XII Simpósio Brasileiro de Melhoramento Animal
Ribeirão Preto, SP – 12 e 13 de junho de 2017

S B M A 2017

Conceitos da teoria de grafo na definição de estratégias de genotipagem para seleção genômica em populações bovinas com aptidão leiteira.

Bruno da C. Perez[1], Júlio C.C Balieiro[2], Juliana Machado[3], Maria Gabriela C.D. Peixoto[4], Ricardo V. Ventura[1,5,6]

[1]Universidade de São Paulo, Faculdade de Zootecnia e Engenharia de Alimentos, Pirassununga, SP, Brasil.
[2]Universidade de São Paulo, Faculdade de Medicina Veterinária e Zootecnia, Pirassununga, SP, Brasil.
[3]Universidade Federal do Rio Grande do Sul, Departamento de Zootecnia, Porto Alegre, RS, Brasil.
[4]Empresa Brasileira de Pesquisa Agropecuária, Embrapa Gado de Leite (CNPGL), Juiz de Fora, MG, Brasil
[5]University of Guelph, Department of Animals and Poultry Science, Guelph, Ontario, Canadá.
[6]Beef Improvement Opportunities, BIO, Elora, Ontario, Canadá.
*Autor correspondente: brunocpvet@usp.br

Resumo: O objetivo do presente estudo foi acessar a aplicação do conceito de comunidade, em pedigrees modelados por teoria de grafos, para definição de estratégias de genotipagem. O estudo foi conduzido por meio de simulação, em cinco réplicas. Uma população contemporânea de 16 mil fêmeas (e 16 mil machos), oriundas de 4 gerações, foi utilizada para amostrar 1.000, 2.000 e 5.000 indivíduos por diferentes métodos para compor a população de referência. A população de validação constava de 8.000 indivíduos oriundos da geração mais recente. Métodos explorados na literatura foram comparados à métodos que incorporam a teoria de detecção de comunidades (em grafos). A comparação entre métodos foi realizada através da correlação linear ($r_{DGV,TBV}$) e do erro quadrático médio (MSE) entre o valor genômico estimado e o valor genético verdadeiro dos animais da população de validação. Métodos baseados na detecção de comunidades obtiveram valores de $r_{DGV,TBV}$ superiores e MSE inferiores ao restante, sugerindo que há potencial em sua aplicação prática.

Palavras–chave: *detecção de comunidades, Python, , redes*

Graph theory concepts for the definition of genotyping strategies for genomic selection in dairy cattle populations.

Abstract: The objective of the present study was to assess the application of the concept of community, in pedigrees modeled by graph theory, for the definition of genotyping strategies. This study was conducted via simulation in five replicates. A contemporary population of 16,000 cows (50% of the total), originated from 4 overlapping generations, was used to sample 1,000; 2,000 and 5,000 individuals by different methods to form the reference population. The validation population contained 8,000 animals originated from the latest generation. Previously explored methods (in literature) were compared to methods that incorporate the theory of community detection (in graphs). The comparison was assessed by the linear correlation ($r_{DGV,TBV}$) and mean squared error (MSE) between estimated genomic values and the true breeding values of animals in the validation population. Mathods based in community detection obtained higher values of $r_{DGV,TBV}$ and lower values of MSE than the others, suggesting potential in its practical application.

Keywords: *community detection, networks, Python*

Introdução

O sucesso da seleção genômica basea-se, entre outros aspectos, na estrutura da população de referência (POP_{REF}) utilizada para se estimar o efeitos dos marcadores, que por sua vez está diretamente relacionada à escolha dos indivíduos a serem genotipados. A obtenção de uma acurácia de predição adequada torna-se um obstáculo principalmente em pequenas populações, onde poucos reprodutores tem resultados de teste de progênie com confiabilidade suficientemente alta (Van Raden et al., 2009). A inclusão de fêmeas na POP_{REF} pode ser uma estratégia factível. Entretanto, o elevado número de candidatas demanda a definição de estratégias que otimizem o custo de genotipagem.

S B M A 2017

Grafos são sistemas matemáticos flexíveis que contém um número finito de pontos (nodos ou vértices) e arestas, que são essencialmente utilizados para representar relacionamentos entre dados estruturados. Quando grafos são aplicados para modelagem de pedigrees, nodos representam indivíduos e arestas representam a relação progenitor – progênie. Comunidades são atributos inerentes da teoria de grafos e podem ser definidas por grupos de nodos fortemente conectados entre sí e esparsamente conectados com o restante dos nodos no sistema. Algoritmos para detecção de comunidades são capazes de dividir um grafo em partições, baseando-se no parâmetro de modularidade que define a qualidade da configuração de comunidades identificada.

Este estudo teve como objetivo avaliar a incorporação do conceito de comunidade em pedigrees modelados como grafo para definição de estratégias de genotipagem de fêmeas em populações bovinas com aptidão leiteira. Os diferentes métodos propostos foram comparados pela correlação linear de Pearson entre o valor genômico estimado (DGV) e o valor genético "verdadeiro" (TBV) para os indivíduos da população de validação (POP_{VAL}).

Material e Métodos

A simulação foi realizada por meio do softwarre QMSim (Sargolzaei and Schenkel., 2006). Foi criada uma população histórica (H_{POP}) contendo 1.500 indivíduos (50% machos – 50% femeas) na primeira geração. Após 1.200 gerações mantendo o mesmo número de animais, H_{POP} foi reduzida a 400 indivíduos durante 550 gerações (mantendo a mesma razão macho/fêmea) de forma a criar um nível de desequilibrio de ligação plausível. A partir da H_{POP}, 400 indivíduos (200 machos e 200 fêmeas) foram selecionados de forma aleatória como fundadores da primeira população recente ($R_{POP}1$), a qual foi formada por 100 gerações de acasalamento aleatório, assumindo taxas de crescimento de 0,10 e 0,40 para machos e fêmeas, respectivamente. Uma segunda população recente ($R_{POP}2$) foi criada, selecionando 100 machos e 8.000 fêmeas (aleatoriamente) oriundos da última geração (100) de $R_{POP}1$. $R_{POP}2$ foi conduzida por 20 gerações sobrepostas, assumindo 1 progênie por vaca por meio de acasalamento aleatório entre os animais selecionados e taxa de reposição de 0,25 e 0,35 para machos e fêmeas. A seleção foi guiada pelo valor genético estimado e a reposição baseada na idade dos animais. Indivíduos das gerações 16 a 20 de $R_{POP}2$ tiveram sua informação de marcadores genéticos registrada e formaram as POP_{REF} e POP_{VAL}. Para cada animal entre as gerações 16 e 20 de ($R_{POP}2$), foram registradas informações referentes a genealogia, valor genético verdadeiro, fenótipo e valor genético estimado em análise baseada na informação de pedigree. Os animais genotipados tinham ao menos 15 gerações de profundidade no pedigree. O pedigree contendo 100 mil animais (gerações 8 à 19) foi analizado como um grafo e uma adaptação do algoritmo de detecção de comunidades (Blondel et al., 2008) foi utilizado para obter o particionamento da população. Comunidades com menos de 30 indivíduos foram desconsideradas da análise. A análise proposta foi implementada utilizando a linguagem de programação Python (v2.7).

Uma característica de herdabilidade 0.25 foi simulada, assumindo distribuição Normal e variância 1. O genoma simulado consistiu de 29 cromossomos de comprimeto (em cM) variado, de acordo com o reportado na literatura, contendo 750 QTL posicionados de forma randômica em uma extensão total de 3.087,4 cM. Um total de 54.609 marcadores bialélicos foram posicionados no genoma, seguindo a distribuição por cromossomo referente a plataforma Ilumina BovineSNP50 BeadChip.

Animais entre as gerações 16 e 19 de $R_{POP}2$ representando uma população ativa de 16 mil fêmeas e 16 mil machos, foram considerados para amostrar populações de referência contendo 1.000, 2.000 ou 5.000 fêmeas. Indivíduos da geração 20 foram considerados como POP_{VAL}. Cinco diferentes estratégias para amostragem da POP_{REF} foram consideradas: escolha aleatória (ALE), maiores EBV (M_{EBV}), valores extremos de EBV (EX_{EBV}), maiores EBV dentro das comunidades ($M_{EBV}C$) e valores extremos de EBV dentro de cada comunidade ($EX_{EBV}C$) detectada.

Resultados e Discussão

O número de comunidades detectadas foi de 92, 90, 88, 90 e 88 nas 5 replicas da simulação analisadas, indicando a consistência do algoritmo proposto . Quando utilizados apenas os machos (650) com mais de 30 filhas na POP_{REF}, o valor de $r_{DGV,TBV}$ foi de 0,53. A média (e respectivos desvios padrão) das acurácias dos DGV obtidos e do o MSE para as situações e estratégias propostas é apresentada na Tabela 1. Em geral, o aumento do número de vacas na POP_{REF} aumentou a $r_{DGV,TBV}$ em todos os cenários.

A estratégia M_{EBV} obteve o maior aumento na acurácia com o aumento do número de fêmeas na POP_{REF} (-0.17 para 0,23), seguida de ALE. Porém, ambas foram inferiores aos métodos baseados no valores extremos de EBV dos indivíduos em todos os cenários. Resultados para as estratégias "tradicionais" estiveram de acordo com os reportados por Jimenez-Montero et al. (2012).

Tabela 1. Média da acurácia dos valores genômicos estimados (em 5 réplicas) considerando populações de referência contendo 1000, 2000 e 5000 vacas pelos diferentes métodos de amostragem propostos.

NRef		ALE	MEBV	EXEBV	MEBVC	EXEBVC
1000	$r_{DGV,TBV}$	0,40 ± 0,06	-0,17 ± 0,04	0,49 ± 0,06	0,46 ± 0,08	0,51 ± 0,04
	MSE	17,47 ± 2,63	17,98 ± 1,37	15,72 ± 3,09	19,64 ± 2,49	11,87 ± 2,25
2000	$r_{DGV,TBV}$	0,53 ± 0,04	-0,01 ± 0,06	0,56 ± 0,04	0,54 ± 0,05	0,58 ± 0,05
	MSE	18,36 ± 1,29	21,67 ± 1,69	19,09 ± 3,10	20,38 ± 1,71	13,69 ± 0,73
5000	$r_{DGV,TBV}$	0,62 ± 0,03	0,23 ± 0,05	0,66 ± 0,04	0,67 ± 0,03	0,69 ± 0,03
	MSE	16,28 ± 2,62	20,09 ± 2,97	13,54 ± 2,20	17,12 ± 2,62	13,19 ± 1,92

NRef = número de fêmeas na população de referência

Em todas as situações, a $r_{DGV,TBV}$ nos indivíduos da população de validação para o método M_{EBV} foi a mais baixa dentre os métodos propostos, sendo inclusive negativa para o cenário com 1000 e 2000 fêmeas na POP_{REF}. No entanto, quando os indivíduos de maior EBV foram selecionados dentro de cada comunidade ($M_{EBV}C$), a $r_{DGV,TBV}$ na POP_{VAL} não apenas foi positiva, como se comportou de forma semelhante aos métodos que exibiram os maiores valores de acurácia de predição. Estratégias baseadas em $M_{EBV}C$ podem ser interessantes em condições de orçamento limitado, uma vez que criadores podem genotipar os indivíduos de EBV superior sem graves impactos negativos na acurácia de predição. É preciso salientar que os maiores valores de EBV para os indivíduos em diferentes comunidades podem, por vezes, representar valor baixos ou intermediários quando considerada toda a população avaliada. A estratégia (EX_{EBV}) apresentou o melhor valor de $r_{DGV,TBV}$ dentre os métodos "tradicionais". Por fim, $EX_{EBV}C$ obteve os melhores resultados de $r_{DGV,TBV}$ em todos os cenários, mostrando potencial como estratégia de genotipagem.

Os valores de MSE para os cenários e estratégias propostas são apresentados na Tabela 1. O erro quadrático foi proporcional dentro de cada cenário quando comparados diferentes tamanhos de amostragem na POP_{REF}. No entanto, o aumento do número de fêmeas afetou de forma expressiva o MSE apenas para EX_{EBV}.

Conclusão

A aplicação da teoria de comunidades mostrou potencial como estratégia para definição de estratégias de genotipagem na formação de populações de referência formadas exclusivamente por vacas em pequenas populações de raças leiteiras.

Agradecimentos

Os autores gostariam de agradecer à CAPES/CNPq pela bolsa de doutorado concedida ao primeiro autor. e à Beef Improvment Opportunities (BIO) pelo tempo cedido.

Literatura citada

BLONDEL V.D.; GUILLAUME J.L.; LAMBIOTTE R.; LEFEBVRE E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics Theory and Experiment, 10, 2008.

VAN RADEN, P.M.; WIGGANS, G.R.; VAN-TASSEL, C.P.; SONSTEGARD, T.S.; SCHENKEL, F. Benefits from cooperation in genomics. Interbull Bulletin 40, 67-72, 2009.

SARGOLZAEI M.; IWAISAKI H.; COLLEAU J.J. CFC: A tool for monitoring genetic diversity. Proc. 8th World Congr. Genet. Appl. Livest. Prod., CD-ROM Communication, 13-18, 2006.

JIMENEZ-MONTERO, J.A.; GONZALEZ-RÉCIO, O.; ALENDA, R. Genotyping strategies for genomic selection in small dairy cattle populations. Animal, 6:8, 1216-1224, 2012.