PRIORIZAÇÃO DE GENES CANDIDATOS UTILIZANDO MINERAÇÃO DE TEXTOS

VINÍCIUS F. **DIAS**[1]**;** ROBERTO HIROSHI **HIGA**[2]**;** MARIA FERNANDA **MOURA**[3]

Nº 13607

RESUMO

O objetivo deste trabalho foi desenvolver e/ou adaptar metodologias de mineração de textos para identificar genes candidatos relacionados a alguma característica fenotípica de interesse econômico para a agricultura brasileira. O projeto consistiu em, inicialmente, adaptar a ferramenta de mineração de dados *eTMLib* (*Embrapa's Text Mining Library*), baixar do site do NCBI (*National Center for Biotechnology Information*) as publicações de genes relacionados à espécie *H. sapiens* e pré-processar os documentos para gerar uma matriz atributo-valor. Em seguida, essa matriz foi utilizada em métodos de validação cruzada e ranqueamento dos valores obtidos. Como experimento de teste, foram utilizados grupos de genes previamente associados a alguma característica fenotípica em meio a outros genes aleatórios, e esperou-se que os primeiros fossem considerados prioritários pelo método. Outro experimento realizado foi a clusterização hierárquica de um conjunto de genes associados a diferentes fenótipos, com a expectativa de que genes associados a um mesmo fenótipo ficassem agrupados. Em ambos os casos, os resultados se mostraram bons para alguns fenótipos, mas ruins para outros, e a qualidade dos mesmos foi variando com os diferentes parâmetros utilizados no pré-processamento para extração e corte de termos. Por se tratar de um grande volume de documentos, entretanto, essa seleção de atributos encontra-se limitada devido ao tempo inviável que a geração da matriz pode tomar. No momento, estão em progresso a paralelização da biblioteca de pré-processamento, para que o tempo de processamento diminua, e o desenvolvimento de novas formas de seleção de atributos, capazes de tratar uma quantidade maior de casos em que ocorre uma boa priorização dos genes.

[1] Bolsista CNPq: Graduação em Eng. de Computação, UNICAMP, Campinas-SP, v.fernandesdias@gmail.com.br.

[2] Orientador: Pesquisador, Embrapa Informática Agropecuária, Campinas-SP, roberto.higa@embrapa.br

[3] Co-orientadora: Pesquisadora, Embrapa Informática Agropecuária,, Campinas-SP, maria-fernanda.moura@embrapa.br.