GENES Hsp20 DE SOJA: ORGANIZAÇÃO NO GENOMA E CATEGORIZAÇÃO ENTRE SUBFAMÍLIAS

SOYBEAN Hsp20 GENES: GENOME ORGANIZATION AND SUBFAMILIES CLASSIFICATION

LOPES, V.S.[1-2]; CARVALHO, M.C.C.G. de[1]; DIAS, W.P.[1]; MARCELINO-GUIMARÃES, F.C.[1].

[1] Empresa Brasileira de Pesquisa Agropecuária. Embrapa Soja, Caixa Postal, 231, 86001-970, Londrina, Paraná; e-mail: *valopes@cnpso.embrapa.br*
[2] Universidade Estadual de Londrina;

Resumo

As proteínas de choque térmico (*Heat Shock Proteins* – HSP) constituem um importante mecanismo de resposta, principalmente para as plantas, ao estresse de calor, e recentemente têm sido associadas a outros estresses. Os genes Hsp20 representam a classe mais abundante dentre os HSPs vegetais, mas ainda pouco se conhece sobre esses em soja. Dessa forma, o presente trabalho realizou a caracterização molecular *in silico* dos genes Hsp20 de soja quanto a sua organização no genoma e distribuição em subfamílias. A partir da prospecção de genes Hsp20 anotados no genoma da soja, foram identificados 76 modelos gênicos. As análises de características estruturais, como a presença do ACD (*alpha chystallin domain*) na C-terminal e peso molecular, além de expressão nos diferentes bancos de dados de expressão digital, demonstraram que apenas 45, dos 76 modelos gênicos iniciais, são potenciais GmHsp20 (*Glycine max*-Hsp20). A análise detalhada de filogenia molecular comparando com membros já identificados em *Arabidopsis* e arroz permitiu a categorização das 45 sequências GmHsp20 em 11 subfamílias, distribuídas no citoplasma: CI, CII, CIII, CIV, CV e CIX com 19, 6 2, 1, 2 e 1 membros, respectivamente; ou em organelas: 4 GmHsp20 de retículo endoplasmático, 3 mitocondriais, 5 cloroplasmáticos e 2 peroxíssomais. A organização no genoma da soja dos 45 Hsp20 sugere que esses estejam presentes em 17, dos 20 cromossomos da espécie. As duplicações em tandem decorridas ao longo da evolução da espécie contribuíram para o grande número de genes membros da subfamília CI.

Introdução

Devido a biologia séssil das plantas, nelas há cerca de 4 vezes mais Hsp20 que nos animais. As HSP20 vegetais são codificadas por famílias nucleares multigênicas e são localizadas em diferentes compartimentos celulares. Em *Arabidopsis*, 19 genes para Hsp20 foram categorizados em 12 subfamílias em função da localização celular e homologia (SIDDIQUE et. al., 2008). A principal característica das HSP20 é a presença de uma sequência de aminoácidos, evolutivamente conservada, de 80-100 bases chamada de domínio α - cristalino (α –*crystallin domain* – ACD), localizado na região C-terminal. As HSP20 são chaperonas moleculares independentes de ATP que agem impedindo a ação das desnaturases nas células (SIDDIQUE et. al., 2008; SARKAR et. al., 2009). As Hsp20s são frequentemente associadas à resposta de plantas ao estresse por calor, e recentemente têm sido associadas a outros estresses abióticos, bem como os bióticos (SUN et al., 2002). Contudo, diferente da família de genes Hsp20 de arroz e *Arabidopsis,* ainda pouco se conhece sobre essa família de genes em soja.

Dessa forma, o presente trabalho realizou a caracterização molecular *in silico* dos genes Hsp20 de soja quanto a sua organização no genoma, características físico-químicas e distribuição em subfamílias, e posteriormente, a determinação de padrões de estrutura secundária para cada subfamília.

Material e Métodos

O HMM (Hidden Markov Model – HMM), que representa o domínio das HSP20 (PF00011), foi obtido pelo site Pfam e empregado para a busca de modelos gênicos anotados no genoma da soja como Hsp20. A submissão do perfil HMM utilizando a ferramenta BlastP contra os bancos de dados Phytozome e Superfamily considerou todos os modelos gênicos com e-value ≤ 0.001. Como estratégia adicional de busca, foi utilizada a palavra-chave "Hsp20" para a identificação de possíveis modelos gênicos com a presença do domínio Hsp20 incompleto. A análises de presença do domínio conservado alfa cristalino (alpha crystallin domain –ACD) nas proteínas foi determinada com o software MEME, utilizando como parâmetros para motivo ideal de 80 a 100 aminoácidos (SIDDIQUE et. al., 2008). A sequência protéica dos genes preditos da família GmHsp20 foi utilizada na ferramenta EXPASY PROTPARAM para obtenção das informações sobre peso molecular.

Para as análises seguintes, foram selecionados os modelos gênicos que obedeciam aos critérios: (a) expressão evidenciada em pelo menos um dos bancos de dados de expressão gênica de soja: Genosoja, SoyBase (Glycine max RNA-Seq Atlas) e Genenvestigator; (b) presença de domínio ACD na região C-terminal da sequência peptídica; (c) intervalo de massa molecular, com limite máximo de 45kDa.

Para as predições de peptídeos sinal, domínio transmembrana e localização subcelular foram utilizados os softwares SignalP, WoLF, PSORT, PREDOTAR, TargetP e TMHMM 2.0. A categorização das HSP20 de soja em subfamílias foi possível através do uso da informação da predição de localização celular e relação filogenética entre os candidatos, utilizando a sequência de aminoácidos correspondente aos ACD identificados pelo MEME, para cada proteína GmHSP20 selecionada para a análise. Também foram utilizadas as sequências dos domínios ACD de HSP20 de diversos organismos para os quais já se conhece a distribuição entre as subfamílias. A partir deste alinhamento múltiplo par-a-par de todas as sequências foi construída uma árvore filogenética (software Mega 5.0) não-enraizada pelo método neighbor-joining com bootstrap (1.000 repetições). A estrutura secundária das possíveis GmHSP20 foi predita (http://www.sbg.bio.ic.ac.uk/phyre2). A partir da predição das estruturas secundárias das HSP20, utilizado o software Phyre2, e da análise filogenética para distribuição em subfamílias, foram construídos modelos específicos para cada subfamília das HSP20 de soja.

Para a análise de organização no genoma e duplicação gênica, os genes GmHsp20 foram plotados nos cromossomos através da informação de suas posições cromossômicas, disponível no Phytozome. Para a análise de duplicação gênica, as sequências proteicas das 47 GmHSP20 candidatas foram analisadas pela ferramenta blastp suite-2sequences-NCBI. Para estabelecer as possíveis duplicações dos genes GmHsp20 no genoma da soja, foram utilizadas informações fornecidas no trabalho de Schmutz e colaboradores (2010) a respeito das duplicações ocorridas durante a evolução da soja, e os resultados de identidade e cobertura obtidos pelo blastp 2seq.

Resultados e Discussão

A partir das duas estratégias de busca por genes Hsp20 no genoma da soja, foram identificados 76 modelos gênicos entre os bancos de dados Phytozome e Superfamily. O resultado da análise da presença do ACD demonstrou que todos os candidatos identificados apresentam o domínio conservado de 80 aminoácidos, similar ao domínio Alfa Cristalino, com valor *E-value* geral de 3.3e-1785. A análise de expressão digital para os 76 candidatos em três bancos de dados de expressão de soja possibilitou a investigação dos reais genes Hsp20, com a eliminação de 24 modelos gênicos. Esses candidatos cuja expressão não tenha sido observada em nenhum outro experimento prévio podem representar erros de montagem do genoma da soja ou genes não funcionais.

Após as análises de predição de domínio ACD, identificação do nível de expressão em experimentos prévios disponíveis em BDs e características, como massa molecular, apenas 45 genes dos 76 inicialmente identificados como genes candidatos foram classificados como GmHsp20.

As análises de peso molecular das sequências proteicas de GmHSP20 resultaram em um intervalo de 15,23 KDa (Glyma13g27590) a 28,63 KDa (Glyma02g45810). Essa observação é relevante, considerando que as classes de HSPs são divididas tendo como critério principal o peso molecular, e as pequenas proteínas de choque térmico HSP20 apresentam predominantemente massa molecular de 15-42 kDa, e o limite máximo das Hsp20 de soja a cerca de 30 kDa. (SUN et al., 2002). Foi verificado que entre regiões codantes de mesmo tamanho ou tamanhos muito próximos houve variação no número e tamanho de íntrons. Dentro da família de Hsp20 de soja caracterizada, as análises indicam que 32 (53,84%) do total de GmHsp20 não apresentam íntrons. Este é um percentual médio próximo ao dos genes Hsp20 de arroz preditos (48,72% de genes sem íntrons) (OUYANG et al., 2009). Entre os genes de soja com íntrons, 10 (35,71%) apresentam um único íntron. O padrão global da posição dos íntrons pode ser utilizado como base no estabelecimento das relações filogenéticas em uma família de genes (OUYANG et al., 2009).

Os resultados obtidos nas análises filogenéticas demonstraram haver identidade mínima de 17,39% e máxima de 98,05% entre as GmHSP20. A inclusão de representantes de HSP20 de diversas espécies para todas as subfamílias já descritas possibilitou a diferenciação das GmHSP20 em 11 das 16 subfamílias já descritas. Os genes foram distribuídos entre as subfamílias CI, CII, CIII, CIV, CV, CVI, CVII, CVIII, CIX, CX e CIX com 19, 6, 2, 1, 2, 0, 0, 0, 0, 0 e 1 membros, respectivamente. Os genes das subfamílias organelares foram distribuídos em 4 GmHsp20 de retículo endoplasmáticos, 3 mitocondriais, 5 cloroplasmáticos e 2 peroxíssomais. No geral, a distribuição das GmHSP20 nas subfamílias foi coerente (para MI e ER) ou idêntica (para MII, P e Px) com a classificação de HSP20 previamente reportada para o arroz.

As análises de predição de localização subcelular identificaram 5 proteínas com localização predita para o cloroplasto (P), 4 proteínas com localização predita para o reticulo endoplasmático, 3 mitocondriais (M) e 31 citoplasmáticas (C), não sendo possível detectar diferenças no endereçamento para proteínas citoplasmáticas e peroxissômica devido aos a inexistência de programas de predição prevendo este parâmetro.detectam). Todas as 45 proteínas foram analisadas pelos programas de predição de peptídeos sinal (SignalP). Os resultados foram positivos para todos os genes agrupados nas análises filogenéticas na subfamília ER. Esse resultado era esperado, já que normalmente as proteínas com função em outras organelas apresentam um sinal de endereçamento, o peptídeo sinal, para serem transportadas.

As HSP20 apresentam uma estrutura secundária bem diferenciada entre as subfamílias. Os modelos de estrutura secundária para cada subfamília de GmHSP20, demostram que as subfamílias CI e CII apresentam diferença no número de estruturas folha-β, sendo 7 segmentos folha-β para CI e 6 segmentos folha-β para CII (Figura 1).

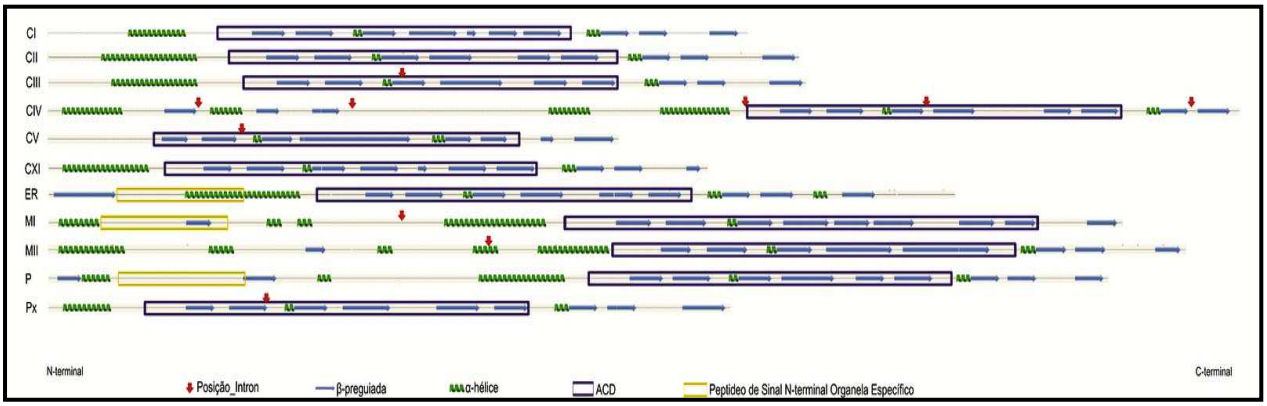

Figura 1: Diagrama ilustrando os padrões de estrutura secundária preditos para as proteínas das subfamílias de GmHSP20. As posições de íntrons estão apontadas por setas em vermelho. As setas horizontais azuis delimitam regiões de formação de estruturas secundárias ß-pregueadas. Espirais em verde são utilizadas para delimitar as regiões de possíveis formações de estruturas secundárias em α-hélices. Os retângulos em azul e amarelo delimitam as regiões preditas para os domínios α-cristalinos e peptídeos sinal, respectivamente.

Esses padrões de estrutura secundária do domínio ACD para CI e CII foram muito similares aos obtidos em *Arabidopsis* por Siddique e colaboradores (2008). Ainda não foram encontrados em *Arabidopsis* genes pertencentes às subfamílias citoplasmáticas CVIII, CIX, CX e CXI, como foram descritas em espécies como *Oryza sativa*, *Zea mays* e outras (SARKAR et al., 2009). Na soja, assim como em Arabidopsis, nenhuma Hsp20 é representante dessas novas subfamílias. A principal característica de estrutura secundária das GmHSP20 da subfamília de retículo endoplasmático é a presença de duas grandes estruturas em folha-β e α-hélice na região N-terminal, na mesma região onde foi predito também um peptídeo sinal.

Os resultados demonstram que as GmHsp20 estão organizadas em 17 dos 20 cromossomos da soja, sendo os cromossomos 7 e 13 os que apresentam maior número de genes. Pautados nos resultados obtidos por SCHMUTZ e colaboradores (2010) e nos resultados das análises de duplicação obtidos neste estudo, possivelmente, a evolução da família de genes Hsp20 no genoma da soja tenha envolvido um total de 21 duplicações gênicas, sendo 5 segmentares em 4 cromossomos.

As duplicações segmentares das GmHsp20 parecem ter contribuído significativamente para o aumento no número de representantes da subclasse CI, que localizam-se nos cromossomos 7, 8, 13 e 14. Em arroz, os membros da subfamília CI encontram-se também distribuídos em aglomerados de duplicações segmentares (OUYANG et al., 2009). Considerando o conceito de parcimônia, essa conservação das proteínas HSP20 duplicadas dentro dos mesmos cromossomos, como observado nos genomas de arroz e da soja, provavelmente tem origem em processos de duplicação segmentar ocorrido na espécie ancestral comum, seguido pelas duplicações cromossômicas dentro das próprias espécies (SCHMUTZ et al., 2010).

Conclusões

- Considerando as características das Hsp20, foram identificados, dentre os 76 genes anotados nos bacos de dados do genoma da soja, 45 GmHsp20, sendo os 31 restantes apenas genes ACD;
- As análises de filogenia, localização subcelular e de estrutura secundária para os 45 GmHSp20 demostraram que os mesmos estão distribuídos em 11 subfamílias, para as quais foi possível estabelecer padrões de estrutura secundária específico;
- Além disso, as análises de duplicação gênica evidenciam que a evolução da família de genes Hsp20 envolveu um total de 21 duplicações gênicas, sendo 5 segmentares em 4 cromossomos.

Referências

OUYANG, Y.; CHEN, J.; XIE, W.; WANG, L.; ZHANG, Q. Comprehensive sequence and expression profile analysis of Hsp20 gene family in rice. **Plant Molecular Biology**. n. 70, p. 341–357, 2009.

SARKAR, N. K.; KIM, Y-K.; GROVER, A. Rice sHsp genes: genomic organization and expression profiling under stress and development. **BMC Genomics.** v. 10, n. 393, p. 1471-2164, 2009.

SCHMUTZ, J.; CANNON, S. B.; SCHLUETER, J. et al. Genome sequence of the palaeopolyploid soybean. **Nature**. v. 463, p. 178-183, 2010.

SIDDIQUE, M.; GERNHARD, S.; VON KOSKULL-DORING, P.; VIERLING, E.; SCHARF, K. D.; The plant sHSP superfamily: five new members in *Arabidopsis thaliana* with unexpected properties. **Cell Stress and Chaperones.** v. 13, n. 2, p. 183-197, 2008.

SUN, W.; MONTAGU, M. V.; VERBRUGGEN, N. Review: small heat shock proteins and stress tolerance in plants**. Biochimica et Biophysica Acta,** v. 1577, p. 1–9, 2002.