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

Visualização de Vias Metabólicas com Banco de Dados em Grafo

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 de Vias Metabólicas com Banco de Dados em Grafo / Gabriella de Oliveira Esteves. Brasília : UnB, 2016.
63 p. : il. ; 29,5 cm.

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

1. Bioinformática, 2. Redes Metabólicas, 3. Banco de Dados Não Relacional, 4. Grafo, 5. OrientDB

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 de Vias Metabólicas com Banco de Dados em Grafo

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: Bioinformática, Redes Metabólicas, Banco de Dados Não Relacional, Grafo, OrientDB

Abstract

Abstract in english

Keywords: Bioinformatics, Metabolic Networks, Non-Relational Database, Graph, OrientDB

Sumário

Lista de Figuras

Lista de Tabelas

Introdução

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, e assim fizeram a grande descoberta que desencadearam várias linhas de pesquisas atuais: 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.

A partir de então, pesquisadores já sequenciaram cadeias de DNA, RNA e proteína de vários organismos, criando uma quantidade de informação tão extensa que apenas ferramentas de Big Data podem ser usadas para análise, visualização, busca, etc, para um tratamento eficiente. Estes dados, denominados dados ômicos, são armazenados em bancos de dados específicos hoje em dia, mais vários deles colaboram entre si. O foco deste trabalho é apresentar os bancos de dados que representam redes metabólicas em modelo de grafo já existentes bem como suas ferramentas de visualização de vias metabólicas, e apresentar a ferramenta de visualização 2Path formularizada e desenvolvida neste projeto.

Justificativa

Atualmente, a quantidade de dados ômicos 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 e a maneira mais natural é representar tais dados em forma de grafo (redes metabólicas). Esta ferramenta deverá permitir que o usuário visualize e interaja com os dados dinamicamente.

Problema

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

Objetivo

Construir 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 a partir de parâmetros informados pelo pesquisador no sistema
- Recuperar a informação desejada e exibi-la para o pesquisador de forma ergonômica

Descrição dos Capítulos

No Capítulo 1 serão descritos os conceitos básicos de biologia molecular, metabolismo primário e secundário e os bancos de dados mais utilizados para armazenar as informações referentes às redes metabólicas. No Capítulo 2 serão apresentados detalhadamente quatro ferramentas de visualização de redes metabólicas: *KEGG Pathway*, *MetaCyc*, *Reactome Browser* e *Cytoscape*.

Os Capítulos 3 e 4 já são relacionados ao sistema 2Path desenvolvido neste projeto. Enquanto o Capítulo 3 aborda o tema de interação humano-computador para a concepção de uma interface auto-explicativa e consistente, o Capítulo 4 descreve as linguagens e ambientes utilizados para a construção de tal. O Capítulo 5 especifica como são feitas as buscas por vias metabólicas no sistema pelo usuário e apresenta os resultados obtidos.

O Capítulo 6 finaliza o trabalho com a conclusão e sugestão de trabalhos futuros.

Capítulo 1

Redes Metabólicas

Neste capítulo serão descritos os conceitos básicos da biologia molecular, metabolismo e bancos de dados específicos para redes metabólicas. A primeira seção detalha a origem das principais estruturas que promovem o metabolismo tais como DNA e enzima, enquanto que a segunda seção descreve de fato como ocorre o processo. Por fim, a última seção apresenta os principais bancos de dados em grafo voltados para redes metabólicas: *KEGG*, *MetaCyc* e *Reactome*.

1.1 Conceitos Básicos de Biologia Molecular

1.1.1 Ácidos Nucléicos

Os ácidos nucleicos são biomoléculas responsáveis pelo armazenamento, transmissão e tradução das informações genéticas dos seres vivos. Isto é possível devido ao processo de síntese 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).

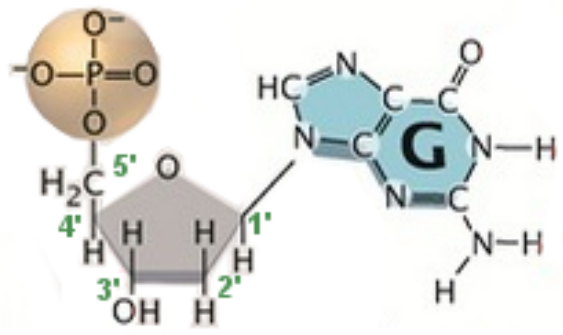

Figura 1.1: imagem de um nucleotídeo e das bases nitrogenadas. Mostrar backbone da pentose 1'...5'. Adaptado de : [18].

Na figura 1.1, 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 nucleico, no processo de polimerização formar-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 [19]. Por definição o sentido da leitura de uma fita de ácido nucleido 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 nucleicos são encontrados nos seres vivos: ácido desoxirribonucleico (DNA ou ADN) e ácido ribonucleico (RNA ou ARN). Eles diferenciam-se tanto na estrutura do *backbone* e nas bases nitrogenadas, quanto em suas funções. Os DNAs 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 [19]. Esta estrutura foi descoberta em 1953, pelo biólogo James Watson e pelo físico Francis Crick [18], porém os ácidos nucleicos 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 1.3. 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 1.3 apresenta a estrutura do DNA como explicada nesta seção.

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 1.1.2. 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.

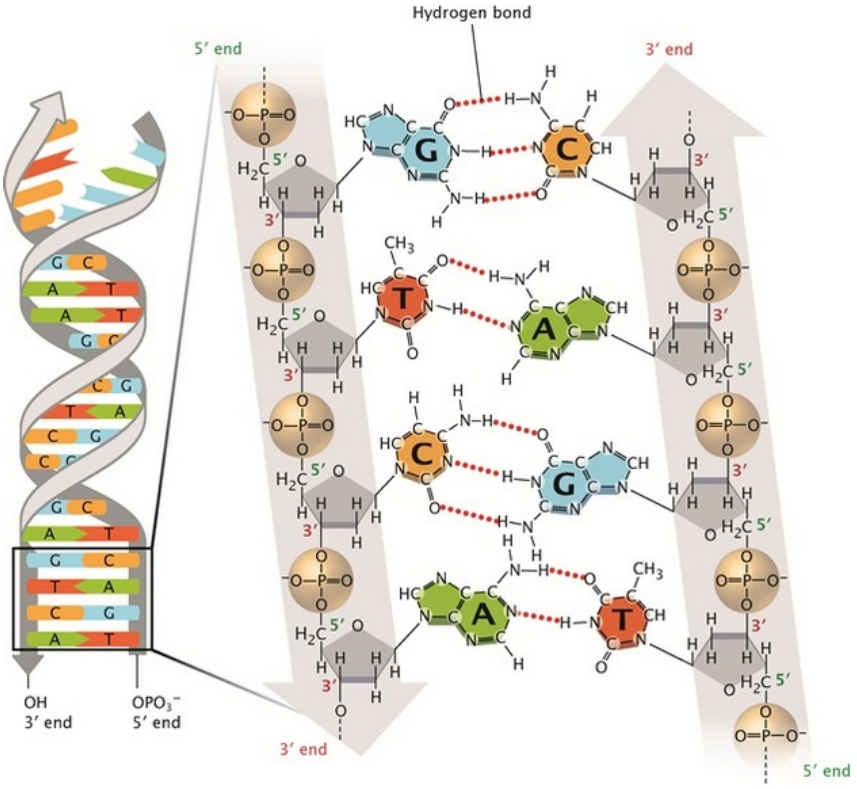

Figura 1.2: Adaptado de : [18]

1.1.2 Síntese de 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 [3]. Cada tarefa é realizada por um proteína com uma estrutura específica e otimizada pra tal.

Assim como os ácidos nucleicos, 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 mais à frente, 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 [19]. 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 (α-Helix) ou em forma de folha (β-Helix). A estrutura terciária é dada pela união de várias estruturas secundárias e, por fim, a estrutura quaternária é composta de múltiplas estruturas terciárias [5].

A **transcrição** é o processo de produção de mRNA a partir do DNA e ele ocorre da seguinte forma: O início de cada gene possui um identificador em uma das fitas para indicar o local da codificação e, a partir dali, uma cópia inversa (A, T, C, G são traduzidos para U, A, G, C respectivamente) do mesmo é feita sob forma de molécula de mRNA que, por consequência, obterá a mesma sequência que a cadeia codificadora (a qual não possui o identificador), porém trocando os U's por Ts.

O mRNA deixa, então, o núcleo celular e inicia a **tradução** 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[1] e no oposto, seu aminoácido correspondente. O processo de tradução se dá da seguinte forma: 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, a molécula de proteína em desenvolvimento é liberada e, com o auxílio da catálise de uma enzima, agregada no aminoácido que estava fixado naquele tRNA. Esta fase é finalmente completa quando o mRNA apresenta um códon de parada, pois nenhum tRNA possui correspondência para tal [19]. Uma proteína simples é, então, formada.

1.2 Conceitos Básicos de Metabolismo

As reações bioquímicas são alterações químicas que fornecem um ou mais produtos a partir de um ou mais substratos. O conjunto de todas as reações bioquímicas que ocorrem dentro de um organismo vivo é chamado de metabolismo, e ele podem ser dividido em dois subconjuntos: catabolismo, quando ocorre a quebra de moléculas complexas produzindo energia, e anabolismo, quando ocorre a síntese de moléculas complexas, o que requer energia. Geralmente a energia liberada pelas reações catabólicas é usada para impulsionar as reações anabólicas[7]. Uma via metabólica é uma sequência de reações bioquímicas, cujo produto e subtrato são denominados metabólitos, que podem ser catalisados por enzimas, as quais muitas vezes necessitam de compostos químicos chamados de co-fatores para realizarem suas atividades na célula. O conjunto de vias metabólicas de um organismo é chamado de rede metabólica.

Enzimas são proteínas responsáveis por auxiliar a realização de biossíntese (construção) e biodegradação de moléculas no metabolismo com o propósito de catalisar (acelerar) reações bioquímicas. As enzimas possuem um local pré-determinado em formato côncavo chamado de sítio ativo, que comporta um ou mais substratos. Se a enzima comporta apenas um substrato, a estrutura que se forma com o preenchimento do sítio ativo é um complexo enzima-substrato, porém se ela comporta mais de um substrato, a estrutura é chamada de complexo ternário intermediário [15]. Quando a atividade catalítica não

[1] Sequência de três nucleotídeos.

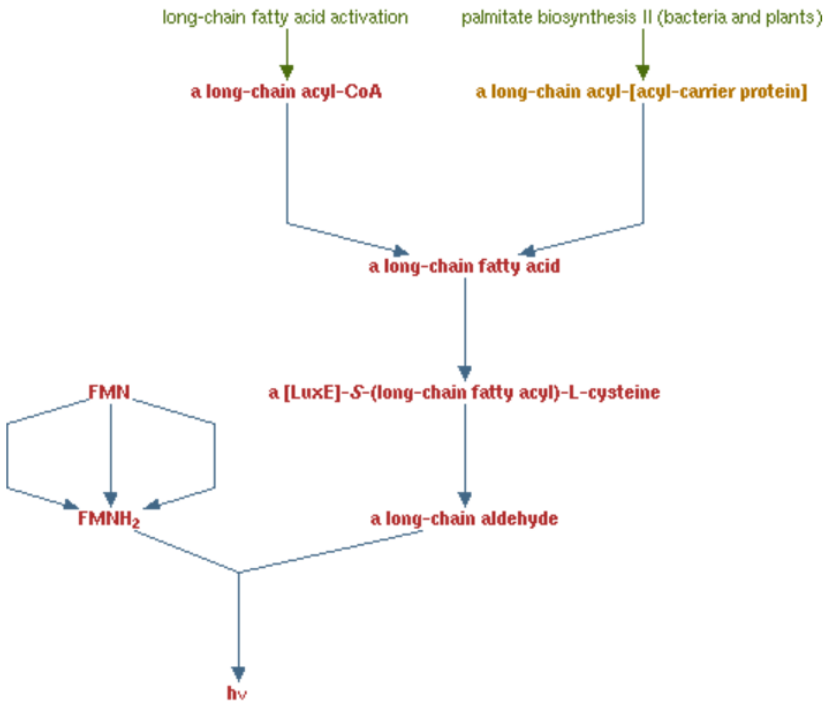

Figura 1.3: Exemplo de via metabólica retirado do site MetaCyC. Esta representa o processo de produção da bioluminescência de bactérias [8].

pode ser realizada apenas pela enzima, co-fatores auxiliam o processo. Eles podem ser coenzimas, associadas momentaneamente às enzima, ou grupos prostéticos, associados firmemente à elas [15]. Quando duas enzimas possuem a mesma atividade enzimática porém estruturas físicas diferentes são chamadas isoenzimas[15].

1.2.1 Metabolismo Primário

O metabolismo, além de ser subdivido em catabolismo e anabolismo, também pode ser classificado em relação à função que exerce no organismo. Se a função realizada é fundamental no organismo, como crescimento, desenvolvimento e reprodução, ele é denominado metabolismo primário [21]. Mitose e meiose são exemplos de metabolismos primários.

1.2.2 Metabolismo Secundário

No caso do metabolismo não realizar função essencial no organismo, ele é classificado como metabolismo secundário. Estes são caracterizados pela vasta diversidade química e, desta forma, são responsáveis pela sobrevivência do organismo em diferentes meio ambientes de acordo com os fatores bióticos (elementos causados pela interação entre organismos como, por exemplo, cadeia alimentar) e abióticos (elementos naturais independente de organismos como, por exemplo, luz e temperatura) [21]. Enquanto 20% dos metabólitos secundário são encontrados em bactérias, fungos e organismos sésseis[2] marinhos, os ou-

[2]Que vivem fixos, sem capacidade de locomoção.

tros 80% encontram-se em plantas vasculares [21] e estes podem ser subdivididos em três classes: terpenóides, alcalóides e fenólicos [13].

- Os **terpenóides** constituem no grupo mais abundante de produtos naturais, apresentando uma grande variedade estrutural e funcional, principalmente no Reino *Plantae*. No metabolismo secundário, possem funções como produção de óleos, esteroides, cera, resinas e borracha natural, produção de compostos usados para defesa contra herbívoros ou aromas usados para atrair polinizadores [21].
- Os **alcalóides** são majoritariamente tóxicos à outros organismos diferentes daquele que os produz. Nesse sentido, eles possuem nas plantas função de defesa contra herbírovos e podem ser encontrados principalmente nos locais mais propícios à ataques como, por exemplo, nas sementes, flores e tecidos periféricos em crescimento. Para o consumo dos seres humanos, são usados na fabricação de estimulantes, como cafeína e nicotina, e drogas, como morfina[21]. Por apresentarem alta diversidade estrutural, é difícil classificá-los; a tentativa mais recente se baseia na semelhanças entre os esqueletos carbônicos [13].
- Os **fenólicos** são caracterizados por suas propriedades anti-oxidantes, anti-inflamatória e anti-cancerígena e muitos deles são bactericidas, antisépticos e vermífugo. Eles estão presentes em praticamente todas as plantas e são utilizados na química, biologia, agricultura e medicina [13].

1.3 Banco de Dados de Redes Metabólicas

Desde o descobrimento da estrutura do DNA por Crick e Watson, o número de sequências de proteínas descobertas cresceu, aumentando 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 construir 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* [16]. 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 possível 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[3]. 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 com o objetivo de catalogar sequências genéticas e coleçoes de anotações de todos os DNAs públicos, agora, com o patrocínio do *National Center for Biotechnology Information*. Os dois bancos são públicos e continuam crescendo exponencialmente [16].

Nos dias de hoje, a quantidade de dados é tão grande e que os biólogos enfrentam dificuldades em tarefas como análise, busca, armazenamento, visualização e atualização de dados. Nesse sentido, eles utilizam agora ferramentas de Big Data em suas pesquisas.

[3] 1 PAM é uma medida de tempo para representar 1 mutação para cada 100 aminoácidos.

Existem grandes áreas da biologia voltadas para estudo deste dados (chamados de dados ômicos), tais como, por exemplo, genoma[4], transcritoma[5], proteoma[6], metaboloma[7] e interactoma[8]. Nesta seção serão apresentados três bancos de dados utilizados no estudo do metaboloma cuja estrutura de dados que representam as redes metabólicas são grafos.

1.3.1 Banco de Dados NoSQL

Para comportar tamanha quantidade de dados e ao mesmo tempo providenciar alta performance, atualmente é possível utilizar um banco de dados NoSQL como alternativa ao tradicional modelo relacional. NoSQL pode significar Não-Relacional ou *Not Only SQL* [9], para ressaltar as vantagens sobre o modelo tradicional, que são rápida leitura e escrita dos dados, suporte à armazenamento em larga escala, fácil escalabilidade, fácil distribuição (replicação ou fragmentação) de dados entre vários servidores, e baixo custo. [11]. Entretanto, uma grande desvantagem é sua falta de suporte à linguagem SQL, o que obriga o usuário a aprender a recuperar seus dados de outra maneira[9].

Em 2000, o Professor Eric Brewer da Universidade da Califórnia em Berkeley estabeleceu o Teorema CAP (*Consistency, Availability and Partition tolerance*), que afirma que quando os dados estão particionados[10] em rede, os BDs NoSQL que são distribuídos devem tolerar tal falha e oferecer, assim, apenas duas escolhas para o usuário: consistência ou disponibilidade de dados, nunca os dois ao mesmo tempo. [11] [9]. Caso o banco decida fornecer mais consistência do que disponibilidade, o sistema tentará sempre retornar a *query* com dados atualizados para o usuário, mesmo que a operação demore. A Figura 1.4 ilustra o teorema também conhecido como Teorema de Brewer. Caso decida fornecer mais disponibilidade do que consistência, o sistema tentará retornar a *query* o mais rápido possível, mesmo se o dados estiverem desatualizados. Esta *trade-off* inerente na maioria dos BDs NoSQL pode, na verdade, ser decidida pelo próprio usuário ao configurar seu banco de dados.

1.3.1.1 Banco de Dados em Grafo

Grafos são as estruturas mais úteis para se representar interações entre objetos, portanto são de grande interesse no campo biológico, na modelagem de dados ômicos [20]. Bancos de dados em grafo podem ser utilizados, por exemplo, quando a finalidade é buscar via *query* a procedência dos objetos[11], pois realizar uma travessia em grafo é bem mais rápido computacionalmente do que realizar múltiplos *joins* num modelo relacional.

Ao se trabalhar com banco de dados em grafos, é importante ter conhecimento das principais terminologias a seguir:

- Adjacência: Nós que dividem uma aresta incidente ou arestas que dividem um nó incidente;

[4] Material genético de um organismo.
[5] Conjunto dos transcritos mRNA, tRNA, rRNA e microRNA.
[6] Conjunto de proteínas e suas variantes em um organismo.
[7] Conjunto de metabólitos de um organismo.
[8] Conjunto de interações moleculares em um organismo.
[9] No banco de dados Neo4j, por exemplo, usa-se a linguagem CYPHER.
[10] Falha em algum dispositivo de rede que causa a separação entre dois conjuntos de computadores.
[11] A procedência é a linhagem do dado, ou seja, os detalhes de sua criação desde a origem conhecida.

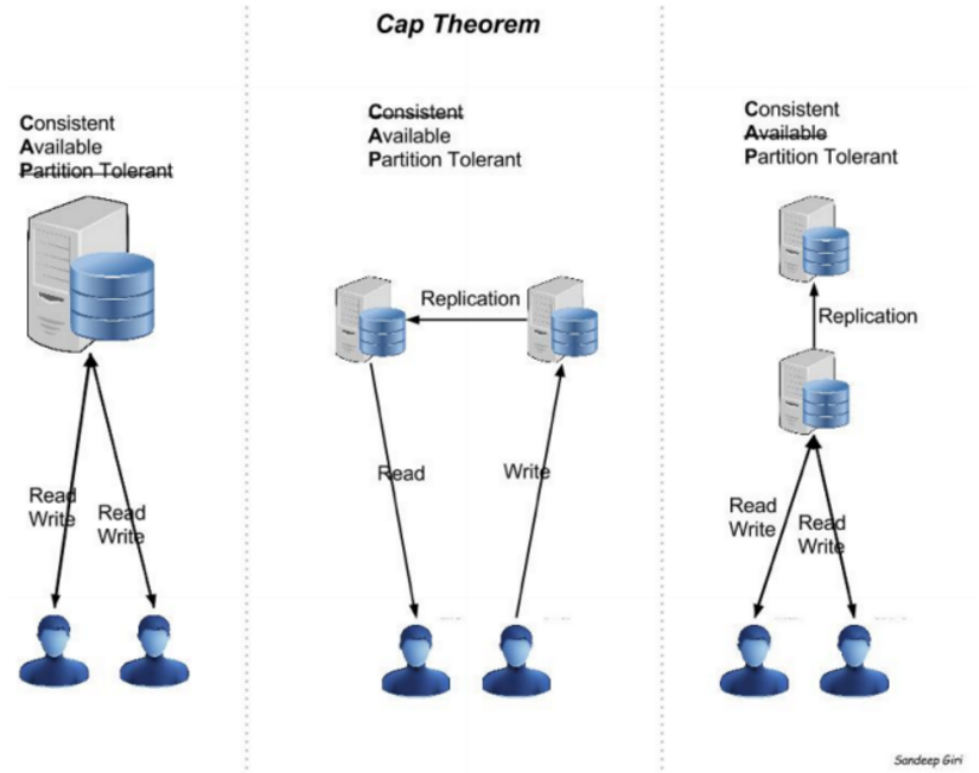

Figura 1.4: Teorema CAP. Na primeira coluna, o banco de dados não suporta partição de rede, na segunda existe replicação parcial de dados entre os servidores e na terceira, replicação total [6].

- Aresta: Uma registro que representa relação entre um par de objetos. Pode possuir direção, quando a relação possui origem e destino, ou não;
- Caminho: Coleção de alternados nós e arestas;
- *Constraint*: Restrição definida pelo usuário que o banco deve impor aos dados;
- DAG: Um grafo direcionado e acíclico;
- Incidência: Aresta adjacente associada à um nó, ou nó associado à uma aresta;
- *Label*: Rótulo de um nó que o agrupa em algum subconjunto;
- Loop: Uma aresta que conecta um nó à ele mesmo;
- Nó: Um registro que representa um objeto e possui um número indefinido de propriedades, labels e arestas incidentes;
- Propriedade: Atributo armazenado em um nó ou aresta;
- Registro: Unidade de armazenamento;
- Vizinho: Nó conectado por uma aresta em comum.

1.3.2 KEGG

O KEGG[12] (*Kyoto Encyclopedia of Genes and Genomes*) é uma base de informações sobre sistemas biológicos em nível molecular, sobretudo sobre conjuntos de dados em larga escala gerados por sequenciamento de genoma [2]. As informações sobre os sistemas podem ser dadas em forma de módulos, unidades funcionais com identificação otimizada para análise dos dados, em forma de *brite*, coleção de arquivos estruturados hierarquicamente sobre as funções das entidades biológicos, ou em forma de vias, mapa de interações moleculares e reações químicas. Dado que o metabolismo é um conjunto de reações e transformações químicas, a maneira natural de representá-lo é por meio de uma rede de interações, ou seja, em forma de vias. O KEGG oferece uma ferramenta de busca de vias metabólicas sobre várias rede metabólica, dos vários organismos que constituem o banco de dados.

1.3.3 BioCyc

O BioCyc[13] é um sistema de coleção de aproximadamente 7 mil bancos de dados chamados PGDBs (*Pathway/Genome Databases*) pois possuem duas maneiras diferentes de representar as informações: modelo de vias metabólicas, que enfatiza as sequências de reações, substratos e produtos de múltiplos organismos, ou modelo de sequência genômica, que destaca a localização e descrição dos genes de cada organismo específico [1]. Os bancos PGDBs são organizado em três camadas de acordo com a frequência de atualizações/refinações e da maneira com que os dados foram obtidos. O BioCyc possui um banco de dados específico para redes metabólicas determinadas experimentalmente, chamado MetaCyc. Este é o único banco de dados multi-organismos do grupo BioCyc e ele é referência na ferramenta gratuita *Pathway Tools* desenvolvida pelo instituto de pesquisa *SRI International*.

1.3.4 Reactome

Reactoma[14] é um banco de dados de reações de mudança de estado, ou seja, além de reações bioquímica, ele também abrange reações de ativação, de degradação e de ligação, por exemplo [4]. Ele faz uma ligação sistemática entre as proteínas de um certo organismo e as funções moleculares do mesmo, fornecendo uma base de funções que pode ser utilizada para pesquisas sobre expressão de genes ou mutações somáticas. O Reactome disponibiliza o *Pathway Browser*, uma rede geral para cada organismo, que representa os vários seus sistemas, como reprodução e metabolismo, por exemplo. Algumas sub-redes estão conectadas (por exemplo, replicação de DNA e ciclo de célula), outra não (por exemplo, contração muscular e reprodução). Nesta rede, cada nó representa uma via cujo número de entidade se reflete no raio do nó, e cada aresta representa a relação entre estas vias. O site ainda possui uma ferramenta de análise de dados baseada nas correspondências entre as reações na redes dos organismos comparados.

[12]Disponível pela *web* através do site *http:www.kegg.jp*.
[13]Disponível pela *web* através do site *http:biocyc.org*.
[14]Disponível pela *web* através do site *http:www.reactome.org*.

Capítulo 2

Ferramentas de visualização de Redes Metabólicas

2.1 KEGG

http:www.kegg.jp ->
Data-oriented entry points ->
KEGG PATHWAY ->
Pathway Maps ->
1. Metabolism ->
1.0 Global and overview maps / Metabolic pathways

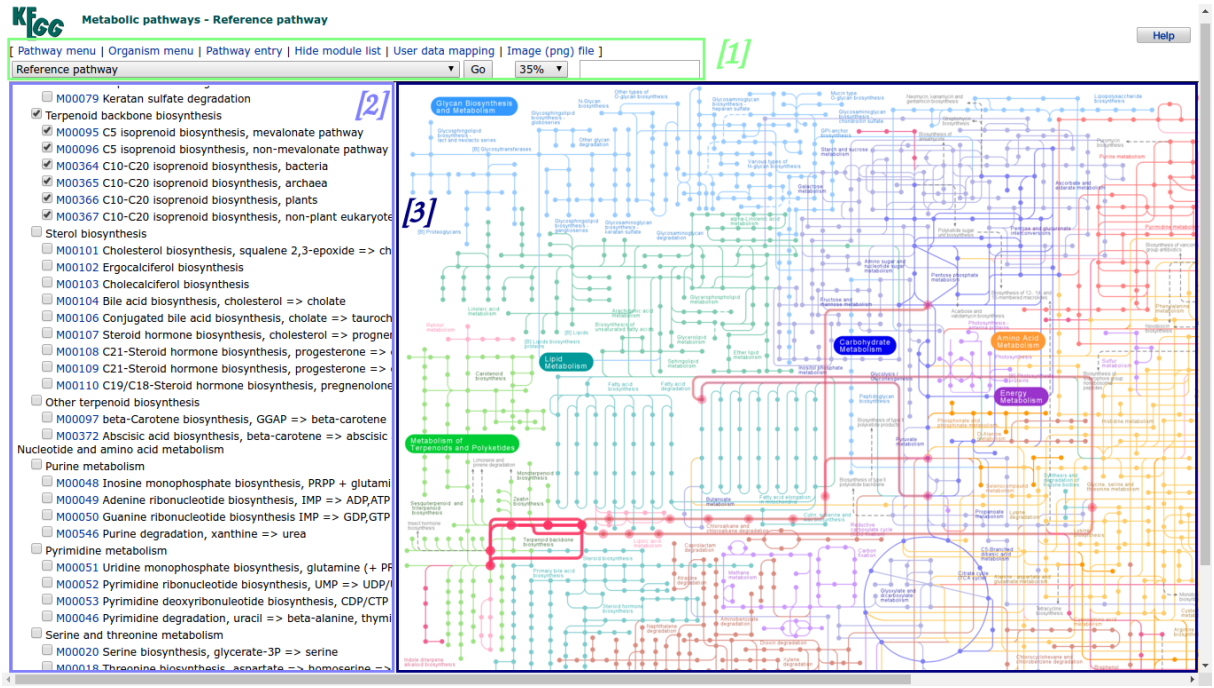

Figura 2.1: Visão geral da rede metabólica de referência do KEGG, com destaque na via metabólica de biossíntese do *backbone* de terpenóides. (1), (2) e (3).

Notação: http:www.genome.jpkeggdocumenthelp_pathway.html

http:www.kegg.jp ->
Data-oriented entry points ->
KEGG PATHWAY ->
Pathway Maps ->
1. Metabolism/Terpenoid/PK ->
1.9 Metabolism of terpenoids and polyketides / Terpenoid backbone biosynthesis

Tipo de arquivo: Objetos clicáveis, setas e containers com significado (Notação: http:www.genome.jp
minimapa ao passar mouse por cima de vias representadas por nós, etc

2.2 MetaCyc

Figura do capitulo anterior: bioluminescencia.

2.3 Reactome Browser

2.4 Cytoscape

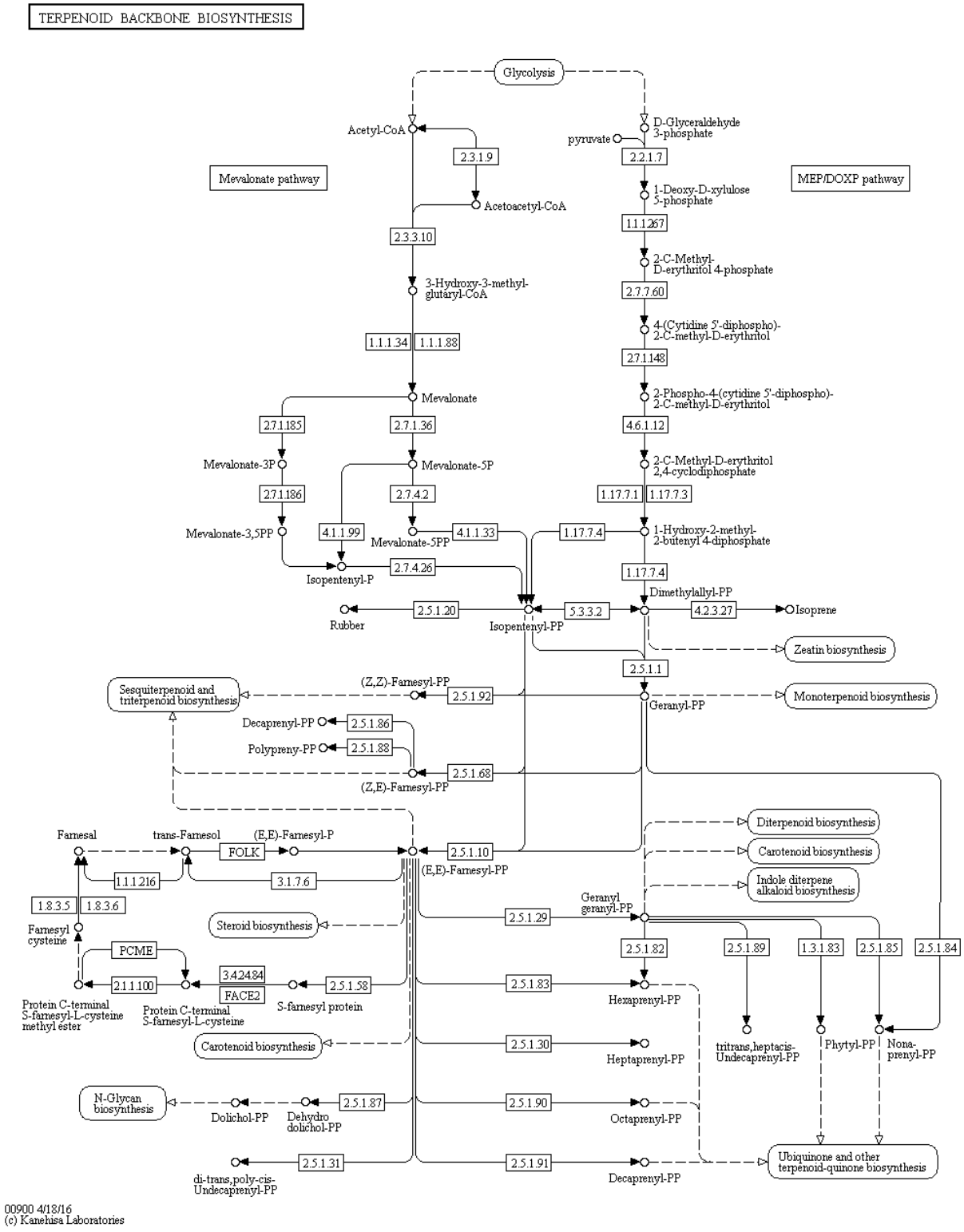

Figura 2.2: Via metabólica de biosíntese do *backbone* de terpenóides

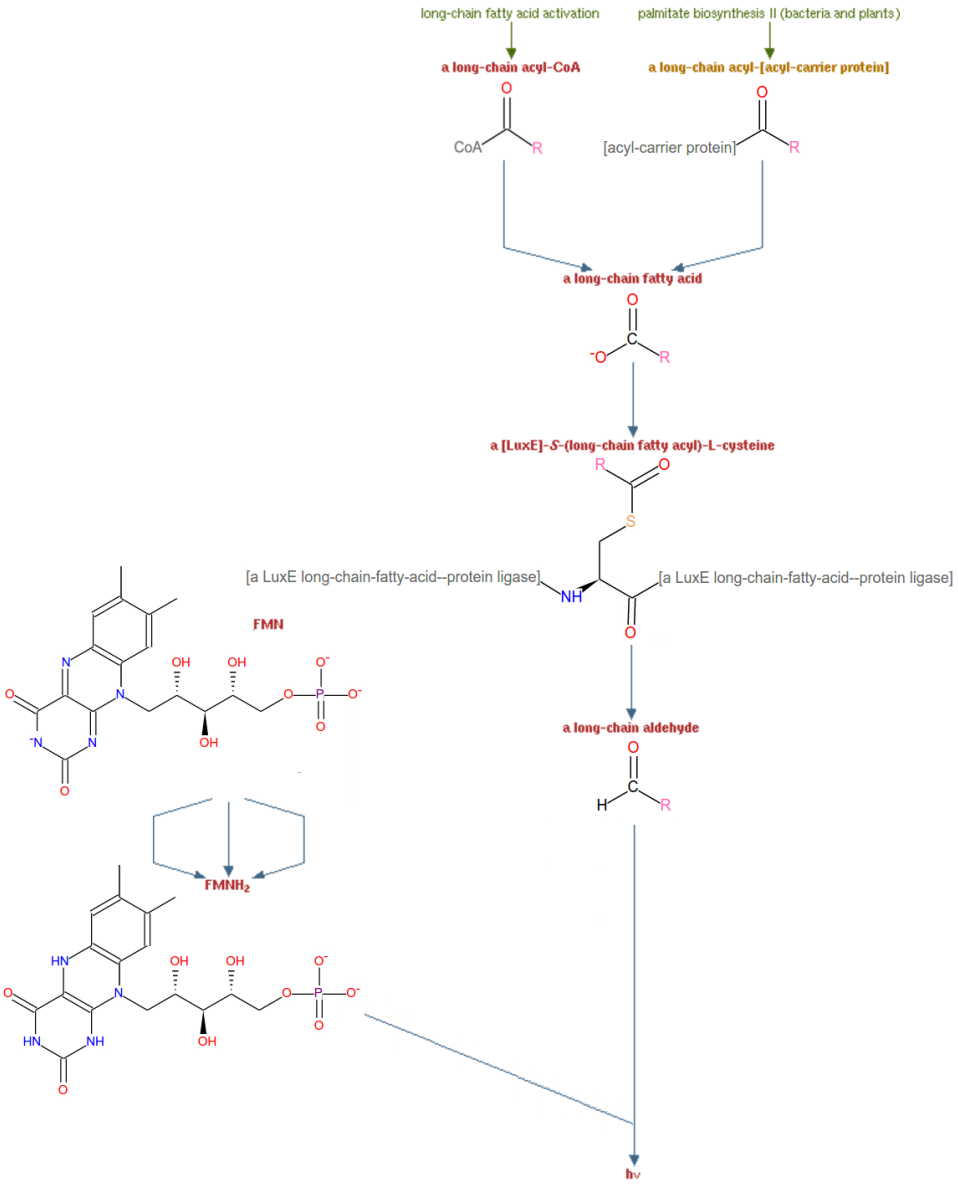

Figura 2.3: Mesa figura do capitulo 1, mas com maior nível de detalhes

Capítulo 3

IHC

Capítulo 4

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.

4.1 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 [10].

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 [17] 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 [12]

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

4.1.1 Banco de Dados OrientDB

sobre ACID
Modelo CAP
JAVA API

4.2 Visualização das redes metabólicas

JAVASCRIPT
ANGULARJS
ORIENTBD

4.3 Desafios

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

Capítulo 5

Método e Resultados

Capítulo 6

Conclusão e Trabalhos Futuros

6.1 Conclusão

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

6.2 Trabalhos Futuros

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

- x

Capítulo 7

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 7.1: 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] Introduction to biocyc. http://biocyc.org/intro.shtml, visitado em 2016-10-01. 11

[2] Kegg overview. http://www.kegg.jp/kegg/kegg1a.html, visitado em 2016-10-01. 11

[3] Proteínas. http://www.professoraangela.net/documents/proteinas.html, visitado em 2016-01-02. 5

[4] Usersguide. http://wiki.reactome.org/index.php/Usersguide, visitado em 2016-10-01. 11

[5] Protein structure, 2009. http://www.particlesciences.com/news/technical-briefs/2009/protein-structure.html, visitado em 2016-01-02. 6

[6] Abhishek Bansal. Nosql, cap theorem (part of mongodb course). https://www.knowbigdata.com/blog/nosql-cap-theorem-part-mongodb-course, visitado em 2016-10-13. vii, 10

[7] Perry Carter. Catabolic and anabolic reactions. http://classes.midlandstech.edu/carterp/courses/bio225/chap05/lecture1.htm, visitado em 2016-10-12. 6

[8] Ron Caspi. Metacyc pathway: bacterial bioluminescence. http://metacyc.org/META/new-image?object=PWY-7723, visitado em 2016-10-13. vii, 7

[9] Rick Cattell. Scalable sql and nosql data stores. *SIGMOD Rec.*, 39(4):12–27, May 2011. 9

[10] 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. 17

[11] Jing Han, E. Haihong, Guan Le, and Jian Du. Survey on NoSQL database. In *Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on*, pages 363–366. IEEE, October 2011. 9

[12] Chua Hock-Chuan. Java web database applications, 2011. https://www.ntu.edu.sg/home/ehchua/programming/java/JavaWebDBApp.html, visitado em 2016-08-19. 18

[13] Justin N. Kabera, Edmond Semana, Ally R. Mussa, and Xin He. Plant secondary metabolites: Biosynthesis, classification, function and pharmacological properties. *Journal of Pharmacy and Pharmacology*, 2:377–392, 2014. 8

[14] Sam Kean. *O Polegar do Violinista - e Outras Histórias da Genética Sobre Amor, Guerra e Genialidade*. Zahar.

[15] Gerhard Michal and Dietmar Schomburg. *The Cell and Its Contents*, pages 14–36. John Wiley & Sons, Inc., 2012. 6, 7

[16] David W. Mount. *Bioinformatics : sequence and genome analysis*. Cold Spring Harbor, N.Y. Cold Spring Harbor Laboratory Press, 2001. 8

[17] Oracle. *Java Platform, Enterprise Edition The Java EE Tutorial, Release 7*, 2014. https://docs.oracle.com/javaee/7/tutorial, visitado em 2016-08-19. 18

[18] 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, 3, 4, 5

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

[20] Chad Vicknair, Michael Macias, Zhendong Zhao, Xiaofei Nan, Yixin Chen, and Dawn Wilkins. A comparison of a graph database and a relational database: A data provenance perspective. In *Proceedings of the 48th Annual Southeast Regional Conference*, ACM SE '10, pages 42:1–42:6, New York, NY, USA, 2010. ACM. 9

[21] Röbbe Wünschiers, Martina Jahn, Dieter Jahn, Ida Schomburg, Susanne Peifer, Elmar Heinzle, Helmut Burtscher, Julia Garbe, Annika Steen, Max Schobert, Dieter Oesterhelt, Josef Wachtveitl, and Antje Chang. *Metabolism*, pages 37–209. John Wiley & Sons, Inc., 2012. 7, 8