Redes de Funções de Base Radial e método Stepwise de redução de dimensionalidade para a predição genômica de caracteres quantitativos

João Guilherme Simões[1], Webster Cristiano dos Reis Teixeira[1], Melquisadec de Souza Oliveira[1], Cíntia Laerzio Trindade[1], Isabela de Castro Sant`Anna[2], Gabi Nunes Silva[1], Laís Mayara Azevedo Barroso[1], Cosme Damião Cruz[2]

1. Introdução

Nas mais diversas áreas do melhoramento genético, animal ou de plantas, o principal objetivo é identificar e recomendar genótipos superiores de modo a aumentar a produtividade e a qualidade do produto. Para tanto, uma gama de metodologias de genética estatística está disponível na literatura, incluindo métodos de experimentação e análise estatística (RESENDE et al., 2014); genética biométrica aliada a conceitos de genética quantitativa; seleção assistida de marcadores e finalmente, métodos baseados em seleção genômica ampla, com o uso de modelos estatísticos paramétricos (MEUWISSEN et al., 2001) ou semi-paramétricos (GIANOLA et al., 2011) de predição.

A seleção genômica ampla permite estimação simultânea dos efeitos genéticos de marcadores dispersos em todo o genoma, sob o pressuposto de que a maioria dos alelos de interesse esteja associado a esses marcadores e possa explicar grande parte da variação genética e predizer o valor genético dos indivíduos que ainda não foram fenotipados (RESENDE et al., 2014). No entanto, os modelos de seleção genômica (SG), de modo geral, negligenciam os efeitos não-aditivos. Além disso, a alta densidade de marcadores moleculares pode levar a problemas de dimensionalidade – uma vez que o número de marcadores (p) é maior que o número de observações (n) (p> n) – e multicolinearidade, dada a alta correlação existente entre estes marcadores.

Neste contexto, pode-se considerar o uso de abordagens não-paramétricas baseadas em inteligência computacional como as Redes de Funções de Base Radial (RBF), que podem capturar relações não lineares que a maioria dos modelos comumente utilizados na SG não conseguem (SILVA et al., 2016). A quantidade de marcadores não gera multicolinearidade mas aumenta a demanda computacional requerida pelas redes neurais, o que acarreta em menor acurácia do valor predito e maior dificuldade para seu treinamento e aprendizado. Para sanar esses problemas relacionados ao tamanho da matriz de marcas, a literatura propõe o Método Stepwise para redução de dimensionalidade.

Diante do exposto, este trabalho foi realizado com o intuito de propor as Redes de Funções de Base Radial, aliadas ao método Stepwise de redução de dimensionalidade para a predição genômica de três caracteres quantitativos com diferentes níveis de dominância como alternativa viável aos métodos convencionais de predição de valor genético em programas de melhoramento genético, a fim de aumentar a acurácia seletiva.

[1]Fundação Universidade Federal de Rondônia. Email: simoesj.guilherme@gmail.com, melquisadec.oliveira@gmail.com, webstercristiano@gmail.com, cintialaerzio@gmail.com, gabi.silva@unir.br, lais.barroso@unir.br.

[2]Universidade Federal de Viçosa. Email: isabelacsantanna@gmail.com, cdcruz@ufv.br.

2. Material e métodos

2.1. Método de Regressão *Stepwise*

A fim de resolver o problema de multicolinearidade enfrentado pela SG e também poupar tempo e demanda computacional ao adotar uma matriz de marcas reduzida, utilizou-se o Método de Regressão Stepwise tal como descrito por James et al., 2013. Basicamente, a Regressão *Stepwise* consiste em uma metodologia para seleção de variáveis combinando dois procedimentos: *Forward* – com a inclusão de variáveis – e *Backward* – com a exclusão de variáveis.

2.2. Determinação do Número de Marcas

Neste trabalho, adaptamos o número de condição da matriz de correlação (NC). Quando o NC resultante dessa divisão foi menor ou igual a 100, considerou-se haver multicolinearidade fraca entre as variáveis explicativas; para 100<NC<1000 multicolinearidade moderada a severa e NC≥1000, considerou-se multicolinearidade severa.

2.3. Predição de valores genéticos por RR-BLUP e RBF

Para estimar os efeitos de marcas e os dos valores genômicos (GEBVs) foi utilizado a metodologia RR-BLUP conforme descrito por Meuwissen et al. (2001).

Foi adotada uma rede RBF de arquitetura do tipo *feedforward,* que consiste de uma camada de entrada, uma camada oculta considerando raios de tamanho r (r variando de 1 a 80) e uma a camada de saída. O número máximo de neurônios fornecidos foi 400 e o critério de parada estabelecido quando o limite mínimo do EQM=0,01 fosse atingido.

As entradas da rede RBF eram as informações dos marcadores selecionados após redução de dimensionalidade pelo Método de regressão *Stepwise*. O treinamento da rede função de base radial foi realizado utilizando-se o procedimento de validação cruzada fundamentado na reamostragem de um grupo de indivíduos via procedimento *k-fold*, adotando k=5. Para avaliar as eficiências dos modelos utilizados RR-BLUP, RBF na seleção genômica (SG) foi utilizada a acurácia preditiva, representada pela estatística raiz do erro quadrático médio (REQM).

2.4. Recursos Computacionais

O procedimento de simulação da população avaliada foi realizado no software GENES (CRUZ, 2016).

O procedimento para determinar o tamanho ideal da matriz reduzida de marcadores, tal como a metodologia de redução de dimensionalidade avaliadas foram implementados no software GENES (CRUZ, 2016), no módulo integração com o software MATLAB (MATLAB, 2011).

O Método do RR-BLUP foi implementado no GENES, no módulo integração com o software R (R Core Team, 2017).

3. Resultados e Discussões

Para o conjunto de dados analisado foi estabelecido que o número adequado de marcadores para fins de predição seria 150 uma vez que para esse número de marcas tem-se um NC < 100; os valores do R^2 tendem a estabilizar e os valores de REQM caem substancialmente (Figura 1).

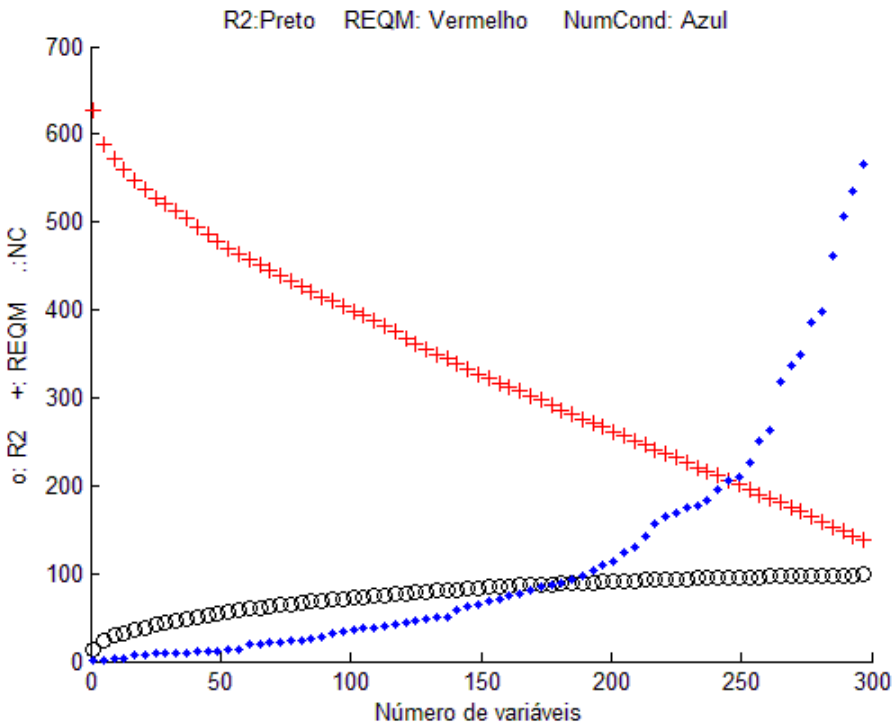

Figura 1: Representação gráfica conjunta dos valores de R^2, *REQM e NC* obtidos pelo Método de *Stepwise* ao incluir de 1 a 300 marcadores moleculares no modelo de regressão múltipla.

Após reduzir a matriz de marcadores 150, procedeu-se à análise de predição, comparando-se as metodologias de RR-BLUP e redes RBF por meio da estatística de REQM, que representa a acurácia preditiva.

Os resultados apresentados na tabela 1 demonstram a superioridade das redes RBF quando comparadas ao RR-BLUP dados os valores inferiores obtidos para o REQM. O REQM reduziu de 160,16 para 3,9613 em C2; e de 144,23 para 4,5473 em C3 na fase de validação (Tabela 1). Mesmo para a característica C1, que representa uma característica menos complexa já que envolve somente efeitos aditivos, as redes neurais foram capazes de reduzir o REQM na fase de validação de 106,47 para 3,60 (Tabela 1).

A superioridade das redes neurais artificiais em problemas de predição já foi verificada por outros autores (VENTURA et al., 2012; NASCIMENTO et al., 2013; SILVA et., 2016; CASTRO et al., 2017). Silva et., 2016 demonstraram que a inclusão de outros parâmetros estatísticos além da

média fenotípica na entrada de redes Perceptron permite que estas realizem a predição do valor genético de indivíduos simulados com maior acurácia.

Tabela 1: Estatísticas obtidas pela Rede RBF e pelo RR-BLUP com 5 validações cruzadas após reduzir a matriz de marcas utilizando o método de regressão *Stepwise*. $REQM_t$ e $REQM_v$: raiz do erro quadrático médio para as fases da treinamento e validação, respectivamente.

Características		RR-BLUP		RBF	
		$REQM_t$	$REQM_v$	$REQM_t$	$REQM_v$
C1 - D0H35Ad	Md	**106,08**	**106,47**	**2,7178**	**3,6023**
	dv	4,48	4,67	0,0027	0,1766
C2 - D60H35Ad	Md	**159,75**	**160,16**	**2,9207**	**3,9613**
	dv	2,999	2,779	0,0038	0,107
C3 - D120H35Ad	Md	**143,34**	**144,23**	**3,3662**	**4,5473**
	dv	7,471	7,789	0,0025	0,0908

4. Conclusões

Os resultados das análises evidenciam a possibilidade que o pesquisador tem para a tarefa de reduzir o número de variáveis explicativas e, também, garantir que em estudos por outras abordagens não enfrente problemas de multicolinearidade e de dimensionalidade, sem que haja perda de informações genéticas relevantes tal como a influência da dominância nas características.

Acredita-se que, com a utilização de procedimento de seleção de variáveis, as técnicas fundamentadas em inteligência computacional possam ser facilmente empregadas sem demandar recursos computacionais sofisticados.

Os resultados demonstraram as grandes potencialidades da metodologia RBF proposta quando comparada ao método RR-BLUP de seleção para estudos que envolvam a predição de valores genéticos em diferentes caracteres quantitativos.

5. Agradecimentos

CNPq pelo apoio financeiro.

6. Referências Bibliográficas

CASTRO, C.A.O.; RESENDE, R.T.; KUKI, K.N.; CARNEIRO, V.Q.; MARCATTI, G.E.; CRUZ, C.D.; MOTOIKE, S.Y. High-performance prediction of macauba fruit biomass for agricultural and industrial purposes using Artificial Neural Networks.Industrial Crops & Products v.108,p.806–813, 2017.

CRUZ, C.D. Genes Software – extended and integrated with the R, Matlab and Selegen. Acta Scientiarum. Agronomy. Maringá, v. 38, n. 4, p. 547-552, Oct.-Dec., 2016.

GIANOLA, D.; OKUT, H.; WEIGEL, K. A.; ROSA, G. J. M. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genetics. p.12-87, 2011.

JAMES, G., WITTEN, D., HASTIE, T. AND TIBSHIRANI, R., 2013. An introduction to statistical learning (Vol. 112). New York: springer.

MATLAB (2010). Matlab Version 7.10.0. Natick, Massachusetts: The Math Works Inc.

MEUWISSEN, T.H.E.; HAYES, B.J.; GODDARD, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4): 1819–1829.

NASCIMENTO, M.; PETERNELLI, L. A.; CRUZ, C. D.; NASCIMENTO, ANA CAROLINA CAMPANA; FERREIRA, R. P.; BHERING, L. L.; SALGADO, C. C. Artificial neural networks for adaptability and stability evaluation in alfafa genotypes. Crop Breeding and Applied Biotechnology, v.12, p.152-156, 2013.

R CORE TEAM. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2017. Available at: <https://www.R-project.org/>.

RESENDE, M.D.V.; SILVA, F.F.; AZEVEDO, C.F. Estatística matemática, biométrica e computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição Sobrevivência. Viçosa: Suprema, 881p. 2014.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I.C.; CARNEIRO, V.Q.; CRUZ, C.D.; NASCIMENTO, M. Evaluation of the efficiency of artificial neural networks for genetic value prediction. Genetic Molecular Research, v.15, p.1-11, 2016. DOI: 10.4238/ gmr.15017676, 2016.

VENTURA, R.; SILVA, M.; MEDEIROS, T.; DIONELLO, N.; MADALENA, F.; FRIDRICH, A.; VALENTE, B.; SANTOS, G.; FREITAS, L.; WENCESLAU, R. Use of artificial neural networks in breeding values prediction for weight at 205 days in Tabapuã beef cattle. Arquivo Brasileiro de Medicina Veterinária e Zootecnia, v.64, n.2, p.411-418, 2012.