$ e $t = (0, -1)$. Já D. Anastassiou associa os seguintes valores complexos $a = (1, 1), c = (-1, -1), g = (-1, 1)$, e $t = (1, -1)$, conforme mostrado na figura 4.1.

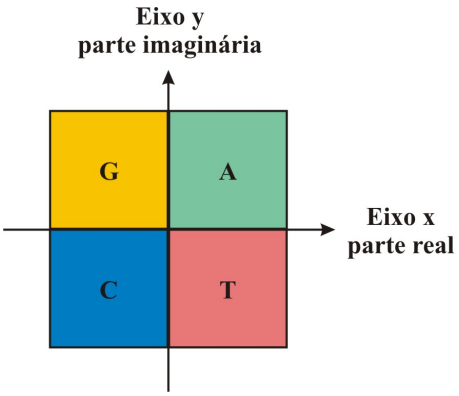


Figura 4.1: Representação de nucleotídeos em um plano cartesiano.

Podemos verificar que as bases nucléicas são representadas em um plano cartesiano bidimensional, ou seja, composto de duas retas, onde o eixo x representa a parte real e o eixo y representa a parte imaginária. Com esse par de valores (x, y) teremos 4 quadrantes, onde cada um será responsável por representar um dado nucleotídeo. Através da figura 4.1 podemos verificar que o primeiro quadrante codifica a base nucléica A, o segundo quadrante codifica a base T e o terceiro e quarto codificam, respectivamente, as bases C e G.

Na representação proposta por D. Anastassiou [7], temos 4 diferentes vetores, denotados cada um por u_A, u_C, u_G e u_T. Cada vetor tem o tamanho equivalente à quantidade de nucleotídeos da seqüência de DNA, por exemplo, se a seqüência possui $N = 1000$ elementos, cada um dos vetores (u_A, u_C, u_G e u_T) terão o mesmo tamanho, ou seja, $N = 1000$.

À medida que os nucleotídeos são encontrados em cada posição n da seqüência eles são codificados para os seus respectivos valores complexos e são associados aos vetores responsáveis por representá-los na posição n. A figura 4.2 mostra este o processo de codificação.

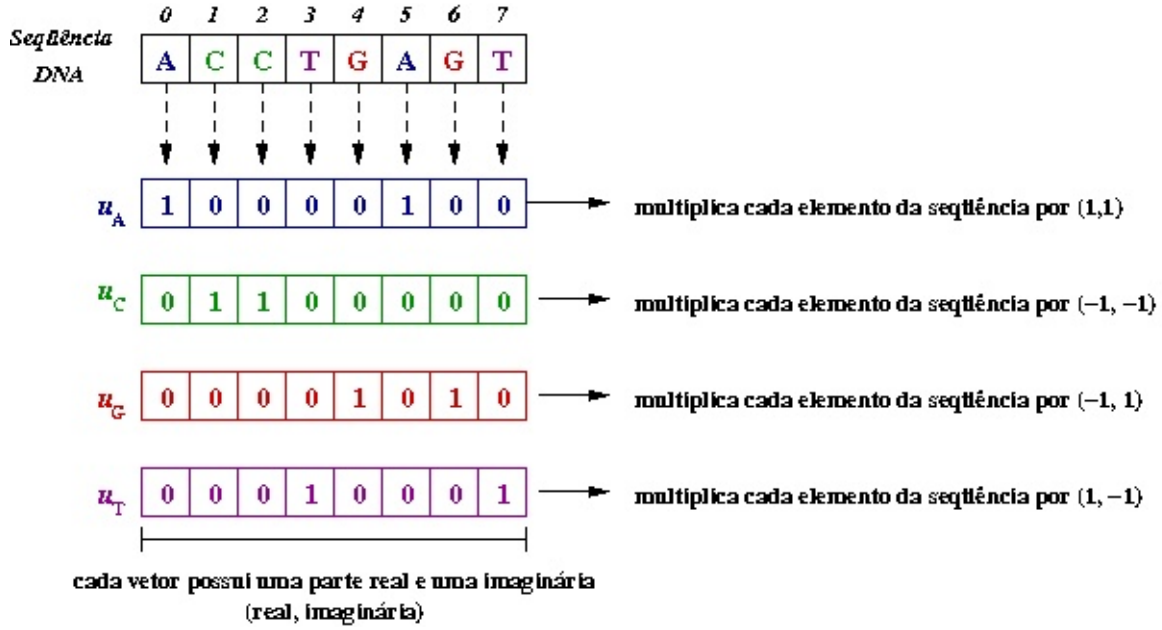


Figura 4.2: Exemplo de processo de codificação de seqüências de DNA proposto por D. Anastassiou [7].

O trabalho de L. Pessoa [23] segue uma idéia análoga ao de D. Anastassiou, entretanto, é utilizado apenas um vetor de valores complexos ao invés de 4, como podemos observar na figura 4.3. Nesta representação, à medida que a seqüência de aminoácidos é percorrida, os nucleotídeos são identificados, codificados para números complexos e associados ao vetor u na posição n, similar a posição da seqüência. Se esta seqüência estiver alinhada, podemos ter a presença dos espaços em branco, motivo pelo qual o vetor u terá o valor equivalente a $(0, 0)$, que determina que nenhum dos 4 nucleotídeos ocorrem na posição n.

No trabalho de P. D. Cristea [11] também encontramos uma representação dos nucleotídeos através de números complexos. Todavia, o autor apresenta uma projeção da seqüência de nucleotídeos em um conjunto de tetraedros aninhados (um no interior do outro). Nesta representação são utilizados 4 vetores com tamanhos equivalentes loca-

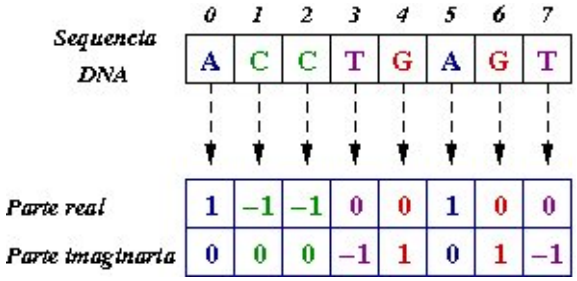


Figura 4.3: Exemplo de processo de codificação de seqüências de DNA proposto por L. Pessoa [23].

lizados simetricamente uns com os outros, conforme mostrado na figura 4.4. Cada vetor é orientado em direção a cada um dos 4 extremos do tetraedro em correspondência com o nucleotídeo definido ao seu respectivo vértice. Os valores associados são apresentados abaixo:

$$\vec{a} = \vec{i} + \vec{j} + \vec{k} \tag{4.1}$$

$$\vec{t} = -\vec{i} + \vec{j} - \vec{k} \tag{4.2}$$

$$\vec{g} = \vec{i} - \vec{j} - \vec{k} \tag{4.3}$$

$$\vec{c} = -\vec{i} - \vec{j} + \vec{k} \tag{4.4}$$

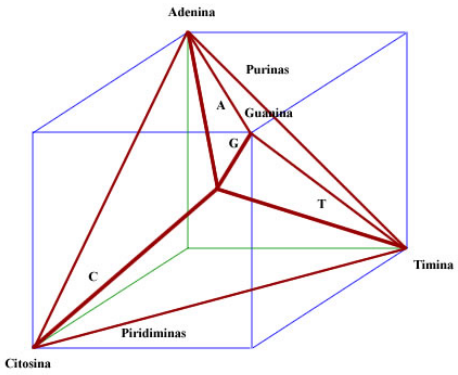


Figura 4.4: Representação de seqüências de DNA através de tetraedros [11].

4.2 Representação numérica de seqüências de proteínas

D. Anastassiou [7] e P. D. Cristea [11] estenderam suas representações para os códons. Como cada códon é composto de três caracteres, os autores desenvolveram uma representação que codifica cada códon em um conjunto de três valores numéricos. Tal característica teve como objetivo manter a estrutura de três elementos particular ao domínio dos códons.

No caso do trabalho de D. Anastassiou, é feito um mapeamento em um plano cartesiano, seguindo o seguinte procedimento: a primeira base do códon, agora denotada por um valor numérico, é utilizada para indicar em qual dos quadrantes está o aminoácido representado pelo códon analisado. As duas últimas bases do códon ficam responsáveis pela localização exata do aminoácido dentro do quadrante previamente identificado. O mapeamento dos aminoácidos é apresentado na figura 4.5.

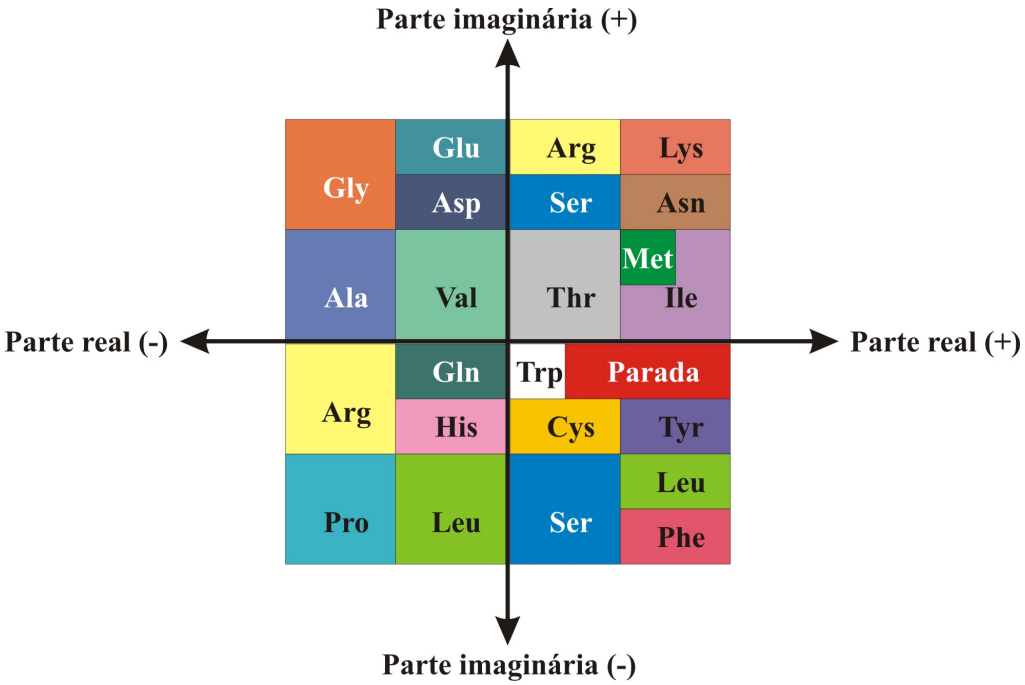


Figura 4.5: Aminoácidos representados em um plano complexo [7].

Todavia, para aplicar corretamente este método, D. Anastassiou descreve a necessidade de associar um peso a cada uma das três bases do códon. Este peso deve estar contido em um vetor de três posições, onde o valor de cada elemento do vetor deverá ser multiplicado pelo valor do elemento do códon nas posições equivalentes. Este vetor de pesos, denotado por h, deve ser ordenado em ordem crescente.

Com base nas características dos parágrafos anteriores, podemos associar a cada ele-

mento do vetor h os valores 1 para $h[0]$, $\frac{1}{2}$ para $h[1]$ e $\frac{1}{3}$ para $h[2]$ (equivalentes a $1, 0.5$ e 0.33). Dado o exemplo do aminoácido Triptofano (W), representado pelo códon TGG, teríamos o seguinte mapeamento: O nucleotídeo T será representado pelo valor complexo $1 - j$, já o nucleotídeo G será representado por $-1 + j$. Sabendo que h_0 corresponde ao valor 1, h_1 ao 0.5 e $h_[2]$ o valor 0.33 o valor é calculado da seguinte forma: $[h_0 * (1 - j)] + [h_1 * (-1 + j)] + [h_2 * (-1 + j)]$, que equivale a $[1 * (1 - j)] + [0.5 * (-1 + j)] + [0.33 * (-1 + j)]$. Como resultado temos que o aminoácido está representando no plano cartesiano na posição $1.17 - 0.17j$.

Encontramos alguns trabalhos na bibliografia que sugerem associarmos valores que expressem numericamente alguma propriedade química dos nucleotídeos (mostrado na seção 4.1) [8] e aminoácidos [27]. Esses autores apresentam o valor da medida do potencial de interação elétrica de íons, ou valor EIIP. Este valor pode ser associado, a exemplo dos nucleotídeos, a cada aminoácido conforme mostrado na figura 4.6 e tabela 4.2.

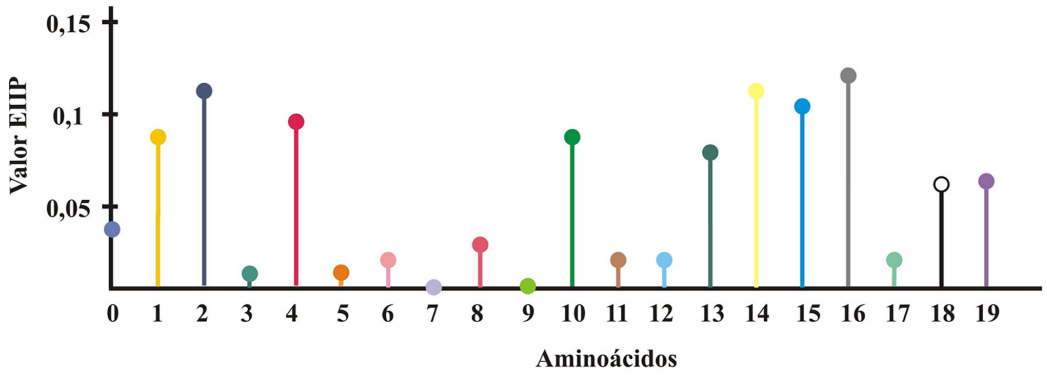


Figura 4.6: Valor do Potencial de Interação Elétrica de Íons de cada um dos 20 aminoácidos.

Com essa representação associamos cada aminoácido ao seu respectivo valor EIIP, seguindo uma forma de codificação baseada em uma característica biológica dos aminoácidos. Entretanto, temos dois grupos de aminoácidos contendo elementos com valores EIIP iguais - leucina (L) com a isoleucina (I); e cisteína (C) com a serina (S). Esta característica pode atrapalhar a identificação de um dos elementos e conseqüentemente a análise.

Nucleotídeo	**Valor EIIP**
A	0.0373
C	0.0829
D	0.1263
E	0.0058
F	0.0946
G	0.0050
H	0.0242
I	0.0000
K	0.0371
L	0.0000
M	0.0823
N	0.0036
P	0.0198
Q	0.0761
R	0.0959
S	0.0829
T	0.0942
V	0.0057
W	0.0548
Y	0.0515

Tabela 4.2: Aminoácidos associados ao seu respectivo valor EIIP (potencial de interação elétrica de íons).

4.3 Representação utilizada para a codificação dos aminoácidos

A representação numérica para seqüências de proteínas é dificultada pelo grande número de símbolos do alfabeto. Tal característica foi constatada na seção anterior (4.2). Trabalhos como o de P. D. Cristea [11] abordam algumas representações que associam escalas de valores que variam de $0 - 19$ para representar os aminoácidos e $0 - 63$ para os nucleotídeos, entretanto, tais valores não são baseados em nenhum fator químico ou biológico dos elementos, tal como a representação sugerida por P. P. Vaidyanathan[27].

A representação numérica proposta para este trabalho consiste na utilização de matrizes binárias (matrizes de bits), denotadas aqui como u. O conteúdo dessas matrizes é formado apenas pelos valores 0 e 1, os quais representam a ausência ou a presença de um aminoácido. Este tipo de representação é bastante utilizada [27], geralmente com algum tipo de valor associado, a exemplo da representação utilizando valores EIIP [8] [27] ou os números complexos para DNA utilizados por [7].

A matriz u tem por definição $M = 20$ linhas com valores no intervalo de $m = 0, 1, \ldots, M-1$, onde cada linha é responsável por representar cada um dos 20 aminoácidos. Já a quantidade de colunas é equivalente a N, que determina o número de elementos contidos na seqüência, onde $n = 0, \ldots, N-1$. Cada instante n, ou seja, cada coluna da matriz u, representa ou um aminoácido ou um espaço em branco. A soma dos elementos de todas as linhas em um instante n deve seguir a restrição apresentada na equação 4.5.

$$u[0][n] + u[1][n] + u[2][n] + \ldots + u[19][n] \in [0, 1] \tag{4.5}$$

A princípio todos os elementos da matriz u possuem seus valores iguais a 0, o que indica a ausência de aminoácidos. Ao percorrermos uma seqüência de proteínas elemento a elemento ($n = 0, \ldots, N-1$) identificamos qual o aminoácido foi encontrado, bem como a linha m responsável por representar este aminoácido e associamos a esta posição na matriz u o valor 1, isto é, $u_{m,n} = 1$. A figura 4.7 mostra um exemplo com uma matriz de bits representando a seqüência $AGATMFWPRF$.

A representação utilizada neste trabalho é a mais simples encontrada na bibliografia, entretanto, não cai nos dois tipos de problemas identificados neste capítulo, tais como a perda de informação, possível através do método apresentado por P. P. Vaidyanathan[27]; e a associação de valores arbitrários, sem algum tipo de fundamentação química ou bioló-

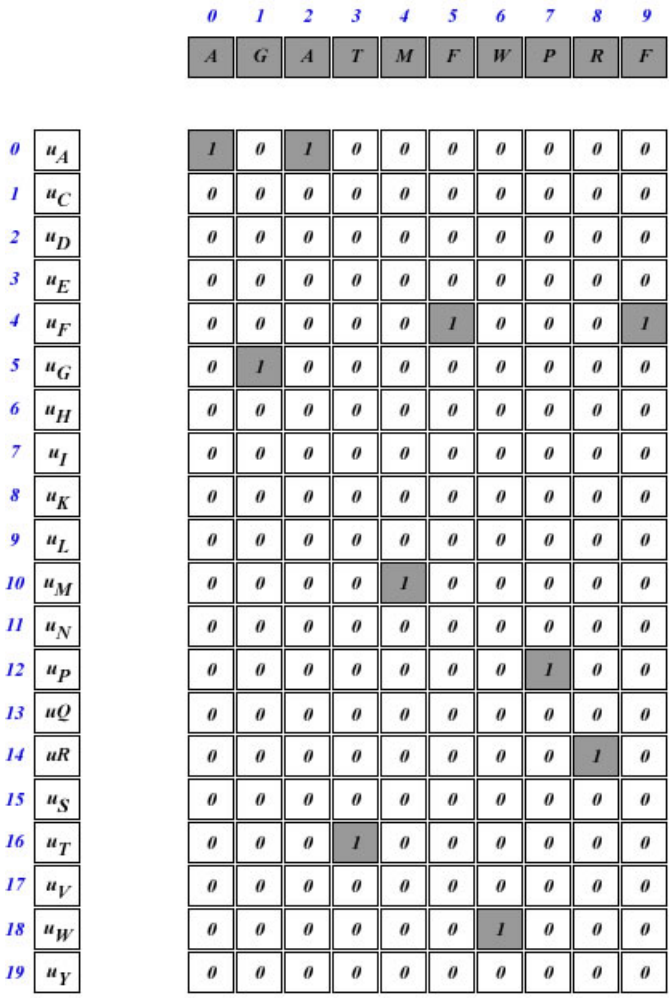


Figura 4.7: Aminoácidos representados em uma matriz de bits contendo em seu interior valores booleanos (0 - representando a ausência e 1 representando a presença de um dado aminoácido).

gica, como as mostradas por [11]. Porém, esta representação tem como grande desvantagem o alto custo computacional, uma vez que trabalhamos com matrizes, o que consumirá, no mínimo, algo na ordem de $O(20 * n)$.

Capítulo 5

Processamento de Sinais Digitais

Neste capítulo vamos abordar os principais conceitos de processamento de sinais apresentando uma visão geral dos conceitos de sinais e sistemas, domínios contínuos e discretos e transformadas de domínios. Vamos analisar a transformada de Fourier, utilizada no método com o propósito de representar as seqüências de aminoácidos no domínio das freqüências, juntamente com sua inversa, o que possibilita recuperarmos os valores originalmente utilizados. Ao final do capítulo, será definida a ferramenta de análise espectral, a qual é utilizada como um complemento aos valores produzidos pela transformada de Fourier.

5.1 Processamento de Sinais

O *Processamento de Sinais Digitais* (DSP) do inglês *Digital Signal Processing* é a disciplina que estuda as regras aplicadas aos sinais como funções de variáveis discretas, bem como os sistemas utilizados para processá-los [13]. Um sinal é tido como um fenômeno que ocorre continuamente no tempo carregando dados e informações [19].

Para uma melhor definição de sinais, J. Y. Stein [26] explica que devemos definir os dois tipos de sinais: o analógico e o digital. O sinal analógico é uma função de valores reais finitos $g(t)$ de uma variável contínua t (chamada tempo), definida para todos os tempos no intervalo de $-\infty < t < +\infty$. O sinal digital é uma seqüência g_n limitada a valores discretos com um índice n (chamado tempo discreto), definidos para todos os tempos $n = -\infty, \ldots, +\infty$.

Os sinais são compostos de uma ou mais ondas, as quais são definidas na física como sendo uma perturbação oscilante no espaço e periódica no tempo. As ondas são definidas a partir de um comprimento, que é a distância entre valores repetidos num padrão de onda

(numa onda senoidal, o comprimento é a distância entre picos); e de uma amplitude, a qual é uma medida escalar não negativa da magnitude de oscilação de uma onda (altura do eixo y).

Os sistemas são as ferramentas que processam os sinais com o objetivo de modificá-los ou de extrair alguma informação [19]. Os sistemas podem ser feitos de componentes físicos ou de ferramentas lógicas, tais como programas que processam um conjunto de sinais de entrada para produzir um outro conjunto de sinais de saída.

O Processamento de Sinais Digitais trabalha com informações contidas no tempo discreto, isto é, um sinal deve ter um período, tempo e/ou freqüências limitados, e que estejam contidos no domínio de valores numéricos. Portanto, para processar qualquer tipo de sinal, torna-se necessário que o mesmo esteja disponível através de uma representação numérica. A equação 5.1 mostra um exemplo de representação válida para o DSP [13]:

$$x(n), n \in \{Z\} \tag{5.1}$$

Baseado nas definições mencionadas nesta seção podemos dizer que um sistema de processamento de sinais é responsável por transformar (processar) uma seqüência de entrada $x(n)$ através de uma função $f\{.\}$ em uma seqüência de saída $y(n)$, tal como mostrado na equação 5.2.

$$y(n) = f\{x(n)\} \tag{5.2}$$

É importante ressaltar que os sinais não são necessariamente variáveis dependentes de uma unidade de tempo. Podemos trabalhar com sinais em razão de unidade de comprimento, de períodos, e no caso deste trabalho em razão da posição de algum elemento em uma dada seqüência.

5.2 Análise de Fourier

A *Transformada de Fourier* é uma ferramenta matemática utilizada na resolução de vários tipos de problemas de processamento de sinais. Ela é geralmente aplicada em valores representados no domínio do tempo, onde os valores de g são submetidos a uma função de um tempo t, por exemplo, $g(t)$; ou senão pela freqüência de domínios, onde o processo é especificado por uma amplitude G, sendo então como uma função de freqüência (espectros)

k, formando $G(k)$. [25].

Vale ressaltar que não precisamos necessariamente trabalhar com valores em função do tempo (segundos, por exemplo), podemos utilizar valores em função de uma posição i, $(g(i))$ ou até mesmo de uma unidade de comprimento. A Transformada de Fourier e sua inversa são definidas através das equações 5.3 e 5.4 [25].

$$G_k = \int_{-\infty}^{+\infty} g_t e^{2\pi i k t} dt \tag{5.3}$$

$$g_t = \int_{-\infty}^{+\infty} G_k e^{-2\pi i k t} dt \tag{5.4}$$

A partir das equações 5.3 e 5.4, a transformada de Fourier e sua inversa, respectivamente, podemos trabalhar com duas formas de representação: a primeira no domínio do tempo que mostra o valor do sinal em cada um dos tempos definidos; e a segunda no domínio das freqüências, ou espectros, onde o conteúdo harmônico do sinal é dado em todas as freqüências.

Estamos trabalhando com o mesmo valor, independente do domínio, porém, representados de forma diferente. Para obtermos valores no domínio das freqüências aplicamos a equação 5.3 e para conseguirmos os valores no domínio do tempo basta aplicarmos a equação 5.4.

5.2.1 Transformada Discreta de Fourier

Ao contrário da transformada de Fourier contínua apresentada acima, a transformada discreta de Fourier tem por objetivo trabalhar com valores em um domínio discreto, contendo um número finito de pontos amostrados denotados por N. Baseado nestes N números de elementos nas amostras consecutivas teremos:

$$g_k \quad \leftrightarrow \quad g(t_k), t_k = k \quad \Delta = k = 0, 1, 2, \ldots, n-1 \tag{5.5}$$

Na equação o elemento Δ representa o espaço de amostragem. Com base nesta primeira definição, o próximo passo é aproximar a integral, utilizada para representação contínua, para uma soma discreta, portanto:

$$G(k) = \int_{-\infty}^{+\infty} g(t)e^{2\pi ikt}dt \quad \rightarrow \quad \sum_{t=0}^{N-1} g_t e^{2\pi ikt}\Delta \quad \rightarrow \quad \sum_{t=0}^{N-1} g_t e^{2\pi itk/N} \tag{5.6}$$

Ao final a equação denominada transformada discreta de Fourier de N pontos G_k é denotada pela equação 5.7.

$$G_k = \sum_{t=0}^{N-1} g_t e^{2\pi itk/N} \tag{5.7}$$

Da mesma forma que a transformada contínua de Fourier, a transformada discreta também tem a capacidade de recuperar o conjunto de entrada através de uma operação de inversão. Tal operação pode ser obtida a partir da equação 5.8.

$$g_t = \sum_{k=0}^{N-1} G_k e^{2\pi ikt/N} \tag{5.8}$$

5.2.2 Análise Espectral

A *análise espectral* tem como objetivo identificar a potência máxima contida no intervalo de freqüências entre k e $k + nk$. Tal definição é considerada levando em conta que trabalhamos com freqüências positivas, isto é, com o valor de k contido no intervalo de 0 a $+\infty$. Para fazer a análise espectral devemos determinar o valor da *densidade de potência espectral* (PSD) de uma função g o qual é dado pela equação 5.9.

$$P_n(k) \leftrightarrow |G(k)|^2 + |G(-k)|^2, \quad 0 \leq k \leq +\infty \tag{5.9}$$

Quando utilizamos a transformada rápida de Fourier, a seqüência produzida pode ser reduzida pela metade, ou seja, terá $k = \frac{N}{2}$ elementos. Tal fato ocorre, pois, os valores da primeira metade da seqüência (após a transformada) são iguais aos da segunda metade, com exceção dos valores contidos em $k = 0$ e $k = \frac{N}{2}$. Portanto, os elementos contidos no intervalo entre $k = 0, \ldots, \frac{N}{2} - 1$ são idênticos aos presentes no intervalo $k = N, \ldots, \frac{N}{2}$.

Observe que o intervalo da primeira metade inicia-se em $k = 0$ (início da seqüência) e termina em $k = \frac{N}{2} - 1$; já a segunda metade, inicia em N, ou seja, do fim da seqüência até a metade $k = \frac{N}{2}$. A partir desta característica, dizemos que temos freqüências positivas e negativas, as quais são representadas, respectivamente por, k e $-k$. Por tal motivo, para calcularmos o espectro de potência, é necessário aplicarmos as equações 5.10, 5.11 e 5.12.

$$|G_0|^2 \quad para, \quad k = 0 \tag{5.10}$$

$$\left[|G_k|^2 + |G_{n-k}|^2\right] \quad para, \quad k = 1, 2, \ldots \frac{n}{2} - 1 \tag{5.11}$$

$$|G_{\frac{n}{2}}|^2 \quad para, \quad k = \frac{n}{2} \tag{5.12}$$

Capítulo 6

Conceitos de Probabilidade e Teoria da Informação

Neste capítulo vamos abordar definições importantes aplicadas nas áreas da teoria das probabilidades e teoria da informação. Tais conceitos são essenciais para a implementação do método proposto neste trabalho, uma vez que, a partir deles extraímos as ferramentas matemáticas utilizadas para efetuar a quantificação da informação nas seqüências biológicas.

6.1 Teoria das probabilidades

A teoria da probabilidade é o ramo da matemática que estuda os fenômenos aleatórios[19]. Através dela podemos calcular a chance de ocorrência de um número em um experimento aleatório. Um experimento aleatório é aquele que, quando executado sobre iguais condições, pode fornecer resultado diferentes. Os experimentos aleatórios possuem um conjunto de resultados possíveis, denotados como espaço amostral (ou S), por exemplo, o espaço amostral de um experimento com um dado é definido por $S = \{1, 2, 3, 4, 5, 6\}$, ou seja, todos os valores das possíveis faces de um dado.

Dado que num evento aleatório as probabilidades são igualmente distribuídas, temos, portanto, que a probabilidade de ocorrer um evento A é dada através da equação 6.1.

$$P(A) = \frac{n(A)}{n(S)} \tag{6.1}$$

onde $n(A)$ define a quantidade de vezes o elemento A ocorreu e $n(S)$ define a quantidade de casos possíveis a partir do espaço amostral.

6.1.1 Probabilidade Condicional

A probabilidade condicional é definida como a probabilidade de um evento X ocorrer dada a prévia ocorrência de um evento Y[22]. Denotamos a probabilidade condicional como a probabilidade de X dado Y, ou, $P(X|Y)$, conforme mostrado na equação 6.2.

$$P(X|Y) = \frac{P(X \wedge Y)}{P(Y)} \tag{6.2}$$

6.1.2 Valor esperado (Média)

Dado que todos os eventos têm igual probabilidade de ocorrer, isto é, $\frac{1}{N}$, onde N representa o número de eventos, então o valor esperado é a *média aritmética*. Através de uma variável aleatória X, contendo os valores $x_0, x_1, \ldots, x_{N-1}$, com suas respectivas probabilidades, denotadas através da função $p(x_i)$, calculamos o valor esperado através da equação 6.3

$$E[X] = \sum_{i=0}^{N-1} x_i p(x_i) \tag{6.3}$$

6.1.3 Variância

A variância é responsável por medir o quão distante em geral os valores de uma variável aleatória X se encontram do valor esperado [22]. Dado que $E[X]$ é o valor esperado de uma variável aleatória X, então a variância é adquirida através da equação 6.4.

$$\hat{X} = E(X^2) - [E(X)]^2 \tag{6.4}$$

6.2 Teoria da informação

A *teoria da informação* tem por objetivo auxiliar na análise da quantidade de informação contida em uma mensagem. O princípio básico da teoria da informação é que o conteúdo de uma mensagem sempre pode ser quantificado [26]. Este conteúdo geralmente é medido em bits, entretanto, existem ferramentas que trabalham com outros tipos de unidades de medidas.

Segundo B. P. Lathi [19], a quantidade de informação recebida está diretamente rela-

cionada com a incerteza ou inversamente relacionada com a probabilidade de sua ocorrência. Quanto menor a probabilidade de ocorrência de uma mensagem (elemento responsável por armazenar alguma informação), mais importante serão as informações associadas a ela. Portanto, se P é a probabilidade de ocorrência de uma mensagem e I a informação obtida através dessa mensagem então quando $P \to 1, I \to 0$ e quando $P \to 0, I \to \infty$. Em geral um pequeno valor de P resulta em um grande valor de I. A equação 6.5 mostra como uma informação pode ser quantificada.

$$I = log_2 \frac{1}{P} bits \tag{6.5}$$

6.3 Entropia

A *entropia* é definida como a medida da quantidade de informação de uma seqüência de valores sucessivos. A principal característica a se observar é o fato de que cada elemento da seqüência é independente dos elementos anteriores. Partindo desse princípio, a entropia é calculada através da equação 6.6.

$$H(m) = \sum_{i=0}^{n-1} P_i I_i \quad H(m) = \sum_{i=0}^{n-1} P_i log \frac{1}{P_i} \quad H(m) = -\sum_{i=0}^{n-1} P_i log P_i \tag{6.6}$$

Autores como T. M. Cover e J. A. Thomas [9] reforçam as definições sobre a entropia definindo-as como: Considerando X uma variável aleatória de um conjunto finito S com probabilidade expressa por $P(X), X \in S$ e seguindo a propriedade expressa pela equação 6.7. Devemos observar que se todas as probabilidades foram concentradas em apenas um elemento contido no conjunto, então X não será um elemento aleatório. Todavia, por outro lado, se todas as n possibilidades são igualmente distribuídas entre todos os elementos do conjunto a incerteza sobre o comportamento de X é alto.

$$P_0 + P_1 + \ldots + P_{n-1} = 1 \tag{6.7}$$

Considerando o valor n como o tamanho do conjunto S teremos $P(X) = \frac{1}{n},\ X \in S,\ H(X) = log\ n$. Uma distribuição uniforme das probabilidades sobre um espaço finito de elementos maximiza a entropia $H(X)$ sobre todos os n pontos. Tal característica é mostrada através da equação 6.8.

$$\max_{P_0,\ldots,P_{n-1}} -\sum_{i=0}^{n-1} P_i \; log \; P_i \quad sujeito\ a \quad \sum_{i=0}^{n-1} P_i = 1 \tag{6.8}$$

6.4 Entropia Condicional

A entropia condicional fornece a entropia sobre um valor X uma vez que conhecemos o valor de uma variável de Y. Portanto, a entropia condicional é denotada por $H(X|Y)$ sobre a distribuição das probabilidades $P(X|Y = y)$, ou seja:

$$H(X|Y) = \sum_{i=0}^{n-1} H(X|Y = y_i)P(Y = y_i) \tag{6.9}$$

Podemos adquirir a quantidade de informação que Y fornece sobre X a partir da entropia condicional, conforme descrito na equação 6.10

$$I(Y \rightarrow X) = H[X] - H[X|Y] \tag{6.10}$$

Segundo l. Pessoa [23] a informação média $I(Y = y_i \rightarrow X)$ também pode ser obtida através do valor genérico esperado de y_i obtido através da equação:

$$I(Y \rightarrow X) = \sum_{i=0}^{n-1} P(Y = y_i)I(Y = y_i \rightarrow X) \tag{6.11}$$

6.5 Informação Mútua

A informação mútua entre duas variáveis aleatórias X e Y mede a queda da aleatoriedade e da incerteza de uma determinada variável quando outra é observada. Esta informação é obtida através da equação;

$$\begin{aligned} \Delta(X,Y) &= H(X) - H(X|Y) && (6.12)\\ &= H(X) + H(Y) - H(X,Y) && (6.13)\\ &= H(Y) - H(Y|X) && (6.14) \end{aligned}$$

Segundo T. M. Corver [9] podemos ver que a informação mútua é simétrica, isto é, a

redução da incerteza de X dado que Y é observado é a mesma redução da incerteza de Y quando X é observado.

Capítulo 7

Método Proposto

Neste capítulo vamos utilizar os conceitos de processamento de sinais digitais e teoria da informação, introduzidos nas seções anteriores, para a construção de uma ferramenta de análise da quantidade de informação em um conjunto de proteínas homólogas. O método proposto neste trabalho é uma extensão da ferramenta de análise da quantidade de informação desenvolvido por L. Pessoa [23], o qual foi aplicado ao domínio das seqüências de DNA. A figura 7.1 mostra uma visão geral do método.

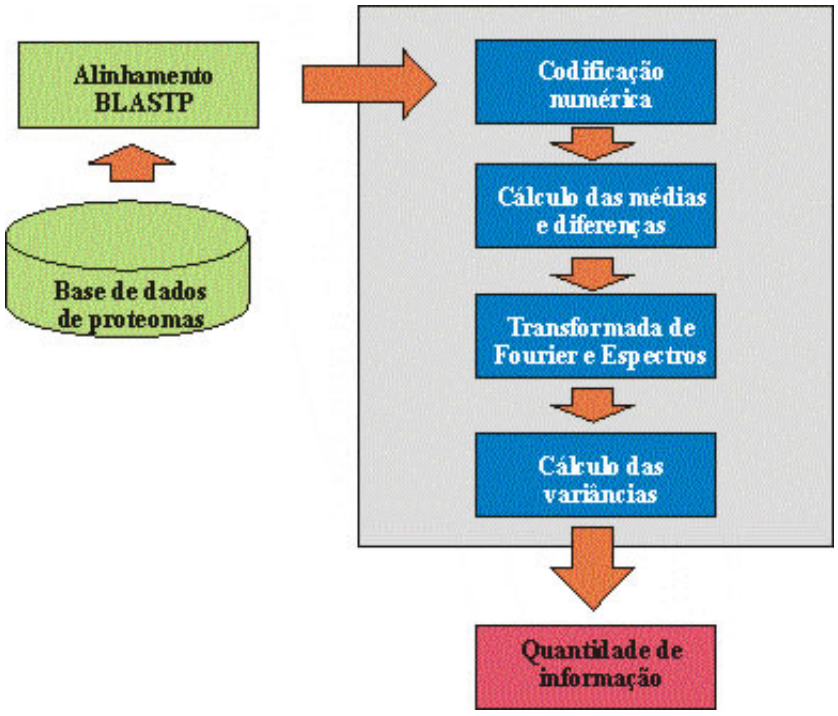


Figura 7.1: Visão geral do método para analisar a quantidade de informação em proteínas homólogas.

Uma vez obtidas seqüências de proteínas homólogas junto às bases de dados de genomas, neste trabalho extraídas do NCBI [5] e do *ExPASy Proteomics Server* [3], alinhamos as seqüências aos pares utilizando a ferramenta de alinhamento BLASTP [5]. Em seguida, aplicamos um mapeamento numérico, transformando cada seqüência de aminoácidos (cadeia de caracteres) alinhadas para codificação numérica apresentada na seção 4.3. Calculamos os sinais numéricos, média e diferença, entre os pares de aminoácidos alinhados. Em seguida, aplicamos a transformada de Fourier convertendo a representação das seqüências e dos sinais média e diferença para o domínio das freqüências. Ao final, calculamos a quantidade de informação mútua dos pares de seqüências homólogas, bem como a informação total de todo o conjunto de pares. Cada uma dessas etapas serão detalhadas e exemplificadas nas seções posteriores.

7.1 Amostras de proteínas homólogas

Para iniciar o processamento deste método é necessário adquirirmos um grupo de proteínas homólogas. Quando trabalhamos com proteínas homólogas supomos que, comparadas, elas têm um grau de similaridade superior à porcentagem de divergência requisitada pela matriz de pontuação (conforme descrito no capítulo 3). Este grau de similaridade pode ser adquirido somente após efetuarmos o alinhamento das seqüências. Portanto, a entrada para o método é um conjunto de seqüências alinhadas.

A figura 7.2 mostra o alinhamento entre dois trechos da proteína *insulin-like growth factor 1 receptor precursor*, a primeira contida na espécie *Homo sapiens* (código indetificador: $NP\_000866.1$) e a segunda da espécie *Gallus gallus* (código: $NP\_990363.1$), ambas extraídas do NCBI [5]. A primeira seqüência é denotada como $a[i]$, e a segunda $b[i]$, onde i representa cada aminoácido da seqüência, a qual possui elementos no intervalor de $i = 0, \ldots, N-1$.

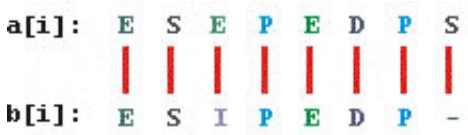


Figura 7.2: Trechos da proteína *insulin-like growth factor 1 receptor precursor* das espécies *Homo sapiens*, denotada como $a[i]$, e *Gallus gallus*, denotada por $b[i]$, alinhadas com o BLASTP [5].

7.2 Codificação numérica das seqüências homólogas alinhadas

Como mostrado na seção 4.3, a codificação numérica proposta para este trabalho utiliza matrizes contendo os valores 0 e 1. Como trabalhamos com seqüências alinhadas aos pares, teremos duas matrizes, cada uma responsável por codificar uma dada seqüência do par. Através da figura 7.3 podemos observar a representação dos trechos de proteínas utilizados na figura 7.2 (seção 7.1).

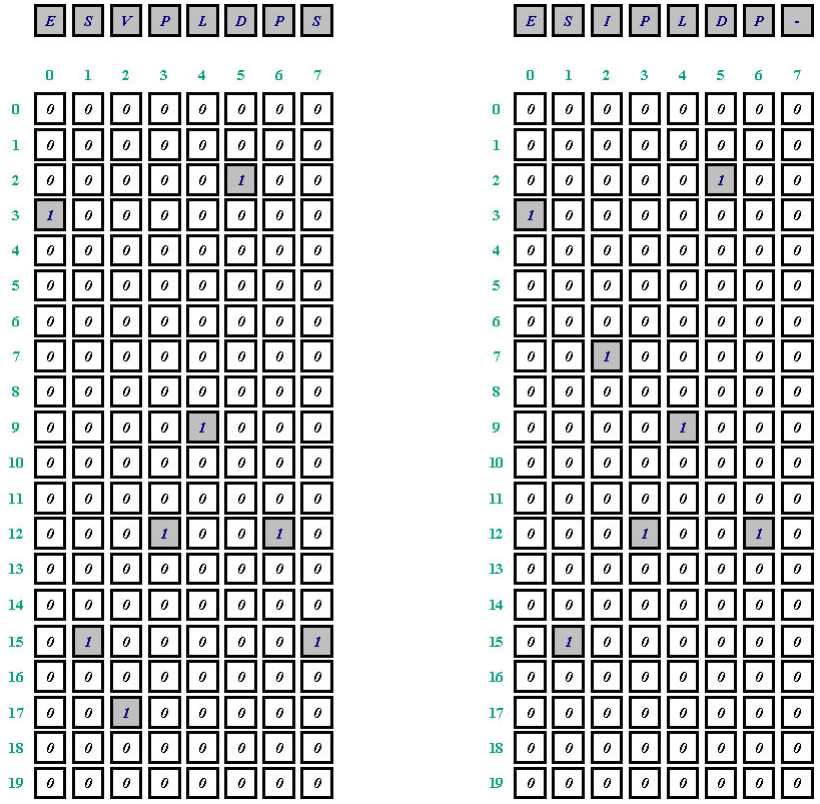


Figura 7.3: Duas seqüências $a(n)$ (matriz à esquerda) e $b(n)$ (matriz à direita) utilizando a codificação numérica em forma de matrizes de bits.

Após codificadas, as seqüências são tratadas como mensagens contendo em seu interior um conjunto de informação na forma de sinais numéricos. O par de seqüências, agora denotadas como mensagens, serão representadas pelas variáveis $a(n)$ e $b(n)$, onde n denota cada aminoácido codificado numéricamente, ou seja, cada coluna da matriz de bits. Portanto, utilizando o exemplo da figura 7.3 determinamos que a mensagem $a(n)$ possui em seu interior a seqüência ESVPLDPS representada numericamente, enquanto a mensagem $b(n)$ tem a representação numérica da seqüência ESIPLDP-.

As seqüências utilizadas neste trabalho, por serem homólogas, possuem uma seqüência ancestral em comum, que denotaremos por $s(n)$. Portanto, as seqüências $a(n)$ e $b(n)$ são

derivadas da seqüência $s(n)$ afetadas por mutações. Essas mutações, são tratadas como ruídos, representados, respectivamente, por $q'(n)$ e $q"(n)$, onde os valores contidos em $q'(n)$ determinam a diferença entre as seqüências $a(n)$ e $s(n)$ e os valores contidos em $q"(n)$ são as diferenças entre $b(n)$ e $s(n)$.

Consequentemente, cada seqüência do par é representada pelo seu ancestral adicionado de ruídos, o que nos leva a seguinte definição: $a(n) = s(n) + q'(n)$ e $b(n) = s(n) + q"(n)$. Esta característica nos habilita a tratar o par de seqüências como mensagens corrompidas por ruídos contendo em seu interior um conjunto de sinais numéricos. Tratando as seqüências como mensagens, torna-se possível aplicarmos conceitos de teoria da informação para calcularmos a quantidade de informação compartilhada por um par de seqüências biológicas.

L. Pessoa [23] propõe a equação 7.3 para calcularmos a quantidade de informação mútua entre um par de mensagens. A autora demonstra em seu trabalho que os coeficientes de Fourier têm uma distribuição gaussiana, motivo pelo qual é feita uma modificação na equação 6.14 (informação mútua), onde substituídos os elementos $H[B]$ e $H[B|A]$, respectivamente, para os apresentados nas equações 7.1 e 7.2. Devemos destacar que os elementos $\hat{Q}$ e $\hat{R}$ são os ruídos entre o sinal original $\hat{S}$ e as seqüências $\hat{A}$ e $\hat{B}$.

$$H[B] = \frac{1}{2} log_2[2\pi e(\hat{S} + \hat{R})] \tag{7.1}$$

$$H[B|A] = \frac{1}{2} log_2 \left[2\pi e \left(\frac{\hat{S}\hat{Q}}{\hat{A}} + \hat{R} \right) \right] \tag{7.2}$$

$$I(A \rightarrow B) = \frac{1}{2} log \left\{ \frac{(\hat{S} + \hat{Q})^2}{2(\hat{S}_k + \hat{Q}_k)\hat{Q}_k} \right\} \tag{7.3}$$

Podemos observar que para obtermos o resultado da equação precisamos do valor das variâncias, definidas através das variáveis $\hat{S}$ e $\hat{Q}$. Essas variâncias são adquiridas a partir das seqüências $s(n)$ (ancestral), $q'(n)$ e $q"(n)$ (ruídos), entretanto, todas essas seqüências devem estar no domínio das freqüências, denotadas neste trabalho como $S(k)$, $Q'(k)$ e $Q"(k)$.

De qualquer forma, a princípio, não temos informação sobre a seqüência ancestral ou sobre os ruídos, contudo, podemos adquirir as variâncias $\hat{S}$ e $\hat{Q}$ a partir das mensagens $a(n)$ e $b(n)$. Isto é possível através do cálculo dos sinais numéricos média e diferença.

Para obtermos esses valores, é necessário aplicarmos as equações 7.4 e 7.5 gerando como resultado, respectivamente, os sinais média $m(n)$ e diferença $d(n)$ das mensagens $a(n)$ e $b(n)$.

$$m(n) = \frac{a(n) + b(n)}{2} \tag{7.4}$$

$$d(n) = a(n) - b(n) \tag{7.5}$$

Devemos transformar os sinais média e diferença para o domínio das freqüências, para que a partir deste ponto possamos adquirir as variâncias $\hat{S}$ e $\hat{Q}$ e desta forma aplicar a equação 7.3 para obtermos a quantidade de informação mútua. A próxima seção detalha como aplicar adquirir os valores das seqüências numéricas no domínio das freqüências.

7.3 Seqüências biológicas como mensagens

Segundo L. Pessoa[23], um dos problemas na quantificação da informação de uma seqüência de DNA é a correlação existente entre os nucleotídeos próximos (responsáveis pela formação dos códons). Com esta característica, não é possível calcularmos a quantidade de informação entre as duas seqüências apenas somando a contribuição individual de cada nucleotídeo, pois a existência da correlação determina que um dado elemento apenas está presente na seqüência por intermédio de outro ocorrido previamente (elementos dependentes).

No caso dos aminoácidos até o presente momento não foram identificadas correlações deste tipo, entretanto, para estendemos a análise da quantidade de informação para o domínio dos aminoácidos, precisamos garantir que estamos trabalhando com elementos independentes. Em seu trabalho, L. Pessoa [23] mostra que os coeficientes de Fourier possuem uma distribução aleatória e independente. Por tal motivo, aplicamos a transformada de Fourier nos sinais numéricos contidos nas mensagens $a(n)$ e $b(n)$, pois, caso os aminoácidos possuam algum tipo de correlação, esta será eliminada, produzindo como resultado coeficientes independentes.

7.3.1 Transformada de Fourier aplicada às mensagens

Para calcularmos a quantidade de informação que uma determinada seqüência biológica dá sobre outra, é necessário representarmos as mensagens $a(n)$, $b(n)$, $m(n)$ e $d(n)$ no domínio das freqüências. Para isso, precisamos aplicar a transformada discreta de Fourier, abordada na seção 5.2.1.

Este procedimento processa cada elemento das mensagens acima mencionadas e gera como resultado novas seqüências, denotadas por $A(k)$, $B(k)$, $M(k)$ e $D(k)$, as quais possuem os valores representados no domínio das freqüências, denominados a partir daqui como coeficientes de Fourier.

As seqüências $A(k)$, $B(k)$, $S(k)$ e $D(k)$, a exemplo das mensagens $a(n)$, $b(n)$, $m(n)$ e $d(n)$, são matrizes com o mesmo número de linhas e colunas, assim como os mesmos elementos, entretanto, representados em um domínio diferente. Os elementos das novas seqüências estão contidos em intervalos de valores denominados de componentes de freqüência, ou seja, k.

Conforme definido na seção 5.2.1, para adquirirmos os sinais numéricos originais, ou seja, $a(n)$, $b(n)$, $m(n)$ e $d(n)$, basta aplicarmos a transformada inversa de Fourier sobre as seqüências $A(k)$, $B(k)$, $M(k)$ e $D(k)$ multiplicadas ao coeficiente de normalização.

As seqüências $A(k)$ e $B(k)$ são equivalentes às seqüências $a(n)$ e $b(n)$, embora representados em domínios diferentes. Desta forma, podemos tratá-las de forma análoga, ou seja, como mensagens corrompidas por ruídos denotadas por $A(k) = S(k) + Q'(k)$ e $B(k) = S(k) + Q"(k)$. Ao término dessa fase é necessário calcularmos o espectro de potência definido na seção 7.3.2.

7.3.2 Cálculo do espectro de potência

Uma vez de posse das mensagens $A(k)$, $B(k)$, $M(k)$ e $D(k)$ aplicamos o cálculo da densidade de potência espectral em cada elemento da seqüência (definido na seção 5.2.2), no intervalo de $k = 0, \ldots, N-1$. Este processo tem como objetivo potencializar o sinal, no intuito de adquirimos uma melhor visualização das senóides geradas a partir dos coeficientes de Fourier das seqüências.

Neste momento as mensagens $A(k)$, $B(k)$, $M(k)$ e $D(k)$ cuja quantidade de elementos é equivalente a N tem seu número reduzido pela metade, isto é, termos elementos no intervalo de freqüência $k = 0, \ldots, \frac{N}{2} - 1$. Isto se deve ao fato dos valores contidos no

intervalo de $k = 0, \ldots, \frac{N}{2} - 1$ serem equivalentes aos contidos no intervalo inverso $k = \frac{N}{2}, \ldots, N-1$, ou seja, $k = N-1, \ldots, \frac{N}{2}$ (denominadas freqüências negativas, ou $-k$).

7.3.3 Cálculo das variâncias e quantidade de informação de proteínas homólogas

Após utilizarmos a transformada de Fourier, temos os valores das seqüências $A(k)$ e $B(k)$, bem como a média e a diferença $M(k)$ e $D(k)$, estas duas últimas, calculadas a partir dos sinais numéricos $m(n)$ e $d(n)$. A média e a diferença dos sinais no domínio das freqüências, $M(k)$ e $D(k)$ são representados, respectivamente, pelas equações 7.6 e 7.7.

$$M(k) = \frac{A(k) + B(k)}{2} \tag{7.6}$$

$$D(k) = A(k) - B(k) \tag{7.7}$$

ou seja,

$$M(k) = S(k) + \frac{Q'(k) + Q"(k)}{2} \tag{7.8}$$

$$D(k) = Q'(k) - Q"(k) \tag{7.9}$$

A partir das equações da média e diferença podemos obter os valores das variâncias $\hat{S}$ e $\hat{Q}$. A variância $\hat{Q}$ representa, por simetria, ambos os ruídos Q' e $Q"$, sendo que fazendo as devidas modificações nas equações 7.6 e 7.7 adquirimos os valores de $\hat{S}$ e $\hat{Q}$ obtemos as equações 7.12 e 7.13.

$$\hat{M}(k) = \hat{S}(k) + \frac{1}{2}\hat{Q}(k) \tag{7.10}$$

$$\hat{D}(k) = 2\hat{Q}(k) \tag{7.11}$$

Com isso, temos:

$$\hat{S}(k) = \hat{M}(k) - \frac{1}{4}\hat{D}(k) \tag{7.12}$$

$$\hat{Q}(k) = \frac{1}{2}\hat{D}(k) \tag{7.13}$$

7.4 Cálculo da quantidade de informação mútua

A partir das variâncias, adquiridas através das equações 7.12 e 7.13, podemos adaptar a equação 7.3 e aplicá-la a fim de adquirir a quantidade de informação mútua entre um par de seqüências (denotadas como mensagens). Antes de procedermos com o cálculo, devemos nos ater a uma característica produzida quando calculamos os espectros, que são as componentes de freqüência $k = 0$, $k = \frac{N}{2} - 1$ e as demais contidas no intervalo $0 < k < \frac{N}{2} - 1$.

As componentes de freqüência devem ser tratadas de forma diferente, pois no caso das freqüências contidas no intervalo $0 < k < \frac{N}{2} - 1$ combinamos as freqüências positivas k com as negativas $-k$. Portanto, quando trabalhamos com as freqüências $k = 0$ e $k = \frac{N}{2} - 1$ utilizamos a equação 7.14, caso contrário, ou seja, $0 < k < \frac{N}{2} - 1$, utilizamos a equação 7.15. Cada um desses valores, ou seja, I em cada componente de freqüência k ($I(k)$), devem ser multiplicados por $\frac{1}{20N}$, onde 20 representa a quantidade de símbolos e N a quantidade total de aminoácidos em uma dada seqüência.

$$I(k) = log\left[\frac{\hat{M}(k)}{\hat{D}(k)} + \frac{1}{2} + \frac{1}{16}\frac{\hat{D}(k)}{\hat{M}(k)}\right] \tag{7.14}$$

$$I(k) = 2log\left[\frac{\hat{M}(k)}{\hat{D}(k)} + \frac{1}{2} + \frac{1}{16}\frac{\hat{D}(k)}{\hat{M}(k)}\right] \tag{7.15}$$

Calculada a informação mútua para cada par de seqüências, podemos agora adquirir a informação mútua total do conjunto de pares de seqüências analisada. Para isso podemos aplicar a equação 7.16.

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

Capítulo 8

Resultados experimentais

Neste capítulo vamos mostrar os resultados obtidos com a aplicação do método proposto neste trabalho. A primeira característica a ser destacada é o fato de trabalharmos com seqüências de todas as espécies de organismos, tanto eucariotos quanto procariotos. Tal característica se deve ao fato de estarmos analisando seqüências de aminoácidos, ou seja, proteínas já codificadas. Diante deste fato, vamos apresentar os grupos de amostras utilizadas para o experimento, bem como descrever detalhadamente os resultados obtidos através de cada um deles.

8.1 Seleção das Amostras

Ao contrário das seqüências de DNA, as proteínas têm seu tamanho reduzido a no máximo milhares de aminoácidos (número dificilmente alcançado). Tal característica dificulta a seleção das amostras, principalmente quando determinamos a escolha baseada em uma dada faixa percentual de similaridade. Por causa do tamanho reduzido não podemos efetuar cortes nas seqüências de aminoácidos, como realizado com cadeias de DNA quando utilizarmos quadros de leitura, pois, as seqüências que são naturalmente pequenas, tornam-se ainda menores. E este número, após calcularmos o espectro de potência, ainda é reduzido pela metade.

Selecionamos as amostras a partir dos grupos de proteínas mostrados na tabela 8.1. Através das proteínas desses grupos, adquirirmos diversas outras proteínas homólogas cuja similaridade mostrou-se acima de 80%. As amostras obtidas foram agrupadas de acordo com o grupo de proteínas. Desta forma, adquirimos proteínas similares à ferritina e agrupamos em um diretório responsável por armazená-las. Este processo foi repetido para cada grupo de proteínas conforme listado na tabela 8.1.

Proteínas	**Proteínas**
Actin	Ferritin Heavy Chain
Aquaporim	Ferritin Ligth Chain
Casein	Glucagon
Creatine Kinase	Insulin
Elastin	Myosin

Tabela 8.1: Grupos de proteínas utilizadas para a análise.

A aplicação processa as proteínas por grupos, sendo que desta forma, a quantidade de informação total será dada a partir do grupo de proteína analisada. Por exemplo, analisamos inicialmente o grupo da proteína actina. Após o processamento, adquirimos uma quantidade de informação igual a x bits. Em seguida passamos para o segundo conjunto de proteínas, aquaporina, onde processamos a análise e geramos a quantidade de informação do grupo.

Este processo de quantificação da informação é executado até serem analisados todos os grupos de proteínas listados na tabela 8.1. Devemos ressaltar que cada grupo de proteína contém uma média de 10 pares de seqüências alinhadas e que após calculada a quantidade de informação de cada grupo de proteínas, procedemos com o cálculo da quantidade de informação média de todos os grupos.

8.2 Características da aplicação

A aplicação foi desenvolvida na linguagem C, utilizando o compilador gnu gcc 4.3, e executada no sistema operacional SUSE LINUX 10.1. Para a implementação da transformada de Fourier utilizamos a biblioteca FFTW 3.1.[4], a qual é responsável por possibilitar o cálculo da transformada de domínios a partir da seqüência valores numéricos. No caso deste trabalho, esta seqüência é composta por um conjunto de aminoácidos codificados numericamente conforme mostrado na seção 4.3.

Os dados numéricos de entrada, utilizados pela FFTW, podem estar representados de duas formas: ou através de números reais ou em números complexos. Da mesma forma a transformada produz qualquer um dos dois tipos de representação como resultado. Independente do tipo de representação ou função utilizada, os algoritmos de transformadas da biblioteca FFTW executam em $O(nlogn)$, sendo n o número de elmentos presentes na seqüência.

No método utilizado, para obtermos a quantificação da informação, trabalhamos com

um conjunto de números reais (seqüências em forma de mensagens contendo sinais numéricos) de entrada e obtemos como saída uma seqüência de números reais. Por tal motivo, optamos pela utilização da transformada de seno/co-seno, com parâmetro denominado REDFT00 (DCT-I).

8.3 Pré-requisitos para a execução do método

A princípio, identificamos qual das seqüências contidas no conjunto analisado apresentava a menor quantidade de aminoácidos. Este procedimento foi necessário para que o tamanho destas proteínas - após o alinhamento - fosse definido como o ponto de corte para as demais, ou seja, nenhuma seqüência poderia ter um tamanho superior ao ponto de corte identificado pelo programa.

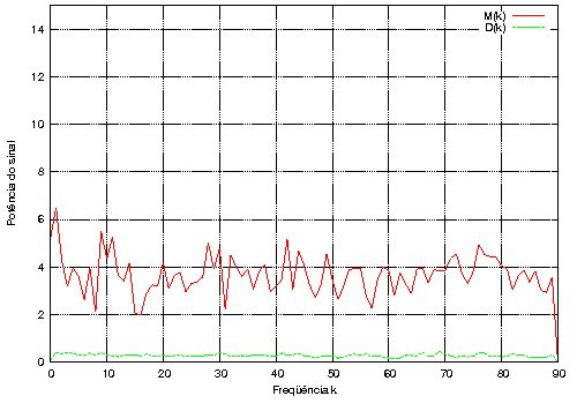
Outra característica importante foi quanto ao ponto de início (partida) da seqüência, isto é, a partir de que posição as seqüências analisadas sejam processadas. Intuitivamente, iniciaríamos a análise da mesma forma que nos alinhamentos (primeira posição). Entretanto, para não cairmos em falsos ótimos ou termos algum tipo de coincidência, o que poderia gerar uma tendência na análise, definimos que o início (ponto de partida) seria gerado de forma aleatória.

8.4 Processo de análise das mensagens

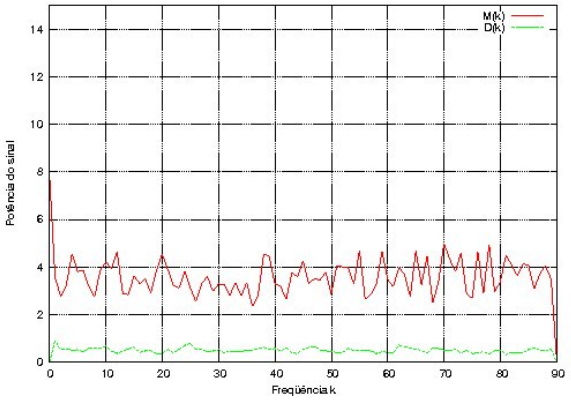
O processo de análise inicia-se com a codificação numérica das seqüências, após termos organizado as amostras em grupos, conforme descrito na seção 8.1 e executados os devidos ajustes quanto ao ponto de partida e o corte, mostrados nas seções 8.2 e 8.3. A codificação numérica gera as mensagens $a(n)$ e $b(n)$, ambas contendo os sinais numéricos na forma de matrizes de bits. A partir desta definição geramos os sinais média e diferença $m(n)$ e $d(n)$ de cada uma das seqüências contidas nos grupos de proteínas.

A partir dos sinais numéricos média e diferença obtivemos os sinais $M(k)$ e $D(k)$, ambos representados no domínio das freqüências. Com base em nossas amostras, geramos os espectros de potência para os conjuntos de proteínas listadas na tabela 8.1, tendo como resultado os gráficos mostrados através das figuras 8.1 e 8.2.

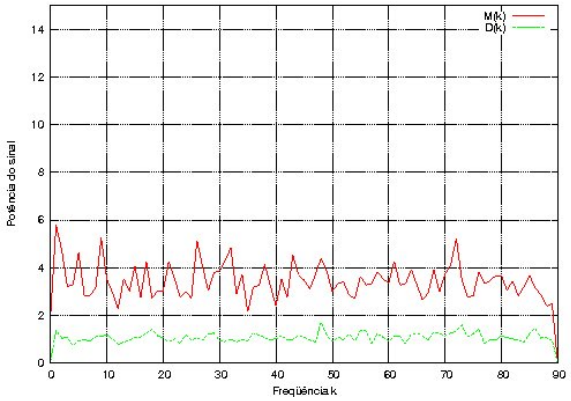
Quando utilizamos o ponto de corte, buscamos encontrar a menor quantidade de aminoácidos, sendo que a partir disto definimos este valor para todas as proteínas. Desta forma, por exemplo, o tamanho da ferritina que é de 90 aminoácidos é definido para todos



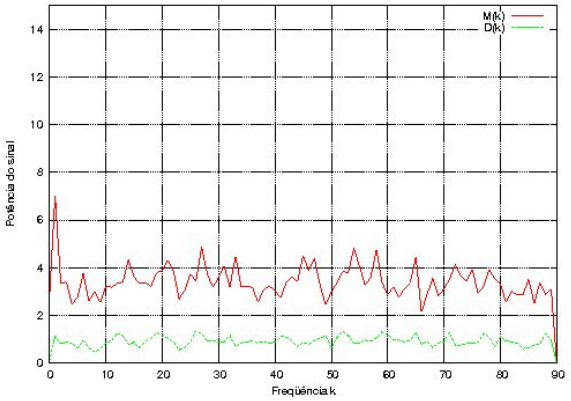
(a) Sinais média e diferença do conjunto de proteínas homólogas geradas a partir da Actina



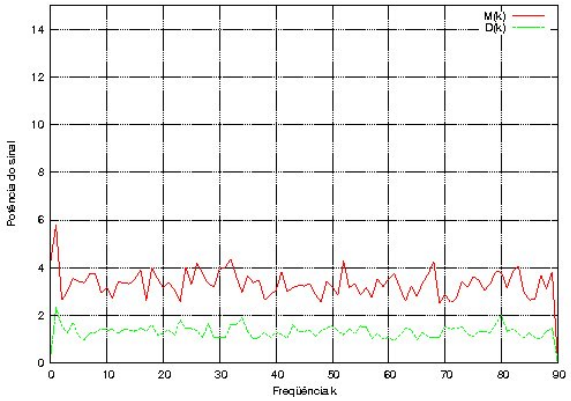
(b) Sinais média e diferença do conjunto de proteínas homólogas geradas a partir da Aquaporina



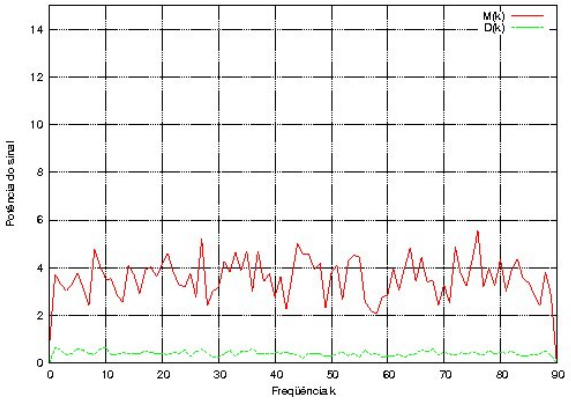
(c) Sinais média e diferença do conjunto de proteínas homólogas geradas a partir da Caseína



(d) Sinais média e diferença do conjunto de proteínas homólogas geradas a partir da Creatina

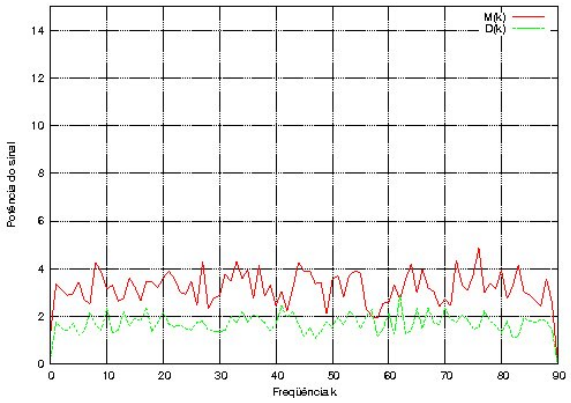


(e) Sinais média e diferença do conjunto de proteínas homólogas geradas a partir da Elastina

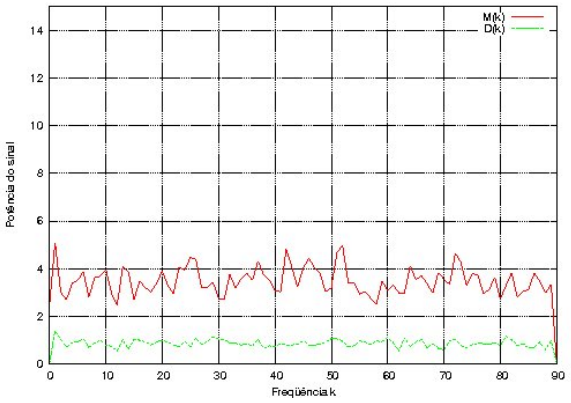


(f) Sinais média e diferença do conjunto de proteínas homólogas geradas a partir da Ferritina (corrente forte)

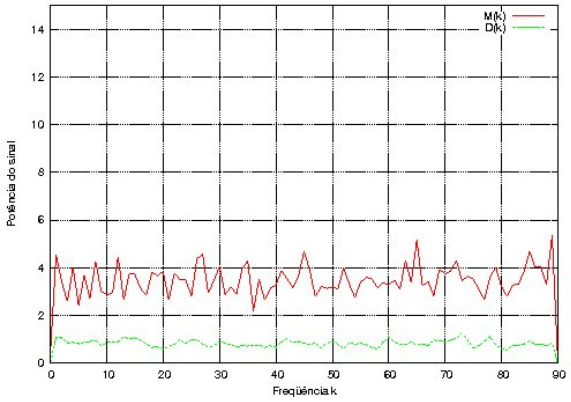
Figura 8.1: Espectros de potência do conjunto de proteínas utilizadas nos experimentos.



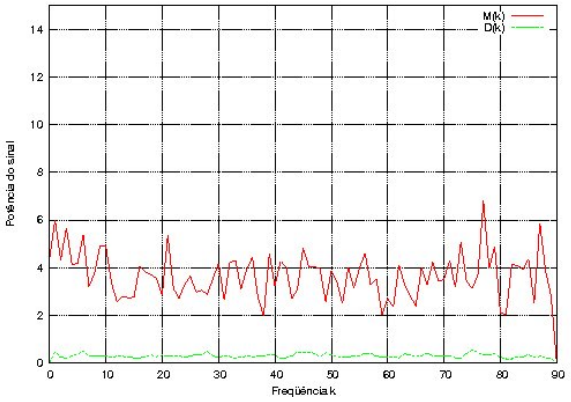
(a) Sinais média e diferença do conjunto de proteínas homólogas geradas a partir da Ferritina (corrente fraca)



(b) Sinais média e diferença do conjunto de proteínas homólogas geradas a partir da Glucagon



(c) Sinais média e diferença do conjunto de proteínas homólogas geradas a partir da Insulina



(d) Sinais média e diferença do conjunto de proteínas homólogas geradas a partir da Miosina

Figura 8.2: Espectros de potência do conjunto de proteínas utilizadas nos experimentos.

os demais (actina, glucagon, insulina e etc).

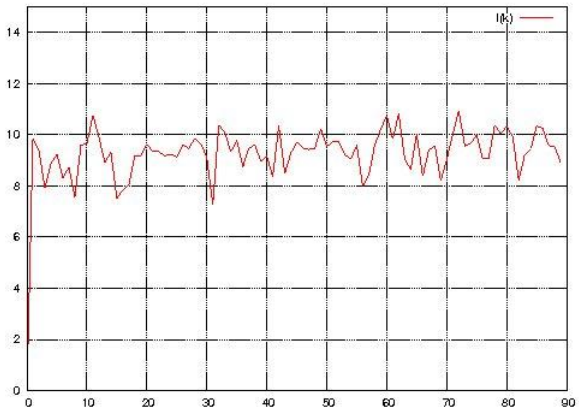
8.5 Quantidade de informação mútua de seqüências de proteínas homólogas

Após calcularmos o espectro de potência, temos por definição que as seqüências - representadas no domínio das freqüências - têm seus tamanhos reduzidos pela metade, isto é, eles passam a ter $\frac{N}{2}$ elementos ao invés de N. Com base nas características apresentadas até este ponto, aplicamos as equações 7.14 e 7.15 para obtermos a quantidade de informação mútua por componente, ou seja, quanto de informação uma primeira seqüência dá sobre uma segunda, alinhada a ela.

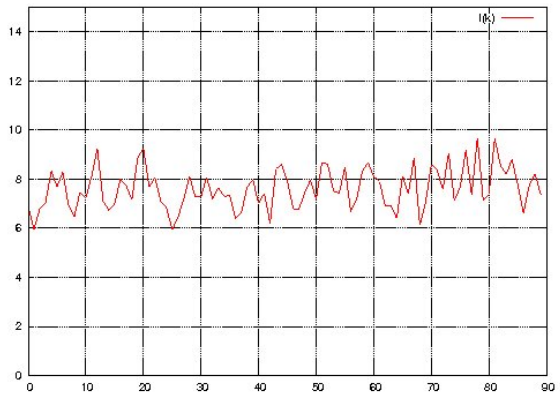
Podemos observar alguns componentes de freqüência com suas respectivas variâncias e informação através das tabelas contidas na figura 8.5. Os valores das distribuições das informações por componente de cada grupo de proteínas utilizados neste trabalho foram plotados e são apresentados através dos gráficos das figuras 8.3 e 8.4.

Com esta definição, temos diferentes quantidades de informação por componente para cada grupo de proteínas, e consequentemente, diferentes quantidades totais de informação. As tabelas da figura 8.5 mostram a quantidade de informação mútua total para cada conjunto de proteínas utilizados nos experimentos, sendo que a partir destes valores adquirimos a quantidade de informação total de todos os pares contidos no grupo de proteínas, bem como a contribuição de cada aminoácido para a aquisição da informação.

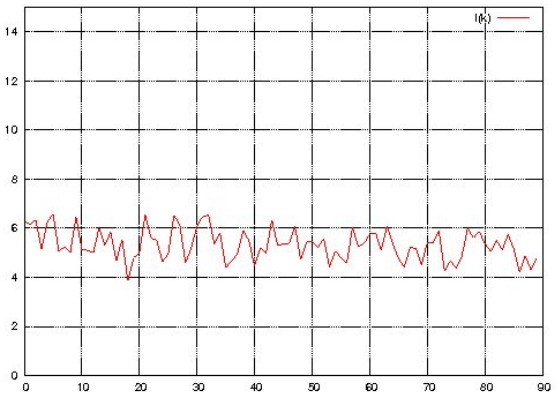
Baseado nas informações adquiridas através das seqüências analisadas distribuímos a informação sobre cada grupo de proteínas conforme a tabela 8.5. Aqui são apresentados os grupos de seqüências biológicas utilizadas nos experimentos (considerando o ponto de corte igual a 180 aminoácidos). Para cada grupo são identificadas a quantidade de informação por componente, ou seja, o quanto de informação uma cadeia dá sobre a outra (por par de seqüências), a contribuição da informação dada por cada aminoácido e a quantidade total de informação produzida por todos os pares contidos em cada grupo.



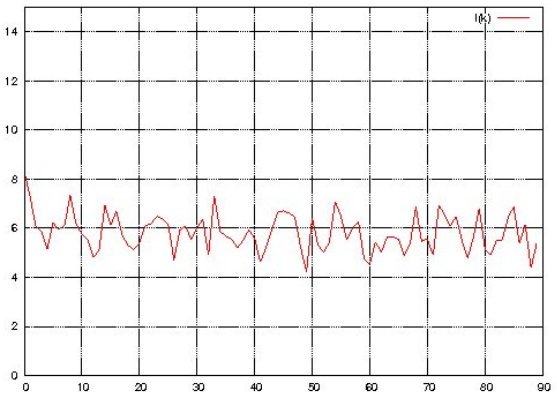
(a) Informação gerada a partir da actina.



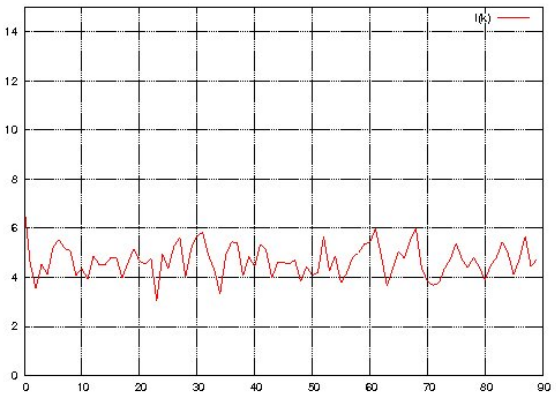
(b) Informação gerada a partir da aquaporina.



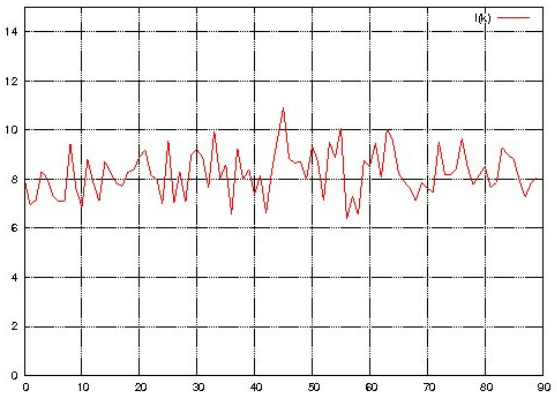
(c) Informação gerada a partir da caseína.



(d) Informação gerada a partir da creatina.

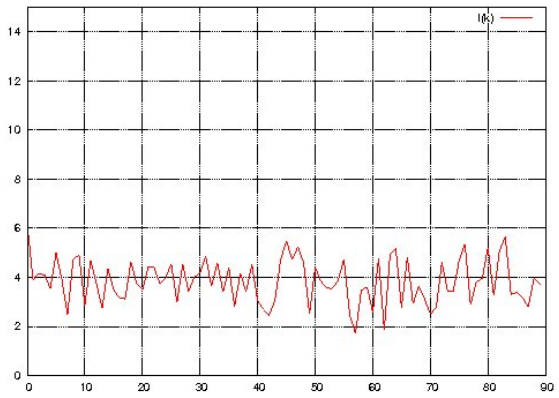


(e) Informação gerada a partir da elastina.

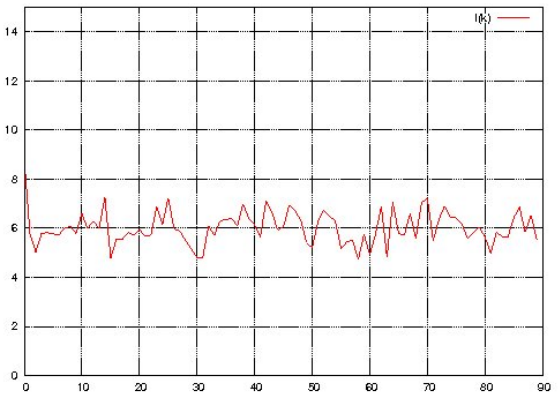


(f) Informação gerada a partir da ferritina (corrente forte).

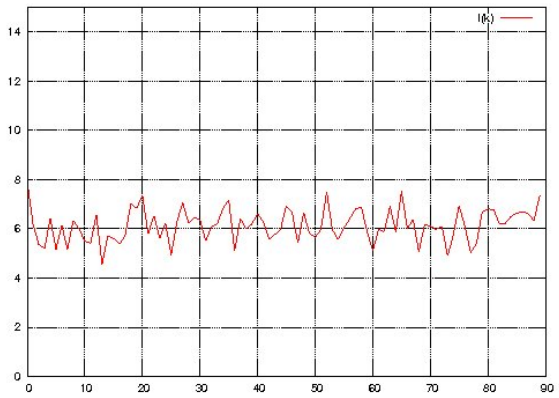
Figura 8.3: Quantidade de informação esperada por componente de freqüência para os conjuntos de proteínas alinhadas.



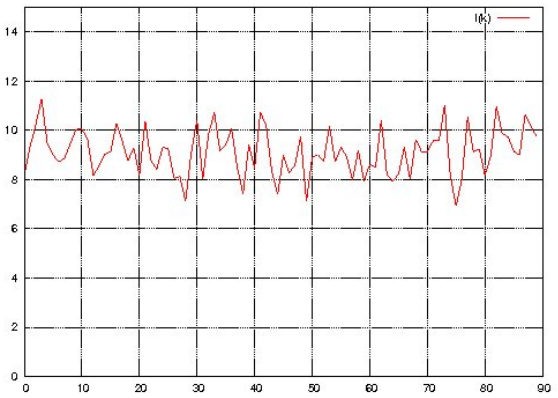
(a) Informação gerada a partir da ferritina (corrente fraca).



(b) Informação gerada a partir da glucagon.



(c) Informação gerada a partir da insulina.



(d) Informação gerada a partir da miosina.

Figura 8.4: Quantidade de informação esperada por componente de freqüência para os conjuntos de proteínas alinhadas.

k	$\hat{M}_k$	$\hat{D}_k$	I_k
...	...	...	...
1	6.41	0.21	9.85
2	4.20	0.17	9.38
3	3.09	0.20	7.93
4	3.87	0.18	8.89
5	3.52	0.15	9.23
...	...	...	...
41	3.38	0.19	8.36
42	5.09	0.14	10.36
43	2.98	0.16	8.49
44	4.59	0.19	9.28
45	4.00	0.14	9.72
...	...	...	...
85	3.32	0.09	10.34
86	3.79	0.11	10.26
87	2.98	0.11	9.58
88	2.89	0.11	9.53
89	3.50	0.16	8.91
Total:			740.01

(a) Quantidade de informação mútua por componente do conjunto de proteínas homólogas geradas a partir da actina.

k	$\hat{M}_k$	$\hat{D}_k$	I_k
...	...	...	...
1	3.35	0.45	5.97
2	2.62	0.26	6.80
3	3.12	0.28	7.05
4	4.43	0.25	8.33
5	3.68	0.26	7.70
...	...	...	...
41	3.07	0.24	7.43
42	2.49	0.31	6.20
43	3.66	0.20	8.42
44	3.51	0.18	8.60
45	4.13	0.27	7.92
...	...	...	...
85	3.93	0.27	7.78
86	2.93	0.31	6.62
87	3.60	0.25	7.75
88	3.94	0.23	8.22
89	3.35	0.27	7.38
Total:			686.13

(b) Quantidade de informação mútua por componente do conjunto de proteínas homólogas geradas a partir da aquaporina.

k	$\hat{M}_k$	$\hat{D}_k$	I_k
...	...	...	...
1	5.46	0.68	6.17
2	4.47	0.53	6.34
3	2.95	0.54	5.16
4	3.11	0.38	6.26
5	4.39	0.47	6.58
...	...	...	...
41	3.24	0.58	5.22
42	2.51	0.49	4.98
43	4.29	0.50	6.35
44	3.41	0.59	5.30
45	3.21	0.55	5.36
...	...	...	...
85	3.38	0.61	5.18
86	2.79	0.74	4.21
87	2.60	0.53	4.88
88	2.12	0.54	4.30
89	2.26	0.47	4.80
Total:			480.29

(c) Quantidade de informação mútua por componente do conjunto de proteínas homólogas geradas a partir da caseína.

Tabela 8.2: Quantidade de informação mútua por componente dos grupos de proteínas: actina, aquaporina e caseína.

k	$\hat{M}_k$	$\hat{D}_k$	I_k
...	...	...	...
1	6.75	0.56	7.30
2	3.15	0.41	6.08
3	3.18	0.45	5.87
4	2.28	0.42	5.16
5	2.61	0.32	6.23
...	...	...	...
41	2.47	0.56	4.62
42	3.09	0.55	5.22
43	3.38	0.46	5.94
44	3.26	0.34	6.65
45	4.29	0.44	6.72
...	...	...	...
85	3.36	0.32	6.88
86	2.31	0.39	5.39
87	3.18	0.40	6.15
88	2.56	0.63	4.38
89	2.88	0.48	5.40
Total:			523.96

(a) Quantidade de informação mútua por componente do conjunto de proteínas homólogas geradas a partir da creatina.

k	$\hat{M}_k$	$\hat{D}_k$	I_k
...	...	...	...
1	5.21	1.16	4.64
2	2.25	0.78	3.53
3	2.69	0.63	4.52
4	3.14	0.86	4.13
5	3.15	0.56	5.22
...	...	...	...
41	3.53	0.60	5.36
42	2.75	0.51	5.12
43	2.76	0.80	3.97
44	2.93	0.66	4.60
45	2.91	0.66	4.60
...	...	...	...
85	2.32	0.64	4.11
86	2.45	0.52	4.76
87	3.44	0.52	5.67
88	2.77	0.67	4.43
89	3.48	0.74	4.75
Total:			422.88

(b) Quantidade de informação mútua por componente do conjunto de proteínas homólogas geradas a partir da elastina.

k	$\hat{M}_k$	$\hat{D}_k$	I_k
...	...	...	...
1	3.56	0.33	6.97
2	3.18	0.28	7.14
3	2.95	0.17	8.32
4	3.21	0.21	8.03
5	3.64	0.30	7.32
...	...	...	...
41	3.51	0.21	8.17
42	2.13	0.23	6.59
43	3.51	0.20	8.35
44	4.93	0.18	9.68
45	4.50	0.10	10.93
...	...	...	...
85	3.30	0.16	8.84
86	2.79	0.19	7.93
87	2.30	0.19	7.28
88	3.72	0.25	7.84
89	2.73	0.17	8.06
Total:			741.86

(c) Quantidade de informação mútua por componente do conjunto de proteínas homólogas geradas a partir da ferritina (corrente forte).

Tabela 8.3: Quantidade de informação mútua por componente dos grupos de proteínas: creatina, elastina e (ferritina corrente forte).

k	$\hat{M}_k$	$\hat{D}_k$	I_k
...	...	...	...
1	2.93	0.88	3.90
2	2.77	0.75	4.15
3	2.55	0.70	4.10
4	2.52	0.86	3.56
5	3.14	0.61	4.99
...	...	...	...
41	2.46	1.22	2.69
42	1.75	0.97	2.46
43	2.62	1.12	3.04
44	3.86	0.84	4.72
45	3.60	0.58	5.49
...	...	...	...
85	2.45	0.90	3.38
86	2.20	0.89	3.18
87	1.97	0.94	2.78
88	3.15	0.91	3.99
89	2.18	0.70	3.71
Total:			346.44

(a) Quantidade de informação mútua por componente do conjunto de proteínas homólogas geradas a paritr da ferritina (corrente fraca).

k	$\hat{M}_k$	$\hat{D}_k$	I_k
...	...	...	...
1	4.76	0.70	5.75
2	2.75	0.53	5.03
3	2.51	0.36	5.80
4	3.17	0.45	5.82
5	3.28	0.48	5.77
...	...	...	...
41	2.82	0.43	5.64
42	4.63	0.41	7.14
43	3.87	0.40	6.66
44	3.04	0.42	5.93
45	3.78	0.50	6.05
...	...	...	...
85	2.96	0.34	6.42
86	3.64	0.35	6.87
87	3.25	0.46	5.84
88	2.82	0.31	6.54
89	3.12	0.49	5.54
Total:			542.26

(b) Quantidade de informação mútua por componente do conjunto de proteínas homólogas geradas a paritr do glucagon.

k	$\hat{M}_k$	$\hat{D}_k$	I_k
...	...	...	...
1	4.29	0.53	6.19
2	3.14	0.53	5.36
3	2.38	0.43	5.21
4	3.80	0.44	6.42
5	2.22	0.41	5.17
...	...	...	...
41	3.67	0.44	6.27
42	3.29	0.51	5.57
43	2.95	0.43	5.77
44	3.42	0.46	5.99
45	4.48	0.42	6.94
...	...	...	...
85	4.44	0.47	6.63
86	3.83	0.39	6.71
87	3.84	0.41	6.62
88	3.11	0.37	6.32
89	5.19	0.42	7.37
Total:			553.29

(c) Quantidade de informação mútua por componente do conjunto de proteínas homólogas geradas a partir da insulina.

Tabela 8.4: Quantidade de informação mútua por componente dos grupos de proteínas: ferritina (corrente fraca), glucagon e insulina.

k	$\hat{M}_k$	$\hat{D}_k$	I_k
...	...	...	...
1	5.88	0.24	9.33
2	4.27	0.12	10.24
3	5.63	0.11	11.29
4	4.06	0.15	9.48
5	4.09	0.18	9.01
...	...	...	...
41	4.23	0.10	10.73
42	3.99	0.12	10.23
43	2.62	0.15	8.31
44	3.00	0.24	7.39
45	4.73	0.21	9.00
...	...	...	...
85	4.28	0.18	9.16
86	2.47	0.11	8.99
87	5.81	0.15	10.65
88	3.84	0.11	10.17
89	2.74	0.10	9.74
Total:			750.60

Tabela 8.5: Quantidade de informação mútua por componente do conjunto de proteínas homólogas geradas a paritr da miosina.

	Informação		
Grupo de proteína	**Componente**	**Aminoácido**	**Total**
Actin	52.02	4.30	740.01
Aquaporim	114.35	3.81	686.13
Casein	30.01	2.66	480.29
Creatine Kinase	87.32	2.91	523.96
Elastin	70.48	2.34	422.88
Ferritin Heavy Chain	123.64	4.12	741.86
Ferritin Ligth Chain	43.30	1.92	346.44
Glucagon	54.26	3.01	542.60
Insulin	92.21	3.07	553.29
Myosin	136.57	4.17	750.60

Tabela 8.6: Quantidade de informação por componente, por aminoácido (contribuição individual de cada elemento) e a informação total adquirida através de cada grupo de proteínas utilizada nos experimentos.

Capítulo 9

Conclusão e Trabalhos Futuros

Neste trabalho estudamos a quantificação de informação mútua em proteínas homólogas, onde tratamos as cadeias de aminoácidos como mensagens corrompidas por ruídos (modificações nas seqüências por conseqüência da evolução). Tal tratamento foi necessário a fim de que fosse possível aplicar os conceitos de teoria da informação, e calcular a quantidade de informações seqüências biológicas, tratadas como mensagens.

Através de conceitos de processamento de sinais foi possível efetuarmos uma análise visual do comportamento das seqüências em cada componente de freqüência. Analisamos os espectros dos coeficientes de Fourier e da informação por componente gerados a partir do grupo de seqüências utilizadas nos experimentos, conforme mostrado no capítulo 8.

Através das equações e métodos apresentadas neste trabalho propomos uma forma de avaliar a quantidade de perda de informação (em bits) que um conjunto de seqüências obteve em anos de evolução em relação a seqüência original. Com isso, por exemplo, levando em consideração que os mesmos aminoácidos tem igual distribuição na seqüências, teríamos uma quantidade de informação de $4,32$ bits por aminoácido, ou seja, $log_2 \frac{1}{20}$. Contudo, observamos através dos resultados que tal valor foi minimizado em alguns casos para até $1,9$ bits, ou seja, ocorreu uma perda menor de informação das seqüências analisadas em relação à original.

Como mencionado, este trabalho é uma extensão do método proposto por L. Pessoa[23] mas, aplicado a um conjunto de aminoácidos em vez de nulceotídeos. Os resultados foram mostrados e discutidos.

Este trabalho tem como maior contribuição a proposta de um novo método para a análise de proteínas homólogas, complementando o trabalho de L. Pessoa[23], que tinha como uma das principais sugestões para trabalhos futuros o desenvolvimento a extensão de sua ferramenta de análise para o domínio das proteínas.

9.1 Trabalhos Futuros

A codificação numérica pode não ser adequada para aplicação do método, vindo a ser interessante a utilização de novas formas de codificação, tais como a apresentada por D. Anastassiou[7]. Entretanto, esta codificação leva em consideração a utilização de seqüências de DNA juntamente com as seqüências de aminoácidos para a representação das proteínas, portanto, para realizar a análise devemos trabalhar com ambas em conjunto. A principal vantagem desta codificação é o fato de podermos aplicar a representação através de números complexos, a qual foi utilizada para a codificação de seqüências de DNA.

Referências

[1] Dna data bank of japan. `http://www.ddbj.nig.ac.jp/`.

[2] European bioinformatics institute. `http://www.ebi.ac.uk/embl/index.html`.

[3] Expasy proteomics server. `http://ca.expasy.org/`.

[4] Fast fourier transform in the west version 3.1. `http://www.fftw3.org`.

[5] National center for biotechnology information. `http://www.ncbi.nlm.nih.gov/`.

[6] Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Robertis, K., and Walter, P. *Essential cell biology: an introduction to the molecular biology of the cell*. Garland Publishing, 1998.

[7] Anastassiou, D. Genomic signal processing. *IEEE Signal Processing Magazine theme article 18*, 4 (2001), 8–20.

[8] Chalco, J. P. M. Identificação de regiões codificadoras de proteína através da transformada modificada de morlet. Tese de Mestrado, Intituto de Matemática e Estatística, Universidade de São Paulo, 2005.

[9] Cover, T. M., and Thomas, J. A. *Elements of Information Theory*. Wiley, 1991.

[10] Cristea, P. D. Complex genomic signals. *VIPromCom, 4th EURASIP - IEEE Region 8 International Symposium on Video/Image Processing and Multimedia Communications* (June 2002), 16–19.

[11] Cristea, P. D. Large scale features in dna genomic signals. *Signal Processing* (2002), 871–888.

[12] de Robertis, E. D. P., and de Robertis Jr., E. M. F. *Fundamentos de Biologia Celular y Molecular*. EL ATENEO - Pedro Garcia S. A., 1989.

[13] Diniz, P. S. R., Silva, E. A. B., and Netto, S. L. *Processamento Digital de Sinais: Projeto e Análise de Sistemas*. Bookman, Porto Alegre, Rio Grande do Sul, Brasil, 2004.

[14] Dougherty, E. R., Shmulevich, I., Chen, J., and Wang, Z. J. Genomic signal processing: Perspectives. In *Genomic Signal Processing and Statistics* (New York, NY, USA, 2005), Hindawi Publishing Corporation, pp. 1–14.

[15] Grosse, I., Buldyrev, S. V., and Stanley, H. E. Average mutual information of coding and noncoding dna. *VIPromCom, 4th EURASIP - IEEE Region 8 International Symposium on Video/Image Processing and Multimedia Communications* (June 2002), 16–19.

[16] Herzel, H., Ebeling, W., and Schimitt, A. O. Entropies of biosequence: The role of repeats. *Physical review E 50*, 6 (1994).

[17] Junqueira, L. C., and Carneiro, J. *Biologia Molecular e Celular.* Guanabara Koogan, 2000.

[18] Koonin, V., E., and Galperin, M. Y. *Computational Approaches in Comparative Genomics.* Norwell (MA): Kluwer Academic Publishers, 2003.

[19] Lathi, B. P. *Modern Digital and Analog Communication Systems.* Oxford, 1998.

[20] Macciato, M. F., Cuomo, V., and Tramontano, A. Determination of the autocorrelation orders of proteins. *Eur. J. Biochem*, 149 (June 1985), 375–379.

[21] Meidanis, J., and Setubal, J. C. *Introduction to Computational Molecular Biology.* PWS Publisher, 1997.

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

[23] Pessoa, L. S. Análise da informação mútua em seqüências de dna homólogas. Tese de Mestrado, Instituto de Computação, Universidade Federal Fluminense, 2004.

[24] Pevsner, J. *Bioinformatics and Functional Genomics.* PWS Publishers, 1998.

[25] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, 1992.

[26] Stein, J. Y. *Digital Signal Processing: A Computer Science Perspective.* Wiley Interscience publication, 2000.

[27] Vaidyanathan, P. P. Genomics and proteomics: A signal processor's tour. *IEEE Circuits and Systems Magazine* (2004).

[28] Watson, J. D., and Crick, F. H. Letters to nature: Molecular structure of nucleic acid. *Nature*, 171 (1953), 737–738.

[29] Weiss, O., Montano, M. A. J., and Herzel, H. Information content of protein sequences. *J. Theor. Biol.* (2000), 379–386.