Ministério da Educação
Universidade Tecnológica Federal do Paraná
Pró-Reitoria de Pesquisa e Pós-Graduação

Relatório Final de Atividades

Integração de dados biológicos para identificação de redes gênicas (GRNs) vinculado ao projeto Bioinformática e Reconhecimento de Padrões

Gabriel Rubino
Voluntário Pibic
Engenharia de Computação
Data de ingresso no programa: 08/2012
Prof. Dr. Fabrício Martins Lopes

Área do Conhecimento: 1.03.04.00-2 Sistemas de Computação

CAMPUS CORNÉLIO PROCÓPIO, 2012

GABRIEL RUBINO
FABRÍCIO MARTINS LOPES

INTEGRAÇÃO DE DADOS BIOLÓGICOS PARA IDENTIFICAÇÃO DE REDES GÊNICAS (GRNS)

Relatório Pesquisa do Programa de Iniciação Científica da Universidade Tecnológica Federal do Paraná com o orientador Prof. Dr. Fabrício Martins Lopes e aluno voluntário Gabriel Rubino.

CORNÉLIO PROCÓPIO, 2012

SUMÁRIO

INTRODUÇÃO

Após a descoberta do DNA em 1953, sugiram vários avanços na área de biológica dentre eles o estudo da genômica. A técnica de *DNA microarray* e *RNA-seq* surgiram para investigar hipóteses geradas nesse contexto, essas técnicas quantificam a expressão de determinados genes ao longo do tempo, com isso pode-se verificar como os genes se comportam em diferentes situações, o que pode gerar bastante informação sobre os genes.

Como existem vários genes e muitos estudos nessa área, a era genômica produz uma enorme quantidade de dados. Esses dados estão distribuídos em diversos bancos de dados como: ***KEGG***, Kyoto Encyclopedia of Genes and Genomes[14], ***TAIR***, The Arabidopsis Information Resource[2], ***NCBI***, National center for biotechnology information[8], os quais são públicos na *internet*. A maioria deles apresenta informações específicas ou parciais sobre as entidades biológicas, como por exemplo: função, localização celular, localização no cromossomo, processo metabólico, via metabólica.

Assim surge a necessidade de integrar os dados disponíveis, pois cada banco tem um tipo de informação diferente e se esses dados conhecidos sobre um gene forem centralizados, tornam o estudo do organismo mais eficiente. Ao conhecer os genes com grande riqueza de informações pode-se começar a inferir redes com maior precisão.

As redes de genes mostram como os genes estão conectados para cumprir determinada tarefa biológica, como: produção de enzimas, proteínas que serão usadas na manutenção do ser vivo.

Nesse contexto este trabalho apresenta, um método de captura dos dados de genes de forma online de diversos bancos biológicos e após sua captura integra esses dados de forma padronizada, de modo a gerar uma rede de genes, que pode ser configurada em relação aos bancos pesquisados ou tipo de informação biológica.

Este trabalho trata da integração de dados biológicos, para identificação dos relacionamentos entres os genes, com auxílio de processamento computacional (bioinformática), por isso é necessário o conhecimento de alguns conceitos das duas áreas, biologia e computação.

Na biologia o DNA (ADN, em português: ácido desoxirribonucleico; ou DNA, do inglês: *deoxyribonucleic acid*) é um composto orgânico cujas moléculas contêm as instruções genéticas que coordenam o desenvolvimento e funcionamento de todos os seres vivos e alguns vírus.

O DNA é responsável pela criação do RNA e de proteínas e é primordial para a divisão celular. O DNA é subdividido em porções menores chamados cromossomos.

A estrutura da molécula de DNA foi descoberta conjuntamente pelo norte-americano James Watson e pelo britânico Francis Crick em 7 de Março de 1953, o que lhes valeu o Prêmio Nobel de Fisiologia/Medicina em 1962, juntamente com Maurice Wilkins[10].

Outra estrutura é o RNA (ARN, em português: ácido ribonucleico; ou RNA, do inglês: *ribonucleic acid*), responsável pela síntese de proteínas da célula e algumas outras funções. O RNA é um polímero de nucleótidos, geralmente em cadeia simples, que pode, por vezes, ser dobrado. As moléculas formadas por RNA possuem dimensões muito inferiores às formadas por DNA.

Para melhor organizar o estudo do DNA é usado o conceito de cromossomo onde uma longa sequência de DNA recebe nomes específicos. Nos eucariontes, os cromossomos encontram-se no núcleo celular, agregado a proteínas estruturais chamada histona, já nos procariontes não existem histonas nem núcleo. O ser humano possui 23 pares de cromossomos sendo uma metade vinda da mãe e a outra do pai do indivíduo.

Os cromossomos são subdivididos em genes que pela genética são considerados a unidade fundamental da hereditariedade, pois correspondem a um código distinto para produzir uma determinada proteína ou controlar uma característica, por exemplo, a cor dos olhos, altura, cor

da pele. Para localizar os genes no cromossomo usa-se o locus que é o endereço de determinado gene.

As informações dos genes podem ser encontradas em bancos de dados online, onde pesquisadores colocam suas descobertas para o público, mostrando onde o gene atua, qual proteína ele ajuda a produzir, em qual função está presente, localização celular, processo metabólico, via metabólica e localização no cromossomo.

Para acessar os bancos de dados é necessário utilizar um navegador (em inglês *browser*) que é um programa de computador que interpreta principalmente instruções em HTML e também XPath que significa XML Path Language que é uma linguagem de consulta (*Query Language*) que permite construir expressões que percorrem e processam um documento XML, uma página na internet pode ser montada com essa estrutura, sendo assim mais fácil a organização e construção do site. O programa proposto nesse projeto utiliza algoritmos para interpretar os códigos Xpath presentes nos bancos de dados.

Um algoritmo é uma sequência finita de instruções bem definidas para executar uma determinada tarefa, por exemplo uma receita culinária é um algoritmo. Os algoritmos podem repetir um conjunto de passos e também tomar decisões lógicas, os computadores foram feitos para executarem algoritmos mas os algoritmos frequentemente são processados por máquinas mecânicas ou pessoas. O conceito de um algoritmo foi formalizado em 1936 pela Máquina de Turing de Alan Turing[6] e pelo cálculo lambda de Alonzo Church[5], que formaram as primeiras fundações da ciência da computação.

O problema da Maior Subsequência Comum (Longest Common Subsequence - LCS)[12] é um exemplo de algoritmo onde se deseja encontrar uma subsequência comum de comprimento máximo de duas sequências, ou seja duas sequências de caracteres são comparadas entre si com o objetivo de encontrar qual a porcentagem de igualdade entre as duas. Esse algoritmo foi usado no nesse projeto para fazer a comparação de informações vindas do banco visto que um dado que representa a mesma coisa pode ser escrita de maneira muito semelhante.

Conjuntos de algoritmos podem ser criados para facilitar a criação de um programa resolvendo determinados problemas ou tarefas, esses conjuntos são denominados bibliotecas (do inglês *library*).

Um exemplo de biblioteca é a Selenium HQ[13] que é usada para automatização dos navegadores, foi criada para testar sistemas web, mas pode ser usada em aplicações onde é preciso usar sistemas online, outro exemplo é a Prefuse[11], conjunto de ferramentas para a visualização de dados em vários formatos, tabelas e gráficos.

METODOLOGIA

Várias etapas são necessárias para a captura dos dados nos bancos até a visualização da rede, a primeira é a busca da informação desejada a partir de uma palavra chave. Depois de capturada a informação, visto que os bancos de dados não possuem um padrão para expor seus dados, é preciso padronizá-la , então é feito um processamento dos dados "brutos". Após a padronização dos dados é possível começar a interpretá-los, e assim gerar uma rede dos genes.

Busca Online.O programa tem como entrada uma lista de locus de genes usado como base para as buscas nos determinados bancos.

Para a captura dos dados foi utilizada a biblioteca SeleniumHQ[13], que disponibiliza um *plugin* para o navegador Mozilla FireFox[3] que auxilia na criação do algoritmo responsável pela captura dos dados na página do banco, pois mostra os padrões da página (Xpath, *XML Path Language*) de forma dinâmica e com eficiente recuperação.

Processamento. Depois de descoberto o Xpath e criado o algoritmo de busca de cada banco, a informação retornada vem de forma "bruta" sendo preciso ser organizada para futuro uso, para isso é necessário o processamento dessa String.

Separação da String. O primeiro deles é separar a informação caso a String contenha mais de um dado específico, para isso foi usado o StringTokenizer[9] que separa palavra por palavra da informação vinda do banco.

Redundância de palavras. Depois de separadas, as palavras precisam ser verificadas de modo que não existam informações redundantes, levanto em conta também, possíveis erros de digitação, por isso foi usado o algoritmo LCS[12] para essa tarefa, pois ele compara as palavras e retorna a porcentagem de igualdade entre elas, assim quando se compara todas as palavras pode se usar uma porcentagem de corte (definida pelo usuário) para definir quais palavras são iguais, que devem ser excluídas, e diferentes, que são mantidas nas informações recuperadas.

Criação da rede. Agora os dados estão filtrados e estão prontos para gerarem a rede, então as informação de cada gene são transformadas em índices e todos os genes são comparados entre si, quando mesmo índices são encontrados representa que os genes tem a mesma informação biológica portanto estão ligados, assim uma tabela de adjacência é criada para representar a rede.

Visualização da Rede. Após a tabela da rede estar pronta, ela é processada pela biblioteca Prefuse[11] que gera a rede de maneira visual.

Figura 1. Etapas do programa Visual Ontograror

RESULTADOS E DISCUSSÕES

Para os resultados foi utilizada a lista de locus a seguir, que foi usada como entrada no Visual Ontograror, assim o programa pesquisou nos bancos de dados online e retornou todas as informações biológicas usadas para a construção das redes.

Tabela 1. Lista de Locus

Locus
AT5G54770
AT4G34200
AT2G36530
AT5G41370
AT1G05055
AT1G03190
AT3G05210
AT1G14030
AT1G67090
AT2G28000
AT2G34590
AT3G55410
AT4G24620
AT2G22480
AT2G01140
AT1G74030
AT4G37870
AT3G04080
AT1G22940
AT3G24030
AT5G65720
AT1G09430
AT2G47510
AT1G01090

Filtro LCS. Segue o resultado dos dados recuperados com todas informações biológicas juntas com o filtro LCS para eliminar redundâncias dos dados:

Tabela 2. Resultados com filtro LCS com 100% de igualdade

100% de igualdade
6-phosphofructokinase complex
apoplast
cellular_component
chloroplast
chloroplast envelope
chloroplast ribulose bisphosphate carboxylase complex
chloroplast stroma
chloroplast thylakoid
chloroplast thylakoid membrane
chromosome: 1
chromosome: 2
chromosome: 3
chromosome: 4
chromosome: 5
cytoplasm
cytosol
cytosolic ribosome
hypocotyl
membrane
mitochondrial envelope
mitochondrion
nucleolus
nucleus
plasma membrane
plasmodesma
plastid
plastoglobule
pollen
primary root elongation zone
root cortex
shoot apex
thylakoid
thylakoid lumen
trichome

Tabela 3. Resultados com filtro LCS com 90% de igualdade

90% de igualdade
6-phosphofructokinase complex
apoplast
cellular_component
chloroplast
chloroplast envelope
chloroplast ribulose bisphosphate carboxylase complex
chloroplast stroma
chloroplast thylakoid
chloroplast thylakoid membrane
chromosome: 1
cytoplasm
cytosol
cytosolic ribosome
hypocotyl
membrane
mitochondrial envelope
mitochondrion
nucleolus
nucleus
plasma membrane
plasmodesma
plastid
plastoglobule
pollen
primary root elongation zone
root cortex
shoot apex
thylakoid
thylakoid lumen
trichome

Ao usar o filtro é necessário a avaliação das informações para o LCS não eliminar informação válidas, como no exemplo, quando o filtro estava a 100% , tabela 2, a informação "chromosome: 2", "chromosome: 3" , "chromosome: 4" , "chromosome: 5" estava presente, se esses dados forem julgados redundantes, é preciso filtrar a informação a 90%, tabela 3, para que o resultado final fique sem esses dados.

Visualização de Redes. Foram geradas as visualizações das tabelas de adjacência para cada tipo de informação biológica a seguir:

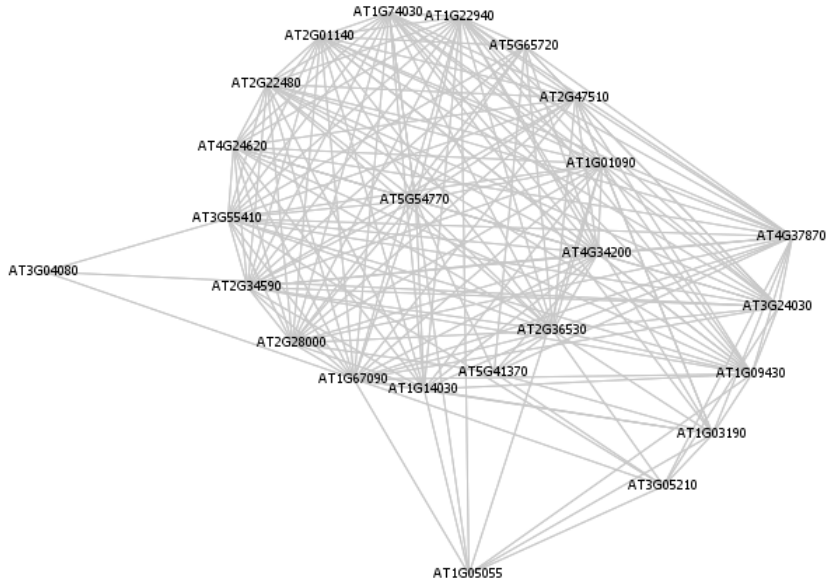

Figura 2. Visualização da rede de todas as informações biológicas

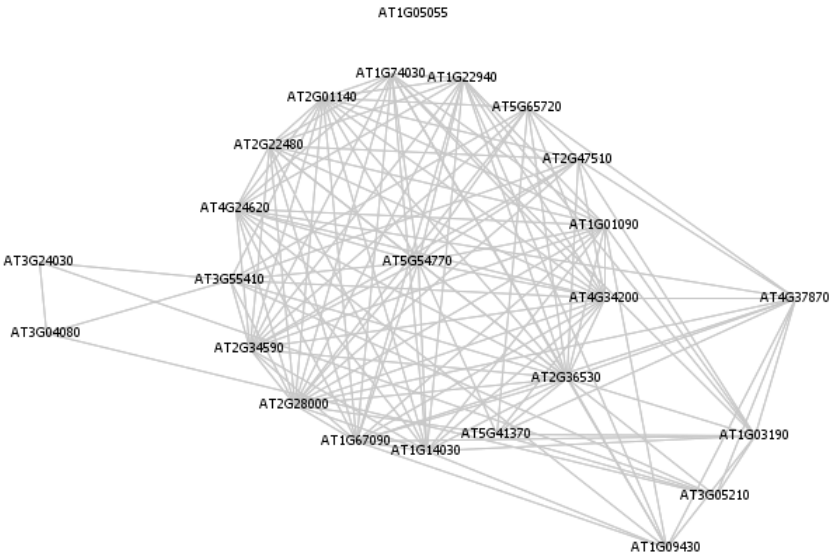

Figura 3. Visualização da rede da função

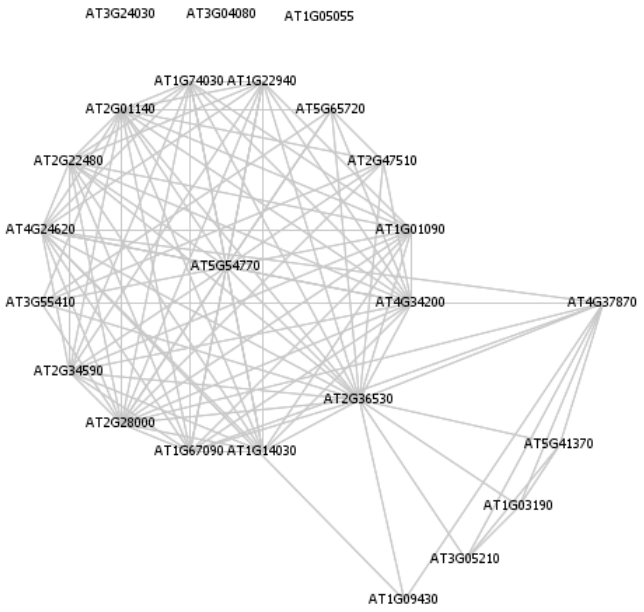

Figura 4. Visualização da rede da localização celular

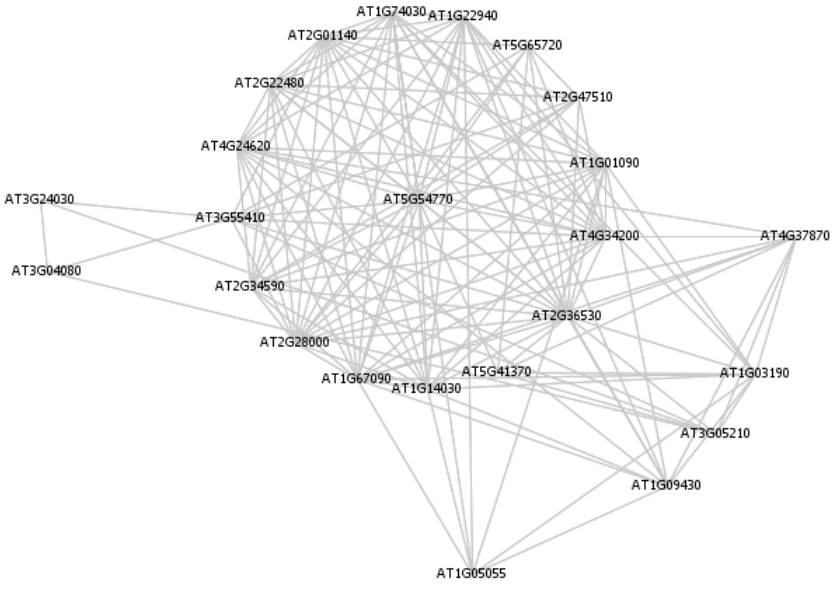

Figura 5. Visualização da rede da localização no cromossomo

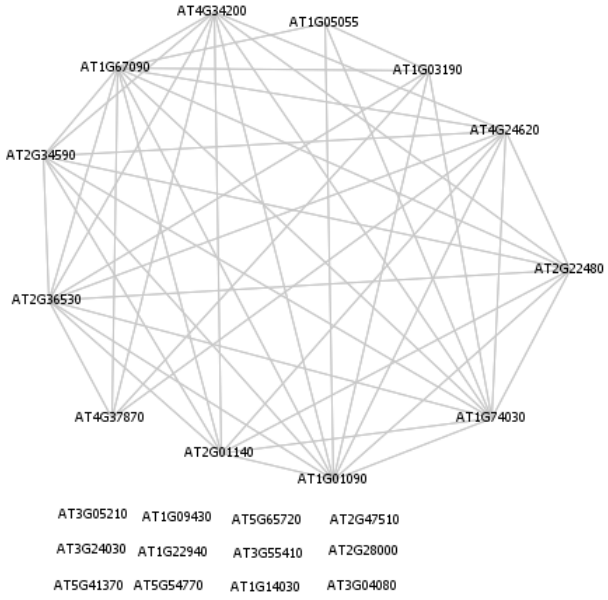

Figura 6. Visualização da rede do processo metabólico

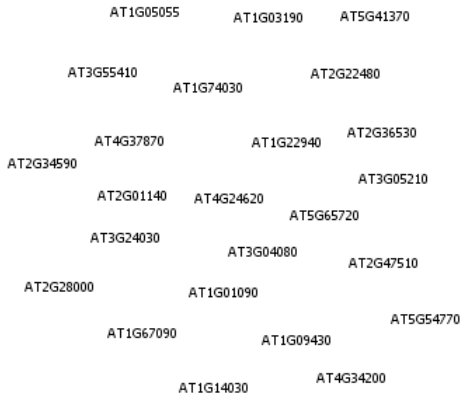

Figura 7. Visualização da rede do produto

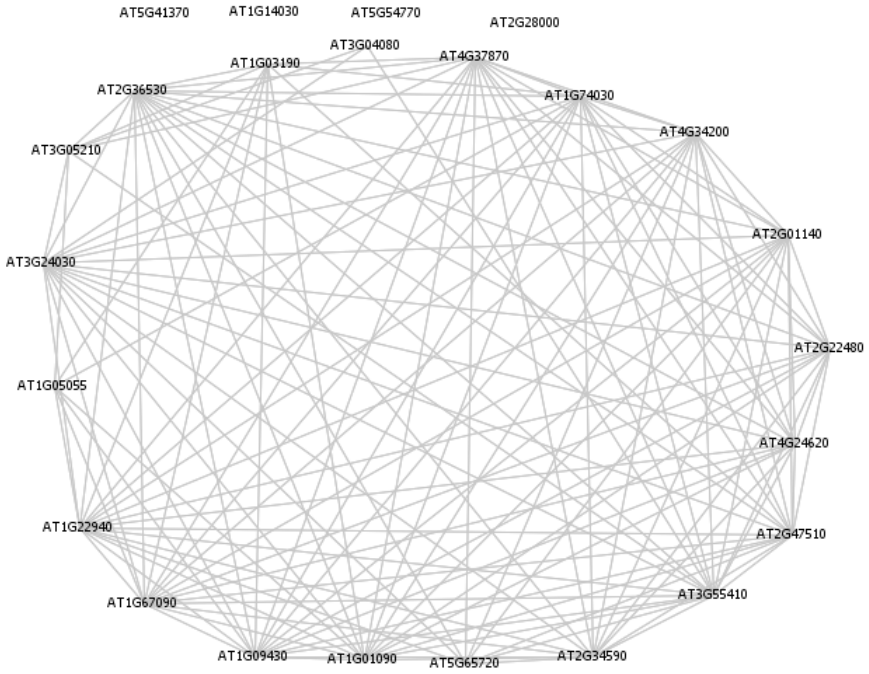

Figura 8. Visualização da rede da via metabólica

Tabela com o número de nós. Analisando a tabela 4, que foi obtida a partir do grafo gerado pelo programa, é possível perceber que o gene com o locus AT2G36530 (ID 2) nas propriedades biológicas função, localização celular, localização no cromossomo e com todas as informações juntas apresenta maior número de ligação mostrando ser um gene importante nessas atividades, outros genes que se destacaram foram os de ID 15, 23, 14, 8 e 1 pois apresentaram grande número de ligações em quase todas as propriedades biológicas.

Já os genes de ID 17, 4, 19, 6 e 3 possuem poucas ligações o que demonstra uma atividade mais baixas nas tarefas das propriedades biológicas selecionadas.

Um fato a se destacar foi a propriedade biológica dos produtos que não apresentou nenhuma conexão entre os genes, mostrando que os genes selecionados possivelmente não trabalham em conjunto nessa atividade.

Tabela 4. Número de ligações dos nós para cada informação biológica

ID	Locus	Função	Loc. Celular	Loc. Crom	Prod. Met.	Produto	Via Met.	Todas
0	AT5G54770	16	15	16	0	0	0	16
1	AT4G34200	17	16	16	9	0	15	19
2	AT2G36530	20	20	20	10	0	17	21
3	AT5G41370	8	4	6	0	0	0	8
4	AT1G05055	0	0	7	4	0	7	8
5	AT1G03190	11	4	11	6	0	10	12
6	AT3G05210	7	4	7	0	0	7	8
7	AT1G14030	14	12	15	0	0	0	15
8	AT1G67090	15	13	16	11	0	16	19
9	AT2G28000	17	16	16	0	0	0	17
10	AT2G34590	15	12	13	8	0	15	18
11	AT3G55410	11	7	10	0	0	17	19
12	AT4G24620	14	14	14	9	0	14	17
13	AT2G22480	13	12	13	8	0	14	17
14	AT2G01140	15	15	15	8	0	15	18
15	AT1G74030	14	12	15	10	0	16	19
16	AT4G37870	11	9	9	5	0	17	19
17	AT3G04080	3	0	3	0	0	3	3
18	AT1G22940	14	12	15	0	0	17	20
19	AT3G24030	3	0	3	0	0	17	17
20	AT5G65720	10	7	8	0	0	9	12
21	AT1G09430	9	3	10	0	0	16	17
22	AT2G47510	11	7	9	0	0	16	18
23	AT1G01090	14	12	15	10	0	16	19
Total de ligações		282	226	282	98	0	274	376

CONCLUSÕES

A partir da lista de locus de genes o programa buscou informações, de forma *online*, nos bancos de dados públicos e isso trouxe mais precisão aos resultados, pois os dados recuperados estão sempre atualizados de acordo com o banco original. O processo de obtenção foi eficiente e automático, atentando que conexões mais rápidas podem acelerar o processo. As informação recuperadas foram filtradas e processadas com base nas análises biológicas definidas pelos usuário, tornando o resultado mais focado.

Após o processamento das informações, foi definido para o algoritmo gerar sete redes, sendo seis para informações biológicas distintas e uma com todas as informações juntas, a partir dessas redes foi possível classificar os genes de acordo com seu número de ligações, assim é mostrado quais genes são mais importantes (mais ligados) e menos importantes (menos ligados). O gene considerado mais ligado de acordo com os resultados obtidos foi o de locus AT2G36530, então esse gene é forte candidato para futuras pesquisas na área biológica. Já o gene de locus AT3G04080 apresentou poucas ligações mostrando não estar muito presente nas propriedades biológicas consideradas.

Ao analisar as redes percebe-se que existem genes que estão mais conectados entre si, mostrando maior interatividade entre eles, esse tipo de dinâmica pode ser chamada de comu-

nidades de genes. Essas comunidades podem ser melhor analisadas, por isso essa é uma proposta para melhoramento do projeto, visto que essas comunidades podem trazer grande quantidade de informações das relações dos genes.

Como uma das características principais desse projeto é a captura de dados *online*, seria viável a criação de um sistema totalmente *web*, pois a conexão com a *internet* é essencial para o pleno funcionamento desse programa.

Uma outra possibilidade de trabalho futuro é a integração dessa ferramenta ao projeto DimReduction[7], e a utilização das informações biológicas indexadas como auxílio na inferência das redes, i.e., as redes passariam a serem inferidas a partir dos perfis de expressão e das informações biológicas conhecidas[1].

Como resultado deste trabalho foi desenvolvido o software VisualOntogrator, o qual está livremente disponível no site https://code.google.com/p/visual-ontogrator/[4].

REFERÊNCIAS

[1] F.F. da Rocha Vicente, F.M. Lopes, and R.F. Hashimoto. Improvement of gns inference through biological data integration. In *Genomic Signal Processing and Statistics (GENSIPS), 2011 IEEE International Workshop on*, pages 70 –73, dec. 2011.

[2] Carnegie Institution for Science Department of Plant Biology. Tair: The arabidopsis information resource. Disponível em: <http://www.arabidopsis.org/> . Acesso em: 6 abr. 2012, 15:36:24.

[3] Mozilla Foundation. Mozzila firefox. Disponível em: <http://www.mozilla.org/pt-BR/about/>. Acesso em: 8 set. 2012, 18:31:22.

[4] RUBINO. Gabriel. Visual ontogrator. Disponível em: <https://code.google.com/p/visual-ontogrator/>. Acesso em: 14 set. 2012, 12:31:42.

[5] Achim Jung. A short introduction to the lambda calculus. Disponível em: <http://www.cs.bham.ac.uk/ axj/pub/papers/lambda-calculus.pdf>. Acesso em: 5 out. 2012, 16:31:42.

[6] WAYNE. Kevin. Turing machines. Disponível em: <http://introcs.cs.princeton.edu/java/74turing/>. Acesso em: 9 abr. 2012, 13:31:22.

[7] Fabricio Lopes, David Martins, and Roberto Cesar. Feature selection environment for genomic applications. *BMC Bioinformatics*, 9(1):451, 2008.

[8] U.S. National Library of Medicine National Center for Biotechnology Information. Ncbi: National center for biotechnology information. Disponível em: <http://www.ncbi.nlm.nih.gov/>. Acesso em: 7 abr. 2012, 11:31:22.

[9] Oracle. String tokenizer. Disponível em: <http://docs.oracle.com>. Acesso em: 9 abr. 2012, 19:31:22.

[10] Nobel Prize ORG. Nobel prize. Disponível em: <http://www.nobelprize.org/nobel_prizes/medicine/laureates/1962/crick-bio.html>. Acesso em: 3 set. 2012, 12:31:22.

[11] Prefuse.org. Prefuse. Disponível em: <http://prefuse.org/>. Acesso em: 4 out. 2012, 13:31:22.

[12] Thomas H. Cormen Ű Charles E. Leiserson Ű Ronald L. Rivest. *Algoritmos: Teoria e Prática*. Elsevier.

[13] Selenium. Selenium hq. Disponível em: <http://seleniumhq.org/about/> . Acesso em: 3 jun. 2012, 10:31:22.

[14] Kyoto University and the University of Tokyo. Kegg: Kyoto encyclopedia of genes and genomes. Disponível em: <http://www.genome.jp/kegg/>. Acesso em: 6 set. 2012, 23:31:22.

Gabriel Rubino

Fabrício Martins Lopes