UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS CORNÉLIO PROCÓPIO

GABRIEL RUBINO

VISUALIZAÇÃO DE REDES GÊNICAS A PARTIR DA INTEGRAÇÃO DE DADOS BIOLÓGICOS

TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO

2016

GABRIEL RUBINO

VISUALIZAÇÃO DE REDES GÊNICAS A PARTIR DA INTEGRAÇÃO DE DADOS BIOLÓGICOS

Trabalho de conclusão de curso apresentado à disciplina Trabalho de Conclusão de Curso 2 da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do título de Engenheiro de Computação

Orientador: Prof. Dr. Fabrício Martins Lopes

CORNÉLIO PROCÓPIO

2016

RESUMO

RUBINO, Gabriel. VISUALIZAÇÃO DE REDES GÊNICAS A PARTIR DA INTEGRAÇÃO DE DADOS BIOLÓGICOS. 38 f. Trabalho de conclusão de curso – Campus Cornélio Procópio, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2016.

Com o surgimento da geração massiva de dados biológicos, que é um caso de *BigData*, existe a necessidade do desenvolvimento de metodologias que possam extrair informações a partir desse grande volume de dados. Nesse contexto, muitas bases de dados biológicos estão disponíveis na internet, as quais tornam possível o uso, atualizações, correções, entre outras ações, que levam a possibilidade de serem processados e analisados por sistemas externos às bases de dados. Entre as fontes de dados biológicos podem se ter como exemplo o TAIR, PO e o GO dedicados a disponibilizar dados sobre genes tais como suas funções e características. Esse trabalho visa a integração de dados biológicos com o objetivo de gerar uma rede de características genéticas. Outro ponto de destaque é a criação de métodos para a visualização das redes e grafos gerados. Para isso algumas ferramentas foram utilizadas. Uma dessas ferramentas é o Neo4j responsável por gerenciar e armazenar o banco de dados de grafos. Outra ferramenta usada foi o JavaScript juntamente com a biblioteca d3.js usada para a representação visual de redes e grafos. Todos os métodos presentes neste trabalho e a consistência do banco foram validados com o uso de informações sobre os genes disponibilizadas em outros trabalhos sobre *Arabdopsis thaliana*. Esse mecanismo de validação foi realizado pela comparação das informações geradas com as informações dos trabalhos relacionados. Como resultado foi desenvolvida uma ferramenta para visualização de redes gênicas a partir da integração de dados biológicos, com interface amigável, possibilidade de selecionar os dados biológicos para a geração da rede e com identificação de cores por função biológica.

Palavras-chave: Redes, Grafos, Visualização, Características de genes, Banco de dados de grafos, Neo4j.

ABSTRACT

RUBINO, Gabriel. VISUALIZATION OF GENE NETWORKS BASED ON THE INTEGRATION OF BIOLOGICAL DATA. 38 f. Trabalho de conclusão de curso – Campus Cornélio Procópio, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2016.

The creation of huge amounts of data is called BigData. One of the problems that BigData brings is the extration of information from this massive database, so it is a challange to create a methodology capable of doing this task. On the context of BigData, a lot of biological databases are avaliable online. These databases let users interact, update and even correct its data. These and other actions make possible its access from external systems. Some biological databases are TAIR, GO and PO. These databases are dedicated to share gene's functions and its characteristics. One of this work's goals is to integrate biological data in a way that is possible to generate a gene network. To help the accomplishment of this goal some tools were used. The first one is Neo4j that is responsable for the data base's management and its storage. Another tool used was JavaScript along with the library d3.js that was used to create the visual representation of the netwrok. All the methods present in this project were tested to guarantee its reliability. To do this task information of genes from other papers were used. These informations were compared with the results from this project to check its consistency. As a result of this project a tool for visualization of a gene network and data integration was developed. This tool has a friendly interface with the possibility of selecting the type of biological data to generate the gene's networks as well as biological's functions with color identification.

Keywords: Networks, Graphs, Visualization, Gene's characteristics, Graphs database, Neo4j.

LISTA DE FIGURAS

LISTA DE SIGLAS

NCBI	*National Center for Biotechnology Information*
GO	*Gene Ontology*
TAIR	*The Arabidopsis Information Resource*
PO	*Plant Ontology*
DNA	*Deoxyribonucleic acid*
RNA	*Ribonucleic acid*

SUMÁRIO

1 INTRODUÇÃO

Com o aumento da disponibilidade dos dados proporcionada pelo desenvolvimento de tecnologias cada vez mais avançadas em diversas áreas, como por exemplo imagens, aúdio, astronomia, biologia, entre outras, dentre as quais muitas delas são disponibilizadas de forma online, houve o surgimento de plataformas para organizar essas informações, sendo essas mantidas por universidades, empresas e governos. Esses dados muitas vezes são confiáveis e são vastamente utilizados para pesquisa e desenvolvimento tecnológico (MARX, 2013).

Neste contexto apresentado, os bancos de dados biológicos se caracterizam como um exemplo dessa prática. Essas entidades recebem vários dados, preenchidos por pesquisadores espalhados por todo o mundo. O acesso se dá através de plataformas online como por exemplo o TAIR (TAIR, 2016), usado para guardar dados de genes de *Arabidopsis thaliana* (INITIATIVE et al., 2000a).

Unir os dados de genes (Seção 2.6) de um mesmo organismo a fim de inferir uma rede gênica é um desafio (KELEMEN et al., 2008). Esse problema acontece devido ao elevado número de genes e a baixa quantidade de experimentos em proporção, esse é um fenômeno conhecido como maldição da dimensionalidade (BISHOP, 1995).

Uma alternativa para contribuir neste cenário é usar a integração de dados biológicos a partir de bancos de dados biológicos públicos na internet. Para abordar o problema da integração dos dados muitas vezes são usados organismos cujo genoma já é totalmente conhecido como é o caso *Arabidopsis thaliana* (INITIATIVE et al., 2000a), pois dessa forma validar as relações inferidas entre os genes torna-se mais adequado devido à disponibilidade de dados sobre esse organismo.

Depois dos dados estarem integrados em forma de um grafo, muito usados para mostrar dependências (PEARL, 2014), sua visualização se torna possível. Para isso o grafo deve representar os seus nós, genes e características, por círculos e suas relações como arestas.

1.1 PROBLEMA

Como descrito na seção anterior existem muitos dados biológicos disponíveis e a integração deles pode ser feita para o estudo do organismo em questão. Assim a resolução do problema de integração dos dados e sua visualização em forma de grafo foi dividida em três etapas. As etapas são interdependentes e devem ser resolvidas na ordem que serão apresentadas a seguir.

1.1.1 SELEÇÃO DE DADOS

Uma das preocupações quando se integra dados é escolher quais fontes de dados serão mais relevantes para conseguir completar os objetivos propostos. Visto que cada base de dado possui perfis diferentes e portanto dados focados em diferentes áreas. Como resumidamente descrito na Etapa 1 da Figura 1.

1.1.2 INTEGRAÇÃO DE DADOS

Depois da obtenção dos dados das fontes de dados é necessário processar esses arquivos de modo a organizá-los para a futura criação da rede. Por isso a definição das informações relevantes deve ser explorada. Como resumidamente descrito na Etapa 2 da Figura 1.

1.1.3 VISUALIZAÇÃO

Os grafos (Seção 2.3.1) são uma boa solução para a visualização de redes, mas seu armazenamento e gerenciamento muitas vezes são complexos devido ao grande número de nós incluídos em sua estrutura. Por isso métodos de visualização foram criados. Como resumidamente descrito na Etapa 3 da Figura 1.

O armazenamento dos grafos pode ser feito em banco de dados relacionais estruturados em tabelas que guardam as entidades e suas relações. Esse tipo de persistência dos dados é eficaz em alguns casos, mas muitas vezes a extração das informações do grafo não é fácil, pois é preciso criar instruções de requisição de dados complexos.

1.2 JUSTIFICATIVA

Este trabalho é proposto com a finalidade de facilitar o estudo dos relacionamentos entre os genes de um organismo, pois irá automatizar muitas etapas da integração de dados disponíveis

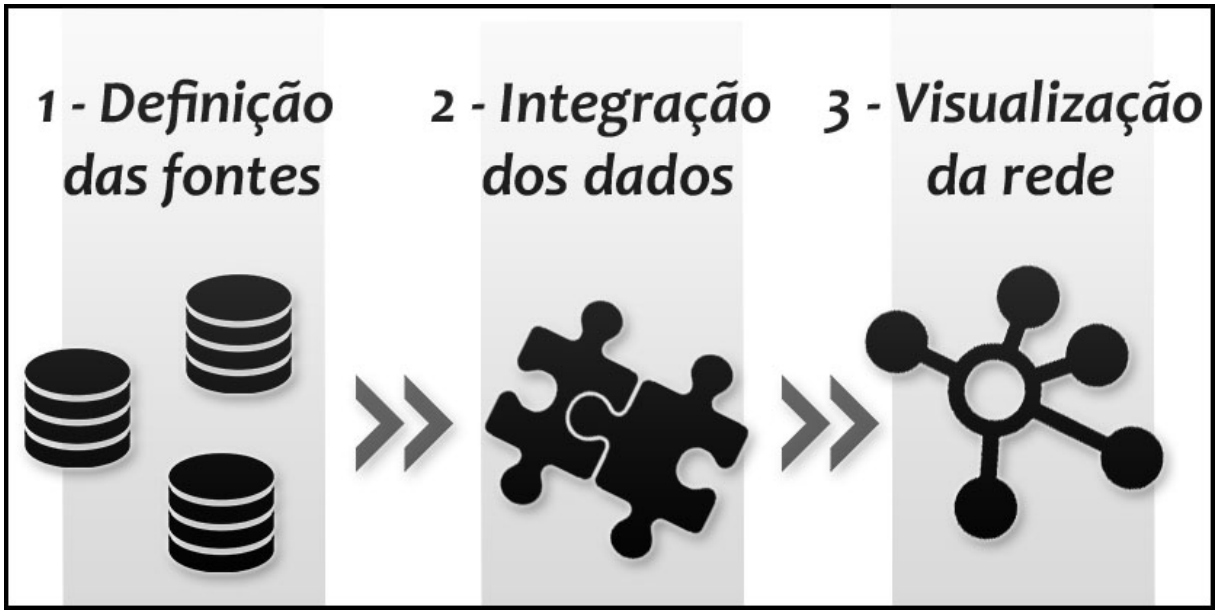


Figura 1: Etapas principais do projeto.

em bancos de dados biológicos. Além disso esse projeto tem como objetivo disponibilizar uma ferramenta de visualização da rede em forma de grafo possibilitando a extração de características do mesmo.

Uma rede de genes auxilia na visualização de relações entre várias entidades direta e indiretamente. Isso é possível, pois medidas de grafos podem ser aplicadas como por exemplo: mensurar os nós mais conectados e as distâncias entre as relações. Com um modelo de grafo simulações podem ser feitas a fim de validar algumas hipóteses sem a necessidade de muitos testes *in vitro*.

1.3 OBJETIVOS

1. Implementação de rotinas de leitura, pré-processamento e visualização de grafos genéricos.
2. Desenvolvimento e implementação de metodologias para a integração de informações biológicas nas redes gênicas.
3. Desenvolvimento e implementação de metodologias para a visualização de redes gênicas.
4. Desenvolvimento e implementação de metodologias para extração de informações de redes gênicas.

1.4 ORGANIZAÇÃO DO TEXTO

O trabalho está dividido em cinco capítulos e referências bibliográficas.

O primeiro capítulo mostra qual o contexto onde o trabalho será desenvolvido tendo como base uma fatia do ambiente ao qual o projeto está inserido. Além disso esta seção descreve o problema a ser resolvido e os objetivos para sua solução.

O segundo capítulo oferece os conceitos necessários para o bom entendimento do trabalho, revisando os principais assuntos que farão parte do desenvolvimento.

O terceiro capítulo traça as etapas necessárias para se executar os objetivos de maneira mais detalhada, para que dessa forma seja possível detalhar o desenvolvimento para se completar o projeto.

O quarto capítulo mostrará os resultados obtidos desse trabalho e como sua validação foi feita.

O quinto e último capítulo irá definir o escopo do presente projeto e também quais assuntos não serão abordados na solução dos problemas existentes. Assim como observações gerais sobre o trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados os tópicos mais relevantes para o entendimento do tema e seus problemas. Essa seção mostrará alguns trabalhos relacionados e fundamentos de biologia e computação.

2.1 MALDIÇÃO DA DIMENSIONALIDADE

A maldição da dimensionalidade (BISHOP, 1995) ocorre quando as amostras usadas para a inferência ou uma classificação são dadas por uma função exponencial da dimensão das características. Esse termo é usado quando existe um aumento muito grande da dimensão e os dados ficam muito separados fazendo com que eles se tornem estatisticamente inadequados dificultando a classificação.

A fim de superar esse obstáculo os dados devem crescer exponencialmente junto com a dimensão de forma consistente. Outra solução é organizar ou procurar locais onde os dados formam grupos com características similares, como exemplificado na Figura 2, para reduzir a dimensão analisada (INDYK; MOTWANI, 1998).

2.2 BIOLOGIA SISTÊMICA

A parte da ciência responsável pelo estudo do organismo como um todo é a biologia sistêmica. Ela é responsável por explicar como todas as partes de um organismo trabalham em conjunto para garantir a vida. Uma parte dessa área tem o foco na genética e interações entre genes (ALON, 2006).

Existem várias ferramentas utilizadas para medir a expressão de genes, uma delas é exibida na Seção 2.6.6. Essas ferramentas sozinhas não conseguem montar toda a rede de interações. Para tentar se obter a rede são utilizados, em conjunto com esses dados, métodos matemáticos, computacionais e biológicos (LOPES, 2011).

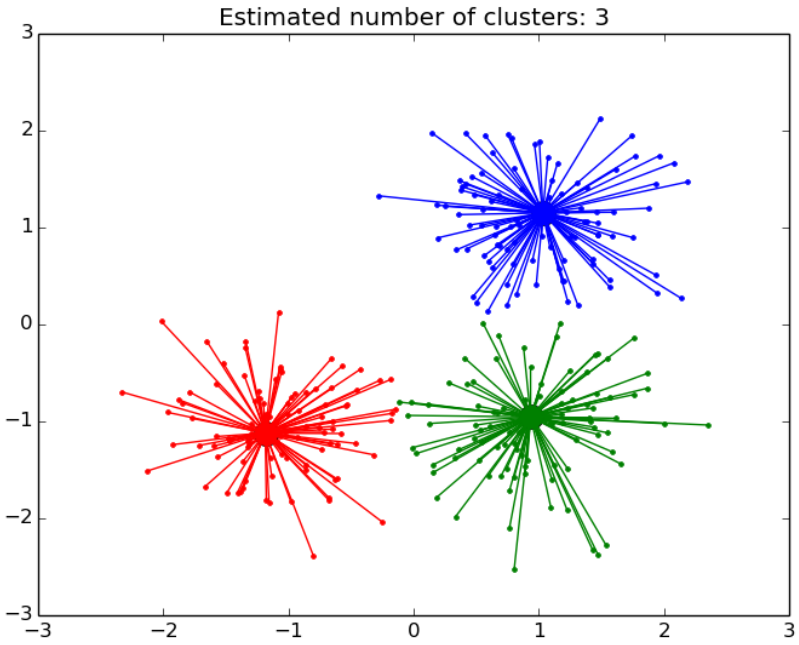


Figura 2: Exemplo de cluster com técnica de *Affinity Propagation* para agrupar os pontos.
Fonte: (LEARN, 2015)

2.3 VISUALIZAÇÃO

As partes fundamentais para a visualização das redes de relações estão na estruturação dos dados. Desse modo esta seção mostrará bases para esse tópico.

2.3.1 GRAFOS

A teoria dos grafos é usada na matemática e computação. Os grafos são considerados estruturas usadas para relacionar objetos assim ele pode ser composto de nós, vértices ou pontos e ligados por linhas, arestas ou setas (WEST et al., 2001).

Os grafos podem ser analisados através de métodos de medida e caracterização sendo algum deles: clusterização e seus coeficientes, distância média, entropia e graus de distribuição, diâmetro e caminho mais curto (BOCCALETTI et al., 2006).

2.3.2 BANCO DE DADOS DE GRAFOS

Os bancos de dados em grafos, como o Neo4j, modelam seus dados usando nós e ligações, onde os nós são as tuplas e suas ligações são as relações existentes entre os dados. Grandes

sistemas usam esse tipo de abordagem para armazenar seus dados, sendo alguns deles o Twitter e Facebook onde os usuários são tratados como os vértices do grafo e as relações entre outros usuários são as arestas desse grafo (ROBINSON et al., 2013). Assim complexas relações entre os usuários são abstraídas. Vários projetos usam essa abordagem para gerar redes de conhecimento (Figura 3) como por exemplo o DBpedia (AUER et al., 2007), PageRank da Google (PAGE et al., 1999) e o Tinkerpop 5 (PENTEADO et al., 2014)

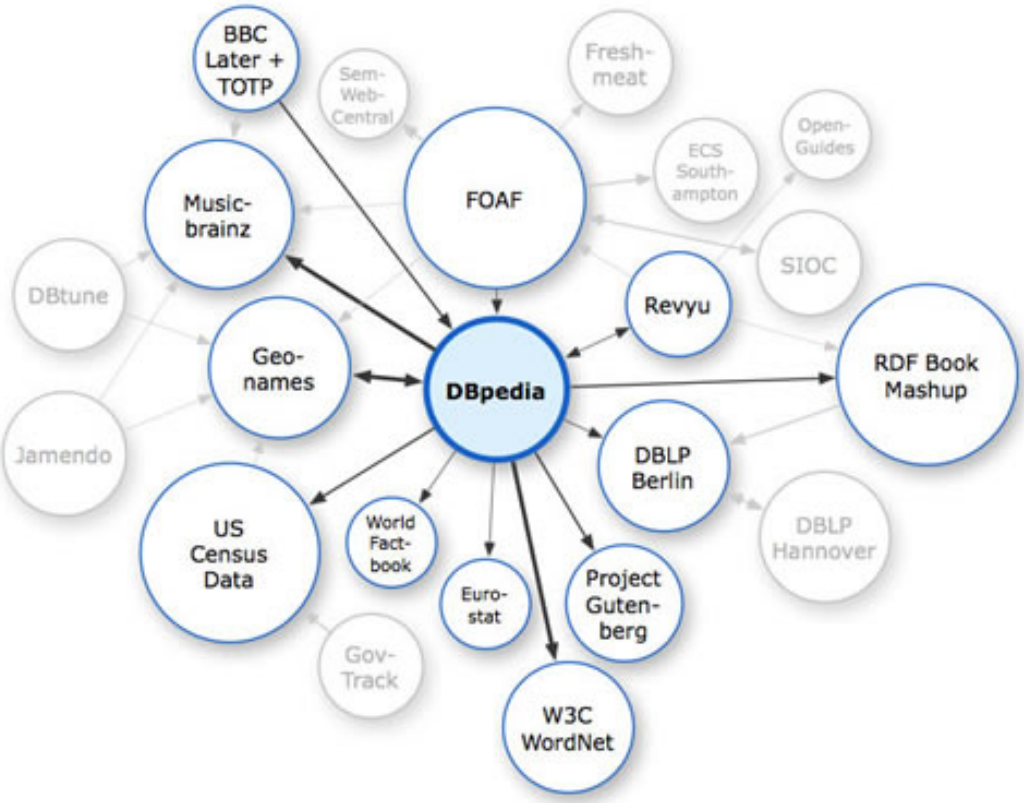


Figura 3: Exemplo de um grafo gerado por informações coletadas do WikiPedia.
Fonte: (SMITH, 2011)

2.3.3 CYPHER

O Cypher (NEO4J, 2016) é uma linguagem delcarativa inspirada em SQL. Ela é usada para descrever padrões em grafos e tem a peculiaridade de fazer o arranjo da sintaxe ser parecido com a forma visual das relações buscadas, chamado de ASCII-Art. A linguagem permite selecionar, inserir, atualizar e excluir dados do grafo.

Um grafo possuí nós e eles são representados na liguagem Cypher por por dois parenteses opostos, ”()”, isso dá a sensação do nó ser um circulo. Esse nó pode ter um nome caso seja

necessário usá-lo em futura referência, nesse caso coloca-se o nome do nó dentro dos parenteses, ”(nome)”.

As relações são representadas por, ”–”, quando não existe relação de origem e destino. Quando se existe relação de origem e destino o simbolo, ”–>”, é usado quando o nó origem está a esquerda. O símbolo, ”<–”, é usado quando a origem vem do nó a direita. Um exemplo de relação com nós pode ser representada desse modo: (origem)–>(destino), onde o nó, ”(origem)”, é o nó de origem e o nó, ”(destino)” é o nó de destino. Depois de executado um comando o banco de dados irá retornar todos os subgrafos que estiverem de acordo com a consulta.

2.3.4 *DATA-DRIVEN DOCUMENTS*

O (D3.JS, 2016) *Data-Driven Documents*, chamada de D3.js, é uma biblioteca em JavaScript para manipulação de documentos que contenham dados. A D3.js ajuda na representação dos dados de maneira visual e usa HTML, SVG e CSS para realizar seus resultados. Ela foi arquitetada para executar nos navegadores mais modernos proporcionando total liberdade para a manipulação dos gráficos gerados. Essa biblioteca pode tornar um site mais dinâmico como por exemplo: gerar tabelas, figuras, gráficos de barra e grafos interativos.

2.4 TRABALHOS RELACIONADOS

A seguir serão apresentados alguns trabalhos que foram considerados no desenvolvimento dos métodos desse projeto.

2.4.1 DBPEDIA

O DBpedia é um projeto criado com o objetivo de unir as informações presentes na WikiPedia em um grafo e disponibilizá-lo na internet (Figura 3). Dessa forma é possível extrair várias relações sobre assuntos diversos. Além disso ele possui acesso a outros dados da web tornando sua rede mais completa. A missão do DBpedia é facilitar o acesso aos dados suas interconexões visando o melhoramento da experiência de pesquisa dos dados (DBPEDIA, 2015).

2.4.2 BIO4J

Bio4j é um sistema de bioinformática que representa seus dados em forma de grafo. Ele é responsável por unir os dados disponíveis no Uniprot KB, Gene Ontology e UniRef, NCBI Taxonomy. Além disso ele possui suporte para gerenciar proteínas.

Sua principal característica é o uso da estrutura de dados baseada em redes. Essa estrutura armazena os dados representando suas próprias características nas estruturas do grafo. Esse tipo de prática difere dos modelos relacionais onde, obrigatoriamente, os dados devem ser estruturados em tabelas e relacionados por identificadores, o que pode a resultar em dificuldades para obter-se relações mais pontuais ou complexas (PAREJA, 2015).

2.5 SELEÇÃO DE DADOS

Alguns dados foram adotados e usados nos processos desse projeto a partir de bancos de dados de acesso público. Portanto, essa seção se dedica em apresentar resumidamente as fontes consideradas relevantes utilizadas nesse trabalho.

2.5.1 *NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION*

O *National Center for Biotechnology Information*, NCBI, é uma instituição dos Estados Unidos e tem como objetivo centralizar vários tipos de informação sobre biologia. (NCBI, 2015)

Uma de suas missões consiste em desenvolver tecnologias da informação que visam entender o funcionamento genético a fim de prevenir e combater doenças. Para isso foram criados sistemas automáticos que ajudam a comunidade médica e de biotecnologia a obter tais dados.

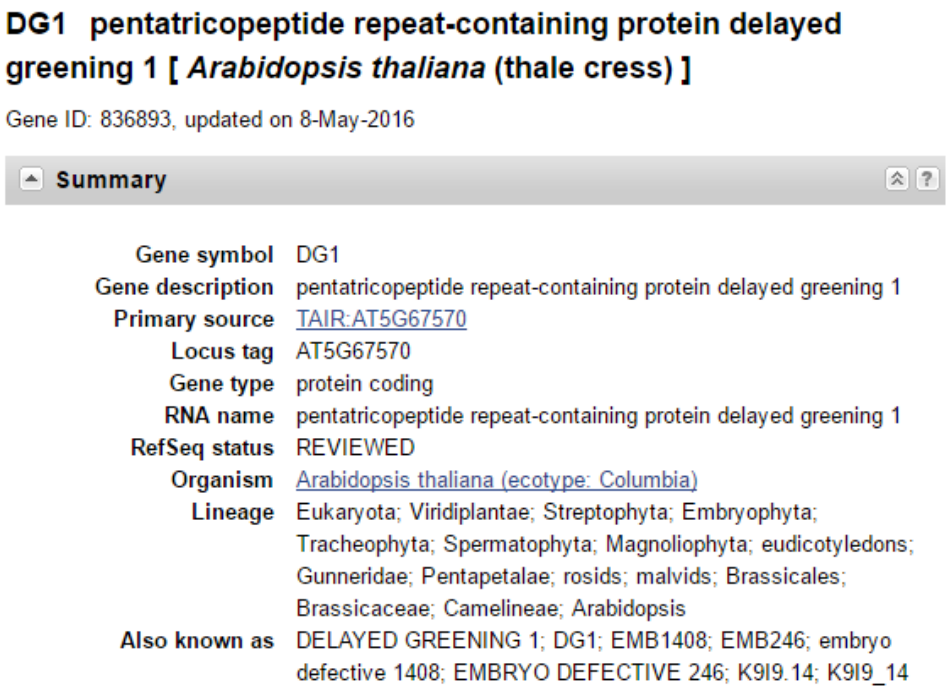
Os seus dados podem ser acessados por sua plataforma online como exemplificado na Figura 4. Dentre esses dados estão sua identificação, tipo de gene, sinônimos do seu nome, nome do RNA dentre outros.

2.5.2 *GENE ONTOLOGY*

O Gene Ontology, GO, mantém e desenvolve toda a linguagem e vocabulário usado para representar os genes e seus produtos e disponibiliza ferramentas que ajudam a manipular todo os dados armazenados por ele caracterizando assim uma ontologia.

De maneira geral o termo ontologia é a representação de algum conhecimento com todos os elementos que os compões e como estão relacionados. Por isso a maioria das coisas que pode ser observada tem sua ontologia. Esse termo é bastante usado no campo da bioinformática (GO, 2015)

Os seus dados podem ser acessados por sua plataforma online como exemplificado na



DG1 pentatricopeptide repeat-containing protein delayed greening 1 [*Arabidopsis thaliana* (thale cress)]

Gene ID: 836893, updated on 8-May-2016

Summary

Gene symbol	DG1
Gene description	pentatricopeptide repeat-containing protein delayed greening 1
Primary source	TAIR:AT5G67570
Locus tag	AT5G67570
Gene type	protein coding
RNA name	pentatricopeptide repeat-containing protein delayed greening 1
RefSeq status	REVIEWED
Organism	Arabidopsis thaliana (ecotype: Columbia)
Lineage	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis
Also known as	DELAYED GREENING 1; DG1; EMB1408; EMB246; embryo defective 1408; EMBRYO DEFECTIVE 246; K9I9.14; K9I9_14

Figura 4: Exemplo de uma tela com as informações de um gene no NCBI

Fonte: (NCBI, 2015)

Figura 5. Dentre esses dados estão sua identificação, tipo de característica, sinônimos do seu nome, definições sobre sua atuação na planta dentre outros.

Outro ponto importante a se destacar sobre os dados do GO é que eles são estruturados de forma hierárquica, como pode ser analisado um exemplo na Figura 6 no qual a ontologia GO:0003700, fator de transcrição, faz parte de vários outros grupos. Dessa forma existem ontologias mais específicas e outras mais genéricas.

2.5.3 *THE ARABIDOPSIS INFORMATION RESOURCE*

O *The Arabidopsis Information Resource*, TAIR (TAIR, 2016), é um banco de dados que mantem informações sobre biologia molecular da *Arabidopsis thaliana*. Os dados disponíveis no TAIR incluem a sequência genômica completa juto com sua estrutura, produto e expressão de genes, mapas genômicos. Os dados das funções dos genes são atualizados semanalmente, com as últimas pesquisas realizadas, através das submissões feitas pela comunidade.

O centro de recursos biológicos da *Arabidopsis thaliana* na universidade de Ohio que coleta, reproduz e preserva sementes e DNA da *Arabidopsis thaliana* está totalmente integrado com a plataforma do TAIR, tornado-a sempre atualizada e confiável.

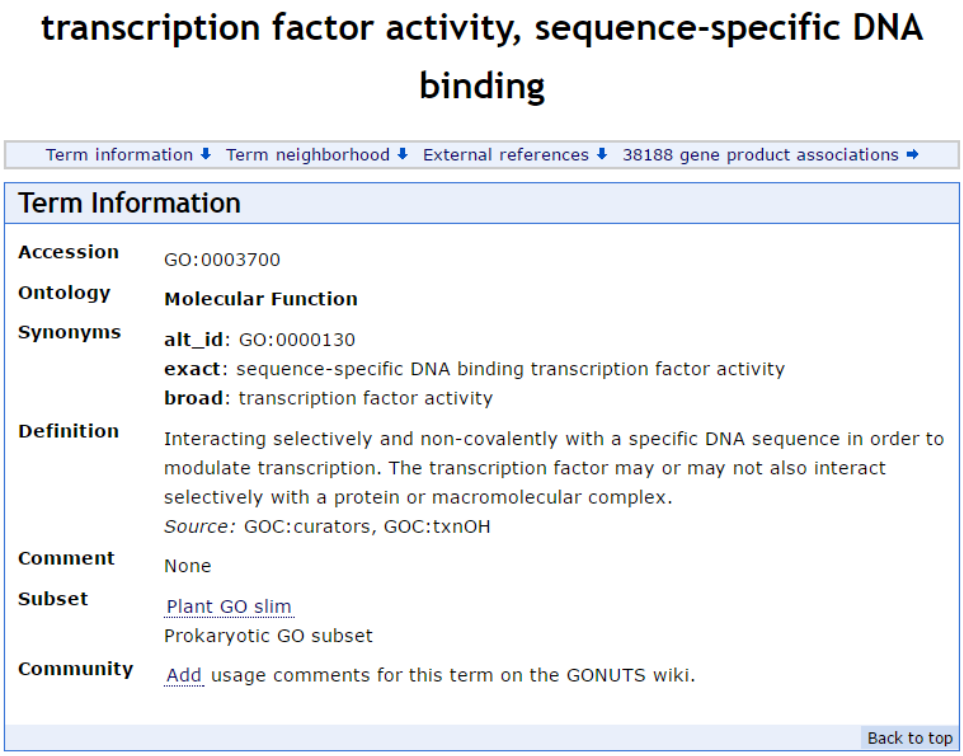


Figura 5: Exemplo de uma tela com as informações de uma ontologia no GO

Fonte: (GO, 2015)

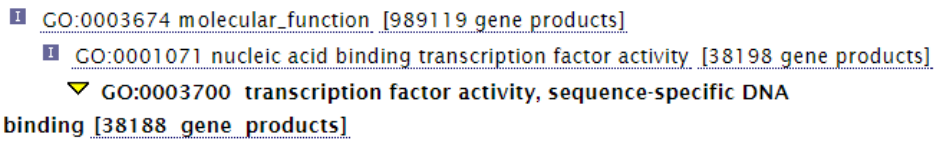


Figura 6: Exemplo de uma árvore hierárquica de uma ontologia do GO

Fonte: (GO, 2015)

Os seus dados podem ser acessados por sua plataforma online como exemplificado na Figura 7. Dentre esses dados estão sua identificação, tipo de gene, sinônimos do seu nome, definições sobre sua atuação na planta e informações sobre suas características.

2.5.4 *PLANT ONTOLOGY*

O *Plant Ontology*, PO (PO, 2016), controla o vocabulário que descreve a anatomia, morfologia e estágios do desenvolvimento de todas as plantas. O objetivo do PO é criar uma *framework* para consultas entre diversas espécies considerando a expressão genética e fenótipo. Desde o inicio de janeiro de 2011 o PO foi unido em uma única base que antes eram sobre a

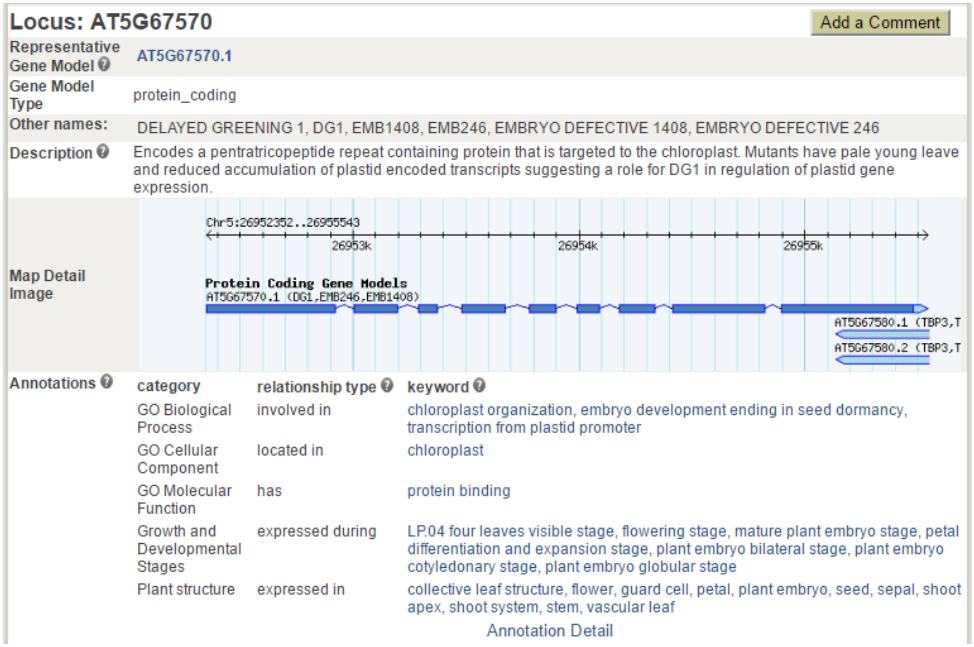


Figura 7: Exemplo de uma tela com as informações de um gene no TAIR

Fonte: (TAIR, 2016)

anatomia e outra sobre o desenvolvimento estrutural das plantas.

Os seus dados podem ser acessados por sua plataforma online como exemplificado na Figura 8. Dentre esses dados estão sua identificação, tipo de característica, sinônimos do seu nome em várias línguas, definições sobre sua atuação na planta e comentários adicionais.

Outro ponto importante a se destacar sobre os dados do PO é que eles são estruturados de forma hierárquica, como pode ser analisado um exemplo na Figura 9 no qual a raiz (*root*) faz parte de vários outros grupos.

2.6 CARACTERÍSTICAS DOS GENES

Essa seção apresenta um breve embasamento teórico para se entender como algumas características dos genes.

2.6.1 DNA

O DNA é uma estrutura orgânica formada por quatro bases nitrogenadas que sempre se organizam aos pares tais como: adenina com timina e citosina com guanina. Um exemplo desta estrutura pode ser observado na Figura 10. O DNA contém toda a informação necessária para o desenvolvimento do ser (KLUG et al., 2010).

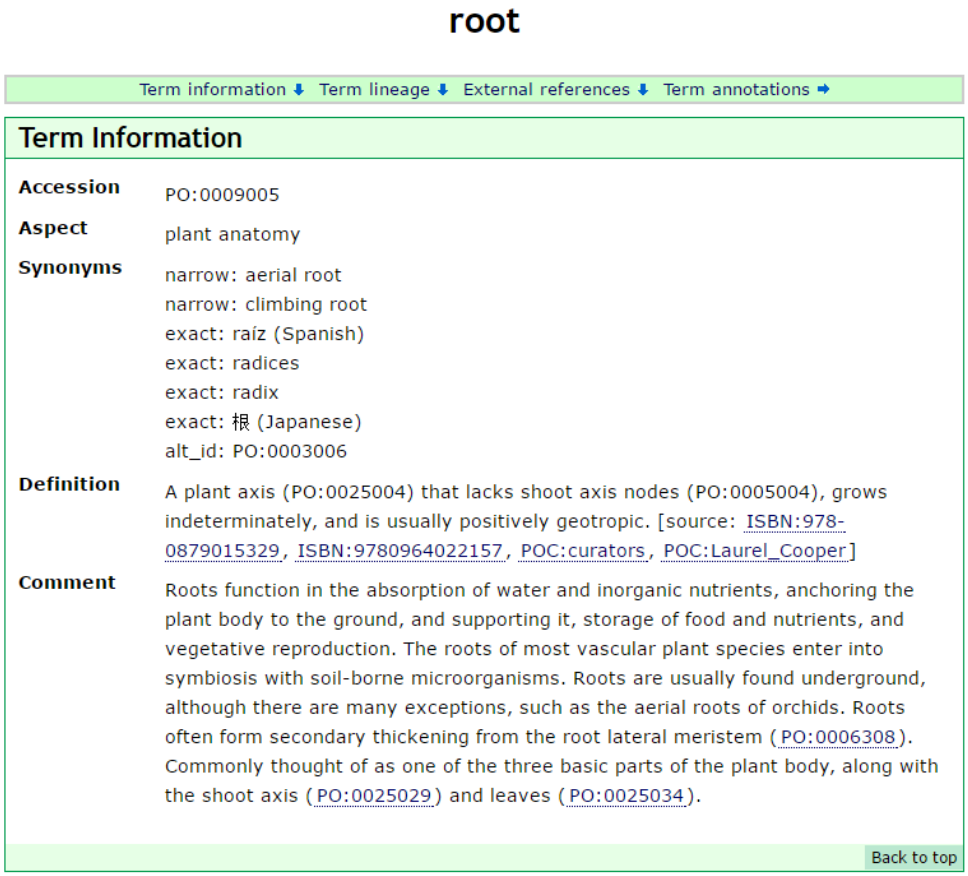


Figura 8: Exemplo de uma tela com as informações de uma ontologia no PO

Fonte: (PO, 2016)

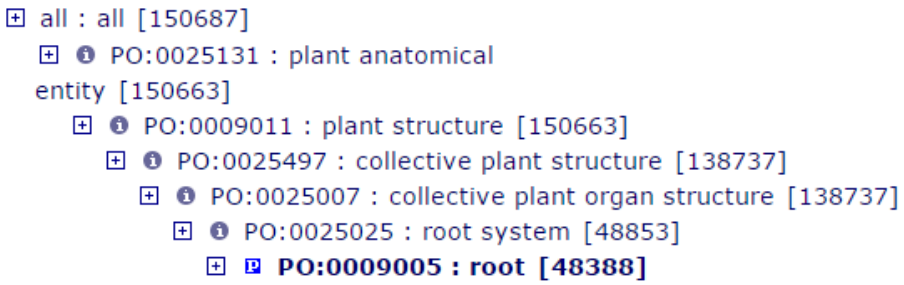


Figura 9: Exemplo de uma árvore hierárquica de uma ontologia do PO

Fonte: (PO, 2016)

2.6.2 RNA

O RNA é uma molécula proveniente do DNA gerada pelo processo de transcrição de um gene. O RNA tem papel fundamental na codificação e decodificação de genes assim como sua regulação e expressão. Ela possui quatro bases nitrogenadas que se unem da seguinte forma:

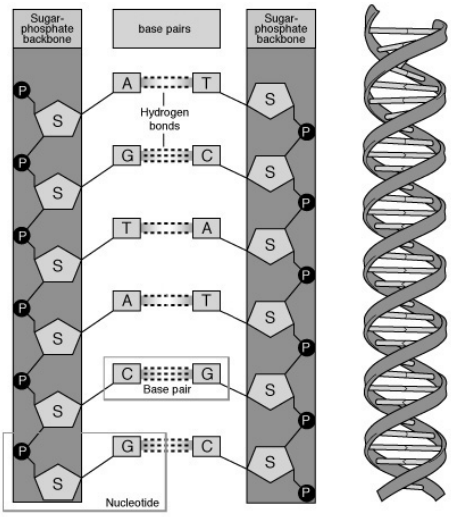


Figura 10: Exemplo de DNA.

Fonte: (NCBI, 2015)

adenina com uracila e citosina com guanina. Um exemplo de RNA pode ser visto na Figura 11 (KLUG et al., 2010).

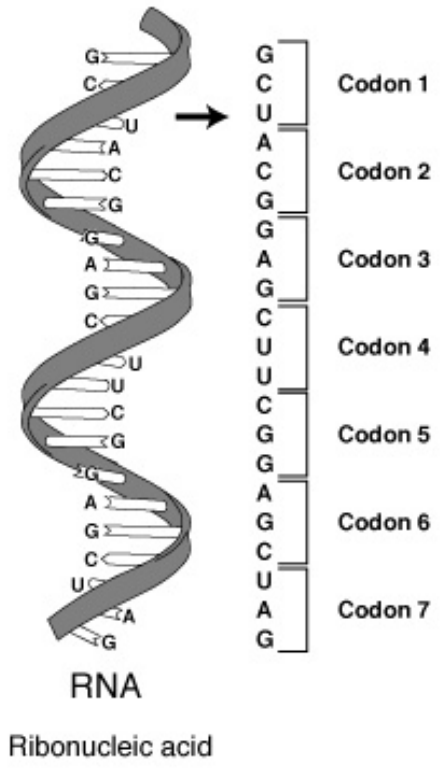


Figura 11: Exemplo de RNA.

Fonte: (NCBI, 2015)

2.6.3 PROTEÍNAS

A síntese proteica é um processo celular complexo que será brevemente explicado. Ele começa com a transcrição de um gene em mRNA o qual sai do núcleo da célula e se desloca até o ribossomo que recebe o mRNA. Através da “leitura” do mRNA o ribossomo gera aminoácidos por meio da identificação dos códons e o agrupamento dos aminoácidos geram as proteínas.(KLUG et al., 2010)

Então as proteínas são moléculas formadas por cadeias de aminoácidos e estão presente nos organismos. Elas são responsáveis pela maioria dos processos celulares sendo muitas vezes usadas como catalisadores, bases estruturais e mecânicas.

2.6.4 GENE

O gene é uma região do DNA que possuí uma função metabólica ou uma característica do ser. Uma ilustração de um gene é mostrada na Figura 12. Para realizarem suas atividades eles devem ser expressos durante uma via metabólica. A expressão de um gene pode ser vista pela produção de RNA e depois proteínas (KLUG et al., 2010).

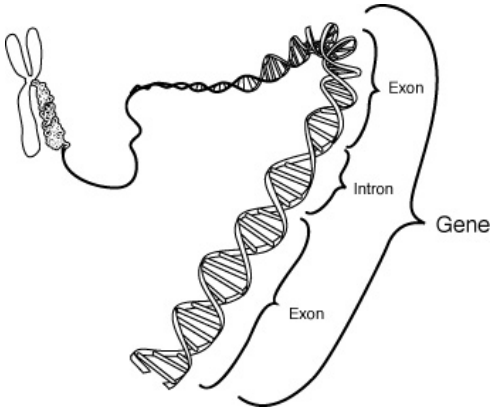


Figura 12: Exemplo de gene.
Fonte: (NCBI, 2015)

2.6.5 PROCESSO DINÂMICO

Uma via metabólica é a realização de várias reações bioquímicas que levam a criação de algum produto celular (Figura 13). Alguns exemplos desses produtos são gorduras e proteínas. Para a via metabólica funcionar de maneira correta um determinado gene pode se expressar ou não em um determinado período de tempo. A via metabólica é regida por estímulos externos,

como frio ou calor, ou internos como é o caso da presença de determinada proteína na célula (DEY; HARBORNE, 1997).

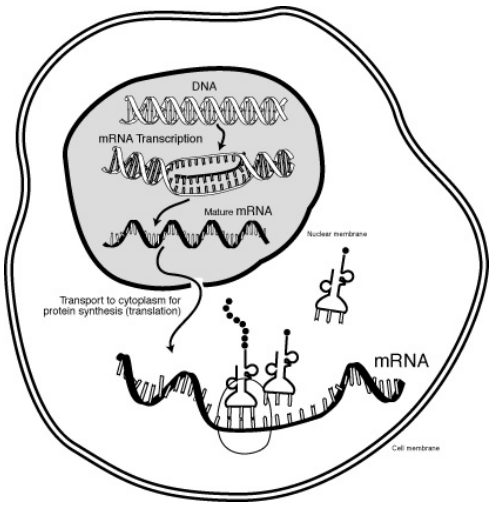


Figura 13: Exemplo de transcrição do DNA.
Fonte: (NCBI, 2015)

As vias metabólicas são a responsáveis por manter a vida em um organismo e são reguladas por redes de expressões dos genes. Essas redes se unem por sinais enviados e recebidos. Entre esses sinais estão as proteínas e RNAs (THORNALLEY, 1990).

Quando dados de expressão de genes, que podem ser capturados pela técnica de DNA *Microarrays* (Seção 2.6.6), são registrados durante um período de tempo é possível observar quais genes participaram de uma determinada via metabólica, podendo essa ser representada em um sistema dinâmico. A rede formada pela expressão de vários genes é composta por vários estágios por isso recebe o nome de sistema de regulação gênica (YOSHIDA et al., 2005).

2.6.6 *MICROARRAY*

Microarray é uma técnica usada na biologia que visa obter o nível de expressão de um gene dado uma amostra o qual pode ser visto na Figura 14. Para se medir a expressão do gene é usada uma matriz, na qual cada posição contém DNA marcados para determinadas moléculas alvo. Desse modo é possível obter resultados quantitativos para cada expressão genética de acordo com a coloração obtida em cada ponto (DUFVA, 2009).

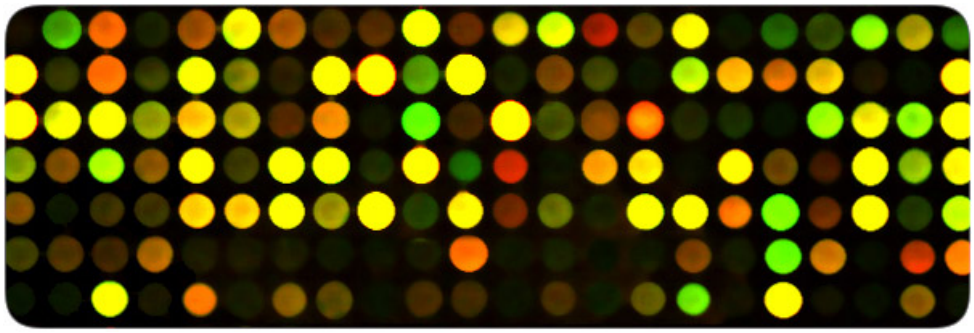


Figura 14: Exemplo de um *microarray*.
Fonte: (CSMBIO, 2015)

2.6.7 *ARABIDOPSIS THALIANA*

Arabidopsis thaliana é a espécie de uma planta da família das Brassicaceae, exibida na Figura 15, cujo genoma já foi totalmente sequenciado. Por isso esse organismo é usado como modelo para identificar relações existentes entre os genes e quais funções elas estão relacionadas. Além disso seu tamanho relativamente pequeno e seu curto ciclo de vida são vantajosos para estudos em laboratórios e para obtenção mais rápida de resultados.

Com isso a popularização do estudo da *Arabidopsis thaliana* foi grande. Com a colaboração de vários países foi criada a iniciativa genoma para *Arabidopsis thaliana* (*The Arabidopsis Genome Initiative, AGI*) que começou sequenciar o genoma desde 1996 (INITIATIVE et al., 2000b).

Figura 15: Flor da *Arabidopsis thaliana*.
Fonte: (NATURE, 2015)

3 DESENVOLVIMENTO

Este capítulo apresenta resumidamente as principais etapas adotadas para o desenvolvimento do projeto. Primeiramente é relatada as escolhas das tecnologias e ferramentas para implementar o trabalho. Posteriormente o foco será dado aos métodos usados para a solução do problema. Depois são apresentados os resultados. Para comprovar a eficácia dos métodos, meios para a validação dos resultados foram adotados e são apresentados.

3.1 TECNOLOGIAS E FERRAMENTAS

Nessa seção são resumidamente descritas algumas ferramentas, tecnologias e como elas contribuíram para a resolução dos problemas expostos na Seção 1.1.

3.1.1 VISUALIZAÇÃO E ARMAZENAMENTO

Para alcançar alguns dos objetivos descritos na Seção 1.3, foi realizada uma pesquisa com o banco de dados Neo4j que é considerado um dos mais utilizados para armazenar e gerenciar grafos (PENTEADO et al., 2014). Essa pesquisa visou identificar o funcionamento desse banco de dados em grafos e como é possível interagir com sua linguagem, gerar consultas e resultados.

A linguagem usada para fazer consultas no Neo4J é a Cypher, explicado na Seção 2.3.3, que tem sua arquitetura voltada para grafos. Uma maneira de interagir com esse banco fazendo consultas em Cypher é através de programas feitos em Java, pois o Neo4J disponibiliza várias bibliotecas com esse objetivo.

Um modo de visualizar os grafos resultantes é através da biblioteca d3.js, descrito na Seção 2.3.4, feita para JavaScript com o objetivo de criar elementos visuais interativos.

3.2 MÉTODOS

O projeto desenvolvido é apresentado em três partes principais como exibido na Figura 16. A primeira etapa é referente a captura dos dados das bases online (TAIR, GO e PO), a segunda parte mostrará como foi feito o processamento dos dados para armazená-lo em forma de grafo e a terceira parte terá como foco a visualização e consulta dos grafos persistidos no banco.

Todos esses métodos em conjunto formam o programa Visual Ontograror, que é um sistema web no qual é possível fazer consultas sobre os genes de *Arabidopsis thaliana* junto com suas relações e características.

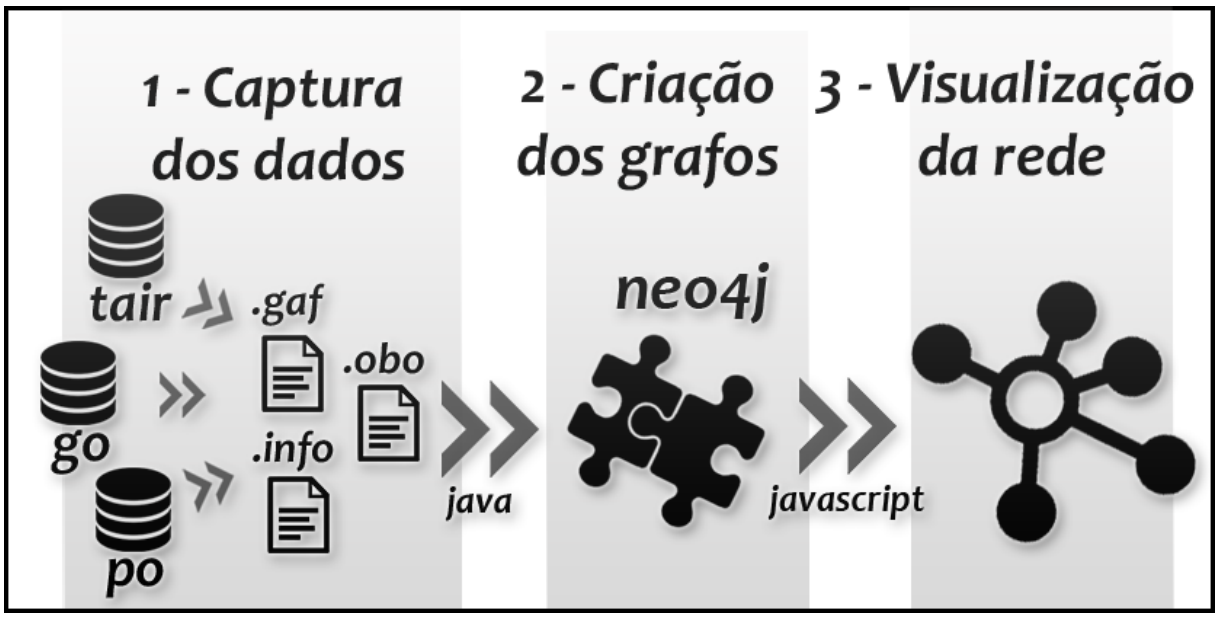


Figura 16: Etapas principais do projeto desenvolvido.

3.2.1 CAPTURA DOS DADOS E PRÉ-PROCESSAMENTO - PRIMEIRA ETAPA

Os bancos usados nesse projeto são o TAIR (TAIR, 2016), GO (GO, 2015) e o PO (PO, 2016) que podem ser acessados através de plataformas online. Seus dados podem ser baixados em forma de arquivos. Esses arquivos possuem três formatos sendo eles, GAF que armazena as relações entre os genes representado na Figura 17, OBO que armazena as ontologias como mostrado na Figura 18 e o INFO que tem o foco na descrição dos genes como pode ser visto na Figura 19.

Depois dos arquivos estarem armazenados eles foram carregados em um programa em Java que os processou de modo a organizá-los criando as relações em um banco de da-

dos MySQL. Os dados foram armazenados no banco de dados MySQL para remover possíveis redundâncias presentes nos arquivos e para facilitar futuras consultas visto que os arquivos estavam estruturados em forma de tabelas.

O banco de dados em MySQL resultou em 3 tabelas. A primeira é sobre os genes, ela contém colunas com a identificação usada no TAIR e no NCBI, o tipo de gene e em qual cromossomo o gene está inserido. A segunda tabela contém as ontologias, nelas se encontram as informações dos banco GO e PO e suas colunas armazenam sua identificação, categoria, definição, palavras chave referentes a ontologia dentre outros dados. A última tabela é responsável por criar as associações entre os genes e as ontologias.

3.2.2 CRIAÇÃO DOS GRAFOS - SEGUNDA ETAPA

A próxima etapa foi criar um algoritmo para a gerar a estrutura do grafo, esse algoritmo utiliza as informações contidas no banco de dados MySQL, preenchidos com os arquivos GAF, OBO e INFO, e os estruturava em forma de grafo utilizando o banco de dados Neo4J e a linguagem Cypher.

O grafo criado possui dois tipos de nós, um deles é o tipo Gene que armazena a descrição do gene, sua identificação de locus em qual cromossomo ele está presente dentre outras. O segundo tipo de nó é referente às características dos genes, esses nós possuem dados sobre a categoria, palavras chaves e as identificações nos bancos de dados GO e PO.

Todas as relações entre os nós são feitas entre um nó do tipo gene e um nó do tipo característica, por isso não é possível um nó do mesmo tipo estar relacionado diretamente um com o outro.

```
taxon:3702  20150606    GOC  info   TAIR:locus:505006114
TAIR    locus:2155593   ICU2     info   GO:0000731
TAIR:Communication:501741973    IBA
PANTHER:PTHR10322_AN3   P   AT5G67100
AT5G67100|ICU2|INCURVATA2|K21H1.14|K21H1_14 protein
taxon:3702  20110729    RefGenome    info
TAIR:locus:2155593
TAIR    locus:2153629   NEDD1    info   GO:0000777
```

Figura 17: Exemplo de parte de um arquivo tipo gaf.

```
[Term]
id: GO:0000076
name: DNA replication checkpoint
namespace: biological_process
def: "A cell cycle checkpoint that prevents the
initiation of nuclear division until DNA replication
is complete, thereby ensuring that progeny inherit a
full complement of the genome." [GOC:curators, GOC:rn,
PMID:11728327, PMID:12537518]
is_a: GO:0031570 ! DNA integrity checkpoint
```

Figura 18: Exemplo de parte de um arquivo tipo obo.

```
3702    836845  ICU2    AT5G67100
INCURVATA2|K21H1.14|K21H1_14    TAIR:AT5G67100  5   -
DNA polymerase alpha catalytic subunit  protein-coding
-   -   -   -   20160103
3702    836846  ALC AT5G67110
```

Figura 19: Exemplo de parte de um arquivo tipo info.

3.2.3 VISUALIZAÇÃO E CONSULTA - TERCEIRA ETAPA

Após todo o processamento descrito nas subseções anteriores foi criado um grafo resultante com 33583 genes. Para acessar esses dados foi criada uma interface na qual o usuário pode interagir com a rede gerada mudando a posição e informações dos nós, a Figura 24 mostra um exemplo de grafo gerado a partir da base criada.

A interface foi feita usando a linguagem Javascript e a biblioteca d3.js especializada em visualização de grafos. Dessa forma o processamento gráfico da rede consultada será feita no computador do cliente.

3.3 VALIDAÇÃO DO MÉTODOS

Para a validação dos dados foi usada a lista de genes disponibilizada no trabalho (WANG et al., 2009) que pode ser baixada no link https://github.com/gabrielrubinobr/VisualOntogrator. Nesse trabalho é afirmado que esse conjunto de genes estão muito conectados em relação a suas características. Portanto é esperado que os métodos do Visual Ontogrator gere uma rede coerente com o descrito. Desse modo será possível verificar se o banco de dados está consistente em conjunto com a interface criada.

No trabalho usado como base (WANG et al., 2009) o autor mostra uma rede de co-

expressão, representada na Figura 20 em conjunto com seu estudo de característica dos genes ilustrado na Figura 21. Portanto os genes usados para a validação dos métodos do Visual Ontogrator terão base biológica.

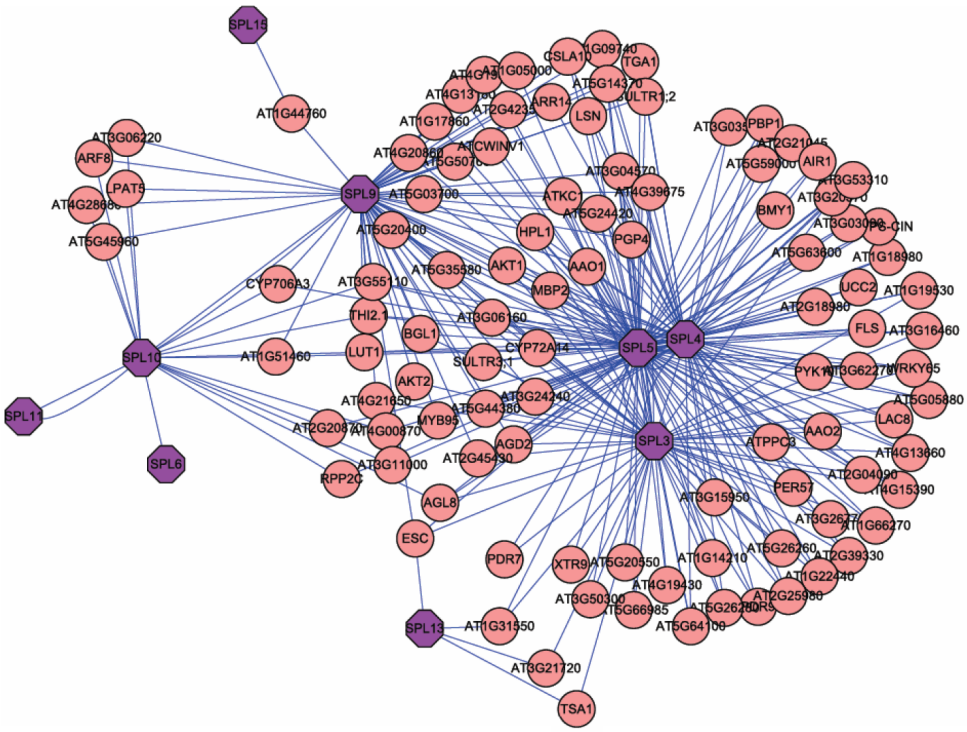


Figura 20: Rede de genes correlacionados.

Fonte: (WANG et al., 2009)

3.3.1 LISTA DE FATORES DE TRANSCRIÇÃO

Na Figura 20 o autor do trabalho (WANG et al., 2009) apresenta uma rede de co-expressão com 112 genes em conjunto com um gráfico, Figura 21 mostrando a ocorrência das características dos genes. Como se trata de uma rede de expressão de proteínas uma das características mais frequente é o fator de transcrição.

Portanto para o teste de validação foi filtrada a lista dos genes que estavam relacionados com o fator de transcrição, representado pelo nome GO:0003700. No total a lista resultante continha 20 genes. Esses dados foram usados como entrada no programa criado (Visual Ontogrator) e foi obtido como resposta o seguinte grafo representado na Figura 22

Analisando o grafo pode-se concluir que a base de dados criada está coerente com os

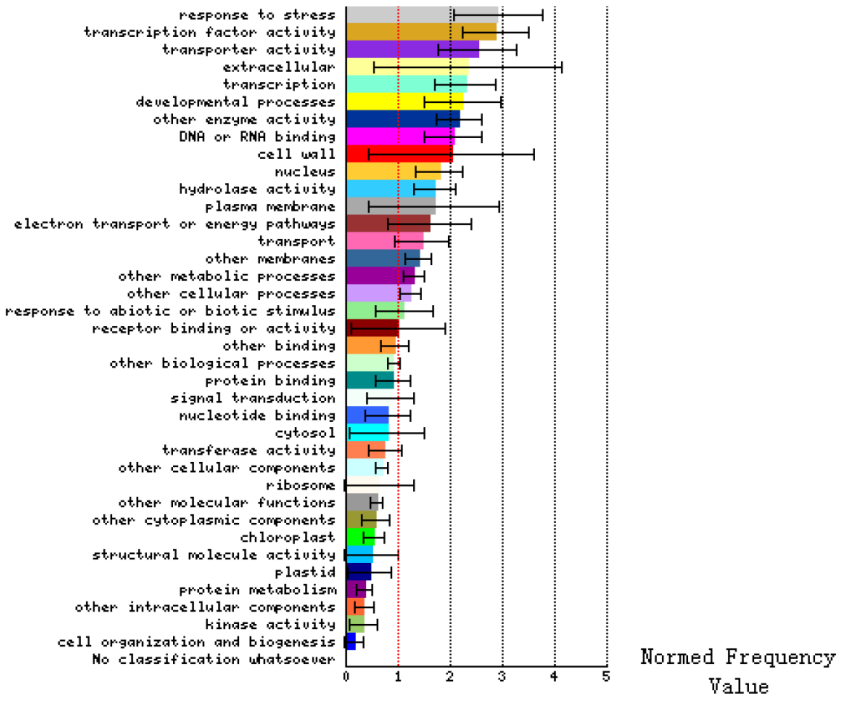


Figura 21: Características dos genes usados na rede de expressão.

Fonte: (WANG et al., 2009)

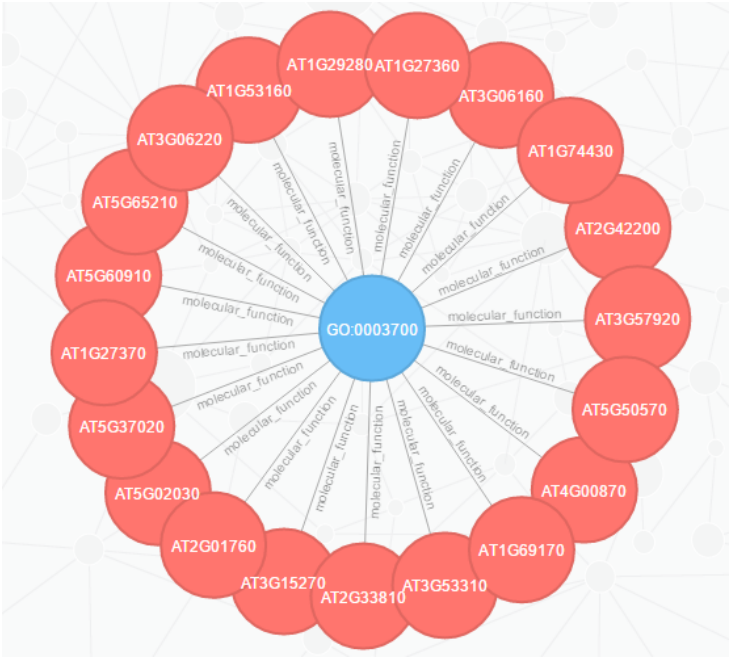


Figura 22: Grafo resultante da lista de fatores de transcrição

resultados do trabalho (WANG et al., 2009).

3.3.2 LISTA DO NÚCLEO

Outra lista gerada, a partir dos genes da rede de co-expressão da Figura 20, foi a dos genes presentes no núcleo da célula da *Arabidopsis thaliana*, esta lista contém 17 genes. Esses dados foram processados e geraram o grafo mostrado na Figura 23.

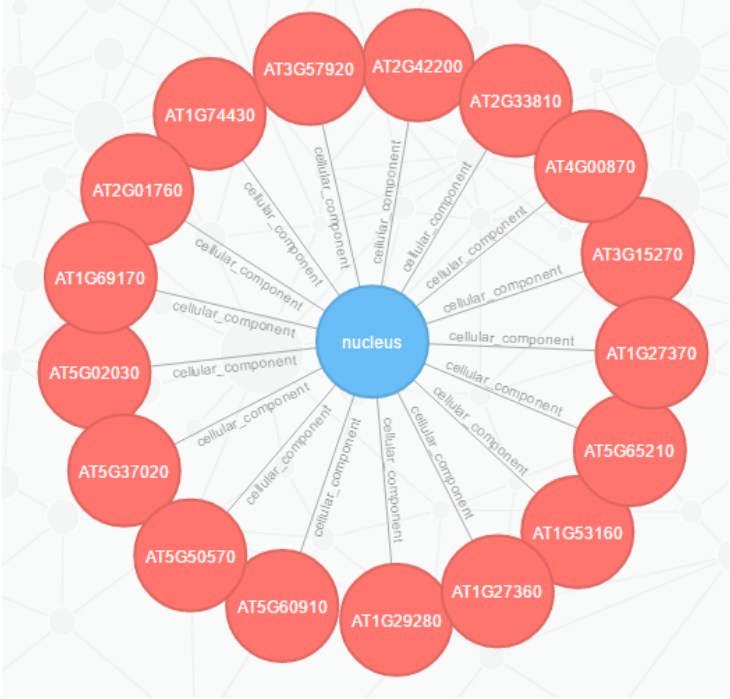


Figura 23: Grafo resultante da lista do núcleo.

Analisando o grafo pode-se concluir que a base de dados criada está coerente com os resultados do trabalho (WANG et al., 2009).

3.4 GRAFOS ADICIONAIS - *PLANT ONTOLOGY*

O trabalho (WANG et al., 2009), usado para a validação dos métodos, utiliza somente a base de dado GO. O Visual Ontogrator, por outro lado utiliza além da base GO a base PO.

Portanto para ilustrar as características originadas do PO foi usada a lista da Seção 3.3. Em conjunto com essa lista foram adicionadas as características da flor (*flower*), raiz (*root*) e do fator de transcrição (identificado por GO:0003677). Como resultado obteve-se o grafo representado na Figura 24.

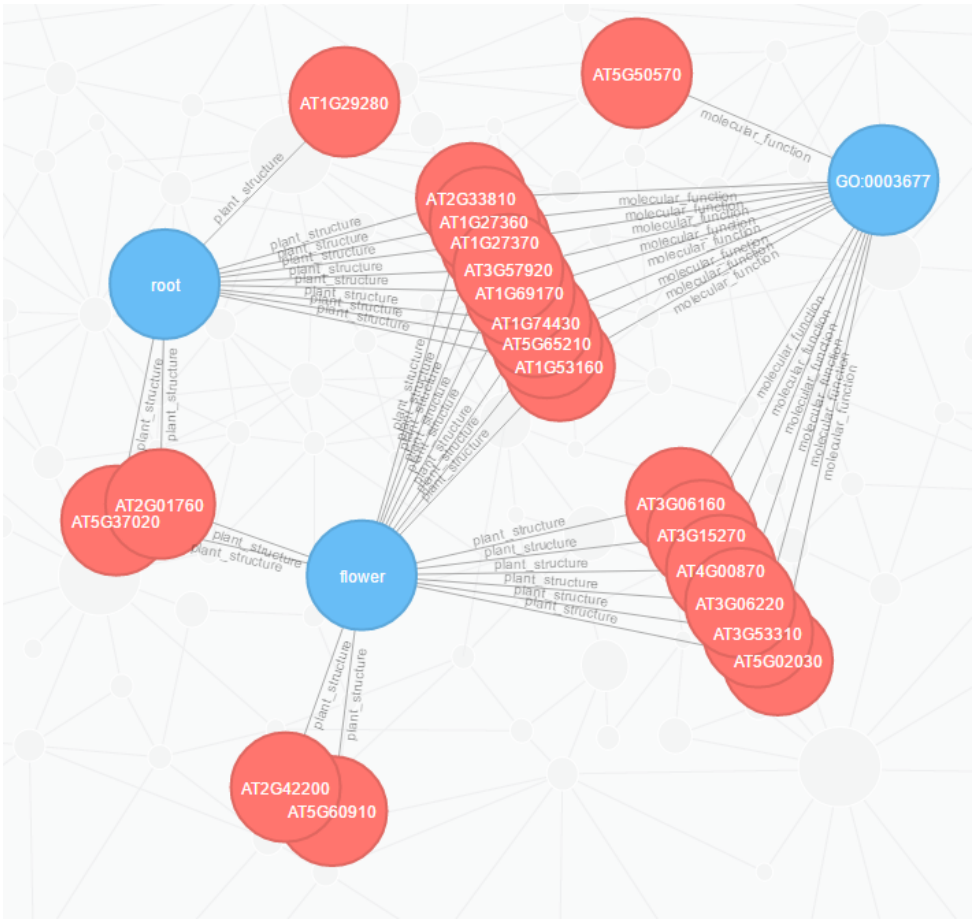


Figura 24: Grafo ilustrando características do *Plant Ontology*.

4 RESULTADOS OBTIDOS

Este capítulo apresenta os resultados desse projeto, com alguns exemplos e descrições do sistema final.

4.1 PROGRAMAS

Como resultado do projeto foram criados 2 programas, o primeiro deles é usado no processamento dos arquivos das bases de dados online e o segundo na consulta e visualização do grafo.

4.1.1 PROCESSAMENTO DE ARQUIVOS

Um dos resultados desse trabalho foi a criação de um programa responsável pelo processamento dos arquivos das bases de dados online (TAIR, GO e PO) e seu armazenamento no banco de dados MySQL, mais detalhes sobre seu uso estão descritos na Seção 3.2.1. Esse programa pode ser baixado no link https://github.com/gabrielrubinobr/VisualOntogrator.

4.1.2 VISUAL ONTOGRATOR

A maior contribuição desse projeto foi a criação do programa Visual Ontogrator. Esse programa é um sistema web que tem como objetivo integrar dados de genes de *Arabidopsis thaliana* e os exibir em forma de grafo.

Na sua página inicial, como mostrado na Figura 25, o usuário pode colocar 3 listas principais. A primeira é a lista “Gene” onde podem ser colocados os genes, a segunda a lista “Relação” contendo as relações e a terceira, lista “Característica”, serão os nós referentes as características dos genes.

Depois de clicar no botão “Gerar Grafo”, apresentado na Figura 25, o usuário é redirecionado para a página onde será mostrado o grafo resultante. Um exemplo de pesquisa pode

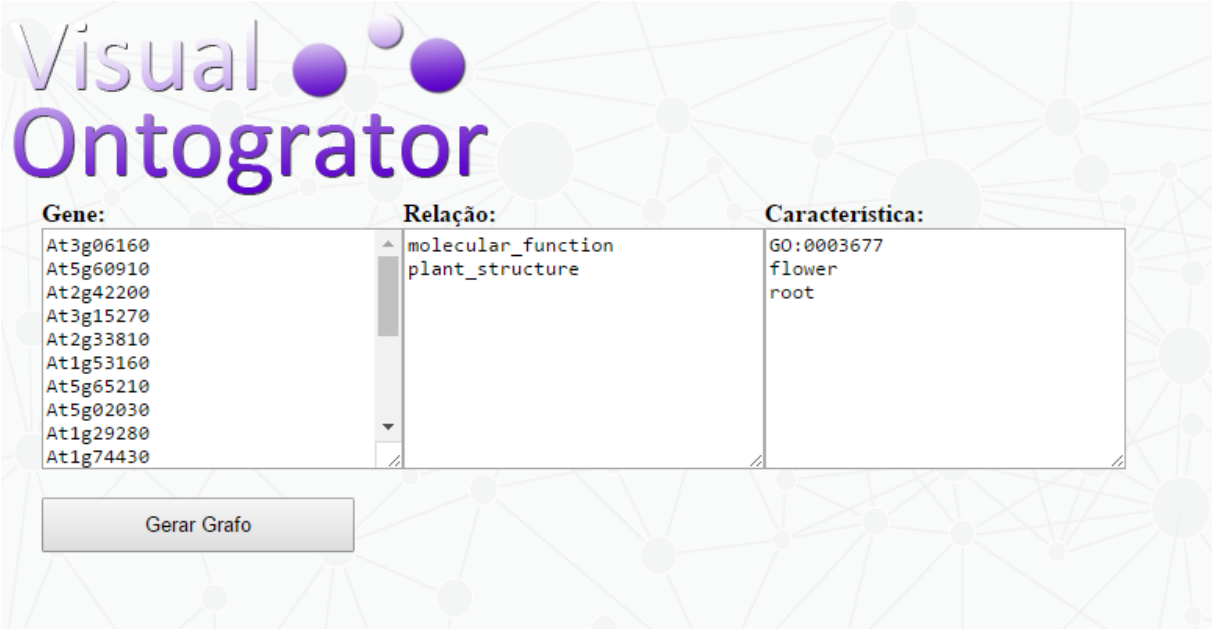


Figura 25: Tela inicial do programa Visual Ontograter.

ser visto na Figura 26

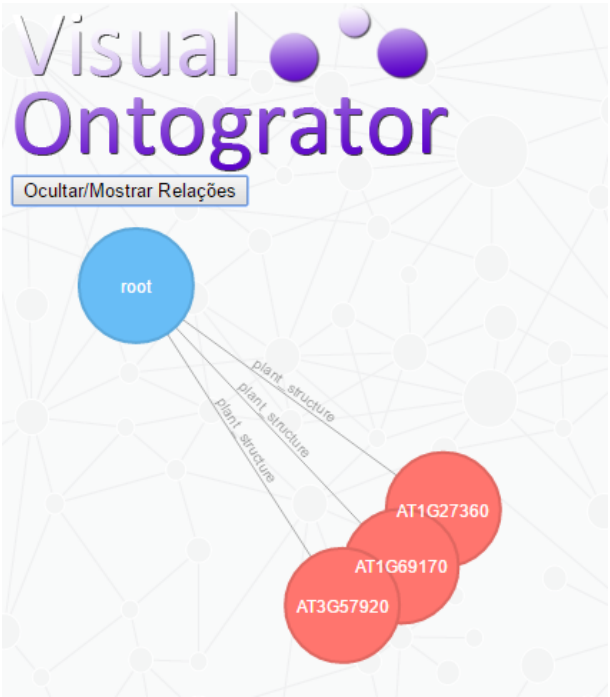


Figura 26: Tela da visualização da rede do programa Visual Ontograter.

Na página de visualização existe a opção para remover ou mostrar os nomes presentes nas relações, para isso basta clicar no botão ”Ocultar/Mostrar Relações”, o resultado desta ação pode ser vista na Figura 27

O programa Visual Ontograter, juntamente com sua base de dados para o Neo4J podem

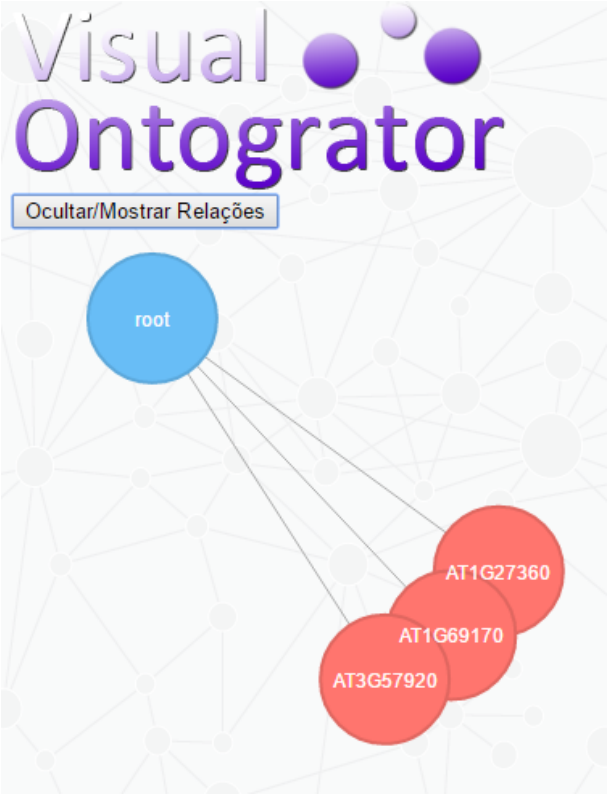


Figura 27: Visualização da rede sem os nomes das relações.

ser baixados no link https://github.com/gabrielrubinobr/VisualOntograter.

5 CONSIDERAÇÕES FINAIS

Neste capítulo são apresentadas quais as limitações que o projeto apresenta para que o escopo do problema fique bem definido permitindo o maior entendimento das tarefas realizadas.

5.1 EXTRAÇÃO DE INFORMAÇÕES DE GRAFOS

Depois da estruturação e armazenamento da rede em forma de grafo é possível medir e analisar suas informações. Por isso métodos de medida de redes complexas poderão ser adicionados em trabalhos futuros a fim de se obter novas informações sobre a rede tais como: clusterização e seus coeficientes, distância média, entropia e graus de distribuição, diâmetro e caminho mais curto.

5.2 LIMITAÇÕES

Este trabalho se dispõe a integrar dados biológicos e gerar sua visualização através do desenvolvimento de metodologias adotadas. Observa-se que organismos possuem muitas peculiaridades tornando extremamente complexa a generalização de métodos eficazes para todos os casos. Nesse sentido o presente trabalho é focado em métodos de integração para um único organismo, a *Arabdopsis thaliana* (INITIATIVE et al., 2000b). Dessa forma a validação dos métodos foi mais eficiente. Assim trabalhos futuros podem tentar generalizar essas metodologias a fim de abranger mais organismos.

Outra ponto a se destacar são os dados que são integrados as redes. Esses dados serão restritos às características dos genes. Futuramente outros tipos dados poderão ser acoplados visando maior redução na dimensionalidade do sistema (BISHOP, 1995) e maior precisão na inferência das redes.

REFERÊNCIAS

ALON, U. **An introduction to systems biology: design principles of biological circuits**. [S.l.]: CRC press, 2006.

AUER, S. et al. **Dbpedia: A nucleus for a web of open data**. [S.l.]: Springer, 2007.

BISHOP, C. M. **Neural networks for pattern recognition**. [S.l.]: Oxford university press, 1995.

BOCCALETTI, S. et al. Complex networks: Structure and dynamics. **Physics reports**, Elsevier, v. 424, n. 4, p. 175–308, 2006.

CSMBIO. **MicroArray**. 2015. Disponível em: <http://pt.dbpedia.org/>. Acesso em: 10 de setembro de 2015.

D3.JS. **Data-Driven Documents**. 2016. Disponível em: <https://d3js.org/>. Acesso em: 24 de maio de 2016.

DBPEDIA. **DBpedia**. 2015. Disponível em: <http://pt.dbpedia.org/>. Acesso em: 10 de setembro de 2015.

DEY, P. M.; HARBORNE, J. B. **Plant biochemistry**. [S.l.]: Academic Press, 1997.

DUFVA, M. Introduction to microarray technology. In: **DNA Microarrays for Biomedical Research**. [S.l.]: Springer, 2009. p. 1–22.

GO. **Gene Ontology**. 2015. Disponível em: <http://geneontology.org/>. Acesso em: 10 de setembro de 2015.

INDYK, P.; MOTWANI, R. Approximate nearest neighbors: towards removing the curse of dimensionality. In: ACM. **Proceedings of the thirtieth annual ACM symposium on Theory of computing**. [S.l.], 1998. p. 604–613.

INITIATIVE, A. G. et al. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. **nature**, v. 408, n. 6814, p. 796, 2000.

INITIATIVE, A. G. et al. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. **nature**, v. 408, n. 6814, p. 796, 2000.

KELEMEN, A.; ABRAHAM, A.; CHEN, Y. **Computational intelligence in bioinformatics**. [S.l.]: Springer, 2008.

KLUG, W. S. et al. **Conceitos de genética**. [S.l.]: Artmed, 2010.

LEARN, S. **Affinity Propagation**. 2015. Disponível em: <http://scikit-learn.org/stable/modules/clustering.html>. Acesso em: 12 de setembro de 2015.

LOPES, F. M. **Redes complexas de expressão gênica: síntese, identificação, análise e aplicações**. Tese (Doutorado) — Universidade de São Paulo, 2011.

MARX, V. Biology: The big challenges of big data. **Nature**, Nature Publishing Group, v. 498, n. 7453, p. 255–260, 2013.

NATURE. **ArabdopsisThaliana**. 2015. Disponível em: <http://www.nature.com/>. Acesso em: 10 de setembro de 2015.

NCBI. **National Center for Biotechnology Information**. 2015. Disponível em: <http://www.ncbi.nlm.nih.gov/>. Acesso em: 10 de setembro de 2015.

NEO4J. **Cypher, SQL-inspired language**. 2016. Disponível em: <http://neo4j.com/developer/cypher-query-language/>. Acesso em: 24 de maio de 2016.

PAGE, L. et al. The pagerank citation ranking: bringing order to the web. Stanford InfoLab, 1999.

PAREJA, P. **bio4j**. 2015. Disponível em: <http://bio4j.com/>. Acesso em: 14 de setembro de 2015.

PEARL, J. **Probabilistic reasoning in intelligent systems: networks of plausible inference**. [S.l.]: Morgan Kaufmann, 2014.

PENTEADO, R. R. et al. Um estudo sobre bancos de dados em grafos nativos. 2014.

PO. **Plant Ontology**. 2016. Disponível em: <http://www.plantontology.org/>. Acesso em: 24 de maio de 2016.

ROBINSON, I.; WEBBER, J.; EIFREM, E. **Graph databases**. [S.l.]: " O'Reilly Media, Inc.", 2013.

SMITH, C. **DBpedia**. 2011. Disponível em: <http://cs.smith.edu/>. Acesso em: 12 de setembro de 2015.

TAIR. **The Arabidopsis Information Resource**. 2016. Disponível em: <https://www.arabidopsis.org/>. Acesso em: 24 de maio de 2016.

THORNALLEY, P. J. The glyoxalase system: new developments towards functional characterization of a metabolic pathway fundamental to biological life. **Biochemical Journal**, Portland Press Ltd, v. 269, n. 1, p. 1, 1990.

WANG, Y. et al. Function annotation of an sbp-box gene in arabidopsis based on analysis of co-expression networks and promoters. **International journal of molecular sciences**, Molecular Diversity Preservation International, v. 10, n. 1, p. 116–132, 2009.

WEST, D. B. et al. **Introduction to graph theory**. [S.l.]: Prentice hall Upper Saddle River, 2001.

YOSHIDA, R.; IMOTO, S.; HIGUCHI, T. Estimating time-dependent gene networks from time series microarray data by dynamic linear models with markov switching. In: IEEE. **Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE**. [S.l.], 2005. p. 289–298.