OS DESAFIOS DO USO DE 'BIG DATA' NO MELHORAMENTO GENÉTICO ANIMAL

Daniela Lourenço[1]

[1] University of Georgia – ATHENS - EUA

RESUMO

O objetivo do melhoramento genético animal é selecionar os melhores animais para serem pais das gerações futuras. Um modelo misto é utilizado para predizer os valores genéticos dos animais, ou a soma dos efeitos dos genes que afetam determinada característica. A predição dos valores genéticos leva em consideração que os animais são aparentados, portanto uma matriz de covariância é utilizada. Essa estrutura de covariância é dada pela inversa da variância genética aditiva da característica, multiplicada pela inversa da matriz de parentesco entre os animais (**A**). A matriz **A** contém a esperança da proporção de genes compartilhada entre indivíduos e tem a dimensão do número de animais no sistema de equações, mas por ser esparsa, um algoritmo simples é utilizado para a construção da inversa. Em 2008, com o surgimento de um chip para genotipar animais, a informação extraída diretamente do genoma passou a ser utilizada em conjunto com a informação de parentesco para a construção da estrutura de covariância entre os animais, com o objetivo de aumentar a acurácia da predição do valor genético. O chip mais comumente utilizado é capaz de ler aproximadamente 55 mil polimorfismos no DNA, chamados *Single Nucleotide Polymorphism* (SNP). Essa informação é utilizada para a construção de uma matriz de parentesco genômico (**G**) entre os indivíduos, que contém a proporção observada de SNP compartilhados. Ao contrário da matriz **A**, a **G** é densa e deve ser invertida diretamente. No entanto, o número de animais genotipados tem crescido rapidamente nos Estados Unidos. A Associação Americana de gado Holandês possui dois milhões de animais genotipados; em segundo lugar está a Associação Americana de Angus com 400 mil. Para manter uma matriz com dimensão de dois milhões de animais genotipados são necessários 2,98 Tb de memória RAM. Quando esse número é maior que 100 mil, a inversão direta da matriz **G** é computacionalmente impraticável. A necessidade da criação de novos algoritmos dado que os existentes não podem ser utilizados é que define *Big Data*. Dessa forma, um algoritmo para a construção da inversa da matriz **G** ($\mathbf{G}^{-1}$) foi proposto e recebeu o nome de algoritmo para animais provados e jovens (*Algorithm for Proven and Young* – APY). O APY usa recursões em um número pequeno de animais, chamado grupo de animais provados, e a teoria é baseada no fato de que a informação genômica tem uma dimensionalidade limitada. Essa dimensionalidade é justificada pelo número de autovalores que explicam 98% da variação presente na matriz **G**. A correlação entre os valores genéticos estimados usando a matriz $\mathbf{G}^{-1}$ obtida diretamente e a $\mathbf{G}^{-1}$ com APY é maior que 0,99, indicando que o método é robusto. Atualmente, testes foram realizados com vários bancos de dados, o maior deles sendo o da Associação Americana de gado Holandês com 81 milhões de animais no pedigree, 32 milhões de animais com fenótipos para três características e 760 mil animais genotipados. As soluções para as equações de modelos mistos foram obtidas com sucesso utilizando-se um critério de convergência de 10^{-15}. O APY foi implementado nos programas da família BLUPF90 e é utilizado na avaliação genética em nível comercial por diversas empresas e associações americanas.