Desenvolvimento de uma ferramenta para análise visual de resultados mineração de textos sobre genes

Marcel dos Santos Toledo [1]
Maria Fernanda Moura[2]
Roberto Hiroshi Higa[2]

O projeto de Prospecção e priorização de genes candidatos por meio de técnicas de mineração de dados e textos - ProsGEN – (03.09.01.025.00.02) objetiva desenvolver e/ou adaptar metodologias de mineração de dados e textos para apoiar a etapa de bioinformática em projetos que utilizem tecnologias de varredura genômica. Essas tecnologias procuram identificar genes candidatos relacionados às características fenotípicas de interesse econômico para a agricultura brasileira, para posterior investigação em bancada e, com essa metodologia estudar os genes candidatos já identificados como relacionados às características fenotípicas de interesse para o melhoramento genético.

A estratégia proposta baseia-se na utilização de informações oriundas de bancos de dados textuais para apoiar a interpretação biológica de genes candidatos. Ela se justifica por relatos na literatura que apontam fontes de dados textuais como as que apresentam melhores resultados na tarefa de priorização semiautomática de genes candidatos seguida pela utilização de informações oriundas de bancos de dados como o Gene *Ontology* (GO) (THE GENE ONTOLGY, 2012). Assim, o projeto pretende disponibilizar metodologias, e as correspondentes ferramentas computacionais, para prospecção e priorização de genes candidatos, a partir de conhecimento já existente, expresso em artigos científicos.

[1] Estudante de Análise e Desenvolvimento de Sistemas, estagiário da área de Inteligência Computacional, marcel_slp@hotmail.com

[2] Embrapa Informática Agropecuária, {maria-fernanda.moura, roberto.higa}@embrapa.br

Os documentos (resumos de artigos científicos), relacionados aos genes do organismo de interesse, são obtidos do sítio da Pubmed (PUBMED, 2012), utilizando a ferramenta *Eutils-search* (TANAKA; HIGA, 2011). Esses documentos são, então, préprocessados, utilizando-se a ferramenta eTMLib (YAMADA et al., 2012), resultando numa matriz atributo-valor onde as linhas representam os genes e as colunas os termos relevantes para a descrição dos genes. Esta matriz é utilizada em 3 diferentes processos para descrição do conjunto de genes analisados:

- no processo de priorização, um conjunto de genes de interesse é utilizado como referência para avaliar a similaridade dos genes em teste, resultando em um ranking;
- no processo de prospecção, os genes são agrupados, formando uma hierarquia, cujos nós são rotulados com os termos mais relevantes para descrição dos genes associados ao ramo definido pelo nó;
- em um terceiro processo, uma hierarquia pré-definida de genes, por exemplo resultante de uma análise de expressão gênica, tem seus nós rotulados com os termos mais relevantes para descrição dos genes associados ao nó.

No primeiro caso, resulta uma lista ordenada de genes, enquanto, nos dois últimos, uma hierarquia com nós rotulados. O objetivo deste trabalho é desenvolver uma ferramenta visual para facilitar a análise dos resultados dos processos acima descritos. Para desenvolver essa ferramenta, pretende-se utilizar a linguagem Java e os componentes gráficos JUNG (JUNG, 2012) para apresentação de hierarquias e JfreeChart (JFREECHART, 2012) para apresentação de rankings. Ambos os componentes são livres e prestam à apresentação de grafos, no caso do JUNG e de diferentes formas de gráficos, no caso do JfreeChart, do qual pretende-se utilizar o componente *scatterplot* para apresentação do *ranking* em escala.

No momento, estão sendo implementados os primeiros protótipos para testar a adequabilidade dos componentes escolhidos e auxiliar o processo de coleta das especificações para a ferramenta. Uma ilustração da interface da ferramenta atualmente em desenvolvimento pode ser observada na Figura 1. Na parte b da Figura 1, tem-se a aba com o componente *ScatterPlot* para a representação do *ranking*; e na parte a, a aba com o componente JUNG para a representação de uma árvore hierarquizada.

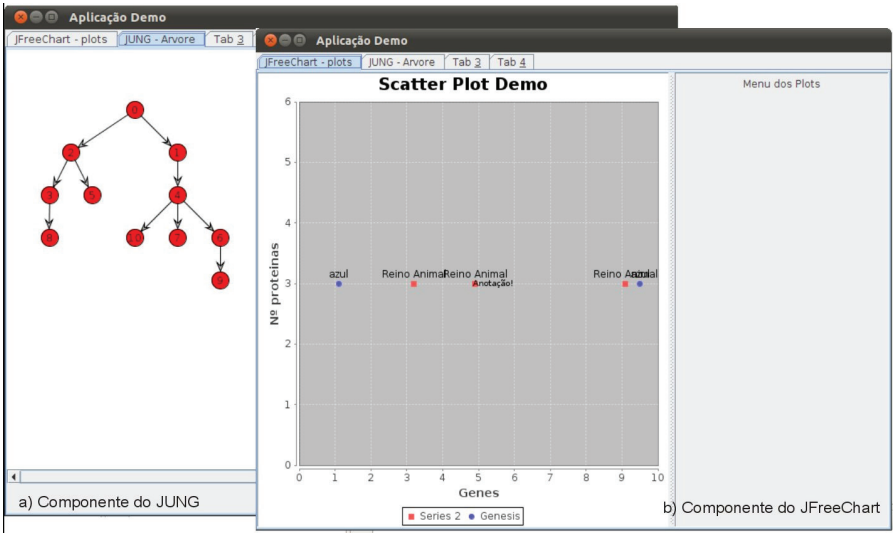

Figura 1. Protótipo de interface para a ferramenta proposta.

Referências

JFREECHART. **JfreeChart**. 2012. Disponível em: < http://www.jfree.org/jfreechart/>. Acesso em: 1 out. 2012.

JUNG. **Jung - Java Universal Network/Graph Framework**. 2012. Disponível em: <http://jung.sourceforge.net/>. Acesso em: 1 out. 2012.

PUBMED. **Pumed**. 2012. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed>. Acesso em: 1 out. 2012.

TANAKA, R. S.; HIGA, R. H. **Eutils-search versão 2.0 - manual do usuário**. Campinas: Embrapa Informática Agropecuária, 2011. 23 p. il. (Embrapa Informática Agropecuária. Documentos, 115). Disponível em: <http://ainfo.cnptia.embrapa.br/digital/bitstream/item/56665/1/Doc115.pdf>. Acesso em: 1 out. 2012.

THE GENE ONTOLGY. **GO**. 2012. Disponível em: < http://www.geneontology.org/>. Acesso em:1 out. 2012.

YAMADA, A. K.; MOURA, M. F.; CRUZ, S. A. B.; HIGA, R. H. Uma solução flexível para a etapa de pré-processamento em mineração de textos. In: CONGRESSO INTERINSTITUCIONAL DE INICIAÇÃO CIENTÍFICA, 6., 2012, Jaguariúna. **Anais**... Campinas: Embrapa; ITAL, 2012. p. 1-12. CIIC 2012. No 12611.