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.
49 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

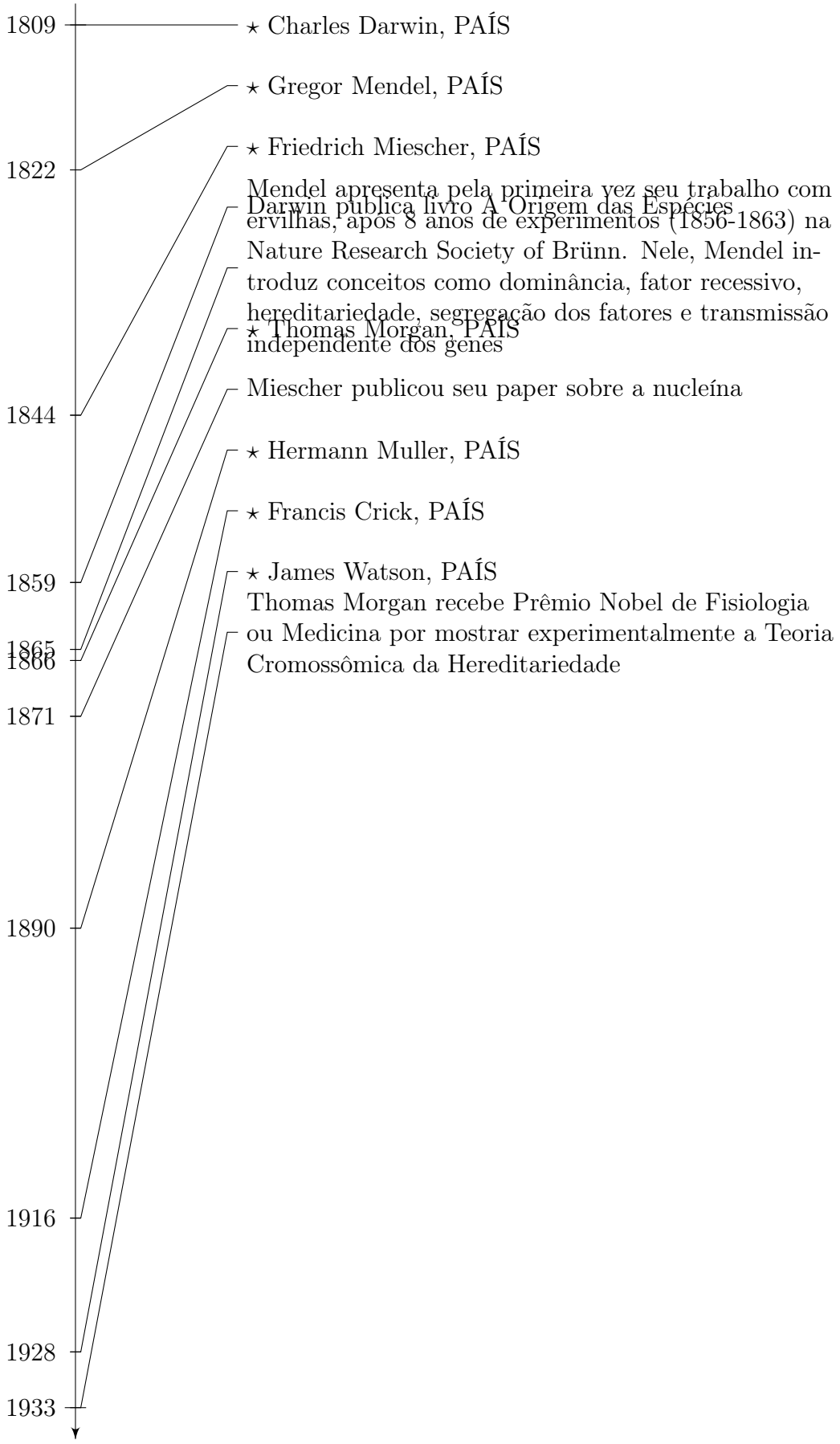
1.1 História da Genética

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. O trabalho de Morgan e sua equipe rendeu-lhe um Prêmio Nobel de Fisiologia ou Medicina em 1933.

Crick e Watson.

Bioinformática
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
⋆ Charles Darwin, PAÍS
⋆ Gregor Mendel, PAÍS
⋆ Friedrich Miescher, PAÍS
1822
Mendel apresenta pela primeira vez seu trabalho com ervilhas, após 8 anos de experimentos (1856-1863) na Nature Research Society of Brünn. Nele, Mendel introduz conceitos como dominância, fator recessivo, hereditariedade, segregação dos fatores e transmissão independente dos genes
Darwin publica livro A Origem das Espécies
⋆ Thomas Morgan, PAÍS
Miescher publicou seu paper sobre a nucleína
1844
⋆ Hermann Muller, PAÍS
⋆ Francis Crick, PAÍS
⋆ James Watson, PAÍS
1859
Thomas Morgan recebe Prêmio Nobel de Fisiologia ou Medicina por mostrar experimentalmente a Teoria Cromossômica da Hereditariedade
1865
1866
1871
1890
1916
1928
1933

1.1.1 Origens da Vida

...

1.1.2 Análise do Núcleo Celular

...

Figura 1.1: Pai da bio

1.1.3 Estudo do genoma

...

1.2 Sequenciamento genético

...

1.3 Definição do Problema

Contruir uma visualização interativa de redes metabólicas aramazenadas em *Graph Databases* que permita ao pesquisador explorar os aspectos biológicos do organismo estudado.

1.4 Justificativa

Atualmente, a quantidade de dados ««»» estudados pelos pesquisadores é extensa e complexa. Uma maneira de amenizar o esforço feito para analisar os dados 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.

1.5 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

1.6 Descrição dos Capítulos

No Capítulo 1 fez-se uma breve introdução aos ... No Capítulo 2 são estabelecidas as principais definições utilizadas neste trabalho mais profundamente, tais como ... Ainda, são apresentados ... Também são descritos ... O Capítulo 3 faz referência à implementação do...

Capítulo 2

Biologia Molecular e Bioinformática

Neste capítulo serão descritos os conceitos básicos da biologia molecular. A seção **??** ...

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).

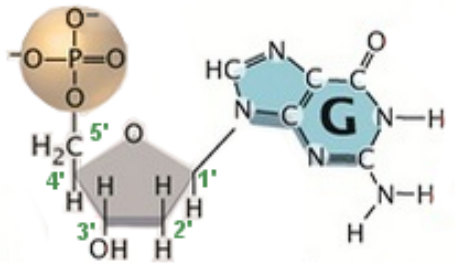


Figura 2.1: imagem de um nucleotídeo e das bases nitrogenadas. Mostrar backbone da pentose 1'...5'. Adaptado de : [3].

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 [4]. 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 [4]. Esta estrutura foi descoberta em 1953, pelo biólogo James Watson e pelo físico Francis Crick [3], 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.

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.

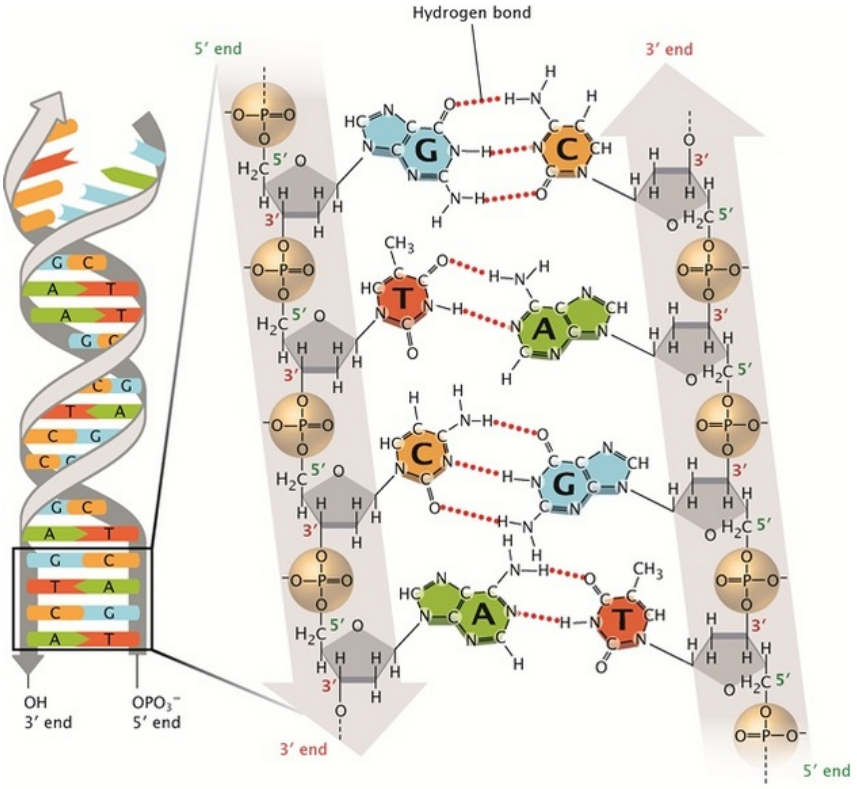


Figura 2.2: Adaptado de : [3]

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 **transcricaoTraducaoSintese**. 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 [4]. 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.

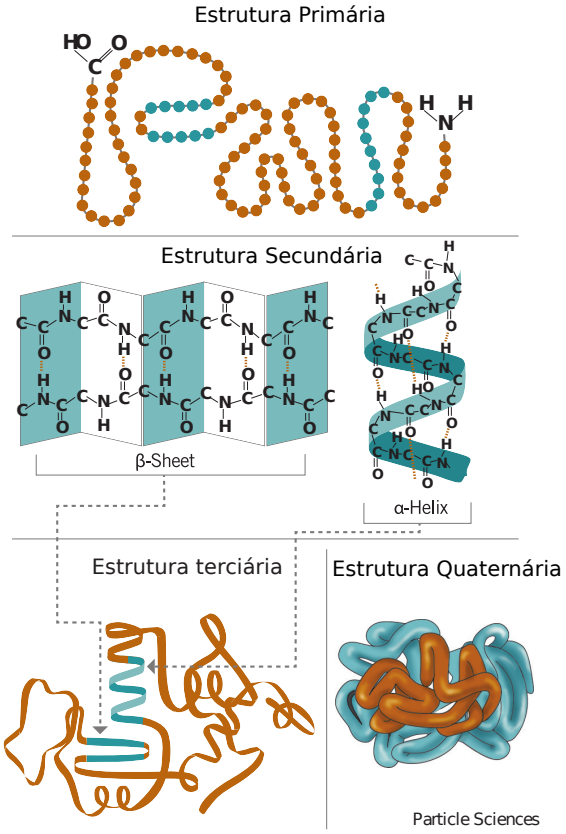


Figura 2.3: Adaptado de : [2]

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. **EXISTE RELAÇÃO ENTRE DUAS ESPÉCIES COM QUASE A MSM QUANTIDADE DE CROMOSSOMOS, THO??**.

Um cromossomo pode ser representado por vários trechos contíguos de DNA, sendo que cada trecho é chamado de **gene**. Portanto, pode-se afirmar que o cromossomo é um conjunto (ou lista) de genes. No caso dos seres humanos, cada cromossomo possui de 20 mil à 25 mil genes, e cada gene possui em média 10 mil pares de base. **BUSCAR A FONTE DISSO. http://www.sobiologia.com.br/conteudos/Corpo/Celula3.php**. 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. A tabela 2.1 que apresenta a correspondência entre códons e aminoácios é chamada representa o **código genético**.

CONCLUIR

[4]

2.2.3 Transcrição e tradução

rRNA, mRNA, tRNA
síntese de proteína

2.3 Bioinformática

2.3.1 Sequenciamento

Mardis 2008

2.3.2 Desafio das ômicas

Genômica
Conceitualização do algortimo.
Artigos: Introdução [introduction to bioinformatics for computer scientists]

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	Asp	Ala	Phe	C
	Cys	Asp	Ala	Phe	U

Capítulo 3

Redes Metabólicas

Dissertação do Waldeyr

Capítulo 4

Banco de Dados NoSQL

NOSQL, NOT ACID, Neo4j, Cypher

Capítulo 5

Resultados

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					
	Jan	Fev	Mar	Abr	Mai	Jun
Revisão bibliográfica	X	X				
Familiaridade com –		X	X			
Implementação			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. vi, 8

[3] 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, 5, 6, 7

[4] João Carlos Setubal and João Meidanis. *Introduction to computational molecular biology*. PWS Publishing Company, 1997. 5, 6, 8, 9