# EXPRESSÃO *IN SILICO* DE GENES MADS-BOX EM *Coffea arabica*

**Raphael Ricon de OLIVEIRA[1]; Antonio CHALFUN-JUNIOR[1,2], E-mail: chalfunjunior@ufla.br; Luciano Vilela PAIVA[1,3], E-mail: luciano@ufla.br**

[1]Universidade Federal de Lavras - UFLA, Lavras, MG, Laboratório Central de Biologia Molecular (LCBM); [2]Departamento de Biologia; [3]Departamento de Química.

**Resumo:**

Genes MADS-box constituem uma família de fatores de transcrição que atuam como reguladores-chave em muitas etapas no desenvolvimento de diversos organismos. Em plantas estão envolvidos em processos de identidade do meristema, tempo de florescimento, caracterização do órgão floral, desenvolvimento do fruto, entre outros. Baseado em regiões conservadas das seqüências de genes MADS, foi realizada uma busca por prováveis membros dessa família dentro do banco de dados CAFEST. Após o processo de busca e seleção de *reads* relacionados ao domínio MADS, foi feita a montagem dos EST-contigs, alinhamento entre seqüências relacionadas publicadas, análise filogenética, e análises de motivos de agrupamento, bem como, do perfil de expressão. Desta forma, foi possível identificar 27 seqüências relacionadas ao domínio MADS-box, sendo todas elas classificadas nos subdomínios pertencentes ao tipo II de proteínas, ou seja, com estrutura do tipo MIKC. A maioria delas foi expressa em bibliotecas de tecidos reprodutivos, mas também foram encontradas sendo expressas em tecidos vegetativos e algumas, em ambas. O presente trabalho é um estudo comparativo entre seqüências que surge como uma importante ferramenta para identificação de genes, baseada em uma característica de interesse, para trabalhos na área de manipulação genética.

Palavras-chave: Genes homeoboxes, Bioinformática, Transcriptoma, Regulação gênica.

## *IN SILICO* EXPRESSION OF MADS-BOX GENES IN *Coffea arabica*

**Abstract:**

MADS-box genes constitute a family of transcription factors that act as regulator proteins in many levels in the development of several organisms. In plants they are involved in processes meristem identity, flowering time, characterization of the floral organ, development of the fruit, amongst others. Based in the conserved regions of MADS sequences, it was realized a search for putative MADS genes in the CAFEST database. After searching and selection of reads related to the MADS domain, it was assembled the EST-contigs, alignment between related sequences published, phylogenetic analysis, and clustering, as well as, the expression profile. It was possible to identify 27 sequences related to the MADS-box domain, being all classified into the subdomains type II of proteins, known as MIKC-type. The majority of the genes were expressed in reproductive tissues libraries and also in vegetative tissues, and some of them, in both. The present work is a comparative study between sequences that appear as an important tool for identification of genes, based on the characteristic of interest, helping experiments in the area of genetic manipulation.

Key words: Homeoboxes genes, Bioinformatics, Transcriptome, Gene regulation.

## Introdução

O café é um *commodity* mundial de extrema importância para o Brasil. O agronegócio mundial de café envolve bilhões de pessoas e movimenta uma quantia gigantesca de divisas para os países produtores. A participação brasileira no mercado internacional, em torno de 30%, demonstra a força econômica do produto cultivado em mais de dois mil municípios em 16 Estados do país. Nos últimos anos, a Organização Internacional do Café (OIC) tem divulgado informações demonstrando sucessivas quedas no preço do café devido ao desequilíbrio entre oferta e demanda. Entre os diversos meios para enfrentar a crise e criar um maior equilíbrio entre consumidores e produtores, estão as propostas de melhoria na qualidade, monitoramento da produção e incentivo ao consumo, incorporando o café em novos mercados.

Um dos problemas na produtividade e qualidade do café consiste no florescimento seqüencial dos botões florais e, conseqüentemente, desuniformidade na maturação dos frutos, dificultando a colheita e prejudicando a qualidade dos grãos. O início do processo de florescimento depende da expressão equilibrada de uma rede complexa de genes, que é regulada por fatores endógenos e ambientais. Fatores ambientais vêm sendo bastante estudados, mas pouco compreendidos na criação de um modelo para o florescimento, sugerindo a necessidade de um melhor entendimento dos fatores genéticos.

Com a criação do banco de ESTs (CAFEST) após o seqüenciamento do transcriptoma do cafeeiro (Vieira et al., 2006), torna-se possível a realização de buscas por seqüências formadoras dos prováveis genes relacionados às características de interesse. Dessa forma, estudos comparativos de genômica funcional surgem como ferramentas poderosas para a identificação de genes envolvidos na regulação de vias metabólicas.

Genes MADS-box constituem uma família de fatores de transcrição que atuam como reguladores-chave em muitos processos no desenvolvimento celular de diversos organismos, tais como fungos, plantas e animais. Genes dessa família

são constantemente relatados estando envolvidos em muitos aspectos no desenvolvimento de plantas, tais como: identidade do meristema, tempo de florescimento, caracterização do órgão floral, fertilidade do pólen, desenvolvimento do óvulo, caracterização e desenvolvimento do fruto, e elongação da raiz lateral. Além disso, a existência de genes MADS-box em gimnospermas, samambaias e musgos, os quais não formam flores ou frutos, demonstra que a rota desses genes em plantas não é restrito ao desenvolvimento de órgãos reprodutivos (Münster et al., 2002 b).

MADS são as iniciais dos primeiros quatro membros descritos encontrados nessa família: MINICHROMOSOME MAINTENANCE 1 (MCM1) de levedura, AGAMOUS (AG) de *Arabidopsis*, DEFICIENS (DEF) de *Antirrhinum*, e SERUM RESPONSE FACTOR (SRF) de humano (Yanofsky et al., 1990). Membros dessa família têm regiões altamente conservadas, em torno de 58 aminoácidos, que ativam os processos de transcrição ligando-se a elementos de reconhecimento chamados CarG boxes (CC (A/T)6 GG), encontrados em genes promotores alvo (Riechmann et al., 1996). Alvarez-Buylla et al. (2000a), sugerem que houve uma duplicação do gene MADS ancestral, antes da divergência de plantas e animais, dando origem a dois grupos de proteínas MADS-box, o tipo I (SRF-like) e o tipo II (MEF2-like), ambas encontradas em animais, fungos e plantas. O único denominador comum de todos os genes MADS-box são os altamente conservados 180 pb, codificando o domínio DNA-binding desses fatores de transcrição.

Proteínas MADS do tipo II são mais comumente encontradas em plantas e apresentam organização estrutural modular e conservada, chamada domínio MIKC-type, devido à presença de quatro domínios característicos do terminal N ao C da proteína: o MADS (M) altamente conservado, o Intervening (I) correspondente a uma região interna que se conecta ao domínio Keratin-like (K), por sua vez, responsável por interações proteína-proteína, seguindo-se então uma porção Carboxy-terminal (C), envolvida na ativação da transcrição. Já as proteínas do tipo I não possuem o domínio K bem definido e são subdivididas em 4 subfamílias, Mα, Mβ, Mγ e Mδ. Esta última atualmente aceita como mais próxima ao tipo II de proteínas (Parenicová et al., 2003).

O presente trabalho teve como objetivo identificar genes da família MADS envolvidos em diversos estágios e expressos em diferentes tecidos, com intuito de melhor compreender os processos de desenvolvimento da planta e principalmente dos órgãos reprodutivos. Com isso, proporcionar um estudo preliminar de possíveis genes alvo para melhoramento de cultivares *Coffea arabica*.

## Material e Métodos

### *Busca por seqüências MADS-box, clusterização e anotação*

O banco de dados CAFEST foi investigado utilizando-se critérios de busca por palavra-chave e similaridade com o domínio MADS descrito e depositado no NCBI (*National Center for Biotechnology Information*). Seqüências que atendiam os requisitos de seleção acima descritos, foram depositadas no sistema de gerenciamento e manipulação de seqüências, o GeneProject. Procurando-se esgotar o banco de dados, uma nova etapa de busca foi realizada, na qual as próprias seqüências encontradas serviram de molde para novas procuras. Após clusterização, os EST-contigs formados foram identificados, por meio da ferramenta InterProScan, e anotados, comparando-os contra bancos de proteínas visando-se obter o maior número de informações relevantes sobre os prováveis genes MADS. Dessa forma, foi possível a validação de 27 seqüências com base em suas regiões conservadas, que foram o alvo do estudo de classificação.

### *Análise Filogenética e de Expressão*

Para classificação das seqüências de EST-contigs, foi realizado um alinhamento múltiplo das regiões do domínio MADS-box utilizando-se o programa CLUSTALW, que reuniu seqüências selecionadas do CAFEST e seqüências homólogas publicadas. As seqüências puderam ser visualmente inspecionadas e manualmente corrigidas, sendo removidos segmentos cuja homologia não pode ser acertada. A árvore filogenética foi feita pelo programa MEGA 3.1 (Kumar et al., 2000) e sua validade pôde ser medida pelo teste de bootstraps (Figura 1). Para descobrir motivos de agrupamento entre as seqüências MADS selecionadas no CAFEST (Figura 2), foi usado o programa MEME (Bailey & Elkan, 1994), sendo os domínios funcionais de proteínas MADS-box encontrados, anotados através do software online SMART (Schultz et al., 1998). Para a análise dos locais de expressão, *Northern Eletrônico*, foi construída uma tabela contendo o número de vezes que cada *read* formador de um EST-contig aparecia expresso em cada biblioteca. Esses dados foram normalizados, para dar uma idéia exata do grau de expressão dos prováveis genes em cada tratamento e local da planta, e seus dados foram lançados em uma matriz relacionando genes e bibliotecas. Os EST-contigs e bibliotecas foram agrupados por agrupamento hierárquico utilizando-se os programas Cluster e TreeView (Eisen et al., 1998), sendo os resultados de expressão apresentados em uma escala cinza, onde expressão zero ou negativa representadas por coloração mais clara sendo aumentada gradativamente até atingir o preto, grau máximo de expressão positiva (Figura 3).

## Resultados e Discussão

Dos 27 prováveis genes MADS-box do cafeeiro considerados em nossos estudos, 12 deles encontravam-se expressos em órgãos reprodutivos, 11 em órgãos vegetativos e 4 em ambos (Figura 3). Analisando-se as relações filogenéticas encontradas, todas as nossas seqüências puderam classificadas como genes do tipo II de proteínas MADS-box (Figura 1). A grande maioria delas pertencentes ao grupo MIKC, em suas diferentes subfamílias, e apenas uma (CaMC13) mostrou-se similar ao grupo dos genes Mδ, atualmente aceitos como genes do tipo II (Parenicová et al., 2003). O fato de

genes do tipo I de proteínas MADS não terem sido encontrados, pode ser explicado pelas hipóteses de terem um baixo nível de expressão ou por serem expressos sob condições não monitoradas em projetos transcriptoma (Dias et al., 2005). Genes do tipo I constituem uma grande família de genes amplamente inexplorada. Os ESTs-contigs agruparam-se da seguinte forma dentro das subfamílias de genes MIKC-*type*: AGL2 (5 genes), AGL6 (1 gene), SQUA (2 genes), FLC (1 gene), TM3 (6 genes), AG (2 genes), D+G (4 genes), STMADS (3 genes), AGL17 (2 genes) e Mδ (1 gene). Essas subfamílias são bastante estudadas, tendo muitos genes de diversas plantas publicados e caracterizados, sendo em alguns casos apresentados mutantes com perda de função, permitindo assim inferir características que auxiliem na escolha de um gene para posteriores trabalhos *in vivo*.

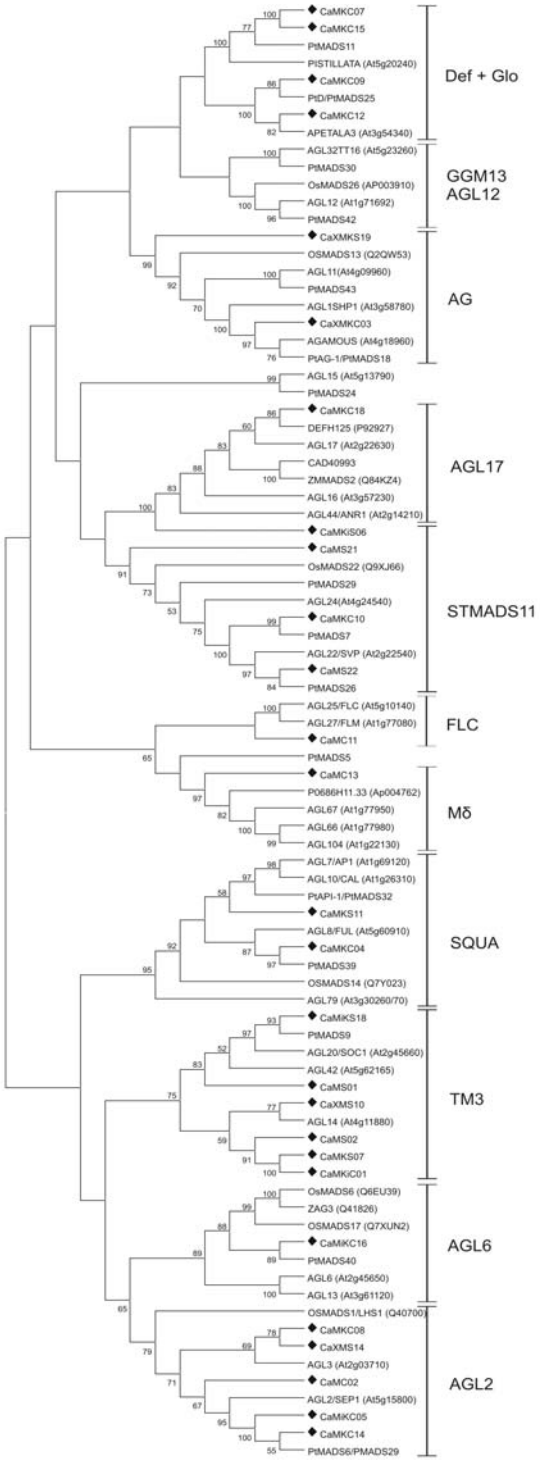


**Figura 1**. Árvore filogenética relacionando seqüências MADS encontradas no CAFEST (♦) e genes MADS MIKC-*type*. Utilizou-se o modelo de comparação *Neighbor-joining* com o método de distância p e *pair-wise deletion*. Valores de *bootstrap* menores que 50% foram omitidos.

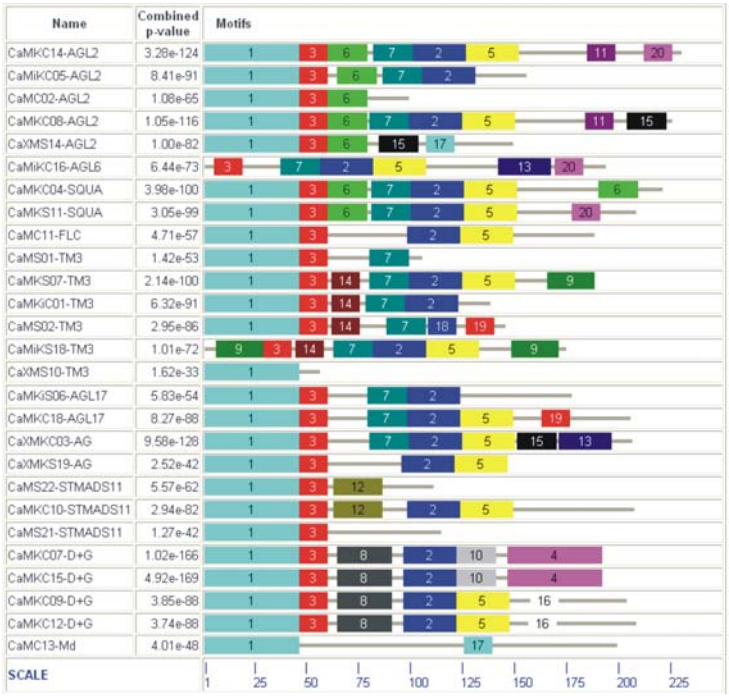


**Figura 2**. MEME (*Multiple Expectation Minimization for Motif Elicitation*, http://meme.sdsc.edu/meme/meme.html). Os parâmetros utilizados foram: número de repetições qualquer, máximo número de motifs 20 e amplitude ótima entre 6 e 200. Os motifs foram anotados pelo SMART (*Simple Motif Architecture Research Tool*, http://smart.embl-heidelberg.de/)**:** Motif 1 - Mads, Motif 3 - Intervening, C-terminal do Motif 3 até C-terminal do Motif 2 - K-box, C-terminal do Motif 2 até final das seqüências - C-domain.

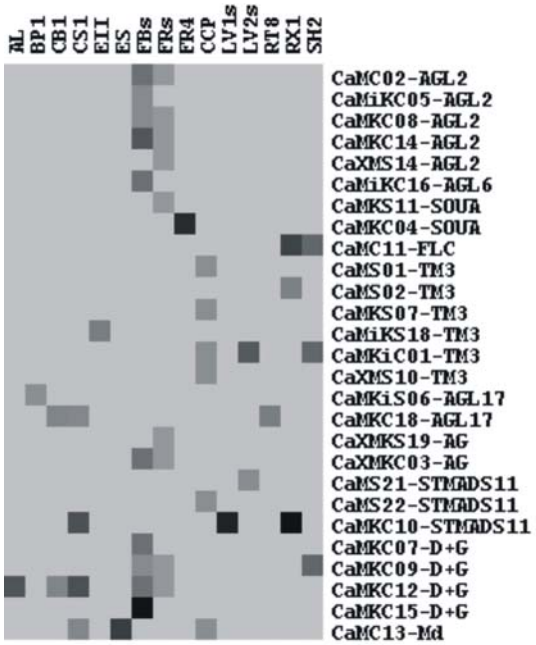


**Figura 3.** Northern Eletrônico representando, por meio de uma escala cinza, os níveis de expressão dos EST-contigs nas diferentes bibliotecas. Quanto mais escuro maior o nível de expressão. Bibliotecas (Vieira et al., 2006): Plântulas e folhas tratadas com ácido araquidônico (AL), Suspensão de células tratadas com acibenzolar-S-methyl (BP1), Suspensão de células tratadas com acibenzolar-S-methyl e brassinoesteróides (CB1), Suspensão de células tratadas com NaCl (CS1), Calo embriogênico (EII), Sementes germinando (ES), Botões florais (FBs), Fruto (FRs), Fruto (*Coffea racemosa*) FR4, Calo não embriogênico com e sem 2,4 D (CCP), Folhas jovens de ramos ortotrópicos (LV1s), Folhas maduras de ramos plagiotrópicos (LV2s), Suspensão de células tratadas estressadas com alumínio (RT8), Talos infectados com *Xylella spp*. (RX1), Plantas estressadas por déficit hídrico (*Pool* de tecidos) (SH2).

**Referências Bibliográficas**


Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS, de Pouplana LR, Martinez-Castilla L, Yanofsk MF (2000a). An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc. Natl. Acad. Sci. USA 97:5328-5333.

Bailey TL and Elkan C (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Proceeding of the Second International Conference on Intelligent Systems for Molecular Biology. (Menlo Park, CA: AAAI Press), pp. 28–36.

Dias BFO, Simões-Araújo JL, Russo CAM, Margis R and Alves-Ferreira M (2005). Unravelling MADS-box gene family in Eucalyptus spp.: A starting point to na understanding of their developmental role in trees. Gen. and Mol. Biol., 28, 3 (suppl):501-510.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95:14863-8.

Kumar S, Tamura K, Jacobsen I and Nei N (2000). MEGA:Molecular Evolutionary Genetics Analysis version 2.0. Pennsylvania and Aryzona State University, University Park, Pennsylvania and Tempe, Aryzona.

Münster T, Faigl W, Saedler H, Theißen G (2002b). Evolutionary aspects of MADS-box genes in the eusporangiate fern *Ophioglossum*. Plant Biol. 4:474-483.

Parenicová L, Folter S, Kieffer M, Horner DS, Favalli C, Busscher J, Cook HE, Ingram RM, Kater MM, Davies B, Angenent GC, Colombo L (2003). Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: New openings to the MADS world. The Plant Cell, Vol. 15, 1538-1551.

Riechmann JL, Krizek BA, Meyerowitz EM (1996). Dimerization specificity of *Arabidopsis* MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA, and AGAMOUS. Proc. Natl. Acad Sci. USA 93:4793-4798.

Schultz J, Milpetz F, Bork P and Ponting CP (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. Proc. Natl. Acad. Sci. USA 95:5857–5864.

Vieira LGE, Andrade AC, Colombo CA et al. (2006). Brazilian coffee genome project: an EST-based genomic resource. Braz. J. Plant Physiol., 18(1):95-108.

Yanofsky MF, Ma H, Bowman JL, Drews GN, Feldmann KA, Meyerowitz EM (1990). The protein encoded by the *Arabidopsis* homeotic gene *AGAMOUS* resembles transcription factors. Nature 346:35-39.