UNIVERSIDADE FEDERAL DE PELOTAS
Programa de Pós-Graduação em Biotecnologia
Centro de Desenvolvimento Tecnológico - CDTec

Dissertação

ESTUDO DOS ELEMENTOS *CIS* ASSOCIADOS À RESPOSTA AO ALAGAMENTO

Lara Isys Dias

Pelotas
Estado do Rio Grande do Sul – Brasil
2011

Lara Isys Dias

Dissertação

ESTUDO DOS ELEMENTOS *CIS* ASSOCIADOS À RESPOSTA AO ALAGAMENTO

Dissertação apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Mestre em Ciências (área de conhecimento: Biotecnologia em Fitomelhoramento).

Orientador: Antônio Costa de Oliveira, PhD. – FAEM/UFPel
Co-orientadores: Eliseu Binneck, Dr. – EMBRAPA Soja
Luciano Carlos da Maia, Dr. – FAEM/UFPel

Pelotas
Estado do Rio Grande do Sul – Brasil
2011

Dados de catalogação na fonte:
Ubirajara Buddin Cruz – CRB 10/901
Biblioteca de Ciência & Tecnologia - UFPel

D541e Dias, Lara Isys
Estudo dos elementos *cis* associados à resposta ao alagamento / Lara Isys Dias. – 98f. – Dissertação (Mestrado). Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas. Centro de Desenvolvimento Tecnológico, 2011. – Orientador Antonio Costa de Oliveira ; co-orientador Eliseu Binneck e Luciano Carlos da Maia.

1.Biotecnologia. 2.Elementos cis. 3.Regulação gênica. 4.Genômica comparativa.. 5.Análises *in silico*. I.Oliveira, Antonio Costa de. II.Binneck Eliseu. III.Maia, Luciano Carlos da. IV.Título.

CDD: 575.1

Banca examinadora:

Prof. Dr. Eliseu Binneck	EMBRAPA Soja
Prof.ª Dra. Caroline Marques Castro	EMBRAPA Clima Temperado
Prof.ª Dra. Denise dos Santos Colares de Oliveira	UFPel

Aos meus pais, Silvana e Waldecy.
Pelo amor e confiança.
À minha sobrinha, Sofia.
Pela inspiração.
Dedico

Agradecimentos

Agradeço à Universidade Federal de Pelotas, e ao Programa de Pós-Graduação em Biotecnologia, pela oportunidade.

Ao professor Dr. Antonio Costa de Oliveira, pela confiança, ensinamentos, amizade e oportunidade de trabalhar sob sua orientação.

Ao Dr. Eliseu Binneck e Dr. Luciano Carlos da Maia, pela co-orientação, idéias, apoio, paciência e ensinamentos.

À Coordenação de Aperfeiçoamento Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudos.

As professoras Dra. Caroline Marques castro e Dra. Denise dos Santos Colares de Oliveira, que aceitaram o convite em participar da banca examinadora deste trabalho.

A todos do Centro de Genômica e Fitomelhoramento (CGF), pela demonstração de trabalho em grupo. Em especial à Thaís, Taci e Tati, pela amizade, planilhas e por fazer meus dias no laboratório mais alegres.

A dupla Alex e Egídio, pelo rock e chimarrão, sem os quais tudo teria sido mais difícil. E ao trio Cati, Fran e Duda, pela amizade fiel e de graça.

À Virginia, por ser minha família, em Pelotas, e por fazer com que eu não consiga, aqui, agradecê-la por completo.

As amigas de sempre, Cynara e Simone, as melhores.

À Milena, pela companhia, paciência, apoio, amor e pela chance de eu não ter que fazer mais nenhum *check-in* sozinha.

À Bel, pelas palavras de estímulo e paz, quando mais precisei. E por acreditar em mim.

Ao Celião, pelos conselhos, abraços, carinho e exemplo.

A toda minha família, pelo conforto e suporte. Principalmente aos meus avós, Lila e Américo, e minha irmã, Patrícia, pelo amor constante.

À minha sobrinha Sofia, pelas ligações surpresa.

Aos meus pais, sem os quais nada disso faria sentido.

Agradeço a todas as pessoas que, de alguma forma, contribuíram para a realização deste trabalho.

"Pra onde vão os trens meu pai?
Para Mahal, Tami, para Camiri, espaços no mapa,
e depois o pai ria: também pra lugar algum, meu filho,
tu podes ir e ainda que se mova o trem
tu não te moves de ti."

Hilda Hilst

Resumo

DIAS, Lara Isys. **Estudo dos elementos cis associados à resposta ao alagamento.** 2011. Dissertação (Mestrado) – Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

Os atuais desafios no melhoramento de plantas são maximizar a produtividade das principais espécies cultivadas e criar meios para explorar uma variedade de ambientes de cultivo. Um desses ambientes são os solos hidromórficos de planícies alagadas. A adaptação de outras culturas a este ambiente poderia reduzir a incidência de doenças, pragas e plantas daninhas, se beneficiando de um sistema de rotação de culturas. Quando uma planta é exposta a estresses abióticos ela tem que lidar com as alterações ambientais através de mudanças fisiológicas e anatômicas, as quais necessitam de rápidas mudanças na expressão gênica, ou seja, no seu estado inicial, bem como nas taxas de transcrição. Elementos regulatórios de ação *cis* têm relação direta com fatores de transcrição (FT) em complexas redes de sinalização. Estes sítios de ligação de FTs são elementos funcionais de DNA que influenciam a atividade transcricional de forma temporal e espacial. Neste trabalho, foram investigados possíveis padrões de seqüências que se possam inferir sobre os mecanismos que as plantas utilizam para se desenvolver na condição do alagamento. Essa busca por possíveis homologias entre os vários elementos *cis* permite realizar análises interativas sobre como as plantas utilizam seus mecanismos moleculares para responder a estresses abióticos. Bancos de dados *online* foram utilizados, na busca de genes previamente descritos em literatura e que são expressos em resposta ao alagamento em *Oryza sativa*, *Arabidopsis thaliana* e seus homólogos em *Glycine max* e *Zea mays*. A porção de 1,0 Kb a montante de cada gene foi extraída e analisada *in silico*. Além disso, todos os promotores das quatro espécies foram submetidos a busca por uma variedade de sinais, com a intenção de encontrar novos padrões de motivos de DNA. Esta análise mostrou que, dos 259 elementos *cis* encontrados em promotores de Arabidopsis e arroz, 12 deles são comuns a ambas as espécies e se distinguem do restante por terem alta freqüência. Utilizando o programa MEME dois motivos consenso foram encontrados entre as espécies *Oryza sativa* e *Zea mays*. Estes podem representar novos padrões de elementos *cis*, pois apresentaram ocorrências relativamente elevada nos promotores de genes e são relacionados com seqüências conservadas em monocotiledôneas. A análise aqui apresentada mostra pontos importantes para futuros estudos relacionados ao estresse por alagamento e auxilia na descoberta de mecanismos moleculares de tolerância ao alagamento nas plantas. A partir dos dados gerados, será possível direcionar experimentos de transformação genética a fim de atribuir alguma característica às plantas, tais como aqueles encontrados no arroz, para que possam se desenvolver em um ambiente com privação de O_2.

Palavras-chave: Elementos *cis*. Regulação Gênica. Gênomica comparativa. Análises *in silico*.

Abstract

DIAS, Lara Isys. **Estudo dos elementos cis associados à resposta ao alagamento.** 2011. Dissertação (Mestrado) – Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

The current challenges in plant breeding are to maximize the productivity of major crop species and to create means for exploring novel crop environments. One of these environments is the lowland hydromorphic soils that are proper for the irrigated rice crop. Adapting other crops to this environment could reduce the incidence of diseases, pests and weeds, therefore benefiting from a crop rotation system. When a plant is exposed to abiotic stresses, it has to cope with environmental changes through physiological and anatomic changes that need quick gene expression responses, i.e., changes in active/silenced status as well as in the rates of transcription. *Cis*-acting regulatory elements have straight relationship with transcription factors (TF) in complex signaling networks. This TF binding sites (*cis*-elements) are the functional DNA elements that influence temporal and spatial transcriptional activity. We investigated possible patterns of sequences that can be inferred about the mechanisms that plants use to develop under flooding stress. This search for possible homologies between the various *cis*-elements would lead us to performed interactive analyses about how plants use their molecular mechanisms responding to abiotic stresses. Online databases were searched, looking for genes previously described in literature which are expressed in response to flooding in *Oryza sativa, Arabidopsis thaliana* and their homologous in *Glycine max* and *Zea mays.* The 1.0 Kb upstream portion of each gene was extracted and analyzed *in silico*. Besides, all the promoters of these four species were subjected to a tool for searching for novel signals, intending to find new motif patterns. Our *in silico* analysis shows that from 259 *cis* elements found in PLACE for all promoters of Arabidopsis and rice, 12 of them are common to both species, and are distinguished by having high frequency. Using the MEME program two consensus motifs could be found among the species *Oryza sativa* and *Zea mays*. These could represent new *cis* elements patterns, because they had relatively high occurrences in the gene promoters and they are related to conserved sequences in monocots.
The analysis here presented shows important points for future studies related to the waterlogging stress and unmasking molecular tolerance mechanisms to this typical stress. From the data generated, it will be possible to direct experiments on genetic transformation with target genes and/or *cis* elements in order to attribute some characteristic in plants, such as those found in rice, so they can develop in an environment with O_2 deprivation.

Keywords: *Cis* elements. Gene regulation. Comparative Genomic. *In silico* analysis.

Sumário

Lista de figuras

Lista de tabelas

Sumário

1. INTRODUÇÃO

O cenário atual de estagnação das áreas destinadas ao cultivo agrícola traz um desafio à comunidade científica: maximizar a produtividade das espécies de interesse e, também, explorar ambientes já utilizados pela agricultura, com novas culturas. É com este enfoque que os campos de cultivo de várzea, de característica hidromórfica, se mostram como uma alternativa.

O Estado do Rio Grande do Sul possui, em média, 5,4 milhões de hectares (ha) de solos de várzea apropriados à lavoura irrigada de arroz. Desses, anualmente, 2 milhões de ha são sistematizados com a finalidade de diminuir a incidência de doenças, pragas e plantas daninhas, e 1 milhão de ha são usados no cultivo de arroz (Gomes et al., 2006). A difícil adaptação de outras culturas, que poderiam dar um maior retorno econômico, diminuindo os riscos acima citados ao entrarem em uma rotação de culturas com o arroz nessas áreas induz esforços na busca de alternativas para o desenvolvimento de novas variedades que comportem características genotípicas e/ou fenotípicas, como as do arroz, para tolerar as condições impostas pelo encharcamento e possuam um elevado potencial de rendimento (Pires et al., 2002).

O processo de aumento da eficiência do sistema produtivo requer o entendimento de como o ambiente anóxico, característico de solo com deficiência de oxigênio (O_2), prejudica o desenvolvimento das plantas e como estas respondem ao estresse. A passagem da hipoxia para a anoxia e a produção de toxinas produzidas pelas bactérias anaeróbicas presentes no solo intensifica o estresse experimentado pelas raízes da planta, podendo matar as mesmas. No entanto, algumas respostas fisiológicas iniciais, antes que a anoxia seja alcançada, permitem que as raízes a evitem, assim como a toxidez, se a inundação persistir (Jackson, 1985). De acordo com Kudahettige et al. (2007), espécies que se originam de ambientes semi-aquáticos são capazes de lidar com o estresse por alagamento, como o exemplo bem conhecido do arroz (*Oryza sativa*). Essas plantas podem sobreviver à completa submergência por semanas e algumas ainda possuem a capacidade de crescer vigorosamente e produzir flores e sementes em solos saturados com água.

Ao nível molecular, sinais de estresses abióticos ativam fatores de transcrição e proteínas, que se ligam a regiões adjacentes a genes-alvo, gerando assim uma resposta fenotípica da planta ao ambiente. Segundo Yamaguchi-Shinozaki & Shinozaki (2005), no DNA de algumas plantas, em regiões *upstream* (a montante) próximas ou dentro do promotor, existem seqüências conservadas que parecem estar envolvidas na regulação deste. Estas pequenas seqüências, de aproximadamente 7 pares de bases (pb), que interagem com diversos fatores de transcrição para formar um complexo de iniciação transcricional, são denominadas elementos *cis*. Estão situadas em doferentes posições, dentro de uma região de aproximadamente 1,0 Kb (mil pares de bases) *upstream* aos sítios de iniciação transcricional.

Os elementos regulatórios de ação *cis* estão envolvidos em vários processos de regulação da transcrição, atuando como interruptores moleculares e controlando processos biológicos de resposta a estresses abióticos, hormonais e processos do desenvolvimento em vegetais (Yamaguchi-Shinozaki & Shinozaki, 2005). Os avanços nas pesquisas e na precisão de seus resultados em perfis de expressão do transcriptoma tem levado à identificação de várias combinações de atuação (*cross-talk*) dos elementos *cis*, nas regiões promotoras de genes induzidos por estresses. E também envolvidos com respostas hormonais. Existem dois principais elementos de ação *cis* que funcionam na regulação da expressão gênica em resposta a estresses osmóticos e de temperatura: ABRE (*ABA responsive element*) e DRE (*dehydration responsive element*) – ABA-dependente e ABA-independente, respectivamente.

O rápido progresso nos projetos de seqüenciamento de genomas de plantas tem produzido muitos resultados de seqüências nucleotídicas e vai continuar produzindo um grande montante de dados nos próximos anos. Porém, muitas destas seqüências geradas não oferecem muita informação, como por exemplo, a função do gene em que elas se encontram. É necessário, portanto, que informações biológicas sejam associadas a estas seqüências por análise computacional (*in silico*), principalmente. Dentre as estratégias utilizadas e bem sucedidas, o acesso a bancos de dados *online*, que oferecem ferramentas de buscas e alinhamentos de sequências de DNA e de proteínas, ajuda na identificação, caracterização e análise das funções biológicas que se deseja estudar.

Neste contexto, o PLACE é o principal banco de dados baseado em informações sobre os elementos regulatórios de ação *cis* para estudos em plantas vasculares. O PLACE consiste de uma ferramenta para buscas por homologias entre seqüências de DNA. É uma ferramenta muito útil para estimar o modo de regulação do gene, as regiões envolvidas em tal regulação, entre outras regiões pertinentes, da sequência submetida (Higo et al., 1998).

Outra abordagem, bastante interessante, é a tentativa de descoberta de novos padrões de ocorrência de elementos regulatórios *cis*, uma vez que o PLACE teve sua última atualização em 2007, o que significa que há um atraso de quatro anos na atualização dos dados. Dessa forma, a suíte de aplicativos MEME, oferece diversas ferramentas de alinhamento de sequências para a descoberta de padrões de motivos repetidos entre elas (Bailey et al., 2009). A partir do alinhamento de regiões promotoras de genes co-regulados e com características em comum, espera-se que alguma inferência possa ser feita sobre a regulação transcricional desses genes. Por exemplo, podem ser identificados novos elementos *cis* ainda não descritos, através da detecção de motivos conservados nos promotores.

Assim a área da Bioinformática ganha, a cada momento, espaço na pesquisa do melhoramento genético de plantas e se faz indispensável para a predição de funções biológicas, bem como dados de metabolismo e *cross-talk* (“conversa cruzada”) entre fatores de transcrição e suas respostas ao ambiente. Tudo isso com dados estatísticos refinados, permitindo-nos fazer extrapolações para os experimentos que possam vir a confirmar as análises computacionais.

2. REVISÃO DE LITERATURA

Esta revisão de literatura foi realizada com a finalidade de conceituar e explorar os itens deste trabalho, caracterizando e situando o tema central. São discutidos temas como a regulação gênica de genes que respondem à anoxia, a caracterização das espécies em estudo e a Bioinformática, área da Biotecnologia, nos estudos *in silico* dos elementos *cis*.

2.1 – Solos Hidromórficos

A característica comum aos solos hidromórficos é a deficiência de drenagem, sendo encontrados nas planícies de rios, lagoas e lagunas. Estes solos de várzea, típicos de regiões de baixas altitudes (0-200 m), ocupam, aproximadamente, 5.400.000 ha do estado do Rio Grande do Sul (20% da área) e chegam a abranger 650.000 ha em Santa Catarina (GOMES et al., 2004).

Nestes solos, desenvolvidos, primariamente, de rochas sedimentares sílticas ou argilosas, a deficiência de drenagem natural se intensifica devido a um perfil de camadas superficial pouco profunda e subsuperficial mais impermeável. Estas são condições favoráveis ao cultivo irrigado e proporcionam baixa susceptibilidade à erosão.

De acordo com Gomes et al. (2006), o sistema produtivo tradicional do Rio Grande do Sul, que utiliza estas áreas alagadas com a prática do cultivo de arroz irrigado e da pecuária de corte, é associado a longos períodos de repouso (2.000.000 ha são sistematizados, anualmente), visando diminuir a incidência de pragas e doenças, a infestação de plantas daninhas (arroz vermelho, principalmente) e promove o descanso do solo para reciclagem de nutrientes; porém, leva ao problema de baixa rentabilidade. Para viabilizar alternativas de produção para exploração mais rentável e sustentável das várzeas arrozeiras, trabalhos vem sendo realizados, para ampliar as possibilidades de rotação de culturas (BARNI et al., 1985; GASTAL et al., 1998).

No que se refere à visão da genômica comparativa, é de interesse científico extrapolar dados obtidos em estudos com espécies modelo para outras espécies próximas (evolutivamente relacionadas), que possuam similaridades entre seus genomas. Nesse contexto, *Arabidopsis thaliana* e *Oryza sativa* se

destacam por serem plantas com o genoma totalmente seqüenciado e terem impacto na importância agronômica (SPANNAGL et al., 2010). *Arabidopsis* enfatizando resultados para as dicotiledôneas e o arroz para as monocotiledôneas.

2.2 – Anoxia

As transformações físicas e bioquímicas que o solo sofre com o alagamento, requerem uma rápida adaptação de todo o sistema envolvido, desde os microorganismos presentes até as plantas que são submetidas à esse ambiente. O arroz é, nesse contexto, uma planta modelo a ser estudada, pois tolera a privação de oxigênio proporcionada pela anoxia do alagamento, e se adapta a essas condições adversas, desenvolvendo-se sem dificuldades. A exposição dos processos bioquímicos dos solos de várzea facilita e guia o entendimento de como a planta tolera este estresse abiótico, a partir de mecanismos fisiológicos e moleculares.

O suprimento de O_2 para o solo passa a ser extremamente lento com o início do alagamento, com uma difusão até 10 mil vezes mais lenta na água (GOMES et al., 2004). De acordo com Greenwood (1961), os microorganismos aeróbicos consomem rapidamente este oxigênio presente e restante, tornando-se inativos ou morrendo com a posterior falta de O_2. Depois disso, ocorre a proliferação de uma microbiota, na sua maioria, anaeróbica facultativa e obrigatória, às custas da matéria orgânica presente (YOSHIDA, 1975).

A atividade anaeróbica destes microorganismos faz com que o acúmulo de ácidos orgânicos, como o CO_2, decorrentes da decomposição da matéria orgânica do solo e do processo de fermentação realizados pelos mesmos, aumente gradativamente (STEVENSON et al., 1967), sendo a composição e concentração destes ácidos importantes, principalmente, pelos seus efeitos sobre a cultura de arroz (YOSHIDA, 1975).

O acúmulo de CO_2 no solo alagado é intensificado pela difusão lenta de gases no meio aquoso para a atmosfera. Apenas uma fração desse acúmulo consegue escapar para a atmosfera, quando a pressão desse gás, no solo, for alta o suficiente para fazer com que borbulhe na superfície, livrando-se da água (STOLZY, 1974). Esta concentração pode atingir valores de 3 ton. ha^{-1} nas primeiras semanas de alagamento (PONNAMPERUMA, 1972). A redução do

CO_2 à CH_4 ocorre logo após este acúmulo do CO_2, e a diminuição da concentração deste gás traz um equilíbrio ao meio (YOSHIDA, 1975). As concentrações de CO_2 normalmente encontradas em solos de cultivo de arroz não alcançam níveis tóxicos (PONNAMPERUMA, 1965).

A partir dos processos de respiração e fermentação dos microorganismos anaeróbicos, ocorre a oxidação da matéria orgânica (C orgânico), pois continua a liberação de elétrons (H^+), com o início do alagamento, e os receptores de elétrons passam a ser outros compostos inorgânicos, como: nitrato, sulfato, acetato, óxido ferroso e glucose (LINDSAY, 1979). Acredita-se, segundo Rowell (1981), que o principal mecanismo de oxidação do C seja acoplado à redução dos compostos inorgânicos, quando estes são absorvidos pela célula, sendo o C reduzido e, por conseguinte, é excretada sua forma reduzida.

Moraes & Freire (1974) relatam um forte decréscimo do pH do solo, nos primeiros dias de alagamento, mas perde essa acidez e se estabiliza, em um período de 30 dias, aproximadamente. O acúmulo de CO_2 é o principal responsável por esta característica ácida inicial (PONNAMPERUMA, 1965), mas posteriormente, com a redução do solo, os valores do pH permanecem ao redor de 6,0 – 6,5, complementando-se ao efeito contrário do CO_2, e entrando em equilíbrio. Em 1972, Ponnamperuma conseguiu elevar o pH de um solo reduzido, de 6,7 para 8,5, eliminando o CO_2 desse solo.

Com o alcance da estabilidade do processo de redução do solo, a atividade dos microorganismos diminui e o suprimento de O_2 atmosférico, obtido por difusão e aquele liberado pela fotossíntese de algas e plantas aquáticas, excedem a quantidade de O_2 consumido pelos microorganismos da superfície da água (YOSHIDA, 1975). Forma-se, a partir disso, uma fina camada superficial oxidada, de aproximadamente 1 cm, que comporta metabolismo microbiológico aeróbico. O gradiente de concentração entre esta camada e a subjacente, reduzida, impulsiona trocas de substâncias por difusão, pois na camada oxidada (mais superficial e em contato com a atmosfera) a concentração de substâncias reduzidas é muito baixa, e a de substâncias oxidadas é alta, em contraponto à camada reduzida (GOMES et al., 2004).

As raízes do arroz tem a capacidade peculiar de oxidar a sua rizosfera, formando, assim, uma camada com propriedades aeróbicas, permitindo a respiração das mesmas (PONNAMPERUMA, 1972). Essa característica

adaptativa dessa cultura de solos alagados é conseguida pelo transporte do O_2 atmosférico, através das folhas até as raízes, pelos aerênquimas, que se constituem de um espaço físico intercelular para a movimentação dos gases. Isso não é restrito à planta de arroz, sendo presente, também, em outras gramíneas e observado, na sua maioria, em condições de solos alagados, como analisado por Pradhan et al. (1973), que encontrou maior porosidade em raízes de arroz cultivado em condições alagadas do que em drenadas.

Esta rizosfera oxidada, imposta em um ambiente reduzido, permite à planta realizar reações inversas ao que ocorre no corpo do solo a que está inserida, assumindo um papel importante na eliminação ou redução de substâncias tóxicas ali produzidas, e que se dirigem à raiz. Exemplos disso são o sulfeto de hidrogênio (H_2S) e o excesso de ferro (Fe) reduzido (GOMES et al., 2004), que tem seus efeitos tóxicos atenuados.

2.3 – Caracterização das espécies

2.3.1 – Arroz

O arroz é uma angiosperma, da classe das monocotiledôneas, ordem Poales, família das gramíneas (Poaceae) e subfamília das *Bambusoideae* ou *Oryzoideae*. Esta família se originou, provavelmente, na era Mesozóica e evidências apontam que seja de clima tropical e que suas linhagens tenham evoluído e se adaptado a vários outros habitats. Com 12 cromossomos (n = 12) de tamanho médio a pequeno, característica considerada primitiva pelos citologistas, a tribo *Oryzeae* provavelmente teve seu genoma derivado por duplicação cromossômica de gramíneas ancestrais com n= 6 (NCBI).

Dentro desta tribo existem duas espécies cultivadas, ambas com um número básico (n) de 12 cromossomos. São elas: *Oryza glaberrima* Steud. e *Oryza sativa* L., africano e asiático, respectivamente. Apesar das semelhanças e possibilidade de cruzamento entre as espécies (alogamia), há diferenças entre os genomas.

Dados históricos indicam que o arroz asiático (*Oryza sativa* subespécie japônica) tenha sua origem no sul da Índia, onde houve condições mais favoráveis para seu cultivo. Expandiu-se para o sudeste asiático, atingindo a China e, daí, foi introduzido na Coréia e Japão. Para o lado ocidental da Ásia,

sua implantação pela Turquia e Pérsia, chegando à Grécia, Irã e Babilônia, se deram por volta do ano 320 a.C., com as invasões de Alexandre Magno (Embrapa Arroz e Feijão, 2010). A expansão árabe chegou ao Marrocos e Espanha, a qual introduziu o arroz na América. A implementação da cultura no Brasil, foi feita através dos portugueses, tornando-se um dos principais alimentos de consumo interno.

Ainda, de acordo com Gomes et al. (2004), o arroz africano (*Oryza glaberrima*) com origem na África Ocidental, teve sua exploração e consumo restritos à sua área de cultivo, mais precisamente no Delta Central do Niger. A introdução do arroz asiático, pelos portugueses e holandeses, na costa da África Ocidental, mostrou uma melhor adaptação deste e cariopse branca, sendo esta vermelha no arroz africano, de maneira geral.

A classificação genética do gênero *Oryza* dificulta o estabelecimento preciso de um agrupamento de espécies, pois os estudos relatam a existência de cinco genomas distintos, em nível diplóide (AA, BB, CC, EE e FF) e dois genomas anfidiplóides (BBCC e CCDD). Percebe-se que o genoma D, não é representado na condição diplóide, o que indica que esta forma possa estar extinta. As espécies *Oryza glaberrima* e *Oryza sativa* são constituídas pelo genoma A, mas devido a problemas de pareamento cromossômicos em seus híbridos, são representados como A^gA^g (*Oryza glaberrima*) e AA (*Oryza sativa*) (NCBI Map Viewer – *Oryza sativa*).

Hoje, o arroz asiático é cultivado em todos os continentes, sendo que a maioria da sua produção (cerca de 95%) é sempre destinada ao consumo local, restando pouco para o comércio internacional. Em uma classificação de produção mundial, a Ásia é representada por 9 países entre os 10 maiores produtores.

O Brasil encontra-se em nono lugar nesta classificação (dados de 2008) sendo o único representante não asiático, dentre os dez maiores produtores mundiais (Figura 1). Cerca de 68% da produção no Brasil vem da Região Sul do país, com concentração nos solos de várzea do estado do Rio Grande do Sul (Figura 2).

Figura 1. Distribuição da produção, em milhões de toneladas, nos principais países produtores de arroz, CGF/UFPel, 2011. Fonte: FAO – FAOSTAT *Database* (2010).

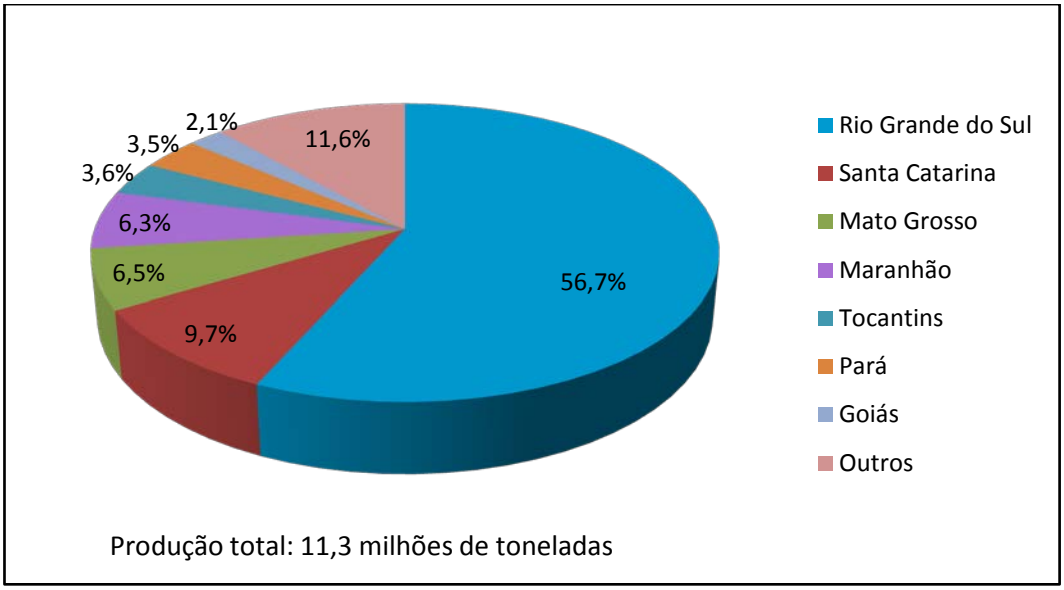


Figura 2. Produção de arroz no Brasil (porcentagem sobre a quantidade) - Safra 2006-07, CGF/UFPel, 2011. Fonte: Conab.

Apesar da alta produção, o país precisou, por sucessivos anos, importar uma quantidade aproximada de 900 mil toneladas/ano de arroz, sendo que no ano de 2007 esse número teve uma queda para 720 mil toneladas importadas principalmente do Uruguai e Argentina que, juntos, somaram 89,5% do total importado, o equivalente a US$235,6 milhões.

2.3.2 – Milho

Assim como o arroz, o milho é uma angiosperma, monocotiledônea, da família das gramíneas (Poaceae), diferindo-se do arroz no táxon da subfamília, sendo representado pela Panicoideae e tribo Maydea. *Zea mays* L.. Parece ter se originado no Golfo do México, tornando-se sustento das civilizações Olmecas, Maias, Astecas e Incas. Há datação de que o milho já era cultivado na América há 4.000 anos (PERRY et al., 2005). Está, hoje, entre os três grãos mais produzidos, ficando atrás do trigo e do arroz.

A origem da conformação genômica do milho (n = 10) parece ter se dado por anfiploidia (fusão de dois genomas diplóides) (HELENTJARIS et al., 1988), ou por alopoliploidia com heterocromatina super enovelada (WILSON et al. 1999), que podem constituir nós na cromatina e interferir no processo de transcrição. As diferenças nos tamanhos das duplicações de elementos, que formam essa heterocromatina característica do milho, levam a uma ampla gama de tamanhos de genomas para esta espécie (NCBI Map Viewer – *Zea mays*). Os seis mapas genéticos existentes estão presentes no MaizeDB (*Maize Genetics and Genomics Database*), um banco de dados que trabalha com grupos de estudos em *Zea mays*, envolvidos na anotação estrutural e funcional do genoma desta espécie.

No que se refere ao panorama econômico mundial, o Brasil ocupa a terceira colocação em um ranking de produção deste grão. Este posto é sustentado por uma característica única do Brasil, em relação a outros países produtores de milho, que é a possibilidade de plantio de duas safras em um ano. Uma safra principal de verão (setembro a novembro) e a segunda (janeiro a abril), chamada safrinha, que é de menor produtividade, por ser plantada no final da época de chuva. Contudo, a safrinha participa significativamente no aumento da produção anual significativamente, com cerca de 28% (COSTA et al., 2008).

A política americana de incentivo à produção de etanol, a partir do grão de milho como matéria prima, impulsionou o mercado da produção desse grão. Essa tendência de utilizar o milho para fins energéticos leva à redução dos estoques internacionais (COSTA et al., 2008), ressaltando a importância da colocação do Brasil como um dos três maiores produtores mundiais e a Argentina, como outro país sul-americano, entre os cinco primeiros (Figura 3).

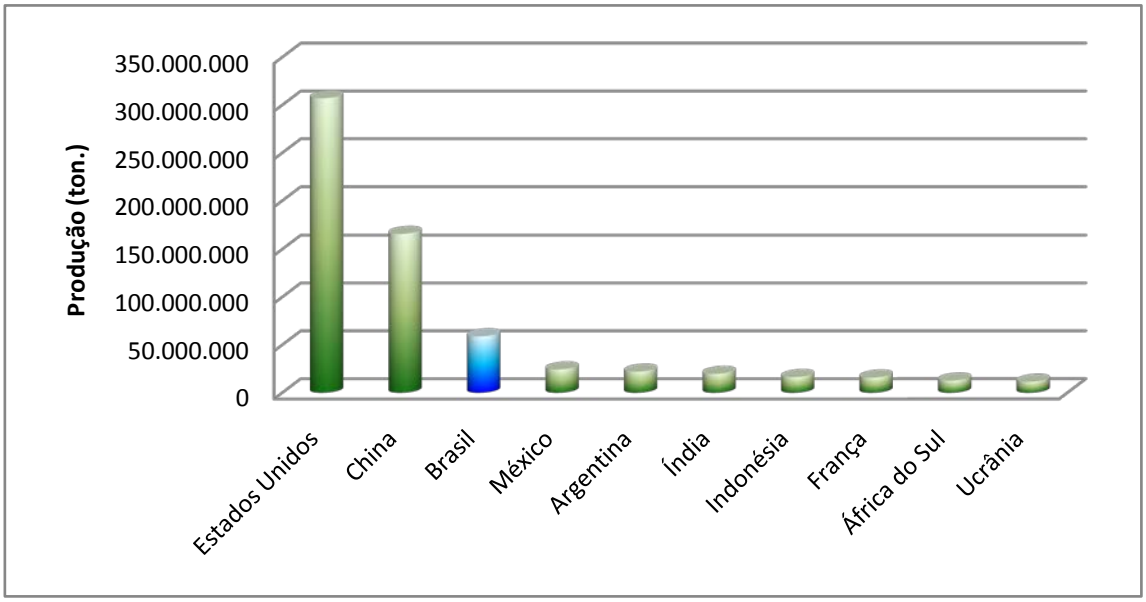


Figura 3. Distribuição da produção, em milhões de toneladas, por país produtor de milho, CGF/UFPel, 2011. Fonte: FAO – FAOSTAT *Database* (2010).

Os maiores importadores de milho do Brasil são países da União Européia, Irã e Coréia do Sul, com uma parcela de 90% das exportações (Figura 4). Tanto a quebra da safra 2006-07 do bloco da União Européia, quanto a demanda de milho 100% livre de Organismos Geneticamente Modificados (OGM) disponível no Brasil, aumento a participação deste bloco de países como destino do produto brasileiro em 2007. Em 2007 a Comissão Técnica Nacional de Biossegurança (CTNBio) aprovou o plantio e a comercialização de uma cultivar de milho transgênico resistente a insetos e uma outra resistente a um herbicida utilizado no controle de plantas invasoras. Esta decisão foi ratificada em 2008, pelo Conselho Nacional de Biossegurança (CNBS).

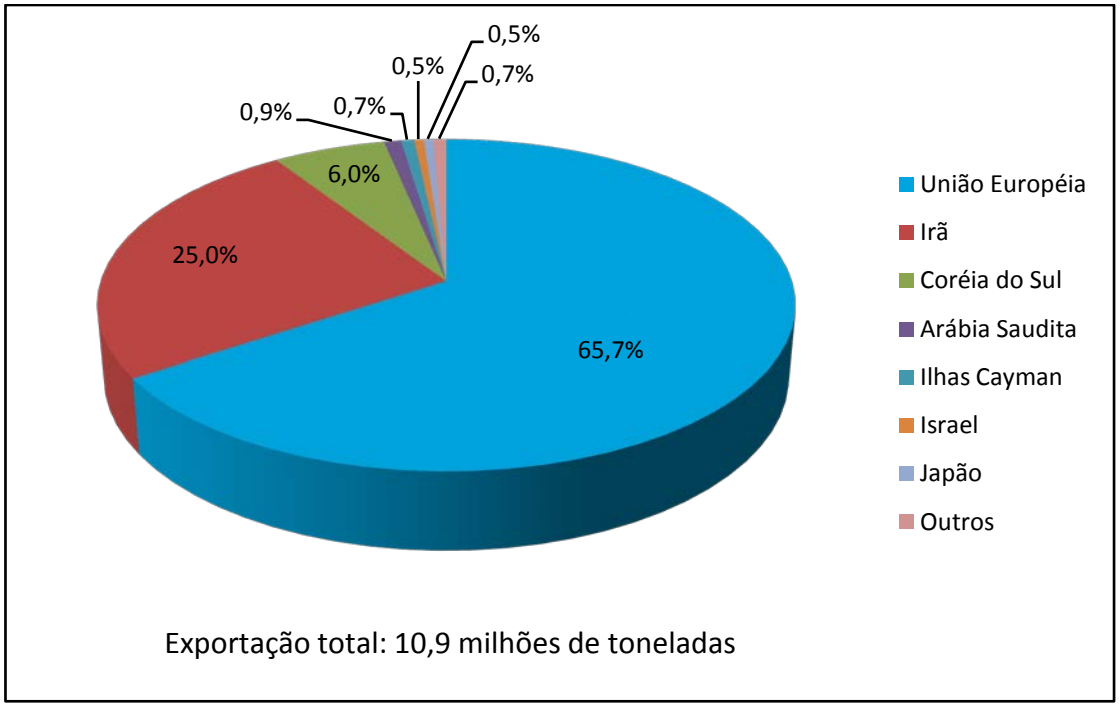


Figura 4. Exportações de milho no Brasil, 2007, CGF/UFPel, 2011. Fonte: USDA.

No Brasil, a Região Sul se destaca como principal produtora, seguida do Sudeste. Nesse contexto o estado do Rio Grande do Sul é o terceiro maior produtor. A produção nacional de milho é voltada, principalmente, para o segmento de alimentação animal.

2.3.3 – Soja

A soja é uma angiosperma, dicotiledônea, da ordem Fabales e pertencente à família Fabaceae (fabáceas ou leguminosas), assim como o feijão, a lentilha e a ervilha. É da subfamília Faboideae e gênero *Glycine*. A espécie cultivada (*Glycine max* (L.) Merr.), descendente do cruzamento de duas espécies selvagens, supostamente rasteiras, tem sua origem na costa leste da Ásia e seu cultivo e produção foram restritos à China até o ano de 1894. Após o final da guerra entre China e Japão, se disseminou nos jardins botânicos da Europa como uma curiosidade. No século XX, com a descoberta dos teores de óleo e proteína do grão, as indústrias mundiais despertaram o interesse para a cultura comercial da soja. A sua disseminação no Brasil ocorreu a partir da imigração japonesa (Embrapa Soja, 2010).

Glycine max possui um número básico de cromossomos igual a 20 e aparece como um modelo de estudo dentre os grãos cultivados na família das

leguminosas. Dois fatos impulsionaram os esforços para o projeto de seqüenciamento desse genoma: primeiramente pelo fato do genoma conter entre 1.000 e 2.000 Mpb (milhões de pares de bases) e, segundo, por ser considerado como sendo um tetraplóide diploidizado. Este é fato sustentado pela Citogenética, que mostrou a soja como única dentre a sua família (Fabaceae), onde todos os outros representantes tem um número haplóide de cromossomos igual a 11 (LACKY, 1980). Opiniões correntes acreditam que *G. max* tenha se originado de uma espécie ancestral leguminosa (n = 11), que passou por uma aneuploidia, perdendo um cromossomo (n = 10), seguido de uma duplicação (n = 20) (PALMER & KILLER, 1987).

O alto valor nutritivo da soja e teor de óleo impulsionam o mercado da produção mundial, no qual o Brasil se destaca como o segundo produtor mundial (Figura 5) e reveza com os Estados Unidos a liderança nas exportações deste produto, nos últimos anos (COSTA et al., 2008).

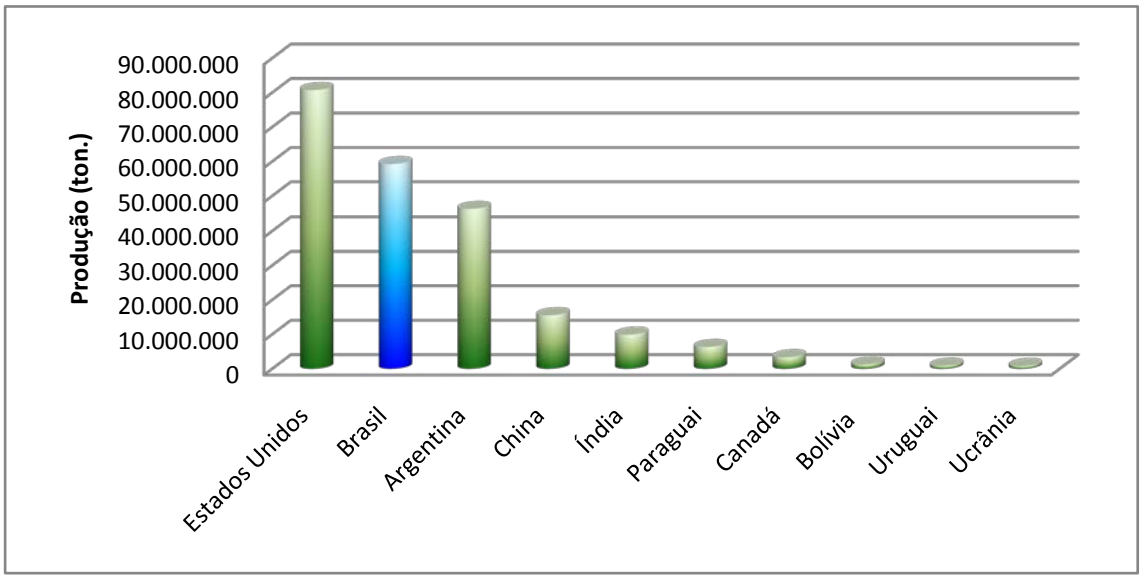


Figura 5. Distribuição da produção, em milhões de toneladas, por país produtor de soja, CGF/UFPel, 2011. Fonte: FAO – FAOSTAT *Database* (2010).

Em território nacional a Região Sul, que possuía 90% da produção na década de 70, foi perdendo espaço para a Região Centro-Oeste que, hoje, colabora com cerca de 45,4% (Figura 6). Contudo, no que se refere ao plantio de soja transgênica, a produção se concentra nos estados do sul do país (Figura 7).

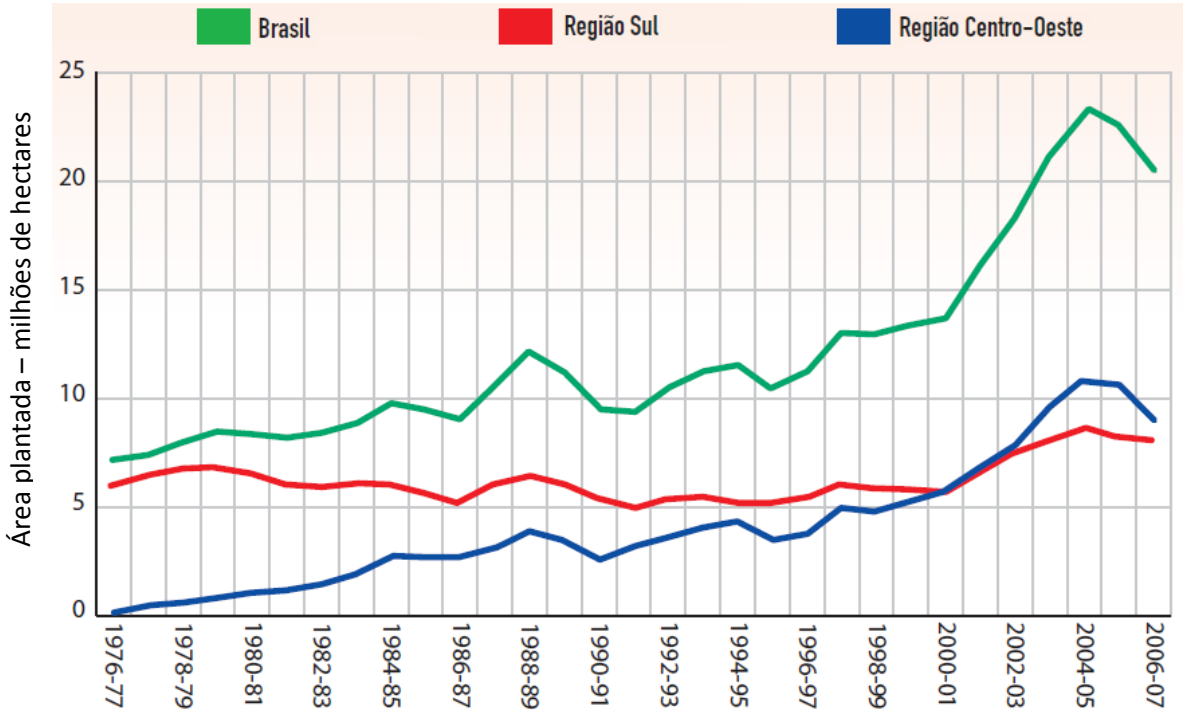


Figura 6. Área plantada com a cultura de soja no Brasil. Safra 2006-07, CGF/UFPel, 2011. Fonte: Conab.

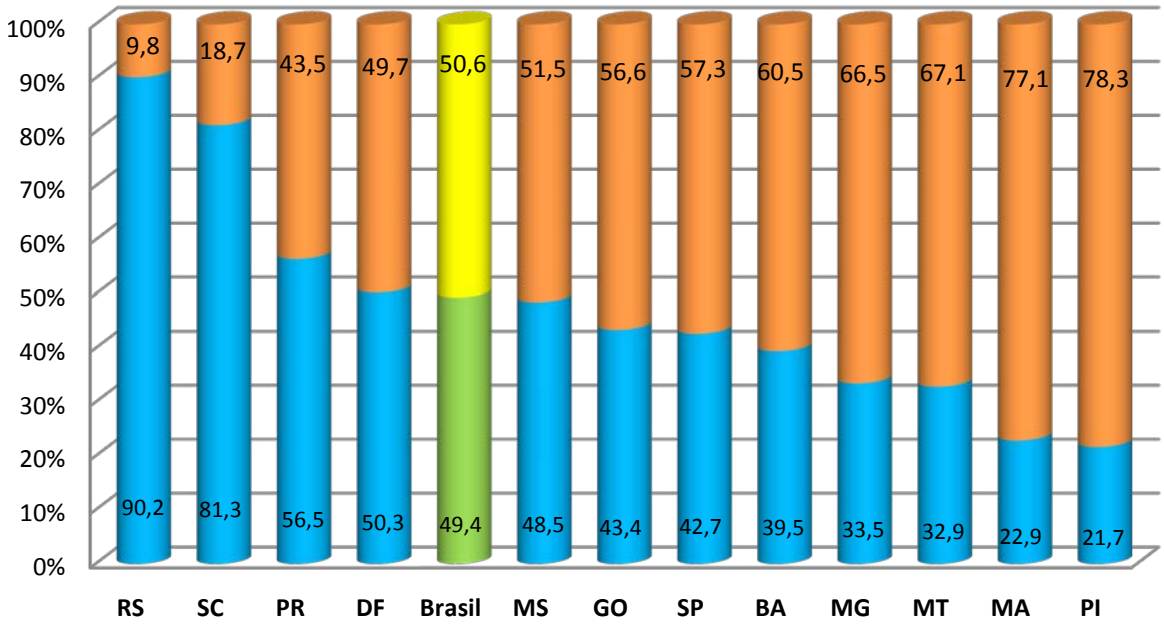
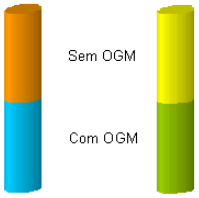


Figura 7. Participação de OGM's (Organismos Geneticamente Modificados) na produção nacional, CGF/UFPel, 2011. Fonte: Céleres.



2.3.4 – Arabidopsis

O modelo de estudos da Botânica para as plantas superiores, *Arabidopsis thaliana* (L.) Heynh., é uma angiosperma, dicotiledônea, da família das Brassicaceae (brassicáceas), assim como a mostarda. Devido à seus amplos recursos genômicos e coleções de mutantes, se tornou uma excelente ferramenta de comparação nas genômicas comparativa e funcional (LU & LAST, 2008).

Esta planta de pequeno porte foi escolhida como modelo em plantas para seqüenciamento pelo fato de ter um genoma pequeno e com estrutura simples, tendo poucas sequências repetidas e alta densidade de genes. Além disso, possui gerações curtas de seis semanas, produz um grande número de sementes e é relativamente fácil sua transformação genética (The *Arabidopsis thaliana* Initiative, 2000), via *Agrobacterium tumefaciens* ou por biobalística (*shotgun*). O sequenciamento foi realizado com um consórcio internacional nomeado *The Arabidopsis thaliana Initiative* (AGI), que consistiu de grupos de pesquisadores nos Estados Unidos, Europa e Japão. O projeto teve início em 1996 e término em 2000.

Dada a vasta gama de dados disponíveis para Arabidopsis para processos biológicos em plantas e a grande importância das espécies produtoras de grãos, torna-se interessante a estratégia da aplicação do conhecimento biotecnológico adquirido em *Arabidopsis thaliana* nessas espécies cultivadas, no intuito de conhecer e, inclusive, melhorar as mesmas (SPANNAGL et al., 2010). Espécies com importância particular para a agricultura são atrativas para programas de transferência de genes que possam conferir resistência a estresses bióticos e abióticos, por exemplo.

Contudo, transferências genéticas são mais efetivas entre genes ortólogos, os quais compartilham, por definição, um ancestral funcional em comum. Há um desbalanço genético entre as espécies, que é causado, na maioria dos casos no genoma de Arabidopsis, por expansões específicas. Exemplo disto são duplicações em *tandem* para esta espécie, que contribuem de forma não proporcional para um aumento no número de cópias de alguns genes (SPANNAGL et al., 2010). Isso tudo leva a um impacto, quando se deseja fazer a transgenia. A sintenia esparsa entre plantas monocotiledôneas e dicotiledôneas se deve a freqüentes rearranjos, translocações, perda de genes

fortemente prejudiciais e redução do número de genes ortólogos que possam ser detectados por regiões conservadas.

2.4 – Genômica Comparativa

A genômica comparativa oferece a base para o entendimento da evolução dos genomas (KELLER et al., 2000), partindo do princípio de que se a organização e a ordem dos genes se mantêm conservadas entre espécies, um genoma menor de referência pode ser usado como modelo, para o isolamento de genes de genomas maiores. Surge, então, o termo sintenia, que é utilizado para referir a conservação de regiões de DNA entre espécies, e essa conservação de regiões e a ordem das mesmas leva à comparação da organização dos genomas de diferentes tamanhos e ao estudo da evolução dos mesmos.

A Paleogenômica, estudo da estrutura de um genoma ancestral, permite a identificação e caracterização de mecanismos (por exemplo, duplicações, translocações e inversões) que tem moldado os genomas das espécies durante suas evoluções. A Paleogenômica pode ser realizada através de análises comparativas em larga escala de espécies atuais e modelagem da estrutura de um genoma ancestral. Diferentes métodos tem sido desenvolvidos para computar um grande número de dados genômicos para a reconstrução de genomas ancestrais. O primeiro passo consiste na identificação de regiões conservadas entre os genomas (SALSE et al., 2009).

Experimentos com fracionamento de DNA e localização gênica sugerem que os genomas das plantas sejam organizados em longos agrupamentos de genes e elementos transponíveis (formando, juntos, o espaço gênico), ocupando cerca de 12-24% do genoma e separados por longos trechos de regiões de “vazio gênico”, que consistem principalmente de sequências repetitivas (BARAKAT et al., 1998). A diferença no tamanho de genomas entre espécies com genomas pequeno e grande se deve, principalmente, às diferenças entre os tamanhos das regiões de vazio gênico.

Em estudos de microcolinearidade (capazes de detectar microrrearranjos, como deleções, inversões e duplicações) no *locus Adh1* entre milho e sorgo, Tikhonov et at. (1999), mostraram que dois genes adjacentes em sorgo também foram encontrados, um próximo ao outro, em Arabidopsis. No

entanto, essa conservação não se estende à genes vizinhos, sugerindo que a colinearidade (conservação de um arranjo linear, no cromossomo) entre dois genes pode ser encontrada entre monocotiledôneas e dicotiledôneas, mas esse tipo de conservação é raro. Apesar disso, a conservação de cinco genes em regiões contíguas em Arabidopsis e arroz, foram relatadas por Van Dodeweerd et al. (1999). Porém, esses genes encontrados são interespaçados por 19 genes não ortólogos em *Arabidopsis thaliana*, o que demonstra que rearranjos devem ter ocorrido na divergência entre monocotiledôneas e dicotiledôneas.

Para tanto, ao estudar espécies que divergem nessa característica (número de cotilédones), principalmente, torna-se interessante ter como base de estudos duas espécies modelo. Segundo Devos (2005), o nível de conservação dos genes é um critério importante na determinação do grau de sintenia, para o qual o conhecimento comparativo pode ser aplicado entre as espécies. Em contraste com esforços baseados em mapas genéticos, que provém uma visão geral dos rearranjos cromossomais, os quais diferenciam espécies relacionadas, as comparações baseadas em sequências determinam se a ordem dos genes permaneceu conservada entre os segmentos de cromossomos ortólogos.

Dada a importância econômica dos cereais, visto que aproximadamente 60% da alimentação mundial é obtida das gramíneas (KELLER et al., 2000), a escolha de uma espécie modelo para estudos biotecnológicos é necessária, e o arroz se encaixa perfeitamente nos pré-requisitos para a base dos estudos da genômica comparativa, no que se refere ao grupo das gramíneas. Possui um genoma relativamente pequeno, de 450 Mb, quando comparado aos genomas do milho, de 2.500 Mb, da cevada (*Hordeum vulgare* L.), de 4.900 Mb, ou do trigo (*Triticum aestivum* L.), com 16.000 Mb (NCBI).

O arroz tem seu genoma totalmente sequenciado e pesquisadores concentram esforços na construção de mapas genéticos, em técnicas de transformações genéticas e construção de mapas comparativos com outras espécies da família Poaceae. Os resultados que se tem obtido nessas análises comparativas, mostram extensas regiões com conteúdo gênico conservadas, tanto na ordem quanto na disposição dos genes. Isso demonstra que o arroz pode prover valiosas informações que podem ser extrapoladas para outras

espécies no que se refere a características semelhantes em outras gramíneas (BENNETZEN, 2002).

Segundo Salse et al. (2009), o conhecimento sobre a extensão da conservação entre os genomas dos cereais e as ferramentas de análises gerados, através da genômica comparativa, estudos podem ser usados para (1) definir estratégias eficientes para estudos genéticos e isolamento de genes através do padrão de grupos de marcadores ortólogos conservados e (2) para melhorar a acuracidade da anotação gênica através do alinhamento de genes ortólogos conservados.

A partir de análises comparativas, ao nível de mapas genéticos, identificam-se segmentos do genoma, ou blocos de ligação, que consistem de dezenas de megabases, as quais são fortemente colineares entre as espécies. Esses estudos de mapas, de baixa e média resolução, não oferecem informação da organização precisa dos genes nesses blocos (DEVOS, 2005). Para obter uma melhor visão do micro-nível de conservação dos genes, muitos laboratórios tem realizado o sequenciamento de BACs (*bacterial artificial chromosome*), que foram selecionados por conterem um gene em particular que esteja presente entre as espécies de cereais (CHEN et al., 1997; TIKHONOV et al., 1999; RAMAKRISHNA et al., 2002; GUYOT et al., 2004, LU et al., 2009).

Chen et al. (1997), clonaram genes de arroz e sorgo homólogos ao locus *sh2* de milho em BACs. Observaram que um homólogo do gene *a1* em milho também estava presente em cada um dos BACs. Encontraram uma alta conservação do arranjo, observado em milho, em arroz e sorgo, principalmente nos éxons. Concluíram, assim, que as três espécies estudadas tem a ordem dos genes conservada, bem como a composição da região sh2-a1, mas adquiriram diferenças quantitativas e qualitativas entre as sequências dos genes (como duplicações), em seus processos evolutivos individuais.

Em outro estudo, Lu et al. (2009), para melhor compreender a evolução do genoma do gênero *Oryza*, sequenciaram e compararam regiões genômicas *MOC1* (*MONOCULM1*), entre 14 genomas de *Oryza*. Localizado no braço longo do cromossomo 6, MOC1 codifica para a família da proteína nuclear GRAS, a qual controla uma característica agronômica importante, a formação de perfilhos, no arroz. O sequenciamento e anotação de 18 BACs para estas espécies revelou uma alta colinearidade na conservação dos genes e na

estrutura da região *MOC1*. As diferenças na amplificação por transposons parecem ser responsáveis pelas diferenças de tamanhos dos genomas atuais do gênero *Oryza*.

O que se obtém a partir da corroboração dos estudos, ao longo das diferentes espécies de gramíneas, é que a colinearidade é mantida ao nível de sequência em diferentes níveis, dependendo da região e da espécie observada. A ruptura da colinearidade é manifestada por pequenas inversões, duplicações em *tandem*, inserções e/ou deleções únicas ou múltipas e translocações gênicas (Figura 8).

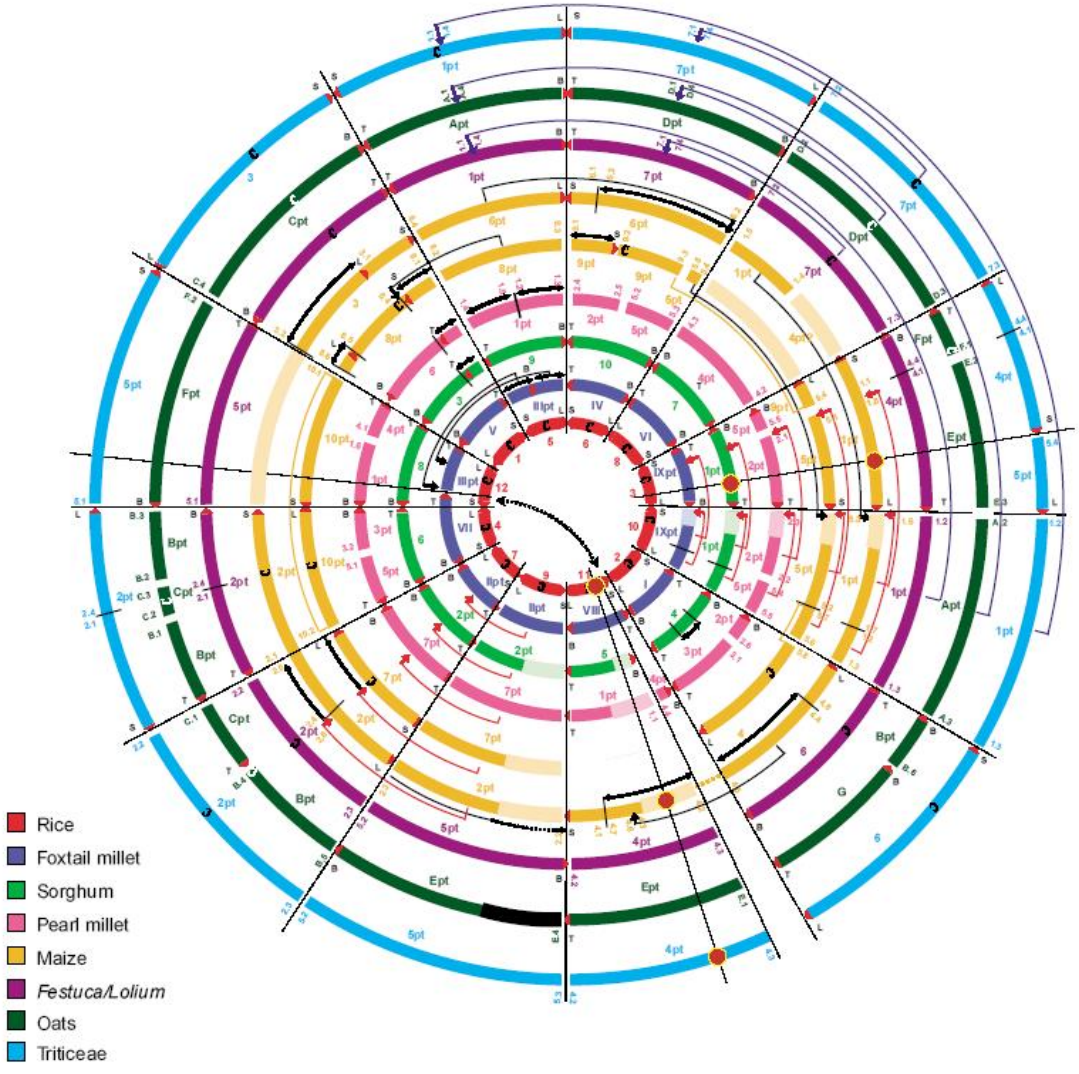


Figura 8. Diagrama em círculo mostrando as relações conhecidas entre os genomas de oito espécies de gramíneas. Fonte: Devos, 2005.

Pela comparação de regiões ortólogas entre as múltiplas espécies, é possível inferir a época relativa dos diferentes rearranjos, e também determinar

a estabilidade relativa dos genomas analisados. O genoma do arroz parece ser mais similar em estrutura a organização do genoma ancestral do que os genomas de sorgo e milho, por exemplo (RAMAKRISHNA et al., 2002; ILLIC et al., 2003).

Regiões que contem agrupamentos de genes em *tandem* são particularmente instáveis. Prova disso é a presença diferencial de grupos de genes de resistência a doenças (GUYOT et al., 2004), genes de pequenos RNAs nucleolares (snoRNA) (RAMAKRISHNA et al., 2002) e genes de proteínas de reserva (SONG et al., 2002) em cereais. Devos (2005) cita que após a amplificação, esses genes devem ter sido retidos em linhagens específicas em resposta às pressões de seleção do ambiente. Os mecanismos que dirigem os rearranjos são geralmente desconhecidos, mas alguma informação pode ser obtida a partir das regiões que flanqueiam estes genes. A inserção de um gene em uma nova posição pode levar a mudança em sua regulação transcricional, que é causada pela mudança do ambiente cromossomal (SONG et al., 2001). A acumulação de mutações degenerativas em elementos regulatórios pode causar a quebra de funções ancestrais (FORCE et al., 1999).

No que diz respeito às dicotiledôneas e a família das Poáceas (leguminosas), particularmente, o mapeamento pela genética comparativa é bem estabelecido (DELSENY et al., 2004), onde estudos iniciais predizem que a sintenia deve facilitar a descoberta de genes entre espécies relacionadas (DEVOS & GALE, 2000). As limitações nas comparações entre as famílias se devem a rearranjos cromossomais que resultam no fracionamento progressivo do genoma em segmentos bem conservados e alta freqüência de perda de genes em segmentos duplicados do genoma. De acordo com Choi et al. (2004), através de estudos de documentos, a conservação substancial entre os genomas de culturas de leguminosas modelo, também revela alguma divergência. O grau com o qual a sintenia do genoma pode facilitar a análise entre espécies sobre a função dos genes dependerá tanto da conservação da ordem e conservação dos genes, quanto da freqüência com as quais uma característica similar tem bases genéticas em comum nas diferentes espécies.

A domesticação é um evento relativamente recente que ocorreu independentemente nos diferentes cereais. As características que tem sido

selecionadas durante a domesticação são a ausência de debulha das sementes, sementes de tamanho maior, tempo de lavoura reduzido e crescimento padronizado (PATERSON et al., 1995). A colinearidade entre as monocotiledôneas e *Arabidopsis thaliana* é fragmentada (DEVOS et al., 1999), mas sabe-se que não há informação de espécies que divergiram entre 60 milhões de anos atrás (surgimento das gramíneas) e 150-200 milhões de anos atrás (divisão das monocotiledôneas e dicotiledôneas). Os genes ortólogos que tem funções conservadas entre espécies produzem fenótipos similares entre elas. Peng et al. (1999), observaram como exemplo que os genes ortólogos do nanismo *GAI, Rht-1* e *D8* reduzem o tamanho da planta em Arabidopsis, trigo e milho, respectivamente.

2.5 – Regulação da expressão gênica

A modelagem contínua da forma de um vegetal é uma integração de uma grande variedade sinais. No desenvolvimento de mudas jovens é particularmente claro. Fatores ambientais, como a luz, alteram profundamente o programa morfogenético inato da planta. Como diversas rotas se mesclam para determinar uma discreta resposta de crescimento celular, em grande parte esse é um processo desconhecido (NEMHAUSER, 2004).

Estresses abióticos como seca, alta salinidade, baixa temperatura e o alagamento são, todos, condições ambientais que tem efeito adverso no crescimento das plantas e na produtividade dos cereais. As plantas tem se adaptado para responder a esses estresses a nível molecular, fisiológico e bioquímico, para tornar viável sua sobrevivência (YAMAGUCHI-SHINOZAKI & SHINOZAKI, 2006). A expressão de uma variedade de genes é induzida por estresses em diversas plantas e os produtos desses genes funcionam não somente na tolerância ao estresse, mas também na regulação da expressão do gene e na transdução de sinais nas respostas aos estresses (BARTELS & SUNKAR, 1995).

Na rede de transdução, desde a percepção dos sinais do estresse até a expressão dos genes a ele responsivos ao estresse, vários fatores de transcrição e elementos de ação *cis*, nos promotores também responsivos ao estresse, funcionam para a adaptação da planta a esses estresses ambientais. A análise das complexas cascatas de sinalização e regulação dos genes no

cross-talk (conversa cruzada) entre os mesmos e seus produtos, bem como a identificação da especificidade da expressão gênica ajudam elucidar como uma planta tolera um estresse e quais meios, sejam eles fisiológicos, bioquímicos ou moleculares, ela utilizou para emitir uma resposta favorável a sua sobrevivência.

Os produtos desses genes induzíveis pelo estresse parecem promover uma resposta de tolerância e regulam a expressão gênica através de rotas transdução de sinais (XIONG et al., 2002). Para melhor compreender os mecanismos moleculares que regulam a expressão gênica em resposta a estresses abióticos, estudos tem sido focados na análise de elementos *cis* e *trans* em Arabidopsis e seu papel na mediação das respostas ao estresse (YAMAGUCHI-SHINOZAKI & SHINOZAKI, 2006).

2.6 – Elementos de ação *cis*

2.6.1 – Caracterização

Um dos mecanismos pelos quais os níveis de proteínas na célula são controlados é a regulação transcricional. Certas regiões no DNA, chamadas de elementos regulatórios *cis,* são locais de ligação para proteínas de ação *trans* envolvidas na transcrição, tanto para o posicionamento da maquinaria transcricional básica, quanto para a regulação.

A maquinaria transcricional básica consiste da RNA polimerase (RNApol), que sintetiza vários tipos de RNA, e promotores essenciais são usados para posicionar a RNApol. Outras regiões próximas vão regular a transcrição: em procariotos, operadores estão envolvidos; em eucariotos, regiões do promotor proximal, potencializadores (*enhancers*), silenciadores (*silencers*) e isoladores (*insulators*) estão presentes (Figura 9) (RIETHOVEN, 2010).

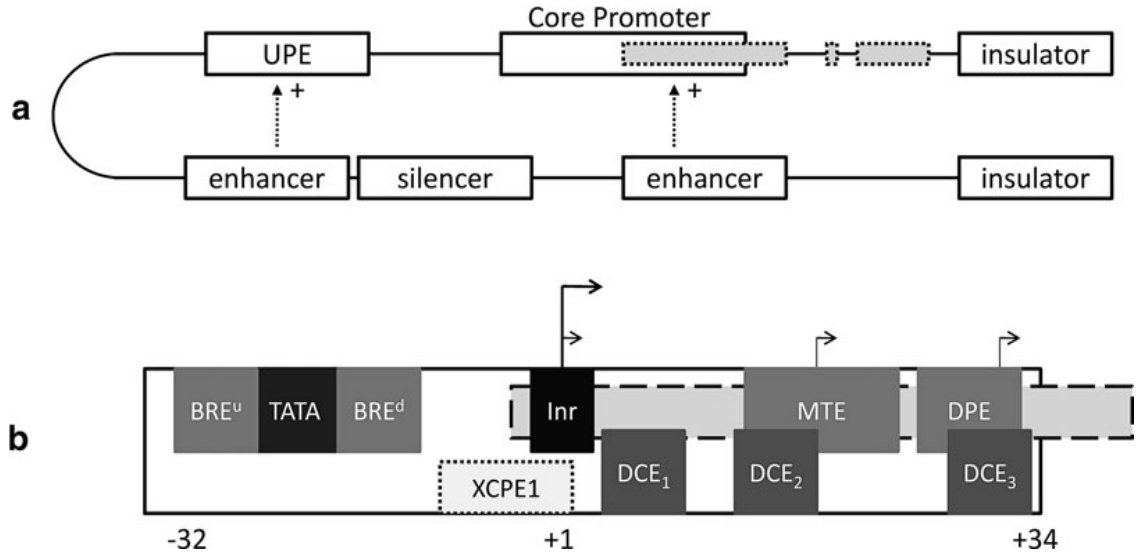


Figura 9. Unidades regulatórias transcricionais. Visão esquemática (**a** não está em escala) dos vários elementos nas unidades transcricionais em eucariotos superiores. Éxons são mostrados em caixas cinzas com linhas tracejadas. **a**, unidade transcricional com um promotor principal grande (maior do que em procariotos) sobrepondo o primeiro exon e elementos do promotor à montante (*upstream*). A alça do DNA (*loop*) mostra que os potencializadores (*enhancers*) podem ser trazidos fisicamente para perto do promotor principal ou aos UPE (*upstream promoter elements* – elementos do promotor à montante). **b**, arquitetura detalhada do promotor constitutivo. Vários elementos, muitos deles opcionais, são mostrados em escala. Elementos do promotor sombreados são mais freqüentes. Abreviações: TATA *box* (TATA), elemento iniciador (Inr), elemento de reconhecimento do complexo iniciador (BRE, *upstream* ou *downstream*), elemento de dez motivos (MTE), elemento constitutivo a jusante (DCE, subunidades 1,2 e 3), elemento 1 do promotor constitutivo X (XCPE1) e o elemento promotor downstream (DPE). Note que o promotor constitutivo pode ser focalizado ou dispersado, aqui é mostrado por um sítio de iniciação transcricional (TSS – *Transcriptional Start Site*) em negrito e outros TSSs menores, respectivamente. Fonte: Riethoven, *in:* Computation Biology of Transcriptional Factor Binding, 2010.

Os sítios de ligação de fatores de transcrição (ou “elementos *cis*”; “motivos”) são elementos funcionais de DNA que influenciam temporal e espacialmente a atividade transcricional. Múltiplos elementos *cis* formam módulos *cis*-regulatórios (MCRs), que integram sinais de múltiplos FTs, resultando no controle combinatório e padrões altamente específicos da expressão gênica (PRIEST et al., 2009).

Os elementos *cis* funcionam como interruptores moleculares em resposta aos sinais de estresse do ambiente. São caracterizados como sequências curtas de DNA, entre 4 a 12 pb, localizadas na região promotora dos genes, com as quais vários FTs interagem para formar um complexo de iniciação transcricional no TATA *box*, logo acima aos sítios de iniciação da transcrição

(YAMAGUCHI-SHINOZAKI & SHINOZAKI, 2005). O complexo de iniciação transcricional ativa, por sua vez, a RNA polimerase a iniciar a transcrição de genes responsivos ao estresse (Figura 10). Nesse processo várias interações entre os elementos *cis* e fatores de transcrição funcionam como reguladores moleculares da transcrição para determinar os eventos de iniciação da transcrição.

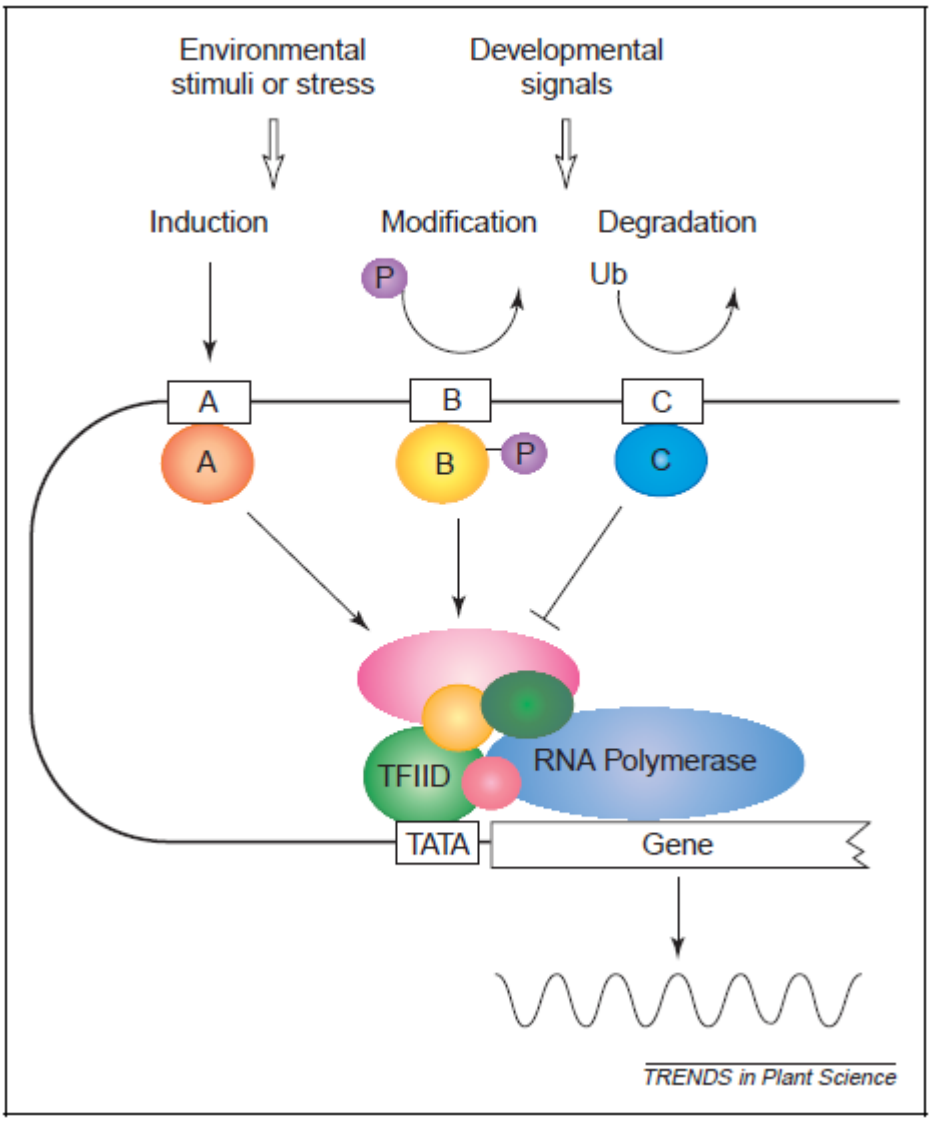


Figura 10. Esquemas das redes de regulação transcricional. O complexo de iniciação transcricional é regulado por fatores de transcrição que são ativados ou reprimidos por estímulos ambientais e sinais de desenvolvimento. As caixas retangulares nomeadas de A, B e C representam fatores de ação *cis* e os círculos A, B e c (vermelho, amarelo e azul, respectivamente) representam os fatores de transcrição. Ub representa ubiquitina. Fonte: *TRENDS in Plant Science*, v.10, n.2, p.89, 2005.

Os sinais de estresses abióticos ativam os fatores de transcrição pela indução dos seus genes, ativação de proteínas (como fosforilação) e degradação de proteínas (Figura 10). De acordo com Yamaguchi-Shinozaki &

Shinozaki (2005), acredita-se que seja importante determinar elementos de ação *cis* nos promotores de genes responsivos a estresses para entender como funcionam esses interruptores moleculares de genes estresse induzíveis. O Quadro 1, contém um resumo das ferramentas utilizadas para determinar elementos *cis*, suas proteínas de ligação ao DNA e atividades de ativação transcricional.

Quadro 1. Procedimentos experimentais para identificar elementos de ação cis e suas proteínas de ligação ao DNA, CGF/UFPel, 2011.

Sistemas de plantas para análise de elementos *cis*

- Plantas transgênicas.
- Expressão transitória usando micro projéteis.
- Expressão transitória usando protoplastos.

Procedimentos experimentais para a determinação de elementos *cis*

- Análise de deleção de regiões promotoras (experimento de perda de função).
- Análise funcional de fragmentos dos promotores (experimento de ganho de função).
- Análise de elementos *cis* usando mutações de ponto em fragmentos dos promotores (substituição de base).
- Imunoprecipitação da cromatina.

Identificação e confirmação de proteínas de ligação ao DNA para elementos *cis*

- Ensaio de mobilidade em gel.
- Análise de marcas (*footprint*) de DNA.
- Interferência de metilação.
- Precipitação de DNA por afinidade (PDNA).

Isolamento de clones de cDNA das proteínas de ligação ao DNA para elementos *cis*

- Varredura de *south-western blotting*
- Varredura de leveduras híbridas.
- Seleção de candidatos baseado em dados genômicos, como análise filogenética e análise de microarranjos.

Fonte: *TRENDS in Plant Science*, v.10, n.2, p.91, 2005.

2.6.2 – Elementos *cis* na regulação transcricional

As interações entre diferentes tipos de elementos *cis* funcionam em um *cross-talk* entre os diferentes sinais. Yamaguchi-Shinozaki & Shinozaki (2005),

observaram interações entre DRE (*Dehydration-responsive element*) e ABRE (*ABA-responsive element*) no promotor do gene *rd29A*; ambos os elementos *cis* são encontrados no promotor *rd29A*. Uma única cópia de DRE é suficiente para a expressão do gene responsivo ao estresse. O promotor *rd29A* tem quatro cópias de sequências do tipo DRE e uma ABRE. Foi demonstrado que um dos DRE funciona em ligação com o elemento ABRE (Narusaka, 2003) e constitui um complexo de resposta ao ABA (ácido abscísico), na expressão do gene induzido por ABA. Isso indica a existência de um *cross-talk* entre diferentes rotas de sinalização e elementos *cis* em promotores sensíveis a estresses.

Os fatores de transcrição são reguladores mestres no controle de grupos de genes. Um único FT pode controlar a expressão de muitos genes alvo através de ligações específicas aos elementos de ação *cis*, presentes nos promotores desses genes alvo. Esse tipo de sistema de regulação transcricional se chama *regulon*. Vários *regulons* tem sido identificados em *Arabidopsis thaliana*. Como exemplo, os *regulons* DREB1 (*Dehydration-responsive element binding protein 1*) / CBF (*C-repeat binding factor*) e DREB2 que funcionam na expressão de genes independentes de ABA. Enquanto que os *regulons* AREB (ABA-responsive element binding protein) / ABF (ABRE binding factor) funcionam na expressão de genes dependentes de ABA (NAKASHIMA et al., 2009) (Figura 11).

Estudos recentes mostram que os *regulons* DREB1 / CBF, DREB2, AREB / ABF e NAC (resposta a desidratação, alta salinidade e frio) tem papéis importantes na resposta a estresses abióticos em *Oryza sativa* (Figura 11), o grão preferido em estudos genéticos e moleculares em respostas a estresses, devido ao seu valor comercial, genoma relativamente pequeno e relação íntima com outros cereais importantes.

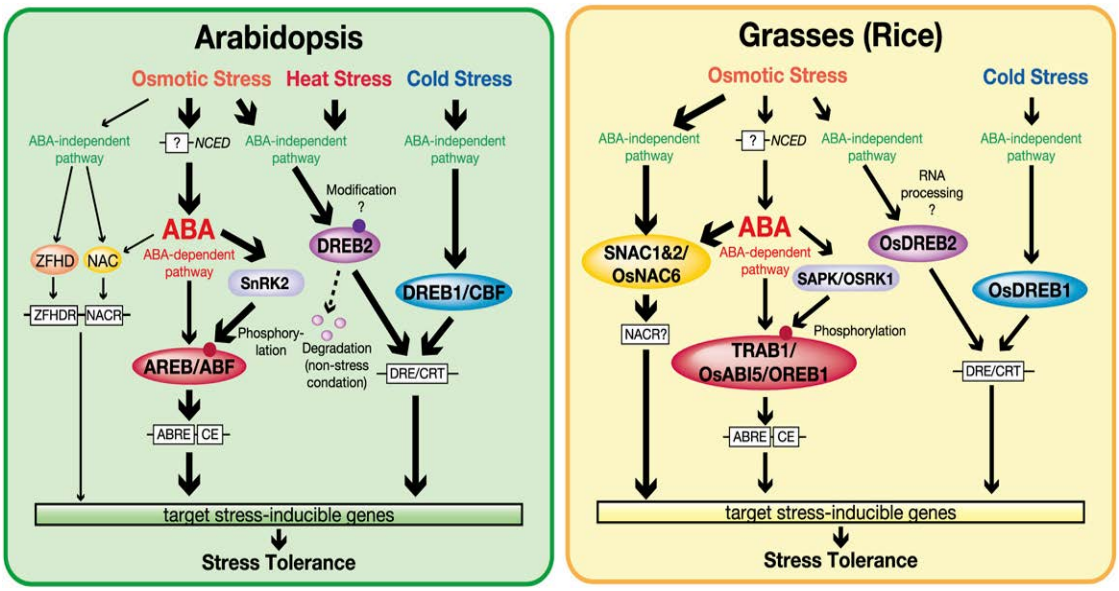


Figura 11. Principais redes de regulação transcricional de elementos *cis* e FTs envolvidos na expressão de genes de resposta a estresses abióticos em Arabidopsis e gramíneas, como o arroz. FTs controlando a expressão de genes estresse induzíveis são mostrados nas elipses. Elementos *cis* envolvidos na transcrição responsiva ao estresse são mostrados nas caixas brancas. Proteínas quinases envolvidas na fosforilação dos FTs estão destacadas em lilás. As elipses pequenas e vermelhas ilustram as modificações dos FTs, por exemplo, quando passa por um processo de fosforilação, em resposta a sinais do estresse. Fonte: Plant Physiology, v.149, p.89, 2009.

Estudos moleculares e genéticos dão evidências de que a dicotiledônea *Arabidopsis thaliana* e a monocotiledônea *Oryza sativa* tem mecanismos de regulação da expressão gênica em comum. FTs exercem um papel importante na regulação da expressão gênica em resposta a estresses abióticos e a maioria dos FTs são comuns entre as gramíneas e Arabidopsis (Figura 11) (NAKASHIMA et al., 2009).

2.7 – A Bioinformática no estudo dos elementos *cis*

Com a recente disponibilidade de muitas sequências de alta qualidade e de genomas de plantas completamente anotados, surgiram grandes bancos de dados públicos de medidas de expressão de transcritos e de fácil acesso às tecnologias de perfis de expressão para laboratórios individuais. Com isso, surge também a necessidade de estudos envolvendo os sítios de ligação dos FTs e seu papel como componentes de uma grande rede transcricional (PRIEST et al., 2009).

Trabalhos recentes em várias espécies eucarióticas tem sido focados em uma abordagem de Biologia de Sistemas (*Systems Biology*), para elucidar as

redes regulatórias e compreender seu contexto biológico (LEVINE & DAVIDSON, 2005). Essas associações podem ser utilizadas em combinação com dados de expressão gênica de experimentos de microarranjo e análise de sequências de promotores de genes co-regulados, para inferir mecanismos dessa regulação e para buscar elementos regulatórios *cis* que podem coordenar essa resposta através da atividade de fatores de transcrição (NERO et al., 2009).

Enquanto as análises de dados obtidos em experimentos de microarranjos podem determinar um grupo de genes que são regulados sob condições experimentais específicas, as mesmas não identificam os componentes de ação *cis* ou *trans* envolvidos nessa regulação. De acordo com Nero et al. (2009), no entanto, esse grupo de genes co-regulados pode ser usado para identificar relações candidatas entre FT e sequências alvo usando associações entre os FTs e seus alvos, baseado na correlação sobre os dados do microarranjo e detecção de possíveis elementos *cis*.

As supostas regiões regulatórias *upstream*, de tamanho arbitrário, são utilizadas para identificar motivos candidatos de DNA. Múltiplos motivos com alta frequência em promotores de um grupo de genes co-expressos podem representar o mesmo módulo *cis* regulatório (MCR), e, portanto, estar agindo em um modo combinatório. Porém esse pressuposto – que genes co-expressos são transcricionalmente co-regulados – pode não ser sempre verdadeiro. Ensaios de microarranjos medem o estado estável dos níveis de transcrito em uma amostra em particular, e não a atividade transcricional *per se*. A maioria dos experimentos de perfil de transcritos baseados em ensaios de microarranjos não pode distinguir entre mudanças nos níveis de transcritos causadas na regulação pós-transcricional (por exemplo, estabilidade do transcrito) e na regulação transcricional mediada por elementos *cis* (PRIEST et al., 2009).

Um número considerável de algoritmos e ferramentas de bioinformática tem sido desenvolvidos para identificar potenciais elementos *cis* nas sequências regulatórias de genes co-expressos (ROMBAUTS et al., 2003; WASSEMAN & SANDELIN, 2004). O princípio fundamental das abordagens computacionais é de que genes co-regulados devem conter elementos *cis* similares nas suas regiões regulatórias *upstream* a níveis estatisticamente significantes (PRIEST,

2009). Independente dos detalhes algorítmicos exatos, em linhas gerais, as abordagens computacionais para a identificação de possíveis elementos *cis* estimam a probabilidade de ocorrência de motivos curtos de DNA pela comparação do número observado de ocorrências de um motivo conservado em um grupo de sequências com o número esperado de ocorrências, baseado em amostras aleatórias ou em algum modelo estatístico de distribuição (MICHAEL et al., 2008).

Walther et al. (2007) usaram o banco de dados *AtGenExpress* (http://www.arabidopsis.org/info/expression/ATGenExpress.jsp), para testar a hipótese de que genes diferencialmente expressos em resposta aos diversos estímulos contém um número maior de elementos *cis* distintos em suas regiões *upstream* do que os genes que respondem a relativamente poucos estímulos. Combinando diferentes padrões de expressão com análise de elementos *cis* em promotores de *Arabidopsis*, encontraram uma correlação positiva entre os genes que respondem a múltiplos estímulos e a densidade de elementos *cis* nas suas regiões promotoras. Talvez, não surpreendentemente, genes preditos com funções na regulação transcricional, na resposta a estresses e em processos de sinalização exibem melhor capacidade regulatória, por terem uma maior quantidade de sítios de ligação a FTs em seus promotores.

Essas anotações de elementos *cis* super-representados em grupos de genes co-expressos oferecem novas e poderosas fontes para elucidar os mecanismos de controle transcricional nas plantas e inferências sobre as informações funcionais dos genes.

2.7.1 – Ferramentas da bioinformática para descoberta de motivos regulatórios em sequências promotoras de genes homólogos e co-regulados

Compreender os mecanismos complexos que regulam a expressão gênica é um dos maiores desafios da biologia molecular moderna. A transcrição é regulada pela interação de FTs com seus sítios de ligação correspondentes (LEVINE & TIJAN, 2003), a maioria localizada próxima ao sítio de iniciação transcricional do gene (por exemplo, região promotora proximal) ou distantes (por exemplo, potencializadores e silenciadores). Então, na busca e descoberta

de elementos *cis*, é desejável que informações biológicas possam ser extraídas de sequências nucleotídicas, por análise computacional.

Há dois tipos de abordagens computacionais para a busca e descoberta de elementos *cis*: uma delas consiste no uso de uma ferramenta de reconhecimento de perfis ou assinaturas frente a um banco de dados com uma coleção de elementos *cis* previamente descobertos e anotados; a outra consiste na submissão de um conjunto de promotores de genes, previamente selecionados como co-regulados ou tendo alguma característica de interesse em comum, à uma ferramenta de busca de motivos conservados *ab initio*.

Para buscas por elementos *cis* já conhecidos e com funções anotadas, em plataformas *online*, o PLACE (http://www.dna.affrc.go.jp/htdocs/PLACE) é uma opção interessante, pois consiste de um banco de dados de elementos regulatórios de ação *cis*, baseado em uma compilação de bancos de dados de elementos cis de DNA de plantas superiores. É o mais atualizado e completo, no momento contem ainda uma breve descrição de cada motivo e literatura no PubMed (HIGO et al., 1998). Contudo, o PLACE teve sua última atualização em 2007, e com o avanço das tecnologias de seqüenciamento de novos genomas e dos dados gerados pela bioinformática, a busca por novos padrões se faz necessária e promissora.

Para a descoberta de novos elementos *cis* uma abordagem possível é a seleção de grupos de genes que sejam co-regulados, e procurar por motivos conservados que possam estar presentes nos seus promotores, os quais podem representar sítios de ligação para os FTs em comum, que possam regulá-los (PAVESI et al., 2004). Alternativamente, um único gene pode ser comparado com seus homólogos em outras espécies (ou, as vezes, parálogos), pela busca de conservação de sequências não codificantes que o flanqueiam ou nos íntrons: regiões preservadas pela evolução parecem exercer algum papel na regulação (PRAKASH & TOMPA, 2005).

Para este segundo tipo de análise de padrões de sequências de DNA, a suíte de aplicativos MEME oferece um *kit* de ferramentas em uma interface unificada (http://meme.sdsc.edu/meme4_5_0/intro.html), que permite realizar quatro tipos de análises: descoberta de motivos, busca em banco de dados motivo-motivo, busca em banco de dados sequência-motivo e atribuição de função (BAILEY et al., 2009). Os passos fundamentais dos processos

utilizados, pelo MEME (e outros programas de busca de padrões de sequências) são: primeiramente, um grupo de oligonucleotídeos, similares o suficiente para serem reconhecidos pelo mesmo FT, é detectado nas sequências submetidas. Estes são os motivos candidatos. Então, o grupo é avaliado de um ponto de vista estatístico. Esta medida deve considerar o tamanho do grupo e quão conservado ele é, isso é, quantas vezes os oligos são encontrados nas sequências e quanto eles diferem um do outro; então, os motivos mais significantes são mostrados, assim como os oligos, onde as sequências foram encontradas, e cada uma é um candidato à sítio de ligação de fatores de transcrição, para o mesmo FT (PAVESI et al., 2004).

Para o reconhecimento de motivos conservados, o MEMEM utiliza um modelo matemático chamado modelo oculto de Markov (HMM – *Hidden Markov Model*). Consiste em um grupo de estados e transições entre esses estados. Cada estado emite um sinal com base em um conjunto de probabilidades de emissão e, em seguida, emite ao acaso transições para outros estados, com base em um conjunto de probabilidades de transmissão. Essas duas distribuições de probabilidade, quando combinadas com o estado inicial de distribuição, caracterizam um HMM por completo (GRUNDY et al., 1997).

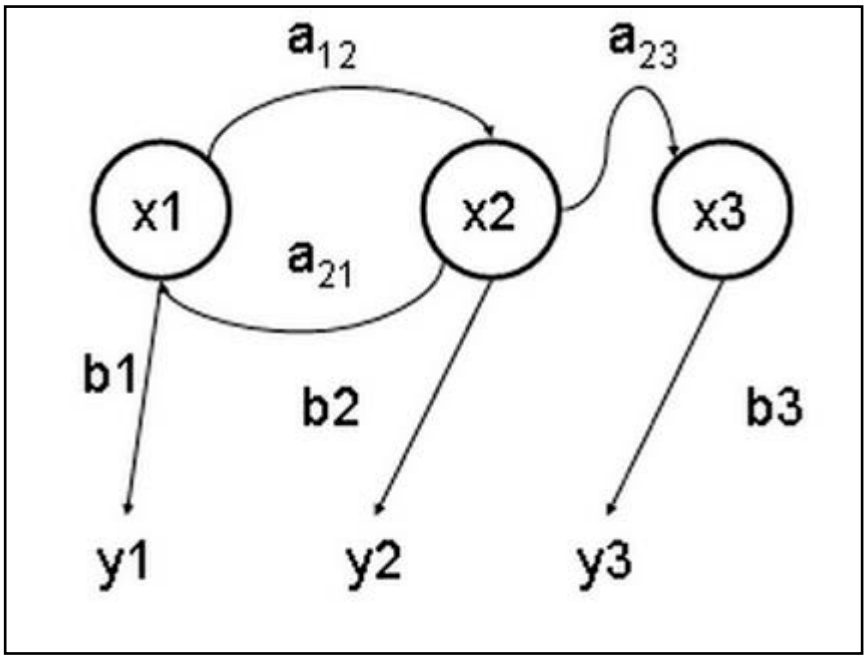


Figura 12. Esquema de um HMM, onde: x1, x2 e x3 representam os estados ocultos; a_{12}, a_{21} e a_{23} são as probabilidades de transição; b1, b2 e b3 probabilidades de saída; y1, y2 e y3 saídas observadas. Fonte: ACADEMIC (www.esacademic.com).

Em linhas gerais, o objetivo de um modelo oculto de Markov é determinar os parâmetros desconhecidos (ou ocultos, daí o nome) da cadeia, a partir dos parâmetros observáveis (Figura 12). Os parâmetros extraídos podem ser usados para realizar análises posteriores, por exemplo, em aplicações de reconhecimento de padrões. Modelos ocultos de Markov são particularmente aplicado ao padrão temporal de reconhecimento em bioinformática.

A predição de elementos regulatórios em estudos *in silico* viabiliza e guia futuros experimentos para a validação de novos padrões de ocorrência destes, e a partir das análises de sequências homólogas, inferências sobre funções genéticas e na regulação da transcrição podem ser feitas de forma mais segura.

3. OBJETIVOS

3.1 – Geral

Identificar padrões de ocorrência dos elementos *cis* em genes de *Oryza sativa* e *Arabidopsis thaliana, Zea mays* e *Glycine Max*, relacionados ao estresse do alagamento.

3.2 – Específicos

Definir grupos de genes que sejam co-regulados pelo estresse da anoxia, com características em comum, para obter possíveis padrões de ocorrência de elementos *cis*.

Descrever padrões de ocorrência de elementos cis, previamente descritos associados aos genes co-regulados.

Descrever novos padrões de ocorrência de elementos *cis* em grupos de genes co-regulados e que tenham características em comum.

4. MÉTODO SUGERIDO

A metodologia é exposta a seguir de forma a descrever os passos e as ferramentas utilizadas nesse trabalho. Uma vez que não há protocolos descritos para tais procedimentos, cada etapa realizada foi detalhada.

4.1 – Busca na literatura por genes que respondem ao estresse do alagamento

Foi realizada uma revisão bibliográfica a fim de listar os genes que tem conhecidamente sua expressão alterada, quando expostos ao estresse abiótico do alagamento (hipoxia ou anoxia).

Os dados extraídos referem-se a análises de plantas de *Arabidopsis thaliana* e *Oryza sativa*, em experimentos com microarranjos, sob o estresse da hipoxia ou anoxia. Essas duas espécies foram escolhidas, pois representam espécies modelo para estudos com genômica comparativa.

Dentre os artigos revisados, destacam-se:

- *Transcript Profiling of the Anoxic Rice Coleoptile. Kudahettige, R.L. et al., 2007.*
- *Genome-Wide Analysis of the Effects of Sucrose on Gene Expression in Arabidopsis Seedlings Under Anoxia. Loreti, E. et al., 2005.*
- *Global Transcription Profiling Reveals Comprehensive Insights into Hypoxic Response in Arabidopsis. Liu, F. et al., 2005.*
- *A Variable Cluster of Ethylene Response Factor-Like Genes Regulates Metabolic and Developmental Acclimation Responses to Submergence in Rice. Fukao, T. et al., 2006.*
- *Expression Profile Analysis of the Low-Oxygen Response in Arabidopsis Root Cultures. Klok, E.J. et al., 2002.*

4.2 – Obtenção das sequências dos genes anotados

Após a anotação dos genes de *Arabidopsis thaliana* e *Oryza sativa*, suas sequências foram obtidas a partir do banco de dados *Entrez Gene* (*Genes and mapped phenotypes*), que consiste em uma completa plataforma *online,* com a compilação de informações sobre sequências gênicas.

Esse processo se deu pela submissão do símbolo do gene, como, por exemplo, *At2g31390* ou *Os11g10480* (uma aquaporina em *Arabidopsis* thaliana e uma álcool desidrogenase em *Oryza sativa*, respectivamente). O resultado do *site* do banco de dados (Figura 13) é um conjunto de informações sobre localização do gene no genoma da planta, descrição, qual fita do DNA genômico (molde ou complementar) que codifica o gene encontra, sequência nucleotídica (com e sem *íntrons*), sequência da proteína codificada, mRNA (RNA mensageiro), NP (numeração de identificação da proteína), fonte, e dados bibliográficos, entre outras.

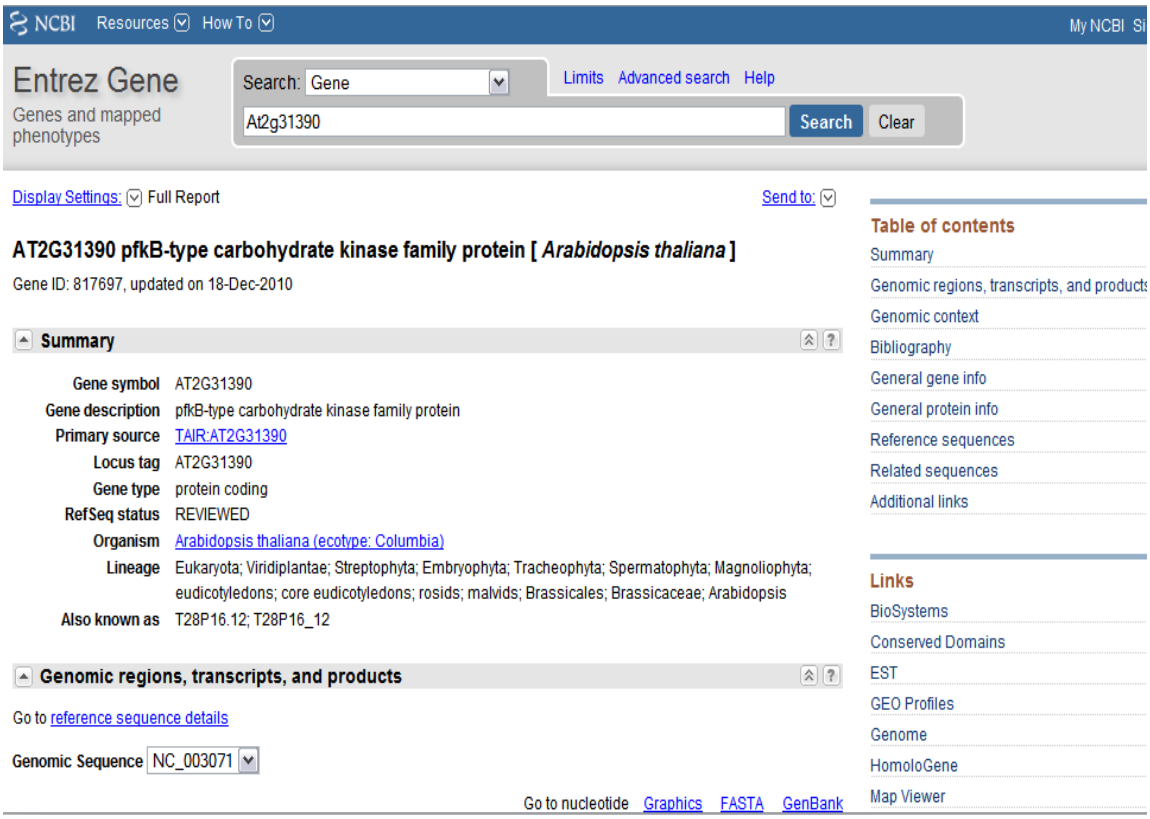


Figura 13. Resultado de uma busca no banco de dados *Entrez Gene*, por sequência genômica de *At2g31390*. Fonte: *Entrez Gene* (http://www.ncbi.nlm.nih.gov/gene/817697).

Os dados anotados para cada gene submetido foram: sua localização (coordenadas de onde o gene inicia e termina); a fita em que se encontra; sua sequência; e o seu NP (Figura 14).

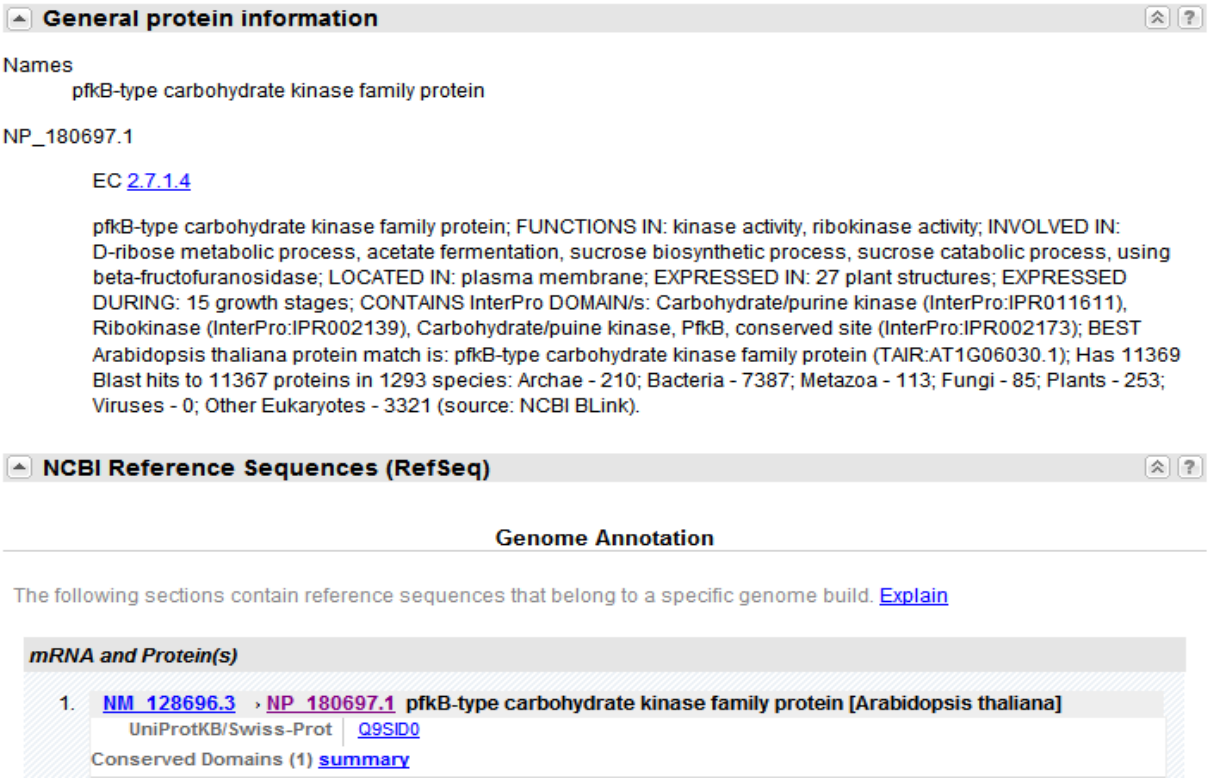


Figura 14. Continuação do resultado da busca por informações sobre o gene *At2g31390*. NP: número de acesso a banco de dados de proteínas. Fonte: *Entrez Gene* (http://www.ncbi.nlm.nih.gov/gene/817697).

A anotação do NP de um gene é importante quando se trabalha com homologia, pois essa nomenclatura dirige a um banco de proteínas primário (*UniProtKB*), onde seus dados são revisados e, por isso, possui maior acuracidade e confiabilidade.

4.2.1 – Obtenção das sequências de aminoácidos

Uma vez que se deseja trabalhar com homologia, é necessário obter dados das sequências protéicas codificadas pelos genes de interesse. Para tanto, o NP de cada gene (anotado juntamente com a sequência nucleotídica e outras informações citadas no item acima) foi submetido na plataforma *online* do *ExPASy – Swiss-Prot* (http://ca.expasy.org/sprot/), na base de dados da *Uniprot* (Figura 15).

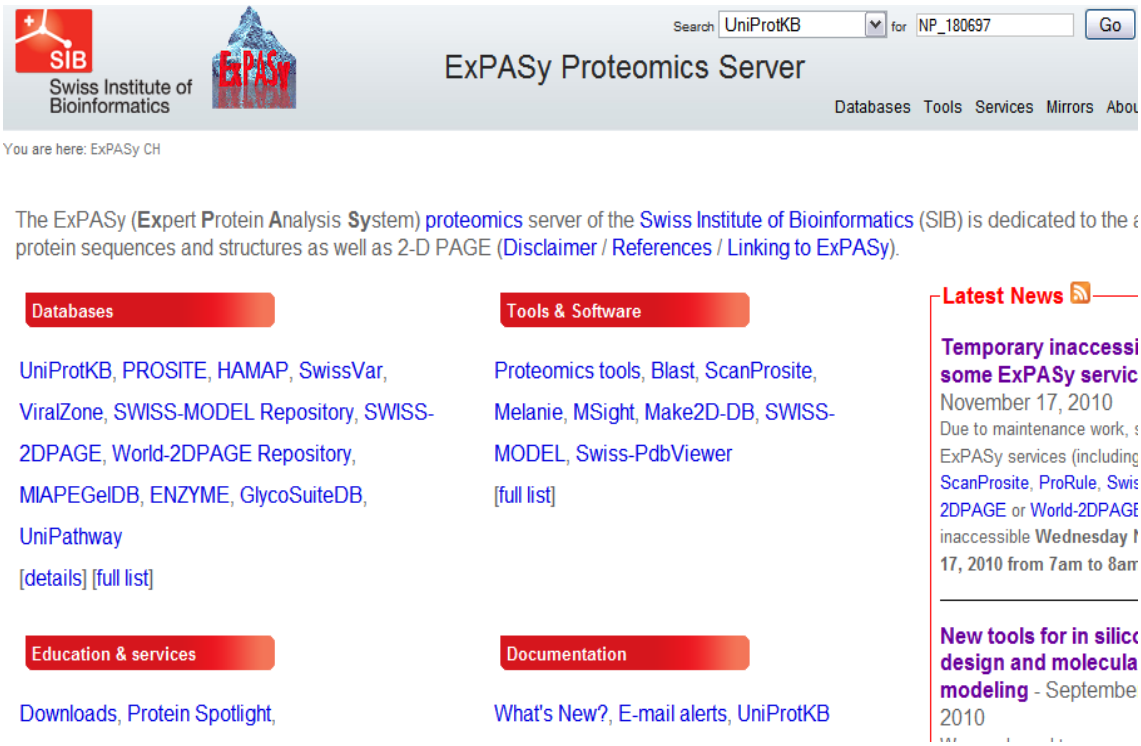


Figura 15. Visão parcial da plataforma *ExPASy.* Submissão do *NP_180697* (obtido na página do banco de sequências nucleotídicas *Entrez Gene*), referente ao gene *At2g31390*, na base de dados *UniProtKB*. Fonte: *ExPASy Proteomics Server* (http://expasy.org/).

Esta busca é encaminhada para o site da *UniProtKB*, onde todas as informações disponíveis (e revisadas) sobre esta proteína estão agrupadas (Figura 16). Dados como o nome da proteína, sequência de aminoácidos, número de acesso, organismo, tamanho da proteína, taxonomia, referências etc., estão disponíveis.

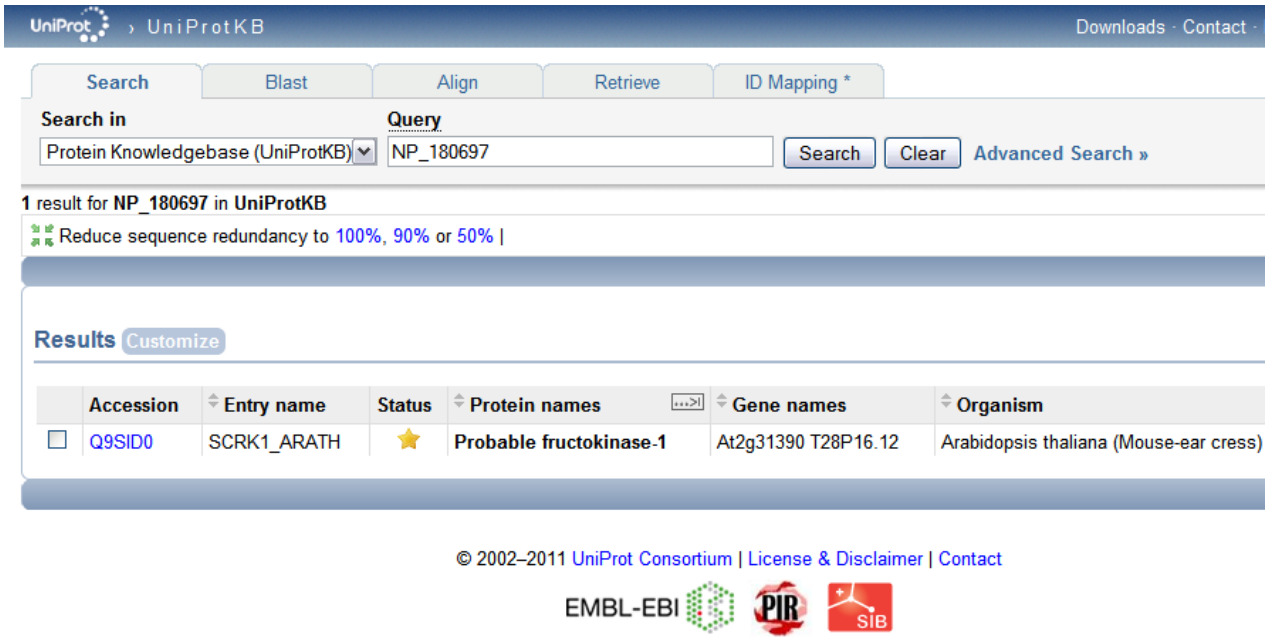
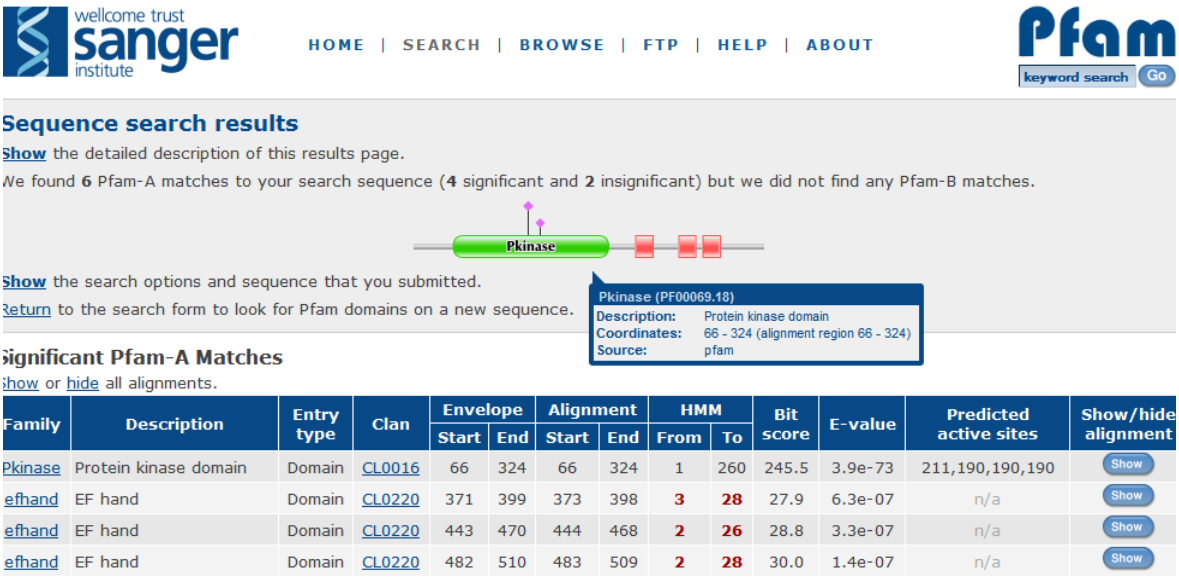


Figura 16. Visão parcial da base de dados de proteínas *UniProtKB*. Resultado da busca do NP_180697, que oferece um *link* (campo: *Acession*), para a obtenção das informações disponíveis sobre a proteína obtida. Fonte: *UniProtKB* (http://www.uniprot.org/uniprot/?query=NP_180697).

Para este trabalho foram anotados: nome da proteína, sequência de aminoácidos e o número de acesso.

4.2.2 – Classificação das famílias protéicas

As sequências protéicas foram submetidas, ao banco de dados *online* de famílias de proteínas, o *Pfam - Sanger Institute* (http://pfam.sanger.ac.uk/). O resultado são informações sobre a família da proteína, descrição, localização, função, entre outras (Figura 17), as quais permitem agrupar os genes estudados de acordo com características mais refinadas, além do fato de todos terem algum tipo de alteração na expressão quando expostos ao alagamento.



Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
Pkinase	Protein kinase domain	Domain	CL0016	66	324	66	324	1	260	245.5	3.9e-73	211,190,190,190	Show
efhand	EF hand	Domain	CL0220	371	399	373	398	3	28	27.9	6.3e-07	n/a	Show
efhand	EF hand	Domain	CL0220	443	470	444	468	2	26	28.8	3.3e-07	n/a	Show
efhand	EF hand	Domain	CL0220	482	510	483	509	2	28	30.0	1.4e-07	n/a	Show

Figura 17. Visão parcial do resultado apresentado, após a submissão da sequência de aminoácidos da proteína codificada pelo gene *At2g31390.* Fonte: *Pfam* (http://pfam.sanger.ac.uk/search/sequence).

A classificação das famílias se fez necessária, uma vez que desejou-se trabalhar com grupos de genes que tem características em comum, ao serem co-regulados, como sob o estresse por alagamento. Análises detalhadas dos genes permitiram uma maior segurança nas escolhas das sequências mais apropriadas ao objetivo do trabalho, diminuindo o risco de eliminar informações importantes.

Optou-se por trabalhar com todos os genes, pois todas as famílias de proteínas encontradas são de grande importância para a tolerância ao estresse abiótico em questão.

4.3 – Obtenção dos homólogos dos genes anotados

Uma vez que os genes de *Arabidopsis thaliana* tiveram todas as anotações feitas (acima citadas), foi realizada a busca por genes homólogos a estes em *Oryza sativa*. Ou seja, para cada gene de Arabidopsis, obteve-se (quando possível) seu homólogo em arroz. Este passo teve o objetivo de aumentar a probabilidade de encontrar um padrão de elemento *cis* quando as sequências foram alinhadas, dado que busca-se trabalhar com genes mais próximos.

Passos realizados na obtenção dos homólogos:

- Copiou-se a sequência protéica de um gene de *Arabidopsis thaliana* (previamente anotada) e realizou-se um alinhamento local, utilizando a ferramenta *BLASTp* (*Basic Local Alignment Search Tool – protein*)

(http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome);

- No BLASTp, selecionou-se "*Oryza sativa*" no campo "*Organism*" (Figura 18);
- No resultado apresentado pelo BLAST, foi selecionado o melhor *hit* o que direcionou para seu "*ref*", que se trata de um *link* que leva a página do *NCBI-Protein* onde o NP da proteína foi copiado;
- Este *Accession NP* foi colado no *site* do *ExPASy – UniProtKB*;
- Isto levou ao *Accession* (do *UniProt*). A partir deste *link*, obteve-se a sequência protéica e gênica de Arabidopsis (homóloga a proteína de arroz).

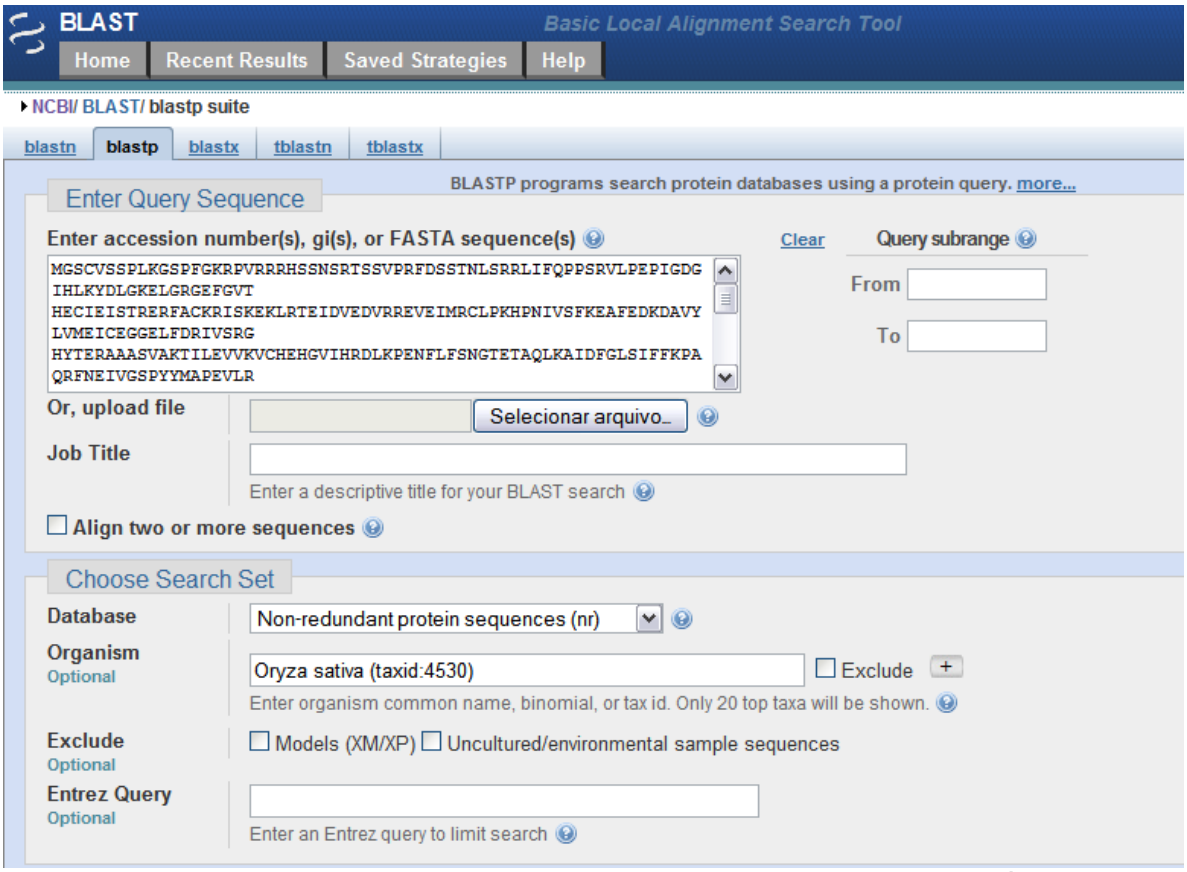


Figura 18. Ferramenta BLASTp, onde foram realizadas as buscas de homólogos. Fonte: BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome)

Este procedimento foi realizado para cada gene de Arabidopsis. O contrário também ocorreu, onde, para cada gene anotado de arroz, seu homólogo foi buscado em Arabidopsis.

A Tabela 1 mostra a classificação dos genes de arroz baseada na identidade de sequências determinada computacionalmente, a partir de um gene conhecido (homologia), proteína ou alguma sequência consenso (como domínio funcional de uma proteína (McCOUCH, 2008).

Tabela 1. Regras para a classificação de sequências gênicas, como sugerido pelo CGSNL (*Committee em Gene Symbolization Nomenclature and Linkage, Rice Genetcs Cooperative*), Rice, 2008, McCOUCH.

Categorias	Classificação	Protocolo padrão	Descrição
Categoria I	Idêntica à proteína do arroz com função conhecida	Identidade > = 98%	Recebe o nome igual ao original
Categoria II	Similar à proteína conhecida	Identidade > = 50%, de uma proteína conhecida	"putativo + nome original do gene"
Categoria III	Proteína com domínio no *InterPro*	Contém domínio *InterPro*	"*InterPro* + nome do domínio da proteína"
Categoria IV	Proteína hipotética conservada	Identidade > = 50%, de uma proteína hipotética	"proteína hipotética conservada"
Categoria V	Proteína hipotética	Se não está em nehuma das categorias	"proteína hipotética"

4.3.1 – Busca por homólogos em outras espécies

A busca de homólogos em outras espécies em estudo (milho e soja) se deu da mesma forma, como acima exemplificado. Assim, para cada gene de *Arabidopsis thaliana* anotado, seu homólogo foi buscado em *Zea mays* e *Glycine max*, quando possível. O mesmo foi feito para cada gene de *Oryza sativa*.

Para os resultados obtidos, somente foram consideradas homólogas as sequências que apresentaram similaridade igual ou maior que 65% (ROST, 1999). O *e-value*, que é o valor (*score*) que indica a probabilidade do alinhamento feito entre duas sequências ser fruto do acaso, foi observado para todos os resultados, pois proporciona uma estimativa dos número de falsos positivos esperados. Quanto menor este valor, maior a probabilidades dos genes serem homólogos e estarem relacionados. Um *e-value* $<10^{-5}$ é um valor moderado e significa que os genes

podem estar relacionados. É importante observar que, quanto maior o banco de dados utilizado na busca, maior é o *e-value* calculado pelo BLAST.

4.4 – Perfil digital de microarranjo dos genes - *GeneVestigator*

Foram realizadas análises de perfis digitais de experimentos de microarranjos dos genes, com a finalidade de corroborar as anotações feitas, à partir da revisão bibliográfica, e confirmar por meio de ferramentas da bioinformática, o fato de que esses genes respondem ao estresse da hipoxia e anoxia.

Desta forma, a ferramenta *Genevestigator* foi utilizada, pois consiste de um banco de alta qualidade para dados de expressão gênica em que por mais de sete anos um grupo de curadores tem coletado dados de milhares de experimentos de microarranjo, controlado sua qualidade, normalizado resultados e anotado cada amostra. E, em paralelo, desenvolveram um sistema sofisticado que simula eficientemente grandes volumes de dados de microarranjo (HRUZ et al., 2008).

Assim, os 64 genes primeiramente anotados para *Arabidopsis thaliana* e 55 genes de *Oryza sativa*, estes últimos selecionados aleatoriamente, foram submetidos ao programa. Além disso, três genes de Ubiquitina foram adicionados a análise, pois se tratam de genes constitutivos da planta e, portanto, não tem a expressão alterada durante seu desenvolvimento ou quando esta é submetida a um estresse.

4.5 – Obtenção das regiões promotoras

De cada gene, uma região de 1,0 Kb (mil pares de bases) *upstream* foi utilizada para as análises. Acredita-se, com base na referência, que neste tamanho de sequência a probabilidade de encontrar resultados para os padrões de ocorrência de elementos *cis* seja maior.

Estas sequências (promotor de 1,0 Kb + gene) foram obtidas utilizando-se a ferramenta de alinhamento local BLAST. Cada sequência consistiu, portanto, do seu promotor (interesse do trabalho) e da região gênica (para que se tenha a certeza de que estamos trabalhando com o promotor correto).

Ainda, para eliminar qualquer possibilidade de erro na obtenção das sequências, os alinhamentos locais foram realizados nos bancos de dados específicos e primários (Figura 19) para cada espécie.

→ *Arabidopsis thaliana*: **TAIR** (*The Arabidopsis Information Resource* - www.arabidopsis.org/).

→ *Oryza sativa* (arroz): **RGAP** (*Rice Genome Anotation Project* - http://rice.plantbiology.msu.edu/)

→ *Zea mays* (milho): **MaizeGDB** (*Maize Genetics and Genomic Database* - http://maizegdb.org/).

→ *Glycine max* (soja): **Phytozome** (http://www.phytozome.net/cgi-bin/gbrowse/soybean/).

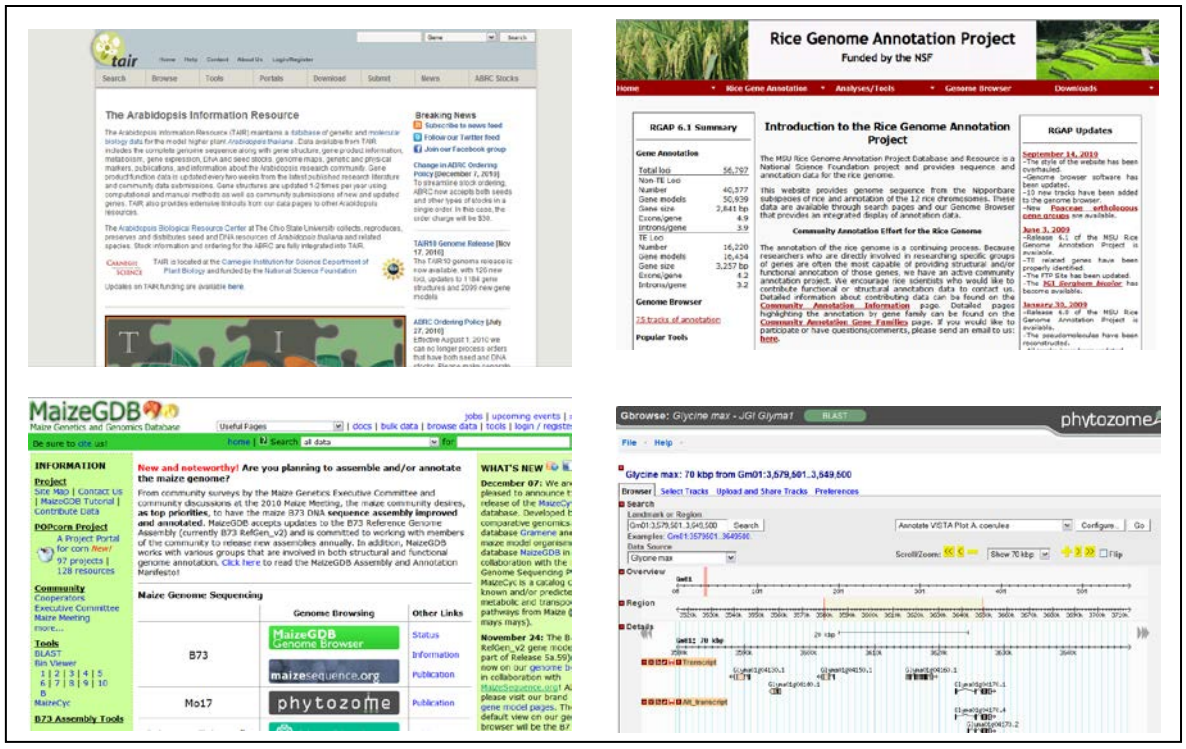


Figura 19. Página inicial dos bancos de dados primários de cada espécie em estudo. Fontes: TAIR, RGAP, MaizeGDB, Phytozome.

Portanto, o procedimento seguiu da seguinte forma:

- A sequência nucleotídica de cada gene foi submetida a uma busca por similaridade utilizando o BLASTx do banco de dados (TAIR para *Arabidopsis thaliana* e RGAP para *Oryza sativa*);
- No resultado, as coordenadas da localização do gene foram alteradas em 1,0 Kb, para se obter a região promotora. Neste passo, desejou-se obter a sequência gênica completa e sua região promotora;
- A fita onde se encontrava o gene foi observada: se fosse na fita *PLUS* (molde), que estava no sentido 5' → 3', bastou voltar para trás 1.000 pares

de bases para obter-se o início da seuqência desejada; se fosse a fita *MINUS* (complementar), que estava no sentido 3' → 5', aumentou-se 1.000 pares de bases, pois o promotor encontrava-se adiante (Figura 20).

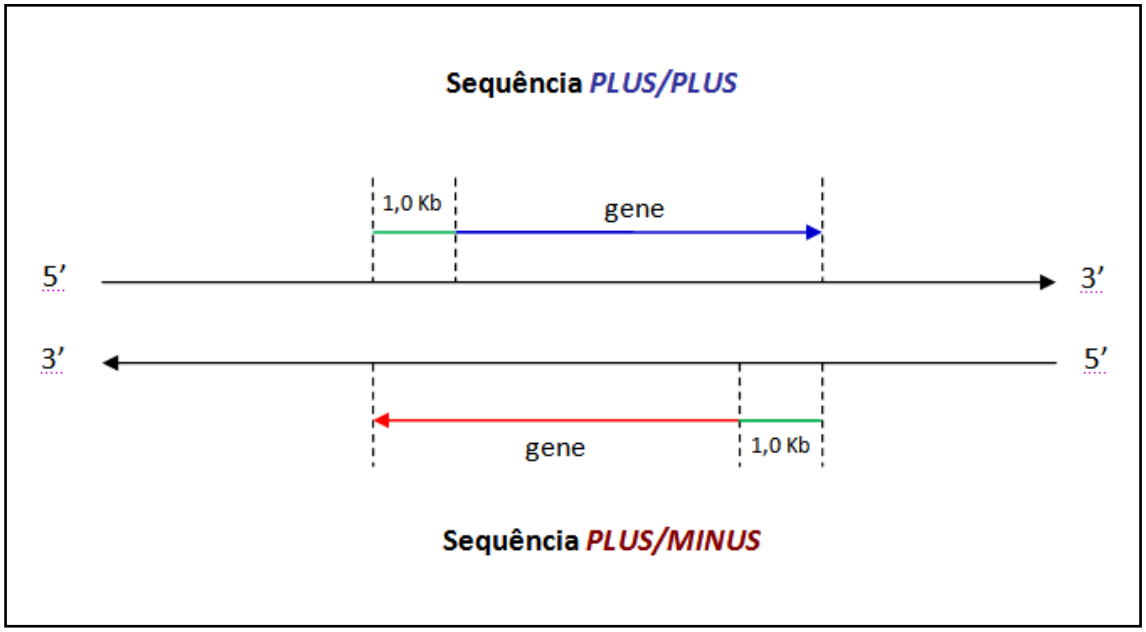


Figura 20. Esquema (sem escala) da localização dos promotores (em verde) nas diferentes fitas do DNA, de acordo com o resultado do BLAST, CGF/UFPel, 2011.

Uma vez que todas as sequências (gene + promotor) foram armazenadas, foi necessário fazer o corte da região promotora, foco das análises sobre os elementos *cis*.

4.5.1 – Cortes das regiões promotoras

Um *script* foi desenvolvidos, em linguagem *Python*, pelo estagiário da Empresa Nacional de Pesquisa Agropecuária – Embrapa Soja, Paulo Silla (Anexo 1).

Nesta etapa, no caso das sequências anotadas como MINUS na etapa anterior, foi necessário obter o reverso complementar da sequência promotora para que todos os promotores alinhassem corretamente na análise de padrões conservados (item 4.6.2). Um esquema deste procedimento é mostrado na Figura 21.

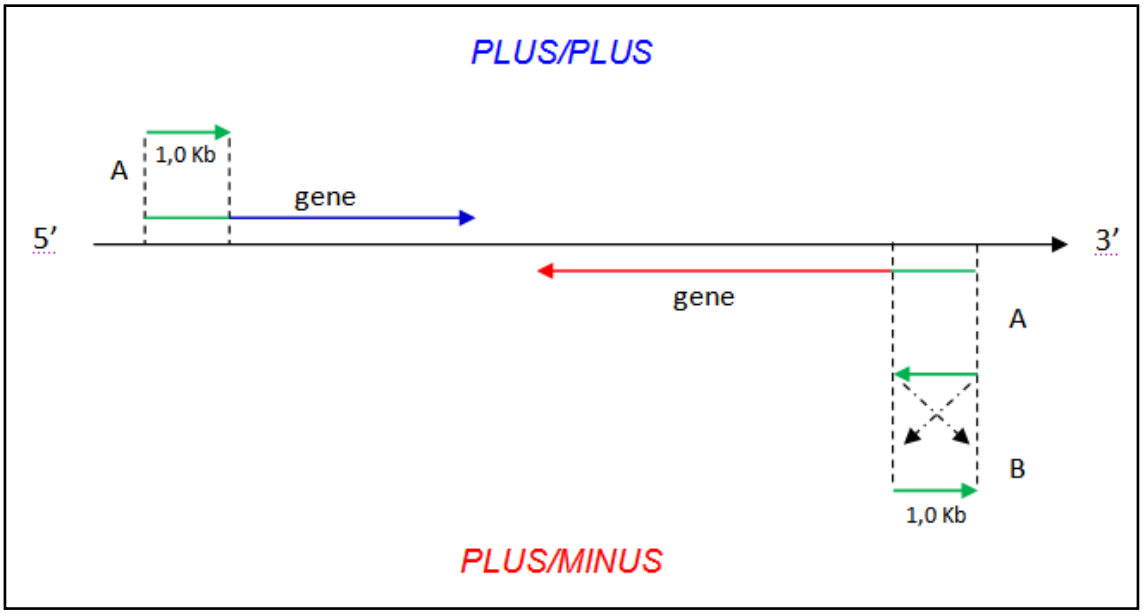


Figura 21. Esquema (sem escala) dos procedimentos para a obtenção dos promotores. Onde, **A** indica o corte das regiões promotoras e **B** a inversão da direção dos promotores obtidos a partir das sequências *reverse*, CGF/UFPel, 2011.

4.6 – Busca por padrões de elementos *cis*

Foram utilizadas duas abordagens no estudo dos elementos *cis*: uma consistiu na submissão de uma sequência promotora ao PLACE, o qual retornou quais elementos *cis* conhecidos e anotados existem naquela região. Outra consistiu da submissão de um grupo de sequências de promotores ao MEME para identificar os motivos conservados. Neste trabalho foram feitas essas duas análises, de modo que a corroboração dos dados pudesse dar mais consistência e singularidade aos resultados obtidos.

4.6.1 – Utilização do banco de dados PLACE para elementos *cis*

Para a obtenção das informações sobre elementos *cis* já relatados e com algum tipo de informação anotada, foi utilizada a plataforma de busca do banco de dados sobre os elementos regulatórios de ação *cis* – PLACE – para plantas vasculares (HIGO et al., 1998).

Para tanto, bastou submeter a sequência de DNA promotora e o resultado foi o número de ocorrências e quais elementos se encontravam em cada promotor (Figura 22).

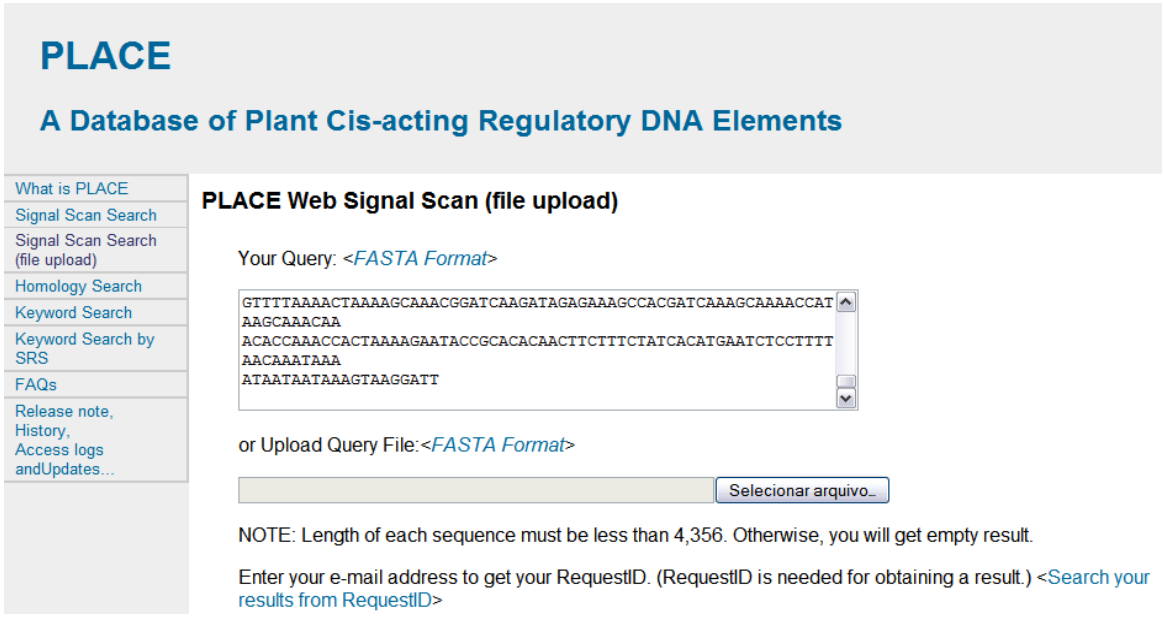


Figura 22. Interface da plataforma PLACE, para o banco de dados sobre elementos cis. Fonte: PLACE.

O banco de dados do PLACE é o mais completo para elementos cis em plantas vasculares, porém sua última atualização foi realizada em fevereiro de 2007. Com o montante de informações geradas pelos projetos de sequenciamento em estudos da Biologia Molecular e Bioinformática, acredita-se que este banco de dados esteja defasado. Também por esse motivo, se fez necessária a busca por novos padrões de ocorrência de sequências curtas, que podem vir a ser reconhecidas como elementos regulatórios de ação *cis*.

4.6.1.1 – Análise dos resultados do PLACE – valor Z

O resultado obtido a partir da busca no banco de dados do PLACE consistiu em uma lista de elementos *cis* preditos para cada promotor submetido a esta análise. Em trabalhos com sequências nucleotídicas muito pequenas (média de 7 pb), estas podem ocorrer de forma aleatória, ou se elas consistem, realmente, alguma importância biológica. Portanto, uma solução foi o cálculo do valor Z (ou *Z score*), para cada elemento *cis* encontrado. Um programa estatístico foi desenvolvido para calcular o valor Z de cada elemento *cis*, partindo do pressuposto (Figura 23):

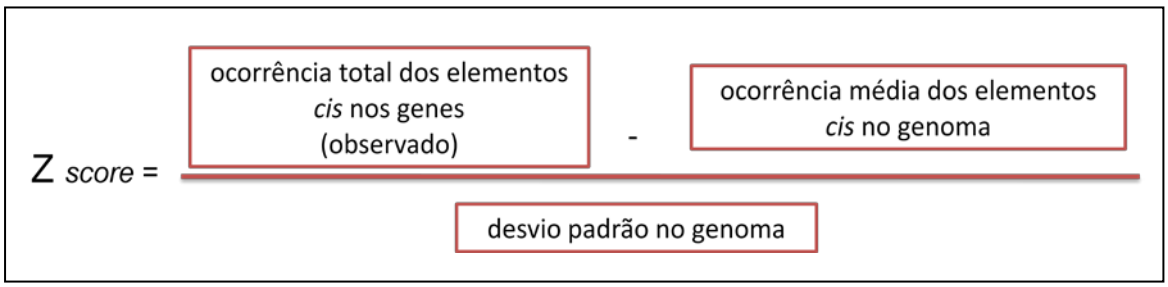


Figura 23. Equação para o cálculo do valor Z, utilizado em cada elemento cis encontrado no banco de dados do PLACE, CGF/UFPel, 2011.

Dessa forma, as probabilidades de ocorrência de todos os elementos *cis* do banco de dados do PLACE foram calculadas, sobre os promotores de todos os genes preditos, no genoma do arroz e de Arabidopsis. Com isso, obteve-se um padrão estatístico de cada elemento *cis* para os promotores desses dois genomas e o desvio padrão do genoma.

Os promotores foram submetidos individualmente no PLACE e as informações sobre os elementos regulatórios para cada sequência foram anotadas em planilha Excel. Após, todos os elementos *cis* terem sido anotados, os mesmos foram submetidos ao programa estatístico para o cálculo do valor Z. Este valor indica a probabilidade do resultado encontrado ser ao acaso. Alguns autores indicam utilizar-se um corte de 0,05 (ou 5%), para eliminar os falso-positivos. Os elementos com valor Z iguais ou maiores que 0,05 foram descartados, pois isto significa que a probabilidade de terem ocorrido ao acaso é igual ou maior que 5%, que consiste em dados não significativos, estatisticamente. Nemhauser et al. (2004), utilizou a mesma estratégia para analisar de elementos *cis* em Arabidopsis, para promotores de genes relacionados com brassinosteróides e auxina.

4.6.2 – Utilização do MEME, uma ferramenta de alinhamento local

Uma análise mais precisa das rotas metabólicas de cada proteína foi realizada com a ajuda do *KEGG (Kyoto Encyclopedia of Genes and Genomes* – http://www.genome.jp/kegg/pathway.html), que é uma fonte de dados, a qual consiste de 16 principais bancos de dados amplamente categorizados em informações de sistemas biológicos, informações genômicas e bioquímicas. Essa ferramenta tem sido muito usada como base de conhecimento para interpretações biológicas em dados de larga escala, gerados pelo sequenciamento de genomas e outras tecnologias experimentais de alto rendimento (KANEHISA et al., 2010).

A partir destas análises de famílias de proteínas e suas rotas, cinco grandes grupos de genes foram formados. São eles:

- Glicólise/Gliconeogênese e Fermentação;
- *Heat shock proteins (HSPs)*;
- Citocromo *P450*;
- Cadeia respiratória (fosforilação oxidativa);
- Outros.

Em cada espécie, portanto, cada gene foi analisado, classificado em uma família protéica (*Pfam*), agrupado em uma das categorias acima e cada grupo foi submetido no MEME.

A suíte MEME consiste em um *software* com um *kit* de ferramentas em uma interface unificada. O algoritmo MEME tem sido muito utilizado para a descoberta de motivos de DNA (BAILEY, 1994). Baseia-se no Modelo Oculto de Markov – HMM (*Hidden Markov Model*), que consiste em um quadro matemático que modela uma série de observações baseado em uma hipótese do processo, porém oculto (Figura 12).

Foram submetidos ao MEME as regiões promotoras dos genes (para este trabalho) que são co-regulados (Figura 24).

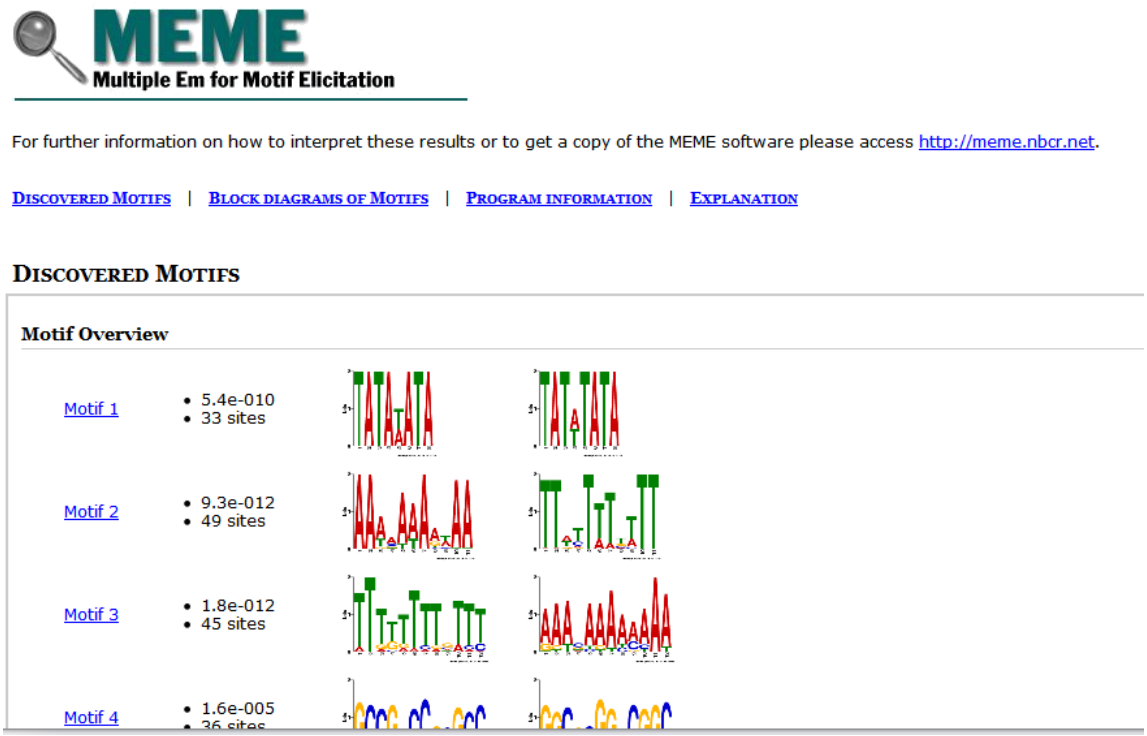


Figura 24. Resultado (vista parcial) do alinhamento de sequências promotoras de *Zea mays* (grupo Fermentação) com os *hits*. Fonte: MEME (http://meme.sdsc.edu/meme4_5_0/intro.html).

O resultado gerado pelo MEME (Figura 24), após a submissão de um grupo de promotores são informações sobre os vários motivos encontrados, como tamanho, *e-value* (probabilidade do resultado ser ao acaso), número de sítios encontrados, localização e número de vezes que cada motivo foi observado.

O *input* (o que se deseja submeter para análise) do MEME permite arquivos com tamanho máximo de 50 Kb (ou seja, 50.000 pares de bases).

Todos os resultados permaneceram disponíveis para acesso, no *link* enviado por *e-mail*, por um período de 30 dias, aproximadamente. Os dados de todos os resultados foram salvos em arquivos de textos.

4.6.2.1 – Análise dos resultados do MEME

Os arquivos foram analisados separadamente e os motivos foram classificados de acordo com:

- Freqüência;
- *Tandem* (repetições em série do motivo);
- Padrões (blocos, duplicados, intercalados).
- Localização no promotor (proximal, medial, distal);

Após análises individuais, os dados foram corroborados para a identificação de possíveis padrões intra e interespecíficos de ocorrência de sequências regulatórias em genes co-regulados, sob anoxia e hipoxia.

5. RESULTADOS E DISCUSSÃO

5.1 – Resultados da busca em literatura por genes responsivos ao alagamento, nas espécies *Arabidopsis thaliana* e *Oryza sativa*

Para as buscas realizadas pela análise de artigos que relatam estudos principalmente com experimentos de microarranjos, da regulação gênica dos genes de *Arabidopsis thaliana* e *Oryza sativa*, quando submetidos ao estresse da hipoxia e anoxia, os resultados foram numerosos. A bibliografia para o estudo do efeito desse estresse abiótico sobre as plantas, e a investigação de quais os mecanismos (fisiológicos, estruturais, genéticos, bioquímicos e moleculares) que a planta faz uso para tolerar a adversidade imposta pelo ambiente (LORETI et al., 2005), é crescente e cada vez mais aprofundada.

Foram anotados, portanto, 64 genes de *Arabidopsis thaliana* que, de alguma forma, tem sua regulação alterada quando expostos ao alagamento e 129 genes para *Oryza sativa*. O dobro de genes encontrados em arroz, em relação a Arabidopsis, não se deu propositalmente, mas, de uma forma coerente, era esperado que mais genes fossem indutíveis pelo estresse nesta planta, que é representante de tolerância ao alagamento e que não tem seu desenvolvimento prejudicado neste tipo de ambiente anóxico (KUDAHETTIGE et al., 2007; JACKSON, 1985).

5.2 – Resultados da busca das sequências nucleotídicas dos genes anotados

Durante a busca pelas sequências nucleotídicas dos genes anotados, algumas sequências não puderam ser obtidas. O acesso pode ser privado devido a falhas de anotação, ou a sequência pode ter sido removida do banco, ou, ainda, a qualidade da sequência pode não ser muito boa, devido a problemas na qualidade do sequenciamento e anotação.

No entanto, este tipo de ocorrência foi pouco freqüente neste trabalho e, de posse de um grande número de genes anotados, os que não puderam ser acessados foram descartados para os próximos passos de análises, com a certeza de que a quantidade de dados que não teríamos acesso, não afetaria o resultado deste estudo. De um total de 64 genes anotados para *Arabidopsis*

thaliana, somente um não foi possível obter sua sequência nucleotídica. Em *Oryza sativa* três genes não tiveram suas informações anotadas, de um total de 129. Todas as sequências foram anotadas no formato *fasta*, que consiste em arquivos de texto simples, em uma linguagem utilizada pela maioria dos programas e ferramentas de alinhamentos, análises e buscas de sequências e informações, na bioinformática (COHEN, 2004).

5.3 – Obtenção das sequências de aminoácidos

A busca das sequências protéicas de cada gene trouxe informações importantes, como o nome da proteína, sua função, local de atuação, em quais processos bioquímicos participa, organismos onde se encontra, tamanho da proteína etc.. Essas informações podem auxiliar na predição e dedução dos mecanismos utilizados pela planta para tolerar um estresse abiótico, por exemplo. Além disso, a posse de uma sequência de aminoácidos (codificada por um determinado gene) permite fazer buscas por homologias em outras espécies.

Assim, todos os genes em estudo tiveram suas sequências de aminoácidos obtidas e, além disso, alguma informação sobre a função da proteína codificada foi anotada (quando disponível), no banco de dados *UniProtKB*.

5.3.1 – Classificação das famílias protéicas e agrupamentos entre funções relacionadas

Para o estudo de padrões de ocorrência de elementos *cis*, é desejável que genes co-regulados que são submetidos a uma análise tenham características em comum. A partir da classificação das famílias protéicas na plataforma *online* do *Pfam*, alguns agrupamentos puderam ser feitos, visando obter resultados mais coerentes e consistentes.

A separação dos genes em cinco grupos (Figura 25) foi realizada com base nas informações retiradas da classificação de suas proteínas, a partir dos bancos de dados *Pfam* e do *KEGG PATAHWAY.*

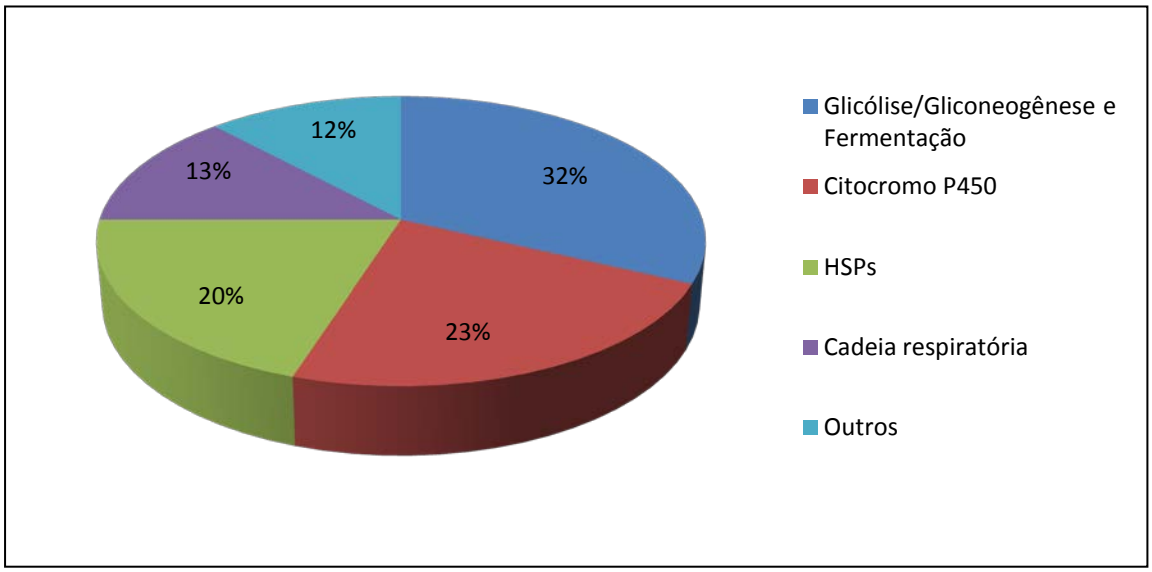


Figura 25. Grupos de genes gerados pela caracterização das famílias protéicas mais freqüentes. Onde: *HSPs* são *as heat shock proteins*; e outros são proteínas de sinalização, transporte e algumas com função desconhecida, CGF/UFPel, 2011.

A partir do nome da proteína e de sua função, banco de dados sobre rotas metabólicas, *KEGG*, foi acessado, para a visualização de cada proteína em seu local de atuação. Dessa maneira, essa ferramenta nos possibilitou situar cada uma delas, quando possível, e caracterizá-la dentro de um grupo de estudos (Figura 25), para posteriores análises.

Os critérios de classificação foram:

- Para glicólise/gliconeogênese e fermentação: proteínas que tinham funções associadas à degradação de carboidratos e quinases, bem como ligadas à fermentação, como álcool desidrogenases, e outras que se localizavam nas rotas que se dirigiam para esse processo;
- Para citocromo P450: proteínas P450;
- Para *HSPs*: proteínas *heat shock* (de choque térmico);
- Para cadeia respiratória: proteínas envolvidas no transporte de oxigênio, carreadoras da membrana da mitocôndria e relacionadas à formação e controle de espécies reativas de O_2;
- Para outros: algumas proteínas transportadoras e outras de ligação. A maioria, porém, sem classificação quanto à sua função.

Levando-se em consideração a função exata de cada proteína, seria possível trabalhar com grupos menores e mais específicos de genes, porém nossa busca por padrões de elementos *cis* com genes que já são ditos, por

bibliografia, como co-regulados e o agrupamento dos mesmos em cinco grandes grupos, nos permite fazer uma abordagem global, porém refinada, no sentido de direcionar, ao menos, um grupo alvo, para posteriores análises com experimentos de validação de prováveis elementos *cis*, em promotores de genes que respondem a anoxia.

5.4 – Busca por homólogos

A homologia entre os genes, em espécies diferentes, traz referências aos mecanismos evolutivos e genéticos da planta. Portanto, no estudo de padrões entre sequências, a obtenção de homólogos se torna interessante, pois nos permite fazer inferências e extrapolações entre as espécies de interesse. Esta busca foi realizada através da ferramenta *BLASTp* para alinhamento de sequências de aminoácidos, e duas sequências foram consideradas homólogas quando sua similaridade foi maior ou igual a 65%, bem como valores *e-value* aceitáveis, na detecção de homologia, de acordo com Rost (1999) e McCouch (2008).

5.4.1 – Resultado da busca de homólogos para *Arabidopsis thaliana*

Durante a busca por homólogos algumas sequências não foram anotadas, devido a baixa similaridade. Observou-se um número considerável de duplicações entre os genomas, onde, para alguns genes diferentes em Arabidopsis, o mesmo homólogo foi observado em arroz (Tabela 2).

Tabela 2. Duplicações observadas entre os genomas de *Arabidopsis thaliana* e *Oryza sativa*, CGF/UFPel, 2011.

Genes de *Arabidopsis thaliana*	Homólogo em *Oryza sativa*
At2g36580 At1g77120 At3g23030	Os01g0360200
At4g25090 At5g47910 At3g27570	Os06g0194900
At2g33380 At1g12240	Os01g0955100
At3g43190 At5g20830	Os02g0579600
At5g07390 At5g20830	Os03g0226200
At1g35580 At2g16060	Os09g0434500

O mesmo aconteceu para os homólogos de soja e milho, em relação ao genoma de Arabidopsis (Tabelas 3 e 4, respectivamente).

Tabela 3. Duplicações observadas entre os genomas de *Arabidopsis thaliana* e *Glycine max*, CGF/UFPel, 2011.

Genes de *Arabidopsis thaliana*	Homólogo em *Glycine max*
At4g25090 At5g47910 At3g27570	Glyma13g17420.1
At2g36580 At3g23030	Glyma06g17030.1
At5g07390 At5g20830	Glyma11g12980.1
At2g34390 At2g19590	Glyma16g01630.1

Spannagl et al. (2010), em estudo entre os genomas de Arabidopsis e culturas de importância agronômica, principalmente, também observou uma quantidade considerável de duplicações no genoma de *Arabidopsis thaliana*, onde, notavelmente, a maioria dos casos esse desequilíbrio é causado por expansões específicas em Arabidopsis. Duplicações em *tandem* (repetitivas) nessa espécie contribuem de forma desproporcional para aumentos nos números de cópias observadas.

Tabela 4. Duplicações observadas entre os genomas de *Arabidopsis thaliana* e *Zea mays*, CGF/UFPel, 2011.

Genes de *Arabidopsis thaliana*	Homólogo em *Zea mays*
At2g17850 At3g27220 At5g54960	si614012d03
At5g07390 At5g20830	LOC541815
At3g02550 At2g14210	pdc3
At2g34390 At2g19590	uaz158(alt)

5.4.2 – Resultado da busca de homólogos para *Oryza sativa*

O mesmo procedimento foi realizado com os genes de arroz e, portanto, para cada gene anotado em *Oryza sativa*, seu homólogo foi buscado nas espécies *Arabidopsis thaliana*, *Glycine max* e *Zea mays.* Sempre respeitando os *scores* (mínimo de 65% de similaridade e *e-value* $<10^{-5}$) sugeridos pela bibliografia, para que o gene fosse considerado homólogo e suas sequências nucleotídica e de aminoácidos pudessem ser anotadas.

O número de duplicações entre estes genomas foi maior do que o observado no item anterior, o que pode indicar que em arroz a quantidade de genes que produzem uma resposta ao ambiente, quando este se encontra anóxico, é maior. O arroz possui mecanismos genéticos e moleculares mais elaborados de tolerância ao alagamento, pois possui um maior número de genes em relação a Arabidopsis (Tabela 5) que tem suas expressões alteradas quando submetidos à submergência.

Tabela 5. Duplicações observadas entre os genomas de *Oryza sativa* e *Arabidopsis thaliana*, CGF/UFPel, 2011.

Genes de *Oryza sativa*	Homólogo em *Arabidopsis thaliana*
Os11g47610 Os11g47560 Os11g47580 Os08g40690 Os11g47510 Os11g47530 Os11g47570 Os11g47550 Os11g47590 Os11g47520 Os07g43820 Os03g40540	At5g24090
Os04g48210 Os04g48200	At1g12740
Os02g54160 Os06g09390 Os09g26420	At1g53910
Os08g39694 Os06g30640 Os06g19070	At2g45550
Os01g43844 Os01g43750 Os01g43720	At3g14690
Os10g30410 Os02g36190 Os10g30380	At5g25140

O mesmo aconteceu para os homólogos de soja e milho, em relação ao genoma de arroz (Tabelas 6 e 7, respectivamente).

Tabela 6. Duplicações observadas entre os genomas de *Oryza sativa* e *Glycine max*, CGF/UFPel, 2011.

Genes de *Oryza sativa*	Homólogo em *Glycine max*
Os02g09400 Os10g30410 Os02g36190 Os02g09240 Os02g09220 Os10g30380	Glyma07g39710.1
Os08g39694 Os02g36110 Os10g08319 Os06g30640	Glyma07g09110.1
Os04g48210 Os04g48200 Os11g18570	Glyma11g02860.1
Os10g40710 Os03g44290	Glyma01g16140.1
Os03g55800 Os02g12690	Glyma04g03740.1
Os11g10480 Os11g10510	Glyma04g39190.1
Os03g04660 Os03g04650	Glyma05g09060.1
Os05g39310 Os01g06660	Glyma07g18570.1
Os05g09500 OsHXK7	Glyma08g03730.1
Os09g35030 Os06g03670	Glyma10g38440.1
Os10g34480 Os02g38290	Glyma11g10640.1
Os01g60770 Os05g39990	Glyma11g26240.1
Os01g63930 Os11g05380	Glyma12g09240.1
Os01g43844 Os01g43750	Glyma13g35230.1
Os04g10160 Os01g72740	Glyma20g00970.1

Tabela 7. Duplicações observadas entre os genomas de *Oryza sativa* e *Zea mays*, CGF/UFPel, 2011.

Genes de *Oryza sativa*	Homólogo em *Zea mays*
Os11g47560 Os11g47550 Os11g47520	LOC100272870
Os03g40540 Os11g18570	LOC100193331
Os04g09920 Os04g10160	LOC100273010
Os02g36190 Os10g30380	LOC100274356
Os08g40690 Os07g43820	LOC100274481
Os02g09240 Os01g72740	LOC100280087
Os02g36110 Os02g09220	LOC100281654

Observa-se uma maior quantidade de duplicações no genoma de arroz (para estes genes), quando comparado aos de Arabidopsis e soja (Tabelas 8 e 9). Isto se deve, provavelmente, a divergência genética entre monocotiledôneas e dicotiledôneas. Porém não se pode afirmar com absoluta certeza que estas duplicações se devem somente a isso. Spannagl et al. (2010) cita que a sintenia esparsa entre mono e dicotiledôneas se deve a frequentes rearranjos, duplicações e perda de genes e que isso diminui fortemente o número de ortólogos detectáveis por conservação posicional.

O que podemos inferir sobre a análise dos homólogos e as respectivas duplicações é que *Oryza sativa* possui uma quantidade maior de genes que respondem ao alagamento em relação às outras espécies aqui estudadas. De fato, o arroz apresenta características morfofisiológicas, bioquímicas e genéticas que dão maior suporte a planta, quando submersa em água. Os eventos de poliploidização podem ter tido conseqüências importantes na evolução das plantas, principalmente na erradicação e adaptação das espécies e para a modulação de suas capacidades funcionais (De Boltd et al., 2005).

Além disso, uma quantidade maior de homólogos foi encontrada entre as duas espécies de dicotiledôneas quando a busca foi realizada a partir das sequências de aminoácidos dos genes de *Arabidopsis thaliana*. Somente três genes em Arabidopsis não tiveram seus homólogos detectados em soja. Contudo, quando a busca foi realizada a partir de *Oryza sativa*, o resultado foi

bem diferente, onde 22 genes em arroz não tendo homólogos encontrados em soja. De acordo com Keller et al. (2000), apesar da sintenia esparsa entre os genomas de arroz e Arabidopsis, a conservação entre essas espécies pode ser útil para a identificação de genes ligados, pois as grandes diferenças existentes entre o tamanho dos genomas das espécies se deve, principalmente, a diferenças de tamanho das regiões de vazio gênico.

Para a espécie *Zea mays* não houve divergências suficientes que pudessem alcançar um nível suficiente para caracterizar diferenças de homologia entre monocotiledôneas e dicotiledôneas.

Em resumo, as quantidades de genes encontrados para cada espécie, descartando-se os que não puderam ter suas sequências nucleotídicas anotadas (como citado no item 5.2 deste trabalho) e outras que não possuíam homólogos, foram:

- *Oryza sativa*: 182 genes;
- *Arabidopsis thaliana*: 155 genes;
- *Glycine max*: 144 genes;
- *Zea mays*: 166 genes.

5.5 – Análise do perfil digital de microarranjo dos genes - *GeneVestigator*

Experimentos de microarranjos de domínio público são muito heterogêneos, dado que a maioria dos dados disponíveis publicamente foram produzidos em uma variedade de plataformas técnicas e diferentes protocolos foram utilizados. Frequentemente faltam informações importantes ou são incompletas e o acesso a informações de qualidade não estão disponíveis. Nesse contexto, o banco de dados *Genevestigator* surge como uma alternativa de análises de níveis de expressões gênicas para várias espécies. É uma ferramenta que possui dados acurados por testes estatísticos e controle de qualidade (HRUZ et al., 2008).

5.5.1 – Perfil digital de microarranjo para os genes de *Arabidopsis thaliana*

O perfil digital de microarranjo obtido para Arabidopsis confirmou os dados obtidos na bibliografia, de que estes genes respondem ao estresse de hipoxia e anoxia (Figuras 26 e 27). O *Genevestigator* possui um banco de

dados de 5.747 experimentos de microarranjos de genes de *Arabidopsis thaliana*, sob várias condições, estágios do desenvolvimento e anatomia.

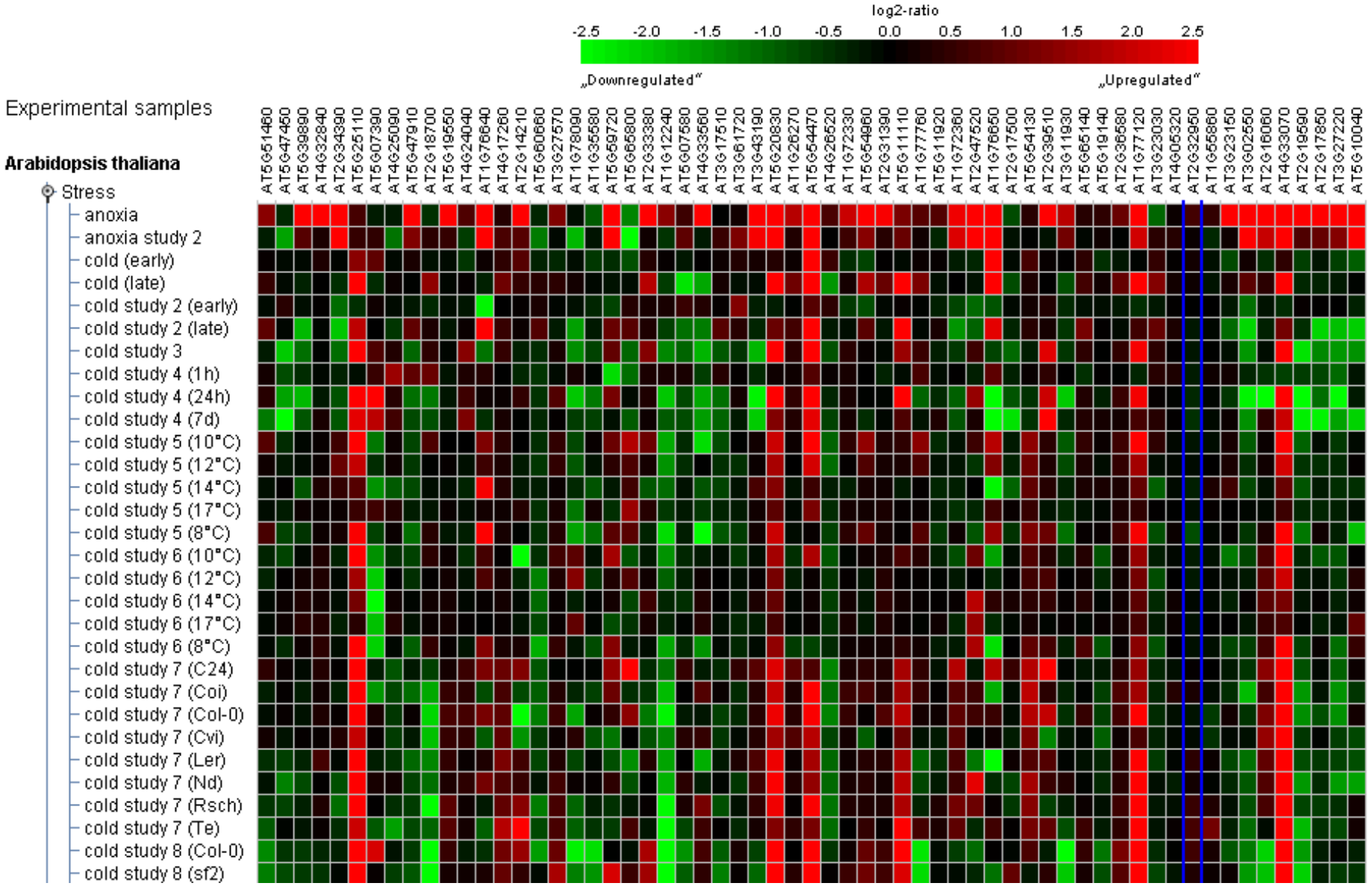


Figura 26. Perfil digital de experimentos de microarranjos em *Arabidopsis thaliana*. O nível de expressão está na razão de $\log_2$, dos genes que tiveram sua expressão suprimida em até 2,5 vezes (verde claro), até os que se mostraram super-expressos em até 2,5 vezes (vermelho). No eixo horizontal se encontram os genes e no eixo vertical as condições de estresse. A coluna destacada em azul e o gene da direita e da esquerda são Ubiquitinas, obtida no Genevestigator, CGF/UFPel, 2011.

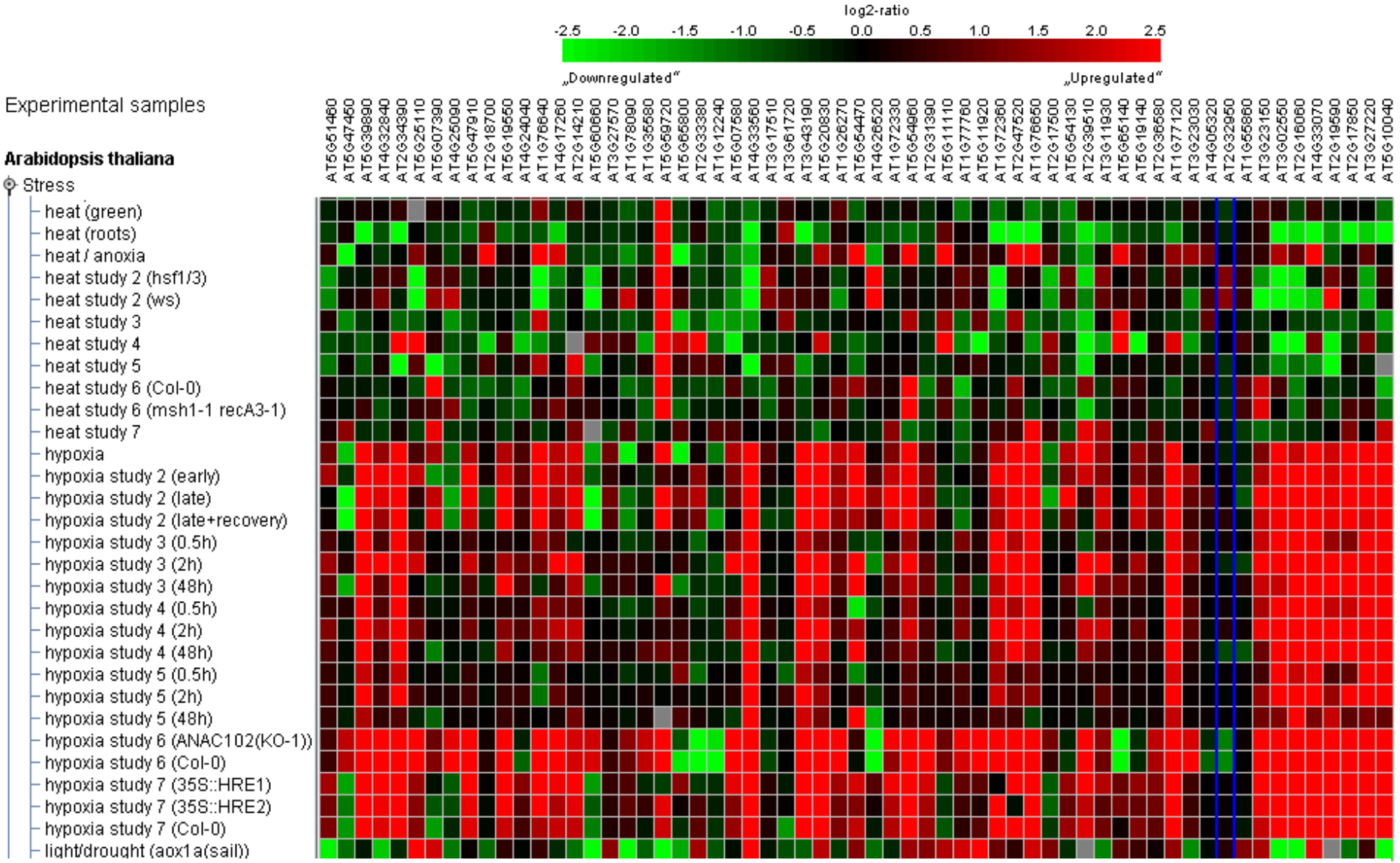


Figura 27. Perfil digital de experimentos de microarranjos em *Arabidopsis thaliana*. O nível de expressão está na razão de $\log_2$, dos genes que tiveram sua expressão suprimida em até 2,5 vezes (verde claro), até os que se mostraram super-expressos em até 2,5 vezes (vermelho). No eixo horizontal se encontram os genes e no eixo vertical as condições de estresse. A coluna destacada em azul e o gene da direita e da esquerda são Ubiquitinas, obtida no Genevestigator, CGF/UFPel, 2011.

A Figura 26 mostra que os genes respondem mais marcadamente ao estresse da anoxia (1ª e 2ª linhas), pois se destacam com a maioria aumentando a expressão, quando comparados a experimentos de plantas expostas ao frio. A Figura 27 mostra 18 tipos de análises com estresse de hipoxia para as quais os genes de Arabidopsis tiveram uma expressão intensa. Destaca, também, o grupo dos outros estresses, como choque térmico, luz e seca. As ubiquitinas (destacadas em linha azul) praticamente não tiveram a expressão alterada, tratando-se de genes constitutivos.

Nos resultados para desenvolvimento, que indicam qual o estágio que o gene é expresso na planta, bem como sua intensidade (em porcentagem de potencial de expressão), o perfil digital mostrou que alguns são expressos durante toda a vida da planta, enquanto outros são expresso em ouças fases ou nenhuma (Figura 28).

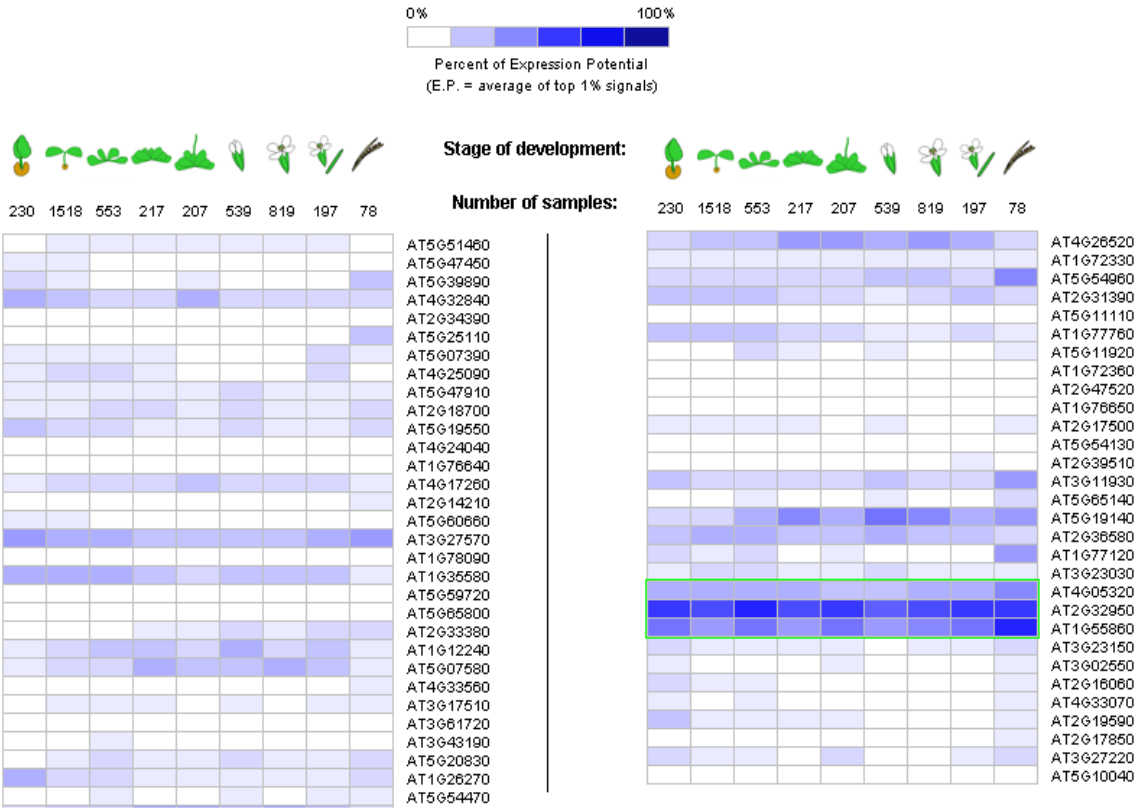


Figura 28. Perfil digital da expressão gênica em *Arabidopsis thaliana* durante seu desenvolvimento. Destacados por linha verde são Ubiquitinas, obtida no Genevestigator, CGF/UFPel, 2011.

Estes dados são úteis para guiar experimentos em busca de genes alvo que respondem a um determinado estresse. Por exemplo, quando um gene é

expresso somente em exposição a uma determinada condição, ele é um gene alvo para estudos, pois, de alguma forma, respondeu à aquela condição, naquele determinado momento. Porém, tratando-se de genes constitutivos, como as ubiquitinas, os quais são expressos durante todas as fases do desenvolvimento da planta, estes não se tornam muito interessantes na busca por genes de tolerância a um estresse.

5.5.2 – Perfil digital de microarranjo para os genes de *Oryza sativa*

O banco de dados do Genevestigator possui um total de 305 experimentos de microarranjos para *Oryza sativa.* Portanto, um volume bem menor de dados sobre esta espécie está disponível. Contudo, o perfil digital revelou um experimento com anoxia e o resultado mostrou que os genes de arroz tem sua expressão fortemente alterada quando expostos a esse tipo de estresse (Figura 29).

Figura 29. Perfil digital de experimentos de microarranjos em *Oryza sativa*. O nível de expressão está na razão de $\log_2$, dos genes que tiveram sua expressão suprimida em até 2,5 vezes (verde claro), até os que se mostraram super-expressos em até 2,5 vezes (vermelho). No eixo horizontal se encontram os genes e no eixo vertical as condições de estresse. A coluna destacada em azul e o gene da direita e da esquerda são Ubiquitinas, obtida no Genevestigator, CGF/UFPel, 2011.

O perfil de expressão dos genes para o desenvolvimento mostrou-se bastante heterogêneo (Figura 30), o que indica que *Oryza sativa* possui mecanismos mais elaborados na tolerância ao alagamento.

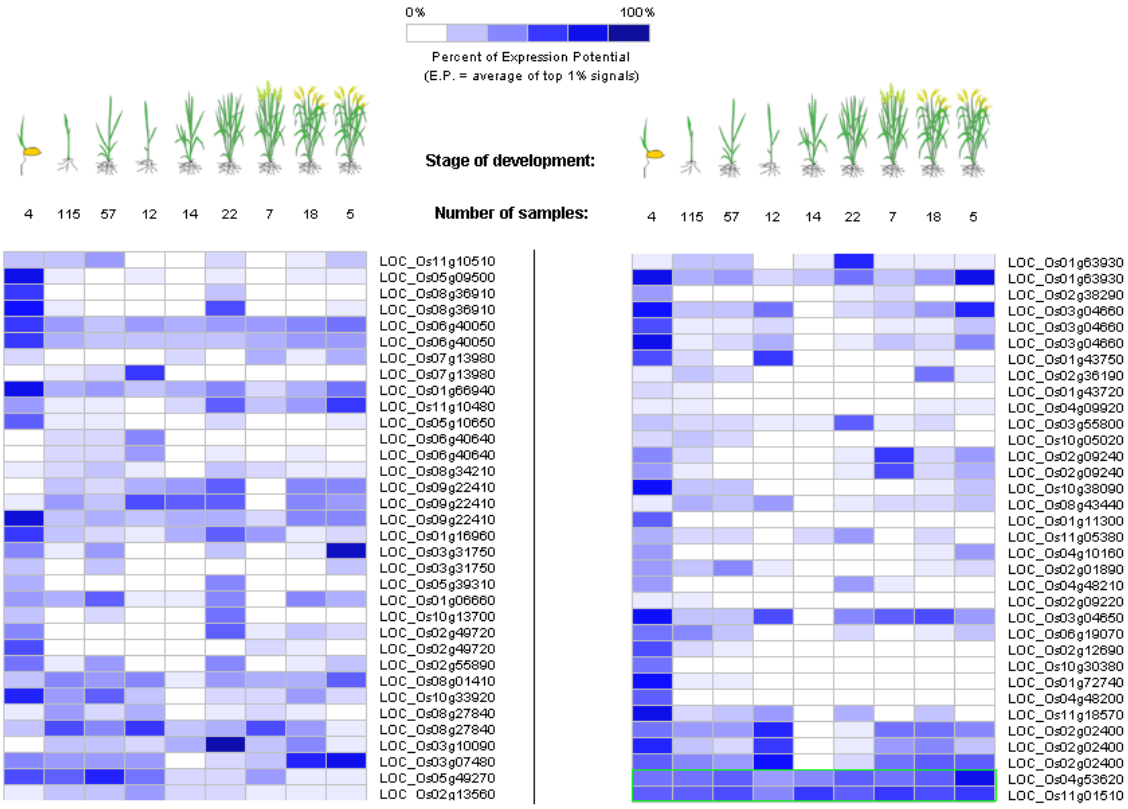


Figura 30. Perfil digital da expressão gênica em *Oryza sativa* durante seu desenvolvimento. Destacados por linha verde são Ubiquitinas, obtida no Genevestigator, CGF/UFPel, 2011.

5.6 – Buscas por padrões de ocorrência de elementos *cis*

Um total de 647 regiões promotoras, prontas para análise e classificadas por espécies, famílias e grupos (exemplificados no item 5.3.1) foram submetidas a dois tipos de análises, para a busca de padrões de elementos *cis*, que possam atuar na regulação dos genes, em resposta ao alagamento.

5.6.1 – Análise dos resultados de elementos *cis* preditos (PLACE)

A partir da submissão de todas as regiões promotoras dos genes anotados neste trabalho para *Arabidopsis thaliana* e *Oryza sativa*, ditas como espécies modelo de estudos para análises de comparações genéticas, na

plataforma do banco de dados PLACE, os elementos cis encontrados em cada promotor foram anotados.

A análise do valor Z de cada elemento *cis* encontrado revelou uma variedade de freqüências de cada elemento regulatório nos promotores. Além disso, considerando-se um valor Z máximo igual a 0,05 alguns padrões puderam ser observados.

5.6.1.1 – *Arabidopsis thaliana*

Foram observadas freqüências bastante distintas entre os elementos *cis* encontrados nos promotores de *Arabidopsis thaliana*. Alguns foram encontrados em apenas um promotor, enquanto que outros foram observados em 38 ou 39 promotores (Tabela 8).

A tabela 9 destaca os 26 elementos *cis* que foram observados em no mínimo 20 promotores. É importante acrescentar que esta tabela representa, unicamente, o número de regiões promotoras em que estes elementos foram encontrados, não se referindo à frequência total dos mesmos. A frequência total leva em consideração o número de vezes que cada elemento *cis* é presente em cada promotor.

Tabela 8. Elementos *cis* mais frequentes para os promotores dos genes que respondem ao estresse do alagamento em *Arabidopsis thaliana*, CGF/UFPel, 2011.

Elemento *cis*	**Número de promotores (Frequência)**
ACGTABOX	39 (25%)
LECPLEACS2	39 (25%)
BOXLCOREDCPAL	38 (24%)
MYBPLANT	35 (22%)
RBCSCONSENSUS	34 (22%)
TATAPVTRNALEU	31 (20%)
MYB2AT	30 (19%)
WBOXNTCHN48	29 (19%)
CURECORECR	28 (18%)
LTRE1HVBLT49	28 (18%)
POLASIG2	27 (17%)
SEF1MOTIF	27 (17%)
SEF4MOTIFGM7S	27 (17%)
SP8BFIBSP8BIB	27 (17%)
BOXIINTPATPB	25 (16%)
SV40COREENHAN	25 (16%)
-10PEHVPSBD	24 (15%)
MYB1AT	24 (15%)
PYRIMIDINEBOXHVEPB1	24 (15%)
TAAAGSTKST1	24 (15%)
CCA1ATLHCB1	23 (14%)
MYB2CONSENSUSAT	22 (14%)
MYB1LEPR	21 (13%)
ACGTATERD1	20 (13%)
CPBCSPOR	20 (13%)
HDZIP2ATATHB2	20 (13%)
HEXMOTIFTAH3H4	20 (13%)

Este tipo de abordagem permite fazermos análises de elementos *cis*, a fim de observar a quantidade de fatores de transcrição que, possivelmente, participam do processo de regulação da expressão gênica sob o contexto do alagamento e ainda, prever os genes que são tidos como peças chave nesse processo. São eles genes cujos promotores contém uma maior quantidade de sítios de ligação de FTs, ou seja, ricos em elementos *cis* (Tabela 9).

Tabela 9. Genes com a maior quantidade de elementos cis em seus promotores, CGF/UFPel, 2011.

Genes	Número de elementos *cis*
At5g51460	27
At1g53310	23
At5g09590	22
At4g34410	21
At2g47520	20
At5g20830	20
At5g20830	20
At5g57260	20
At1g19840	20
At4g33560	20
At2g47520	19
At2g17500	19
At1g77120	19
At5g39890	19
At2g36580	19
At2g36580	19
At2g24270	19
At3g27570	18
At1g68550	18
At2g18700	18
At5g01320	18
At1g72360	18
At3g27220	18
At2g14210	18
At4g12320	18

A partir da classificação das famílias destes genes, pode-se gerar um gráfico (Figura 31), que representa os grupos gênicos mais representativos, em Arabidopsis, e que tem sua regulação alterada, quando expostos ao estresse da anoxia.

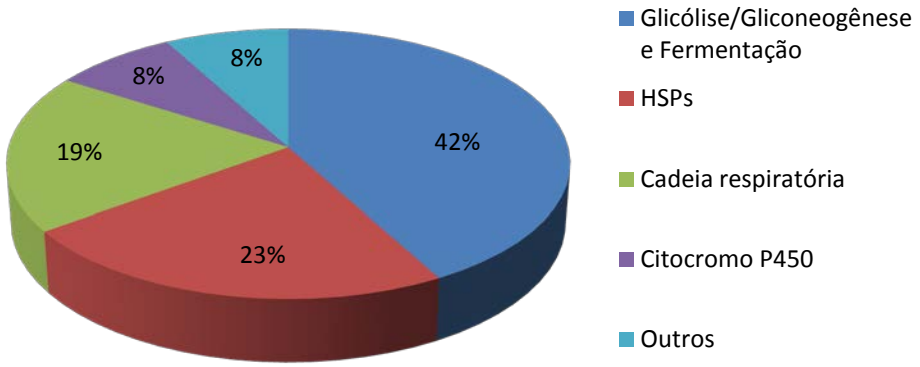


Figura 31. Porcentagem de cada grupo, em relação aos genes listados na Tabela 9. *HSPs* são as *heat shock proteins* Outros são proteínas de sinalização, transporte e algumas com função desconhecida, CGF/UFPel, 2011.

Este resultado indica que estes genes responsivos ao estresse do alagamento (Tabela 12) devem ser mais sensíveis a uma maior quantidade de sinais, pois suas regiões promotoras possuem uma variedade de elementos *cis* e, por isso, participam de uma gama de rotas metabólicas no *cross talk* entre FTs, para promover a alteração na regulação de vários genes e, então, a resposta da planta ao ambiente. Ainda, a figura 31 traz uma idéia de quais são os principais mecanismos que esses mesmos genes participam e que, provavelmente, os processos de glicólise, gliconeogênese, fermentação, as HSPs e proteínas da cadeia respiratória se tratam de pontos chave na descoberta da tolerância à hipoxia/anoxia.

5.6.1.2 – *Oryza sativa*

Os elementos cis encontrados no PLACE para *Oryza sativa* são mais numerosos, em relação à *Arabidopsis thaliana*. Este fato pode indicar uma regulação mais refinada dos genes de arroz para a tolerância ao alagamento, dando aos genes de arroz, uma maior versatilidade na percepção de sinais, *cross talk* entre FTs e na resposta ao ambiente. A Tabela 10 mostra os elementos cis mais frequentes nos genes induzidos pela anoxia, em arroz. Da mesma forma como na Tabela 8, esta tabela representa, unicamente, o número

de regiões promotoras em que estes elementos foram encontrados, não se referindo à frequência total dos mesmos.

Tabela 10. Elementos *cis* mais frequentes para os promotores dos genes que respondem ao estresse do alagamento em *Arabidopsis thaliana*, CGF/UFPel, 2011.

Elemento *cis*	**Número de promotores (Frequência)**
ACGTABOX	53 (29%)
ACGTTBOX	42 (23%)
TATABOX3	42 (23%)
LTRE1HVBLT49	41 (22%)
ACGTATERD1	38 (21%)
HEXMOTIFTAH3H4	38 (21%)
T/GBOXATPIN2	37 (20%)
TGACGTVMAMY	36 (29%)
GT1CORE	34 (19%)
ACGTCBOX	32 (17%)
MYB2AT	32 (17%)
AMYBOX1	31 (17%)
ARFAT	31 (17%)
REBETALGLHCB21	31 (17%)
BS1EGCCR	30 (16%)
CRTDREHVCBF2	30 (16%)
QELEMENTZMZM13	30 (16%)
RHERPATEXPA7	30 (16%)
TATCCAYMOTIFOSRAMY3D	30 (16%)
SEF1MOTIF	29 (16%)
DRE2COREZMRAB17	27 (15%)
SV40COREENHAN	27 (15%)
MYBGAHV	26 (14%)
SP8BFIBSP8BIB	26 (14%)
TAAAGSTKST1	26 (14%)
RYREPEATGMGY2	25 (14%)
TATAPVTRNALEU	25 (14%)
AACACOREOSGLUB1	24 (13%)
CURECORECR	24 (13%)
PYRIMIDINEBOXOSRAMY1A	24 (13%)
ASF1MOTIFCAMV	23 (13%)
IBOXCORENT	23 (13%)
NODCON1GM	22 (12%)
OSE1ROOTNODULE	22 (12%)
ABRERATCAL	21 (11%)
ACGTABREMOTIFA2OSEM	21 (11%)
PRECONSCRHSP70A	21 (11%)
CCAATBOX1	20 (10%)
MYB1LEPR	20 (10%)

Um total de 39 genes de arroz possui pelo menos 20 elementos regulatórios de ação *cis* em seus promotores. Esse número é mais expressivo do que os 26 em Arabidopsis, o que pode indicar que quanto maior a quantidade de elementos *cis*, maior seria a capacidade de tolerar o estresse.

A tabela 11 mostra a lista de genes com maior quantidade de elementos *cis* em seus promotores. Quase o dobro de genes (48) em *Oryza sativa* tem, pelo menos, 18 elementos *cis* em seus promotores em relação à *Arabidopsis thaliana* (25). Este fato reforça a idéia de que o arroz, comparado à Arabidopsis, possui muito mais recursos genéticos para tolerar a hipoxia/anoxia.

Tabela 11. Genes com a maior quantidade de elementos cis em seus promotores, CGF/UFPel, 2011.

Genes	Número de elementos *cis*
Os08g3692	31
Os06g03670	29
Os02g0753000	27
Os07g0485000	27
Os02g36110	26
Os06g30640	25
Os10g38090	24
Os09g31480	24
Os03g60720	23
Os11g10510	22
Os11g47610	22
Os06g0283400	22
Os10g0437500	22
Os11g0147800	22
Os06g10990	22
Os02g12690	22
Os11g47590	21
Os05g10650	20
Os05g49270	20
Os01g21120	20
Os09g22410	20
Os03g0727600	20
Os04g09920	19
Os01g60770	19
Os10g13700	19
Os02g55890	19
Os06g0298200	19

Os03g60120	19
Os08g43210	19
Os03g0226200	19
Os11g47510	18
Os10g30380	18
Os02g54160	18
Os10g05020	18
Os10g0210500	18
Os02g43930	18
Os02g0725100	18
Os04g01140	18
Os08g39690	18
Os03g08490	18
OsHXK	18
Os09g0554300	18
Os03g22170	18
Os01g0360200	18
Os10g0390500	18
Os01g0760600	18
Os03g31750	18
Os02g46640	18

A Figura 32 representa a porcentagem dos grupos destes genes, indicando os principais processos utilizados pela planta na tolerância ao estresse em estudo.

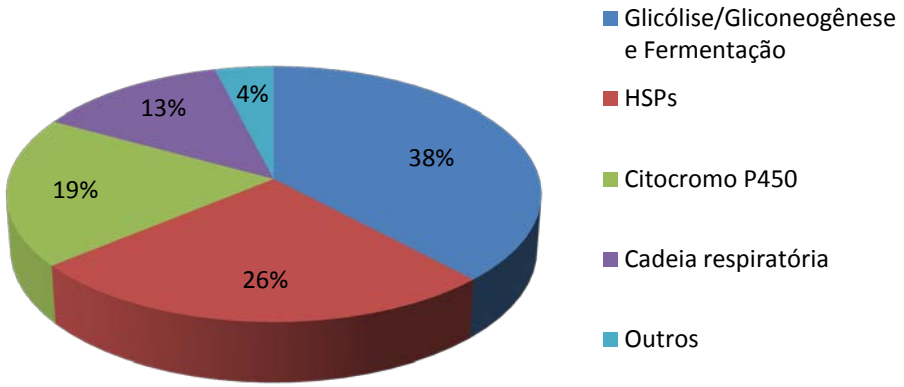


Figura 32. Porcentagem de cada grupo, em relação aos genes listados na tabela 11. *HSPs* são *as heat shock proteins*. Outros são proteínas de sinalização, transporte e algumas com função desconhecida, CGF/UFPel, 2011.

Este resultado condiz com o observado em Arabidopsis, onde os genes relacionados à glicólise, gliconeogênese, fermentação e HSP despontam nos processos mais representativos quanto ao número de elementos regulatórios em seus promotores. Esses genes provavelmente são peças chave no direcionamento da descoberta dos principais genes de tolerância ao alagamento e de um padrão de elementos *cis* que faça a regulação, juntamente com os respectivos FTs, dessa rota de tolerância à hipoxia/anoxia.

Foi possível, a partir da observação das Tabelas 8 e 10, obter os elementos *cis* em comum entre estas duas espécies que se mostraram mais presentes em seus promotores (Tabela12).

Tabela 12. Nomes e sequências dos elementos cis mais comuns entre Arabidopsis thaliana e Oryza sativa, CGF/UFPel, 2011.

Elemento *cis*	**Sequência**
ACGTABOX	TACGTA
ACGTATERD1	ACGT
CURECORECR	GTAC
HEXMOTIFTAH3H4	ACGTCA
LTRE1HVBLT49	CCGAAA
MYB2AT	TAACTG
PYRIMIDINEBOXHVEPB1	TTTTTTCC
SEF1MOTIF	ATATTTAWW
SP8BFIBSP8BIB	TACTATT
SV40COREENHAN	GTGGWWHG
TAAAGSTKST1	TTTATATA
TATAPVTRNALEU	TTTATATA

Estes, portanto, são os elementos *cis* mais representativos, encontrados nos genes que tem suas expressões alteradas quando a planta está em ambiente alagado (anóxico), existentes em ambas as espécies *Arabidopsis thaliana* e *Oryza sativa*. Tratam-se de pontos importantes de estudo para a compreensão dos mecanismos utilizados pela planta, para que a mesma possa se desenvolver sob esse tipo de estresse.

Entre os elementos *cis* com número de ocorrência maior ou igual a 20, a conservação entre os grupos foi analisada. As tabelas 13 e 14 referem-se aos cinco elementos mais conservados e suas ocorrências nos grupos em *Arabidopsis thaliana* e *Oryza sativa*, respectivamente. Nestas tabelas as cores indicam os elementos *cis* conservados entre as espécies e também entre os grupos.

Tabela 13. Elementos cis mais conservados entre os grupos, em *Arabidopsis thaliana*, CGF/UFPel, 2011

Elemento *cis*	Sequência	Grupo	Ocorrência no grupo
MYB2AT	TAACTG	GGF	13
RBCSCONSENSUS	AATCCAA	GGF	11
BOXLCOREDCPAL	ACCWWCC	GGF	10
MYBPLANT	MACCWAMC	GGF	10
SV40COREENHAN	GTGGWWHG	GGF	10
LECPLEACS2	TAAAATAT	P450	12
ACGTABOX	TACGTA	P450	11
MYBPLANT	MACCWAMC	P450	11
BOXLCOREDCPAL	ACCWWCC	P450	10
CURECORECR	GTAC	P450	9
ACGTABOX	TACGTA	HSP	11
TATAPVTRNALEU	TTTATATA	HSP	11
LECPLEACS2	TAAAATAT	HSP	9
SEF1MOTIF	ATATTTAWW	HSP	8
SP8BFIBSP8BIB	TACTATT	HSP	7
LECPLEACS2	TAAAATAT	RC	7
POLASIG2	AATTAAA	RC	6
BOXLCOREDCPAL	ACCWWCC	RC	5
SEF4MOTIFGM7S	RTTTTTR	RC	5
ACGTATERD1	ACGT	RC	5

Tabela 14. Elementos cis mais conservados entre os grupos, em *Oryza sativa*, CGF/UFPel, 2011

Elemento *cis*	Sequência	Grupo	Ocorrência no grupo
ACGTABOX	TACGTA	GGF	16
HEXMOTIFTAH3H4	ACGTCA	GGF	16
T/GBOXATPIN2	AACGTG	GGF	14
CRTDREHVCBF2	GTCGAC	GGF	13
ACGTTBOX	AACGTT	GGF	12
LTRE1HVBLT49	CCGAAA	P450	15
TATABOX3	TATTAAT	P450	14
ACGTABOX	TACGTA	P450	13
SEF1MOTIF	ATATTTAWW	P450	13
ACGTTBOX	AACGTT	P450	12
ACGTCBOX	GACGTC	HSP	14
BS1EGCCR	AGCGGG	HSP	12
ACGTABOX	TACGTA	HSP	9
TATABOX3	TATTAAT	HSP	9
TGACGTVMAMY	TGACGT	HSP	9
ACGTABOX	TACGTA	RC	6
TGACGTVMAMY	TGACGT	RC	6
GT1CORE	GGTTAA	RC	6
ARFAT	TGTCTC	RC	6
ACGTTBOX	AACGTT	RC	5

A partir destes dados é possível observar quais são os elementos *cis* mais conservados dentro de cada grupo, nas espécies. O elemento cis ACGTABOX é observado em ambas as espécies, estando ausente apenas entre os cinco mais frequentes para os grupos GGF e RC de Arabidopsis. Assim, pode se tratar de um importante motivo para o controle da regulação gênica, quando estas plantas estão submetidas ao alagamento.

Outros elementos estão mais representados em apenas uma das espécies, como o MYBPLANT em Arabidopsis e o ACGTTBOX do arroz, ests parecem ser motivos que participam da regulação de genes nas plantas de uma forma mais restrita, ou seja, controlar alguns genes em um família específica de plantas, talvez exclusiva a monocotiledôneas ou dicotiledôneas.

5.6.2 – Análise dos resultados da busca por novos padrões de ocorrência de motivos de DNA (MEME)

Para cada espécie em estudo, suas regiões promotoras foram agrupadas de acordo com a função de cada gene (estudo das famílias gênicas) e, portanto, buscaram-se padrões de ocorrência de motivos de DNA conservados intra e interespecíficos, de acordo com as características de freqüência, repetições, agrupamentos e concentração dos motivos.

5.6.2.1 – *Arabidopsis thaliana*

No grupo dos promotores de Arabidopsis relacionados à glicólise, gliconeogênese e fermentação, dois motivos mostraram-se bastante abundantes, porém sem padrão algum de organização: A[GA]AAAAA[AG]AAAA e [TC]TT[CT][TC]TTTTTCT. Estes promotores continham poucos motivos diferentes (somente 5), o que pode indicar que pertencem a genes que tem a regulação muito específica e que podem responder a poucos sinais.

Referente aos promotores do agrupados como possuindo funções na cadeia respiratória e estresse oxidativo, apesar de abundantes, os motivos também não apresentaram qualquer padrão de ocorrência. Exceto pelo motivo [CT]CAGC[TC]C[GA][TC]C[CG]A que foi observado 14 vezes no gene AAF79618. Desses 14 sítios nesta região promotora, 9 encontram-se em repetidos

seguidamente. Isso pode ser um indicativo de que este gene responde fortemente a algum sinal.

Apesar de não apresentarem padrões de ocorrência, os grupos de promotores associados às *HSPs* e ao citocromo P450 contém uma variedade de motivos. Este dado pode indicar um refinamento na regulação gênica, pois os genes podem ser controlados por um número proporcional de fatores de transcrição (FTs), sendo, assim, sensíveis a um maior número de sinais na rota de transdução.

5.6.2.2 – *Oryza sativa*

Apenas 4 motivos foram apresentados nos promotores dos genes do grupo relacionado à glicólise, gliconeogênese e fermentação. Destes, foi observado que o motivo CG[GC]C[GC][CA][CG]G[GC][GC][GC][GC] apareceu algumas vezes em agrupamentos (sem padrão) de 4 a 7 motivos, separados apenas por alguns pares de bases, nos genes Os05g10650, Os11g47550, Os10g0437500 e Os01g0730300. O motivo [GA][AG]G[AG][GA][GA]G[AG] G[GA]A[GA] foi representado em “duplas”, ou seja, tem duas repetições, em alguns casos. Porém, nada pode ser inferido sobre este caso, devido ao fato de esse mesmo motivo ocorrer, também, de forma aleatória no mesmo grupo.

Para os genes com relação a cadeia respiratória, os promotores mostraram-se bastante abundantes para possíveis sítios de ligação de FTs. Um total de 20 motivos, com diversas distribuições, foi obtido e foram observados promotores com apenas 4 tipos de motivos diferentes (Os03g07480) e outros com até 8 tipos diferentes (Os06g13720). Estes genes parecem fazer parte de vários processos de regulação da expressão em arroz, devido ao fato de apresentarem tantos motivos.

Para as *HSPs* observou-se uma prevalência dos motivos [CG][GC]GCG[GC]C[GC][GC][CT][CG][CG], [GA][GA][AG]GA[GT][GA][GA][AG]GAG, G[GC]CCCAC[ACG][TC]G[TAG]C e GCC[AC]G[TG][GC]CGTG[CG] nos genes Os03g60620, Os09g30418, Os06g50300, Os03g22170, Os07g47790 e Os01g21120, o que indica que estes genes participam, provavelmente e

conjuntamente, de funções similares na planta e que respondem aos mesmos fatores.

Os promotores de genes do citocromo P450 de arroz mostraram um arranjo interessante dos motivos descobertos. O motivo [CG][CG][CG][GC][CG]C[GC][CG]CGC[CG] teve uma distribuição considerável entre os promotores, porém chegou a um número de 12 vezes na região medial/distal do gene Os04g48200. O mesmo pode ser observado para outros motivos que, apesar de bem distribuídos entre os promotores, tiveram alguma concentração em determinados genes, como o motivo [CG][GC][GC] [GAT][GC][GC]CGG[CT]GG que teve 9 repetições na região proximal do promotor do gene Os03g55240. Também o motivo T[CGT]AAAACACT[AT][TA] aparece 6 vezes na porção medial do promotor do gene Os01g63930.

5.6.2.3 – *Glycine Max*

O resultado para os promotores dos genes relacionados à glicólise, gliconeogênese e fermentação não apresentou padrão para motivos. Nos genes Glyma14g01470.1, Glyma16g06390.1, Glyma06g10990.1 e Glyma10g37200.1 observou-se apenas um motivo, o que indica que estes não se constituem de genes-chave, para o processo de resposta a vários sinais da célula/planta.

Para o grupo de promotores relacionados a funções da cadeia respiratória, os motivos encontrados mostraram-se dispersos e sem algum padrão aparente, com exceção do motivo GTCTTGCATTTT que teve uma representação de 10 vezes no gene Glyma08g41350.1. Algumas sequências apresentaram apenas 2 motivos diferentes, como Glyma02g03080.1 e Glyma08g40980.1, contudo outras apresentaram até 10 motivos diferentes, são eles Glyma11g12980.1 e Glyma07g18570.1, indicando que são genes que participam de poucos e muitos mecanismos de regulação, respectivamente.

Não foram observados padrões e/ou características, como as descritas acima, para os grupos das *HSPs* e do citocromo P450.

5.6.2.4 *Zea mays*

Apesar de não apresentar diferenças ou padronização na distribuição dos 4 motivos encontrados para o grupo de glicólise, gliconeogênese e fermentação, o motivo GTCTTGCATTTT foi observado com 6 repetições na

porção distal do promotor do gene U31451 e 5 repetições, também em série, para o gene si614012d03.

Apesar de muitos motivos (20) encontrados para o grupo relacionado à cadeia respiratória, e de estes estarem representados numerosas vezes nos promotores, nenhuma observação que destacasse algum dos itens em análise foram observadas.

No grupo das *HSPs*, o motivo AAGCTCCGGCT[GC] foi observado na porção distal de três promotores (LOC100283876, LOC100283876 e umc1588), com 3 repetições seguidas. Em dois casos (LOC100283876 e LOC100283876), acompanhado por 3 repetições do motivo C[CGA]CGCCGC[CG][GT]C[GC] e na outra (umc1588) pelo motivo GGGGT[GC]G[GC][GT]T[CGA]G. Isto pode indicar que esses três genes dispõem de vários sítios de ligação, para os mesmos FTs, e que estes motivos tem participação importante na regulação destes genes. Segundo Nakashima et al. (2009), um único FT pode controlar a expressão de muitos genes alvo, através da ligação específica do FT ao elemento cis nos promotores dos respectivos genes, é o chamado *regulon.* Assim, um único FT pode controlar a regulação de muitos arranjos de genes que respondem ao estresse, os quais são controlados pelo FT.

Os promotores do grupo de genes do citocromo P450 mostraram alguns padrões interessantes, como: 7 repetições seguidas e sem espaço entre elas do motivo CATACATACAT[AG] no gene LOC100273457, muito semelhante a provável um TATA*box* (TT[AT]TTATTATTA), que teve 6 repetições como esta, no promotor do gene LOC100272865. Os motivos [CA]TTTGC[CT]GAGTG e A[CA]TCGG[CT]AAAG, apesar de serem observados em algumas vezes, estão presentes em conjunto, separados por apenas algumas bases, na região distal dos promotores dos genes cl377_-3 e LOC100272865, com uma representação de no mínimo 3 vezes em cada promotor.

Para todas as espécies estudadas, o grupo “Outros”, que continha regiões promotoras de genes não muito relacionados, como transportadores e outras proteínas de ligação, além de uma variedade com funções não muito esclarecidas, os motivos encontrados não apresentaram qualquer relação de padrões e foram pouco frequentes. Isto enfatiza o fato de que genes co-regulados e com funções semelhantes tem mecanismos de regulação semelhantes e respondem a sinais em comum.

Além das características acima citadas, os motivos podem se destacar pelo número de vezes que foram observados dentro de cada grupo e/ou espécie. As tabelas 15, 16, 17 e 18 resumem os motivos mais frequentes nos promotores dos genes estudados, dentro de cada grupo.

Tabela 15. Distribuição dos motivos mais frequentes no grupo dos promotores relacionados à Glicólise, Gliconeogênese e Fermentação (GGF), obtidos pelos resultados da ferramenta MEME, CGF/UFPel, 2011.

Motivo	Número de ocorrências	Espécie
A[GA]AAAAA[AG]AAAA	46	*Arabidopsis thaliana*
[TC]TT[CT][TC]TTTTTCT	50	*Arabidopsis thaliana*
[CG]TC[TG][TG]TCTCT[TC][CT]	29	*Arabidopsis thaliana*
C[TC]C[CT]CC[ATC][CT][CA]CC	32	*Oryza sativa*
CG[GC]C[GC][CA][CG]G[GC][GC][GC][GC]	**50**	***Oryza sativa***
AAAAAAAAAGAA	46	*Oryza sativa*
[GA][AG]G[AG][GA][GA]G[AG]G[GA]A[GA]	39	*Oryza sativa*
TTTTTTTT[CTA]TTT	44	*Oryza sativa*
C[TC]C[CT]CC[ATC][CT][CA]CC	32	*Glycine max*
[CG]GT[GC]G[GT][AGC][GT][TA]GG[CTG]	23	*Glycine max*
AAAAA[AT]AAAA	48	*Glycine max*
TTTT[TC][TCA]TT[TC]T[TC]T	47	*Glycine max*
AA[AT][AC]A[AT]A[AG][TA]AA	49	*Zea mays*
TTT[TG]TTTT[GC]TTT	45	*Zea mays*
GC[CG]G[GC][CG]C[GT][CG]G[CT]C	**36**	***Zea mays***
GG[CT]G[GA][CA]G[GA]CG[GA][CG]	**36**	***Zea mays***

Há uma forte semelhança entre os motivos CG[GC]C[GC][CA][CG]G[GC][GC][GC][GC], GC[CG]G[GC][CG]C[GT][CG]G[CT]C e GG[CT]G[GA][CA]G[GA]CG[GA][CG] (Tabela 15), onde pode-se gerar um motivo consenso, como: **[GC][GC]C[GC]GC[GC]GCG[GC][CG]**, gerado a partir da combinação dos três anteriores, presentes em *Oryza sativa* e *Zea mays*. Esse

provável elemento de DNA pode estar relacionado a mecanismos de regulação da expressão gênica dos genes que foram encontrados.

Tabela 16. Distribuição dos motivos mais frequentes no grupo dos promotores relacionados à Cadeia Respiratória (CR), obtidos pelos resultados da ferramenta MEME, CGF/UFPel, 2011.

Motivo	Número de ocorrências	Espécie
[GC][GTA][TC]T[CG][CT]C[TG][CT]CT	20	*Arabidopsis thaliana*
CG[CG][GC][GC][CG]C[GC]GCG	**27**	***Oryza sativa***
[TC]TTTTTTTT[CAT][TG]T	26	*Oryza sativa*
[AT]AT[AT]AAAA[AG]AA	27	*Oryza sativa*
[AC]G[AG][AG]G[GAC][GA][GA]AG[AG][GA]	25	*Oryza sativa*
AAAA[AG][AG]AA[AG][CA]A	20	*Glycine max*
[CT]T[CT][TC]C[TC][TC][TC][CT][TC]C[CT]	32	*Glycine max*
GAA[AC]AAA[AG]AAA	20	*Glycine max*
G[GC]CG[CG]CGGC[GC][GCA]	**33**	***Zea mays***
T[TA][AGTCC]TTT[TC]T[TC]TT	30	*Zea mays*
AAA[AGC]AA[AC]A	24	*Zea mays*
CCT[TG][CG][CG][CT]CTCC	22	*Zea mays*

Por serem bastante semelhantes, os motivos CG[CG][GC][GC][CG]C[GC]GCG e G[GC]CG[CG]CGGC[GC][GCA], relacionado à cadeia respiratória e encontrados nas espécies *Oryza sativa* e *Zea mays* (em vermelho na Tabela 16), podem se tratar do mesmo motivo – **[CG]GCG[GC]C[CG]G[GC]CG** – e terem a mesma função em relação à sua funções na ligação dos FTs. Os FTs podem se ligar a uma variedade de sítios de ligação, dependendo do sinal recebido e da resposta que a planta deve ter ao ambiente. Porém como se tratam de novos motivos descobertos, nenhuma referência quanto aos processos que possivelmente exercem pode ser feita até o momento.

Priest et al. (2009), citam que fundamentalmente, a regulação transcricional da expressão gênica em eucariotos é mediada pelo recrutamento

de FTs aos elementos regulatórios *cis*. FTs interagem com elementos de DNA específicos, com outros FTs e com a maquinaria basal para regular a expressão de genes alvo. Em plantas, a regulação transcricional é mediada por mais de 1.500 FTs e cada FT controla a expressão de milhares de genes alvo em redes de sinalização complexas (REICHMANN et al., 2000)

Tabela 17. Distribuição dos motivos mais frequentes no grupo dos promotores relacionados as Heat Shock Proteins (HSP), obtidos pelos resultados da ferramenta MEME, CGF/UFPel, 2011.

Motivo	**Número de ocorrências**	**Espécie**
A[GA]A[AG][AC][AC]AAAA[AC]A	45	*Arabidopsis thaliana*
[TG][TC]T[TC]TT[CT]T[TC]TTT	48	*Arabidopsis thaliana*
[CG][GC]GCG[GC]C[GC][GC][CT][CG][CG]	50	*Oryza sativa*
T[CT][TC]TTTTTTTTT	34	*Oryza sativa*
[TA]AAA[AT][AT]AAA[GT]A	41	*Oryza sativa*
[GA][GA][AG]GA[GT][GA][GA][AG]GAG	43	*Oryza sativa*
C[CGA]CGCCGC[CG][GT]C[GC]	29	*Zea mays*
C[TG]GC[GT]GC[GA]GC[GA]	29	*Zea mays*
[TC]C[TC][TC][TC]TT[CT]TT[TC]	29	*Zea mays*
AA[AG]A[AC][AG][AC][ATG][AC]AAA	39	*Zea mays*

Tabela 18. Distribuição dos motivos mais frequentes no grupo dos promotores relacionados ao citocromo P450 (P450), obtidos pelos resultados da ferramenta MEME, CGF/UFPel, 2011.

Motivo	Número de ocorrências	Espécie
[TC]TT[TGC]T[TC]T[CG]TT[TC]	37	*Arabidopsis thaliana*
[CG][CG][CG][GC][CG]C[GC][CG]CGC[CG]	50	*Oryza sativa*
C[CT][CT][TC][CG][CT]CTC[CT][CT][CT]	50	*Oryza sativa*
[CG][GC][GC][GAT][GC][GC]CGG[CT]GG	41	*Oryza sativa*
T[TA]TTTTTT	43	*Oryza sativa*
[AT][TA]TAAAAAA[AT][AT][AT]	50	*Oryza sativa*
A[AG][AG][AG][GA]AAG[AG]AAA	26	*Glycine max*
[GC][CG][CG]CG[CG]C[AG]CG[GC]C	27	*Zea mays*

Não foram encontrados motivos em comum, referente aos grupos HSP e P450, entre as espécies. Apesar dos genes serem, na sua maioria, homólogos, a expressão dos mesmos, segundo os dados obtidos, parece depender de mecanismos isolados de regulação, ou seja, mais específicos dentro de cada espécie.

6. CONCLUSÃO

A busca na literatura por genes responsivos ao estresse abiótico do alagamento mostrou uma maior quantidade de genes (dobro) em *Oryza sativa*, em ralação aos encontrados em *Arabidopsis thaliana*. Isto pode indicar que o arroz possui uma maquinaria genética mais especializada na tolerância a essa adversidade imposta pelo ambiente.

Um total de 12 elementos *cis* do banco de dados PLACE são comuns às duas espécies, entre os mais frequentes. A maioria dos genes que possuem um grande número de elementos *cis* se encaixa nos grupos com mecanismos ligados à glicólise, gliconeogênese, fermentação e as *HSPs*.

A análise dos elementos *cis* mais frequentes entre os grupos, revelou uma certa conservação, como o elemento ACGTABOX, que está presente nas espécies *Arabidopsis thaliana* e *Oryza sativa*. Estes dados podem revelar mecanismos importantes da regulação gênica para genes co-regulados pelo estresse do alagamento nas plantas.

Com a utilização do programa MEME, dois motivos consenso puderam ser observados entre as espécies *Oryza sativa* e *Zea mays*, são eles: **[GC][GC]C[GC]GC[GC]GCG[GC][CG]** e **[CG]GCG[GC]C[CG]G[GC]CG.** Sendo o primeiro relacionado a processos de glicólise, gliconeogênese e fermentação e o segundo da cadeia respiratória. Podem se tratar de novos elementos *cis,* pois tem ocorrências relativamente altas, nos promotores dos genes.

As análises aqui realizadas são de importância para estudos posteriores que se referem ao estresse do alagamento e descobertas de mecanismos moleculares de tolerância a esse ambiente típico de planícies alagadas. A partir dos dados gerados, é possível guiar experimentos com transformação genética com genes alvo e elementos *cis* preditos como importantes e frequentes a fim de tentar conferir alguma característica às plantas, como as existentes no arroz, para que possam se desenvolver em ambiente com privação de O_2.

Os estudos *in silico*, bem como a utilização das ferramentas disponíveis são, portanto, de grande utilidade para a comunidade científica, onde uma grande quantidade de dados é gerada a cada dia e os mesmo precisam ser tratados, classificados e organizados, para que possamos fazer inferências sobre funções biológicas com mais precisão, cada vez mais.

7. REFERÊNCIAS

BAILEY, T.L.; ELKAN, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. **Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology**, pp. 28-36, 1994.

BAILEY, T. L.; BODEN, M.; BUSKE, F.A.; FRITH, M.; GRANT, C.E.; CLEMENTI, L.; REN, J.; LI, W. W.; NOBLE, W. S. MEME Suite: tools for motif discovery and searching. **Nucleid Acids**, n.37, p. 202-208, 2009.

BARAKAT, A.; MATASSI, G.; BERNARDI, G. Distribution of genes in the genome of Arabidopsis thaliana and its implications for the genome organization of plants**. Proceedings of the National Academy of Sciences**, v.95, p.10044-10049, 1998.

BARNI, N. A.; LOPES, M. S.; HILGERT, E. R.; SARTORI, G.; GONÇALVES, J. C.; GOMES, J. E. da S. Avaliação agronômica de cultivares de soja (*Glycine max* (L.) Merril) em solos hidromórficos. **Agronomia Sul-riograndense**, v.21, n.2, p.189-207, 1985.

BARTELS, D.; SUNKAR, R. Drought and salt tolerance in plants. **Critical Reviews in Plant Science**, v.24, p.23-58, 1995.

BENNETZEN, J.L. The rice genome. Opening the door to comparative plant biology. **Science**, v.296, n.5565, p.60-63, 2002.

CHEN, M.; SAN MIGUEL, P.; COSTA DE OLIVEIRA, A.; WOO, S-S.; ZHANG, H.; WING, R.A.; BENNETZEN, J.L. Microcolinearity on *sh2*-homologous regions of the maize, rice, and sorghum genomes. **Proceedings of the National Academy of Sciences, Plant Biology,** v. 94, pp.3431-3435, 1997.

CHOI, H-K.; MUN, J-H.; KIM, D-J.; ZHU, H.; BAEK, J-M.; MUDGE, J.; ROE, B.; ELLIS, N.; DOYLE. J.; KISS, G.B.; YOUNG, N.D.; COOK, D.R. Estimating

genome conservation between crop and model legumes species**. Proceedings of the National Academy of Sciences**, v.101, n.43, p.15289-15294, 2004.

CHOWDHARY, R.; WONG, L.; BAJIC, V. B. Finding functional promoter motifs by computacional methods: a word of caution. **International Journal Bioinformatics Research and Applications**, v.2, n.3, p.282-288, 2006.

COHEN, J. Bioinformatics – An introduction of computer scientists. **ACM Computing Surveys**, v.36, n.2, pp. 122-158, 2004.

COSTA, A.C.P.B.; MACÊDO, F.-S.; GREGORY, H. Agronegócio Brasileiro. Características, Desempenho, Produtos e Mercados. **FIESP**, 2008.

DE BODT, S.; MAERE, S.; VAN DE PEER, Y. Genome duplication and the origin of angiosperms. **Trends in Ecololgy & Evolution,** v.20, pp.591-597, 2005.

DELSENY, M. Re-evaluating the relevance of ancestral shared synteny as a tool for crop improvement. **Current Opinion in Plant Biology**, v.27, p.126-131, 2004.

DEVOS, K.M.; BEALESM J.; NAGAMURA, Y.; SASAKI, T. Arabidopsis-rice: Will colinearity allow gene prediction across the eudicot-monocot divide? **Genome Research**, v.9, p.825-829, 1999.

DEVOS, K.N.; GALE, M. Genome Relationships: the grass model in current research. **The Plant Cell**, v.12, p.637-646, 2000.

DEVOS, K. Updating the "Crop Circle." **Current Opinion in Plant Biology**, v.8, n.2, p.155-162, 2005.

EMBRAPA ARROZ E FEIJÃO – Centro Nacional de Pesquisa em Arroz e Feijão. Origem e História do Arroz. Disponível em: http://www.cnpaf.embrapa.br/arroz/historia.htm Acesso em: 17 dez. 2010.

EMBRAPA INFORMÁTICA AGROPECUÁRIA – Centro Nacional de Pesquisa em Informática Agropecuária. Cultivo do arroz irrigado no Brasil. Disponível em: http://sistemasdeproducao.cnptia.embrapa.br/FontesHTML/Arroz/ArrozIrrigadoBrasil/index.htm Acesso em: 17 dez. 2010.

ENTREZ GENE. Genes and mapped phenotypes. Disponível em: http://www.ncbi.nlm.nih.gov/gene Acesso em: durante todo o período de desenvolvimento do projeto.

FAOSTAT – Food and Agriculture Organization of the United States. Disponível em: http://faostat.fao.org Acesso em: 18 e 19 dez. 2010.

FORCE, A.; LYNCH, M.; PICKETT, F.B.; AMORES, A.; YAN, Y-L, POSTLETHWAIT, J.H. Preservation of duplicate genes by complementary, degenerative mutations. **Genetics**, v.151, p.1531-1545, 1999.

FUKAO, T.; XU, K.; RONALD, P. C.; SERRES, J. B. A variable cluster of ethylene response factor-like genes regulates metabolic and developmental acclimation responses to submergence in rice. **The Plat Cell**, v.18, p.2021-2934, 2006.

GASTAL, M. F. da C.; BRANCÃO, N.; VERNETTI, F. de J. Indicação de cultivares de soja para terras baixas. **Agropecuária de Clima Temperaro**, v.1, n.1, p.95-99, 1998.

GENEVESTIGATOR. Disponível em: https://www.genevestigator.com/gv/index.jsp Acesso em: set. 2010.

GOMES, A. S., MAGALHÃES JÚNIOR, A. M. Arroz irrigado – Brasil. **Embrapa Informação Tecnológica**, 2004.

GOMES, A. S.; SILVA, C. A. S.; PARFITT, J. M. B.; PAULETTO, E. A.; PINTO, L. F. S. Caracterização de indicadores de qualidade de solo, com ênfase às

áreas de várzea do Rio Grande do sul. **Embrapa Clima Temperado Documentos**, v.169, p40, 2006.

GREENWOOD, D.J. The effect of oxygen concentration on the decomposition of organic materials in soil. **Plant and soil**, The Hague, v.14, p.360-376, 1961.

GRUNDY, W.N.; BAILEY, T.L.; ELKAN, C.P.; MICHAEL, E.B. Meta-MEME: Motif-based hidden Markov models of protein Families. **Computer Applications in the Biosciences**, 1997.

GUYOT, R.; YAHIAOUI, N.; FEUILLET, C.; KELLER, B. In silico comparative analysis reveals a mosaic conservation of gens within a novel collinear region of wheat chromosomes 1AS and rice chromosome 5S. **Functional & Integrative Genomics**, v.4, p.47-58, 2004.

HELENTJARIS, T.; WEBER, D.; WRIGHT, S. Identifications of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polimorphisms. **Genetics**, v.118, p.185-188, 1988.

HIGO, K.; UGAWA, Y.; IWAMOTO, M; HIGO, H. PLACE: a database of plant *cis*-acting regulatory DNA elements. **Nuclieic Acids Research**, v. 26, n.1, 1998.

HOUDE, M.; BELCAID, M.; OUELLET, F.; Danyluk, J.; Monroy, A. F.; Dryanova, A.; Gulik, P.; Berferon, A.; Laroche, A.; Links, M. G.; MacCarthy, L.; Crosby, W. HRUZ, T.; LAULE, O.; WESSENDORP, F.; BLEULER, S.; OERTLE, L.; WIDMAYER, P.; GRUISSEM, W.; ZIMMERMANN, P. Genevestigator V3: a reference expression database for the meta analysis of transcriptomes. **Advances in Bioinformatics**, 420747, 2008.

L.; Sarhan, F. Wheat EST resources for functional genomics of abiotic stress. **BMC Genomics**, v.7, p.149, 2006.

ILIC, K.; SANMIGUEL, P.J.; BENNETZEN, J.L. A complex history of rearrangement in a orthologous region of maize, sorghum, and rice genomes.

Proceedings of the National Academy of Sciences, v. 100, p.12265-12270, 2003.

JACKSON, M.B. Ethylene and responses of plants to soil waterlogging and submergence. **Annual Review of Plant Physiology and Plant Molecular Biology**, v.36, p.145-174, 1985.

KANEHISA, M.; GOTO, S.; FURUMICHI, M.; TANABE, M.; HIRAKAWA, M. .; KEGG for representation and analysis of molecular networks involving diseases and drugs. **Nucleic Acids Researches**. v.38, pp.355-360, 2010.

KEGG PATHWAY Database. Disponível em: http://www.genome.jp/kegg/pathway.html Acesso em: jun. 2010.

KELLER, B.; FEUILLET, C. Colinearity and gene density in grass genomes. **Trends in Plant Science Reviews**, v. 5, n. 6, p.246-251, 2000;

KLOK, E. J.; WILSON, I. W.; WILSON, D.; CHAPMAN, S. C.; EEING, R. M.; SOMERVILLE, S. C.; PEACOCK, W. J.; DOLFERUS, R.; DENNIS, E. S. Expression profile analysis of the low-oxygen response in Arabidopsis root cultures. **The Plant Cell**, v.14, p.2481-2494, 2002.

KUDAHETTIGE, R. L.; MAGNESCHI, L.; LORETI, E.; GONZALI, S.; LICAUSI, F.; NOVI, G.; BERETTA, O.; VITULLI, F.; ALPI, A.; PERATA, P. Transcript profiling of the anoxic rice coleoptiles. **Plant Physiology**, v.144, p.218-231, 2007.

LACKY, J.A. Chromosome numbers in the *Phaseoleae* (Fabaceae: Faboideae) and their relation to taxonomy. **American Journal of Botany**, v.67, p. 595-602, 1980.

LESCOT, M.; DÉHAIS, P.; THIJS, G.; MARCHAL, K.; MOREAU, Y.; PEER, Y. V.; ROUZÉ, P.; ROUMBAUTS, S. PlantCARE, a databade of *cis*-acting

regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. **Nucleid Acids Research**, v.30, n.1, p.325-327, 2002.

LEVINE, M.; TIJAN, R. Transcription regulation and animal diversity. **Nature**, v.424, p.147-151, 2003.

LEVINE, M.; DAVIDSON, E.H. Gene regulatory networks for development. **Proceedings of the National Academy of Sciences**, v.102, p.4936-4942, 2005.

LINDSAY, W.L. **Chemical equilibria in soils**, 449p., 1979.

LIU, F.; VANTOAI, T.; MOY, L. P.; BOCK, G.; LIDORF, L. D.; QUACKENBUSH J. Global transcription profiling reveals comprehensive insights into hypoxic response in Arabidopsis. **Plant Physiology**, v.137, p.1115-1129, 2005.

LORETI, E.; POGGI, A.; NOVI, G.; ALPI, A.; PERATA, P. A genome-wide analysis of the effects of sucrose on gene expression in Arabidopsis seedlings under anoxia. **Plant Physiology**, v.137, p.1130-1138, 2005.

LU, F.; AMMIRAJU, J.S.S.; SANYAL, A.; ZHANG,S.; SONG, R.; CHEN, J.; LI, C.; SUI, Y.; SONG, X.; CHENG, Z.; COSTA DE OLIVEIRA, A.; BENNETZEN, J.L.; JACKSON, S.A.; WING, R.A.; CHEN, M. Comparative sequence analysis of *MONOCULM1*-orthologous regions in 14 *Oryza* genomes. **Proceedings of the National Academy of Sciences,** v.106, n.6, p.2071-2076, 2009.

LU, Y.; LAST, R.L. Web-based Arabidopsis functional and structural genomics resources. **The American Society of Plant Biologis**ts, doi: 10.1199/tab.0118, 2008.

MAIZEDB – Maize Genetics and Genomics Database. Disponível em: http://www.maizegdb.org/ Acesso em: mar. a mai. 2010 e 11 jan. 2011.

MATSON, G. A.; EVANS, S. K.; GREEN, M. R. Transceiptional regulatory elements in the human genome. **Annual Review Genomics Human Genetics**, v.7, p.29-59, 2009.

McCOUCH, S.R. Gene Nomenclature System for Rice. **Rice**, v.1, p.72-84, 2008.

MICHAEL, T.P.; MOCKLER, T.C.; BRETON, G.; McENTEE, C.; BYER, A.; TROUT, J.D.; HAZEN, S.P.; SHEN, R.; PRIEST, H.D.; SULLIVAN, C.M. *et al*.. Network discovery pipeline elucidates conserved time of day specific *cis*-regulatory modules. **PLoS Genetics**, v.4, p.14, 2008.

MORAES, J.F.V.; FREIRE, C.J.S. Variação do pH, da condutividade elétrica e da disponibilidade dos nutrientes nitrogênio, fósforo, potássio, cálcio e magnésio, em quatro solos submetidos a inundação. **Pesquisa Agropecuária Brasileira**, v.9, n.9, p.35-43, 1974.

NAKASHIMA, K.; YUSUKE, I.; YAMAGUCHI-SHINOZAKI, K. Transcriptional regulatory networks on response to abiotic stresses in Arabidopsis and grasses. **Plant Physiology**, v.149, p.88-95, 2009.

NARUSAKA, Y.; NAKASHIMA, K.; SHINWARI, Z.K.; SAKUMA, Y.; FURIHATA, T.; ABE, H.; NARUSAKA, M.; SHINOZAKI, K.; YAMAGUCHI-SHINOZAKI, K. Interaction between two *cis*-acting elements. ABRE and DRE, is ABA-dependent expression of Arabidopsis rd29A gene in response to dehydration and high-salinity stresses. **Plant Journal,** v.34, p.137-148, 2003.

NCBI Map Viewer – National Center of Biotechnology Information. Disponível em: http://www.ncbi.nlm.nih.gov/mapview/ Acesso em: durante todo o período de desenvolvimento do projeto.

NEMHAUSER, J.L.; MOCKLER, T.C.; CHORY, J. Interdependency of brassinosteroid and auxin signaling in Arabidopsis. PLoS Biology, v.2, n.9, p.e259, 2004.

NERO, D.; KATARI, M. I.; KELFER, J.; TRANCHINA, D.; CORUZZI, G. M. In silico evaluation of predicted regulatory interactions in *Arabidopsis thaliana*. **BMC Bioinformatics**, v.10, p.435, 2009.

PALMER, R.G.; KILIEN, T.C. Qualitative genetics and cytogenetics. In: WILCOX, J.R., editor. Soybeans: Improvement, Production, and Uses. **Madison (WI): American Society of Agronomy, Inc., Crop Science Society of America, Inc., and Soil Science Society of America, Inc.**; e.2, p. 135-209, 1987.

PATERSON, A.H.; LIN, H-R.; LI, Z.; SCHERTZ, K.F.; DOEBLEY, J.F.; PINSON, S.R.M.; LIU, S-C.; STANSEL, J.W.; IRVINE, J.E. Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. **Science**, v.269, p.1714-1718, 1995.

PAVESI, G.; MAURI, G.; PESOLE, G. *In silico* representation and Discovery of transcriptional factor binding sites. **Briefings in Bioinformatics**, v.5, p.217-236, 2004.

PAVESI, G.; MEREGHETTI, P.; ZAMBELLI, F.; STEFANI, M.; MAURI, G.; PESOLE, G. MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes**. Nucleid Acids**, v.34, p.566-570, 2006.

PENG, J.R.; RICHARDS, D.E.; HATLEY, N.M.; MURPHY, G.P.; DEVOS, K.M.; FLINTHAM, J.E.; BEALES, J.; FISH, L.J.; WORLAND, A.J.; PELICA, F. *et al*.. "Green Revolution" genes encode mutant giberellin response modulators. **Nature**, v.400, p.256-261, 1999.

PERRY, L. SANDWEISS, D.H.; PIPERNO, D.R.; RADEMARKER, K.; MALPASS, M.A.; UMIRE, A.; DE LA VERA, P. Early maize culture and interzonal interaction in southern Peru. **Nature**, 440, p.76-79, 2006.

Pfam database. Disponível em: http://pfam.sanger.ac.uk/ Acesso em: jan. e fev. 2010.

PHYTOZOME. Disponível em: http://www.phytozome.net/cgi-bin/gbrowse/soybean/ Acesso em: mai. a jun. 2010.

PIRES, J.L.F.; SOPRANO, E.; CASSOL, B. Adaptações morfofisiológicas da soja em solo inundado. **Pesquisa Agropecuária Brasileira**, v.37, n.1, p.41-50, 2002.

PLACE – A Database os Plant Cis-acting Regulatory DNA Elements. Disponível em: http://www.dna.affrc.go.jp/PLACE/ Acesso em: jul. a ago. 2010.

PONNAMPERUMA, F.N. The chemistry of submerged soils. **Advances in Agronomy**, v.24, p.29-96, 1972.

PONNAMPERUMA, F.N. Dinamics aspects offlooded soils and nutrition of the rice plant. In: SYMPOSIUM ON THE MINERAL NUTRITION OF THE RICE PLANT, 1694, Los Banos. **Proceedings**, p.295-328, 1965.

PRAKASH, A.; TOMPA, M. Discovery of regulatory elements in vertebrates through comparative genomics. **National Biotechnology**, v.23, p.1249-1256, 2005.

PRADHAN, S.K.; VARADE, S.B; KAR, S. Influence of soil water conditions on growth and root hydroxide systems. **Soil Science**, v.103, p.374-382, 1967.

PRIEST, H. D., FILICHKIN, S. A., MOCKLER, T. C. *cis*-Regulatory elements in plant cell signaling. **Current Opinion in Plant Biology**, v.12, p.1-7, 2009.

RAMAKRISHNA, W.; DUBCOVSKY, J.; PARK, Y-J.; BUSSO, C.; EMBERTON, J.; SANMIGUEL, P.J.; BENNETZEN, J.L. Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. **Genetics**, v.162, p.1389-1400, 2002.

REICHMANN, J.L.; HEARD, J.; MARTIN, G.; REUBER, L.; JIANG, C-Z.; KEDDIE, J.; ADAM, L.; PINEADA, O.; RATCLIFFE, O.; SAMAHA, R.R.; CREELMAN, R.; PILGRIM, M.; BROUN, P.; ZHANG, J.Z.; GHANDEHARI, D.; SHERMAN, B.K.; YU, G-L. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. **Science**, v.209, pp.2105-2110, 2000.

RGAP – Rice Genome Annotation Project. Disponível em: http://rice.plantbiology.msu.edu/ Acesso em: fev. a mar. 2010.

RIETHOVEN, J-J.M. Regulatory regions in DNA: Promoters, Enhancers, Silencers, and Insulator *in:* **Computation Biology of Transcriptional Factor Binding.** Edit by LADUNGA, I. v.1, p.33.42, 2010.

ROMBAUTS, S.; FLORQUIN, K.; LESCOT, M.; MARCHAL, K.; ROUZE, P.; VAN DE PEER, Y. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. **Plant Physiol**ogy, v.132, p. 1162-1176, 2003.

ROST, B. Twilight zone of protein sequence alignments. **Protein Engineering**, v.12, n.2, p.85-94, 1999.

ROWELL, D.L. Oxidation and reduction. In: GREENLAND. D.J.; HAYES, M.H.B., eds. **The chemistry of soil processes**, p.401-461, 1981.

SALSE, J.; ABROUK, M.; MURAT, F.; QURAISHI, U.M.; FEUILLET, C. Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. **Briefings in Bioinformatics**, v.10, n.6, p.619-630, 2009.

SILVA, C. F. L. Variabilidade genética para o caráter tolerância ao encharcamento em aveia. **Pelotas, Universidade Federal de Pelotas**, 1999.

SONG, R.; LLACA, V.; LINTON, E.; MESSING, J. Sequence, regulationm and evolution of the maize 22-kD zein gene family. **Genome Research**, v.11, p.1817-1825, 2001.

SONG, R.; LLACA, V.; MESSING, J. Mosaic organization of orthologous sequences in grass genomes. **Genome Research**, v.12, p.1549-1555, 2002.

SPANNAGL, M.; MAYER, K.; DURNER, J.; HABERER, G.; FRÖHLICH, A. Exploring the genomes: From Arabidopsis to crops. **Journal of Plant Physiology**, v.168, p.3-8, 2010.

STOLZY, L.H. Soil atmosphere. In: CARSON, E.W. ed, **The plant root and its environment**, p.355-361, 1974.

TAIR – The Arabidopsis Information Resource. Disponível em: www.arabidopsis.org/ Acesso em: jan. a fev. 2010.

THE ARABIDOPSIS GENOME INITIATIVE. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. **Nature**, v.48, p.796-815, 2000.

The NCBI Entrez Taxonomy Homepage (http://www.ncbi.nlm.nih.gov/taxonomy/).

TIKHONOV, A.P.; SANMIGUEL, P.J.; NAKAJIMA, Y.; GORENSTEIN, N.M.; BENNETZEN, J.L.; AVRAMOVA, Z. Colinearity and its exceptions in orthologous adh regions of maize and sorghum. **Proceedings of the National Academy os Sciences**, v.96, p.7409-7414, 1999.

VAN DODEWEERD, A.M.; CAROLINE, C.R.; BENT, E.G.; JOHNSON, S.J.; BEVAN, M.W.; BANCROFT, I. Identification and analysis of homoeologus segments of the genomes of rice and Arabidopsis thaliana. **Genome**, v.42, p.887-892, 1999.

WASSERMAN, W.W.; SANDELIN, A. Applied bioinformatics for the identificantion of regulatory elements. **National Review Genetics**, v.5, p.276-287, 2004.

WILSON, W.A.; HARRINGTON, S.E.; WOODMAN, W.L., LEE, M.; SORRELLS, M.E.; McCOUCH, S.R. Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize and domesticated panicoids. **Genetics**, v.153. 453-73, 1999.

XIONG, L.; SCHUMAKER, K.S.; ZHU, J.K. Cell signaling during cold, drought, and salt stress. **Plant Cell**, v.14, p.S165-S183, 2002.

YAMAGUCHI-SHINOZAKI, K.; SHINOZAKI, K. Organization of *cis*-acting regulatory elements in osmotic- and cold-stress-responsive promoters. **TRENDS in Plant Science**, v.10, n.2, 2005.

YAMAGUCHI-SHINOZAKI,K.; SHINOZAKI, K. Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. **Annual Review Plant Biology**, v.57, p.781-803, 2006.

YOSHIDA, T. Microbial metabolism of flooded soils. In: PAUL, E.A.; McLAREN, A.D., eds. **Soil Biochemistry**, v.3, p.083-122, 1975.

Apêndice 1

Artigo gerado pelo trabalho de dissertação, submetido na revista *BMC Research Notes* em fevereiro de 2011

CHARACTERIZATION OF *CIS*-REGULATORY ELEMENTS IN PROMOTER REGIONS OF FLOODING RESPONSIVE GENES

Lara Isys Dias[1], Eliseu Binneck[2], Luciano Carlos da Maia[1], Antonio Costa de Oliveira[1].

[1]Plant genomics and Breeding Center, Graduate Program in Biotechnology, Federal University of Pelotas (UFPel), Pelotas, RS, Brazil.

[2]Empresa Brasileira de Pesquisa Agropecuária – Embrapa Soja, Londrina, PR, Brazil.

E-mail adresses:

LID: laraisysdias@gmail.com

EB: eliseubinneck@gmail.com

LCM: lucianoc.maia@gmail.com

ACO: acostol@gmail.com

Abstract

Background: The current challenges in plant breeding are to maximize the productivity of major crop species and to create means for exploring novel crop environments. One of these environments is the lowland hydromorphic soils that are proper for the irrigated rice crop. Adapting other crops to this environment could reduce the incidence of diseases, pests and weeds, therefore benefiting from a crop rotation system. When a plant is exposed to abiotic stresses, it has to cope with environmental changes through physiological and anatomic changes that need quick gene expression responses, i.e., changes in active/silenced status as well as in the rates of transcription. *Cis*-acting regulatory elements have straight relationship with transcription factors (TF) in complex signaling networks. This TF binding sites (*cis*-elements) are the functional DNA elements that influence temporal and spatial transcriptional activity. Some *cis* elements occurrence patterns usually indicates how molecular mechanisms in plants are modulated under stress conditions, like flooding.

Findings: We investigated possible patterns of sequences that can be inferred about the mechanisms that plants use to develop under flooding stress. This search for possible homologies between the various *cis*-elements would lead us to performed interactive analyses about how plants use their molecular mechanisms responding to abiotic stresses. Online databases were searched, looking for genes previously described in literature which are expressed in response to flooding in *Oryza sativa, Arabidopsis thaliana* and their homologs in *Glycine max* and *Zea mays.* The 1.0 Kb upstream portion of each gene was extracted and analyzed *in silico*. Besides, all the promoters of these four species were subjected to a tool for searching for novel signals, intending to find new motif patterns.

Conclusions:

Our *in silico* analysis shows that from 259 *cis* elements found in PLACE for all promoters of Arabidopsis and rice, 12 of them are common to both species, and are distinguished by having high frequency. Using the MEME program two consensus motifs could be found among the species *Oryza sativa* and *Zea mays*. These could

represent new *cis* elements patterns, because they had relatively high occurrences in the gene promoters and they are related to conserved sequences in monocots.

The analysis here presented shows important points for future studies related to the waterlogging stress and unmasking molecular tolerance mechanisms to this typical stress. From the data generated, it will be possible to direct experiments on genetic transformation with target genes and/or *cis* elements in order to attribute some characteristic in plants, such as those found in rice, so they can develop in an environment with O_2 deprivation.

Findings

Background

The current scenario of stagnation of the area used for agriculture poses a challenge to the scientific community, which is to maximize the productivity of major crop species and also to create means for exploring novel crop environments. One of these environments is the lowland hydromorphic soils, and this leads for the search for alternatives to the development of new varieties/species containing genotypic and/or phenotypic characteristics similar to the rice crop. These novel crops would need to tolerate the conditions imposed by flooding and maintain a high yield potential [1].

At the molecular level, abiotic stress signals activate transcription factors and proteins that bind to regions adjacent to target genes, thereby generating a plant response to the environment. In the upstream DNA regions near or within the promoter of these genes there are conserved sequences that appear to be involved in their regulation [2-3]. These short sequences, nearly 7 base pairs (bp), interact with various transcription factors (TFs) to form a transcriptional initiation complex and are called *cis* elements. These *cis*-acting elements are involved in multiple transcription regulatory processes, acting as molecular switches and controlling biological processes in response to abiotic stress, hormonal and developmental processes in plants [4].

Advances in research and the accuracy of their results in the transcriptome expression profiling has led to identification of various combinations (cross-talk) of *cis*-acting elements in promoter regions of stress-inducible genes, also involved in hormonal responses. Previous reports showed that there are two main *cis*-acting elements that work in the regulation of gene expression in response to osmotic stress and temperature, they are ABRE (ABA responsive element) and DRE (dehydration responsive element) - ABA-dependent and ABA-independent, respectively [3].

Molecular and genetic studies provide evidence that the monocot *Oryza sativa* and the dicot *Arabidopsis thaliana* have common mechanisms for regulating gene expression. TFs play an important role in regulating gene expression in response to abiotic stresses and most PTFs are common among the grasses and Arabidopsis [5].

The recent availability of many high quality sequences, fully annotated plant genomes, and large public databases of global expression measures and accessible technologies for expression profiling of individual laboratories, has translated into many studies involving binding sites of TFs and their role as components of a large transcriptional network [6]. Recent studies in several eukaryotic species have been focused on a Systems Biology approach to elucidate the regulatory networks and to understand its biological context [7]. These associations can be used in combination with gene expression data from microarray experiments and sequence analysis of co-regulated gene promoters, to infer mechanisms of this regulation and to seek *cis* regulatory elements, which can coordinate the response through the activity of TFs [8].

A considerable number of algorithms and bioinformatics tools have been developed to identify potential *cis* elements in the regulatory sequences of genes co-expressed [9-10]. The basic principle of computational approaches is that co-regulated genes should contain similar *cis* elements in their upstream regulatory regions in statistically significant levels [6]. Regardless of the exact algorithmic details, in general, the computational approaches to the identification of potential *cis* elements estimate the probability of short DNA motifs occurrences by comparing an observed number of a particular motif in a set of sequences with expected occurrences number based on random samples or some statistical distribution model [11].

Material and Methods

Literature search and clustering of genes that respond to the flooding stress

We performed a review in order to obtain a considerable number of genes whose expression is altered when exposed to waterlogging stress. The extracted data refer to analysis of microarray experiments in *Arabidopsis thaliana* plants and *Oryza sativa* under the stress of hypoxia/anoxia. These two species were chosen because they represent model species, as regards the studies with comparative genomics. We also get the homologs in *Zea mays* and *Glycine max*, using the amino acid sequences of each gene previously annotated from Arabidopsis and rice.

Protein sequences allowed us to cluster genes into groups according to their functions and then alignments of the promoter regions could be carried out in co-regulated genes with characteristics in common.

Cis element pattern searches

There are two approaches widely used in the study of *cis* elements: one is the submission of a promoter sequence in the database, which returns the known and annotated *cis* elements in that region; another way is an alignment program which will recognize and identify repeating patterns of short sequences. In this work these two tests were made, so that the data corroboration could provide us more consistency and uniqueness of the results obtained.

To obtain information about reported *cis* elements, we used the platform of PLACE database on the *cis*-acting regulatory elements for vascular plants [12]. We calculated the Z score for each *cis* element found. This value indicates the probability of the result found to be random. A cutoff of 0.05 (or 5%) is commonly used to eliminate false positives [9].

To search for new patterns of occurrence of *cis* elements, the MEME tool was used. Its algorithm has been used for the discovery of DNA motifs [13]. The MEME works with the promoter regions of genes that are co-regulated. It constructs motif models of the most conserved regions and then combines these models with a linear framework of

HMM (Hidden Markov Model) [14], being the most conserved regions given as hits from patterns of occurrence of motifs.

Analysis

The predicted *cis* elements with greater frequency obtained from the PLACE database for *Arabidopsis thaliana* were compared with those obtained for more frequent *Oryza sativa* by the same process. Thus, it was possible to observe if there was any pattern of occurrence of predicted *cis* elements in genes whose expression is altered under flooding stress.

From the results obtained by the alignment of sets of promoters for the four species: *Arabidopsis thaliana*, *Oryza sativa*, *Glycine max* and *Zea mays*, the *cis* element occurrence patterns were sorted among groups of genes with common features, as follows: 1) glycolysis, gluconeogenesis and fermentation, 2) heat shock proteins (HSPs), 3) cytochrome P450, 4) respiratory chain, and 5) other (with unknown function, protein binding and signaling)

Results and Discussion

Gene annotation and homologous searching

The number of genes meeting our set standards, i.e., discarding those that might not have their nucleotide sequences annotated and others that had no counterpart found in each species, were: *Oryza sativa*, 182 genes; *Arabidopsis thaliana*, 155 genes; *Glycine max*, 144 genes; *Zea mays*, 166 genes.

During the search for homologous some sequences were not included due to low similarity and low e-value. There was a considerable amount of duplication between the genomes, especially among the monocot *Oryza sativa* and the dicotyledons *Arabidopsis thaliana* and *Glycine max*. We can infer from these duplications most observed in rice is that this species probably has a larger number of genes that respond to flooding in relation to the others in this study. In fact, rice has morphological, physiological, biochemical, and genetic characteristics, which strongly support the plant when it's submerged in water. Polyploidization events may have had

important consequences in plant evolution, particularly in the eradication and adaptation of species and for the modulation of functional capacities [15].

Digital microarray profile for *Arabidopsis thaliana* and *Oryza sativa* genes

The digital microarray profile obtained for Arabidopsis and rice confirmed the data obtained in the literature that these genes respond to the stress of hypoxia and anoxia. For this analysis we used the Genevestigator tool which has a database of 5,747 genes from microarray experiments from Arabidopsis thaliana and 305, under various conditions, the developmental stages and anatomy.

Analysis of the predicted *cis* element results

Quite different frequencies were observed between the *cis* elements in *Arabidopsis thaliana* promoters. Some were found in only one promoter, while others were observed in 38 or 39 promoters (Table 1). This approach allows analysis of *cis* elements in order to observe the amount of TFs that possibly influence the regulation of gene expression under flooding, also providing the key genes in this process. They are genes whose promoters contain a greater amount of binding sites of TFs, in other words rich in *cis* elements.

Table 1. Frequently found *cis* elements in promoters of 155 flooding stress responsive genes in *Arabidopsis thaliana*.

***Cis* elements**	**Number of promoters (frequency)**
ACGTABOX	39 (25%)
LECPLEACS2	39 (25%)
BOXLCOREDCPAL	38 (24%)
MYBPLANT	35 (22%)
RBCSCONSENSUS	34 (22%)
TATAPVTRNALEU	31 (20%)
MYB2AT	30 (19%)
WBOXNTCHN48	29 (19%)
CURECORECR	28 (18%)
LTRE1HVBLT49	28 (18%)
POLASIG2	27 (17%)
SEF1MOTIF	27 (17%)
SEF4MOTIFGM7S	27 (17%)
SP8BFIBSP8BIB	27 (17%)

BOXIINTPATPB	25 (16%)
SV40COREENHAN	25 (16%)
-10PEHVPSBD	24 (15%)
MYB1AT	24 (15%)
PYRIMIDINEBOXHVEPB1	24 (15%)
TAAAGSTKST1	24 (15%)
CCA1ATLHCB1	23 (14%)
MYB2CONSENSUSAT	22 (14%)
MYB1LEPR	21 (13%)
ACGTATERD1	20 (13%)
CPBCSPOR	20 (13%)
HDZIP2ATATHB2	20 (13%)
HEXMOTIFTAH3H4	20 (13%)

These results suggest that these flooding responsive genes should be more sensitive to a greater number of signals, because their promoter regions have a variety of *cis*-acting elements and, therefore, participate in a range of metabolic pathways. Besides, the cross talk between TFs that promote changes in the regulation of several genes may enable plants to respond to the environment. The protein family classification of the 25 genes with higher number of *cis* elements in their promoters gives an idea of what are the key mechanisms that these same genes are involved and that the processes of glycolysis, gluconeogenesis, fermentation, HSPs and proteins of the respiratory chain are likely to be key points in the discovery of flooding tolerance.

The *cis* elements found for *Oryza sativa* are more numerous in comparison to *Arabidopsis thaliana* (Table 2). This may indicate a more refined regulation of the rice genes for flooding tolerance, leading to a greater versatility in signal perception. A total of 39 genes from rice have at least 20 *cis* regulatory elements in their promoters. This number is more expressive than the 26 found for the same number in Arabidopsis promoters, which may indicate that the greater the amount of *cis* elements, the greater the ability to tolerate stress. Furthermore, almost twice (48) the genes in *Oryza sativa* have at least 18 *cis* elements in their promoters when compared to *Arabidopsis thaliana* (25).

Table 2. Frequently found *cis* elements in promoters of 182 flooding stress responsive genes in *Oryza sativa*.

***Cis* elements**	**Number of promoters (frequency)**
ACGTABOX	53 (29%)
ACGTTBOX	42 (23%)
TATABOX3	42 (23%)
LTRE1HVBLT49	41 (22%)
ACGTATERD1	38 (21%)
HEXMOTIFTAH3H4	38 (21%)
T/GBOXATPIN2	37 (20%)
TGACGTVMAMY	36 (29%)
GT1CORE	34 (19%)
ACGTCBOX	32 (17%)
MYB2AT	32 (17%)
AMYBOX1	31 (17%)
ARFAT	31 (17%)
REBETALGLHCB21	31 (17%)
BS1EGCCR	30 (16%)
CRTDREHVCBF2	30 (16%)
QELEMENTZMZM13	30 (16%)
RHERPATEXPA7	30 (16%)
TATCCAYMOTIFOSRAMY3D	30 (16%)
SEF1MOTIF	29 (16%)
DRE2COREZMRAB17	27 (15%)
SV40COREENHAN	27 (15%)
MYBGAHV	26 (14%)
SP8BFIBSP8BIB	26 (14%)
TAAAGSTKST1	26 (14%)
RYREPEATGMGY2	25 (14%)
TATAPVTRNALEU	25 (14%)
AACACOREOSGLUB1	24 (13%)
CURECORECR	24 (13%)
PYRIMIDINEBOXOSRAMY1A	24 (13%)
ASF1MOTIFCAMV	23 (13%)
IBOXCORENT	23 (13%)
NODCON1GM	22 (12%)
OSE1ROOTNODULE	22 (12%)
ABRERATCAL	21 (11%)
ACGTABREMOTIFA2OSEM	21 (11%)
PRECONSCRHSP70A	21 (11%)
CCAATBOX1	20 (10%)
MYB1LEPR	20 (10%)

The same way as found in Arabidopsis, rice genes related to glycolysis, gluconeogenesis, fermentation, and HSP emerge to be the more representative processes regarding the number of common regulatory elements found in their promoters.

From the comparison of data obtained from the analysis of Arabidopsis and rice promoters, we got the most common *cis* elements between these two species (Table 3, they are. These are therefore the most representative *cis* elements found in genes that have their expression altered, when the plant is under flooding stress (anoxic), existing in both species *Arabidopsis thaliana* and *Oryza sativa*. These are important points of study to understand the mechanisms used by the plant when developing under such stress.

Table 3. Most common *cis* elements names and sequences between Arabidopsis and rice,

***Cis* element**	**Sequence**
ACGTABOX	TACGTA
ACGTATERD1	ACGT
CURECORECR	GTAC
HEXMOTIFTAH3H4	ACGTCA
LTRE1HVBLT49	CCGAAA
MYB2AT	TAACTG
PYRIMIDINEBOXHVEPB1	TTTTTTCC
SEF1MOTIF	ATATTTAWW
SP8BFIBSP8BIB	TACTATT
SV40COREENHAN	GTGGWWHG
TAAAGSTKST1	TTTATATA
TATAPVTRNALEU	TTTATATA

For the 25 genes of *Arabidopsis thaliana* and *Oryza sativa* with the promoter regions richer in *cis* elements, those that fit the groups related to glycolysis, gluconeogenesis, fermentation and HSPs, stood out, reaching 65% in Arabidopsis and 64% in rice (Figure 1).

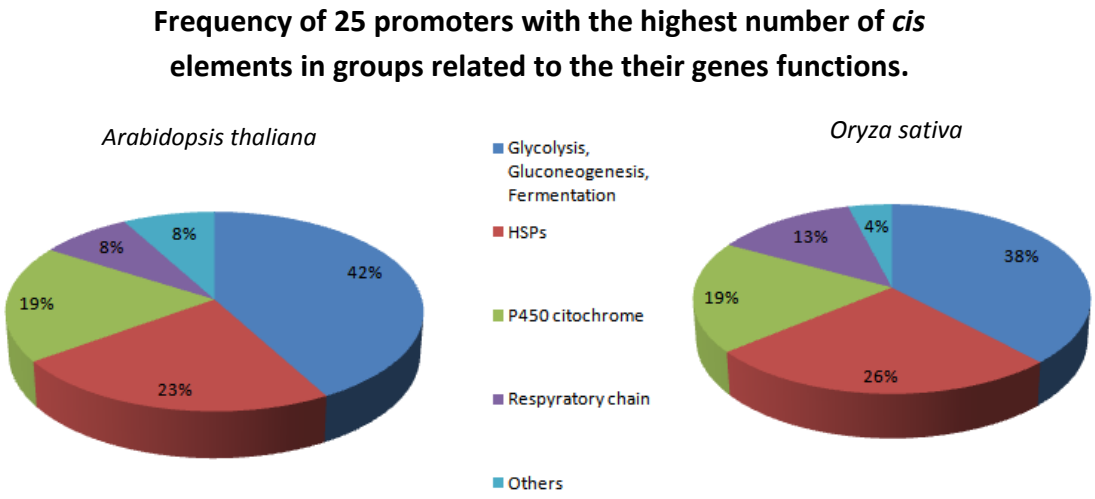


Figure 1. Distribution of 25 of *Arabidopsis thaliana* and *Oryza sativa* promoters with the largest number of *cis* elements in groups related to their gene functions.

Therefore, these groups are most representative genes in Arabidopsis and rice that have their gene regulation altered when exposed to the stress of anoxia. These data might suggest that these pathways may play a role in gene regulation mechanisms in response to flooding.

Among the *cis* elements that had a number of occurrences equal to or greater than 20, we analyzed its conservation within each class. Tables 4 and 5 refer to the five most preserved elements, and its occurrence in the group, in *Arabidopsis thaliana* and *Oryza sativa*, respectively.

Table 4. Most conserved *cis* elements in the classes for *Arabidopsis thaliana.*

***Cis* element**	**Sequence**	**Group**	**Ocorrence in the group**	**Species**
MYB2AT	TAACTG	GGF	13	*Arabidopsis thaliana*
RBCSCONSENSUS	AATCCAA	GGF	11	*Arabidopsis thaliana*
BOXLCOREDCPAL	ACCWWCC	GGF	10	*Arabidopsis thaliana*
MYBPLANT	MACCWAMC	GGF	10	*Arabidopsis thaliana*
SV40COREENHAN	GTGGWWHG	GGF	10	*Arabidopsis thaliana*
LECPLEACS2	TAAAATAT	P450	12	*Arabidopsis thaliana*
ACGTABOX	TACGTA	P450	11	*Arabidopsis thaliana*
MYBPLANT	MACCWAMC	P450	11	*Arabidopsis thaliana*
BOXLCOREDCPAL	ACCWWCC	P450	10	*Arabidopsis thaliana*
CURECORECR	GTAC	P450	9	*Arabidopsis thaliana*
ACGTABOX	TACGTA	HSP	11	*Arabidopsis thaliana*
TATAPVTRNALEU	TTTATATA	HSP	11	*Arabidopsis thaliana*
LECPLEACS2	TAAAATAT	HSP	9	*Arabidopsis thaliana*
SEF1MOTIF	ATATTTAWW	HSP	8	*Arabidopsis thaliana*
SP8BFIBSP8BIB	TACTATT	HSP	7	*Arabidopsis thaliana*
LECPLEACS2	TAAAATAT	RC	7	*Arabidopsis thaliana*
POLASIG2	AATTAAA	RC	6	*Arabidopsis thaliana*
BOXLCOREDCPAL	ACCWWCC	RC	5	*Arabidopsis thaliana*
SEF4MOTIFGM7S	RTTTTTR	RC	5	*Arabidopsis thaliana*
ACGTATERD1	ACGT	RC	5	*Arabidopsis thaliana*

Note: GGF refers to glycolysis, gluconeogenesis and fermentation, P450 to P450 cytochrome, HSP to heat shock protein and RC to respiratory chain.

From these data it is possible to observe the most conserved *cis* elements among groups and species. The cis element ACGTABOX is observed in both specie, regarding only the groups related to glycolysis, gluconeogenesis and fermentation (GGF) and respiratory chain. Thus, it might be an important motif in controlling the gene regulation when these plants are submitted do flooding.

Others elements are more represented only in one species, like MYBPLANT in Arabidopsis and ACGTTBOX in rice, these are likely to be motifs that participate in the plant gene regulation in a restricted way, i.e., controlling some genes on specific plant family, maybe like exclusive to monocots or dicot.

Table 5. Most conserved *cis* elements in the classes for *Oryza sativa.*

***Cis* element**	**Sequence**	**Group**	**Ocorrence in the group**	**Species**
ACGTABOX	TACGTA	GGF	16	*Oryza sativa*
HEXMOTIFTAH3H4	ACGTCA	GGF	16	*Oryza sativa*
T/GBOXATPIN2	AACGTG	GGF	14	*Oryza sativa*
CRTDREHVCBF2	GTCGAC	GGF	13	*Oryza sativa*
ACGTTBOX	AACGTT	GGF	12	*Oryza sativa*
LTRE1HVBLT49	CCGAAA	P450	15	*Oryza sativa*
TATABOX3	TATTAAT	P450	14	*Oryza sativa*
ACGTABOX	TACGTA	P450	13	*Oryza sativa*
SEF1MOTIF	ATATTTAWW	P450	13	*Oryza sativa*
ACGTTBOX	AACGTT	P450	12	*Oryza sativa*
ACGTCBOX	GACGTC	HSP	14	*Oryza sativa*
BS1EGCCR	AGCGGG	HSP	12	*Oryza sativa*
ACGTABOX	TACGTA	HSP	9	*Oryza sativa*
TATABOX3	TATTAAT	HSP	9	*Oryza sativa*
TGACGTVMAMY	TGACGT	HSP	9	*Oryza sativa*
ACGTABOX	TACGTA	RC	6	*Oryza sativa*
TGACGTVMAMY	TGACGT	RC	6	*Oryza sativa*
GT1CORE	GGTTAA	RC	6	*Oryza sativa*
ARFAT	TGTCTC	RC	6	*Oryza sativa*
ACGTTBOX	AACGTT	RC	5	*Oryza sativa*

Note: GGF refers to glycolysis, gluconeogenesis and fermentation, P450 to P450 cytochrome, HSP to heat shock protein and RC to respiratory chain.

Analysis of new *cis* elements patterns results

For each species studied, their promoter regions were grouped according to the function of each gene and therefore we searched patterns of occurrence of intraspecific and interspecific conserved DNA motifs, according to frequency, repetition, and focus groups of motives. Besides, promoters can be highlighted by the number of times they were observed within each group and/or species.

Among the most frequent motifs in the promoters group related to glycolysis, gluconeogenesis and fermentation, we observed a strong similarity between the motifs CG[CG]C[CG][AC][CG]G[GC][GC][GC][GC], GC[CG]G[CG][CG]C[GT][CG]G[CT]C and GG[CT]G[GA][CA]G[GA]CG[GA][CG], which can generate a consensus motif, such as: [GC][GC]C[CG]GC[GC]GCG[GC][CG], generated from the combination of the three previous, present in *Oryza sativa* and *Zea mays* (Table 5). This DNA pattern may be related to mechanisms of regulation of gene expression of genes that were found.

Table 5. Distribution of most common motifs in promoters of genes belonging to Glycolysis, Gluconeogenesis and Fermentation (GGF).

Motif	Number of occurrences	Group	Species
A[GA]AAAAA[AG]AAAA	46	GGF	*Arabidopsis thaliana*
[TC]TT[CT][TC]TTTTTCT	50	GGF	*Arabidopsis thaliana*
[CG]TC[TG][TG]TCTCT[TC][CT]	29	GGF	*Arabidopsis thaliana*
C[TC]C[CT]CC[ATC][CT][CA]CC	32	GGF	*Oryza sativa*
CG[GC]C[GC][CA][CG]G[GC][GC][GC][GC]	**50**	**GGF**	***Oryza sativa***
AAAAAAAAAGAA	46	GGF	*Oryza sativa*
[GA][AG]G[AG][GA][GA]G[AG]G[GA]A[GA]	39	GGF	*Oryza sativa*
TTTTTTTT[CTA]TTT	44	GGF	*Oryza sativa*
C[TC]C[CT]CC[ATC][CT][CA]CC	32	GGF	*Glycine max*
[CG]GT[GC]G[GT][AGC][GT][TA]GG[CTG]	23	GGF	*Glycine max*
AAAAA[AT]AAAA	48	GGF	*Glycine max*
TTTT[TC][TCA]TT[TC]T[TC]T	47	GGF	*Glycine max*
AA[AT][AC]A[AT]A[AG][TA]AA	49	GGF	*Zea mays*
TTT[TG]TTTT[GC]TTT	45	GGF	*Zea mays*
GC[CG]G[GC][CG]C[GT][CG]G[CT]C	**36**	**GGF**	***Zea mays***
GG[CT]G[GA][CA]G[GA]CG[GA][CG]	**36**	**GGF**	***Zea mays***

Despite being quite similar, the motifs CG[CG][CG][CG][CG]C[GC] and GCGG[CG]GC[CG]CGGC[GC][GCA], related to the respiratory chain and found in the species *Oryza sativa* and *Zea mays* may have a consensus motif – [CG]GCG[GC]C[CG]G[CG]CG – and maybe have the same function in relation of possible TFs binding relationship (Table 6).

Table 6. Distribution of most common motifs in promoters of genes belonging to the respiratory chain (RC).

Motif	Number of occurrences	Group	Species
[GC][GTA][TC]T[CG][CT]C[TG][CT]CT	20	CR	*Arabidopsis thaliana*
CG[CG][GC][GC][CG]C[GC]GCG	**27**	**CR**	***Oryza sativa***
[TC]TTTTTTTT[CAT][TG]T	26	CR	*Oryza sativa*
[AT]AT[AT]AAAA[AG]AA	27	CR	*Oryza sativa*
[AC]G[AG][AG]G[GAC][GA][GA]AG[AG][GA]	25	CR	*Oryza sativa*
AAAA[AG][AG]AA[AG][CA]A	20	CR	*Glycine max*
[CT]T[CT][TC]C[TC][TC][TC][CT][TC]C[CT]	32	CR	*Glycine max*
GAA[AC]AAA[AG]AAA	20	CR	*Glycine max*
G[GC]CG[CG]CGGC[GC][GCA]	**33**	**CR**	***Zea mays***
T[TA][AGTCC]TTT[TC]T[TC]TT	30	CR	*Zea mays*
AAA[AGC]AA[AC]A	24	CR	*Zea mays*
CCT[TG][CG][CG][CT]CTCC	22	CR	*Zea mays*

Conclusions

Our *in silico* analysis showed that there is greater homology between monocots (*Oryza sativa* and *Zea mays*), and between dicotyledonous (*Arabidopsis thaliana* and *Glycine max*) genes responsive to flooding stress. There is also a greater amount of duplication of *Oryza sativa* flooding tolerance genes, reinforcing the idea that this species has many complex genetic mechanisms, probably because during its evolution this plant acquired the ability to develop under the flooding fields.

From 259 *cis* elements found in PLACE for all promoters of Arabidopsis and rice, 12 (4.6%) of them are common to both species, and are distinguished by having high frequency. These *cis* elements should be focus of studies for understanding the mechanisms of gene regulation under the hipoxia/anoxia stress.

Using the MEME program, and alignment of gene promoters regions of the four species studied, two consensus motifs could be found among the species *Oryza sativa* and *Zea mays*. These could represent new *cis* elements patterns, because they had relatively high occurrences in the gene promoters and they are related to conserved sequences in monocots.

The analysis here presented show important points for future studies related to the waterlogging stress and unmasking molecular tolerance mechanisms to this typical stress. From the data generated, it is possible to guide experiments on genetic transformation with target genes and/or *cis* elements and to try to attribute some characteristic in plants, like those found in rice, so they can develop in an environment with O_2 deprivation.

In silico studies, as well as the use of the available tools are therefore useful for the scientific community, where a large amount of data is generated every day and they need to be treated, sorted and organized so that we can infer about biological functions with more precision.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

LID performed the experiments and wrote the manuscript. ACO designed the experiments and supervised the project. EB and LCM provided expert support for the algorithms and software used. All authors read and approved the final manuscript.

Acknowledgments

We thank the Federal University of Pelotas (UFPel) and the Graduate Program in Biotechnology for the opportunity to realize the project.

References

1. Pires JLF, Soprano E, Cassol B: **Adaptações morfofisiológicas da soja em solo inundado.** Pesquisa Agropecuária Brasileira, 2002, **37**:41-50, 2002

2. Yamaguchi-Shinozaki K, Shinozaki K: **Organization of *cis*-acting regulatory elements in osmotic- and cold-stress-responsive promoters.** TRENDS in Plant Science, 2005, **10**.

3. Bartels D, Sunkar R: **Drought and salt tolerance in plants.** Critical Reviews in Plant Science, 1995, **24**:23-58.

4. Yamaguchi-Shinozaki K, Shinozaki K: **Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses.** Annual Review Plant Biology, 2006, **57**:781-803.

5. Nakashima K, Yusuke I, Yamaguchi-Shinozaki K: **Transcriptional regulatory networks on response to abiotic stresses in Arabidopsis and grasses.** Plant Physiology, 2009, **149**:88-95.

6. Priest HD, Filichkin SA, Mockler TC: ***cis*-Regulatory elements in plant cell signaling.** Current Opinion in Plant Biology, 2009, **12**:1-7.

7. Levine M, Davidson EH: **Gene regulatory networks for development.** Proceedings of the National Academy of Sciences, 2005, **102**:4936-4942.

8. Nero D., Katari MI, Kelfer J, Tranchina D, Coruzzi GM: ***In silico* evaluation of predicted regulatory interactions in *Arabidopsis thaliana*.** BMC Bioinformatics, 2009, **10**:435.

9. Rombauts S, Florquin K, Lescot M, Marchal K, Rouze P, Van de Peer Y: **Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiology**, 2003, **132**:1162-1176.

10. Wasserman WW, Sandelin A: **Applied bioinformatics for the identificantion of regulatory elements.** National Review Genetics, 2004, **5**:276-287.

11. Michael TP, Mockler TC, Breton G, McEentee C, Byer A, Trout JD, Hazen SP, Shen R, Priest HD, Sullivan CM *et al*.: **Network discovery pipeline elucidates conserved time of day specific *cis*-regulatory modules. PLoS Genetics,** 2008, **4**:14.

12. Higo K, Ugawa Y, Iwamoto W, Higo H: **PLACE: a database of plant *cis*-acting regulatory DNA elements.** Nuclieic Acids Research, 1998, **26**: 1.

13. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, 2994, 28-36.

14. Grundy WN, Bailey TL, Elkan CP, Michael EB: **Meta-MEME: Motif-based hidden Markov models of protein Families.** Computer Applications in the Biosciences, 1997.

15. De Bodt S, Maere S, Van de Peer Y: **Genome duplication and the origin of angiosperms.** Trends in Ecololgy & Evolution, 2005, **20**:591-597.

Figures

Figure 1 - Distribution of 25 of *Arabidopsis thaliana* and *Oryza sativa* promoters with the largest number of *cis* elements in groups related to their gene functions.

Tables

Table 1 - Frequently found *cis* elements in promoters of flooding stress responsive genes in *Arabidopsis thaliana*.

Table 2 - Frequently found *cis* elements in promoters of flooding stress responsive genes in *Oryza sativa*.

Table 3 - Most common *cis* elements names and sequences between Arabidopsis and rice,

Table 4 - Most conserved *cis* elements in the classes for *Arabidopsis thaliana.*

Table 5 - Most conserved *cis* elements in the classes for *Oryza sativa.*

Table 6 - Distribution of most common motifs in promoters of genes belonging to Glycolysis, Gluconeogenesis and Fermentation (GGF).

Table 7 - Distribution of most common motifs in promoters of genes belonging to the respiratory chain (RC).

Anexo 1

Script em linguagem *Phyton* para o corte e reversão das regiões promotoras.

```
#!/usr/bin/python

##########################################################################
##############
#                                                                        #
#     AUTOR.........:   Paulo R. Silla - paulo.silla@gmail.com
         #
#     DESCRICAO.....:   Script que recebe como entrada um arquivo de
sequencias no #
#                  formato .fasta e escreve em um arquivo de saida,
tambem no #
#                  formato .Fasta,   com as mesmas sequencias, porem
limitadas  #
#                  a uma quantidade de nucleotideos, informado na
entrada.       #
#     COMO UTILIZAR.:   O script deve ser executado em linha de
comando, do          #
#                  seguinte modo:                                      #
#                  python corte.py arquivo_entrada arquivo_saida S N
#
#                  Onde:                                               #
#                        S - sentido da sequencia, admitindo 'p' (plus)
ou  #
#                            'm' (minus);                              #
#                        N - numero de nucleotideos que serao inseridos
em  #
#                            arquivo_saida, para cada sequencia.
#
#                                                                        #
##########################################################################
##############

import sys

#Metodo que obtem a sequencia de saida e a organiza em linhas de 70
colunas (formato Fasta)
def trata_sequencia(sequencia, corte, sentido):
      sequencia = sequencia.replace('\t','')
      sequencia = sequencia.replace('\n','')
      sequencia = sequencia.replace('\r','')
      tamanho = len(sequencia)
      sequencia = sequencia[tamanho-corte:]

      #obtem a sequencia da fita complementar, caso o sentido seja
Minus
      if sentido == 'M':
            sequencia = sequencia[::-1]
            sequencia = sequencia.replace("A","t")
            sequencia = sequencia.replace("T","a")
            sequencia = sequencia.replace("G","c")
            sequencia = sequencia.replace("C","g")
            sequencia = sequencia.upper()
      count = 0
      seq = ''
      for i in range(corte / 70):
            inicio = (i * 70)
            fim = inicio + 70
            seq += sequencia[inicio:fim]+'\n'
            count += 70
      inicio = count
      fim = tamanho
```

```
	seq += sequencia[inicio:fim]+'\n'
	return seq

#Realiza a leitura do arquivo de entrada e grava os dados no arquivo
de saida
def leitura_arquivo(entrada, saida, corte, sentido):
	file_in = open(entrada, 'r')
	file_out = open(saida, 'w')
	sequencia = ''
	try:
		for line in file_in:
			if line[0] != 'A' and line[0] != 'T' and line[0] !=
'C' and line[0] != 'G':
				if sequencia != '':
					seq = trata_sequencia(sequencia, corte,
sentido)
					file_out.write(seq)
					sequencia = ''
					info = ''
				file_out.write(line)
			else:
				sequencia += line
	finally:
		file_in.close()
		seq = trata_sequencia(sequencia, corte, sentido)
		file_out.write(seq)
		file_out.close()

def main(argv):
	sentido_fita = 'MP'
	if ( (len(argv) != 5) or (sentido_fita.find(argv[4].upper()) ==
-1) or (argv[3].isdigit() == False) ):
		print "Parametro(s) invalido(s) na linha de comando."
		print "Utilize:"
		print "python corte.py arquivo_entrada arquivo_saida S N"
		print "Onde:"
		print "S - sentido da sequencia, admitindo 'p' (plus) ou
'm' (minus);"
		print "N - numero de nucleotideos que serao inseridos em
arquivo_saida, para cada sequencia."
		return 0
	corte = int(argv[3])
	sentido = argv[4].upper()
	entrada = argv[1]
	saida = argv[2]
	dados = leitura_arquivo(entrada, saida, corte, sentido)
	return 1

if __name__ == "__main__":
	main(sys.argv)
```