

CARACTERIZAÇÃO *IN SILICO* DE GENES RELACIONADOS A BIOSÍNTESE DE ÓLEO EM GERGELIM

Autores: Luziane Brandão Alves¹, Rebeca Almeida Gonçalves¹, Gabriella Reis Zaccaron¹, Mikael de Paula Brandão¹, Rodrigo Monte Lorenzoni², Edilson Marques Junior², Lucimara Cruz de Souza¹.

¹Universidade Federal de Lavras/Departamento de Biologia, Trevo Rotatório Professor Edmir Sá Santos, s/n - 37203-202 – Lavras – MG, Brasil, luzianealves@gmail.com, rebeca.goncalves@estudante.ufla.br, grzaccaron@gmail.com, mikael.brandao1@estudante.ufla.br, lucimaracruz@ufla.br.

²Sebra Agrícola S.A., Departamento de Pesquisa e Desenvolvimento, Lote Canarana I, Chácara G08, 78640-000 – Canarana – MT, rodrigo@sebra.ag, edilsonj@sebra.ag.

Resumo

O gergelim se destaca como uma das oleaginosas mais comercializadas no mundo e rica em ácidos graxos benéficos à saúde. A composição de lipídios é um dos principais objetivos no melhoramento genético da cultura. Nesse intuito, realizou-se a caracterização morfofuncional e estrutural de 12 genes, com base nas sequências obtidas a partir de marcadores microssatélites. Observou-se grande diversidade entre as proteínas, como em tamanho, peso, estabilidade e localização. Foram identificados domínios conservados exclusivos e um *motif* único (PVYPPVIPKPP) em uma proteína. Oito proteínas foram reveladas com códigos *KEGG Orthology* (KOs associadas) as vias metabólicas de lipídios. As informações obtidas possibilitarão uma base no aprimoramento de marcadores moleculares e estudos de expressão gênica voltados para o teor de óleo no gergelim.

Palavras-chave: *Sesamum indicum*. Lipídios. Proteínas. Oleaginosas.

Área do Conhecimento: Ciências Biológicas – Biologia Molecular.

Introdução

O gergelim (*Sesamum indicum* L.) é uma espécie diploide pertencente à família Pedaliaceae e está entre as principais culturas oleaginosas com expressivo crescimento nos últimos anos. Essa cultura se destaca por ser fonte de proteínas, carboidratos, minerais essenciais e, principalmente, pela alta composição de óleo presente em suas sementes (Tesfaye et al., 2022). Aproximadamente 85% dos ácidos graxos presentes no gergelim são insaturados, o que confere grande valor nutricional e importância para a saúde humana (Teklu et al., 2022).

De acordo com os dados da FAO, em 2023 a área cultivada de gergelim obteve um alcance de 13.052.038 ha e produção estimada em 6.788.881,07 toneladas. O continente africano e asiático detém a maior parte dessa produção com 55,8% e 37,4%, respectivamente (FAO, 2025). Nos últimos anos, a produção em outros continentes tem avançado, em especial na América, onde o Brasil se destacou como o sétimo maior produtor mundial, atingindo uma produção de 289 mil toneladas. As regiões Norte e Centro-oeste do país constituem-se as maiores produtoras de gergelim e o estado do Mato Grosso ocupa a liderança nacional (CONAB, 2025).

O óleo de gergelim é reconhecido pela ótima qualidade em decorrências dos ácidos graxos poli-insaturados, sendo estes 43% de ácidos graxos oleico, 43% de ácidos linoleico, além de 9% de ácidos palmítico e 4% de esteáricos (Biswas et al., 2018). O teor e a composição desses ácidos graxos estão diretamente associados ao mecanismo de genes-chaves envolvidos na via biossintética de óleo, permitindo inferir sobre a diversidade genética e suas interações funcionais.

Entre os principais objetivos dos programas de melhoramento genético de gergelim está em associar produtividade e qualidade do óleo nas sementes, o que envolve tecnologias e a caracterização de genes que atuam nestas rotas metabólicas de forma detalhada. Nesse contexto, o objetivo deste trabalho foi caracterizar a nível estrutural e funcional, proteínas codificadas por genes envolvidos no

metabolismo lipídico em gergelim a partir de análises físico-químicas e vias metabólicas associadas a fim de subsidiar o desenvolvimento de marcadores moleculares e estratégias de melhoramento genético voltadas ao aumento do teor de óleo em gergelim.

Metodologia

Marcadores microssatélites em genótipos de gergelim foram associados os seus IDs no banco *Sesamum indicum Genetic Discovery Database* (SiGeDiD) (<https://sigedid.ucad.sn/>). Com base nas anotações funcionais, 12 genes relacionados ao teor e composição de óleo, conforme descrito por Wei et al. (2015), foram selecionados. Para as análises *in silico* foram utilizadas somente as sequências de proteínas codificadas por esses genes.

A predição de regiões de peptídeo sinal foi realizada por meio do programa Signal IP v. 6.0 (<https://services.healthtech.dtu.dk/services/SignalP-6.0/>). A caracterização físico-química das proteínas foi conduzida pela ferramenta ProtParam do servidor ExPASy (<https://web.expasy.org/protparam/>) (Gasteiger et al. 2005), considerando número de aminoácidos, peso molecular, ponto isoelétrico, índice de instabilidade, índice alifático, média de hidropaticidade (GRAVY) e estabilidade.

As hélices transmembranares foram preditas pelo TMHMM 2.0 (<https://services.healthtech.dtu.dk/services/TMHMM-2.0/>) com base no modelo oculto de Markov (KROGH et al., 2001). A provável localização celular foi determinada pelo software DeepLoc2 v. 2.1 (<https://services.healthtech.dtu.dk/services/DeepLoc-2.1/>) utilizando parâmetros padrão.

A análise de domínios conservados e suas posições foram realizados pelo do Conserved Domain Database (CDD-Search) do NCBI (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>). A predição de motivos conservados foi feita com o pacote MEME suite (*Multiple Em for Motif Elicitation*) buscando até 10 motivos (<https://meme-suite.org/meme/tools/meme>; BAILEY et al., 2009).

As possíveis modificações pós-traducionais, considerando os sítios de fosforilação em resíduos de serina, tirosina e treonina, foram analisadas pelo NetPHOS 3.1 (<https://services.healthtech.dtu.dk/services/NetPhos-3.1/>).

O mapeamento das vias metabólicas foi realizado a partir de uma análise de enriquecimento funcional com o eggNOG-mapper v.2 (<http://eggnog-mapper.embl.de/>) e os códigos KO obtidos foram submetidos ao BlastKOALA (<https://www.kegg.jp/blastkoala/>) para identificação das vias associadas.

Resultados

Foram avaliadas as propriedades físico-químicas das 12 proteínas envolvidas no metabolismo de óleos em gergelim (Tabela 1).

Tabela 1: Caracterização físico-químicas de proteínas em gergelim.

ID	aa	pI	PM (Kda)	II	IA	GRAVY	PS
SIN_1024652	548	8,55	59,303	49,3	74	-0,143	Não
SIN_1003248	351	6,1	39,18	49,38	86,41	-0,21	Não
SIN_1009923	835	5,86	91,18	49,53	77,78	-0,346	Não
SIN_1007213	191	8,88	19,93	67,17	103,51	0,243	Não
SIN_1007671	472	5,85	53,33	48,63	88,11	-0,387	Não
SIN_1008977	383	6,18	43,07	32,51	78,22	-0,395	Não
SIN_1022133	386	6,2	44,17	37,65	80,54	-0,547	Não
SIN_1010870	353	5,88	40,16	50,04	84,7	-0,382	Não
SIN_1015471	175	6,04	19,69	31,77	66,34	-0,225	Não
SIN_1025734	506	6,96	57,24	39,09	96,15	-0,158	Não
SIN_1021891	294	10,01	32,18	35,98	90,99	0,102	Não

A Ciência do NANO e seu impacto transformador no MACRO

SIN_1002815	245	5,59	27,63	40,91	79,76	-0,225	Não
-------------	-----	------	-------	-------	-------	--------	-----

ID: identificação do gene; aa: número de aminoácidos; pI: Ponto Isoelétrico; PM (Kda): Peso Molecular em quilodalton; II: Índice de Instabilidade; IA: Índice Alifático; GRAVY: Média de hidropaticidade; PS: Peptídeo Sinal.

As sequências analisadas (Tabela 1) variaram de 175 a 835 aminoácidos, com peso molecular entre 19,69 a 91,18 kDa. Os valores de ponto isoelétrico (pI) ficaram entre 5.59 a 10.01, classificando 75% das proteínas como ácidas e as demais 25% como básicas. A análise de hidrofobicidade das proteínas, baseada no índice GRAVY (*Grand Average of Hydropathy*), demonstrou que a maioria das proteínas possuem natureza hidrofílicas, com exceção de SIN_1007213 e SIN_1021891 que apresentaram valores positivos de GRAVY 0,243 e 0,102, respectivamente.

Quanto ao índice de instabilidade e a sua provável estabilidade metabólica (Tabela 01), SIN_1008977, SIN_1022133, SIN_1015471, SIN_1025734 e SIN_1021891 foram classificadas como estáveis (índice <40). A termoestabilidade, avaliada pelo índice alifático, variou de 66,34 a 103,51. SIN_1007213 se destacou com o maior índice alifático e GRAVY positivo, denotando uma possível correlação entre a proporção de resíduos alifáticos com a hidrofobicidade. Já SIN_1021891, apesar de ter valor positivo para GRAVY, mostrou índice alifático mediano, indicando uma possível relação e contribuição de outros resíduos não polares para o nível de hidrofobicidade desta proteína. Nenhuma proteína apresentou peptídeo sinal, sugerindo que não há endereçamento para outros compartimentos celulares.

No aspecto funcional e estrutural das proteínas (Tabela 2), a predição de localização subcelular indicou três proteínas no plastídio (SIN_1024652, SIN_1008977 e SIN_1022133), duas no citoplasma (SIN_1003248 e SIN_1010870), duas no núcleo (SIN_1009923 e SIN_1007671), duas como secretadas (SIN_1007213 e SIN_1015471) e três presentes no retículo endoplasmático (SIN_1025734, SIN_1021891 e SIN_1002815).

Tabela 2: Caracterização funcional e estrutural das proteínas envolvidas no metabolismo de óleo em gergelim

ID	ST	Ser	Tre	Tir	Loc.	Ass. Mem.	Dom. Cons.	Int. Dom.
SIN_1024652	0	30	17	6	Plastídio	Solúvel	3-oxoacil [proteína transportadora de acila] sintase II (KasII)	1-509
SIN_1003248	0	9	6	2	Citoplasma	Solúvel	dobra de alfa/beta hidrolase	83-321
SIN_1009923	0	50	26	8	Núcleo	Solúvel	Domínio START	339-569
							Homeodomínio	139-194
							Domínio de zíper de leucina básico (bZIP)	188-221
SIN_1007213	0	6	9	0	Extracelular	Solúvel	Proteína hidrofóbica	107-190
SIN_1007671	0	34	8	6	Núcleo	Solúvel	Histona acetiltransferase HAT1 N-terminal	41-146
							MOZ/SAS	260-277
							Histona acetiltransferas	222-369
SIN_1008977	0	10	11	3	Plastídio	Solúvel	Acil-proteína transportadora	1-383

A Ciência do NANO e seu impacto transformador no MACRO

							dessaturase	
SIN_1022133	0	17	16	2	Plastídio	Solúvel	Acil-ACP tioesterase (FAT)	96-383
SIN_1010870	0	36	15	5	Citoplasma	Periférica; Solúvel	Subunidade C da V-ATPase;	5-346
SIN_1015471	0	6	5	4	Extracelular	Solúvel	Proteína semelhante a Dirigente (DIR)	25-167
SIN_1025734	1	26	11	3	Ret. End.	Transm.	Citocromo P450 (CYP450)	68-486
SIN_1021891	4	10	12	3	Ret. End.	Transm.	-	-
SIN_1002815	1	11	6	2	Ret. End.	Transm.	Caleosina	66-233

ST: Segmentos transmembranares; SER, TRE, TIR: Sítios de fosforilação serina, treonina e tirosina; Loc: Localização; Ret. End.: Retículo Endoplasmático; Ass. Mem.: Associação com membrana; Transm.: Transmembranar; Dom. Cons.: Domínio Conservado; Int. Dom.: Intervalo de Domínio.

As proteínas associadas ao retículo endoplasmático apresentaram regiões transmembranares, caracterizando-as como transmembrananas. As demais foram preditas como solúveis, com destaque para a SIN_1010870, que apresentou associação simultaneamente como solúvel e periférica.

Todas as proteínas apresentaram resíduos de fosforilação (Tabela 2), sendo de forma geral a frequência maior para os resíduos de serina, seguidos de treonina e tirosina. Apenas três proteínas (SIN_1007213, SIN_1008977 e SIN_1021891) apresentaram resíduos de treonina superiores ao de serina.

Foram identificados domínios conservados em quase todas as proteínas, variando de 1 a 3 por sequência, conforme demonstrado na tabela 2, exceto para SIN_1021891 que não foram descritos domínios pelo NCBI - CDD. Os domínios identificados desde regiões iniciais até a porção final das sequências. SIN_1008977 foi a única proteína que apresentou o domínio acil-proteína transportadora dessaturase abrangendo toda a sequência de aminoácidos. Duas proteínas apresentaram três domínios cada, SIN_1009923 e SIN_1007671. Além disso, cada domínio identificado foi único a cada proteína.

A análise de *motifs* conservados realizados pelo MEME, identificou até 10 regiões conservadas por proteínas (Figura 1). SIN_1007213 apresentou um *motif* exclusivo (PVYPPVIPKPP). As proteínas SIN_1024652, SIN_1003248, SIN_1022133 e SIN_1010870 foram as mais diversas na quantidade de regiões *motifs* apresentadas. Em contrapartida, SIN_1007671, SIN_1015471, SIN_1025734, SIN_1021891 e SIN_1002815 apresentaram apenas um *motif*.

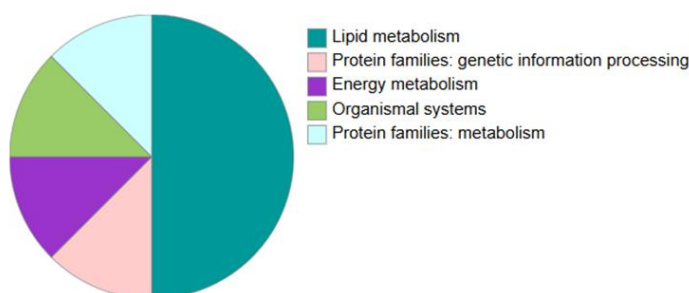
Figura 1: Frequência e localização dos *motifs* em cada uma das proteínas analisadas



A anotação funcional via eggNOG e BlastKoala atribuiu códigos KO para oito proteínas, a qual representa 66,7% do conjunto avaliado (Figura 2), 50% deles são relacionados a composição lipídica

em gergelim enquanto os demais estão envolvidas no metabolismo energético, sistemas do organismo e de famílias proteicas.

Figura 2: Relação dos processos metabólicos associados as proteínas com KO



SIN_1025734 apresentou múltiplos códigos KO, e todos os identificados foram utilizados na busca por vias metabólicas. No total foram identificadas 38 vias metabólicas, incluindo biossíntese de ácidos graxos, metabolismo de ácidos graxos, biossíntese de metabólitos secundários e biossíntese de ácidos graxos insaturados.

Discussão

Existe uma diversidade tanto nas características físico-químicas quanto nas funções metabólicas desempenhadas pelas proteínas. As proteínas com regiões transmembranares podem atuar no transporte de lipídios para a membrana e aquelas solúveis participam de processos cruciais da síntese lipídica. Nawade et al. (2022) identificaram regiões transmembranares variando de duas a nove em cinco famílias ligadas à via biossintética do óleo.

Não foram encontradas proteínas com peptídeos sinal, o que indica ausência de secreção extracelular ou direcionamento para organelas específicas, sugerindo que estas proteínas desempenham suas funções predominantemente em compartimentos intracelulares.

A anotação funcional revelou que algumas proteínas não apresentaram código KO, possivelmente devido à ausência de categorização completa desses genes nas bases de dados. As categorias funcionais aos códigos KO demonstram a participação das proteínas em no funções complementares e de suporte à biossíntese e regulação dos lipídios no gergelim. A análise de *motifs* conservados também demonstrou diversidade funcional. O *motif* PVYPPVIPKPP pode indicar uma adaptação específica e com uma função particular no metabolismo lipídico de gergelim.

Em relação aos domínios conservados, alguns pertencem a superfamílias e outros para proteínas específicas, proporcionando essa especialização funcional. Destaca-se a presença de domínios relacionados às enzimas KAS e acil-ACP tioesterase (FAT). Em gergelim, genes que codificam para a KAS foram identificados em regiões próximas de SNPs significativos associados a ácido oleico, um importante ácido graxo vegetal (Elsafy et al., 2024). A FAT atua na etapa final de ácidos graxos e determina o tamanho da cadeia de ácidos graxos que será formada e qual o tipo, podendo ser FATA com especificidade para grupos acil insaturados ou FATB, mais ativa em cadeias de acil saturados (Zhou et al. 2024)

Conclusão

Os resultados revelaram uma ampla diversidade estrutural e funcional entre as proteínas envolvidas na biossíntese de óleo em gergelim. A caracterização *in silico* proporcionou informações importantes para futuros estudos de diferenciação de cultivares, polimorfismos e expressão gênica. Destacam-se SIN_1007213, pela estabilidade, e SIN_1025734, com múltiplos KOs, como potencial regulador do metabolismo lipídico e da constituição genotípica da espécie.

Referências

BAILEY, T. L.; BODEN, M.; BUSKE, F. A.; FRITH, M.; GRANT, C. E.; CLEMENTI, L.; NOBLE, W. S. MEME SUITE: tools for motif discovery and searching. **Nucleic acids research**, v. 37, n. 2, p. 202-208, 2009. doi.org/10.1093/nar/gkp335

BISWAS, S.; NATTA, S.; RAY, D. P.; MONDAL, P.; SAHA, U. Til (Sesamum indicum L.) - An Underexploited but Promising Oilseed with Multifarious Applications: A Review. **International Journal of Bioresource Science**, v. 5, n. 2, p. 127-139, 2018, DOI: 10.30954/2347-9655.02.2018.8.

CONAB - **Companhia Nacional de Abastecimento**. <https://www.gov.br/conab/pt-br/atuacao/informacoes-agropecuarias/safras>. Acesso em 07 de agosto de 2025.

ELSAFY, M.; BADAWEI, W.; IBRAHIM, A.; HASSAN, A. B.; WANG, E. S.; BAILLO, E. H.; ABDELHALIM, T. S.; BAJGAIN, P.; RAHMATOV, M. Genome-wide association study of oil content and fatty acid composition in sesame (Sesamum indicum L.) under diverse environmental conditions. **Crop Science**, v. 65, n. 70099, p. 1 – 16, 2025. DOI: 10.1002/csc2.70099.

FAOSTAT. **Organização das Nações Unidas para a Alimentação e a Agricultura**; Serviço Estatístico Online FAOSTAT. <https://www.fao.org/faostat/en/#home>. Acesso em 07 de agosto de 2025.

KROGH, A.; LARSSON, B.; VON HEIJNE, G.; SONNHAMMER, E. L. L. Predicting transmembrane protein topology with a Hidden Markov Model: Application to complete genomes. **Journal of Molecular Biology**, v.305, n. 3, p. 567–580, 2001.

NAWADE, B.; KUMAR, A.; MAURYA, R.; SUBRAMANI, R.; YADAV, R.; SINGH, K.; RANGAN, P. Longer Duration of Active Oil Biosynthesis during Seed Development Is Crucial for High Oil Yield—Lessons from Genome-Wide In Silico Mining and RNA-Seq Validation in Sesame. **Plants**, v. 11, n. 2980, p. 1-24, 2022. <https://doi.org/10.3390/plants11212980>.

TEKLU, D. H.; SHIMELIS, H.; TESFAYE, A.; SHAYANOWAKO, A. I. T. Analyses of genetic diversity and population structure of sesame (Sesamum indicum L.) germplasm collections through seed oil and fatty acid compositions and SSR markers. **Journal of Food Composition and Analysis**, v. 110, n. 104545, p. 1-12, 2022. doi.org/10.1016/j.jfca.2022.104545.

TESFAYE, T.; TESFAYE, K.; KENEN, G.; ZIYOMO, C.; ALEMU, T. Genetic diversity of Sesame (Sesamum indicum L) using high throughput diversity array technology. **Journal of Crop Science and Biotechnology**, v. 25, 359–371, 2022. doi.org/10.1007/s12892-021-00137-x.

ZHOU, L.; WU, Q.; YANG, Y.; LI, Q.; LI, R.; YE, J. Regulation of oil biosynthesis and genetic improvement in plants: advances and prospects. **Genes**, v. 15, n. 9, 2024. <https://doi.org/10.3390/>.

Agradecimentos

A Sebra Agrícola pela disponibilidade das informações dos marcadores e apoio financeiro para a realização da pesquisa e a Universidade Federal de Lavras pelo suporte estrutural para o desenvolvimento do estudo.