Programa de Pós Graduação em Modelagem Computacional

Defesa de Tese de Doutorado em Modelagem Computacional

DATA: 26/11/2015
HORÁRIO: 13h
LOCAL: Auditório 2 do Prédio Engenheiro Itamar Franco/Faculdade de Engenharia

"Um método para seleção de atributos em dados genômicos"

Doutorando: Fabrízzio Condé de Oliveira
Orientador: Prof. Carlos Cristiano Hasenclever Borges

BANCA EXAMINADORA:
Prof.Carlos Cristiano Hasenclever Borges (Presidente/Orientador) – UFJF, D. Sc.
Prof. Wagner Antônio Arbex (Coorientador) - UFJF, D. Sc.
Prof.ª Priscila Vanessa Zabala Capriles Golliat - UFJF, D. Sc.
Prof. Raul Fonseca Neto, UFJF, D.Sc.
Prof. Fabyano Fonseca e Silva- UFV, D. Sc.
Prof. Moysés Nascimento - UFV, D. Sc.

RESUMO:
Estudos de associação em escala genômica buscam encontrar marcadores moleculares do tipo SNP que estão associados direta ou indiretamente ao fenótipo em questão, o qual pode ser uma doença ou uma característica benéfica. O SNP pode ser a própria mutação causal ou pode estar correlacionado com a mesma por serem herdados juntos. Para possibilitar a captura da região que possui a mutação causal, a qual não é conhecida *a priori*, milhares ou milhões de SNPs são genotipados em amostras compostas de centenas ou milhares de indivíduos. Com isso, surge o grande desafio de selecionar os SNPs mais informativos no conjunto de dados de genótipo-fenótipo onde o número de atributos é muito superior ao número de indivíduos, além da existência de atributos altamente correlacionados, podendo existir também interações entre pares, trios ou combinações de SNPs de quaisquer ordens. Os métodos mais usados em estudos de associação em escala genômica utilizam o valor-p de cada SNP em testes estatísticos de hipóteses baseados em regressão para fenótipos contínuos e baseados nos testes Qui-Quadrado ou similares em classificação para fenótipos binários como filtro para selecionar os SNPs mais significativos. Entretanto, esses métodos capturam somente SNPs com efeitos aditivos, pois a relação adotada é linear. Na tentativa de superar as dificuldades anteriores, este trabalho propõe um novo método de seleção de SNPs baseado em técnicas não-paramétricas de aprendizado de máquina e inteligência computacional denominado SNP *Markers Selector* (SMS). O método foi construído a partir da aplicação sequencial de *Random Forests*, *Support Vector Machine* e Algoritmos Genéticos com o objetivo de capturar efeitos aditivos e/ou não-aditivos moderados com interações entre pares e trios de SNPs, ou, até mesmo, para interações de ordens superiores com efeitos suficientemente grandes, sem especificar a quantidade de SNPs interagindo, nem o modelo matemático referente ao tipo interação e sem premissas sobre a distribuição dos dados de genótipo e fenótipo, podendo ser usado para problemas de regressão e classificação. O método foi aplicado em sete conjuntos de dados simulados e em uma base de dados real fornecida pela Embrapa, onde a produção de leite foi medida como fenótipo contínuo. O método proposto foi comparado com os métodos baseados no valor-p e com o Lasso Bayesiano apresentando, de forma geral, melhores resultados nos dados simulados com efeitos aditivos juntamente com interações entre pares e trios de SNPs. Nos dados reais, o método identificou 245 QTLs associados à produção e à composição do leite e 90 genes candidatos associados à mastite, à produção e à composição do leite, sendo esses QTLs e genes identificados por estudos anteriores utilizando outros métodos de seleção. Assim, o método demonstrou ser competitivo frente aos métodos utilizados para comparação em cenários complexos, com dados simulados ou reais, o que indica seu potencial para estudos de associação em escala genômica em humanos, animais e vegetais.