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.
59 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 concentração 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.

Figura 1.1: James Watson e Francis Crick

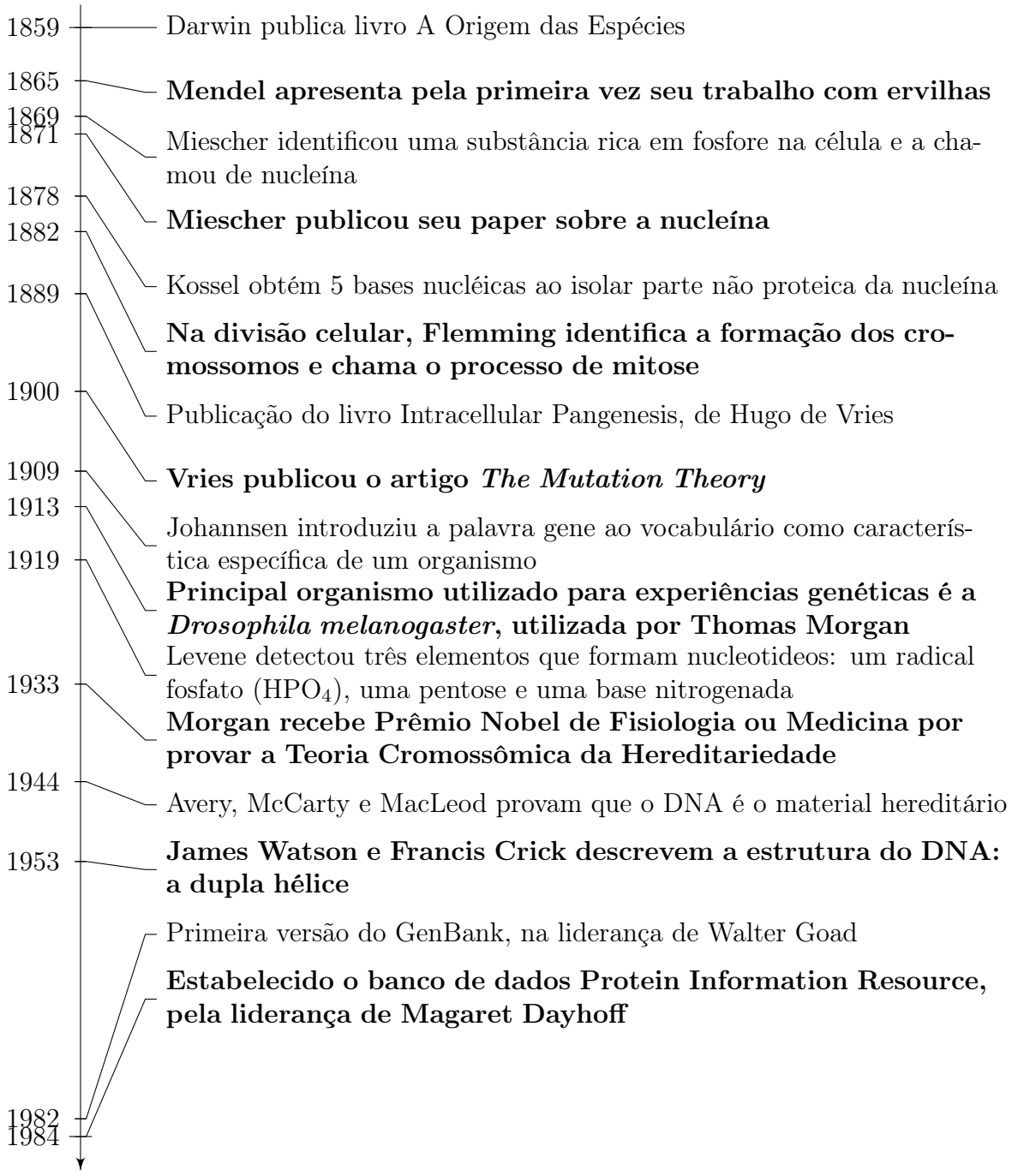
desenvolviento de métodos de SEQUENCIAMENTO de proteínas

Conforme o número de sequências de proteínas cresciam, aumentava também a necessidade de criar-se um banco de dados para indexá-las. A físico-química norte-americana Margaret Dayhoff, com colaboração de alguns membros do *National Biomedical Research Foundation* em Washington, foi a primeira a contruir um banco de dados com este propósito em um tipo de atlas de proteínas na década de 60. Somente em 1984 esta coleção foi intitulada de *Protein Information Resource* [9]. Os dados eram organizados de acordo com o grau de similaridade das sequências, onde o agrupamento das mesmas era dado em forma de árvore filogenética representando famílias e superfamílias de proteínas. Caso a semelhança seja alta, é provável que tenham as mesmas funções bioquímicas e estrutura tridimensional. A partir da árvore gerada, foi possívei calcular as mutações que ocorreram nos aminoácidos durante a evolução genética e, então, produzir uma tabela utilizada até hoje, chamada PAM (*percent acept mutation*), que apresenta tais dados. Outro banco de dados de grande porte e bastante utilizado nos dias de hoje é o GenBank, estabelecido em 1982 por Walter Goad e demais colaboradores e, agora, com o patrocínio do *National Center for Biotechnology Information*. Os dois bancos são públicos e continuam crescendo exponencialmente [9].

problemas ôminicos

web search database david lipman

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 os maiores marcos dos principais pesquisadores da área.



1.2 Problema e Hipótese

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 OrientDB.

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).

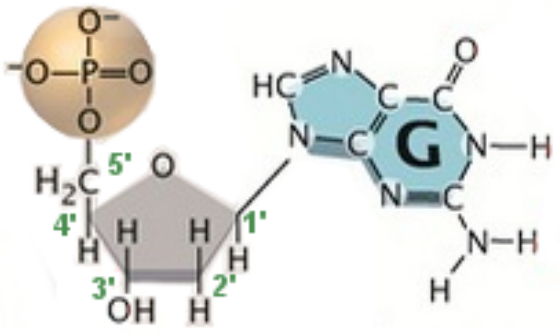


Figura 2.1: imagem de um nucleotídeo e das bases nitrogenadas. Mostrar backbone da pentose 1'...5'. Adaptado de : [13].

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 [14]. Por definição o sentido da leitura de uma fita de ácido nucléido é 5' → 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 [14]. Esta estrutura foi descoberta em 1953, pelo biólogo James Watson e pelo físico Francis Crick [13], 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' → 3', mas no caso do DNA, as hélices são dispostas de maneira antiparalela, ou seja, uma é lida de 5' → 3' e a outra, de 3' → 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.

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.

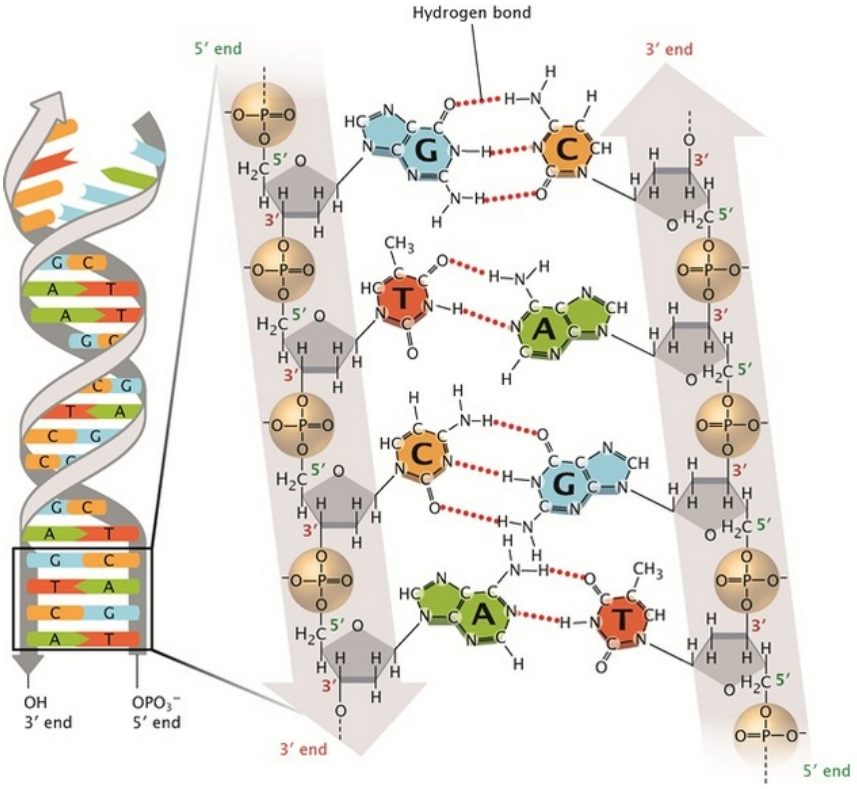


Figura 2.2: Adaptado de : [13]

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 2.2.3. 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 [1]. 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 [14]. 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 [2]. A figura 2.3 ilustra os quatro tipos de proteínas descritos.

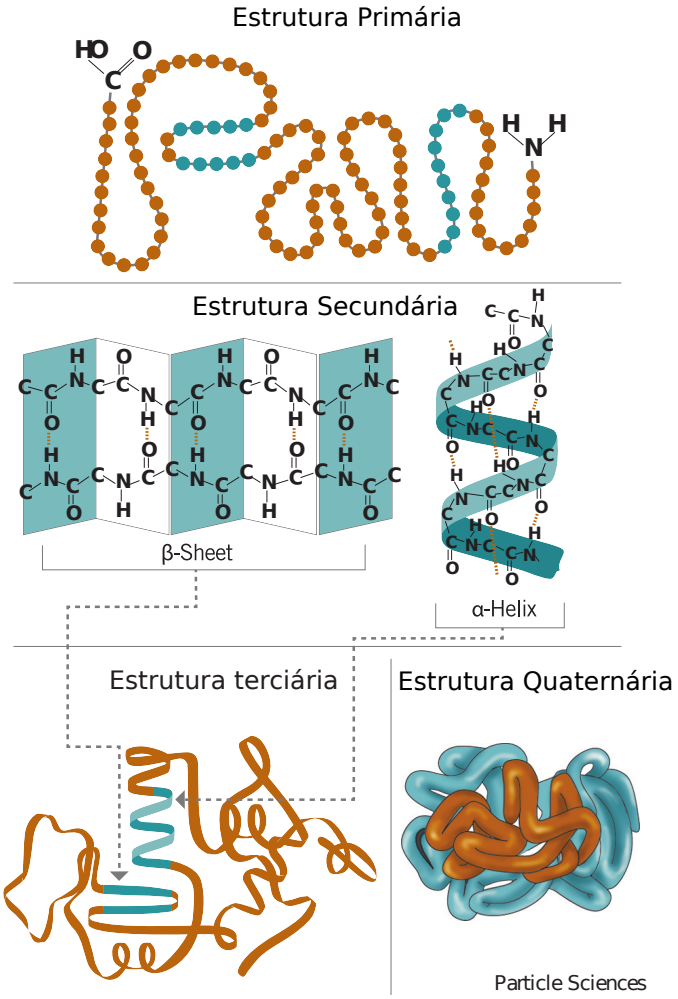


Figura 2.3: Adaptado de : [2]

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 [12], 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 [10]. Um gene, por sua vez, pode ser representado por vários trechos de três pares de base, sendo que cada trecho é chamado de **códon**.

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 [14], 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 subseção 2.2.3.

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

2.2.3 Transcrição e tradução

Para finalizar esta seção, serão descritos os processos de transcrição, tradução e síntese de proteína. O início de cada gene é reconhecido ou em uma região chamada de *promotor* ou, alternativamente, pelo códon AUG. A partir de então, uma cópia deste gene, ou *clusters* de genes, é feita sobre uma molécula de mRNA que, por consequência, obterá a mesma sequência que uma das cadeias do gene, porém trocando os U's por Ts, através deste processo chamado de **transcrição**. A cadeia que se assemelha ao produto mRNA é chamado de cadeia codificadora (lida de 5' para 3'), enquanto a cadeia oposta é chamada de

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

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 de molde (lida de 3' para 5'). Para denotar a orientação de cada cadeia codificadora,

os termos *upstream* e *downstream* são utilizados. Observe que o promotor é o *upstream* do gene, pois reconhece seu início.

No caso dos organismos eucariotas, em que o núcleo da célula está elvolto na membrana nuclear, os fragmentos do gene que não serão utilizados para síntese de proteína, íntrons, são descartados do mRNA após a transcrição e aqueles que serão usados, éxons, são mantidos. Observe que a partir do momento em que o gene foi dividido entre íntrons e éxons, ele muda sua denominação e passa a se chamar DNA genômico. Nesse sentido, a sequência de somente éxons também recebe outro nome: DNA complementar, ou cDNA.

Após esta separação, o mRNA deixa o núcleo celular e inicia a **transcrição reversa** no citoplasma. O processo ocorre no interior de uma organela celular chamada de ribossomo, constituído de proteínas e rRNA e cuja função é construir a molécula de proteína a partir de duas entradas, o mRNA e tRNA. A estrutura do tRNA é tal que de um lado se encaixa exatamente um códon e no oposto, seu aminoácido correspondente. O processo de **tradução** se dá da seguinte maneira: a medida em que o mRNA passa pelo interior do ribossomo, este atrai quaisquer tRNAs das proximidades cujos códons sejam correspondentes ao da subsequência corrente do mRNA. No momento em que o códon do tRNA se conecta com um dos códons do mRNA, o aminoácido que estava fixado naquele é liberado e agregado, com o auxílio da catálise de uma enzima, na molécula de proteína em desenvolvimento. Esta é finalmente completa quando o mRNA aprensenta um códon de parada, pois nenhum tRNA possui correspondência para tal [14].

2.3 Bioinformática

O que é, quando surgiu
Quero amplificar um fragmento pra estudar : PCR
Quero comparar dois transcriptomas : Microarry
Quero sequenciar : Sanger (antigo) e Illumina (novo)
[8], Annotations (interseções)

2.3.1 Desafio das ômicas

Nas últimas duas décadas, o progresso tecnológico no sequenciamento genômico ultrapassou o crescimento previsto pela Lei de Moore [3], a qual afirma que as capacidades de processamento e armazenamento de um computador dobram a cada 18 meses. Nesse sentido, algumas áreas de sufixo "ômica", como, por exemplo, genômica, transcriptômica, interatômica, enfrentam desafios computacionais atualmente.

Em relação ao procesamento, armazenamento e recuperação de dados, existem três problemas principais sendo enfrentados. O primeiro se refere a construção do genoma, que requer a associação se milhões de pequenas sequências e memória na ordem de terabytes; O segundo desafio é o alinhamento de sequências, que também é um processo oneroso se utilizado a técnica de combinação de cadeias, letra por letra; existem técnicas de compressão de dados que encurtam as sequências e outros de programação paralela que aceleram este processo. O terceiro desafio é a (des)compressão de dados para armazenamento e processamento, visto que a análise computacional do material requer as sequências genômicas originais.

Em relação aos desafios da área transcriptômica, referente aos RNA's, os problemas são a identificação de expressões, ou características fenótipas, de células específicas, especialmente tecidos, pois são difíceis de de interpretar; identificação de genes e módulos regulatórios, um grupo de gene com funções biológicas específicas; identificação das alterações nas expressões genéticas em doenças [3].

Já em relação à área de interatoma, referente à interações entre proteínas que ocorrem nas células do organismo [4], os desafios são analisar os conjuntos de dados genômicos heterogêneos e elaborar análise interatoma de conjunto de dados de doenças. Geralmente as redes interatômicas são representadas por grafos, onde os nós são componentes biológicos e as arestas são as interações entre eles.

2.3.2 Visualização de dados ômicos

Capítulo 3

Redes Metabólicas

3.1 Descrição das entidades

3.2 Modelagem de Metabolismo usando grafos

3.2.1 Grafo

[7]

Capítulo 4

Banco de Dados NoSQL

NoSQL
Comparação com SQL

4.1 Propriedades

ACID, BASE
Consistency, availability and tolerance of network partition (consistência, disponibilidade e tolerancia de partição de redes)

4.2 OrientDB

sobre ACID

Modelo CAP

JAVA API

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 [5].

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 [11] 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 Visualização das redes metabólicas

JAVASCRIPT
ANGULARJS
ORIENTBD

5.3 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] Proteínas. http://www.professoraangela.net/documents/proteinas.html, visitado em 2016-01-02. 7

[2] Protein structure, 2009. http://www.particlesciences.com/news/technical-briefs/2009/protein-structure.html, visitado em 2016-01-02. vii, 8

[3] Bonnie Berger, Jian Peng, and Mona Singh. Computational solutions for omics data. *Nature reviews. Genetics*, 14(5):333–346, May 2013. 11, 12

[4] Laura Bonetta. Protein-protein interactions: Interactome under construction. *Nature*, 468(7325):851–854, Dec 2010. 12

[5] 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. 15

[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. 16

[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. 13

[8] Elaine R Mardis. Next-generation dna sequencing methods. *Annu Rev Genomics Hum Genet*, 9:387–402, 2008. 11

[9] David W. Mount. *Bioinformatics : sequence and genome analysis*. Cold Spring Harbor, N.Y. Cold Spring Harbor Laboratory Press, 2001. 2

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

[11] Oracle. *Java Platform, Enterprise Edition The Java EE Tutorial, Release 7*, 2014. https://docs.oracle.com/javaee/7/tutorial, visitado em 2016-08-19. 16

[12] 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. 9

[13] 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. vii, 5, 6, 7

[14] João Carlos Setubal and João Meidanis. *Introduction to computational molecular biology*. PWS Publishing Company, 1997. 6, 8, 9, 11