PROPOSTA DE UTILIZAÇÃO DE MINERAÇÃO DE TEXTOS PARA ANÁLISE E PRIORIZAÇÃO DE GENES CANDIDATOS DE INTERESSE PARA A AGRICULTURA

Roberto Hiroshi Higa[1]
Rodrigo Shizuo Yasuda[2]
Maria Fernanda Moura[3]
Poliana Fernanda Giachetto[4]

RESUMO: Neste artigo é apresentado um projeto para utilização de técnicas de mineração de textos para análise e priorização de genes candidatos em experimentos de varredura genômica, visando a inclusão de novos genes em programas de melhoramento animal e vegetal. Embora no presente estágio do projeto, os resultados ainda sejam preliminares, é apresentado um experimento que demonstra como a descrição da função biológica de genes pode ser capturada a partir de textos que descrevem sua função biológica, que é o primeiro passo no sentido de comparar genes funcionalmente, visando priorizá-los segundo um critério específico.

PALAVRAS-CHAVE: mineração de textos, genes candidatos, priorização de genes, melhoramento genético.

A PROPOSAL FOR USING TEXT MINING IN THE ANALYSIS AND PRIORITIZING OF CANDIDATE GENES IN AGRICULTURE

ABSTRACT: This paper presents the current status of a project whose aim is to make use of text mining to analyse and prioritize candidate genes in genome wide screening experiments. The candidate genes should be included in animal and plant breeding programs. Although the project is still a working in progress and present results are not complete, they illustrate how genes biological functions could be captured from texts describing it, which is the first step to be able to compare genes based on their function and then prioritize them.

KEYWORDS: text mining, candidate gene, gene priorization, genetic improvement.

1. INTRODUÇÃO

A identificação dos genes, que contribuem para a variação de características fenotípicas de interesse econômico, é o primeiro passo para o desenvolvimento de métodos de seleção de genótipos superiores em programas de melhoramento genético de espécies vegetal e animal de interesse econômico para a agricultura. Grosso modo, existem duas estratégias para o estudo da genética envolvida na manifestação de características complexas: a de varredura de todo o genoma e a de identificação de genes candidatos. Avanços tecnológicos recentes na área de genômica, que contemplam um alto grau de automatização em plataformas de alta capacidade de processamento e baixo custo operacional, tem favorecido, cada vez mais, a estratégia de varredura de todo o genoma. Como exemplo, tem-se os estudos de associação genótipo vs fenótipo, utilizando polimorfismos de base individual (SNP – *Single Nucleotide Polymorphism*) como marcadores moleculares, que

[1] Doutor em Engenharia Elétrica. Embrapa Informática Agropecuária. E-mail:roberto@cnptia.embrapa.br
[2] Estudante de Engenharia da Computação. Universidade Estadual de Campinas. E-mail: rodrigoyasuda@gmail.com
[3] Doutora em Ciências Matemáticas e da Computação. Embrapa Informática Agropecuária. E-mail: fernanda@cnptia.embrapa.br.
[4] Doutora em Melhoramento Animal. Embrapa Informática Agropecuária. E-mail: poliana@cnptia.embrapa.br.

resultam em um refinamento expressivo na região de QTLs (*Quantitative Trait Locus*) identificada. Outra estratégia de varredura genômica consiste em comparar o perfilamento de expressão gênica entre representantes de valores fenotípicos extremos para uma característica em estudo, por exemplo, utilizando a tecnologia de microarranjo. Essas tecnologias geram um volume extraordinário de dados, capaz de fornecer informações valiosas sobre diferentes aspectos da biologia do organismo estudado e sua relação com a manifestação das características fenotípicas de interesse. Contudo, paradoxalmente, identificam centenas e, às vezes, milhares de genes candidatos relacionados com a característica fenotípica estudada em experimentos de bancada

Em paralelo a este cenário, observa-se um enorme acúmulo de conhecimento sobre genes relacionados a diversos organismos, tanto na literatura científica quanto em bancos de dados moleculares. Hoje, existem mais de 2000 genomas entre finalizados, *drafts* e em progresso, resultando em mais de 5 milhões de entradas na coleção de sequências de referência *Entrez Gene* (NCBIa, 2011) e mais de um milhão de citações de literatura biomédica a partir do MEDLINE (NLM, 2011), coleção de periódicos relacionados à ciência da vida, e livros *online* contendo a palavra "gene". Esses números refletem o paradoxo a que estão submetidos o(a)s pesquisadore(a)s da área de genética, genômica e biologia molecular: ao mesmo tempo em que as tecnologias de varredura genômica estão gerando dados moleculares a uma taxa nunca vista, para dar sentido biológico (e biotecnológico) a esses dados, ele(a) precisa minerar essas informações em bancos de dados estruturados ou textuais ainda maiores que o conjunto de dados para o qual ele(a) quer dar sentido biológico.

É praticamente impossível manter-se atualizado com todas as novas descobertas e teorias, mesmo considerando o campo de pesquisa do próprio(a) pesquisador(a). Além disso, muito dessa riqueza de informação pode não ter sido completamente capturada por revisores e curadores de banco de dados, permanecendo na forma textual no corpo da literatura das ciências biológica. Assim, para enfrentar este desafio é fundamental lançar mão de ferramentas computacionais, baseadas em técnicas de aprendizado de máquina, capazes de automatizar a transformação e sumarização de grandes volumes de dados em informação útil. Essa tarefa é conhecida como mineração de dados , e, em particular para dados textuais, mineração de textos (Witten e Frank, 2005).

Para auxiliar o trabalho desses pesquisadores, no projeto Prosgen – Prospecção e priorização de genes candidatos por meio de técnicas de mineração de dados e textos – vem sendo construídas ferramentas baseadas em mineração de textos, para utilização na fase de análises que visem a priorização de genes candidatos para estudo em bancada e, posterior inclusão em programas de. melhoramento genético animal e vegetal. Embora a utilização deste tipo de ferramenta já venha sendo aplicada na área de biomedicina (Ananiadou e McNaught, 2006), ela ainda é uma novidade na área de melhoramento genético de interesse agropecuário. Assim, neste trabalho, é apresentada a metodologia em desenvolvimento no Prosgen e uma exemplificação de uso, resultados e possibilidades.

2. MATERIAL E MÉTODOS

A proposta do projeto Prosgen para análise e priorização de genes candidatos utiliza informações oriundas do banco de dados de artigos científicos denominado Pubmed (NCBIb, 2011) e técnicas de mineração de textos na para apoiar a interpretação biológica de genes candidatos. Ela se justifica por relatos na literatura (Aerts et al., 2006) que apontam fontes de dados textuais como as que apresentam melhores resultados na tarefa de priorização semi-automática de genes candidatos complementadas ou melhoradas pela utilização de ontologias como a *Gene Ontology – GO* (GO, 2011), embora a utilização combinada de fontes de dados adicionais como KEGG (KEGG, 2011) ou Interpro (EMBL-EBI, 2011), também tenham melhorado o resultado da priorização (Aerts et al., 2006).

A estratégia de priorização proposta baseia-se no trabalho desenvolvido por Yu et al. (2010), que apresentou um estudo para priorização de genes candidatos relacionados a doenças em humanos, combinando diferentes vocabulários e uma estratégia de fusão de resultados baseado em combinação de kernels (De Bie et al., 2007).

O primeiro passo para implementação da metodologia proposta é a construção de uma coleção de documentos, em formato XML, contendo descrições funcionais de genes encontrados no genoma do organismo em estudo. Na presente proposta, considera-se como estudo de caso o organismo *b.taurus*. Para cada gene é criado um documento contendo os títulos e resumos de artigos relacionados, recuperados, via *ftp*, do sítio (NCBIc, 2011), utilizando o conjunto de programas *Entrez Programming Utilities*. Também são incluídas as anotações da característica GeneRIF do Entrez Gene para aqueles genes que possuem essa anotação.

No segundo passo, esses documentos são submetidos a um processo de filtragem para eliminação de *stopwords*, seguido por um processo de identificação e remoção de inflexões (*stemming*). Em seguida, obtém-se um conjunto de unigramas que, neste caso, corresponde aos *stems* encontrados. Então, elimina-se os *stem*s muito raros e muito frequentes, considerados estatisticamente não informativos e realiza-se um processo de identificação de bigramas e trigramas estatisticamente mais significativos, de acordo com a metodologia proposta por Moura et al. (2008). Esses bigramas e trigramas correspondem às combinações sequenciais, por ordem de ocorrência dos *stems* já filtrados. Este mesmo pré-processamento será utilizado para os vocabulários controlados (*Gene Ontology* – GO, *Phenotypic Quality Ontology* – PATO (PATO, 2011), *Mammalian Phenotype Ontology* – MP (MP, 2011), *Environment Ontology* – EO (EO, 2011), *Trait Ontology* – TO (GRAMENE, 2011), *Vertebrate Trait Ontology* – VT (VT, 2011) e *Plant Ontology* (PO, 2011)), de acordo com a aplicação. O conjunto de unigramas, bigramas e trigramas em comum entre o conjunto de documentos que representam os genes e os vocabulários controlados constituem o conjunto de atributos a ser utilizado para representar os documentos analisados, visando comparar, agrupar e priorizar genes candidatos. Para isso, esses documentos são representados em um espaço n-dimensional, formado pelo conjunto de atributos identificados, onde os valores dos vetores são calculados como o inverso da frequência de ocorrência nos documentos, baseado no fato de que Yu et al. (2008) relatam essa representação como a que apresenta os melhores resultados para priorização de genes relacionados a doenças em humanos.

Os genes, representados por seus documentos associados, serão analisados segundo duas estratégias. A primeira estratégia tem por objetivo priorizar, automaticamente, o conjunto de genes, comparando-os como um conjunto de genes que, sabe-se *a priori,* estão relacionados com a característica fenotípica de interesse. Desta forma, genes apresentando maior similaridade de descrições funcionais com o conjunto de genes previamente relacionados com a característica fenotípica ocuparão o topo da lista de genes priorizados. Dentre as possíveis estratégias para se priorizar genes candidatos, será utilizada a estratégia de classificação com uma classe (OCC) (Tax, 2001), onde um gene é priorizado de acordo com sua distância para o centro de uma bola n-dimensional envolvendo todo o conjunto de genes já relacionados com a característica fenotípica de interesse. Essa escolha é justificada pelos resultados obtidos por Yu et al. (2008), que indicam essa como sendo a melhor estratégia. A segunda estratégia para análise do conjunto de genes candidatos consiste em agrupar hierarquicamente o conjunto de documentos representando genes, utilizando o cosseno como medida de dissimilaridade e a estratégia

linkage para construção da hierarquia; descrever cada ramo da hierarquia com os termos mais relevantes para o conjunto de genes associados; e apresentar a hierarquia rotulada ao usuário, utilizando uma ferramenta de visualização gráfica do agrupamento. Neste caso, ao contrário da primeira abordagem, supõe-se que nenhum conjunto de genes relacionados à característica fenotípica de interesse é conhecido *a priori*.

3. EXPERIMENTOS E RESULTADOS

Atualmente, as ferramentas propostas pelo projeto Prosgen encontram-se em fase de construção, com os seguintes resultados parciais: o conjunto de resumos de artigos científicos relacionados ao organismo *b.taurus* e as respectivas anotações GeneRIF foram baixados do Pubmed e armazenadas em um banco de dados local. A partir destes dados, foram gerados documentos em formato XML para realização de pré-processamento utilizando uma versão beta da ferramenta TaxEdit (Moural et al., 2010). Esta ferramenta permite que padrões relativos aos assuntos ou tópicos tratados pelos

documentos sejam evidenciados. No presente estágio do projeto, ainda não foram incorporadas as informações de vocabulário controlado, oriundos de bio-ontologias e, embora os documentos obtidos possam ser associados com genes do organismo estudado, os algoritmos para análise, clusterização e priorização de genes candidatos ainda não foram implementados.
Para ilustrar o potencial da abordagem por mineração de textos na análise de conjuntos de genes, analisou-se o conjunto de documentos relacionados a um conjunto de genes diferencialmente expressos em um experimento que comparou as expressões de grupos de bovinos resistentes e suscetíveis a carrapatos (Ibelli et al., 2010).
Os resultados apresentados na Figura 1 evidenciam alguns dos padrões extraídos dos documentos associados a este conjunto de documentos (descrição de genes). Eles evidenciam a existência de genes associados a diferentes famílias, como as chemoquinas, proteínas de membrana localizadas na superfície molecular e envolvidas com a interação com outras células (*cell adhesion molecule*); e genes envolvidos com o mecanismo de controle da concentração de cálcio na célula. Todos esses genes são relevantes para a resposta de bovinos à infestação de carrapatos, sendo que é conhecido o envolvimento de todos eles com a resposta do sistema imune a processos inflamatórios.

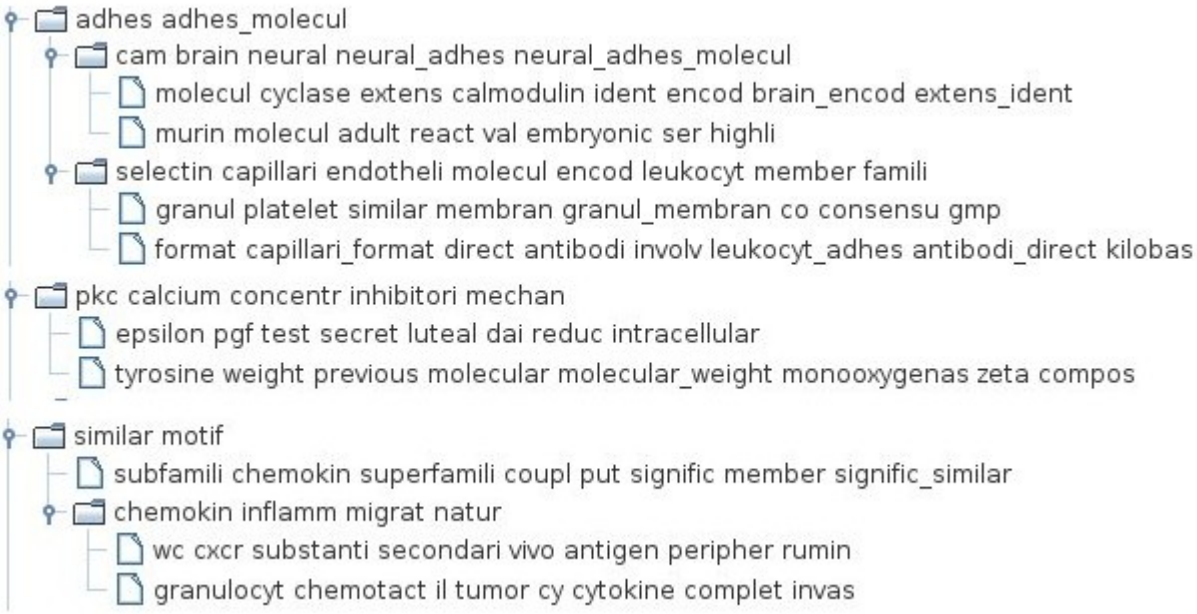

Figura 1: Ilustração de resultado de utilização de mineração de textos na análise de genes diferencialmente expressos entre bovinos resistentes e suscetíveis a carrapatos.

4. CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, apresentou-se o estágio atual do projeto Prosgen, que introduz a utilização de mineração de textos na análise e priorização desses genes para inclusão em programas de melhoramento genético animal e vegetal.
No experimento apresentado, analisou-se um conjunto de documentos relacionados aos genes diferencialmente expressos em um experimento de resistência a carrapatos em bovinos. Os padrões extraídos (tópicos que caracterizam as descrições funcionais dos genes) evidenciam a presença de genes envolvidos com a resposta do sistema imune a processos inflamatórios, o que demonstra a capacidade das técnicas utilizadas em capturar a função biológica dos genes analisados.
Embora, os resultados apresentados ainda sejam parciais, pois muitas das técnicas que se planeja implementar ainda se encontram em processo de experimentação e validação; eles permitem inferir a factibilidade de se capturar de textos a funcionalidade biológica de genes candidatos, que é o primeiro passo para que se possa comparar genes funcionalmente e, por fim, priorizá-los.

5. REFERÊNCIAS BIBLIOGRÁFICAS

AERTS, E.A.; ADAMS, R.R.; EVANS, K.L.; PORTENOUS, D.J.; PICKARD, B.S. SUSPECTS: enabling fast and effective priorization of positional candidates. Bioinformatics, v. 22, n. 6, january, 2006, 773-774.

ANANIADOU, S.; MCNAUGHT, J. (Eds) Text Mining for Biology and Biomedicine. Artech House, Inc. Norwood, MA. 2006.

DE BIE, T.; TRANCHEVENT, L.C.; VAN OEFFELEN, L.M.M.; MOREAU, Y. Kernel-based data fusion for gene priorization. Bioinformatics, v. 23 (ISMB/ECCB 2007), 2007, i125-i132.

EMBL-EBI; Interpro protein sequence analysis & classification. In: http://www.ebi.ac.uk/interpro/, acessado em maio de 2011.

EO. Environment Ontology. In: http://www.gramene.org/, acessado em maio de 2011.

GO; The Gene Ontology. In: www.geneontology.org, acessado em maio de 2011.

GRAMENE; Gramene search. In: http://www.gramene.org/db/ontology/search?id=TO:0000387, acessado em maio de 2011.

IBELLI, A. M. G.; HIGA, R. H.; GIACHETTO, P. F.; YAMAGISHI, M. E. B.; OLIVEIRA, M. C. S.; CARDOSO, F. F.; ALENCAR, M. M.; REGITANO, L. C. A. Genes e vias metabólicas envolvidos nos mecanismos de resistência e susceptibilidade de bovinos infestados com carrapato Rhipicephalus microplus. In: CONGRESSO BRASILEIRO DE GENÉTICA, 56., 2010, Guarujá. Resumos. Ribeirão Preto: Sociedade Brasileira de Genética, 2010. p. 74.

KEGG; The Kyoto Encyclopedia of Genes and Genomes. In: http://www.genome.jp/kegg/, acessado em maio de 2011.

MP. Mammalian Phenotype Ontology. In: http://www.informatics.jax.org/searches/MP_form.shtml, acessado em maio de 2011.

MOURA, M. F.; NOGUEIRA, B. M.; CONRADO, M. S.; SANTOS, F. F.; Rezende, S. O. Making Good Choices of Non-Redundant N-gram Words. In: INTERNATIONAL WORKSHOP ON DATA MINING AND ARTIFICIAL INTELLIGENCE, 1.; INTERNATIONAL CONFERENCE ON COMPUTER AND INFORMATION TECHNOLOGY, 11., 2008, Khulna, Bangladesh. Khulna, Bangladesh: IEEE Faculty of Electrical and Electronic Engineering and Khulna University of Engineering and Technology, 2008. v. 1. p. 64-71.

MOURA, M. F.; MERCANTI, E.; PEIXOTO, B. M.; MARCACINI, R. M. TaxEdit - Taxonomy Editor. Versão 1.0. Campinas: Embrapa Informática Agropecuária, 2010. 1 CD-ROM.

NCBIa; Entrez Gene. In: http://www.ncbi.nlm.nih.gov/gene/, acessado em maio de 2011.

NCBIb; PubMed. In: http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed&itool=toolbar, acessado em maio de 2011.

NCBIc; Entrez Gene. In: ftp://ftp.ncbi.nih.gov/gene/, acessado em maio de 2011.

NLM; MELINE. In: http://www.nlm.nih.gov/pubs/factsheets/medline.html, acessado em maio de 2011.

PATO. Phenotipic Quality Ontology. In: http://obofoundry.org/wiki/index.php/PATO:Main_Page, acessado em maio de 2011.

PO. Plant Ontology. In: http://www.plantontology.org/, acessado em maio de 2011.

TAX, D.M.J. One-class classification; Concept-learning in the absence of counter-examples, Ph.D. thesis Delft University of Technology, ASCI Dissertation Series, 65, Delft, 2001, June 19, 1-190.

VT. Vertebrate Trait Ontology. In: http://www.genome.iastate.edu/cgi-bin/amion/browse.cgi, acessado em maio de 2011.

YU, S.; TRANCHEVENT, L.C.; MOOR, B.D.; MOREAU, Y. Gene priorization and clustering by multi-view text mining, v. 11:28, 2010, http://biomedcentral.com/1471-2105/11/28.