Mineração de textos aplicada à análise de dados de expressão gênica por microarranjos

Rodrigo Shizuo Yasuda[1]
Roberto Hiroshi Higa[2]

O estudo de perfis de expressão gênica relacionados a manifestações de diferentes fenótipos pode fornecer informações importantes para a compreensão da biologia desses processos. Em particular, na agricultura, a identificação dos genes mais relevantes para manifestações de fenótipos de interesse econômico constitui uma etapa importante do processo de melhoramento genético animal e vegetal.

O projeto "Rede Genômica Animal" utiliza a tecnologia de microarranjos para realizar análises do perfil de expressão gênica, medidas em diferentes condições, com o objetivo de prospectar genes relevantes para manifestações de fenótipos de interesse econômico para a pecuária brasileira, como aqueles relacionados à resistência a carrapatos, à resistência a mastite e à maciez de carne.

Uma parte crucial na análise de expressão gênica consiste em relacionar o conjunto de genes, cujo perfil de expressão tenha se mostrado interessante (ex: genes diferencialmente expressos), com o conhecimento biológico relacionado a eles e que encontra-se armazenado em bancos de dados especializados e em publicações científicas.

[1] *Universidade Estadual de Campinas; rodrigosy@cnptia.embrapa.br*

[2] *Embrapa Informática Agropecuária; roberto@cnptia.embrapa.br*

Este trabalho se insere no projeto “Rede Genômica Animal” e tem por objetivo construir uma ferramenta que utilize técnicas de mineração de textos para apoiar a interpretação biológica de dados de experimentos de expressão gênica. Por isso, os dados de expressão gênica a serem utilizados para validação da ferramenta são aqueles gerados no escopo do projeto “Rede Genômica Animal”, referente ao organismo Bos taurus.

Os dados textuais serão obtidos do Pubmed (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2010a), um banco de dados público que agrega informações relevantes sobre artigos técnico-científicos da área de medicina e bioquímica. Para essa tarefa será utilizada a ferramenta E-Utils (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2010b), que permite realizar buscas e *downloads* no banco de dados Pubmed.

Após o *download* do Pubmed, esses dados textuais serão tratados utilizando as ferramentas de mineração de textos Pretext (SOARES et al., 2008) e Taxtools (MOURA; REZENDE, 2010). A ferramenta Pretext permite que se calcule as frequências de cada termo dos textos obtidos; enquanto a ferramenta TaxTools permite obter *clusters* de documentos de acordo com a frequência com que cada termo aparece em todos os artigos obtidos. Isso permite identificar os termos-chaves para cada *cluster,* relacionando-os aos correspondentes genes. A ferramenta, propriamente dita, será desenvolvida utilizando as linguagens de programação Python e Java.

Até o momento, foram realizadas as tarefas de *download* dos dados textuais, incluindo o seu armazenamento em um banco de dados local, e de pré-processamento, utilizando técnicas de mineração de textos. Os dados obtidos do banco de dados Pubmed incluem resumos, títulos, data de publicação dos artigos, descrições dos genes e a característica RIF do Entrez Gene (este consiste de uma sentença com a descrição funcional de um gene) vinculados ao organismo Bos taurus. O *download* desses dados foi feito com cautela, pois o sítio do Pubmed impõe restrições quanto à quantidade de dados que pode ser obtida no decorrer de um dia.

Toda a plataforma de coleta dessas informações foi desenvolvida em linguagem Java.

Em seguida, os dados coletados foram armazenados na forma de documentos XML e analisados utilizando as ferramentas de mineração de textos PreText e TaxTools. Considerando o conjunto de documentos relacionados ao organismo Bos taurus (estudo de caso), foram encontrados 6548 artigos, dos quais 6158 possuem resumo e 2476 possuem RIF. Além disso, foram encontrados 23628 genes relacionados ao Bos taurus, dos quais 23617 possuem símbolo e descrição. Esses números (26% dos genes do organismo Bos taurus possuem publicações associado) estão em acordo com a proporção de genes anotados do genoma do Bos taurus.

As próximas atividades consistem (i) na definição dos parâmetros mais adequados para utilização das ferramentas PreText e TaxTools; e (ii) na construção da ferramenta de análise propriamente dita, utlizando a linguagem java, contemplando a visualização dos diferentes *clusters* contidos no conjunto de genes em análise e as correspondentes palavras-chaves que os descrevem.

Referências

MOURA, M. F.; REZENDE, S. O. A simple method for labeling hierarchical document clusters. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 10., 2010, Innsbruk, Austria. **Proceedings**... Anaheim, Calgary; Zurich: Acta Press, 2010. p. 336-371. v. 1.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **Pubmed**. 2010a. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed>. Acesso em: 20 jun. 2010.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **Entrez programming utilities.** 2010b. Disponível em**:** <http://eutils.ncbi.nlm.nih.gov/entrez/query/static/ eutils_help.html>. Acesso em: 20 jun. 2010.

SOARES, M. V. B.; PRATI, R. C.; MONARD, M. C. **PreTexT II**: descrição da reestruturacão da ferramenta de pré-processamento de Textos. São Carlos, SP: USP, ICMC, 2008. (Relatório técnico, n. 333).