Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Visualização em Grafos de Redes Metabólicas via Web

Gabriella de Oliveira Esteves

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr.ª Maria Emília Machado Telles Walter

Brasília
2016

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Bacharelado em Ciência da Computação

Coordenador: Prof. Dr. Rodrigo Bonifácio de Almeida

Banca examinadora composta por:

Prof. Dr.ª Maria Emília Machado Telles Walter (Orientador) — CIC/UnB
Prof. Dr. Professor I — CIC/UnB
Prof. Dr. Professor II — CIC/UnB

CIP — Catalogação Internacional na Publicação

Esteves, Gabriella de Oliveira.

Visualização em Grafos de Redes Metabólicas via Web / Gabriella de Oliveira Esteves. Brasília : UnB, 2016.
63 p. : il. ; 29,5 cm.

Monografia (Graduação) — Universidade de Brasília, Brasília, 2016.

1. Biologia Molecular, 2. Bioinformática, 3. Redes Metabólicas,
4. Banco de Dados Não Relacional, 5. Grafo, 6. neo4j

CDU 004.4

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília–DF — Brasil

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Visualização em Grafos de Redes Metabólicas via Web

Gabriella de Oliveira Esteves

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr.ª Maria Emília Machado Telles Walter (Orientador)
CIC/UnB

Prof. Dr. Professor I
CIC/UnB

Prof. Dr. Professor II
CIC/UnB

Prof. Dr. Rodrigo Bonifácio de Almeida
Coordenador do Bacharelado em Ciência da Computação

Brasília, 08 de Julho de 2016

Dedicatória

Dedicatória

Agradecimentos

Agradecimento

Resumo

Resumo em português

Palavras-chave: Biologia Molecular, Bioinformática, Redes Metabólicas, Banco de Dados Não Relacional, Grafo, neo4j

Abstract

Abstract in english

Keywords: Molecular Biology, Bioinformatics, Metabolic Networks, Non-Relational Database, Graph, neo4j

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 História da Genética

O estudo do núcleo celular começou no século XIX, em um laboratório na Alemanha, com o objetivo de catalogar as substâncias químicas presentes nas células sanguíneas do ser humano. Como naquela época as pesquisas eram mais voltadas ao citoplasma - fluido pastoso que constitui a célula, o bioquímico suíço Friedrich Miescher foi o pioneiro no estudo do núcleo. Ele quem descobriu a substância nucleína composta por carbono, hidrogênio, oxigênio, nitrogênio e fósforo (ausênte nas proteínas), que mais tarde chamaram de ácido desoxirribonucléico, ou DNA.

No início do século XX, o geneticista estadunidense Thomas Morgan liderou uma equipe de estudantes e realizou vários experimentos em *Drosophila melanogaster* - espécie de mosca, com a finalidade de compreender a hereditariedade a partir de genes transmitidos aos organismos em desenvolvimento. Esta pesquisa foi fundamental para demonstrar exerimentalmente a Teoria Cromossômica da Hereditariedade (Sutton-Boveri, 1902), que assumem várias suposições como verdade, dentre elas: Os genes estão localizados em cromossomos; Os cromossomos formam pares de homólogos; Destes pares, um tem origem paterna, o outro tem origem materna. Tais hipóteses são baseadas nos experimentos caseiros do botânico Gregor Mendel, que após 8 anos de experimentos (1856-1863), publicou seu paper na Nature Research Society of Brünn. Nele, Mendel introduz conceitos como dominância, fator recessivo, hereditariedade, segregação dos fatores e transmissão independente dos genes. O trabalho de Morgan e sua equipe rendeu-lhe um Prêmio Nobel de Fisiologia ou Medicina em 1933.

No início dos anos 50, uma química britânica chamada Rosalind Frankling usou a técnica de difração de raios-X para determinação da estrutura da biomolécula do DNA e concluiu que sua forma era helicoidal. Seu trabalho foi empregado nos experimentos de dois pesquisadores, Francis Crick e James Watson, em um laboratório em Cambridge, na Inglaterra. No mesmo ano, a dupla decifrou a estrutura do DNA: duas longas fitas enroladas uma na outra em espiral para a direita, ligadas por pares de bases complementares, formando o que chamaram de dupla-hélice. Apesar da grande descoberta, isto não era o suficiente para entender como eram produzidas as proteínas, portanto os cientistas mudaram o foco das pesquisas para o RNA, uma vez que sabiam o quanto sua concentra-

ção aumentava sempre que as células começavam a produzir proteínas. Em 1958, Crick e Watson anunciaram mais uma descoberta: A partir do DNA, o processo de *transcrição* fornece uma fita de RNA, que por sua vez, a partir do processo de *tradução*, fornecem a proteína. Esta sequência de processos ficou conhecida como Dogma Central da biologia molecular.

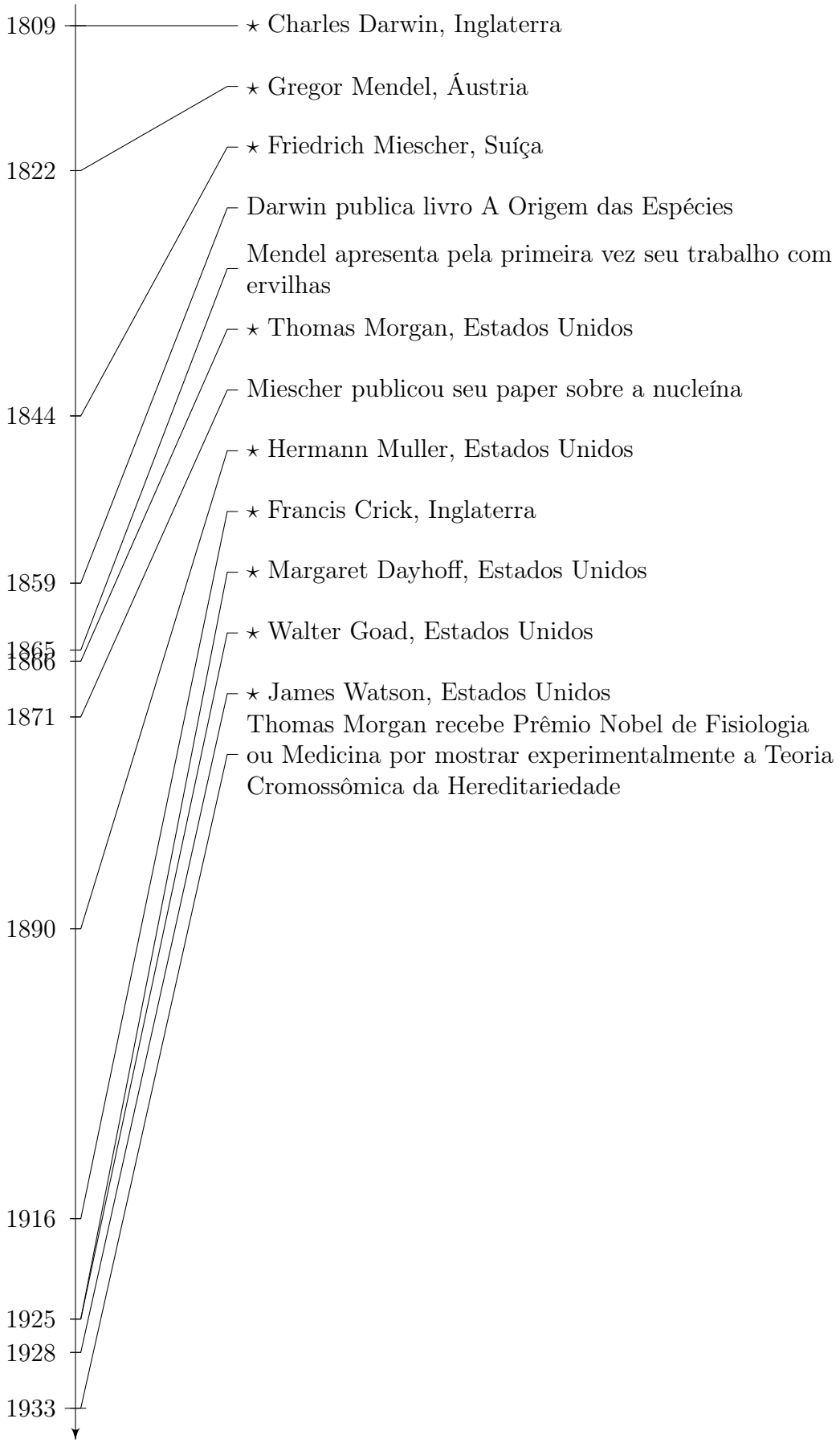
CITAR: O polegar do violinista

Figura 1.1: James Watson e Francis Crick

Bioinformática. Margaret Dayhoff e Walter Goad

.
.
.
.
.
.
.

A linha do tempo abaixo tem o objetivo de auxiliar na localização temporal da história da biologia molecular e da bioinformática ao passo que apresentam as datas de nascimento dos principais pesquisadores da área.


1809
1822
1844
1859
1865
1866
1871
1890
1916
1925
1928
1933
⋆ Charles Darwin, Inglaterra
⋆ Gregor Mendel, Áustria
⋆ Friedrich Miescher, Suíça
Darwin publica livro A Origem das Espécies
Mendel apresenta pela primeira vez seu trabalho com ervilhas
⋆ Thomas Morgan, Estados Unidos
Miescher publicou seu paper sobre a nucleína
⋆ Hermann Muller, Estados Unidos
⋆ Francis Crick, Inglaterra
⋆ Margaret Dayhoff, Estados Unidos
⋆ Walter Goad, Estados Unidos
⋆ James Watson, Estados Unidos
Thomas Morgan recebe Prêmio Nobel de Fisiologia ou Medicina por mostrar experimentalmente a Teoria Cromossômica da Hereditariedade

1.2 Definição do Problema

Contruir uma visualização interativa de redes metabólicas aramazenadas em banco de dados de grafos que permita ao pesquisador explorar os aspectos biológicos do organismo estudado.

1.3 Justificativa

Atualmente, a quantidade de dados «»» estudados pelos pesquisadores é extensa e complexa. Uma maneira de amenizar o esforço feito para analisá-los e compreendê-los é oferecer uma ferramenta que aproxime o usuário (pesquisador) e os dados em forma de grafo(redes metabólicas). Esta ferramenta deverá permitir que o usuário visualize e interaja com os dados dinamicamente, além de disponibilizar mecanismos de busca em grafos, úteis para sua pesquisa.

1.4 Objetivo

Constrir um sistema que acesse redes metabólicas armazenadas em bancos de dados em grafo e gere uma visualização interativa

- Implementar uma busca das vias metabólicas de interesse a apartir de parâmetros informados pelo pesquisador no sistema
- Recuperar a informação desejada e exibí-la para o pesquisador de forma ergonômica
- Implementar algoritmos de busca em grafos para recuperar a informação solicitada e/ou sugerir informação relevante

1.5 Descrição dos Capítulos

No Capítulo 1 fez-se uma breve introdução à história da biologia molecular e da bioinformática. No Capítulo 2 são estabelecidas as principais definições utilizadas neste trabalho mais profundamente, tais como ácidos nucléidos, biomoléculas gerais que originam o DNA e o RNA; a proteína, macromolécula extensa, formada por um processo complexo chamado síntese de proteína; código genético, listagem do arranjo de bases nitrogenadas que formam aminoácidos, que por sua vez compõem a proteína; Neste capítulo também são descritos os processos de sequenciamento de proteínas, na subseção de bioinformática e os desafios enfrentados nessa área.

O Capítulo 3 apresenta uma estrutura chamada Redes metabólicas, estrutura de dados extremamente complexas que existem para auxiliar o pesquisador biólogo a entender reações intracelulares, bem como determinar propriedades fisiológicas e bioquímicas das células. A construção destas redes é possível pos existe sequenciamento do genoma do organismo estudado. O Capítulo 4 propõe um banco de dados não relacional (NoDB) em grafos como maneira de armazenar estas redes metabólicas. Nele é descrito todo o

conceito de NoDB, e é apresentado aquele utilizado neste trabalho: banco de dados neo4j.

No Capítulo 5 são exibidos os resultados da implementação do programa e no Capítulo 6, as conclusões tiradas a parir da análise dos dados. O Capítulo 7 expõe os problemas enfrentados, bem como sugestões de melhorias e trabalhos futuros. Por fim, o Capítulo 8 apresenta uma tabela do cronograma da execução deste trabalho.

Capítulo 2

Biologia Molecular e Bioinformática

Neste capítulo serão descritos os conceitos básicos da biologia molecular. A seção 2.1 define tais estruras e diferencia DNA de RNA por suas configurações e funções. A seção 2.2 define as proteínas, apresenta seus quatro tipos diferentes e descreve o processo de sintetização de proteína. Por fim, a seção 2.3 estabelece os conceitos básicos dessa área, além de apontar os problemas atuais enfrentados nela.

2.1 Ácidos Nucléicos

Os ácidos nucléicos são biomoléculas responsáveis pelo armazenamento, transmissão e tradução das informações genéticas dos seres vivos. Isto é possível devido ao processo de síntede de proteínas que permite, assim, a base da herança biológica.Os acidos nucléicos são polímero, macromoléculas formadas por estruturas menores chamadas monômeros, que nesse caso são nucleotídeos. Nucleotídeos são compostos de três elementos: um radical fosfato (HPO_4), uma pentose, ou seja, um monossacarídeo formado por cinco átomos de carbono, e uma base nitrogenada. Existem cinco tipos de bases nitrogenadas que podem compor um nucleotídeo: Adenina(A), Timina(T), Citosina(C), Guanina(G) e Uracila(U).

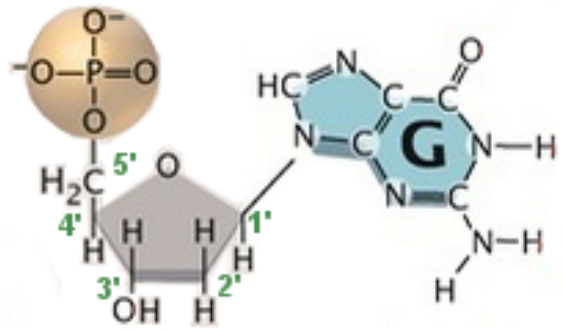


Figura 2.1: imagem de um nucleotídeo e das bases nitrogenadas. Mostrar backbone da pentose 1'...5'. Adaptado de : [10].

Na figura 2.1, observa-se que no *backbone* do nucleotídeo existe uma numeração de 1' à 5', que representam os carbonos presentes na pentose. Para a criação de uma fita de ácido nucléico, no processo de polimerização fomar-se uma ligação fosfodiéster entre o carbono da posição 5' do *backbone* de um nucleotídeos e o carbono de posição 3' do *backbone* de outro [11]. Por definição o sentido da leitura de uma fita de ácido nucléido é 5' $\rightarrow$ 3', o que é deve ser levado em consideração ao se fazer interpretação de dados do material genético.

Dois tipos de ácidos nucléicos são encontrados nos seres vivos: ácido desoxirribonucleico (DNA) e ácido ribonucleico (RNA). Eles diferenciam-se tanto na estrutura do *backbone* e nas bases nitrogenadas, quanto em suas funções. A seguir serão apresentadas as definições de DNA e RNA.

2.1.1 DNA

Os DNAs (ou ADN - Ácido Desoxirribonucleico) são as biomoléculas que armazenam as informações referentes ao funcionamento de todas as células dos seres vivos de maneira específica: sequências de pares de bases nitrogenadas. Nesse sentido, além de haver a ligação fosfodiéster entre os nucleotídeos, cada um também se liga a partir de suas bases nitrogenadas, formando assim um eixo helicoidal tridimensional chamada de dupla hélice [11]. Esta estrutura foi descoberta em 1953, pelo biólogo James Watson e pelo físico Francis Crick [10], porém os ácidos nucléicos já eram estudado desde 1869 na Suíça pelo químico-fisiológico Friedrich Miescher.

Em relação à estrutura dos monômeros do DNA, o *backbone* dos nucleotídeos é uma desoxirribose, indicada na figura 2.2. Para a formação da dupla hélice, os pares são feitos com uma base nitrogenada do grupo de purinas, composto orgânico que possui um anel duplo de carbono, e outra base do grupo de pirimidinas, composto orgânico que possui um anel simples de carbono. No caso do DNA, somente quatro das cinco bases são empregadas: as purinas Adenina(A) e Guanina(G), que se ligam com as pirimidinas Timina(T) e Citosina(C) respectivamente. Desta forma, A e T são bases complementares, assim como G e C. Uma fita de DNA pode conter centenas de milhões de nucleotídeos.

A representação do DNA, seja nos livros ou computacionalmente, é dada por um par em paralelo de strings de letras A, T, G e C. Como explicado no início dessa seção, o sentido padrão da leitura de uma fita é de 5' $\rightarrow$ 3', mas no caso do DNA, as hélices são dispostas de maneira antiparalela, ou seja, uma é lida de 5' $\rightarrow$ 3' e a outra, de 3' $\rightarrow$ 5'. Observa-se que a partir de uma hélice, pode-se inferir a sequência de sua hélice complementar. Seja, por exemplo, uma hélice H1 igual a AGTAAGC; então H2 em seu sentido oposto é H2' igual a TCATTCG, e no sentido regular, igual a GCTTACT. A figura 2.2 apresenta a estrutura do DNA como explicada nesta seção.

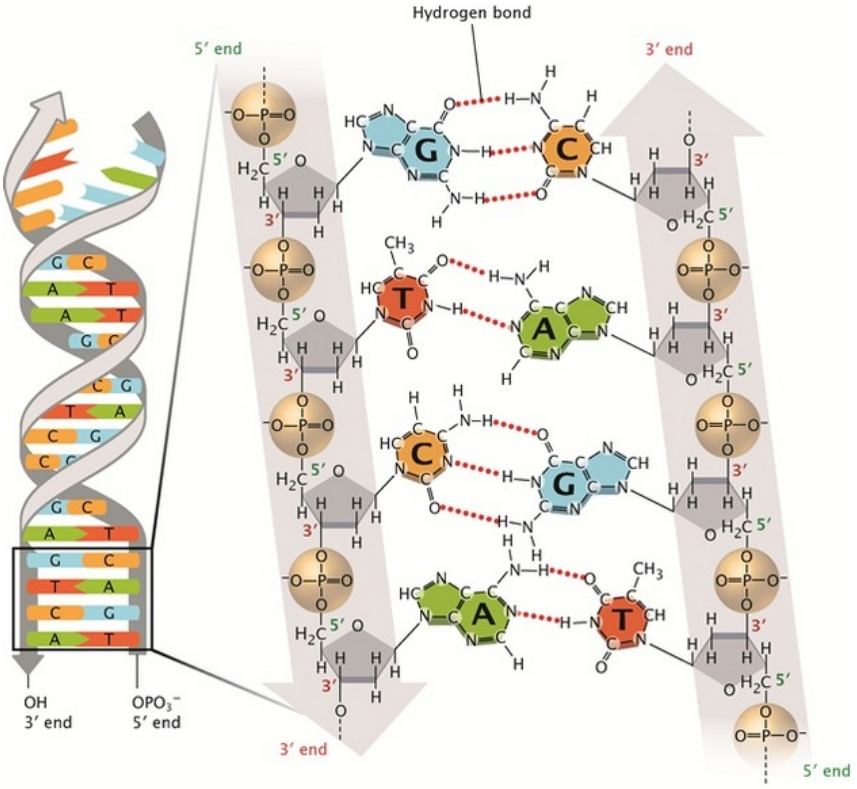


Figura 2.2: Adaptado de : [10]

2.1.2 RNA

Os RNAs são biomoléculas semelhantes ao DNA, porém contam com três diferenças básicas. A primeira é a estrutura do *backbone* dos nucleotídeos, que é composta por uma ribose ao invés de um desoxirribose. A segunda difereça é em relação às bases nitrogenadas, onde a pirimidina Uracila(U) substitui a Timita(T). Por fim, o RNA é formado por apenas uma hélice tridimensional.

Existem três tipos de RNAs presentes no citoplasma - espaço entre a membrama plasmática e o núcleo da célula. Cada um possui funções específicas que serão detalhadas na seção **??**. Em suma, O RNA mensageiro (mRNA) é responsável pela transferência de informação do DNA para o RNA ribossômico (rRNA), que por sua vez irá desanexar a proteína do RNA transportador (tRNA) combinando-o com o rRNA, executando assim, a síntese de proteína.

2.2 Síntese de Proteína

2.2.1 Proteína

As proteínas são biomoléculas com diversas responsabilidades no corpo dos seres vivos. Se fizerem parte do no grupo de proteínas fibrosas, como o colágeno, irão compor a estrutura do corpo e para isso precisam ser resistentes e insolúveis em água. Caso estejam

no grupo de proteínas globulares, como a hemoglobina, realizarão processos dinâmico pelo corpo tais como transportações e cataliações [2]. Cada tarefa é realizada por um proteína com uma estrutura específica e otimizada pra tal.

Assim como os ácidos nucléicos, as proteínas são polímeros, macromoléculas cujos monômeros são aminoácidos. Aminoácidos são moléculas que possuem cinco componentes: amina (NH_2), carbono (C), hidrogênio (H), ácido carboxílico (COOH) e uma cadeia lateral que funciona como identificador de cada um dos 20 tipos de aminoácidos presentes nos seres vivos. A maneira como eles são criados será explicada com mais detalhes na subseção 2.2.3, pois envolve um processo complexo de síntese de proteína executado pelo ribossomo. A ligação, ou polimerização, de dois aminoácidos é feita unindo a amida de um com o ácido carboxílico do outro, liberando uma molécula de água (H_2O) e formando uma cadeia chamada de dipeptídeo. Como houve liberação de água na ligação, o dipeptídeo não é formado por aminoácidos, mas sim resíduos dos mesmos. Nesse sentido, cadeias peptídicas de 100 à 5000 diferentes resíduos aminoácidos, ou cadeia polipeptídicas, constituem a proteína.

Existem quatro estruturas para caracterização de uma proteína [11]. A mais simples é chamada de estrutura primária e é composta por uma sequência linear de resíduos aminoácidos. A estrutura secundária é tridimensional e estabiliza-se por meio de ligações de hidrogênio na cadeia principal, chamada de *backbone*. Dependendo da disposição dos resíduos de aminoácidos, esta cadeia pode se dar forma de hélice ou em forma de folha. A estrutura terciária é dada pela união de várias estruturas secundárias e, por fim, a estrutura quaternária é composta de múltiplas estruturas terciárias [3]. A figura 2.3 ilustra os quatro tipos de proteínas descritos.

2.2.2 Código Genético

No núcleo de cada célula eucariota, ou no citoplasma das células procariotas, estão localizados as moléculas de DNA, chamadas individualmente por **cromossomo**. O número de cromossomos em cada célula varia por espécie. No caso dos chimpanzés, o núcleo das células possui 48 cromossomos e no caso dos seres humanos, 46. Note que não existe relação entre o grau evolutivo das espécies e o número de cromossomos nas células.

Um cromossomo pode ser representado por vários trechos contíguos de DNA, sendo que cada trecho é chamado de **gene**, ou locus - local fixo no cromossomo. Portanto, pode-se afirmar que o cromossomo é um conjunto (ou lista) de genes. No caso dos seres humanos, o número de genes em cada célula gira em torno de 22 mil [9], e o genoma humano possui em média 3 bilhões de pares de bases. Poderíamo inferir, então, que a média de pares de bases por gene é de $\frac{3.000.000.000}{22.000} \simeq 136.000$, porém este cálculo é muito generalizado e equivocado, uma vez que os genes possuem tamanhos diferentes, onde o maior possui 250 milhões de pares, enquanto o menor possui apenas 50 milhões, no caso dos seres humanos [8]. Um gene, por sua vez, pode ser representado por vários trechos de três pares de base, sendo que cada trecho é chamado de **códon**.

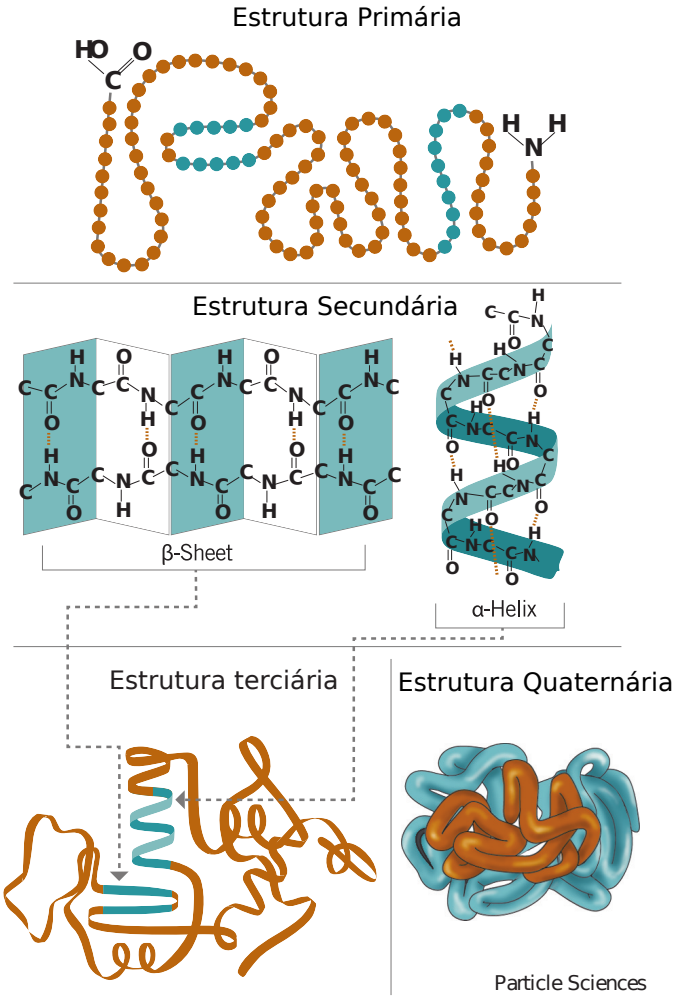


Figura 2.3: Adaptado de : [3]

Normalmente cada proteína é formada a partir de um gene particular. Mais especificamente, cada aminoácido da proteína é formado a partir de um códon do gene. Entretanto, existem 64 códon possíveis ($4^3_{ParesDeBase}$) mas somente 20 aminoácidos a serem codificados. Nesse sentido, é comum haver mais de um códon correspondendo à um aminoácio. Além disso, 3 destes códons são responsáveis por indicar o final de uma proteína. O mRNA é encarregado de transportar a informação da sequência correta para construção de proteína, em forma de sequência de códons. A tabela 2.1 que apresenta a correspondência entre códons e aminoácios é chamada de **código genético** [11], e a tabela 2.2 apresenta o código genético codificado em letras do alfabeto utilizado atualmente para comparação entre proteínas. Note que as bases nitrogenadas são do RNA, e não do DNA, pois é a molécula do primeiro que faz a conexão entre DNA e a proteína, num processo que será explicado na próxima subseção. **BUSCAR FONTE DISSO**

A partir destas tabelas, podemos montar o seguinte exemplo: Suponha que a palvra GENETICA seja uma proteína. Então existe uma configuração de aminoácidos que forma essa proteína, e ela pode ter a forma:

GENETICA ← Glicina - Glutamano - Metionina - Glutamano
Treonina - Isoleucina - Cisteina - Alanina
GENETICA ← Gly - Glu - Asn - Glu - Thr - Ile - Cys - Ala
GENETICA ← GGG - GAG - AAC - GAA - ACG - AUC - UGC - UCC

Tabela 2.1: Código Genético

Primeira Posição	Segunda Posição				Terceira Posição
	G	A	C	U	
G	Gly	Glu	Ala	Val	G
	Gly	Glu	Ala	Val	A
	Gly	Asp	Ala	Val	C
	Gly	Asp	Ala	Val	U
A	Arg	Lys	Thr	Met	G
	Arg	Lys	Thr	Ile	A
	Ser	Asn	Thr	Ile	C
	Ser	Asn	Thr	Ile	U
C	Arg	Gln	Pro	Leu	G
	Arg	Gln	Pro	Leu	A
	Arg	His	Pro	Leu	C
	Arg	His	Pro	Leu	U
U	Trp	**FIM**	Ser	Leu	G
	FIM	**FIM**	Ser	Leu	A
	Cys	Tyr	Ala	Phe	C
	Cys	Tyr	Ala	Phe	U

2.2.3 Transcrição e tradução

Finalmente, agora será descrita a forma como as informações contidas no DNA resulta em proteínas. Um mecanismo de célula reconhece o início de um gene ou aglomerado de genes graças ao *promotor*. O promotor é uma região antes de cada gene no DNA que serve como uma indicação para o mecanismo celular que é um gene está a frente. O códon AUG (que é codificado para metionina) também indica o início de um gene. Tendo reconhecido o início de um gene ou *cluster* de genes, uma cópia do gene é feita sobre uma molécula de RNA. Este RNA resultante é o RNA mensageiro, ou mRNA, e terá exatamente a mesma sequência que uma das cadeias do gene, mas substituindo U por T. Este processo é chamado de **transcrição**. O mRNA, então, será utilizado em estruturas celulares chamados ribossomas para a fabricação de uma proteína.

Uma vez que o RNA é de cadeia simples e DNA é de cadeia dupla, o mRNA produzido é idêntico em sequência à apenas uma das cadeias do gene, sendo complementar à outra

Tabela 2.2: Aminoácidos codificados

	Aminoácido	Abreviação	Código no alfabeto
1	Alanina	Ala	A
2	Cisteina	Cys	C
3	Aspartano ou Ácido aspártico	Asp	D
4	Glutamato ou Ácido glutâmico	Glu	E
5	Fenilalanina	Phe	F
6	Glicina ou Glicocola	Gly	G
7	Histidina	His	H
8	Isoleucina	Ile	I
9	Lisina	Lys	K
10	Leucina	Leu	L
11	Metionina	Met	M
12	Asparagina	Asn	N
13	Prolina	Pro	P
14	Glutamina	Gln	Q
15	Arginina	Arg	R
16	Serina	Ser	S
17	Treonina	Thr	T
18	Valina	Val	V
19	Tripofano	Trp	W
20	Tirosina	Tyr	Y

cadeia - lembrando que T é substituído por U no RNA. A vertente que se parece com o produto mRNA é chamado de anti-sentido ou cadeia codificadora, e o outro é o sentido ou anticodificação ou cadeia de molde. A cadeia de molde é aquele que é transcrita, pois o mRNA é composto por ligações de ribonucleotidos complementares a esta vertente. O processo sempre contrói moléculas de mRNA da extremidade 5' até a 3', ao passo que a cadeia de molde é lida de 3' para 5 '. Note também que a cadeia de molde não é sempre a mesma; Por exemplo, a cadeia de molde para um determinado gene A pode ser uma das fitas, e a cadeia de molde para outro gene B pode ser a outra fita. Para um dado gene, a célula é capaz de reconhecer os correspondentes moldes de cadeia graças ao promotor. Mesmo que o complemento reverso do promotor aparece na outra cadeia, ele não é um promotor e por isso não irá ser reconhecido como tal. Uma consequência importante deste fato é que os genes do mesmo cromossomo tem uma orientação com relação ao outro: Dado dois genes, se eles aparecem na mesma vertente eles têm a mesma orientação; caso contrário, eles têm orientação oposta. Finalmente, note que os termos *upstream* e *downstream* são usados para indicar as posições do DNA em referência à orientação da cadeia codificadora, com o promotor sendo o *upstream* do gene.

A transcrição descrita é válida para os organismos classificados como procariontes. Estes organismos têm o seu DNA livre na célula, pois não há uma membrana nuclear. Exemplos de procariontes são bactérias e algas azuis. Todos os outros organismos, categorizados como eucariotas têm um núcleo separado do resto da célula por uma membrana nuclear, e o seu DNA é mantido no interior do núcleo. Nestes organismos a transcrição

genética é mais complexa. Muitos genes eucarióticos são compostos por partes alternadas dos chamados *introns* e *exons*. Após a transcrição, os *introns* são unidos fora do mRNA. Isto significa que os *introns* são partes de um gene que não são utilizados na síntese de proteínas. Depois que os *introns* são unidos, o mRNA encurtado (contendo cópias apenas dos *exons* e de regiões reguladoras nas extremidades), deixa o núcleo, uma vez que os ribossomos estão fora.

Devido ao fenômeno de *introns/exons*, serão usados nomes diferentes para se referir ao gene como um todo encontrado no cromossomo e a sequência de *splicing*, *exons* unidos. O primeiro é chamado de DNA genômico e o segundo de DNA complementar ou DNAc. Os cientistas podem fabricar DNAc sem saber o seu homólogo genômico. Eles primeiro capturam o mRNA fora do núcleo no seu caminho para os ribossomos. Em seguida, num processo denominado **transcrição reversa**, que produzem moléculas de DNA utilizando o mRNA como um molde. Uma vez que o mRNA contém apenas os *exons*, esta é também a composição do DNA produzido. Assim, eles podem obter cDNA sem sequer olhar para os cromossomos. Ambos transcrição e transcrição reversa são processos complexos que necessitam a ajuda de enzimas. Transcriptase e transcriptase reversa são os enzimas que catalisam a estes processos na célula. Há também um fenômeno chamado de *splicing* alternativo. Isto ocorre quando o mesmo DNA genômico pode dar origem a duas ou mais moléculas de mRNA diferentes, escolhendo os *introns* e *exons* de maneiras diferentes. Eles em geral produzem proteínas diferentes.

Voltando ao mRNA e síntese protéica, dois outros tipos de moléculas RNA desempenham papéis muito importantes. Como já mencionado, a síntese de proteína é realizada dentro de estruturas celulares chamadas ribossomos. Os ribossomas são feitos de proteínas e uma estrutura de RNA denominada RNA ribossômico, ou rRNA. Os ribossomos funcionam como uma linha de montagem em uma fábrica usando como *inputs* uma molécula de mRNA e outro tipo de molécula de RNA chamado RNA de transferência, ou tRNA.

tRNAs são as moléculas que realmente implementam o código genético em um processo chamado de **tradução**. Elas fazem a conexão entre um códon e o aminoácido específico este códons codifica. A medida que o mRNA passa através do interior do ribossomo, um tRNA correspondente ao códon corrente - o códon no mRNA atualmente no interior do ribossomo - se liga a ele, trazendo o aminoácido correspondente (vários aminoácidos livres estão sempre em torno da célula). A posição tridimensional de todas essas moléculas neste momento é tal que, assim que o tRNA se liga ao seu códon, o aminoácido ligado vai para o lado do aminoácido precedente na cadeia de proteína que está sendo formada. Uma enzima adequada catalisa então a adição deste aminoácido corrente para a cadeia de proteína, libertando-o do tRNA. Uma proteína é construída, resíduo por resíduo, desta forma. Quando um códon de parada aparece, nenhum tRNA é associado à ele e a síntese termina. O RNA mensageiro é liberado e degradados por mecanismos celulares em ribonucleotidos, que serão em seguida reciclados para fazer outros RNAs.

[11]

2.3 Bioinformática

...

2.3.1 Sequenciamento

Mardis 2008, Annotations

2.3.2 Desafio das ômicas

⇒ Processamento, armazenamento e recuperação ⟶ Construção do genoma ⟶ Alinhamento de sequências genéticas ⟶ Compressão, armazenamento e pesquisas em genomas de grande porte
⇒ *Data mining* para transcriptomas ⟶ Identificar expressão genéticas de células específicas ⟶ Identificar genes e módulos regulatórios ⟶ Identificar alterações nas expressões genéticas em doenças
⇒ Interactomas integrativos ⟶ Análise de conjuntos de dados genômicos heterogêneos ⟶ Análise interactoma de conjunto de dados de doenças
BIBLIOGRAFIA: Computational solutions for omics data

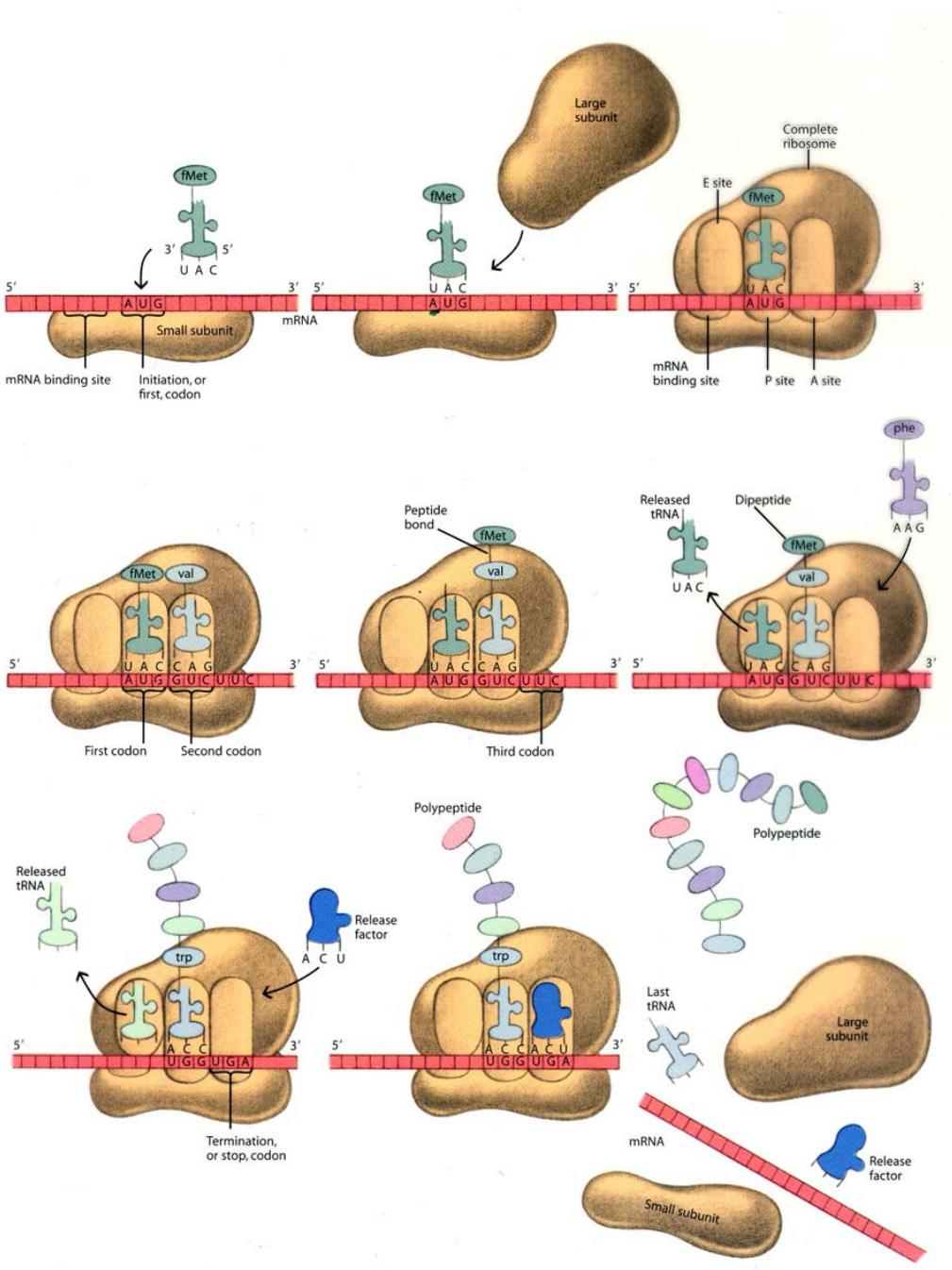


Figura 2.4: Adaptado de : [1]

Capítulo 3

Redes Metabólicas

A rede metabólica pode ser definida formalmente como uma coleção de objetos e as relações entre eles. Os objetos correspondem a compostos químicos, reações bioquímicas, enzimas e genes.

compostos químicos, também chamados metabólitos, são pequenas moléculas que são importados / exportados e / ou sintetizadas / degradadas dentro de um organismo. Para a maioria dos metabolitos, a quantidade observada varia de acordo com o compartimento da célula e tecido no interior do qual o composto está presente. Os tecidos e células de facto conter um número de compartimentos líquidos separados uns dos outros por membranas de permeabilidade selectiva.

Reações bioquímicas produzem um conjunto de um ou mais compostos (chamado de produtos) a partir de um outro conjunto de um ou mais compostos (chamados os substratos). Em teoria, uma reacção química pode ocorrer em ambos os sentidos. No entanto, sob determinadas condições fisiológicas, algumas reações ocorrem em apenas uma direção. Neste caso, eles são definidos como sendo irreversível, se todas as outras condições permanecem constantes. Dentro de uma célula, algumas reacções são espontâneas, mas a maioria são catalisada por uma ou várias enzimas que aceleram fortemente a sua velocidade. Uma enzima é uma proteína ou um complexo de proteína, codificada por um ou vários genes. Uma única enzima pode aceitar substratos distintas e pode catalisar a várias reacções, e, inversamente, uma única reacção pode ser catalisada por diversas enzimas. Elucidar as ligações entre genes, proteínas e reações (o chamado relacionamento GPR) não é uma tarefa trivial e é uma grande preocupação na reconstrução metabólica, como é discutido na próxima seção.

A descoberta de enzimas por Eduard Buchner no início do século 20 separados o estudo das reacções químicas que compõem o metabolismo de um organismo a partir do estudo da biologia das suas células. Tais reacções são tradicionalmente agrupados em chamados vias metabólicas, o que pode por sua vez ser classificados como anabólico ou catabólico. Anabolismo, é a síntese de moléculas através do uso de energia e no consumo de agentes redutores (um agente de redução é uma substância que quimicamente reduz outras substâncias doando um ou vários electrões), enquanto o catabolismo corresponde à degradação de moléculas de rendimento de energia e a produção de redução agentes. As vias podem ser estudados, quer isoladamente, ou, uma vez que são sobrepostas, ser

combinados em conjunto para produzir o que é referido como uma rede metabólica. Os benefícios de estudar toda a rede, em vez de percursos individuais são numerosas e incluem, por exemplo, a possibilidade de explorar alternativa vias.

[7]

Capítulo 4

Banco de Dados NoSQL

NOSQL, NOT ACID, Neo4j, Cypher

Capítulo 5

2Path: Aplicação Web

O sistema desenvolvido para este projeto é uma aplicação web chamada *2Path*. O usuário deve se cadastrar no *website* para ter acesso às redes metabólicas do banco de dados do sistema, bem como pesquisar por palavras chaves no mesmo. Neste caítulo serão apresentadas as linguagens e ferramentas utilizadas no desenvolvimento do *website*, as características, funcionalidades e limItes do sistema e, por fim, as dificuldades enfrentadas na implementação do projeto.

5.1 Implementação

O sistema foi desenvolvido no amibiente de desenvolvimento integrado *open source* Eclipse Java EE - *Java Platform, Enterprise Edition*, versão Mars 4.5.2. A plataforma Eclipse foi projeta com o objetivo de agilizar o desenvolvilmento de recursos integrados baseando-se em um modelo de *plug-in*. Na *workbench* no Eclipse, cada *plug-in* é responsável por pequenas tarefas, tais como compilar, testar ou debugar [4].

Para simplificar a obtenção das dependências do projeto, ou seja, pacotes de arquivos java (extensão .jar), foi utilizada o Apache Maven, *software* de gerenciamento de projeto e ferramenta de compreensão de programa. Este *software* opera sobre o arquivo *pom.xml*, onde POM significa *Project Object Model* e contém as especificações de cada projeto que se tornará dependência do sistema em desenvolvimento, além de outros aspectos do código. No exemplo abaixo, o fragmento do *pom.xml* indica o *groupID* - código único entre a organização ou projeto, *artifactId* - nome do projeto, *version* - versão do projeto que será baixada e *scope* - escopo em que o projeto será necessário no sistema (compilação, execução ou teste).

```
<dependencies>
(...)
        <!-- PrimeFaces (biblioteca de componentes) -->
        <dependency>
                <groupId>org.primefaces</groupId>
                <artifactId>primefaces</artifactId>
                <version>3.5</version>
                <scope>compile</scope>
```

```
        </dependency>
(...)
</dependencies>
```

O servidor selecionado para «<» o sistema na rede, *localhost* porta 8080, foi o Apache TomCat versão 7.0. Este software é uma implementação *open source* das quatro tecnologias [5] a seguir:

- *Java Servlet*:
- *JavaServer Pages*:
- *Java Expression Language*:
- *Java WebSocket*:

COLOCAR IMAGEM DA ARQUITETURA MVC : An MVC EBookShop with Servlets, JSPs, and JavaBeans Deployed in Tomcat [6]

As quatro páginas da aplicação foram desenvolvidas na linguagem de marcação XHTML, *Extensible Hypertext Markup Language*, e a estilzação em CSS, *Cascading Style Sheets*. Com a primeira é possível criar objetos na página *web* através de componentes nativos e não nativos da linguagem chamados *tags*. As principais *tags* são apresentadas na Tabela **??**. Já com CSS é possível customizar cada objeto da página web, alterando seu tamanho, posição, cor, fonte, e várias outras características. Para tal, o objeto por ser alterado individualmente através de seu ID; em conjunto, com objetos da mesma classe ou *tag*.

JSF PRIMEFACES

5.2 Desafios

O que foi o trabalho. Decrever todo o ambiente usado Neste capítulo serão apresentados os primeiros resultados experimentais obtidos.

Capítulo 6

Conclusão

Neste capítulo serão apresentadas as considerações finais do trabalho, assim como as limitações e dificuldades encontradas.

Capítulo 7

Trabalhos Futuros

A partir deste trabalho, foi possível identificar os seguintes pontos a serem melhorados:

- x

Capítulo 8

Cronograma

O cronograma está apresentado na Tabela a seguir, mostrando o inicio das atividades em Janeiro de 2016 com a revisão literária e com término previsto para Junho de 2016, juntamente com a defesa do Trabalho de Conclusão de Curso.

Tabela 8.1: Cronograma

Atividades	2016					
	Jul	Ago	Set	Out	Nov	Dez
Revisão bibliográfica	X	X				
Familiaridade com ambiente de desenvolvimento		X	X			
Implementação da aplicação		X	X	X		
Interpretação dos resultado				X	X	X
Defesa						X

Referências

[1] Protein syntesis steps. vi, 15

[2] Proteínas. http://www.professoraangela.net/documents/proteinas.html, visitado em 2016-01-02. 9

[3] Protein structure, 2009. http://www.particlesciences.com/news/technical-briefs/2009/protein-structure.html, visitado em 2016-01-02. vi, 9, 10

[4] Eclipse Foundation, Inc., 102 Centrepointe Drive, Ottawa, Ontario,. *Eclipse documentation - Current Release*, 4.6 edition, 2016. http://help.eclipse.org/neon/index.jsp?topic=%2Forg.eclipse.platform.doc.isv%2Fguide%2Fint_eclipse.htm, visitado em 2016-07-01. 19

[5] Ian Evans Kim Haase William Markito Eric Jendrock, Ricardo Cervera-Navarro. *Java Platform, Enterprise Edition The Java EE Tutorial, Release 7*. Oracle, 2014. https://docs.oracle.com/javaee/7/tutorial, visitado em 2016-08-19. 20

[6] Chua Hock-Chuan. Java web database applications, 2011. https://www.ntu.edu.sg/home/ehchua/programming/java/JavaWebDBApp.html, visitado em 2016-08-19. 20

[7] Vincent Lacroix, Ludovic Cottret, Patricia Thébault, and Marie-France Sagot. An introduction to metabolic networks and their structural analysis. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 5(4):594–617, 2008. 17

[8] R. Nussbaum. *Genética Médica*. Elsevier Editora Ltda., 2008. 9

[9] Mihaela Pertea and Steven L. Salzberg. Between a chicken and a grape: estimating the number of human genes. *Genome Biology*, 11(5):1–7, 2010. http://dx.doi.org/10.1186/gb-2010-11-5-206, visitado em 2016-04-04. 9

[10] Leslie A. Pray. Discovery of dna structure and function: Watson and crick, 2008. http://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397, visitado em 2016-01-15. vi, 6, 7, 8

[11] João Carlos Setubal and João Meidanis. *Introduction to computational molecular biology*. PWS Publishing Company, 1997. 7, 9, 10, 13