SP 7184
2016
SP-PP-SP 7184

Sodebras

Volume 11 - N° 129 - Setembro/ 2016
XXXV International Sodebras Congress
01 a 03 de junho de 2016 - Foz do Iguaçu - PR

REGRESSÃO COM MÁQUINAS DE VETORES SUPORTE E SELEÇÃO DE ATRIBUTOS VIA ALGORITMO GENÉTICO APLICADA EM SELEÇÃO GENÔMICA

BRUNO ZONOVELLI[1]; CARLOS CRISTIANO HASENCLEVER BORGES[1]; WAGNER ANTONIO ARBEX[1, 2]; FABRIZZIO CONDÉ DE OLIVEIRA[3]; IGOR MAGALHÃES RIBEIRO[1]
1 – UNIVERSIDADE FEDERAL DE JUIZ DE FORA - UFJF; 2 – EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA – EMBRAPA; 3 – UNIVERSIDADE SALGADO DE OLIVEIRA - UNIVERSO.
wagner.arbex@embrapa.br

***Resumo** - A seleção genômica busca predizer os valores fenotípicos dos indivíduos através de modelos. O processo de construção desses modelos é feita com a definição da população de referência, a escolha da ferramenta e a montagem do mesmo, contudo alguns fatores podem dificultar a obtenção de um modelo preciso. Nesse trabalho foi analisado o impacto da presença de epistasia, ou seja, a interação entre os marcadores ou variáveis bem como o comportamento do SVR (ferramenta de predição) quando apresentada a uma amostra pequena. Também foi alvo de estudo a consequência da seleção de atributos. Os resultados mostram que o uso da seleção de atributos trouxe melhorias na obtenção de modelos mais eficientes utilizando o SVR.*

***Palavras-chave:** Bioinformática. Seleção Genômica. Aprendizado de Máquina, Inteligência Computacional.*

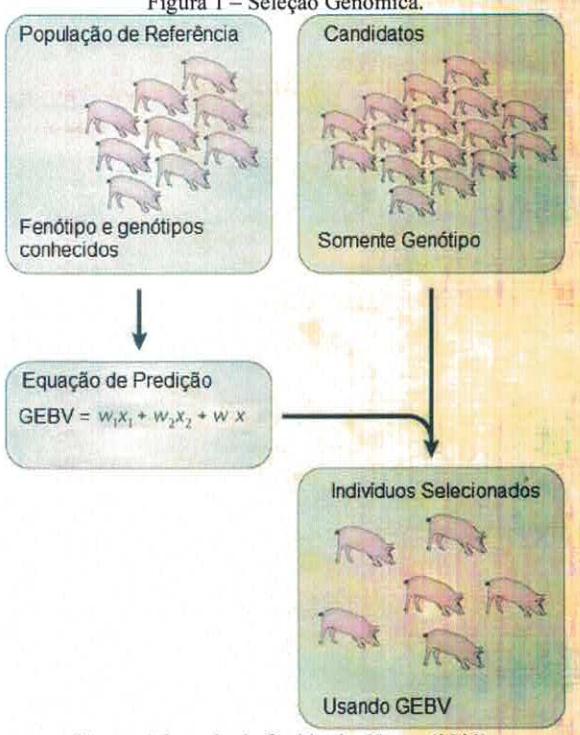
I. INTRODUÇÃO

O melhoramento genético animal ou vegetal consiste, basicamente, em um conjunto de técnicas e métodos com a finalidade de melhorar o valor genético médio da população de interesse. O método utilizado consiste em selecionar e acasalar indivíduos superiores de forma que suas características sejam transmitidas às gerações futuras.

A seleção dos mais aptos consiste em definir quais indivíduos serão os progenitores da próxima geração, o que visa aumentar os genes desejáveis nas gerações futuras. Contudo, o processo de seleção pode ser lento e custoso, nesse ponto entra a seleção genômica, que consiste em construir um modelo para calcular o valor genômico previsto (do inglês, *genomic estimated breeding values* - GEBV) dos animais a serem avaliados.

A predição do valor genômico de um animal permite selecioná-lo no início da vida, ou antes, mesmo do seu nascimento (SCHAEFFER, 2006). O GEBV da próxima geração é calculado tendo como base uma população de referência, dessa forma é possível efetuar a seleção via dados genômicos (MEUWISSEN; HAYES; GODDARD, 2001; HAYES; GODDARD, 2010). A Figura 1 mostra, de forma geral, o processo de seleção genômica onde o GEBV é dado pelo somatório dos efeitos de cada marcador, o que permite a predição do valor genômico da população futura (GEBV = $w1x1 + w2x2 + w3x3...$), sendo w o efeito do marcador e x o seu valor genômico (MEUWISSEN; HAYES; GODDARD, 2001). A redução do custo operacional para a obtenção dos marcadores aliada à precisão dos GEBV obtidos levou a rápida adoção da seleção genômica por parte das empresas e profissionais do meio (SCHAEFFER, 2006).

Figura 1 – Seleção Genômica.



Fonte: Adaptada de Goddard e Hayes (2009).

O sucesso na seleção genômica depende, em geral, de três itens: o tamanho da população de referência; a herdabilidade; e o tamanho do desiquilíbrio de ligação entre os marcadores e o *locus* de características quantitativas (do inglês, *Quantitative Trait Locus* - QTL) (GODDARD; HAYES, 2009). Em geral, a seleção genômica tem por objetivo estimar com precisão o GEBV e, para isso, necessita de um grande conjunto de dados para o treinamento, conforme sugerem Meuwissen, Hayes e Goddard (2001), Hayes e Goddard (2010). A precisão do GEBV, segundo a fórmula de Daetwyler, Villanueva e Woolliams (2008), é diretamente proporcional à hereditariedade, onde traços com maior herdabilidade geram GEBVs mais precisas do que aqueles com herdabilidades

ISSN 1809-3957

SP 7184

menores. O aumento do desequilíbrio de ligação entre os marcadores e o QTL, segundo Goddard e Hayes (2009), também geram uma maior precisão no cálculo do GEBV.

A eficiência do processo de seleção genômica depende da correta identificação dos indivíduos geneticamente superiores. Muitos fatores podem dificultar a construção de um modelo eficiente para o cálculo do GEBV. Diversos são os desafios encontrados no processo de seleção genômica. Neste trabalho foram analisados, inicialmente, dois tópicos: o impacto da redução do tamanho da amostra ou população de referência; e a dificuldade da obtenção de um modelo para predição quando existe epistasia ou interação entre os genes (GODDARD; HAYES, 2009; HAYES *et al.*, 2009).

Algumas populações podem não possuir o tamanho necessário para satisfazer os requisitos mínimos para a obtenção de resultados precisos, como os presentes em raças de grande porte. Deste modo, torna-se necessário o estudo do impacto das populações pequenas em seleção genômica.

Em estudos iniciais, Mészáros *et al.* (2015) utilizou três bases de dados pequenas e discutiu o efeito da junção das mesmas, obtendo uma melhora de aproximadamente 10% em uma delas.

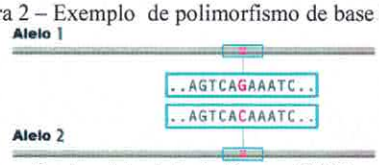
As células possuem pares de genes, sendo uma cópia da mãe e outra do pai. Cada gene possui uma sequência de DNA que é conhecida como alelos. A relação entre os genes é chamada de epistasia, que consiste na interação entre genes de diferentes locos. A interação entre alelos é conhecida como dominância ou recessividade.

O estudo com epistasia é definido por Hayes *et al.* (2009) como um dos desafios para a seleção genômica. Ele tem sido utilizado para justificar uma série de fenômenos, tais como, a interação funcional entre os genes, o resultado de mutações genéticas que atuam dentro da mesma via metabólica e o desvio estatístico da ação aditiva (PHILLIPS, 2008).

A combinação desses diferentes desafios em um mesmo conjunto de dados pode vir a dificultar a obtenção de um GEBV mais preciso. Neste sentido, este trabalho visa avaliar se a identificação e seleção de atributos largamente informativos, associado ao uso de técnicas de inteligência computacional, podem influenciar de forma positiva na melhora da predição do valor genômico.

Em geral, as "regras" que regem o estudo do genoma podem ser aplicadas a qualquer espécie viva, diferenciando-se apenas os organismos procariotos dos eucariotos. Uma das muitas variações e particularidades do genoma, humano ou de qualquer espécie, são os polimorfismos de base única (do inglês, *Single Nucleotide Polymorphisms* - SNPs). Os SNPs são modificações de um único nucleotídeo, em uma dada sequência, quando comparada a outra (Figura 2). Ou seja, são pares de bases em uma única posição no DNA genômico, que se apresentam com diferentes alternativas nas sequências, em uma porção significativa da população, ou seja, $\geq$ 1%, e podem ser encontrados no genoma de indivíduos normais em algumas populações ou grupos (HAPMAP, 2003). São encontrados em vasto número e em qualquer genoma (BROWN, 2006).

Figura 2 – Exemplo de polimorfismo de base única.



Fonte: Adaptado de Brown (2006)

Assim, tais diferenças são importantes no estudo da variabilidade das espécies, pois podem provocar alterações funcionais ou fenotípicas, que, por sua vez, podem implicar em consequências evolutivas ou bioquímicas nos indivíduos em que os SNPs se manifestam (ARBEX, 2009).

II. MÉTODO PROPOSTO

O método proposto consiste em uma combinação eficiente entre um processo de seleção de características e um procedimento para a determinação do GEBV com o máximo de acurácia. Como visto na introdução o cálculo do GEBV consiste no somatório dos marcadores e seus pesos, Equação 1. Nesse trabalho usaremos a Regressão com Máquina de Vetores Suporte (do inglês, *Support Vector Regression* - SVR) para esse calculo, principalmente porque ele pode ou não ter um comportamento linear, assim sendo o calculo do GEBV será dado por uma função genérica $f(x)$, Equação 2.

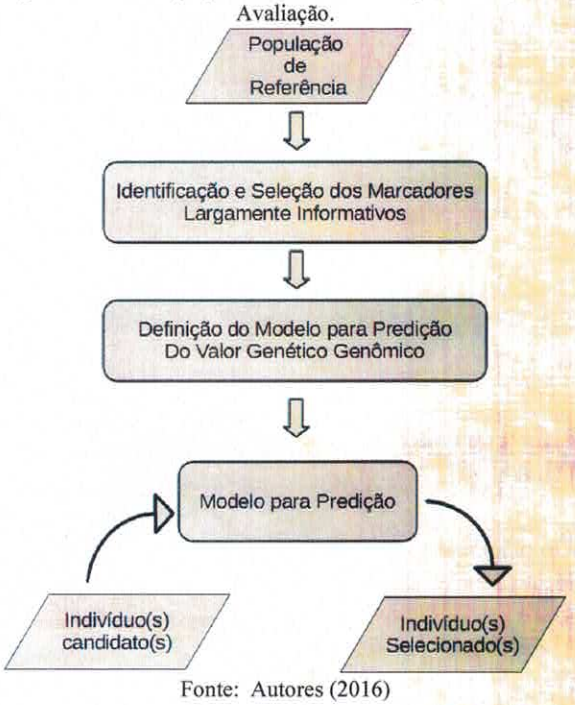
$$GEBV = \sum_{j=1}^{n} w_j x_j \quad (1)$$

$$GEBV = f(x) \quad (2)$$

onde n é igual ao número de marcadores.

A Figura 3 mostra o método proposto que consiste em duas etapas a de seleção e a de avaliação. A primeira fase seleciona as variáveis mais relevantes por meio de seleção de atributos via *SNP Markers Selector* (SMS). A segunda etapa é a avaliação de cada subconjunto pelo SVR, comparando-os com o grupo completo de marcadores e o grupo com somente os SNPs causais. A seleção genômica busca o modelo com maior correlação ou acurácia, contudo a dimensão dos dados pode impactar diretamente na capacidade de obtenção de um modelo eficiente. Os *chips* de genotipagem atuais possuem um número elevado de marcadores, muitas vezes maior que a quantidade de indivíduos na amostra. A seleção dos atributos mais relevantes visa diminuir o problema da dimensionalidade e facilitar a obtenção de um modelo mais eficaz.

Figura 3 - Método proposto em suas duas etapas a Seleção e a Avaliação.



Fonte: Autores (2016)

ISSN 1809-3957

Em muitas tarefas de classificação ou regressão, o número total de possíveis atributos associados às instâncias que definem a base de dados é relativamente alta (STANCZYK; JAIN, 2015). Esta alta dimensionalidade tende a dificultar o processamento, ou até mesmo torná-lo impraticável.

A seleção de atributos em seleção genômica, em geral, consiste em calcular o impacto de cada marcador e após ordená-los, cada um então é incluído no modelo e o subgrupo escolhido é que possui maior acurácia. Esse modelo pode não ser capaz de capturar o efeito dos marcadores quando a ação gênica não for aditiva. Long *et al.* (2011) em seu trabalho utilizou técnicas de redução da dimensão multivariadas, como regressão componentes principais (PCR) e regressão de mínimos quadrados parciais (PLS), obtendo um aumento de acurácia com a seleção de um subconjunto. O trabalho de Granato et al. (2013) afrouxou o critério de seleção do BLASSO para definir o subconjunto e avaliou o impacto de cada um, escolhendo o subconjunto com maior acurácia, obtendo novamente uma melhora no resultado com a seleção de um subconjunto. De forma semelhante Usai, Carta e Casu (2012) utilizou uma técnica para selecionar subconjuntos utilizando o LASSO obtendo novamente resultados interessantes com essa seleção.

A seleção de atributos neste trabalho será feito de forma diferente visando obter o menor subgrupo com maior correlação e para isso será utilizado a metodologia SMS (SNP *Marker Selector*). O método SMS, cuja tradução livre é Seletor de Marcadores SNP, teve sua primeira versão publicada por Oliveira et al. (2014a) Oliveira et al. (2014b).

O método busca combinar técnicas da inteligência computacional de forma otimizada, para selecionar os marcadores SNPs mais informativos para um dado fenótipo considerando efeitos isolados e de interação entre os SNPs. Na versão atual, o SMS utiliza: Floresta Randômica (do inglês, *Random Forests* - RF), Máquinas de Vetores Suporte (do inglês, *Support Vector Machine* - SVM) e Algoritmos Genéticos (do inglês, *Genetic Algorithms* - GA). Cada técnica foi combinada de forma a se obter o máximo de eficiência em cada etapa. O SMS combina a seleção de atributos por filtro e por encapsulamento em etapas distintas e complementares.

A primeira etapa utilizada é o método de seleção de atributos baseado em filtro, por ser menos custoso computacionalmente. Ela consiste em utilizar a RF para ordenar os marcadores por sua relevância em relação ao fenótipo, em seguida é utilizado o SVM/SVR que incrementa o conjunto de teste de n em n elementos (em geral, $n = 10$), selecionando o subconjunto com menor erro. Essa primeira seleção só é possível, porque, em geral, o número de marcadores causais e bem menor que o sequenciado. A segunda seleção é um encapsulamento combinando o GA com o SVM/SVR, onde o GA seleciona o melhor subconjunto seguindo a avaliação feita pelo SVM/SVR. O resultado final do SMS é o menor subconjunto de marcadores com a maior informação relativa ao fenótipo estudado.

A etapa de avaliação do método proposto é executada após a seleção dos SNPs mais informativos para o fenótipo e consiste em utilizar o SVR. Bem como, avaliar o impacto da redução da população de referência associada à possível presença de epistasia entre determinados genes, os quais são marcados pelos SNPs.

O SVM é uma técnica de aprendizado supervisionado que analisa padrões entre os dados de entrada, caraterizados por variáveis numéricas contínuas ou discretas, com os dados de saída, designados por um atributo dicotômico (problema de classificação). Esse modelo foi desenvolvido por Cortes e Vapnik (1995) e é baseado na ideia de encontrar o hiperplano ótimo que separa as duas classes por meio da maximização da margem. A primeira versão do SVM com regressão foi proposta em 1997 por Drucker et al. (1997), e foi denominada como Regressão com Máquina de Vetores Suporte (SVR – *Support Vector Regression*). Dentre as vantagens do SVR, vale citar que este método não pressupõe linearidade do modelo, desde que se adote função *kernel* não-linear, não necessita de normalidade dos resíduos e adapta-se facilmente a dados de alta dimensionalidade (número de instâncias menor que o número de atributos). O *kernel* é uma função K tal que para todo $x;z \in X$ satisfaz $K(x,z) = \langle \phi(x), \phi(z) \rangle$ onde ϕ é uma função de X para um espaço de características com produto interno F, onde $\varphi x \in X \mapsto \varphi\ x\ \in F$. A dimensão do espaço F é superior à do X ($n > p$), pois o objetivo é aumentar a probabilidade de separação entre as classes pelo hiperplano ótimo. Para padrões que não sejam linearmente separáveis, ϕ é não linear.

O SVR foi implementado utilizando o software R (R Core Team, 2015), e os pacotes e1071 de Meyer et al. (2014), além dos pacotes doParallel e foreach para o processamento paralelo (ANALYTICS; WESTON, 2014a; ANALYTICS; WESTON, 2014b). O kernel utilizado foi o Gaussiano, gama = 0,01, Custo = 1,0 e Épsilon = 0,1.

III. DADOS SIMULADOS

Esta secção descreve como os dados utilizados nos experimentos computacionais foram simulados, e as características de cada conjunto de dados. Os estudos foram feitos com dados simulados, pois permitem avaliar de forma eficaz a metodologia aplicada, bem como a eficiência da seleção em relação aos falsos positivos e negativos.

O simulador utilizado foi desenvolvido por Schwender (2007) como um pacote para o software R e é conhecido como SCRIME. A sigla SCRIME faz referência a (*Statistical Complexity Reduction In Molecular Epidemiology*), que em tradução livre é redução da complexidade estatística em epidemiologia molecular.

Os dados são simulados a partir de uma matriz *NxM* com N observações e M SNPs. Outro parâmetro a ser informado é a *Minor Allele Frequency* (MAF) de cada SNP. Os SNPs são simulados de forma independente, logo estão ligados (em desequilíbrio de ligação). Um modelo de regressão é utilizado para determinar o fenótipo, de acordo com os parâmetros definidos na lista de SNP, e a lista de preferências, bem como os valores do *beta* para cada SNP e do *beta0* (efeito fixo do modelo) da regressão. O modelo de regressão depende do fenótipo de interesse, sendo escolhida a regressão logística para fenótipos binários e a linear para contínuos. Para fenótipos binários o valor é obtido por sorteio, a partir de uma distribuição Bernoulli (SCHWENDER, 2007; NUNKESSER *et al.*, 2007).

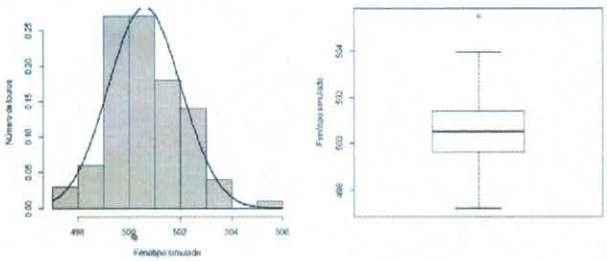
A seguir será explicado como cada conjunto de dados foi simulado, os parâmetros utilizados, bem como o teste de normalidade de Shapiro e Wilk (1965).

ISSN 1809-3957

O efeito aqui analisado é o de interação entre os marcadores, de forma que a contribuição de cada marcador pode ser dada de forma isolada ou em conjunto com outros, de acordo com os parâmetros da simulação. O efeito simulado nos conjuntos de dados a seguir foi a epistasia. Foram simulados dados com 100, 500 e 1000 indivíduos e 100 e 2000 marcadores. Os SNPs causais são os 1,2,3,4 e 5 com uma interação entre o 1 e 2, bem como outra entre o 4 e 5. Os efeitos ou *betas* utilizados foram 2, 1,2 e 1,5 respectivamente. A MAF gerada para cada SNP é simulada por meio de uma distribuição contínua uniforme com mínimo igual a 0,1, e máximo 0,4. E uma função de erro normal com média 1 e desvio padrão 0. Com esses parâmetros foram gerados seis conjuntos de dados simulados.

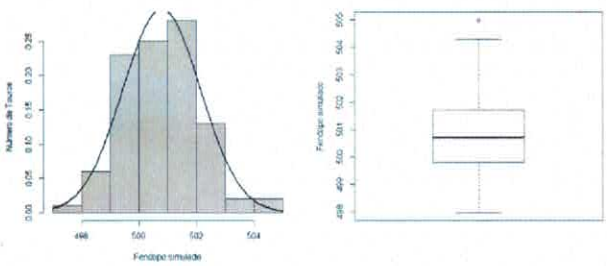
As Figuras 4 e 5 mostram a distribuição para os dados simulados com 100 indivíduos e 100 e 2000 marcadores respectivamente. Aplicando o teste de Shapiro e Wilk nos conjuntos de dados com 100 indivíduos obtém-se um W = 0,9847 e valor-p = 0,2997 para o conjunto com 100 marcadores, e W = 0,9871, valor-p = 0,4444 para o com 2.000 marcadores.

Figura 4 – Base com 100 indivíduos e 100 marcadores.



Fonte: do Autores (2016)

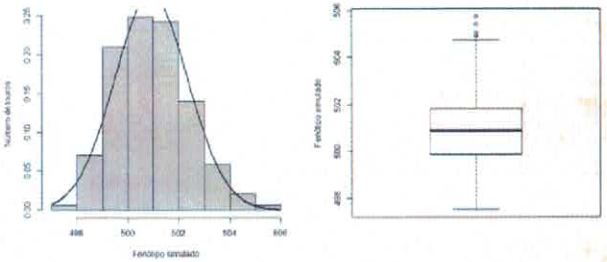
Figura 5 – Base com 100 indivíduos e 2000 marcadores.



Fonte: do Autores (2016)

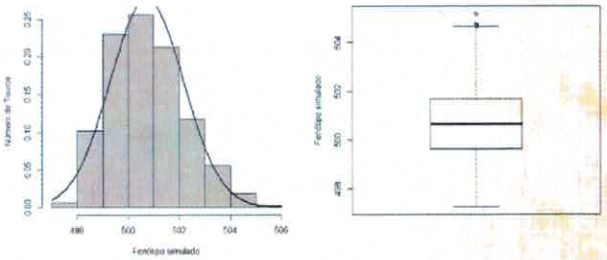
As Figuras 6 e 7 exibem a distribuição para os dados simulados com 500 indivíduos e 100 e 2000 marcadores respectivamente. O teste de normalidade para o conjunto com 500 indivíduos obteve W = 0,9881 e valor-p = 0,0004133 para 100 e W = 0,9884, valor-p = 0,0005262 para 2.000.

Figura 6 – Base com 500 indivíduos e 100 marcadores.



Fonte: do Autores (2016)

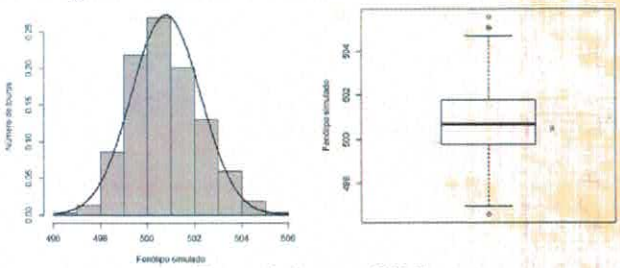
Figura 7 – Base com 500 indivíduos e 2000 marcadores.



Fonte: do Autores (2016)

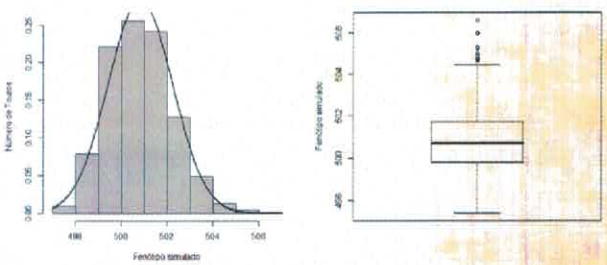
As Figuras 8 e 9 mostram a distribuição para os dados simulados com 1000 indivíduos e 100 e 2000 marcadores respectivamente. Os testes de normalidade dos conjuntos com 1.000 indivíduos resultaram em W = 0,9926 e valor-p = 7,08e-05 para 100 marcadores e W = 0,9887, valor-p = 5,536e-07 para o com 2.000.

Figura 8 – Base com 1000 indivíduos e 100 marcadores.



Fonte: do Autores (2016)

Figura 9 – Base com 1000 indivíduos e 2000 marcadores.



Fonte: do Autores (2016)

ISSN 1809-3957

IV. RESULTADOS

Esta seção exibe os resultados obtidos com a execução dos experimentos computacionais. Cada conjunto de dados foi submetido à seleção pelo SMS e cada subconjunto encontrado foi avaliado pelo SVR. Os resultados de cada etapa são mostrados a seguir.

5.1 – Etapa de Seleção

A Tabela 1 apresenta o resultado obtido pelo SMS na etapa de seleção, como é possível verificar o aumento no número de indivíduos gerou uma melhoria na acurácia (ACC). O SMS encontrou todos os verdadeiros positivos (VP), com exceção da população (POP) com 100 indivíduos. A taxa de falso positivo (FP) caiu com o aumento populacional. Os resultados obtidos pelo SMS mostram que a ferramenta é um pouco permissiva, pois permitiu a seleção de alguns falsos positivos. Mesmo não obtendo uma seleção ótima, a ferramenta mostrou-se robusta com a variação da população.

Tabela 1 – Resultados da seleção obtida com o software SMS no conjunto com 100 marcadores.

POP	Total	VP	FP	ACC
100	8	2	6	0,91
500	13	5	8	0,92
1000	8	5	3	0,97

A Tabela 2 mostra o resultado obtido pelo SMS no conjunto com 2000 marcadores, onde é possível verificar que o comportamento na população com 1000 indivíduos é muito diferente nos conjuntos com 100 e 500 indivíduos. O aumento populacional torna a seleção mais próxima do ótimo. O SMS somente conseguiu selecionar os 5 SNPs causais com uma população de 1000 indivíduos, mostrando que o aumento no número de marcadores levou a uma maior complexidade na obtenção de uma seleção efetiva.

Tabela 2 – Resultados da seleção obtida com o software SMS no conjunto com 2000 marcadores.

POP	Total	VP	FP	ACC
100	100	2	98	0,950
500	34	4	30	0,990
1000	9	5	4	0,998

Os valores de acurácia da seleção não sofreram alterações nítidas com o aumento da dimensionalidade, contudo é necessário avaliar esse impacto na etapa de avaliação, pois com o aumento no número de marcadores e com uma presença maior de falsos positivos o modelo de seleção genômica pode ou não ser o mais eficiente.

5.2 – Etapa de Avaliação

O objetivo dessa etapa é verificar se a seleção de atributos gerou melhoria na correlação entre os marcadores e o fenótipo. O uso das bases de dados simulados permite conhecer os SNPs causais *a priori*, de forma que na etapa de seleção essa capacidade foi avaliada. Contudo, vale ressaltar que a melhor "seleção" pode não ser o subgrupo com maior correlação. Com o objetivo de se obter uma medida de correlação mais confiável, foi utilizado o processo de validação cruzada com 10-*folds*, dessa forma a correlação obtida será mais genérica que o método *holdout*, o qual utiliza somente um subconjunto para o treino e o complementar do treino para o teste. Logo, o resultado final será a média e o desvio padrão das 10 correlações obtidas a partir dos 10 *fold*.

A Tabela 3 e a Tabela 4 mostram os resultados obtidos pelo SVR nos conjuntos de dados com efeito de interação entre os SNPs causais. A melhor seleção não obteve a maior correlação, mostrando a dificuldade de se fazer uma boa seleção em populações pequenas. Vale ressaltar que apesar da presença de falsos positivos no subconjunto selecionados as correlações obtidas são maiores que a encontrada no grupo com somente os SNPs causais.

Tabela 3 – Correlações obtidas pelo SVR utilizando os subconjuntos com efeito de interação e 100 marcadores.

POP	Inicial	VP	SMS
100	-0,04(0,30)	0,42(0,38)	0,60(0,23)
500	0,43(0,15)	0,57(0,09)	0,59(0,09)
1000	0,41(0,04)	0,57(0,07)	0,58(0,07)

A presença de epistasia dificulta a obtenção de um modelo eficiente para a seleção genômica, de forma que existe uma diferença entre a acurácia obtida pelo conjunto inicial e o maior valor entre os subconjuntos. A seleção de atributos se mostra com uma possível solução, pois consegue aumentar a acurácia da ferramenta sem a necessidade de um aumento na população de referência.

Tabela 4 – Correlações obtidas pelo SVR utilizando os subconjuntos com efeito de interação e 2000 marcadores.

POP	Inicial	VP	SMS
100	-	0,45(0,22)	0,83(0,06)
500	-	0,54(0,08)	0,58(0,07)
1000	-	0,61(0,04)	0,61(0,04)

A acurácia nos conjuntos iniciais (Tabela 4) não foi calculada pelo SVR, pois o mesmo ao utilizar o conjunto completo calculou um GEBV igual para todas as entradas, dessa forma é possível dizer que a seleção de atributos se torna uma necessidade para que a ferramenta tenha um funcionamento eficiente, com os parâmetros utilizados. O SVR permite a variações de múltiplos parâmetros e *kernels*. O *kernel* gaussiano e os parâmetros utilizados não conseguiram obter um modelo eficiente no conjunto de dados analisados.

A análise dos conjuntos com somente 100 marcadores geraria a falsa conclusão de que a seleção de atributos é benéfica somente para populações pequenas, porém quando há um aumento no número de marcadores a seleção se torna interessante mesmo para populações com 1000 indivíduos.

V. CONCLUSÃO

A seleção genômica é um campo de estudo amplo e com muitos desafios. Nesse trabalho foi avaliado o impacto na diminuição populacional e da presença de epistasia entre os SNPs causais. O método proposto apresentou resultados promissores nas bases com as características descritas gerando um aumento na correlação final. O SMS manteve um comportamento estável mesmo com as variações no tamanho da população de referência e do número de marcadores.

A dimensionalidade das bases de dados mostrou-se um problema. Os conjuntos com 100 marcadores possuem 1 SNP causal a cada 20 e nos dados com 1000 tem-se 1 a cada 200 marcadores. Os chips atuais vão de 3000 até 1 milhão de marcadores o que aumenta ainda mais a

ISSN 1809-3957

dimensionalidade do dados, indicando a necessidade da utilização de métodos que sejam robustos na seleção desses marcadores.

VI. REFERÊNCIAS BIBLIOGRÁFICAS

ANALYTICS, R.; WESTON, S. **doParallel: Foreach parallel adaptor for the parallel package.** [S.l.], 2014. R package version 1.0.8. Disponível em: <http://CRAN.R-project.org/package=doParallel>.

ANALYTICS, R.; WESTON, S. **foreach: Foreach looping construct for R.** [S.l.], 2014. R package version 1.4.2. Disponível em: <http://CRAN.R-project.org/package=foreach>.

ARBEX, W. A. **Modelos Computacionais para Identificação de Informação Genômica Associada à Resistência ao Carrapato Bovino.** Tese (Doutorado) UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2009.

BROWN, T. A. **Genomes.** [S.l.]: Garland science, 2006.

CORTES, C.; VAPNIK, V. **Support-vector networks. Machine learning,** v. 20, n. 3, p. 273 - 297, 1995.

DAETWYLER, H. D.; VILLANUEVA, B.; WOOLLIAMS, J. A. **Accuracy of predicting the genetic risk of disease using a genome-wide approach**. PLoS One, v. 3, n. 10, p.e3395, 2008.

DRUCKER, H. et al. **Support vector regression machines. Advances in neural information processing systems,** Morgan Kaufmann Publishers, v. 9, p. 155 - 161, 1997.

GODDARD, M. E.; HAYES, B. J. **Mapping genes for complex traits in domestic animals and their use in breeding programmes.** Nature Reviews Genetics, Nature Publishing Group, v. 10, n. 6, p. 381 - 391, 2009.

GRANATO, I. et al. **Seleção de marcadores para os métodos rr-blup e blasso na seleção genômica ampla**. In: IN: CONGRESSO BRASILEIRO DE MELHORAMENTO DE PLANTAS, 7., 2013, UBERLÂNDIA. VARIEDADE MELHORADA: A FORÇA DA NOSSA AGRICULTURA: ANAIS. VIÇOSA, MG: SBMP, 2013. Embrapa Florestas-Artigo em anais de congresso (ALICE). [S.l.], 2013.

HAPMAP, C. I. **The international hapmap project**. Nature, v. 426, n. 6968, p. 789 - 96, 2003. Disponível em: <http://dx.doi.org/10.1038/nature02168>.

HAYES, B. et al. **Invited review: Genomic selection in dairy cattle: Progress and challenges**. Journal of dairy science, Elsevier, v. 92, n. 2, p. 433 - 443, 2009.

HAYES, B.; GODDARD, M. **Genome-wide association and genomic selection in animal breeding**. Genome, NRC Research Press, v. 53, n. 11, p. 876 - 883, 2010.

LONG, N. et al. **Dimension reduction and variable selection for genomic selection: application to predicting milk yield in holsteins**. Journal of Animal Breeding and Genetics, Wiley Online Library, v. 128, n. 4, p. 247 - 257, 2011.

MEYER, D. et al. e1071: **Misc Functions of the Department of Statistics (e1071)**, TU Wien. [S.l.], 2014. R package version 1.6-3. Disponível em: <http://CRAN.R-project.org/package=e1071>.

MÉSZÁROS, G. et al. **Genomic analysis for managing small and endangered populations: A case study in tyrol grey cattle**. Frontiers in Genetics, Frontiers, v. 6, p. 173, 2015.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. **Prediction of total genetic value using genome-wide dense marker maps**. Genetics, v. 157, n. 4, p. 1819 - 1829, 2001. Disponível em: <http://www.genetics.org/content/157/4/1819.abstract>.

NUNKESSER, R. et al. **Detecting high-order interactions of single nucleotide polymorphisms using genetic programming**. Bioinformatics, v. 23, n. 24, p. 3280 - 3288, 2007. Disponível em: <http://bioinformatics.oxfordjournals.org/content/23/24/3280.abstract>.

OLIVEIRA, F. C. de et al. **Metodologia para seleção de marcadores com máquina de vetores suporte com regressão**. In: . [S.l.]: Embrapa, 2014. p. 101 - 126. ISBN 978-85-7035-382-5.

OLIVEIRA, F. C. de et al. **Snps selection using support vector regression and genetic algorithms in gwas**. BMC genomics, BioMed Central Ltd, v. 15, n. Suppl 7, p. S4, 2014.

PHILLIPS, P. C. **Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems**. Nature Reviews Genetics, Nature Publishing Group, v. 9, n. 11, p. 855 - 867, 2008.

R Core Team. R: **A Language and Environment for Statistical Computing.** Vienna, Austria, 2015. Disponível em: <http://www.R-project.org/>.

SCHAEFFER, L. **Strategy for applying genome-wide selection in dairy cattle.** Journal of Animal Breeding and Genetics, Wiley Online Library, v. 123, n. 4, p. 218 - 223, 2006.

SCHWENDER, H. **Statistical Analysis of Genotype and Gene Expression Data.** Tese (Doutorado) the Department of Statistics of the University of Dortmund, 2 2007.

SHAPIRO, S. S.; WILK, M. B. **An analysis of variance test for normality (complete samples).** Biometrika, v. 3, n. 52, 1965.

STANCZYK, U.; JAIN, L. C. **Feature Selection for Data and Pattern Recognition.** [S.l.]: Springer, 2015.

USAI, M. G.; CARTA, A.; CASU, S. **Alternative strategies for selecting subsets of predicting snps by lasso-lars procedure.** In: BIOMED CENTRAL LTD. BMC proceedings. [S.l.], 2012. v. 6, n. Suppl 2, p. S9.

VII. COPYRIGHT

Direitos autorais: O(s) autor(es) é(são) o(s) único(s) responsável(is) pelo material incluído no artigo.

dimensionalidade do dados, indicando a necessidade da utilização de métodos que sejam robustos na seleção desses marcadores.

VI. REFERÊNCIAS BIBLIOGRÁFICAS

ANALYTICS, R.; WESTON, S. **doParallel: Foreach parallel adaptor for the parallel package.** [S.l.], 2014. R package version 1.0.8. Disponível em: <http://CRAN.R-project.org/package=doParallel>.

ANALYTICS, R.; WESTON, S. **foreach: Foreach looping construct for R.** [S.l.], 2014. R package version 1.4.2. Disponível em: <http://CRAN.R-project.org/package=foreach>.

ARBEX, W. A. **Modelos Computacionais para Identificação de Informação Genômica Associada à Resistência ao Carrapato Bovino.** Tese (Doutorado) UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2009.

BROWN, T. A. **Genomes.** [S.l.]: Garland science, 2006.

CORTES, C.; VAPNIK, V. **Support-vector networks. Machine learning,** v. 20, n. 3, p. 273 - 297, 1995.

DAETWYLER, H. D.; VILLANUEVA, B.; WOOLLIAMS, J. A. **Accuracy of predicting the genetic risk of disease using a genome-wide approach**. PLoS One, v. 3, n. 10, p.e3395, 2008.

DRUCKER, H. et al. **Support vector regression machines. Advances in neural information processing systems**, Morgan Kaufmann Publishers, v. 9, p. 155 - 161, 1997.

GODDARD, M. E.; HAYES, B. J. **Mapping genes for complex traits in domestic animals and their use in breeding programmes.** Nature Reviews Genetics, Nature Publishing Group, v. 10, n. 6, p. 381 - 391, 2009.

GRANATO, I. et al. **Seleção de marcadores para os métodos rr-blup e blasso na seleção genômica ampla**. In: IN: CONGRESSO BRASILEIRO DE MELHORAMENTO DE PLANTAS, 7., 2013, UBERLÂNDIA. VARIEDADE MELHORADA: A FORÇA DA NOSSA AGRICULTURA: ANAIS. VIÇOSA, MG: SBMP, 2013. Embrapa Florestas-Artigo em anais de congresso (ALICE). [S.l.], 2013.

HAPMAP, C. I. **The international hapmap project**. Nature, v. 426, n. 6968, p. 789 - 96, 2003. Disponível em: <http://dx.doi.org/10.1038/nature02168>.

HAYES, B. et al. **Invited review: Genomic selection in dairy cattle: Progress and challenges**. Journal of dairy science, Elsevier, v. 92, n. 2, p. 433 - 443, 2009.

HAYES, B.; GODDARD, M. **Genome-wide association and genomic selection in animal breeding**. Genome, NRC Research Press, v. 53, n. 11, p. 876 - 883, 2010.

LONG, N. et al. **Dimension reduction and variable selection for genomic selection: application to predicting milk yield in holsteins**. Journal of Animal Breeding and Genetics, Wiley Online Library, v. 128, n. 4, p. 247 - 257, 2011.

MEYER, D. et al. e1071: **Misc Functions of the Department of Statistics (e1071)**, TU Wien. [S.l.], 2014. R package version 1.6-3. Disponível em: <http://CRAN.R-project.org/package=e1071>.

MÉSZÁROS, G. et al. **Genomic analysis for managing small and endangered populations: A case study in tyrol grey cattle**. Frontiers in Genetics, Frontiers, v. 6, p. 173, 2015.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. **Prediction of total genetic value using genome-wide dense marker maps**. Genetics, v. 157, n. 4, p. 1819 - 1829, 2001. Disponível em: <http://www.genetics.org/content/157/4/1819.abstract>.

NUNKESSER, R. et al. **Detecting high-order interactions of single nucleotide polymorphisms using genetic programming**. Bioinformatics, v. 23, n. 24, p. 3280 - 3288, 2007. Disponível em: <http://bioinformatics.oxfordjournals.org/content/23/24/3280.abstract>.

OLIVEIRA, F. C. de et al. **Metodologia para seleção de marcadores com máquina de vetores suporte com regressão**. In: . [S.l.]: Embrapa, 2014. p. 101 - 126. ISBN 978-85-7035-382-5.

OLIVEIRA, F. C. de et al. **Snps selection using support vector regression and genetic algorithms in gwas**. BMC genomics, BioMed Central Ltd, v. 15, n. Suppl 7, p. S4, 2014.

PHILLIPS, P. C. **Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems**. Nature Reviews Genetics, Nature Publishing Group, v. 9, n. 11, p. 855 - 867, 2008.

R Core Team. R: **A Language and Environment for Statistical Computing.** Vienna, Austria, 2015. Disponível em: <http://www.R-project.org/>.

SCHAEFFER, L. **Strategy for applying genome-wide selection in dairy cattle.** Journal of Animal Breeding and Genetics, Wiley Online Library, v. 123, n. 4, p. 218 - 223, 2006.

SCHWENDER, H. **Statistical Analysis of Genotype and Gene Expression Data.** Tese (Doutorado) the Department of Statistics of the University of Dortmund, 2 2007.

SHAPIRO, S. S.; WILK, M. B. **An analysis of variance test for normality (complete samples).** Biometrika, v. 3, n. 52, 1965.

STANCZYK, U.; JAIN, L. C. **Feature Selection for Data and Pattern Recognition.** [S.l.]: Springer, 2015.

USAI, M. G.; CARTA, A.; CASU, S. **Alternative strategies for selecting subsets of predicting snps by lasso-lars procedure.** In: BIOMED CENTRAL LTD. BMC proceedings. [S.l.], 2012. v. 6, n. Suppl 2, p. S9.

VII. COPYRIGHT

Direitos autorais: O(s) autor(es) é(são) o(s) único(s) responsável(is) pelo material incluído no artigo.

REVISTA

SOLUÇÕES PARA O DESENVOLVIMENTO DO PAÍS

Atendimento:
sodebras@sodebras.com.br
Acesso:
http://www.sodebras.com.br

ISSN 1809-3957