**ALBERTO SOUZA BOLDT**

# COLEÇÕES NUCLEARES E ASSOCIAÇÃO DO TEOR DE ÓLEO DE CÁRTAMO COM VARIÁVEIS ECOGEOGRÁFICAS POR INTELIGÊNCIA COMPUTACIONAL

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de Doctor Scientiae.

**VIÇOSA**
**MINAS GERAIS - BRASIL**
**2014**

**Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa**

T
B687c
2014

Boldt, Alberto Souza, 1987-
Coleções nucleares e associação do teor de óleo de cártamo com variáveis ecogeográficas por inteligência computacional / Alberto Souza Boldt. – Viçosa, MG, 2014.
vii, 58f. : il. (algumas color.) ; 29 cm.

Inclui apêndice.
Orientador: Sérgio Yoshimitsu Motoike.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Plantas Oleoginosas. 2. *Carthamus tinctorius*. 3. Açafrão. 4. Germoplasma vegetal. 5. Coleção nuclear. I. Universidade Federal de Viçosa. Departamento de Fitotecnia. Programa de Pós-graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 633.85

**ALBERTO SOUZA BOLDT**

# COLEÇÕES NUCLEARES E ASSOCIAÇÃO DO TEOR DE ÓLEO DE CÁRTAMO COM VARIÁVEIS ECOGEOGRÁFICAS POR INTELIGÊNCIA COMPUTACIONAL

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de Doctor Scientiae.

APROVADA: 08 de abril de 2014.

| | |
|---|---|
| Paulo Roberto Cecon | Rogério Oliveira de Sá |
| Luiz Antônio dos Santos Dias | Cosme Damião Cruz<br>(Coorientador) |

Sérgio Yoshimitsu Motoike
(Orientador)

# AGRADECIMENTOS

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Genética e Melhoramento pela oportunidade de realização do curso de Doutorado.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio financeiro, que possibilitou a realização deste trabalho.

Ao professor Tuneo Sediyama, pelos ensinamentos e valores transmitidos durante o meu treinamento em melhoramento de plantas.

Ao professor Sérgio Yoshimitsu Motoike pela orientação, paciência, amizade, incentivo e confiança.

Ao professores Cosme Damião Cruz e Luiz Antônio dos Santos Dias pela coorientação, sugestões, atenção e ensinamentos.

Aos Dr. Rogério Oliveira de Sá e Dr. Paulo Roberto Cecon, pela disponibilidade, participação e sugestões na banca examinadora.

Ao Instituto Matogrossense do Algodão (IMAmt), pela cessão dos genótipos de cártamo e das áreas experimentais para realização do trabalho.

Ao Departamento de Agricultura dos Estados Unidos (USDA), pela disponibilização de dados históricos de melhoramento de cártamo.

Ao Laboratório de Agroenergia do Departamento de Fitotecnia da Universidade Federal de Viçosa, coordenado pelo professor Luiz Antônio dos Santos Dias, pela disponibilidade e apoio na realização das análises químicas.

Ao Laboratório de Biotecnologia e Melhoramento Vegetal, pela disponibilidade e apoio na realização de análises moleculares.

Aos funcionários do Programa de Melhoramento Genético de Soja, pela ajuda prestada.

Aos meus pais, Alberto Francisco Boldt e Ana Gizelli Dias de Souza Boldt, pelo apoio, educação e incentivo constantes, e aos meus irmãos, Gustavo, Marcela e Francielli.

À minha namorada Giuliana Cristina Mourão Soares, por estar sempre ao meu lado, sendo minha amiga e companheira.

Aos amigos Lívio, Vitor e Thiago, pelo apoio e companheirismo.

# SUMÁRIO

# RESUMO

BOLDT, Alberto Souza, D.Sc., Universidade Federal de Viçosa, abril de 2014. **Coleções nucleares e associação do teor de óleo de cártamo com variáveis ecogeográficas por inteligência computacional.** Orientador: Sérgio Yoshimitsu Motoike. Coorientadores: Tuneo Sediyama e Cosme Damião Cruz.

Cártamo (Carthamus tinctorius L.) é uma espécie oleaginosa com um grande potencial genético confinado nos bancos de germoplasma. Fonte de características relevantes, os bancos de germoplasma de cártamo tem apresentado uso limitado devido ao grande número de acessos disponíveis nas coleções. O presente trabalho objetivou explorar a diversidade genética de cártamo por meio do estabelecimento de coleções nucleares mais expressivas utilizando as estratégias de maximização e estratificação de genótipos em grupos genéticos conhecidos. O trabalho também objetivou investigar a existência de associação preditiva entre teor de óleo e variáveis ecogeográficas da origem de acessos de cártamo, utilizando a estratégia de identificação focada de germoplasma para explorar a associação e aumentar as chances de encontrar genótipos de cártamo com alto teor óleo. No estabelecimento das coleções nucleares foram utilizados caracteres fenotípicos, qualitativos e quantitativos, de 1640 acessos de cártamo provenientes de 48 países. Os acessos foram estratificados nos grupos genéticos de acordo com país de origem e amostrados segundo a estratégia de maximização. As coleções nucleares estabelecidas foram comparadas com a coleção base utilizando estatísticas de validação adequadas. As magnitudes das estimativas das estatísticas de validação indicaram que a variabilidade genética dos acessos da coleção base foi preservada nas coleções nucleares estabelecidas. As coleções nucleares estratificadas por grupos genéticos apresentaram aproximadamente 60 genótipos, com diferença média de apenas 7% em relação a coleção base e com taxa de coincidência de superior a 94%. O uso conjunto da estratégia de maximização e da estratificação dos genótipos em grupos genéticos maximizou a captação da variabilidade genética e introduziu maior eficiência no estabelecimento das coleções nucleares ao selecionar uma quantidade reduzida de acessos. As coleções nucleares estabelecidas incluíram aproximadamente 3.75% dos acessos conservados na coleção base. Para estabelecer coleções nucleares expressivas é necessário que os acessos sejam selecionados da coleção base de maneira apropriada. A estratégia de identificação focada de germoplasma é um método eficiente de otimizar a seleção de acessos presentes nos bancos de germoplasma. A FIGS faz uso da associação preditiva entre características e variáveis ambientais na busca de genótipos com maior probabilidade de conter a característica

de interesse. Florestas aleatórias, máquinas de vetor de suporte e redes neurais artificias foram utilizadas para modelar a associação entre teor de óleo de 100 genótipos cártamo e 56 variáveis ecogeográficas. As acurácias dos modelos utilizados mostraram que a distribuição de genótipos de cártamo com alto teor de óleo não é aleatória mas ligada a fatores ambientais, mesmo com certo grau de sobreposição entre os teores de óleo em alguns ambientes. Os resultados finais sugerem que explorar a associação preditiva entre o teor de óleo e as características ecogeográficas do local de origem do germoplasma aumenta as chances de encontrar genótipos com alto teor óleo.

# ABSTRACT

BOLDT, Alberto Souza, D.Sc., Universidade Federal de Viçosa, April 2014. **Core collections and association between safflower oil and ecogeographic data by computational intelligence.** Adviser: Sérgio Yoshimitsu Motoike. Co-advisers: Tuneo Sediyama and Cosme Damião Cruz.

Safflower (Carthamus tinctorius L.) is an oilseed species with a large genetic potential available in genebanks. Source of relevant characters, safflower germplasm banks have shown limited use due to the large number of accessions available in collections. The present study aimed to explore the genetic diversity of safflower through the establishment of more expressive core collections using maximization strategy and the stratification of genotypes in genetic groups. The study also aimed to investigate the existence of predictive association between oil content and ecogeographic parameters of the original site of safflower accessions, using the focused identification of germplasm strategy to explore the association and increase the chances of finding safflower genotypes with high oil content. Core collections were established using phenotypic qualitative and quantitative traits data of 1640 safflower accessions from 48 countries. The accessions were stratified into genetic groups according to country's origin and sampled according to the maximization strategy (M strategy) . The established core collections were compared with the base collection using the following validation statistics: chi-square test, mean difference, difference of variances, coincidence rate, variable rate and Shannon index. Magnitude estimates of validation statistics indicated that base collection's genetic variability was preserved in the core collections based on safflower centers of similarity. Core collections stratified by genetic groups consisted in about 60 genotypes, with a mean difference of 7% over the base collection and coincidence rate above 94%. The combined use of the maximization strategy and stratification of genotypes in genetic groups maximized the capture of genetic variation and introduced more efficiency, establishing core collections with a fewer number of accessions. The core collections included approximately 3,75% accessions conserved in safflower base collection. To establish expressive core collections is necessary selecting accessions properly from base collection. The Focused Identification Germplasm Strategy (FIGS) is an efficient method to optimize the selection of useful accessions kept in collections. The FIGS makes use of predictive association between characteristics and environmental variables in the search for genotypes with high probability of containing the trait of interest. The present study aimed to investigate the existence of predictive association between oil content and ecogeographic parameters of

the original site of safflower genotypes using the FIGS based on machine learning approaches. Random forests, support vector machines and artificial neural networks were used to model the association between oil content of 100 safflower genotypes and 56 ecogeographic parameters. The models accuracies indicated that the distribution of safflower genotypes with high oil content is not random but associated to environmental factors, even with some degree of overlap between the oil content in some environments. The final results suggest that exploring the predictive association between oil content and ecogeographic parameters of original collection site of safflower germplasm increases the chances of finding genotypes with high oil content.

# 1. INTRODUÇÃO GERAL

A busca por novas fontes de energia, renováveis e ecologicamente corretas tem estimulado a utilização dos óleos vegetais como alternativa para produção de energia, principalmente na forma de biocombustíveis.

Nos últimos anos, estudos de fontes alternativas aos derivados de petróleo tem sido realizados em diversos centros de pesquisa. Considerando o caráter finito do petróleo, carvão e gás natural, tais estudos buscam intensificar o uso de fontes renováveis de energia e otimizar a utilização das fontes não renováveis. Há um conjunto de fatores que motivam a adoção de tais medidas, destacando-se os benefícios ambientais, econômicos e sociais gerados pela utilização mais racional dos recursos naturais. Os preços elevados do petróleo no mercado mundial e a pressão internacional para a redução da emissão de gases de efeito estufa também aceleraram a expansão dos cultivos de oleaginosas destinadas a produção de biocombustíveis.

O Brasil é um dos países com maior potencial para a produção de combustíveis a partir de vegetais. A exploração de menos de um terço de sua área agriculturável constitui a maior fronteira para expansão agrícola do mundo, com cerca de 150 milhões de hectares. O país ainda é capaz de incorporar novas áreas à agricultura para geração de energia sem competir com a agricultura para alimentação, e com impactos ambientais limitados. A extensa área geográfica, clima tropical e subtropical do Brasil também favorecem uma ampla diversidade de matérias-primas para a produção de biodiesel (TRZECIAK et al., 2008).

Entre as espécies oleaginosas já estabelecidas e potenciais para a produção de biocombustível no Brasil, destaca-se o cártamo (Carthamus tinctorius L.). O cártamo é uma planta oleaginosa, anual, altamente adaptada às condições semiáridas, com elevada resistência à falta d'agua, à baixa umidade relativa do ar, à elevadas temperaturas e aos ventos fortes e quentes. Exibe grande capacidade de adaptação nas regiões semiáridas de baixa altitude, mostrando-se pouco sensível às variações de fotoperíodo e desenvolvendo-se bem nos diferentes tipos de solos (SINGH e NIMBKAR, 2007).

O cártamo é cultivado em mais de 60 países no mundo, em uma área de menos de 1 milhão de hectares. Em 2010, a produção mundial de cártamo foi de 634.604 toneladas, em uma área de 772.705 hectares (FAOSTAT, 2012). Os principais países produtores são Índia, Cazaquistão, Estados Unidos, México, Argentina e China, responsáveis por aproximadamente 90% da produção mundial.

A produtividade potencial de grãos do cártamo é similar a produtividade da soja, aproximadamente 3000 $kg.ha^{-1}$. No entanto, o teor médio de óleo extraído das sementes de variedades de cártamo é superior, variando de 35 a 46%. O óleo do cártamo possui altos teores de ácidos linoleico e/ou oleico. Uma das características químicas mais importantes deste óleo é a poliinsaturação, o que lhe confere um alto valor dietético pela redução do nível de colesterol e das doenças circulatórias e cardíacas (DAJUE e MÜNDEL, 1996).

A alta qualidade do óleo extraído das sementes de cártamo faz com que seu uso seja muito diversificado, permitindo também sua utilização na indústria alimentícia e farmacêutica. No entanto, o sucesso da capacidade produtiva dessa oleaginosa depende, em grande parte, da disponibilidade de tecnologias que contribuam para eficiência produtiva. Nesse contexto, as variedades melhoradas de cártamo são imprescindíveis para alto desempenho produtivo de óleo.

Os programas de melhoramento contribuem para o desenvolvimento de variedades mais produtivas, com maior teor de óleo, tolerantes a fatores bióticos e abióticos e adaptadas às condições das regiões de cultivo. No entanto, são capazes de atingir sucesso somente com a existência de potencial produtivo e suficiente diversidade genética.

Dessa forma, torna-se indispensável investigar o potencial e a diversidade genética de cártamo para subsidiar e otimizar o desenvolvimento de materiais genéticos pelos programas de melhoramento. A diversidade genética do cártamo, expressa na forma de acessos conservados em bancos de germoplasma, pode ser avaliada por meio da caracterização do germoplasma, utilizando características morfológicas, agronômicas, moleculares e ecogeográficas de origem.

Estudos demonstraram que há grande variabilidade entre os genótipos de cártamo, apenas ao se considerar características morfológicas e agronômicas, principalmente entre os acessos depositados no banco de germoplasma do Departamento de Agricultura dos Estados Unidos (USDA) (BRADLEY e JOHNSON, 2001; JOHNSON et al., 1999; JOHNSON et al., 2001).

Os acessos de cártamo caracterizados, conservados e distribuídos nos bancos de germoplasma devem ser, portanto, utilizados com o propósito de fornecer alelos para caracteres de interesse do melhoramento genético. Porém, frequentemente a grande quantidade de acessos disponíveis impossibilita o uso imediato e eficiente do germoplasma pelos melhoristas. Nessa condição o estabelecimento de coleções nucleares é indicado como a forma mais apropriada de utilizar o germoplasma disponível (VAN HINTUM et al., 2000).

O estabelecimento de coleções nucleares consiste na seleção de uma quantidade limitada de acessos representativos da coleção base, com o objetivo de representar grande fração da diversidade genética da coleção base (BROWN, 1995). A menor quantidade de acessos na coleção nuclear proporciona manuseio e utilização mais eficiente dos acessos, de modo que possibilita a concentração de esforços na caracterização mais completa dos acessos presentes na coleção. Os procedimentos básicos utilizados no estabelecimento de coleções nucleares envolvem a utilização das características morfológicas, agronômicas, moleculares e ecogeográficas para seleção de acessos. A primeira coleção nuclear de cártamo, desenvolvida por Johnson et al. (1993), foi estabelecida a partir de caracteres morfológicos e agronômicos, como hábito de crescimento, cor de flor e teor de óleo de 2000 acessos provenientes de 50 países.

No entanto, coleções são capazes de representar melhor a estrutura da variabilidade genética existente na coleção base quando os componentes de adaptação também são considerados, por meio da utilização das características ecogeográficas do local de origem dos acessos (GALWEY, 1995). A associação preditiva entre determinadas características e variáveis ecogeográficas da origem dos acessos também tem sido explorada na seleção de amostras de coleções de germoplasma de forma a aumentar a probabilidade de "captura" de características de interesse. Essa estratégia é denominada Estratégia de Identificação Focada de Germoplasma (FIGS) (MACKAY e STREET, 2004).

A FIGS se baseia na premissa de que o ambiente influencia a seleção natural e consequentemente a distribuição geográfica dos genótipos, para "minerar" características úteis em bancos de germoplasma. Diversos estudos reconheceram o potencial dessa técnica na rápida identificação de genótipos com tolerância a seca e calor, resistência a insetos pragas e resistência a doenças (KHAZAEI et al., 2013; EL-BOUHSSINI et al., 2010; ENDRESEN et al. 2011; ENDRESEN et al., 2012; BARI et al., 2012). A FIGS é fundamentada em técnicas de modelagem para associar determinadas características à variáveis ecogeográficas (ou ambientais). Especificamente os métodos baseados em aprendizado de máquina tem sido utilizados com frequência na FIGS devido a sua capacidade de modelar padrões complexos e não lineares de interação (ICARDA, 2013).

A necessidade de genótipos superiores, com características desejadas, em menor intervalo de tempo e com menor custo, exige que a exploração dos recursos genéticos seja mais eficiente principalmente para cultura emergentes como o cártamo.

O presente trabalho objetiva explorar a diversidade genética de cártamo por meio do estabelecimento de coleções nucleares mais robustas utilizando as estratégias de maximização e estratificação de genótipos em grupos genéticos conhecidos. O trabalho também investiga a existência de associação preditiva entre teor de óleo e variáveis ecogeográficas da origem de acessos de cártamo, utilizando a FIGS baseada em abordagens de aprendizado de máquina, com o objetivo de explorar a associação para aumentar as chances de encontrar genótipos de cártamo com alto teor óleo.

# REFERÊNCIAS


BARI, A.; STREET, K.; MACKAY, M.; ENDRESEN, D. T. F.; DE PAUW, E.; AMRI, A. Focused identification of germplasm strategy (FIGS) detects wheat steam rust resistance linked to environmental variables. **Genetic Resources and Crop Evolution**, v. 59, p. 1456-1481. 2012.

BRADLEY, V. L.; JOHNSON, R. C. Managing the U.S. safflower collection. In: BERGMAN, J. W.; MUNDEL, H. H. (Ed.). **Proceedings of the 5th International Safflower Conference**, Williston and Sidney, United States. 2001. p. 143–147.

BROWN, A. H. D. The core collection at the crossroads. In: HODGKIN, T. B., A. H. D.;HINTUM, T. J. L. VAN;MORALES, E. A. V. (Ed.). **Core collections of plant genetic resources**. 1995. p. 3-20.

DAJUE L.; MÜNDEL, H. H. **Safflower. Carthamus tinctorius L**. Promoting the conservation and use of underutilized and neglected crops 7. Institute of Plant Genetics and Crop Plant Research, Gatersleben/International Plant Genetic Resources Institute, Rome, Italy. 1996. 83 p.

EL-BOUHSSINI, M.; STREET, K.; AMRI, A.; MACKAY, M.; OGBONNAYA, F. C.; OMRAN, A.; ABDALLA, O.; BAUM, M.; DABBOUS, A.; RIHAWI, F. Sources of resistance in bread wheat to Russian wheat aphid (Diuraphis noxia) in Syria identified using the Focused Identification of Germplasm Strategy (FIGS). **Plant Breeding**, v. 130, p. 96-97. 2010.

ENDRESEN, D. T. F.; STREET, K.; MACKAY, M; BARI, A; DE PAUW, E. Predictive association between biotic stress traits and ecogeographic data for wheat and barley landraces. **Crop Science**, v. 51, p. 2036-2055. 2011.

ENDRESEN, D. T. F.; STREET, K.; MACKAY, M.; BARI, A.; DE PAUW, E.; NAZARI, K.; YAHYAOUI, A. Sources of Resistance to Stem Rust (Ug99) in Bread Wheat and Durum

Wheat Identified Using Focused Identification of Germplasm Strategy (FIGS). **Crop Science**, v. 52, n. 2, p. 764-773. 2012.

FAOSTAT Database, Food and Agriculture Organization of the United Nations. **"World Safflower Production and Import/Export Data"**. Disponível em: <http://faostat.fao.org/site/567/default.aspx#ancor> Acesso em 21 de março de 2012.

GALWEY, N. W. Verifying and validating representativeness of a core collection. In: HODGKIN, T.; BROWN, A. H. D.; VAN HINTUM, T. J. L.; MORALES, E. A. V. (Eds.). **Core collections of plant genetic resources.** New York: J. Wiley. 1995. p. 187-198.

INTERNATIONAL CENTER FOR AGRICULTURAL RESEARCH IN THE DRY AREAS (ICARDA). A new approach to mining agricultural gene banks – to speed the pace of research innovation for food security. 'FIGS' - the Focused Identification of Germplasm Strategy. **Research to Action 3**. ICARDA: Beirut, Lebanon. 2013. 22p.

JOHNSON, R.C.; STOUT, D. M.; BRADLEY, V. L. The US Collection: a rich source for safflower germplasm. In: DAJUE, L.; YUANZHOU, H. (Eds.). **Proceedings of the Third International Safflower Conference**, Beijing Botanical Garden, Institute of Botany. Chinese Academy of Sciences, Beijing, China. 1993. p. 202–208.

JOHNSON, R.C.; GHORPADE, P. B.; BRADLEY, V. L. Evaluation of the USDA core safflower collection for seven quantitative traits. In: BERGMAN, J. W.; MUNDEL, H. H. (Eds.). **Proceedings of the 5th International Safflower Conference**, Williston and Sidney, United States. 2001. p.149–152.

KHAZAEI, H.; STREET, K.; BARI, A.; MACKAY, M.; STODDARD, F. L. The FIGS (Focused Identification of Germplasm Strategy) Approach Identifies Traits Related to Drought Adaptation in Vicia faba Genetic Resources. **PLoS ONE 8(5): e63107. doi:10.1371/journal.pone.006310**. 2013.

MACKAY, M. C.; STREET, K. Focused identification of germplasm strategy – FIGS. In: BLACK, C. K.; PANOZZO, J. F.; REBETZKE, G. J. (Eds). **Proceedings of the 54th**

**Australian Cereal Chemistry Conference and the 11th Wheat Breeders' Assembly**. Royal Australian Chemical Institute, Melbourne, Australia. 2004. p. 138-141.

SINGH, V.; NIMBKAR, N. Safflower (Carthamus tinctorius L.). In: SINGH, R. J. (Ed.). **Oilseed Crops. Genetic Resources, Chromosome Engineering, and Crop Improvement Vol. 4.** Boca Raton: CRC Press, Taylor & Francis Group. 2007. p.167-194.

TRZECIAK, M. B.; NEVES, M. B.; VINHOLES, P. S.; VILLELA, F. A. Utilização de sementes de espécies oleaginosas para produção de biodiesel. **Informativo ABRATES**, v.18, p.30-38. 2008.

VAN HINTUM, T. J. L.; BROWN, A. H. D.; SPILLANE, C.; HODGKIN, T. **Core Collections of Plant Genetic Resources**. IPGRI Technical Bulletin No. 3. International Plant Genetic Resources Institute: Rome. 2000. 51 p.

# CAPÍTULO 1. ESTABELECIMENTO DE COLEÇÕES NUCLEARES DE CÁRTAMO

## RESUMO

Cártamo (Carthamus tinctorius L.) é uma espécie oleaginosa com um grande potencial genético disponível nos bancos de germoplasma. A solução para o uso mais eficiente e racional do germoplasma consiste no estabelecimento de coleções nucleares representativas da variabilidade genética existente. Para estabelecer coleções nucleares expressivas é necessário que os acessos sejam selecionados da coleção base de maneira apropriada, utilizando estratégias e abordagens oportunas às informações disponíveis. O presente estudo teve como objetivo estabelecer coleções nucleares de cártamo utilizando a estratégia de maximização e a estratificação dos genótipos em grupos genéticos. No estabelecimento das coleções nucleares foram utilizados caracteres fenotípicos, qualitativos e quantitativos, de 1640 acessos de cártamo provenientes de 48 países. Os acessos foram estratificados nos grupos genéticos de acordo com país de origem e amostrados segundo a estratégia de maximização (estratégia M). As coleções nucleares estabelecidas foram comparadas com a coleção base utilizando as seguintes estatísticas de validação: teste de qui-quadrado, diferença de médias, diferença de variâncias, taxa de coincidência, taxa variável e índice de Shannon. As magnitudes das estimativas das estatísticas de validação indicaram que a variabilidade genética dos acessos da coleção base foi preservada nas coleções nucleares estabelecidas de acordo com grupos genéticos fundamentados nos centros de similaridade do cártamo. As coleções nucleares estratificadas por grupos genéticos apresentaram aproximadamente 60 genótipos, com diferença média de apenas 7% em relação a coleção base e com taxa de coincidência de superior a 94%. O uso conjunto da estratégia de maximização e da estratificação dos genótipos em grupos genéticos maximizou a captação da variabilidade genética e introduziu maior eficiência no estabelecimento das coleções nucleares ao selecionar uma quantidade reduzida de acessos. As coleções nucleares estabelecidas incluíram aproximadamente 3,75% dos acessos conservados na coleção base.

**Palavras-chave:** Coleção nuclear, Cártamo, Estratégia de maximização, Grupos genéticos.

# ESTABLISHMENT OF SAFFLOWER CORE COLLECTIONS

## ABSTRACT

Safflower (Carthamus tinctorius L.) is an oilseed species with a large genetic potential available in genebanks. Establishing core collections representative of genetic variability is the most efficient and rational use of safflower germplasm. To establish expressive core collections is necessary to select properly accessions from base collection, using appropriate strategies and approaches. The present study aimed to establish core collections of safflower using maximization strategy and the stratification of genotypes in genetic groups. Core collections were established using phenotypic qualitative and quantitative traits data of 1640 safflower accessions from 48 countries. The accessions were stratified into genetic groups according to country's origin and sampled according to the maximization strategy (M strategy) . The established core collections were compared with the base collection using the following validation statistics: chi-square test, mean difference, difference of variances, coincidence rate, variable rate and Shannon index. Magnitude estimates of validation statistics indicated that base collection's genetic variability was preserved in the core collections based on safflower centers of similarity. Core collections stratified by genetic groups consisted in about 60 genotypes, with a mean difference of 7 % over the base collection and coincidence rate above 94%. The combined use of the maximization strategy and stratification of genotypes in genetic groups maximized the capture of genetic variation and introduced more efficiency, establishing core collections with a fewer number of accessions. The core collections included approximately 3,75% accessions conserved in safflower base collection .

**Keywords:** Core collection, Safflower, Maximization Strategy, Genetic Groups.

# 1. INTRODUÇÃO

Cártamo (Carthamus tinctorius L.) é uma espécie da família Compositae, anual, diploide (2n = 2x = 24) e predominantemente autógama. Embora o cártamo seja uma espécie com uma ampla variedade de usos, tem sido cultivado principalmente em função dos vários usos do óleo extraído de suas sementes. A alta qualidade desse óleo faz com que esse seja utilizado na indústria alimentícia, na indústria farmacêutica e na produção de biocombustível (MÜNDEL e BERGMAN, 2009). Para atender a essa demanda diversificada de usos é indispensável que variedades desenvolvidas de cártamo sejam mais apropriadas às necessidades de cada situação.

Os bancos de germoplasma tem por objetivo ser a principal fonte de variação genética para impulsionar o desenvolvimento de diversas variedades de cártamo. Os acessos conservados e distribuídos nesses bancos são utilizados principalmente com o propósito de fornecer novos alelos para caracteres de interesse. Hamdan et al. (2009) demonstraram a utilização de acessos de cártamo no desenvolvimento de linhagens com altos teores de ácidos graxos saturados, potencialmente úteis para indústria alimentícia. Linhagens com altos teores de ácidos palmítico e esteárico, foram desenvolvidas a partir dos acessos PI 20686 e PI 198990. Outras linhagens com maior teor de alfa-tocoferol (vitamina E) também foram derivadas de acessos (PI 304597 e PI 406001) disponíveis na coleção de germoplasma do Departamento de Agricultura dos Estados Unidos (USDA) (VELASCO e FERNÁNDEZ-MARYINEZ, 2004).

As coleções de germoplasma de cártamo contém uma grande quantidade de acessos, distribuídos em 18 diferentes coleções em 14 países. As maiores coleções de cártamo encontram-se na China, Índia e Estados Unidos. Essas coleções contém mais de 10.000 acessos, representados por genótipos selvagens, "landraces", linhagens e variedades (DAJUE e MÜNDEL, 1996). No entanto, o grande número de acessos impossibilita o uso prático e imediato desse germoplasma (ZHANG e JOHNSON, 1999).

O estabelecimento de coleções nucleares apresenta-se como a forma mais adequada de utilizar o germoplasma disponível. O conceito de coleção nuclear foi proposto por Frankel (1984) com o objetivo de representar a diversidade da coleção total (ou coleção base) através de uma amostra representativa dessa coleção. O estabelecimento de coleções nucleares consiste na seleção de uma quantidade limitada de acessos representativos da coleção base, com o objetivo de representar grande fração da diversidade genética. De fato, o

estabelecimento de coleções nucleares também representa uma estratégia de baixo custo para melhorar a caracterização e utilização do germoplasma (FRANKEL e BROWN, 1984).

Os procedimentos básicos utilizados no estabelecimento da coleção nuclear são: identificação e caracterização dos acessos, estratificação dos acessos em grupos, escolha do número de acessos a serem selecionados de cada grupo e seleção dos acessos que serão incluídos na coleção nuclear (VAN HINTUM et al., 2000).

A estratificação dos acessos em grupos representativos é um procedimento essencial para o estabelecimento da coleção nuclear. A estratificação consiste na separação dos acessos em grupos contrastantes, de modo a maximizar a variação entre grupos e minimizar a variação dentro dos grupos. Esse procedimento é indispensável para amostragem de acessos distintos e que preservam as características do grupo de origem. Diferentes tipos de caracteres podem ser utilizados na estratificação dos acessos, porém esses devem refletir diferenças genéticas consistentes. Por isso, diversos estudos sugerem a estratificação dos acessos em grupos (ou "pools") genéticos baseados em diferenças geográficas pois esses refletem a adaptação dos acessos às diversas condições ambientais (ORTIZ, et al., 1998; VAN HINTUM et a., 2000; ESCRIBANO et al., 2008; OLIVEIRA et al., 2010).

No estabelecimento de coleções nucleares de cártamo existe a possibilidade de utilizar os grupos genéticos, fundamentados nos "centros de similaridade", para estratificação dos acessos. Os centros de similaridade foram identificados e sugeridos por Knowles (1969) segundo a uniformidade da expressão fenotípica de determinados caracteres de acessos coletados em regiões específicas. Sete centros foram definidos: Extremo oriente, Índia-Paquistão, Oriente médio, Egito, Sudão, Etiópia e Europa. Nesses centros há considerável diversidade fenotípica em caracteres de grande importância, como altura de planta, produtividade, ramificação e peso de sementes. Apenas com a utilização de marcadores genéticos evidenciaram a existência de oito grupos genéticos distintos provenientes dos centros de similaridade (JOHNSON et al., 2007). No entanto, Chapman et al. (2010) analisaram a representatividade desses centros de similaridade e determinaram a existência de apenas cinco grupos genéticos. A discordância entre os estudos anteriores foi resultante do uso de diferentes genótipos e marcadores moleculares. Embora o verdadeiro número de centros de similaridade ainda não seja conhecido, existe o consenso de que esses centros contém diversidade genética passível de exploração (CHAPMAN et al., 2010).

Após a estratificação em grupos distintos, é necessário que os acessos sejam amostrados de maneira eficiente. A estratégia de seleção utilizada na amostragem deve ser

capaz de representar em poucos acessos a variabilidade existente nos grupos. Existem várias estratégias de seleção, baseadas em número de acessos, distâncias genéticas, riqueza genotípica e conhecimento informal (VAN HINTUM et al., 2000). Atualmente, a estratégia de maximização (M) é uma das estratégias mais eficientes na amostragem de acessos. Essa estratégia consiste na seleção de combinações específicas de acessos que contém o máximo da diversidade da coleção base. Por meio de um processo iterativo de seleção, conjuntos de acessos são formados segundo o critério de riqueza de classes, ou seja, cobertura da amplitude de variação das características dos acessos. Dessa forma, essa estratégia procura maximizar o número de classes de amostradas em uma combinação de acessos, mantendo um mínimo de redundância (KIM et al. 2007).

Estudos realizaram a comparação da estratégia de maximização com estratégias baseadas em dendogramas e distâncias genéticas, os resultados apresentados mostraram maior representatividade genética nas coleções estabelecidas pela estratégia de maximização (ESCRIBANO et al., 2008; GOPAL et al., 2013). Estudos de simulação também indicaram que a estratégia de maximização apresenta melhor performance em situações de autofecundação e fluxo gênico restrito (BATAILLON et al., 1996).

Uma vez que as coleções nucleares de cártamo desenvolvidas até o momento não exploraram o potencial de diferentes estratégias de amostragem e estratificação (DWIVEDI et al., 2005), o presente estudo teve como objetivo o estabelecimento de coleções nucleares de cártamo utilizando a estratégia de maximização e a estratificação dos genótipos em grupos genéticos propostos por Johnson et al. (2007) e Chapman et al. (2010).

## 2. MATERIAIS E MÉTODOS

Os acessos utilizados no presente estudo compõe a coleção de germoplasma (coleção base) de cártamo do Departamento de Agricultura dos Estados Unidos (USDA). Toda a coleção consiste de aproximadamente 2400 acessos, porém menos de 2000 acessos estão totalmente caracterizados. Assim, apenas 1640 acessos (genótipos) de diversas origens e devidamente caracterizados foram utilizados para o estabelecimento das coleções nucleares (Tabela 1). As características avaliadas nos genótipos são resultado de estudos de caracterização de germoplasma conduzidos entre 1988 e 2010 (JOHNSON et al., 1999; DAJUE et al., 1993; USDA, 2013).

**Tabela 1**. Origem e número de acessos de cártamo na coleção de germoplasma de cártamo

| Origem | Número | Origem | Número | Origem | Número | Origem | Número | Origem | Número |
|---|---|---|---|---|---|---|---|---|---|
| Índia | 690 | Espanha | 17 | Quênia | 8 | Japão | 3 | Dinamarca | 1 |
| Iran | 194 | Austrália | 13 | Marrocos | 8 | Estados Unidos | 3 | Grécia | 1 |
| China | 192 | Etiópia | 13 | Hungria | 6 | Argentina | 2 | Líbano | 1 |
| Turquia | 110 | Bangladesh | 12 | Itália | 6 | Armênia | 2 | Líbia | 1 |
| Paquistão | 85 | Síria | 11 | Cazaquistão | 6 | Bulgária | 2 | México | 1 |
| Egito | 63 | Jordânia | 10 | Polônia | 5 | Romênia | 2 | Holanda | 1 |
| Afeganistão | 29 | Alemanha | 9 | Reino Unido | 5 | Tajiquistão | 2 | Suíça | 1 |
| Portugal | 28 | Uzbequistão | 9 | Iraque | 4 | Tailândia | 2 | Ucrânia | 1 |
| Israel | 26 | Rússia | 8 | Áustria | 3 | Argélia | 1 | Desconhecida | 7 |
| Sudão | 24 | França | 8 | Eritreia | 3 | Canadá | 1 | **Total** | **1640** |

**Tabela 2.** Caracteres qualitativos e classes avaliadas nos acessos de cártamo

| Caracteres | Classes |
|---|---|
| Ângulo do ramos | 0 = Ausência, 3 = Apresso (15º a 20º), 6 = Intermediário (20º a 60º), 7 = Difuso (60º a 90º), 9 = Pendente ( > 90º) |
| Ramificação | 0 = Ausência, 1 = Basal, 2 = No terço inferior da planta, 3 = No terço médio, 4 = Da base ao ápice da planta |
| Cor da corola no florescimento | 1= Branco, 2 = Amarelo claro, 3 = Amarelo, 4 = Amarelo-laranja, 5 = Laranja-vermelho, 6 = Vermelho, 7 = Roxo, 8 = Outros |
| Forma do capítulo | 1 = Cônico, 2 = Oval, 3 = Achatado |
| Borda foliar | 1 = Lisa , 2 = Serrada ou dentada, 3 = Profundamente Serrada, 4 = Partida |
| Forma da folha | 1 = Ovalada, 2 = Obovada, 3 = Lanceolada, 4 = Ovalada-oblonga, 5 = Oblonga |
| Distribuição espinhos nas folhas | 0 = Ausência, 3 = Poucos, 5 = Intermediário, 7 = Muitos |
| Localização de espinhos em brácteas | 1 = Na ponta, 2 = Na ponta-apical, 3 = Na ponta-basal, 4 =Na ponta e nas margens, 5 = Apenas nas margens |
| Salinidade | 3 = Baixa suscetibilidade, 5 = Média suscetibilidade, 7 = Alta suscetibilidade |
| Forma da semente | 1 = Oval, 2 = Cônica, 3 = Crescente |
| Dormência de semente | 1 = Presente, 2 = Ausente |
| Tipo de tegumento | 1 = Normal, 2 = Estriado, 3 = Reduzido, 4 = Fino, 5 = Parcial |
| Comprimento de espinhos nas brácteas | 0 = Ausência, 3 = Curtos, 5 = Intermediários, 7 = Longos |
| Quantidade de espinhos nas brácteas | 0 = Ausência, 3 = Poucos, 5 = Intermediário, 7 = Muitos |
| Intensidade de espinhos nas folhas | 0 = Ausência, 3 = Poucos, 5 = Intermediário, 7 = Muitos |

Nos estudos os genótipos de cártamo foram avaliados em experimentos a campo no Centro de Pesquisa Agrícola Oriental da Universidade do Estado de Montana em Sidney, Montana (47º43'34''N, 104º09'W) e no Jardim Botânico de Beijing - Academia Chinesa de Ciências (39º33'N; 116º16'E). Os genótipos foram plantados em fileiras únicas de 6 metros de comprimento com espaçamento de 60 centímetros. Os caracteres avaliados consistiram de descritores morfológicos e caracteres agronômicos, classificados em duas categorias, caracteres qualitativos e quantitativos.

Os caracteres qualitativos e suas classes estão apresentados na Tabela 2. Os caracteres quantitativos foram: tamanho de capítulo (mm), número de capítulos por planta, tegumento na semente (%), teor de ácido linoleico (%), teor de ácido oleico (%), teor de ácido palmítico (%), teor de ácido esteárico (%), teor de óleo (%), teor de proteína (%), altura de planta (cm), dias para maturidade, peso de 100 sementes (gramas) e produção de planta (gramas). No presente estudo foram utilizados as médias das características quantitativas de cada acesso para estabelecimento das coleções nucleares.

## Estratificação e Amostragem

Antes do estabelecimento das coleções os acessos foram estratificados nos grupos genéticos propostos por Johnson et al. (2007) e Chapman et al. (2010), de acordo com os países de origem (Tabela 3). Os acessos provenientes de regiões não observadas pelos estudos anteriores e de origem desconhecida foram mantidos nas análises e classificados no grupo "Outros".

**Tabela 3**. Número de acessos de cártamo nos grupos genéticos propostos por Johnson et al. (2007) e Chapman et al. (2010)

| **Johnson et al. (2007)** | | | **Chapman et al. (2010)** | | |
|---|---|---|---|---|---|
| **Grupos** | | **Número de acessos** | **Grupos** | | **Número de acessos** |
| 1 | Oriente Médio | 187 | 1 | Oriente Próximo | 49 |
| 2 | Egito, Sudão e Quênia | 95 | 2 | Iran, Afeganistão, Iraque e Turquia | 337 |
| 3 | Etiópia e Eritreia | 16 | 3 | Egito, Etiópia e Sudão | 103 |
| 4 | Afeganistão | 29 | 4 | Extremo Oriente | 984 |
| 5 | Europa | 86 | 5 | Europa | 100 |
| 6 | Índia e Bangladesh | 702 | 6 | Outros | 67 |
| 7 | Paquistão, Iran e Iraque | 283 | | | |
| 8 | China | 192 | | | |
| 9 | Outros | 50 | | | |
| **Total** | | **1640** | **Total** | | **1640** |

Dessa forma, foram empregadas as seguintes formas de estratificação dos genótipos para o estabelecimento de coleções nucleares: a) Sem estratificação (CC1); b) Estratificação por países de origem dos acessos (CC2); c) Estratificação por grupos genéticos propostos por Chapman et al. (2010) (CC3); d) Estratificação por grupos genéticos propostos por Johnson et al. (2007) (CC4); e) Estratificação combinada por grupos genéticos (JOHNSON et al., 2007; CHAPMAN est al., 2010) (CC5); f) Amostragem aleatória de acessos (Aleatória). A ausência de estratificação e a amostragem aleatória foram utilizadas para avaliar a eficiência das estratificações por grupos genéticos.

Após a estratificação dos acessos foram realizadas as construções das coleções utilizando a estratégia de maximização (estratégia M) proposta por Shoen e Brown (1993). Essa estratégia busca maximizar a diversidade genotípica na coleção nuclear pela seleção de acessos com alta riqueza de classes e baixa redundância. Originalmente desenvolvida para marcadores moleculares, a estratégia M pode ser estendida para caracteres qualitativos e quantitativos pela simples categorização das observações desses caracteres em classes discretas. Desse modo, o número de classes representadas entre os acessos define a riqueza da coleção nuclear.

Os caracteres quantitativos foram convertidos em categóricos utilizando a regra de Sturges (STURGES, 1926). A distribuição contínua dos caracteres é dividida em $k$ intervalos de classes (ou categorias) pela seguinte fórmula:

$$k = 1 + \log_2 n$$

onde n é o número de acessos da coleção.

Realizada a categorização dos caracteres, a construção de cada coleção foi iniciada utilizando o seguinte algoritmo heurístico baseado em maximização iterativa: i) Inicialmente um conjunto de acessos de tamanho n é amostrado ao acaso da coleção base de tamanho N; ii) Posteriormente, cada acesso é retirado do conjunto (n-1) e a riqueza de classes no conjunto remanescente é determinada. O conjunto com maior nível de riqueza total é mantido; iii) Em seguida, entre os acessos remanescentes do conjunto, o acesso que fornece o maior aumento na riqueza é retido na coleção, resultando em uma coleção de tamanho n. Esses passos são repetidos até a formação de uma coleção cuja riqueza não pode ser mais melhorada (BATAILLON et al., 1996).

**Estatísticas de validação**

As coleções nucleares estabelecidas foram validadas pela comparação com a coleção base utilizando todos os caracteres qualitativos e quantitativos. No processo de validação das coleções foram utilizados parâmetros estatísticos para estimar a acurácia com que as coleções nucleares representaram a coleção base. Esses parâmetros foram baseados na análise das médias, variâncias e distribuições das características nas coleções. Para verificar se a frequência dos caracteres nos acessos das coleções nucleares permaneceu similar à frequência observada na coleção base foi utilizado o teste de qui-quadrado (SNEDECOR e COCHRAN, 1980). Os caracteres quantitativos, após a conversão em categóricos, também foram submetidos ao teste de qui-quadrado. Os valores de qui-quadrado foram calculados da seguinte forma:

$$\chi^2 = \sum_{i=1}^{k} \frac{(CQt_i - WCQt_i)^2}{WCQt_i}$$

onde $CQt_i$ é a quantidade relativa do acesso da categoria i (i = 1, 2, ...., k) na coleção nuclear e $WCQt_i$ é a quantidade relativa dos acessos da categoria i na coleção base. O número de graus de liberdade utilizado foi o número de classes menos um.

As médias e variâncias de cada um dos 13 caracteres quantitativos, não convertidos em categóricos, das coleções nucleares foram comparadas com as médias e variâncias observadas na coleção base, utilizando os testes t e F (SNEDECOR e COCHRAN, 1980). Além dessa comparação individual, também foram calculadas as diferenças entre as coleções utilizando os caracteres quantitativos de forma conjunta. As seguintes estatísticas foram estimadas: diferença média (MD,%), diferença de variância (VD,%), taxa de coincidência (CR,%) e taxa variável (VR,%). Essas foram calculadas conforme as seguintes fórmulas:

$$MD\ (\%) = \frac{1}{m}\sum_{j=1}^{m}\frac{|Me-Mc|}{Mc}\ x\ 100 \qquad VD\ (\%) = \frac{1}{m}\sum_{j=1}^{m}\frac{|Ve-Vc|}{Vc}\ x\ 100$$

$$CR\ (\%) = \frac{1}{m}\sum_{j=1}^{m}\frac{Rc}{Re}\ x\ 100 \qquad VR\ (\%) = \frac{1}{m}\sum_{j=1}^{m}\frac{CVc}{CVe}\ x\ 100$$

onde Me: média na coleção base; Mc: média da coleção nuclear; Ve: variância da coleção base; Vc: variância da coleção nuclear; Re: amplitude das classes na coleção base; Rc: amplitude das classes na coleção nuclear; CVe: coeficiente de variação na coleção base; CVc: coeficiente de variação na coleção nuclear; m: número de caracteres (HU et al., 2000).

Para os caracteres qualitativos o critério utilizado na avaliação das coleções foi o índice de diversidade de Shannon (SHANNON, 1948) calculado da seguinte forma:

$$H = -\sum_{i=1}^{k} p_i \log(p_i)$$

onde $p_i$ é a frequência da classe e k é o total de classes. Os valores do índice H de cada um dos caracteres das coleções nucleares foram comparados com o máximo valor possível para o caráter (log(n), onde n é o número de classes do caráter na coleção base).

Todas as análises foram realizadas nos aplicativos computacionais R (R DEVELOPMENT CORE TEAM, 2013) e PowerCore (KIM et al., 2007).

## 3. RESULTADOS

Todas as coleções nucleares estabelecidas pela estratégia de maximização apresentaram menos de 10% do total de acessos da coleção base (Tabela 4).

**Tabela 4.** Estimativas das estatísticas de validação das coleções nucleares

| Grupo | Número de acessos | MD (%) | VD(%) | CR(%) | VR(%) |
|---|---|---|---|---|---|
| CC1 | 62 | 7.33 | 57.63 | 94.96 | 157.41 |
| CC2 | 89 | 7.00 | 49.14 | 95.64 | 140.86 |
| CC3 | 61 | 6.52 | 57.36 | 94.97 | 155.53 |
| CC4 | 60 | 6.90 | 57.21 | 94.95 | 155.08 |
| CC5 | 91 | 5.54 | 47.90 | 95.27 | 140.62 |
| Aleatória | 118 | 4.04 | 41.88 | 95.14 | 131.45 |

Diferença média (MD,%), diferença de variância (VD,%), taxa de coincidência (CR,%) e taxa variável (VR,%).

A estratificação dos genótipos em grupos genéticos permitiu o estabelecimento de coleções nucleares (CC3 e CC4) representativas da coleção base, com um menor número de acessos. As médias dos caracteres quantitativos dessas coleções foram similares ao observado

na coleção base, com diferença média (MD,%) menor que 7%. As estimativas das diferenças de variâncias (VD,%) e taxas variáveis (VR,%) também indicaram a presença de maior diversidade nessas coleções. As taxas de coincidência (CR,%) também demonstraram uma ampla cobertura das classes presentes na coleção base.

As coleções nucleares CC3 e CC4 tiveram uma performance média similar a coleção nuclear estabelecida sem estratificação (CC1). A diversidade captada pelas coleções foi equivalente, mas as diferenças médias (MD,%) e a quantidades de acessos foram menores em CC3 e CC4. A coleção nuclear estratificada por países de origem dos acessos (CC2) apresentou a maior taxa de coincidência devido à amostragem de acessos de todos os países. Porém, essa maior representatividade não refletiu em maior diversidade (VD e VR). Da mesma forma, a coleção nuclear estabelecida pelo uso conjunto dos grupos genéticos e dos países (CC5) foi similar à coleção CC2, apresentando o maior número de acessos e a menor diferença média em relação à coleção base.

A maioria das médias de cada um dos 13 caracteres quantitativos das coleções nucleares não diferiu das respectivas médias da coleção base (Tabela 5). A exceção do caráter tamanho de capítulo, as variâncias apresentaram-se significativas com magnitudes superiores à coleção base. Em todas as coleções nucleares, apenas os caracteres teor de ácido linoleico (%) e teor de ácido oleico (%) tiveram médias e variâncias estatisticamente diferentes da coleção da base. De modo geral, as médias e variâncias apresentaram magnitudes coerentes com a estratégia de maximização, ou seja, nas coleções nucleares os caracteres apresentaram médias similares e variâncias superiores às encontradas na coleção base.

Os caracteres qualitativos das coleções nucleares foram avaliados pelo índice de diversidade de Shannon. Os índices dos caracteres nas coleções nucleares foram maiores que os observados na coleção completa (Tabela 6). No entanto, a análise dos índices não deve ser realizada pela simples comparação com os índices da coleção base. A magnitude dos índices de Shannon nas coleções CC3 e CC4 indica a superioridade dessas coleções em relação as outras. Em contraste, os menores índices na coleção aleatória atestam o desequilíbrio na frequência das classes amostradas e a possibilidade de perda de diversidade com a seleção aleatória de acessos.

A análise dos desvios de distribuição de frequências de classes também é uma forma de avaliar a qualidade das coleções. A significância dos testes de qui-quadrado mostraram que as distribuições de frequências das classes nas coleções nucleares foram diferentes das observadas na coleção base, para a maioria dos caracteres (Tabela 7). A ocorrência de desvios

significativos contraria as estatísticas apresentadas anteriormente e indica que as coleções nucleares estabelecidas não representaram a coleção base em termos de distribuição de classes. De fato, esses resultados eram esperados pois a estratégia de maximização seleciona acessos de modo a maximizar a frequência das classes de cada caráter. Em outras palavras, as classes menos representativas dos caracteres tiveram suas frequências aumentadas nas coleções nucleares e causaram maior desvio em relação a coleção base. Esses resultados podem ser constatados pelos maiores valores de qui-quadrado de alguns caracteres.

A inspeção visual da distribuição de frequências de alguns caracteres permite examinar as distribuições de forma mais clara do que a simples avaliação dos resultados dos testes de qui-quadrado (Figura 1). A distribuição das frequências das classes de teor de óleo e de produtividade das coleções nucleares foram condizentes com a distribuição observada na coleção base, mas foram consideradas significativas e não significativas, respectivamente, pelo teste de qui-quadrado. As causas dos desvios significativos foram a menor redundância de classes amostradas e o aumento da frequência de classes menos representativas, constatadas pela forma dos gráficos de distribuição de frequências, com densidades menores em torno da média e extremidades maiores.

**Tabela 5.** Número de acessos, médias e variâncias de 13 caracteres quantitativos de cartámos na coleção base e nas seis coleções nucleares

| Caracteres | Coleção completa | | | CC1 | | | | CC2 | | | | CC3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acessos | Média | Variância | Acessos | Média | | Variância | | Acessos | Média | | Variância | | Acessos | Média | | Variância | |
| Tamanho de capítulo (mm) | 1588 | 19.395 | 15.718 | 60 | 19.492 | ns | 18.487 | ns | 88 | 19.674 | ns | 18.086 | ns | 59 | 19.328 | ns | 18.791 | ns |
| Capítulos (quantidade/planta) | 1634 | 20.862 | 120.604 | 62 | 24.131 | ns | 331.083 | ** | 88 | 23.539 | ns | 254.069 | ** | 61 | 23.426 | ns | 322.315 | ** |
| Tegumento na semente (%) | 1421 | 51.013 | 62.129 | 54 | 51.918 | ns | 97.143 | ** | 77 | 51.787 | ns | 86.670 | ** | 53 | 51.443 | ns | 111.584 | ** |
| Ácido linoleico (%) | 1640 | 75.344 | 50.103 | 62 | 69.332 | * | 276.939 | ** | 89 | 71.615 | * | 183.092 | ** | 61 | 70.191 | * | 241.445 | ** |
| Maturidade (dias) | 1624 | 121.438 | 27.078 | 61 | 119.672 | ns | 90.557 | ** | 87 | 120.326 | ns | 72.540 | ** | 61 | 119.787 | ns | 87.004 | ** |
| Teor de óleo (%) | 1614 | 28.402 | 17.837 | 61 | 28.033 | ns | 38.632 | ** | 86 | 27.146 | ns | 28.603 | ** | 60 | 27.738 | ns | 35.863 | ** |
| Ácido oleico (%) | 1638 | 15.733 | 43.100 | 62 | 20.574 | * | 251.127 | ** | 89 | 18.605 | * | 156.312 | ** | 61 | 19.674 | * | 210.068 | ** |
| Ácido palmítico (%) | 1620 | 7.390 | 2.919 | 62 | 7.987 | ns | 14.973 | ** | 89 | 7.766 | ns | 10.309 | ** | 61 | 8.011 | ns | 14.309 | ** |
| Altura planta (cm) | 1589 | 67.962 | 267.797 | 62 | 66.869 | ns | 568.649 | ** | 88 | 70.792 | ns | 470.505 | ** | 61 | 67.918 | ns | 558.610 | ** |
| Teor de proteína (%) | 1491 | 17.595 | 3.024 | 56 | 17.851 | ns | 6.200 | ** | 83 | 17.408 | ns | 5.271 | ** | 55 | 17.715 | ns | 5.994 | ** |
| Peso de 100 sementes (g) | 1631 | 4.801 | 1.393 | 62 | 4.643 | ns | 2.854 | ** | 89 | 4.474 | * | 2.161 | ** | 61 | 4.652 | ns | 2.912 | ** |
| Ácido esteárico (%) | 1365 | 1.527 | 0.451 | 56 | 1.771 | ns | 1.211 | ** | 82 | 1.763 | ns | 1.273 | ** | 55 | 1.793 | ns | 1.362 | ** |
| Produtividade (g/planta) | 1621 | 17.767 | 180.283 | 61 | 21.515 | ns | 351.973 | ** | 87 | 21.819 | ns | 347.570 | ** | 60 | 20.503 | ns | 348.979 | ** |

| Caracteres | Coleção completa | | | CC4 | | | | CC5 | | | | Aleatória | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acessos | Média | Variância | Acessos | Média | | Variância | | Acessos | Média | | Variância | | Acessos | Média | | Variância | |
| Tamanho de capítulo (mm) | 1588 | 19.395 | 15.718 | 59 | 19.800 | ns | 17.925 | ns | 91 | 19.560 | ns | 20.516 | ns | 115 | 19.525 | ns | 19.431 | ns |
| Capítulos (quantidade/planta) | 1634 | 20.862 | 120.604 | 60 | 23.450 | ns | 319.133 | ** | 91 | 22.747 | ns | 256.635 | ** | 118 | 21.890 | ns | 223.278 | ** |
| Tegumento na semente (%) | 1421 | 51.013 | 62.129 | 52 | 51.750 | ns | 106.462 | ** | 78 | 51.451 | ns | 83.761 | ** | 103 | 51.966 | ns | 87.264 | ** |
| Ácido linoleico (%) | 1640 | 75.344 | 50.103 | 60 | 70.474 | * | 244.370 | ** | 91 | 71.778 | * | 172.693 | ** | 118 | 72.827 | * | 146.926 | ** |
| Maturidade (dias) | 1624 | 121.438 | 27.078 | 60 | 119.850 | ns | 89.316 | ** | 90 | 120.648 | ns | 73.942 | ** | 117 | 120.805 | ns | 57.577 | ** |
| Teor de óleo (%) | 1614 | 28.402 | 17.837 | 60 | 27.300 | ns | 36.553 | ** | 88 | 27.121 | ns | 27.374 | ** | 117 | 27.788 | ns | 26.049 | ** |
| Ácido oleico (%) | 1638 | 15.733 | 43.100 | 60 | 19.357 | * | 207.868 | ** | 91 | 18.453 | * | 147.657 | ** | 118 | 17.890 | * | 128.622 | ** |
| Ácido palmítico (%) | 1620 | 7.390 | 2.919 | 60 | 8.040 | ns | 14.380 | ** | 91 | 7.777 | ns | 10.619 | ** | 118 | 7.613 | ns | 7.876 | ** |
| Altura planta (cm) | 1589 | 67.962 | 267.797 | 60 | 69.217 | ns | 536.139 | ** | 91 | 69.451 | ns | 468.561 | ** | 115 | 70.280 | ns | 349.519 | ** |
| Teor de proteína (%) | 1491 | 17.595 | 3.024 | 53 | 17.795 | ns | 5.903 | ** | 85 | 17.477 | ns | 5.001 | ** | 110 | 17.630 | ns | 4.945 | ** |
| Peso de 100 sementes (g) | 1631 | 4.801 | 1.393 | 60 | 4.732 | ns | 2.987 | ** | 91 | 4.463 | * | 2.097 | ** | 118 | 4.598 | ns | 2.038 | ** |
| Ácido esteárico (%) | 1365 | 1.527 | 0.451 | 56 | 1.783 | ns | 1.339 | ** | 80 | 1.679 | ns | 1.072 | ** | 104 | 1.601 | ns | 0.859 | ** |
| Produtividade (g/planta) | 1621 | 17.767 | 180.283 | 59 | 21.165 | ns | 414.380 | ** | 90 | 19.958 | ns | 305.728 | ** | 115 | 19.854 | ns | 281.023 | ** |

ns – não significativo ao nível P > 0.05; * significativo a P<0.05 pelos testes F e t de Student; ** significativo a P<0.01 pelos testes F e t de Student.

**Tabela 6**. Índice de Shannon dos caracteres qualitativos nas coleções nucleares e na coleção base

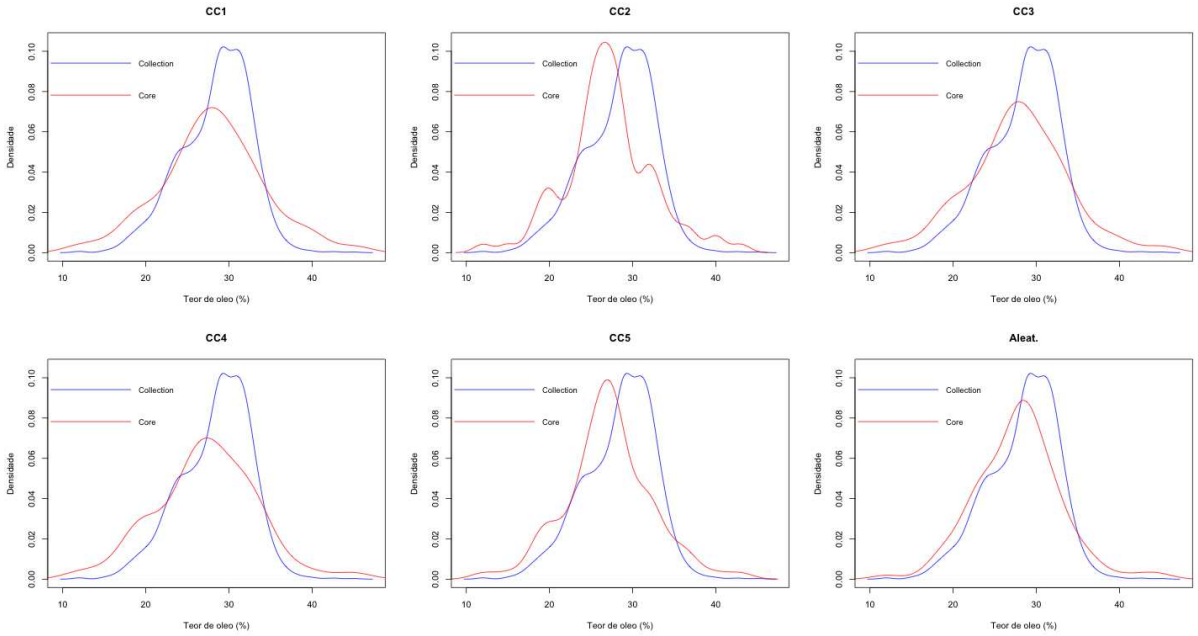
| Caracteres | Índice de Shannon | | | | | | |
|---|---|---|---|---|---|---|---|
| | Coleção base | CC1 | CC2 | CC3 | CC4 | CC5 | Aleatório |
| Ângulo do ramos | 0.885 | 1.180 | 1.096 | 1.207 | 1.230 | 1.111 | 0.980 |
| Ramificação | 0.900 | 1.040 | 0.904 | 1.019 | 0.989 | 0.936 | 0.900 |
| Cor da corola no florescimento | 0.996 | 1.395 | 1.311 | 1.398 | 1.435 | 1.339 | 1.256 |
| Forma do capítulo | 0.295 | 0.471 | 0.471 | 0.418 | 0.444 | 0.483 | 0.414 |
| Borda foliar | 0.660 | 0.880 | 0.854 | 0.842 | 0.834 | 0.935 | 0.730 |
| Forma da folha | 0.944 | 1.007 | 1.015 | 1.024 | 1.133 | 1.090 | 0.941 |
| Distribuição de espinhos nas folhas | 0.676 | 0.867 | 0.906 | 0.901 | 0.855 | 0.898 | 0.792 |
| Localização de espinhos em brácteas | 0.638 | 0.830 | 0.840 | 0.844 | 0.790 | 0.816 | 0.698 |
| Salinidade | 1.993 | 1.977 | 1.889 | 1.973 | 2.009 | 1.948 | 1.789 |
| Forma da semente | 0.085 | 0.305 | 0.155 | 0.354 | 0.246 | 0.142 | 0.114 |
| Dormência de semente | 0.209 | 0.283 | 0.237 | 0.242 | 0.275 | 0.291 | 0.180 |
| Tipo de tegumento | 0.490 | 0.675 | 0.452 | 0.571 | 0.577 | 0.491 | 0.469 |
| Comp. de espinhos nas brácteas | 1.285 | 1.357 | 1.352 | 1.377 | 1.361 | 1.355 | 1.330 |
| Quantidade de espinhos nas brácteas | 1.238 | 1.465 | 1.462 | 1.494 | 1.507 | 1.443 | 1.389 |
| Intensidade de espinhos nas folhas | 1.190 | 1.196 | 1.230 | 1.201 | 1.231 | 1.250 | 1.219 |

**Tabela 7**. Valores dos testes de qui-quadrado comparando a distribuição dos caracteres de cada coleção nuclear com a coleção base

| Caracteres | CC1 | | CC2 | | CC3 | | CC4 | | CC5 | | Aleatória | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ângulo do ramos | 83.56 | ** | 55.31 | ** | 86.44 | ** | 87.88 | ** | 54.90 | ** | 40.25 | ** |
| Ramificação | 27.18 | ** | 17.26 | ** | 26.61 | ** | 26.63 | ** | 16.92 | ** | 12.96 | ** |
| Cor da corola no florescimento | 35.77 | ** | 26.79 | ** | 37.15 | ** | 38.22 | ** | 32.14 | ** | 25.82 | ** |
| Forma do capítulo | 3.83 | ns | 6.25 | * | 3.36 | ns | 2.31 | ns | 6.25 | * | 3.62 | ns |
| Tamanho de capítulo | 14.58 | ns | 14.50 | ns | 17.32 | ns | 10.60 | ns | 27.39 | ** | 8.00 | ns |
| Capítulos | 53.61 | ** | 36.83 | ** | 52.40 | ** | 51.10 | ** | 37.43 | ** | 25.20 | ** |
| Tegumento na semente | 43.20 | ** | 30.10 | ** | 44.11 | ** | 45.04 | ** | 30.83 | ** | 21.03 | * |
| Borda foliar | 6.36 | ns | 8.78 | * | 5.28 | ns | 4.84 | ns | 17.53 | ** | 4.65 | ns |
| Forma da folha | 1.10 | ns | 1.98 | ns | 1.57 | ns | 3.97 | ns | 5.07 | ns | 7.16 | ns |
| Ácido linoleico | 106.99 | ** | 69.66 | ** | 105.28 | ** | 103.81 | ** | 67.45 | ** | 56.72 | ** |
| Maturidade | 48.34 | ** | 47.24 | ** | 47.16 | ** | 47.92 | ** | 48.66 | ** | 24.98 | ** |
| Distribuição de espinhos nas folhas | 27.66 | ** | 28.24 | ** | 31.25 | ** | 28.01 | ** | 29.89 | ** | 14.53 | ** |
| Localização de espinhos em brácteas | 9.97 | * | 9.11 | ns | 14.12 | ** | 4.25 | ns | 7.23 | ns | 6.05 | ns |
| Teor de óleo | 72.28 | ** | 66.19 | ** | 44.96 | ** | 51.47 | ** | 47.46 | ** | 37.38 | ** |
| Ácido oleico | 78.36 | ** | 52.49 | ** | 73.37 | ** | 67.01 | ** | 43.12 | ** | 49.26 | ** |
| Ácido palmítico | 61.00 | ** | 34.96 | ** | 54.64 | ** | 49.78 | ** | 38.80 | ** | 20.17 | * |
| Altura planta | 29.59 | ** | 33.48 | ** | 29.19 | ** | 27.23 | ** | 26.56 | ** | 14.69 | ns |
| Teor de proteína | 68.91 | ** | 53.88 | ** | 64.54 | ** | 67.96 | ** | 50.75 | ** | 33.86 | ** |
| Salinidade | 3.10 | ns | 5.43 | ns | 9.15 | ns | 8.22 | ns | 2.71 | ns | 12.15 | ns |
| Forma da semente | 16.30 | ** | 0.80 | ns | 23.47 | ** | 7.10 | * | 0.71 | ns | 0.23 | ns |
| Peso de 100 sementes | 100.27 | ** | 91.04 | ** | 97.02 | ** | 100.55 | ** | 72.61 | ** | 48.34 | ** |
| Dormência de semente | 0.54 | ns | 0.19 | ns | 0.02 | ns | 0.60 | ns | 1.92 | ns | 0.38 | ns |
| Tipo de tegumento | 31.54 | ** | 25.37 | ** | 15.31 | ** | 15.55 | ** | 22.44 | ** | 16.88 | ** |
| Comp. de espinhos em brácteas | 6.28 | ns | 10.48 | * | 11.45 | ** | 3.70 | ns | 11.02 | * | 4.75 | ns |
| Quantidade de espinhos nas brácteas | 66.50 | ** | 49.87 | ** | 68.40 | ** | 66.89 | ** | 48.23 | ** | 33.39 | ** |
| Intensidade de espinhos nas folhas | 8.71 | * | 7.23 | ns | 9.27 | * | 6.87 | ns | 7.12 | ns | 4.79 | ns |
| Ácido esteárico | 58.99 | ** | 86.00 | ** | 98.99 | ** | 95.11 | ** | 46.36 | ** | 30.80 | ** |
| Produtividade | 18.28 | ns | 16.59 | ns | 15.76 | ns | 24.01 | * | 12.49 | ns | 16.48 | ns |

ns – não significativo ao nível P>0.05; * significativo a P<0.05 pelo teste de qui-quadrado; ** significativo a P<0.01 pelo teste de qui-quadrado.

(A)




(B)

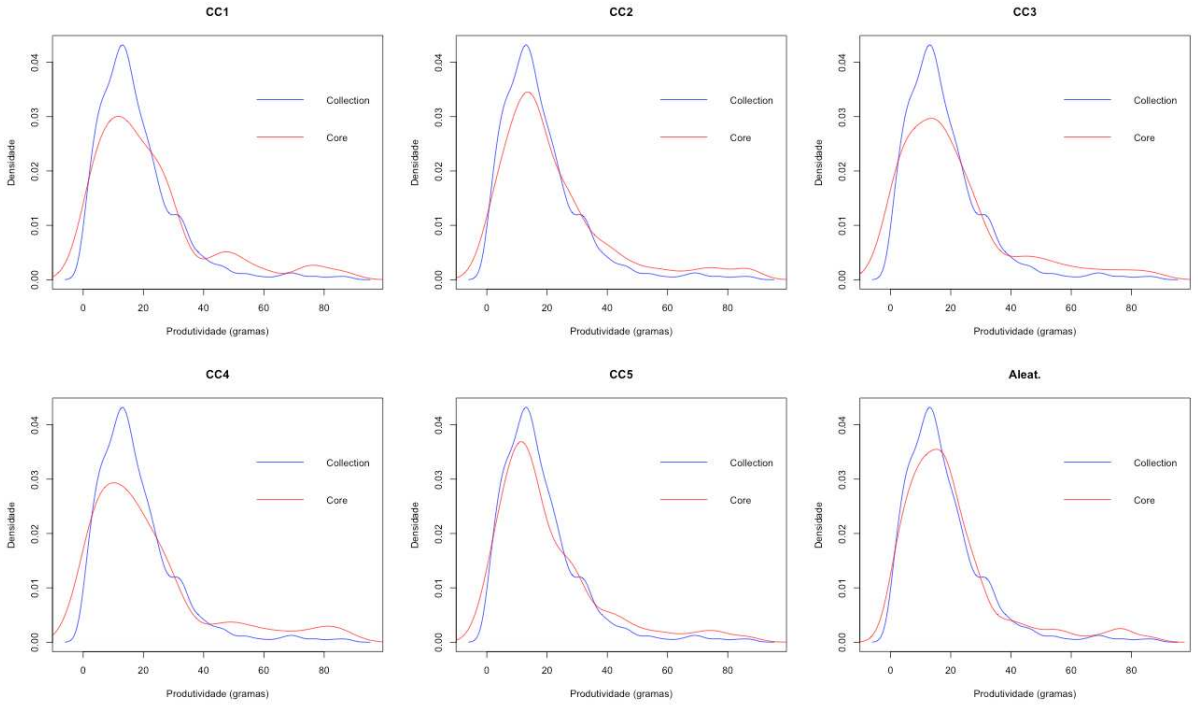



**Figura 1.** Distribuição das frequências de teor óleo (A) e produtividade de planta (B) dos acessos de cártamo na coleção completa (Collection) e nas seis coleções nucleares (Core).

# 4. DISCUSSÃO

Para o estabelecimento de coleções nucleares de cártamo representativas da variabilidade genética existente no banco de germoplasma do USDA, o presente estudo propôs a utilização conjunta da estratégia de maximização e da estratificação dos acessos em grupos genéticos. Os resultados apresentados mostraram que a estratificação foi adequada para selecionar acessos distintos e representativos da coleção base. De fato, a estratificação bem sucedida dos acessos em grupos geneticamente expressivos é essencial para que a estratégia de amostragem seja efetiva (VAN HINTUM et al., 2000).

A ampla distribuição geográfica e a presença de "centros de similaridade" de cártamo requerem uma amostragem abrangente do germoplasma. A estratificação dos acessos de cártamo em grupos garante que, no processo de amostragem, a maioria dos componentes (alelos) de adaptação dos genótipos a "habitats" ou regiões ecogeográficas estejam devidamente representados. De fato, as coleções CC3 e CC4 conseguiram representar a maioria das classes dos caracteres (CR = 94.97% e 94.95%) presentes na coleção base e mantiveram altos níveis de variabilidade. O elevado potencial dessa abordagem também foi constatado em outras espécies, como quinoa (ORTIZ et al., 1998), milho (MALOSETTI e ABADIE, 2001; LI et al., 2004), soja (OLIVEIRA et al., 2010) e batata (GOPAL et al., 2013).

A performance similar das coleções CC3 e CC4 mostra que a estratificação dos acessos nos grupos genéticos definidos por Chapman et al. (2010) e Johnson et al. (2007), ainda que contrastante, mantém a representatividade da coleção base. Atualmente não há consenso quanto aos verdadeiros centros de similaridade de cártamo pois esses foram definidos a partir de diferentes hipóteses, métodos e genótipos. Os centros fundamentados em caracteres fenotípicos foram baseados em poucos caracteres, como altura de planta, produtividade, ramificação e peso de sementes. Por outro lado, na definição dos centros de similaridade por marcadores moleculares foram considerados diferentes marcadores (dominantes e codominantes), genótipos distintos e estruturas populacionais contrastantes. Além disso, muitas vezes acessos classificados em um centro são mais similares a acessos classificados em outro centro devido à redundância na classificação dos centros de similaridade (KNOWLES, 1969; ASHRI, 1975; JOHNSON et al., 2007; CHAPMAN et al., 2010).

A presença de redundância foi observada no presente estudo quando os acessos foram estratificados nos dois grupos genéticos (Coleção CC5). A coleção CC5 apresentou mais

acessos do que as coleções CC3 e CC4, porém não houveram incrementos na variabilidade (VD,%) e na representatividade das classes dos caracteres estudados (VR,%). A ausência de magnitudes mais elevadas nos índices de Shannon dos caracteres qualitativos da coleção CC5 também indicaram a presença de redundância (Tabela 5). De fato, a proximidade geográfica e a troca de materiais genéticos ("landraces") entre agricultores são as principais causas da redundância observada em alguns centros de similaridade de cártamo (CHAPMAN et al., 2010).

Em outras espécies, como linho e soja, redundâncias também ocorrem porque os grupos morfológicos de acessos nem sempre correspondem a grupos genéticos e poucos grupos genéticos correspondem apropriadamente às origens geográficas (FU, 2005; LI et al., 2008). Embora essas discordâncias também sejam, em certa medida, observadas em cártamo, a efetividade da estratificação em grupos genéticos não deve ser desconsiderada. Escribano et al. (2007) relataram que a falta de grupos genéticos bem definidos limita as estratégias de amostragem em representar a diversidade total da coleção base. Os autores observaram que nessa situação alelos raros poderiam ser perdidos durante o processo de amostragem, mesmo com a utilização de métodos baseados em dendogramas (estratégia logaritma e agrupamento em "stepwise").

A utilização da estratégia de maximização tem sido a abordagem mais apropriada para amostrar todas as classes da coleção base, incluindo as classes raras, e para conservar a frequência original dessas classes. Em todas as coleções nucleares de cártamo estabelecidas por essa estratégia as taxas de coincidência (CR,%) superiores a 94% mostraram uma extensa preservação das classes presentes na coleção base. A diversidade da coleção base também foi conservada, conforme indicado pelas magnitudes das estimativas das estatísticas VD(%) e VR(%). As magnitudes de todas estatísticas de validação da coleção nuclear estabelecida somente pela estratégia de maximização (CC1) reforçam a efetividade dessa estratégia ao se mostrar tão eficiente quanto a estratificação em grupos genéticos. De fato, a estratégia M fornece uma combinação específica de acessos que maximiza a riqueza alélica ou de classes, principalmente em espécies autógamas e com fluxo gênico restrito (BATAILLON et al., 1996).

A redução das coleções nucleares a menos de 10% da coleção original também mostrou o potencial da estratégia de maximização em captar a riqueza de classes com uma quantidade menor de acessos. As coleções nucleares CC1, CC3 e CC4 incluíram acessos representando aproximadamente 3.75% dos acessos conservados na coleção base. A extensão

dessas coleções pode tornar sua utilização mais acessível em programas de melhoramento. De fato, diferentes configurações e tamanhos de coleções nucleares, de 0.3% a 50% da coleção base, são utilizadas com diferentes propósitos em atividades de conservação e melhoramento (VAN HINTUM et al., 2000).

O grande potencial da associação entre a estratégia de maximização e a estratificação dos acessos está representado na performance superior das estatísticas de validação das coleções CC3 e CC4 em relação às outras coleções nucleares estabelecidas. Coleções nucleares que fornecem uma boa representatividade da coleção base apresentam pequena diferença média (MD,%) e maiores CR% (> 80%) e VR% (> 100%) (FRANKEL e BROWN, 1984; HU et al., 2000). No entanto, as coleções nucleares não objetivam somente a consistência estatística mas buscam representar a diversidade genética da coleção base.

De modo geral, estabelecer coleções nucleares pelas estratégias de maximização e estratificação em grupos genéticos é uma abordagem conveniente para aumentar a acessibilidade e a utilidade das coleções de germoplasma de cártamo, não apenas por causa da vantagem de amostrar todos os alelos ou classes da coleção, mas também pela oportunidade de manter a maior variabilidade possível com um número mínimo de acessos.

# 5. CONCLUSÃO

As coleções nucleares de cártamo estabelecidas representaram de forma eficiente a diversidade genética da coleção base.

As coleções nucleares estratificadas por grupos genéticos apresentaram aproximadamente 60 genótipos, com diferença média de apenas 7% em relação a coleção base e com taxas variável e de coincidência superiores a 150% e 94%, respectivamente.

O uso conjunto da estratégia de maximização e da estratificação dos genótipos em grupos genéticos maximizou a captação da variabilidade genética e introduziu maior eficiência no estabelecimento das coleções nucleares ao selecionar uma menor quantidade de acessos. As coleções nucleares estabelecidas incluíram aproximadamente 3.75% dos acessos conservados na coleção base.

# 6. REFERÊNCIAS


ASHRI, A. Evaluation of germplasm collection of safflower, Carthamus tinctorius L. V. Distribution and regional divergence for morphological characters. **Euphytica**, v. 24, p. 651-659. 1975.

BATAILLON, T. M.; DAVID, J. L.; SCHOEN, D. J. Neutral genetic markers and conservation genetics: a simulated germplasm collections. **Genetics**, v. 144, p. 409-417. 1996.

BROWN, A. H. D. The case for core collections. In: BROWN, A. H. D.; FRANKEL, O. H.; MARSHALL, D. R.; WILLIAMS, J. T. (Eds.), **The Use of Plant Genetic Resources**. University Press Cambridge: Cambridge, pp. 136–156. 1989.

CHAPMAN, M. A.; HVALA, J.; STREVER, J.; BURKE, J. M. Population genetic analysis of safflower (Carthamus tinctorius; Asteraceae) reveals a near eastern origin and five centers of diversity. **American Journal of Botany**, v. 97, n. 5, p. 831-840. 2010.

DAJUE, L.; ZHOU, M.; RAO, V. R. **Characterization and Evaluation of Safflower Germplasm**. Geological Publishing House, 1993. 260 p.

DWIVEDI, S. L.; UPADHYAYA, H. D.; HEDGE, D. M. Development of core collection using geographic information and morphological descriptors in safflower (Carthamus tinctorius L.) germplasm. **Genetic Resources and Crop Evolution**, v. 52, p. 821–830. 2005.

ESCRIBANO P.; VIRUEL M. A.; HORMAZA J. I. Molecular analysis of genetic diversity and geographic origin within an ex situ germplasm collection of cherimoya by using SSRs. **Journal of the American Society for Horticultural Science**, v. 132, p. 357–367. 2007.

ESCRIBANO, P.; VIRUEL, M. A.; HORMAZA, J. I. Comparison of different methods to construct a core germplasm collection in woody perennial species with simple sequence repeat markers. A case study in cherimoya (Annona cherimola, Annonaceae), an underutilised subtropical fruit tree species. **Annals Applied Biology**, v. 153, p. 25–32. 2008

FRANKEL, O. H. Genetic perspectives of germplasm conservation. In: ARBER, W.; LLIMENSEE, K.; PEACOCK, W. J.; STARLINGER, P. (Eds.), **Genetic Manipulation: Impact on Man and Society**. Cambridge University Press, Cambridge, pp. 161-170. 1984

FRANKEL, O. H.; BROWN, A. H. D. Plant genetic resources today: a critical appraisal. In: HOLDE, J. H. W.; WILLIAMS, J. T. (Eds.), **Crop Genetic Resources: Conservation and Evaluation**. Allen and Unwin: Winchester, pp. 249-257. 1984.

FU, Y. B. Geographic patterns of RAPD variation in cultivated flax. **Crop Science**, v. 45, p. 1084-1091. 2005.

GOPAL, J.; KUMAR, V.; KUMAR, R.; MATHUR, P. Comparison of different approaches to establish a core collection of andigena (Solanum tuberosum group andigena) potatoes. **Potato Research**, v. 56, p. 85-98. 2013.

HAMDAN, Y. A. S.; PÉREZ-VICH, B.; FERNÁNDEZ-MARTÍNEZ, J. M.; VELASCO, L. Novel safflower germplasm with increased saturated fatty acid content. **Crop Science**, v. 49, p. 127–132. 2009.

HU, J. J.; ZHU, J.; XU, H. H. M. Methods of constructing core collections by stepwise clustering with three sampling strategies based on genotypic values of crops. **Theoretical and Applied Genetics**, v. 101, p. 190-196. 2000.

JOHNSON, R. C.; BERGMAN, J .W.; FLYNN, C. R. Oil and meal characteristics of core and non-core safflower accessions from the USDA collection. **Genetic Resources and Crop Evolution**, v. 46, p. 611-618. 1999.

JOHNSON, R. C.; KISHA, T. J.; EVANS, M. A. Characterizing safflower germplasm with AFLP molecular markers. **Crop Science**, v. 47, p. 1728-1736. 2007.

KIM, K. W.; CHUNG, H. K.; CHO, G. T.; MA, K. H.; CHANDRABALAN, D.; GWAG, J. G.; KIM, T. S.; CHO, E. G.; PARK, Y. J. PowerCore: a program applying the advanced M

strategy with a heuristic search for establishing core sets. **Bioinformatics**, v. 23, n. 16, p. 2155-2162. 2007.

KNOWLES, P. F. Centers of plant diversity and conservation of crop germplasm – Safflower. **Economic Botany**, v. 23, p. 324-329. 1969.

LI, Y.; SHI, Y.; CAO, Y.; WANG, T. Establishment of a core collection for maize germplasm preserved in Chinese National Genebank using geographic distribution and characterization data. **Genetic Resources and Crop Evolution**, v. 51, p. 845-852. 2004.

LI, Y. H.; GUAN, R. X.; LIU, Z. X.; MA, Y. S.; WANG, L. X.; LI, L. H.; LIN, F. Y.; LUAN, W.; CHEN, P.; YAN, Z.; GUAN, Y.; ZHU, L.; NING, X.; SMULDERS, M. J.; LI, W.; PIAO, R.; CUI, Y.; YU, Z.; GUAN, M.; CHANG, R.; HOU, A.; SHI, A.; ZHANG, B.; ZHU, S.; QIU, L. Genetic structure and diversity of cultivated soybean (Glycine max (L.) Merr.) landraces in China. **Theoretical and Applied Genetics**, v. 117, p. 857-871. 2008.

MALOSETTI, M.; ABADIE, T. Sampling strategy to develop a core collection of Uruguayan maize landraces based on morphological traits. **Genetic Resources and Crop Evolution**, v.48, p. 381-390. 2001.

MÜNDEL, H.; BERGMAN, J. W. Safflower. In: VOLLMANN J.; RAJCAN, I. (Eds.) **Handbook of plant breeding: oil crops**. Springer: New York, pp. 423–447. 2009.

OLIVEIRA, M. F.; NELSON, R. L.; GERALDI, I. O.; CRUZ, C. D.; TOLEDO, J. F. F. Establishing a soybean germplasm core collection. **Field Crops Research**, v.119, p. 277-289. 2010.

ORTIZ, R.; RUIZ-TAPIA, E. N.; MUJICA-SANCHEZ, A. Sampling strategy for a core collection of Peruvian quinoa germplasm. **Theoretical and Applied Genetics**, v. 96, p. 475-483. 1998.

R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing. Disponível em: <http://www.R-project.org>. Acesso em: 20 out. 2013.

SHANNON, C. E. A mathematical theory of communication. **The Bell System Technical Journal**, v. 27, p. 379-423. 1948.

SHOEN, D. J.; BROWN, H. D. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. **Proceedings of the National Academy of Sciences of the United States**, v. 22, p. 10623-10627. 1993.

SNEDECOR, G. W.; COCHRAN, W. G. **Statistical methods**. Iowa State University Press: Ames. 1980. 503p.

STURGES, H. The choice of a class-interval. **Journal of the American Statistical Association**, v. 21, p. 65-66. 1926.

USDA, ARS, National Genetic Resources Program. Germplasm Resources Information Network - (GRIN). [Online Database] National Germplasm Resources Laboratory, Beltsville, Maryland. Disponível em: <http://www.ars-grin.gov/cgi-bin/npgs/html/crop.pl?108>. Acesso em: 03 out. 2013.

VAN HINTUM, T. J. L.; BROWN, A. H. D.; SPILLANE, C.; HODGKIN, T. **Core Collections of Plant Genetic Resources**. IPGRI Technical Bulletin No. 3. International Plant Genetic Resources Institute: Rome. 2000. 51 p.

VELASCO, L.; FERNÁNDEZ-MARYINEZ, J. M. Registration of CR-34 and CR-81 safflower germplasms with increased tocopherol. **Crop Science**, v.44, p. 2278. 2004.

ZHANG, Z., JOHNSON, R. C. **Safflower germplasm collection directory**. IPGRI Office for East Asia: Beijing. 1999. 18p.

# CAPÍTULO 2. ASSOCIAÇÃO PREDITIVA ENTRE TEOR DE ÓLEO E VARIÁVEIS ECOGEOGRÁFICAS DE ORIGEM DE GENÓTIPOS DE CÁRTAMO

## RESUMO

O melhoramento de cártamo com o propósito de aumentar o teor de óleo requer a busca constante de genótipos portadores de alelos que condicionam ao alto teor de óleo. Essas fontes de alelos estão ao alcance dos melhoristas em extensos bancos de germoplasma, porém o grande número de acessos nas coleções frequentemente limita seu uso imediato. A estratégia de identificação focada de germoplasma (FIGS) é um método eficiente de otimizar a seleção de acessos úteis presentes em bancos de germoplasma. A FIGS faz uso da associação preditiva entre características e variáveis ambientais na busca de genótipos com maior probabilidade de conter a característica de interesse. O presente estudo teve como objetivo investigar a existência de associação preditiva entre teor de óleo e variáveis ecogeográficas da origem dos genótipos de cártamo, utilizando a FIGS baseada em abordagens de aprendizado de máquina. Florestas aleatórias, máquinas de vetor de suporte e redes neurais artificias foram utilizadas para modelar a associação entre teor de óleo de 100 genótipos cártamo e 56 variáveis ecogeográficas. As acurácias dos modelos utilizados mostraram que a distribuição de genótipos de cártamo com alto teor de óleo não é aleatória mas ligada a fatores ambientais, mesmo com certo grau de sobreposição entre os teores de óleo em alguns ambientes. Os resultados finais sugerem que explorar a associação preditiva entre o teor de óleo e as características ecogeográficas do local de origem do germoplasma aumenta as chances de encontrar genótipos com alto teor óleo.

**Palavras-chave:** Estratégia de identificação focada de germoplasma (FIGS), Aprendizado de máquina, Cártamo, Óleo, Germoplasma.

# PREDICTIVE ASSOCIATION BETWEEN OIL CONTENT AND ECOGEOGRAPHIC PARAMETERS OF THE ORIGINAL COLLECTION SITE OF SAFFLOWER GENOTYPES


## ABSTRACT

Safflower breeding aiming to increase oil content requires endless search for genotypes carrying alleles that determine high oil content. These allele sources are available to safflower breeders in large germplasm banks, but the enormous number of accessions in collections often limits its immediate use. The Focused Identification Germplasm Strategy (FIGS) is an efficient method to optimize the selection of useful accessions kept in genebanks. The FIGS makes use of predictive association between characteristics and environmental variables in the search for genotypes with high probability of containing the trait of interest. The present study aimed to investigate the existence of predictive association between oil content and ecogeographic parameters of the original site of safflower genotypes using the FIGS based on machine learning approaches. Random forests, support vector machines and artificial neural networks were used to model the association between oil content of 100 safflower genotypes and 56 ecogeographic parameters. The models accuracies indicated that the distribution of safflower genotypes with high oil content is not random but associated to environmental factors, even with some degree of overlap between the oil content in some environments. The final results suggest that exploring the predictive association between oil content and ecogeographic parameters of original collection site of safflower germplasm increases the chances of finding genotypes with high oil content.



**Keywords:** Focused Identification Germplasm Strategy (FIGS), Machine Learning , Safflower Oil , Germplasm.

# 1. INTRODUÇÃO

O cártamo (Carthamus tinctorius) é uma espécie da família Compositae cuja sementes são utilizadas como fonte de óleo na alimentação humana e de pássaros, indústria farmacêutica e na produção de biocombustível. A produção mundial de óleo de cártamo é pequena em relação a outras espécies oleaginosas anuais. Em 2012 a produção mundial de óleo de cártamo representou menos de 1% da produção mundial de óleo vegetais (FAOSTAT, 2013). No entanto, a demanda crescente por espécies alternativas para produção de biocombustíveis tem despertado o interesse pelo cártamo. O potencial produtivo, a tolerância a estresses abióticos e o alto teor de óleo são características notáveis do cártamo que fazem com que seja considerado uma cultura de grande potencial.

A variação do teor de óleo de genótipos de cártamo entre 12 e 45% oferece a possibilidade de aumento da produtividade de óleo até a 1200 kg/ha, quantidade maior do que a produtividade de outras espécies oleaginosas anuais. De fato, os avanços obtidos pelos programas de melhoramento nos últimos 30 anos mostram que o teor de óleo dos genótipos pode ser efetivamente incrementado (DAJUE e MÜNDEL, 1996). Variedades desenvolvidas nos Estados Unidos alcançaram grande melhoria no teor de óleo, contendo até 45% (variedade Oker). Da mesma forma, na Índia, os híbridos de cártamo mostraram aumento de 20 a 25% na produtividade de óleo (BERGMAN et al., 1985; SINGH et al., 2003).

Entretanto, há necessidade constante de variedades mais produtivas e com alto teor de óleo para substituir e melhorar as variedades presentes no mercado. A utilização de genótipos conservados em bancos de germoplasma pode certamente fornecer novos alelos para incrementar o teor de óleo (PRADA, 2009). Essa afirmação também é sustentada por Chapman et al. (2010) que sugerem explorar genótipos provenientes do grupo genético do Oriente Próximo como uma forma de melhorar as variedades americanas de cártamo. Esse grupo abriga uma diversidade genética única que deve ser utilizada pois contém alguns dos genótipos com os maiores teores de óleo já registrados em bancos de germoplasma (ASHRI et al., 1977). Os locais de coleta do germoplasma também são providos de uma variação de teores de óleo passível de uso (CLAASEN e KIESSELBACH, 1945).

A existência de certo indício de "seleção geográfica" desperta a atenção para uma possível associação dessa característica com variáveis ecogeográficas. De fato, essa hipótese foi estudada em genótipos cevada (Hordeum vulgare) para verificar associação preditiva entre tolerância à salinidade e padrões de precipitação (PEETERS et al. 1990). Hijmans et al.

(2003) também exploraram a existência de associação entre tolerância ao congelamento e temperatura do local de coleta de acessos de batata (Solanum spp.). Em ambos os estudos os autores utilizaram técnicas de predição baseadas em modelos lineares e reconheceram que havia associação entre as características estudadas e as variáveis ecogeográficas, porém complexa e de difícil modelagem. No entanto, recentemente a estratégia de identificação focada de germoplasma (FIGS) desenvolvida por Mackay e Street (2004) tem se mostrado uma técnica eficaz na descoberta de associações entre determinadas características e variáveis ecogeográficas.

A FIGS busca explorar a associação entre o ambiente e as propriedades adaptativas de caracteres para auxiliar na "mineração" de características úteis em bancos de germoplasma. Diversos estudos reconheceram o potencial dessa técnica na rápida identificação de genótipos com as seguintes características: tolerância a seca e calor (KHAZAEI et al., 2013), resistência a insetos pragas (EL-BOUHSSINI et al., 2010) e resistência a doenças (ENDRESEN et al. 2011; ENDRESEN et al., 2012; BARI et al., 2012). Basicamente a FIGS é fundamentada em técnicas de modelagem para associar determinadas características à variáveis ambientais, baseada na premissa de que o ambiente influencia fortemente na seleção natural e consequentemente na distribuição dos genótipos (ICARDA, 2013). Os modelos construídos na FIGS são utilizados para identificar genótipos com características desejáveis originários de ambientes que com pressões seletivas similares.

Métodos baseados em aprendizado de máquina tem sido utilizados com frequência na FIGS devido a sua capacidade de modelar padrões complexos e não lineares de interação. Técnicas como florestas aleatórias, máquinas de vetor de suporte e redes neurais artificiais são empregados há muito tempo em informática ecológica, principalmente em estudos de distribuição de populações, dinâmica de dispersão e modelagem de habitats (OLDEN et al., 2008). No contexto da FIGS, essas técnicas foram aplicadas em trigo (Triticum aestivum) para verificar a associação da resistência à ferrugem do colmo com variáveis ambientais. Os modelos obtidos apresentaram acurácias elevadas que permitiram identificar genótipos resistentes confinados em certos ambientes (BARI et al., 2012).

Conhecido o potencial da FIGS, o presente estudo teve como objetivo determinar a existência associação entre teor de óleo e variáveis ecogeográficas representativas do ambiente de origem dos acessos de cártamo. A hipótese testada foi que genótipos com alto teor óleo são provenientes de locais que possuem características ambientais similares. Para

comprovar esta hipótese foi empregada a estratégia de identificação focada de germoplasma (FIGS) utilizando técnicas de modelagem baseadas em aprendizado de máquina.

## 2. MATERIAIS E MÉTODOS

Os dados utilizados neste estudo foram as médias de teor de óleo de genótipos de cártamo e os dados ecogeográficos do local de coleta desses genótipos. O aplicativo computacional R foi utilizado no preparo e análise dos dados (R DEVELOPMENT CORE TEAM, 2013).

**Dados Fenotípicos**

Os dados de teor de óleo utilizados neste estudo são provenientes da Rede de Informação de Recursos Genéticos (GRIN) do Departamento de Agricultura dos Estados Unidos (USDA). O teor de óleo dos genótipos foi determinado em estudos de caracterização de germoplasma conduzidos entre 1988 e 2010 (JOHNSON et al., 1999; DAJUE et al., 1993; USDA, 2013).

Nos estudos os genótipos de cártamo foram avaliados em experimentos a campo no Centro de Pesquisa Agrícola Oriental da Universidade do Estado de Montana em Sidney - Montana (47º43’34’’N, 104º09’W), na Estação Regional de Introdução de Plantas em Pullman - Washington (46º43’N, 117º10’W) e no Jardim Botânico de Beijing - Academia Chinesa de Ciências (39º33’N; 116º16’E ). Os genótipos foram plantados em fileiras únicas de 6 metros de comprimento com espaçamento de 60 centímetros. No momento da colheita, amostras de 16 gramas de sementes foram coletadas de cada genótipo, secas à 60ºC for 4h em estufa e submetidas ao processo de extração de óleo via Soxhlet. O óleo extraído foi pesado e o teor determinado.

Cem genótipos provenientes de locais de coleta com coordenadas geográficas registradas foram utilizados no estudo. Esses genótipos representam “landraces” (cultivares tradicionais) de cártamo e compõe a coleção de germoplasma do USDA. A distribuição geográfica dos genótipos envolve diversas regiões da Europa, Ásia e Norte da África (Figura 1).

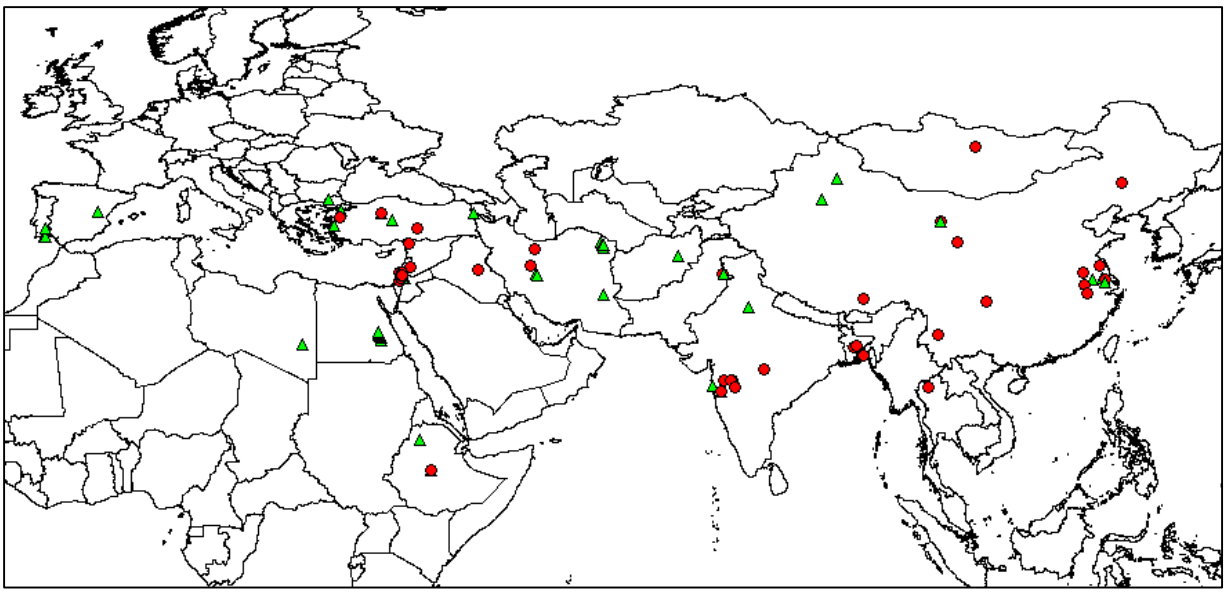


**Figura 1**. Distribuição geográfica dos genótipos de cártamo com alto teor (triângulos verdes) e baixo teor de óleo (círculos vermelhos).

**Dados Ecogeográficos**

Para o presente estudo os dados ecogeográficos foram extraídos do local de coleta de cada um dos 100 genótipos usando as coordenadas de latitude e longitude registrados. Esses dados foram obtidos do banco de dados "WorldClim" com o auxílio do programa DIVA GIS (HIJMANS et al., 2005).

Cada local foi representado pela altitude, precipitação média mensal, temperaturas máxima, média e mínima mensal, além de 19 variáveis bioclimáticas (BUSBY, 1991). Essas últimas variáveis são derivadas dos dados de temperatura e precipitação mensais com o intuito de gerar variáveis biologicamente significativas (Tabela 1). Usadas rotineiramente em modelagem de nichos ecológicos, representam informações de tendências anuais, sazonalidade e fatores ambientais extremos. Todas as variáveis foram utilizadas na resolução espacial de 30 arc segundos (aproximadamente 1 $km^2$ de resolução), resolução mais refinada para captar variabilidade ambiental. No total foram obtidas 56 variáveis ecogeográficas.

As variáveis ecogeográficas são fontes de informação confiáveis pois o conjunto de dados utilizado se refere a um compilado de médias mensais de variáveis climáticas medidas em estações meteorológicas de um grande número de fontes globais, regionais, nacionais e locais, principalmente no período de 1950-2000.

**Tabela 1**. Variáveis ecogeográficas utilizadas no estudo

| Código | Descrição |
|---|---|
| prec | Precipitação média mensal (Janeiro a Dezembro) |
| tmax | Temperatura máxima mensal (Janeiro a Dezembro) |
| tmin | Temperatura mínima mensal (Janeiro a Dezembro) |
| alt | Altitude |
| bio1 | Temperatura média anual |
| bio2 | Intervalo médio diurno (média mensal(Temp. máx. - Temp. mín.)) |
| bio3 | Isotermalidade (BIO2/BIO7)*(100) |
| bio4 | Sazonalidade de temperatura (desvio padrão * 100) |
| bio5 | Temperatura máxima no mês mais quente |
| bio6 | Temperatura mínima no mês mais frio |
| bio7 | Intervalo de temperatura anual (BIO5 - BIO6) |
| bio8 | Temperatura média do trimestre mais úmido |
| bio9 | Temperatura média do trimestre mais seco |
| bio10 | Temperatura média do trimestre mais quente |
| bio11 | Temperatura média do trimestre mais frio |
| bio12 | Precipitação anual |
| bio13 | Precipitação no mês mais úmido |
| bio14 | Precipitação no mês mais seco |
| bio15 | Sazonalidade de precipitação (coeficiente de variação) |
| bio16 | Precipitação no trimestre mais úmido |
| bio17 | Precipitação no trimestre mais seco |
| bio18 | Precipitação no trimestre mais quente |
| bio19 | Precipitação no trimestre mais frio |

**Processamento e transformação dos dados**

Os dados de teor óleo dos 100 genótipos foram transformados em variáveis binárias para classificação dos genótipos. Genótipos com teor de óleo menor que 30% foram classificados como genótipos de baixo teor de óleo (receberam a codificação binária 0) e os genótipos com teor maior que 30% classificados com genótipos de alto teor (codificação binária 1). O critério de escolha do limiar de classificação foi o teor médio de óleo dos genótipos de cártamo depositados nos bancos de germoplasma ao redor do mundo (DAJUE e MÜNDEL, 1996).

Os dados das variáveis ambientais foram centralizados e padronizados antes da realização das análises. A centralização e padronização de variáveis consiste em subtrair das observações individuais a média de cada variável e dividir pelo seu desvio padrão. Como

resultado, as variáveis transformadas terão média zero e desvio padrão igual a um. A utilização dessas técnicas de transformação melhora a estabilidade numérica dos cálculos em modelos de aprendizado de máquina, elimina o viés introduzido pelas diferenças nas escalas da variáveis usadas e iguala a variância de cada variável (HASTIE et al., 2001).

Antes de realizar as análises os dados foram separados aleatoriamente em dois conjuntos: dados de treinamento e dados de teste. Aproximadamente 2/3 dos dados foram utilizados para o treinamento dos modelos e 1/3 para validação. A divisão dos dados foi realizada de forma a manter a mesma proporção de classes de teor de óleos nos dois conjuntos.

**Abordagens**

Para verificar se há a associação entre o alto teor de óleo e variáveis ecogeográficas foram utilizados três técnicas de modelagem baseadas em aprendizado de máquina: Máquinas de Vetor de Suporte (SVM) e Redes Neurais Artificiais (NN) e Florestas Aleatórias (RF). Essas técnicas são capazes de identificar estrutura em dados complexos, não lineares, e gerar modelos preditivos acurados sem ter que satisfazer as restritivas suposições requeridas por abordagens paramétricas convencionais. Todas as técnicas tem sido utilizados com eficiência na resolução de problemas complexos em ecologia e melhoramento de plantas (OLDEN et al., 2008; ENDRESEN et al., 2011).

**Máquinas de Vetor de Suporte**

Máquinas de vetor de suporte (SVM) é uma abordagem de aprendizado para classificação de dados baseada no mapeamento dos vetores de entrada em um espaço característico de alta dimensionalidade e separação da observações em classes definidas. Nesse espaço é construído uma superfície de decisão constituindo um hiperplano de separação ótima das classes (JAMES et al., 2013). O hiperplano de separação é capaz de generalizar sem problemas de super-parametrização através do controle de suas margens. Para realizar esse controle as máquinas de vetor de suporte utilizam uma relação funcional, conhecida como núcleo, para mapear os dados em um novo hiperespaço no qual padrões mais complexos podem ser simplesmente representados.

O modelos de máquinas de vetor de suporte foram testados com duas funções núcleo: linear e função de base radial. O modelo com função núcleo linear deve ser entendido como um classificador de vetor de suporte de margem máxima para dados linearmente separáveis no espaço de características. Esse modelo fornece um ótimo ponto de partida para modelos mais complexos.

Inicialmente os dados de treinamento consistiram de $n$ pares de observações arranjados na seguinte forma $\{(x_i, y_i)\}_i^n$, onde $x_i$ indica o vetor de entrada (observações ecogeográficas) e $y_i$ valores de saída (classe binária de teor de óleo). A regra de classificação é dada pela função:

$$f(x) = x^T\beta + \beta_0$$

Onde $x$ é o vetor de entrada, $\beta$ é o vetor de pesos ajustáveis e $\beta_0$ é o viés.

Com essa função é possível encontrar o hiperplano com a maior margem entre classes de treinamento. De modo geral, o hiperplano é solução do seguinte problema de otimização:

$$\text{Maximizar } M, \text{ sujeito a } \begin{cases} \sum_{j=1}^{p} \beta_j^2 = 1\,,\ j = 1, \dots, p \\ y_i(x_i^T\beta + \beta_0) \geq M, i = 1, \dots, n \end{cases}$$

Onde $M$ representa a distância à margem do hiperplano e o problema de otimização define $\beta_0$, $\beta_1, \dots, \beta_p$ ($p$ variáveis ecogeográficas) que maximizam $M$.

No entanto, as classes podem se sobrepor no espaço de características. A maneira de lidar com essa sobreposição consiste em maximizar $M$, permitindo que alguns pontos (observações) sejam alocados em lados errados do hiperplano. Esse procedimento consiste na utilização de variáveis de folga $\xi = (\xi_1, \xi_2, \dots, \xi_n)$. Assim, a solução do problema de otimização fica:

$$y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i)$$

$$\text{para } \forall_i\ \xi_i \geq 0,\ \sum_{i=1}^{n} \xi_i \leq C.$$

$C$ é o parâmetro de custo que delimita a soma de $\xi$ e por isso determina o número e a severidade das violações do hiperplano. O problema de otimização é então solucionado por meio de multiplicadores de Lagrange. Porém é conveniente expressar o problema na seguinte forma

$$\text{Minimizar } \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i \text{ , sujeito a } \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Assim a função de Lagrange fica

$$L_p = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(x_i^T\beta + \beta_0) - (1-\xi_i)] - \sum_{i=1}^{n}\mu_i\xi_i$$

onde $\alpha_i \geq 0$ e $\mu_i \geq 0$ são os multiplicadores de Lagrange. Diferenciando a função em relação a $\beta, \beta_0 \; e \; \xi_i$ e substituindo em $L_p$ obtém-se a função objetivo dual Lagrangiana:

$$L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_{i'}y_iy_{i'}x_i^Tx_{i'}$$

Ao maximizar $L_D$ sujeita a $0 \leq \alpha_i \leq C$ e $\sum_{i=1}^{n}\alpha_i y_i = 0$, utilizando técnicas de otimização, é possível encontrar o hiperplano com a margem máxima de separação das classes de observações. A solução desse problema retorna valores de $\alpha_i$ e $\xi_i$ utilizados na solução de $\hat{\beta} = \sum_{i=1}^{n}\widehat{\alpha_i}y_ix_i$.

O único parâmetro livre é o parâmetro de custo $C$. Um valor ótimo pode ser escolhido variando $C$ através de um conjunto de valores pré-definido, monitorando a performance do classificador por meio de validação cruzada.

Em problemas mais complexos com margens não lineares entre classes pode-se aumentar o espaço de características usando função núcleo de base radial. Essa função é aplicada no produto interno dos vetores de entrada da seguinte forma:

$$\begin{aligned} f(x) = &\quad h(x)^T\beta + \beta_0 \\ = &\quad \sum_{i=1}^{n}\alpha_i\, y_i\langle h(x_i), h(x_{i'})\rangle + \beta_0 \end{aligned}$$

Substituindo na função acima, $K(x_i, x_{i'}) = \langle h(x_i), h(x_{i'})\rangle) = \exp\left(-\gamma\sum_{j=1}^{p}(x_{ij} - x_{i'j})^2\right)$.

Nos modelos avaliados os parâmetros de sintonia, custo ($C$) e gama ($\gamma$) (somente para o núcleo radial) foram selecionados por 10 rodadas de validação cruzada repetidas cinquenta

mil vezes. Os melhores valores dos parâmetros foram definidos quando a soma dos erros dos modelos foi minimizada no conjunto de dados de treinamento.

**Redes Neurais Artificiais**

Redes neurais artificiais (NN) são modelos computacionais inspirados no forma como o sistema nervoso biológico processas informações e reconhece padrões complexos. Basicamente NN são formadas por um sistema de processamento composto de um grande número de elementos interconectados chamados neurônios, os quais trabalham em conjunto na resolução de problemas específicos. Existem muitos tipos de modelos NN de aprendizado supervisionado ou não supervisionado. O método supervisionado utilizado no presente estudo foi o perceptron multicamadas, representado na Figura 2.

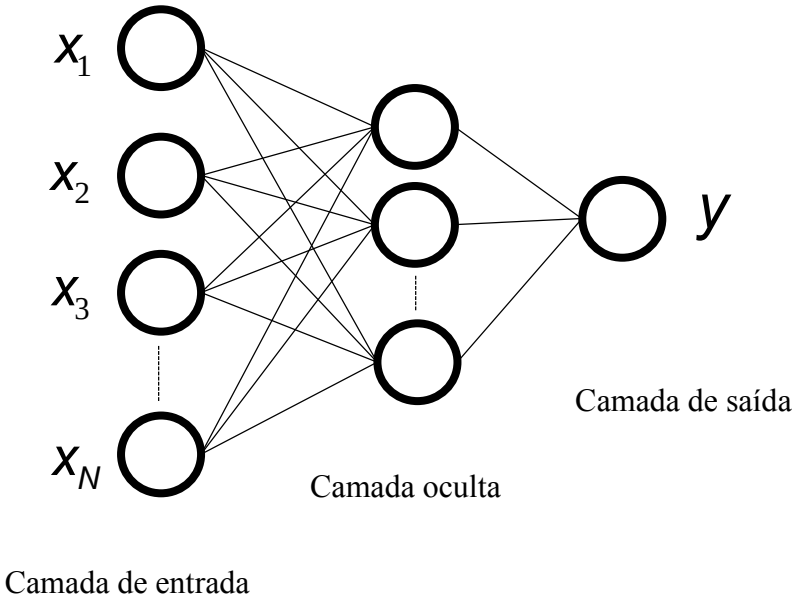



**Figura 2**. Diagrama da rede perceptron multicamadas com uma única camada oculta.

O perceptron multicamadas é composto de um conjunto de neurônios arranjados em camadas, especificamente em camada de entrada, camada oculta e camada de saída. Os dados de entrada (variáveis ecogeográficas) são introduzidas na rede por meio dos neurônios da camada de entrada, um para cada variável. Essa informação é retroalimentada através da rede, primeiramente para os neurônios da camada oculta e depois para os neurônios da camada de saída (BISHOP, 2006).

No processo de treinamento da rede neural, os pesos associados às ligações entre os neurônios da camada de entrada e os da camada oculta, e os valores de ativação para cada neurônio são calculados da seguinte forma:

$$a_j = \sum_{i=1}^{n} w_{ji}^{(1)} x_i + w_{j0}^{(1)}$$

Onde j = 1, ... , M é número de neurônios da primeira camada oculta, $x_i$ são as observações das variáveis de entrada ($x_1$, ..., $x_n$), $w_{ji}^{(1)}$ e $w_{j0}^{(1)}$ são os pesos e os parâmetros de viés correspondentes à primeira camada (1) e $a_j$ são as ativações. Essas últimas são transformadas usando uma função de ativação $z = h(\alpha_j)$, do tipo logística sigmóide ou tangente hiperbólica. Os valores das ativações são novamente combinados fornecendo as ativações das unidades de saída:

$$a_k = \sum_{j=1}^{n} w_{kj}^{(2)} z_j + w_{k0}^{(2)},$$

Onde k = 1, ... , K é o número total de saídas. Essa transformação corresponde à segunda camada da rede e $w_{k0}^{(2)}$ é o viés. As ativações das unidades de saída são transformadas usando as funções de ativação apropriadas para fornecer as $y_k$ saídas da rede.

Para problemas de classificação as ativações das unidades de saída podem ser transformadas usando a função logística sigmoide

$$y_k = \varphi(a_k) = \frac{1}{1 + \exp(-a_k)}$$

Desta forma, todos os estágios de confecção da rede podem ser combinados em uma única função de ativação sigmoidal com a seguinte forma

$$y_k(x, \boldsymbol{w}) = \varphi\left(\sum_{j=1}^{M} w_{kj}^{(2)}\, h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$

Onde o conjunto de todos os parâmetros de pesos e viés foram agrupados em um único vetor $\boldsymbol{w}$. O modelo da rede é simplesmente uma função não linear das variáveis de

entrada ($x_i$) para um conjunto de variáveis de saída ($y_k$) controlado por um vetor $\boldsymbol{w}$ de parâmetros ajustáveis.

Para a definição da estrutura mais adequada, a rede é treinada com um algoritmo de treinamento que ajusta os pesos e viés em função do erro quadrático médio obtido ao final de cada rodada (ou época) de treinamento. Um dos algoritmos mais eficientes é o "back-propagation"(BASHEER et al., 2000).

No presente estudo a foi estabelecida rede perceptron multicamadas constituída de apenas uma camada oculta. De modo que as redes treinadas possuíram apenas uma camada de entrada, uma camada oculta e uma camada de saída. A camada de entrada foi composta de 56 neurônios, referentes às variáveis ecogeográficas do local de origem dos acessos. O número de neurônios da camada oculta foi definido pelo algoritmo de treinamento "back-propagation" e as funções de ativação testadas foram logística e tangente hiperbólica. O critério de performance utilizado foi o erro quadrático médio ($e \leq 0{,}001$) obtido pela comparação das saídas obtidas (classificação binária) com valores conhecidos dos exemplos de treinamento. Para o estabelecimento da rede foram utilizados 50000 épocas (iterações) de treinamento.

**Florestas Aleatórias**

Florestas aleatórias (RF) é algoritmo de particionamento recursivo que combina predições feitas por várias árvores de decisão (BREIMAN, 2001). As Florestas aleatórias são baseadas na metodologia de árvores de decisão, na qual os dados são particionados recursivamente em regiões (ou nós) de classificação homogêneas ou quase homogêneas.

As florestas aleatórias são desenvolvidas a partir de uma coleção de centenas a milhares de árvores, onde cada uma é criada usando apenas uma amostra aleatória dos dados. Além disso, em cada nó de cada árvore é utilizada apenas uma amostra das variáveis na partição de classes. As árvores são "cultivadas" até um tamanho máximo sem "poda" e agregadas formando as "florestas".

A classificação de uma observação é feita por todas as árvores da floresta. A decisão final da classe a qual pertence a observação é feita por maioria de votação, ou seja, a classe mais "votada" pelas árvores é atribuída à observação (STROBL et al. 2009). Para a construção das florestas aleatórias dois parâmetros devem ser especificados: número de variáveis disponíveis para divisão em cada nó de cada árvore (mtry) e número de árvores na floresta aleatória (ntree). Esses dois parâmetros são escolhidos por processo de otimização

que busca minimizar o erro quadrático médio ($e \leq 0{,}001$) observado na classificação de amostras não utilizadas na construção das árvores.

**Avaliação dos modelos**

A avaliação do desempenho dos modelos foi realizada pela comparação dos parâmetros obtidos pelas estatística Kappa e área abaixo da curva (AUC) de características de operação do receptor (ROC).

A estatística Kappa é utilizada para comparar a habilidade de diferentes modelos em classificar corretamente variáveis categóricas. Kappa leva em consideração tanto a acurácia observada quanto a esperada simplesmente por acaso, ambas baseadas na matriz de confusão gerada pelo modelo. A estatística pode assumir valores entre -1 e 1, o valor 1 indica uma concordância perfeita entre o modelo e as classes observadas. Porém, dependendo do contexto e acurácia esperada valores de Kappa entre 0.3 e 0.5 podem indicar uma boa concordância (KUHN e JOHNSON, 2013).

A curva de características de operação do receptor é uma técnica para caracterizar a performance de um modelo de classificação. Trata-se de uma representação gráfica que ilustra a performance do classificador binário quando seu limiar de discriminação é variado. A curva é criada quando a taxa de verdadeiros positivos é plotada no eixo vertical contra a taxa de falsos positivos no eixo horizontal, em diferentes limiares. No gráfico da curva ROC um bom classificador produz uma curva próxima ao canto superior esquerdo ou da coordenada (0,1), representando baixa taxa de falsos positivos e falsos negativos. Por outro lado, um classificador fraco produz uma curva ROC que segue a linha diagonal do canto inferior esquerdo para canto superior direito (WEBB e COPSEY, 2011).

Como diferentes classificadores produzem diferentes curvas ROC, uma maneira de comparar sua performance se dá pelo cálculo da área abaixo da curva ROC (AUC). Áreas menores ou iguais a 0.5 representam aleatoriedade e com valores acima de 0.7 representam alta performance do classificador (FAWCETT, 2006).

**Importância de variáveis**

A importância relativa das variáveis ecogeográficas utilizadas no estabelecimento dos modelos foi determinada segundo as particularidades de cada abordagem. Nos modelos obtidos pelas máquinas de vetor de suporte (SVM), o critério empregado na quantificação da importância relativa foi a extração recursiva de características (SMV-RFE). Esse método classifica todas as características de acordo com seus pesos (scores) e elimina uma ou mais características com os menores pesos por um processo iterativo até a obtenção da máxima acurácia (GUYON et al, 2002).

Para as Florestas aleatórias a medida importância utilizada foi o aumento da taxa de classificações incorretas de uma árvore da floresta aleatória quando os valores observados da característica foram permutados aleatoriamente com características não utilizadas na construção das árvores (STROBL et al., 2009). Da mesma forma que nas máquinas de vetor de suporte, as características são ordenadas segundo a magnitude do número de classificações incorretas.

Para redes neurais a análise de sensibilidade foi empregada para determinar a importância das características. A análise de sensibilidade consiste em provocar "perturbações" nas características (variáveis de entrada da rede) para estimar mudanças na sensibilidade da estrutura e dos parâmetros da rede. As características submetidas a pequenas variações e que causam grande mudanças na classificação da rede neural são consideradas de maior importância (ENGELBRECHT et al., 1995).

**Software**

O processamento e preparo dos dados foi realizado com pacote caret do aplicativo computacional R (KUHN, 2008). Nas análises e avaliações dos modelos foram utilizados os seguintes pacotes: randomForest (Florestas aleatórias), nnet (Redes neurais artificiais), e1071 (Máquinas de vetor de suporte) e ROCR (estatística Kappa e AUC).

# 3. RESULTADOS

Todos os modelos utilizados no presente estudo foram capazes de classificar corretamente os genótipos de cártamo utilizando as variáveis ecogeográficas, com uma acurácia superior a 72% (Tabela 2). Todos os modelos tiveram performance similar em relação as estimativas dos parâmetros de desempenho.

**Tabela 2.** Estimativas das estatísticas de avaliação de desempenho e acurácia dos modelos

| Modelo | AUC | AUC_L | AUC_U | K | K_L | K_U | Acurácia |
|---|---|---|---|---|---|---|---|
| RF | 0.7220 | 0.5366 | 0.9093 | 0.435 | 0.2938 | 0.5761 | 0.7550 |
| SVM radial | 0.8234 | 0.6912 | 1.0000 | 0.349 | 0.2218 | 0.4761 | 0.7280 |
| SVM linear | 0.7861 | 0.5883 | 0.9661 | 0.412 | 0.2712 | 0.5527 | 0.7410 |
| NN | 0.7891 | 0.6107 | 0.9674 | 0.390 | 0.2528 | 0.5271 | 0.7140 |

RF: Florestas aleatórias; SVM radial e SVM linear: máquinas de vetor de suporte com núcleo do tipo radial e linear; NN: Redes neurais artificiais; AUC: área abaixo da curva ROC; K: estatística Kappa; L e U: limites superior e inferior do intervalo de confiança a 95%.

A extensão dos intervalos de confiança das estatísticas dos modelos indicaram que não houve superioridade de nenhum modelo em relação ao demais. Nos modelos SVM a mudança da função núcleo não provocou ganhos em termos de acurácia. A hipótese de que com o uso da função núcleo de base radial haveria uma melhor classificação não foi comprovada, mesmo com um parâmetro de custo (C) quatro vezes maior que do que na função linear (radial C=1 vs linear C=0.25). A simplicidade observada nos modelos RF (mtry = 2 e ntree = 500) e NN (3 camadas e 1 neurônio na camada oculta) acompanhada das acurácias elevadas não fornece indicativos de que o problema de classificação seja de alta complexidade.

Os valores das estatísticas Kappa e AUC sustentam essa observação. Os modelos exibiram valores de AUC iguais ao que seria esperado de um modelo robusto, com valores maiores que 0.7. Valores de Kappa acima de 0.4 também indicam boa concordância entre os modelos e dados de teste. Ainda que os valores de Kappa dos modelos SVM radial e NN tenham ficado abaixo de 0.4, os valores observados indicam considerável concordância de classificação. Ademais, deve se considerar que a estatística Kappa é reconhecidamente uma medida conservadora de concordância (SIM e WRIGHT, 2005).

A acurácia dos modelos também é ilustrada pelos curvas ROC (Figura 3a). As curvas do modelos se mostraram acima da linha diagonal, tendendo ao canto superior esquerdo. Uma vez que não seguiram a linha diagonal, reforça a ideia de que os modelos tiveram uma boa performance, classificando os genótipos corretamente e com poucos falsos positivos. A linha diagonal representada nos gráficos indica o limiar para um modelo ser considerado aleatório.

A curva ROC de um modelo que segue essa linha diagonal ou que se encontra abaixo dela não deve ser utilizado para classificação (FAWCETT, 2006).

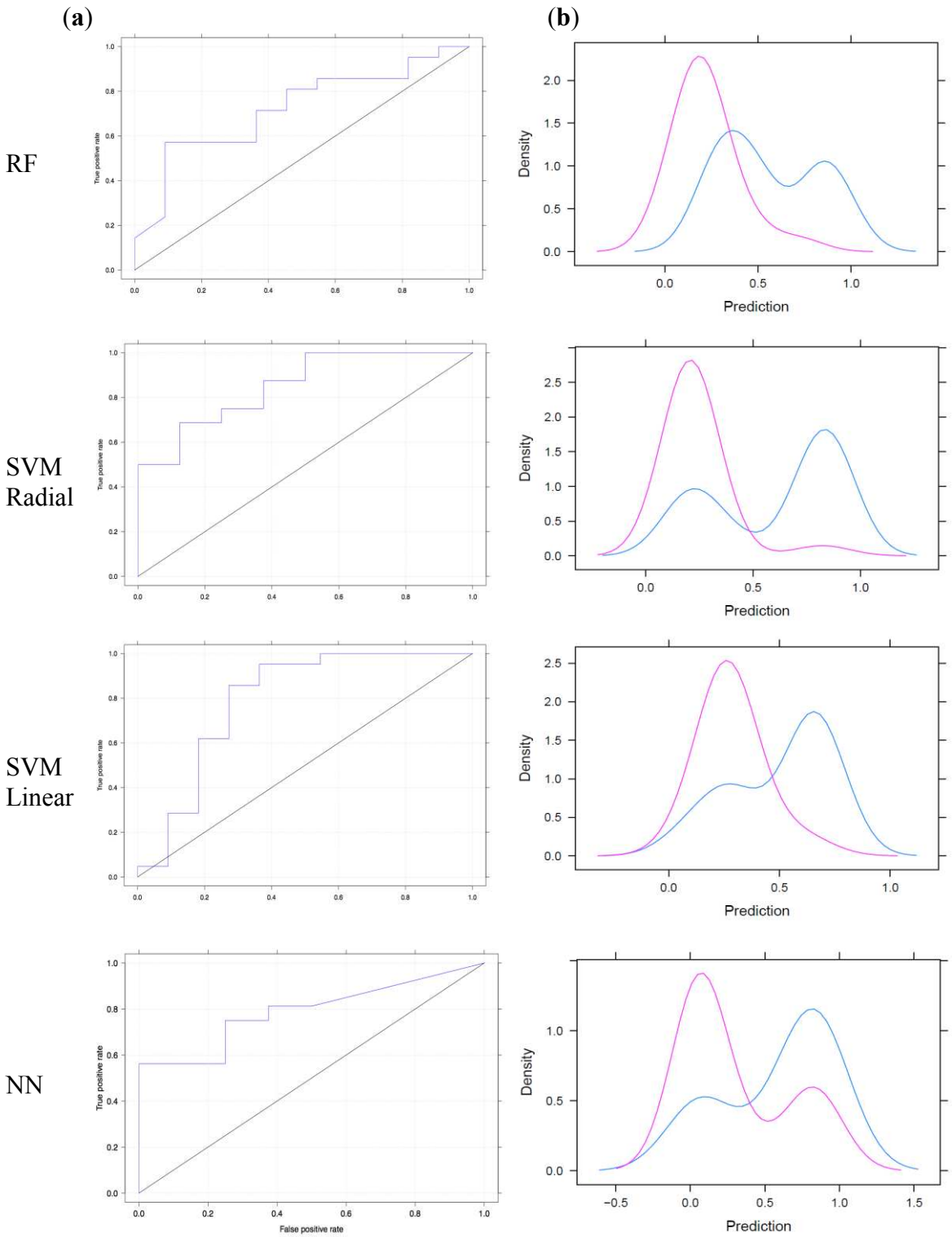



**Figura 3**. Curvas ROC (**a**) e gráficos de densidade de predição (**b**) das classes de teor de óleo dos quatro modelos de classificação. RF: Florestas aleatórias; SVM radial e SVM linear: máquinas de vetor de suporte com núcleo do tipo radial e linear; NN: Redes neurais artificiais.

Os gráficos de densidade de predição construídos com os dados de teste suportam ainda mais as estatísticas Kappa e AUC (Figura 3b). Em cada um dos modelos, as densidades de predição das classes de teor de óleo representadas pelas curvas azuis e púrpuras mostraram que os modelos classificaram corretamente os genótipos nos grupos correspondentes ao seu teor de óleo. Há uma certa sobreposição entre as classes, porém nos modelos SVMradial e SVMlinear o grau de sobreposição é menor.

A importância relativa das características ecogeográficas apresentaram magnitudes semelhantes nos diferentes modelos (Figura 4). Dentre as 56 características utilizadas, as mais importantes foram bio2 (intervalo médio de temperatura diurna no mês de fevereiro), prec2 (precipitação no mês de fevereiro), prec11 (precipitação no mês de novembro), alt (altitude), bio19 (precipitação no trimestre mais frio), prec12 (precipitação no mês de dezembro) e prec1 (precipitação no mês de janeiro).

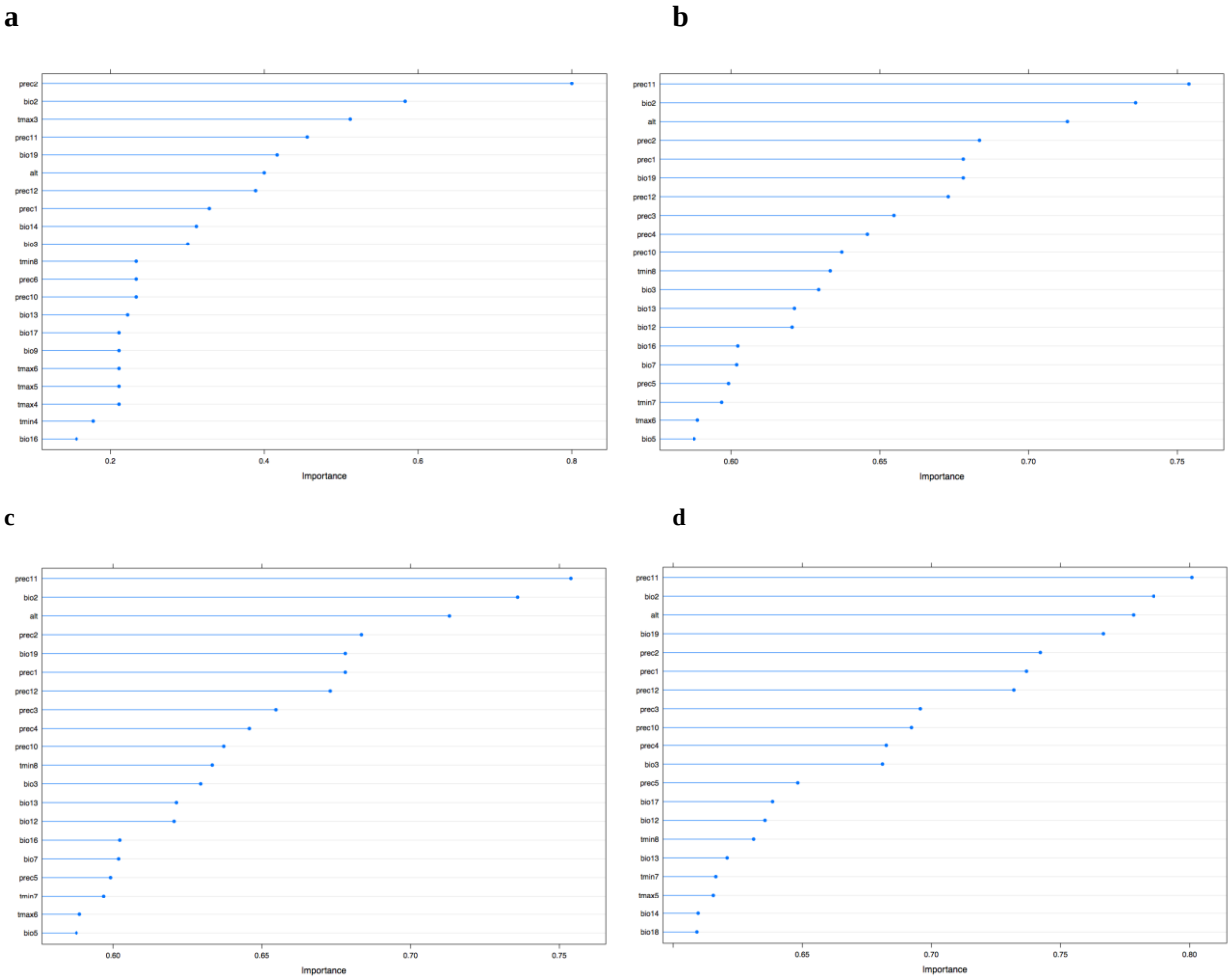



**Figura 4**. Importância das variáveis ecogeográficas nos modelos: (a) Florestas aleatórias (RF); (b) Máquinas de vetor de suporte com núcleo radial (SVM radial); (c) Máquinas de vetor de suporte com núcleo linear (SVM linear); (d) Redes neurais artificiais (NN).

# 4. DISCUSSÃO

O presente estudo mostrou que em genótipos de cártamo há associação preditiva entre teor de óleo e características ecogeográficas. Em outras palavras, a distribuição geográfica de genótipos de cártamo com alto teor de óleo não é aleatória mas ligada, em certa medida, a fatores ambientais. Os modelos apresentados confirmam a hipótese de que os genótipos com características em comum tendem a refletir as pressões de seleção do ambiente de origem. Estudos da variação do teor de óleo em bancos de germoplasma já forneciam indícios de associação entre o teor de óleo e o local de origem dos genótipos. Mesmo com grande diversidade de teores (16 a 38%), genótipos provenientes de determinadas regiões apresentavam maior teor de óleo, variando entre 33 e 37% por exemplo (CLAASSEN et al., 1945; ASHRI et al., 1977).

A ação do ambiente sobre determinadas características também é significativa em outras espécies. Em cevada (Hordeum vulgare L.) caracteres agronômicos e de resistência à patógenos podem ser preditos com segurança por variáveis ambientais (ENDRESEN, 2010). Caracteres relacionados à adaptação de feijão fava (Vicia faba L.) a estresses abióticos também são muito influenciados pelo ambiente, havendo uma relação preditiva muito forte, principalmente para os caracteres temperatura do dossel e conteúdo relativo de água (KHAZAEI et al., 2013). Diversos estudos com trigo (Triticum aestivum ssp aestivum) também comprovam a forte associação da resistência a patógenos e insetos praga com características ecogeográficas (EL-BOUHSSINI et al., 2010; ENDRESEN et al., 2011; BARI et al., 2012).

A extensa distribuição nos continentes europeu, asiático e africano faz com que o cártamo esteja sujeito a várias condições ambientais. Especialmente o clima e a geografia tem grande influência na dispersão, no fluxo gênico e na seleção, levando ao desenvolvimento de diversos ecótipos de cártamo. Esses podem ser definidos como variedades ou populações adaptadas a condições ambientais específicas, cuja manifestação de determinado fenótipo/genótipo é resultado da seleção local (TURESSON, 1922). Ecótipos de cártamo apresentam grande diversidade de características, assim como diferentes teores de óleo (GAO et al., 2001). Essa variabilidade foi um dos fatores motivadores do presente estudo.

A possibilidade de associar caracteres de semente, como o teor de óleo, a condições ambientais de ecótipos também tem originado outras pesquisas. Em um estudo que teve como objetivo investigar a previsibilidade do teor de óleo e a composição de ácidos de 360 ecótipos

de Arabidopsis thaliana, O'Neill et al. (2003) relataram extensa variação natural consistente com origem geográfica, suficiente para formação de coleções nucleares. Em um estudo similar Endresen (2010) relatou a associação preditiva entre características ecogeográficas e peso de sementes de variedades tradicionais ("landraces") de cevada nórdica. Nessa situação as "landraces" devem ser vistas como o equivalente agrícola dos ecótipos, cuja adaptação à geografia local é resultado não apenas da seleção do ambiente mas também da pressão de seleção humana.

A hipótese de associação entre o teor de óleo e características ambientais está fundamentada em princípios biológicos e ecológicos. O teor de óleo da semente pode ser considerado uma característica adaptativa com potencial de garantir vantagens em determinadas condições ambientais. O óleo é uma fonte de reserva de energia para o processo de germinação e seus constituintes (lipídeos) atendem efetivamente aos requerimentos de carbono e energia necessários para o início do processo. Plantas com sementes ricas em óleo e mais pesadas investem mais energia na progênie do que plantas com sementes mais leves e com maior conteúdo de amido, aumentando assim o potencial de estabelecimento e sobrevivência em ambientes estressados e com pouco recursos. Além disso, o teor de óleo é herdável e controlado por vários genes, portanto, está sujeito a mudanças evolucionárias. Em plantas selvagens o teor de óleo pode representar um ótimo seletivo ou um compromisso seletivo com o local de origem (LEVIN, 1974; O'NEILL et al., 2003).

Como o cártamo está distribuído do Mediterrâneo ao oceano Pacífico em latitudes entre 20ºS e 40ºN, está exposto a uma grande variação de condições climáticas (DAJUE e MÜNDEL, 1996). A adaptação a essas condições também pode levar a modificações dos constituintes, ácidos graxos, do óleo de cártamo segundo relatado por Ladd e Knowles (1971). Os autores observaram maior porcentagem de ácido linoleico em genótipos provenientes de locais com clima frio. Yang et al. (1993) apoiaram essa observação e adicionaram que o alto conteúdo de ácido linoleico também está associado com altas latitudes e altitudes, além de grande flutuação de temperaturas diurnas e noturnas.

Entre os vários fatores de seleção que impulsionam a evolução da composição de ácidos graxos Linder (2000) sugere a temperatura na germinação como o fator de maior importância. A temperatura no momento da germinação seleciona as proporções de ácidos graxos em sementes oleaginosas, de forma a otimizar a reserva e a taxa de produção de energia durante o processo germinativo. Sementes de plantas de altas latitudes (e altitudes) tendem a ter mais ácidos graxos insaturados. O clima mais fresco desses ambientes torna o

catabolismo de ácidos graxos insaturados mais viável, de forma que sementes com maior proporção desse ácido são capazes de germinar mais cedo, crescer rapidamente e estabelecer plantas mais cedo. Evidencias biogeográficas da performance de germinação no gênero Helianthus suportam essa teoria.

Como o cártamo está adaptado a uma ampla variação de temperaturas, outras ações de seleção, além da temperatura, são responsáveis por selecionar o teor de óleo. No presente estudo foi demonstrado que características ecogeográficas baseadas em temperatura e precipitação estão associadas ao teor óleo de cártamo. Principalmente a intensidade de precipitação nos meses de novembro a fevereiro, em locais com diferentes altitudes, atua como um fator determinante no conteúdo de óleo (Figura S1).

Os gráficos de densidade de predição traçados junto aos gráficos ROC (Figura 3b) demonstram que as classes de alto e baixo teor de óleo incluem genótipos com certo grau de distinção, e as bases dessa diferença são as variáveis ecogeográficas que atuaram como critério de seleção nos locais de coleta. De fato, a verificação dessa associação foi feita de maneira apropriada pela estratégia de identificação focada de germoplasma (FIGS).

Os modelos baseados em aprendizado de máquina utilizados pela FIGS podem ser aplicados como um padrão de busca para identificar outros genótipos de cártamo com alto teor de óleo em localidades com geografia similar. Diversos estudos reconheceram o potencial dessa estratégia na rápida identificação de genótipos com as seguintes características: tolerância a seca e calor (KHAZAEI et al., 2013), resistência a insetos pragas (EL-BOUHSSINI et al., 2010) e resistência à doenças (ENDRESEN et al., 2011; ENDRESEN et al., 2012; BARI et al., 2012). No entanto, a principal contribuição dos modelos está na utilização eficiente do bancos de germoplasma porque aumentam a probabilidade de encontrar genótipos (selvagens ou "landraces") com alto teor de óleo. Na prática esses genótipos poderiam ser selecionados pelos modelos e utilizados como candidatos em triagens para alto teor de óleo. Assim a frequência de identificação de genótipos com alto teor de óleo seria maior que a seleção aleatória ou via coleção nuclear.

A performance dos modelos de aprendizado de máquina utilizados transmite confiança aos resultados alcançados pelo presente estudo. Esses modelos são flexíveis o suficiente para lidar com problemas complexos, não lineares, e apropriados para predizer padrões e processos complexos. Os modelos Florestas Aleatórias (RF), Máquinas de Vetor de Suporte (SVM) e Redes Neurais Artificiais (NN) tiveram performance similar, sobretudo nas estatísticas Kappa e AUC. Bari et al. (2012) também utilizaram as mesmas técnicas e observaram acurácias de

até 77%. Em um estudo com o propósito de identificar características relacionadas à tolerância a seca Khazaei et al. (2013) obtiveram modelos de SVM e RF com alta acurácia mesmo com poucos genótipos nos conjuntos de treinamento. Da mesma forma que o presente estudo, os estudos anteriores também condensaram as classes da característica estudada em apenas duas classes, e os modelos atingiram performances excelentes (ENDRESSEN, 2010; ENDRESSEN et al., 2011).

O número e a forma de classificação dos acessos utilizados nos estudos de associação não é um fator limitante para o estabelecimento de modelos informativos quando as associações entre características fenotípicas e ecogeográficas foram estabelecidas por processos de seleção genética e adaptação. A associação preditiva envolvendo poucos genótipos também se mostra significativa sob condições de fluxo gênico restrito e adaptação a ambientes específicos (ENDRESEN, 2010; LARJAVAARA, 2014).

O grande potencial da estratégia FIGS e a robustez dos modelos em captar as variações ambientais podem expandir as possibilidades de explorar bancos de germoplasma de cártamo em busca de genótipos com alto teor de óleo, novas fontes de alelos para o melhoramento, formação de coleção temáticas e estudos de genética de associação para o teor de óleo.

# 5. CONCLUSÃO

Há associação preditiva entre o teor de óleo e variáveis ecogeográficas do local de origem dos genótipos. Os resultados dos modelos reforçam o fato de que a estratégia de identificação focada de germoplasma (FIGS) pode aperfeiçoar a identificação de genótipos de cártamo com maior teor de óleo. A associação preditiva entre o teor de óleo e as características ecogeográficas do local de origem do germoplasma aumenta as chances de encontrar genótipos com alto teor óleo distribuídos em bancos de germoplasma.

# 6. REFERÊNCIAS


ASHRI, A.; KNOWLES, P. F.; URIE, A. L.; ZIMMER, D. E.; CAHANER, A.; MARANI, A. Evaluation of the germplasm collection of safflower, Carthamus tinctorius. III. Oil content and Iodine value and their associations with other characters. **Economic Botany**, v. 31, p. 38-46. 1977.

BARI, A.; STREET, K.; MACKAY, M.; ENDRESEN, D. T. F.; DE PAUW, E.; AMRI, A. Focused identification of germplasm strategy (FIGS) detects wheat steam rust resistance linked to environmental variables. **Genetic Resources and Crop Evolution**, v. 59, p. 1456-1481. 2012.

BERGMAN, J .W.; CARLSON, G.; KUSHNAK, G.; RIVELAND, N. R,; STALLKNECHT, G. 1985. Registration of 'Oker' safflower. **Crop Science**, v. 25, p. 1127-1128. 1985.

BUSBY, J. R. BIOCLIM - a bioclimatic analysis and prediction system. In: Margules, C.R., Austin, M.P. (Eds.). **Nature Conservation**. Australia: CSIRO, 1991. p. 64-68.

CHAPMAN, M. A.; HVALA, J.; STREVER, J.; BURKE, J. M. Population genetic analysis of safflower (Carthamus tinctorius; Asteraceae) reveals a near eastern origin and five centers of diversity. **American Journal of Botany**, v. 97, n. 5, p. 831-840. 2010.

CLAASEN R. E.; KIESSELBACH, T.A. Experiments with safflower in western Nebraska. **Bulletin 376**, Lincoln, Nebraska, 1945. 28 p.

DAJUE L.; MÜNDEL, H. H. **Safflower. Carthamus tinctorius L**. Promoting the conservation and use of underutilized and neglected crops 7. Institute of Plant Genetics and Crop Plant Research, Gatersleben/International Plant Genetic Resources Institute, Rome, Italy. 1996. 83 p.

DAJUE, L.; ZHOU, M.; RAO, V. R. **Characterization and Evaluation of Safflower Germplasm**. Geological Publishing House, 1993. 260 p.

EL-BOUHSSINI, M.; STREET, K.; AMRI, A.; MACKAY, M.; OGBONNAYA, F. C.; OMRAN, A.; ABDALLA, O.; BAUM, M.; DABBOUS, A.; RIHAWI, F. Sources of resistance in bread wheat to Russian wheat aphid (Diuraphis noxia) in Syria identified using the Focused Identification of Germplasm Strategy (FIGS). **Plant Breeding**, v. 130, p. 96-97. 2010.

ENDRESEN, D. T. F. Predictive association between trait data and ecogeographic data for nordic barley landraces. **Crop Science**, v. 50, p. 2418-2430. 2010.

ENDRESEN, D. T. F.; STREET, K.; MACKAY, M; BARI, A; DE PAUW, E. Predictive association between biotic stress traits and ecogeographic data for wheat and barley landraces. **Crop Science**, v. 51, p. 2036-2055. 2011.

ENDRESEN, D. T. F.; STREET, K.; MACKAY, M.; BARI, A.; DE PAUW, E.; NAZARI, K.; YAHYAOUI, A. Sources of Resistance to Stem Rust (Ug99) in Bread Wheat and Durum Wheat Identified Using Focused Identification of Germplasm Strategy (FIGS). **Crop Science**, v. 52, n. 2, p. 764-773. 2012.

ENGELBRECHT, A. P.; CLOETE, I.; ZURADA, J. M. Determining the significance of input parameters using sensitivity analysis. **Lecture Notes in Computer Science**, v. 930, p. 382-388. 1995.

FAOSTAT. FAO Statistical Databases, Disponível em : <http://faostat.fao.org/site/636/DesktopDefault.aspx?PageID=636#ancor>. Acesso em: 25 de Out. de 2013.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, p. 861-874. 2006.

GAO, W.; RAMANATHA, R.; ZHOU, M. Eds. 2001. Plant genetics resources conservation and use in China. In: National Workshop on Conservation and Utilization of Plant Genetics Resources, 2001, Beijing. **Proceedings of National Workshop on Conservation and**

**Utilization of Plant Genetics Resources**. Beijing: Institute of Crop Germplasm Resources, CASS and IPGRI Office for East Asia. 2001.

GUYON, I.; WESTON, J.; BARNHILL, S.; VAPNIK, V. Gene selection for cancer classification using support vector machines. **Machine Learning** 2002, v. 46, p. 389-422. 2002.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference and Prediction**. Springer-Verlag: New York, 2009. 745 p.

HIJMANS, R. J.; CAMERON, S. E.; PARRA, J. L.; JONES, P. G.; JARVIS, A. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, v. 25, p. 1965-1978. 2005.

INTERNATIONAL CENTER FOR AGRICULTURAL RESEARCH IN THE DRY AREAS (ICARDA). A new approach to mining agricultural gene banks – to speed the pace of research innovation for food security. 'FIGS' - the Focused Identification of Germplasm Strategy. **Research to Action 3**. ICARDA: Beirut, Lebanon. 2013. 22p.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to statistical learning with applications in R**. Springer: New York. 2013, 426p.

JOHNSON, R.C.; BERGMAN, J .W.; FLYNN, C. R. Oil and meal characteristics of core and non-core safflower accessions from the USDA collection. **Genetic Resources and Crop Evolution**, v. 46, p. 611-618. 1999.

KHAZAEI, H.; STREET, K.; BARI, A.; MACKAY, M.; STODDARD, F. L. The FIGS (Focused Identification of Germplasm Strategy) Approach Identifies Traits Related to Drought Adaptation in Vicia faba Genetic Resources. **PLoS ONE 8(5): e63107. doi:10.1371/journal.pone.006310**. 2013.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. Springer:New York. 2013, 620p.

KUHN, M. Building predictive models in R using the caret package. **Journal of Statistic Software**, v. 28, n. 5, p. 1-26. 2008.

LARJAVAARA, M. The world's tallest trees grow in thermally similar climates. **New Phytologist**, v. 202, p. 344-349. 2014.

LEVIN, D. A. The oil content of seeds: an ecological perspective. **American Naturalist**. v. 108, p. 193-206. 1974.

LINDER, C. R. Adaptive evolution of seed oils in plants: accounting for the biogeographic distribution of saturated and unsaturated fatty acids. **American Naturalist**. v. 156, n. 4, p. 442-458. 2000.

MACKAY, M. C.; STREET, K. Focused identification of germplasm strategy – FIGS. In: 54th Australian Cereal Chemistry Conference and the 11th Wheat Breeders'Assembly, 2004, Melbourne. **Proceedings of the 54th Australian Cereal Chemistry Conference and the 11th Wheat Breeders' Assembly**. Melbourne: Royal Australian Chemical Institute. 2004. p. 138-141.

OLDEN, J. D.; LAWLER, J. J.; POFF, N. L. Machine learning methods without tears: a primer for ecologists. **The quarterly review of biology**. v. 83, n. 2, p. 171-193. 2008.

PEETERS, J. P.; WILKES, H. G.; GALWEY, N. W. The use of ecogeographic data in the exploitation of variation from gene banks. **Theoretical and Applied Genetics**, v. 80, p. 110-112. 1990.

PRADA, D. Molecular population genetics and agronomic alleles in seed banks: search for a needle in a haystack? **Journal of Experimental Botany**, v. 60, n. 9, p. 2541-2555. 2009.

R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing. Disponível em: <http://www.R-project.org>. Acesso em: 20 out. 2013.

SIM, J.; WRIGHT, C. C. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. **Physical Therapy**, v. 85, p. 257-268. 2005.

SINGH, V.; DESHPANDE, M. B.; NIMBKAR, N. NARI-NH-1: the first non-spiny hybrid safflower released in India. **Sesame Safflower Newsletter**, v. 18, p. 77-79. 2003.

STROBL, C.; MALLEY, J.; TUTZ, G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. **Psychological Methods**, v.14, p. 323–348. 2009.

TURESSON, G. The genotypical response of the plant species to the habitat. **Hereditas**, v. 3, p. 211-350. 1922.

USDA, ARS, National Genetic Resources Program. Germplasm Resources Information Network - (GRIN). [Online Database] National Germplasm Resources Laboratory, Beltsville, Maryland. Disponível em: <http://www.ars-grin.gov/cgi-bin/npgs/html/crop.pl?108>. Acesso em: 03 out. 2013.

WEBB, A. R.; COPSEY, K. D. **Statistical Pattern Recognition**. John Wiley & Sons: Chichester. 2011. 666 p.

YANG, J.; JIANG, Y.; LIU, X.; ZHAO, Y. The research on the germplasm resources of safflower with different contents of fatty acids. In: Third International Safflower Conference, 1993, Beijing. **Proceedings of the Third International Safflower Conference**. Beijing: Beijing Botanical Garden, Institute of Botany, Chinese Academy of Sciences. 1993. p. 358-365.

# APÊNDICES

**Tabela A**. Matrizes de confusão do conjunto de dados de treinamento e teste

**Florestas aleatórias**

| Conjunto de Treinamento | | | | |
|---|---|---|---|---|
| Classe Original | Classe predita | | Precisão das classes | Acurácia |
| | Baixo | Alto | | |
| Baixo teor | 22 | 2 | 0.9617 | 0.9017 |
| Alto teor | 1 | 42 | 0.9767 | |

| Conjunto de Teste | | | | |
|---|---|---|---|---|
| Classe Original | Classe predita | | Precisão das classes | Acurácia |
| | Baixo | Alto | | |
| Baixo teor | 3 | 8 | 0.2727 | 0.5938 |
| Alto teor | 6 | 16 | 0.7272 | |

**Máquinas de vetor de suporte radial**

| Conjunto de Treinamento | | | | |
|---|---|---|---|---|
| Classe Original | Classe predita | | Precisão das classes | Acurácia |
| | Baixo | Alto | | |
| Baixo teor | 14 | 9 | 0.6086 | 0.8000 |
| Alto teor | 2 | 42 | 0.9545 | |

| Conjunto de Teste | | | | |
|---|---|---|---|---|
| Classe Original | Classe predita | | Precisão das classes | Acurácia |
| | Baixo | Alto | | |
| Baixo teor | 4 | 4 | 0.500 | 0.7500 |
| Alto teor | 5 | 20 | 0.800 | |

**Máquinas de vetor de suporte linear**

| Conjunto de Treinamento | | | | |
|---|---|---|---|---|
| Classe Original | Classe predita | | Precisão das classes | Acurácia |
| | Baixo | Alto | | |
| Baixo teor | 16 | 8 | 0.6667 | 0.7910 |
| Alto teor | 6 | 37 | 0.8605 | |

| Conjunto de Teste | | | | |
|---|---|---|---|---|
| Classe Original | Classe predita | | Precisão das classes | Acurácia |
| | Baixo | Alto | | |
| Baixo teor | 7 | 4 | 0.6364 | 0.8125 |
| Alto teor | 2 | 20 | 0.9090 | |

**Redes Neurais**

| Conjunto de Treinamento | | | | |
|---|---|---|---|---|
| Classe Original | Classe predita | | Precisão das classes | Acurácia |
| | Baixo | Alto | | |
| Baixo teor | 21 | 2 | 0.9130 | 0.8933 |
| Alto teor | 3 | 41 | 0.9318 | |

| Conjunto de Teste | | | | |
|---|---|---|---|---|
| Classe Original | Classe predita | | Precisão das classes | Acurácia |
| | Baixo | Alto | | |
| Baixo teor | 11 | 4 | 0.7500 | 0.7500 |
| Alto teor | 4 | 14 | 0.7500 | |

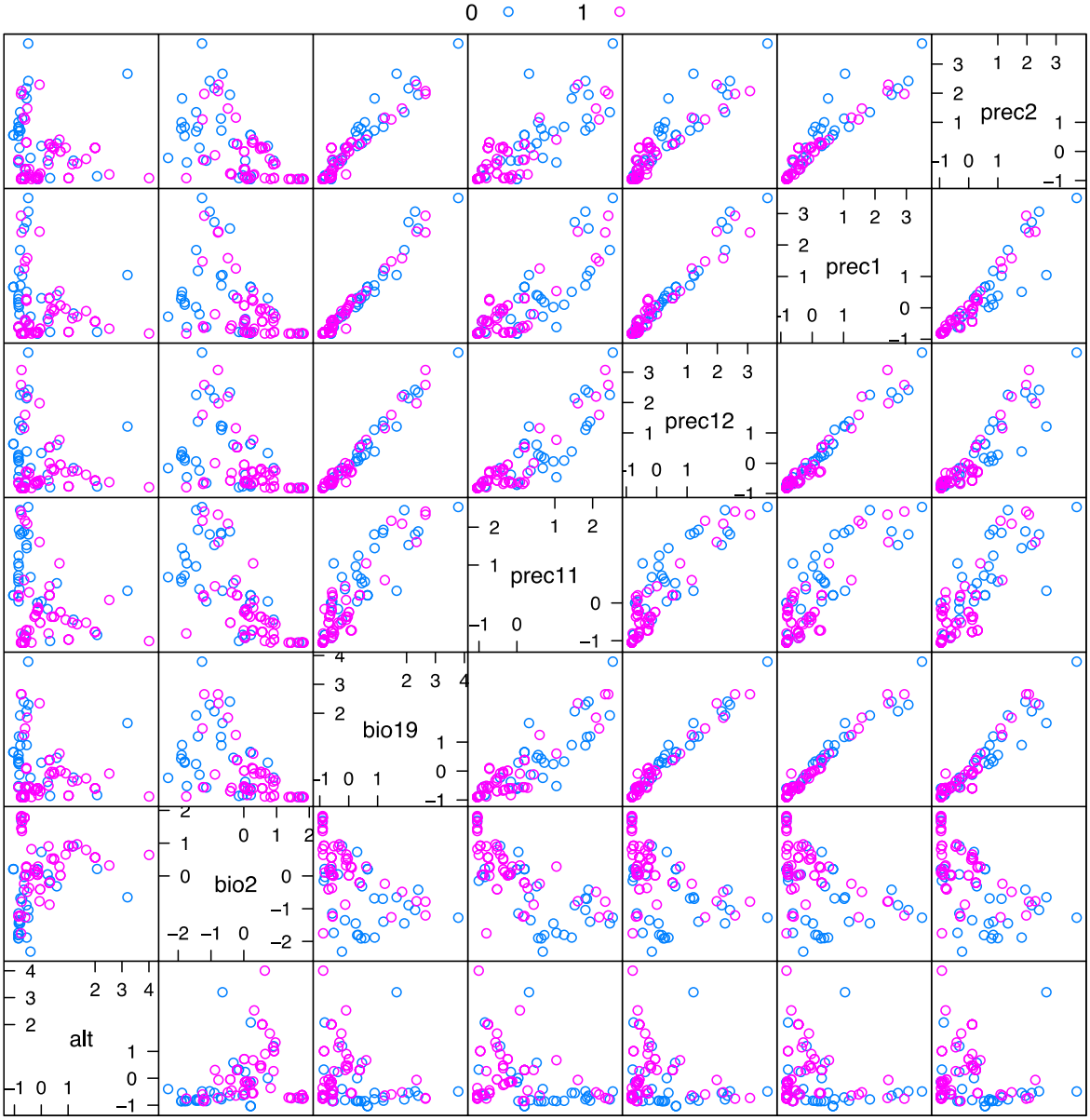



**Figura S1**. Distribuição das classes de teor óleo de genótipos cártamo nas características ecogeográficas de maior importância indicadas pelos modelos.