

Identificação de Marcadores SNPs para Teor de Amilose em Arroz pela Estratégia de *Machine Learning*⁽¹⁾

Douglas Eterno Louza Sartor², Aliny Heloísa Alcântara Rodrigues³, Ricardo Cerri⁴, Priscila Zaczuk Bassinello⁵, Marcelo Gonçalves Narciso⁶, Claudio Brondani⁷ e Rosana Pereira Vianello⁸

¹ Pesquisa Financiada pelo CNPq.

² Estudante de graduação em Ciências Biológicas pela Pontifícia Universidade Católica de Goiás, estagiário da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO

³ Estudante de graduação em Agronomia da Universidade Federal de Goiás, estagiária da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO

⁴ Doutor em Ciência da Computação, professor adjunto da Universidade Federal de São Carlos, São Carlos, SP

⁵ Engenheira-agrônoma, doutora em Ciência de Alimentos, pesquisadora da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO

⁶ Engenheiro Eletrônico, doutor em Computação Aplicada, pesquisador da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO

⁷ Engenheiro-agrônomo, doutor em Biologia Molecular, pesquisador da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO

⁸ Bióloga, doutora em Biologia Molecular, pesquisadora da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO

Diversos e complexos são os aspectos associados à qualidade de grão de arroz. O conteúdo de amilose, que integra a composição do amido, é considerado um importante parâmetro para a qualidade tecnológica e de consumo do arroz. Grãos com teor de amilose (TA) intermediário apresentam textura mais firme após o cozimento, o que os torna mais atrativo para o mercado consumidor brasileiro. O objetivo deste estudo foi identificar, por meio da estratégia *Machine Learning*, marcadores SNPs relacionados com TA para uso na seleção assistida por marcadores no programa de melhoramento genético de arroz irrigado. Essa estratégia se baseia no reconhecimento de padrões associados entre os marcadores SNPs e o respectivo fenótipo de TA, correspondendo ao aprendizado computacional em inteligência artificial. Essa análise tem sido utilizada em humanos para identificação de genes associados a doenças, mas de uso ainda não descrito em arroz para a identificação de genes de interesse para o melhoramento. Foram avaliados 550 acessos da Coleção Nuclear de Arroz da Embrapa, fenotipados quanto ao TA em sete experimentos de campo, compreendendo as regiões Norte, Nordeste, Centro-Oeste e Sul. O TA foi determinado na Embrapa Arroz e Feijão para todos os acessos, oriundos dos sete experimentos. Os marcadores SNPs dos 550 acessos foram obtidos por meio da análise GBS (Genotyping By Sequencing), e resultou, inicialmente, em mais de 450.000 SNPs. Desse total, foram selecionados 4.079 SNPs, regularmente espaçados e distribuídos ao longo dos 12 cromossomos do arroz. A identificação dos SNPs foi realizada utilizando ferramentas de *análise de big data* e algoritmos para a identificação de padrões de SNPs que explicassem o teor de amilose. O total de acessos com TA intermediário variou de 24% para o ambiente de Teresina (PI) a 66% em Pelotas (RS), e 72% no ambiente de Uruguaiana (RS). Para os dois experimentos conduzidos em Goiânia, genótipos com TA intermediários foram identificados em ~40% dos acessos. Dos 550 acessos avaliados, sete apresentaram TA intermediários e coincidentes nos sete experimentos. A análise conjunta, calculada a partir da média dos sete ambientes, revelou 91% dos acessos com TA intermediário. A análise de *machine learning*, através da integração dos dados fenotípicos e moleculares, identificou padrões de SNPs para prever as respostas e gerar o melhor resultado preditivo com menor chance de erro. Desse modo, foi possível chegar a uma combinação de dez SNPs que diferenciam acessos agrupados pelas classes de TA. Nesse momento, ensaios TaqMan estão em fase de desenvolvimento e serão utilizados para validar esses resultados para cada local avaliado, e pela análise conjunta. Os SNPs ou conjunto de SNPs relacionados ao TA intermediário (seleção a favor) ou relacionados ao TA alto ou baixo (seleção contra) serão úteis para a rotina do programa de melhoramento genético do arroz.