UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

Renato Vicentini dos Santos

Localização subcelular de proteínas de cana-de-açúcar (*Saccharum* spp.): caracterização *in silico* e avaliação funcional

Tese apresentada ao Instituto de Biologia para obtenção do Título de Doutor em Genética e Biologia Molecular, na área de Genética Vegetal e Melhoramento

Orientador: Prof. Dr. Marcelo Menossi Teixeira

Campinas, 2008

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DO INSTITUTO DE BIOLOGIA – UNICAMP

Sa59L

Santos, Renato Vicentini dos
Localização subcelular de proteínas de cana-de-açúcar (*Saccharum* spp.): caracterização *in silico* e avaliação funcional / Renato Vicentini dos Santos. – Campinas, SP: [s.n.], 2008.

Orientador: Marcelo Menossi Teixeira.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Localização subcelular. 2. Proteínas – Tradução. 3. Cana-de-açúcar. I. Teixeira, Marcelo Menossi. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

(rcdt/ib)

Título em inglês: Subcellular localization of sugarcane (*Saccharum* spp.) proteins: *in silico* characterization and functional evaluation.
Palavras-chave em inglês: Subcellular localization; Proteins - Translation; Sugarcane.
Área de concentração: Genética Vegetal e Melhoramento.
Titulação: Doutor em Genética e Biologia Molecular.
Banca examinadora: Marcelo Menossi Teixeira, Michel Georges Albert Vincentz, Carlos Augusto Colombo, Marcio de Castro Silva Filho, Marco Aurélio Takita.
Data da defesa: 03/06/2008.
Programa de Pós-Graduação: Genética e Biologia Molecular.

Campinas, 03 de junho de 2008

BANCA EXAMINADORA

Prof. Dr. Marcelo Menossi Teixeira (Orientador) _____
Assinatura

Prof. Dr. Michel Georges Albert Vincentz _____
Assinatura

Prof. Dr. Carlos Augusto Colombo _____
Assinatura

Prof. Dr. Marcio de Castro Silva Filho _____
Assinatura

Prof. Dr. Marco Aurelio Takita _____
Assinatura

Prof. Dr. José Andrès Yunes _____
Assinatura

Prof. Dr. Celso Eduardo Benedetti _____
Assinatura

Prof. Dr. Marcelo Carnier Dornelas _____
Assinatura

À meu filho Pedrinho
com amor e agradecimento
pelos bons-dias
repletos de alegria
e pelas boas-noites
recheadas de gratidão.

Agradecimentos

Gostaria de expressar minha profunda gratidão às várias pessoas que me deram o prazer da convivência e de seu apoio durante os três anos em que trabalhei nesta tese. Não tenho como citar todos, mas entre eles estão Rodrigo Drummond, Marcelo Mattioli, Eduardo Mariano, Juliana Felix, Antonio Paulino da Costa Netto, e a todos os membros e ex-membros do Laboratório de Genoma Funcional. Esta tese deve algo aos vários amigos do Centro de Biologia Molecular e Engenharia Genética, por terem tido a generosidade de compartilhar aquela época comigo. A alegria e o prazer foram todos meus. Este ainda insipiente cientista agradece ao prof. Paulo Arruda, a Felipe Rodrigues da Silva, Andre Vettore, Edson Kemper, e aos que estavam presentes naquele raro momento no final da década de noventa e início dos anos dois mil. Aquele ambiente inspirador e estimulante define muito do que sou hoje, e faço valer a prerrogativa de autor para não estragar a poesia vivida na época; como diria García Márquez: "A vida não é a que a gente viveu, e sim a que a gente recorda".

Aos meus queridos amigos e jovens cientistas Almir Zanca, Thaís Rezende e Silva, Natalia Cristina Verza, Sylvia Souza, Mario Paniago, e Eduardo Kiyota, por possibilitarem as mais elevadas discussões a cerca do nosso mundo. E à meu irmão de coração, Evandro Luis Salvador, a quem sempre recorri quando faltava um pouco de humanidades em minha universidade.

Agradeço em especial ao prof. Marcelo Menossi, que mais uma vez, provou ser não só um grande mentor, mas também um grande amigo, oferecendo-me suas opiniões e seu apoio. E principalmente, me empurrando montanha abaixo quanto ainda não sabia que tinha asas para voar. Um agradecimento sempre em débito aos diversos professores do Instituto de Biologia, que nunca fizeram ressalvas ao fato de um ainda estudante de Engenharia da Computação cursar regularmente suas excelentes disciplinas.

Por vários motivos sou grato a Eugênio Ulian, Sabrina Chabregas e Maria Cristina Falco, que além de serem minhas referências em cana-de-açúcar me apresentaram a um mundo que não imaginava ser ainda possível. Agradeço aos professores Michel Vincentz, José Andrès Yunes, e Marcio de Castro pelas discussões e sugestões valiosas durante o transcorrer do projeto ou na banca prévia. Assim como aos professores Carlos Colombo, Marco Aurelio Takita, Vicente Eugenio De Rosa Junior, Antonio Vargas de Oliveira Figueira, e novamente aos professores Michel Vincentz, José Andrès Yunes, e Marcio de Castro pela disposição em participar da banca de tese.

Enfim, minha gratidão vai a meus pais, pelo apoio irrestrito, pelo porto seguro, e pela transmissão do atualmente desvalorizado valor da educação. Assim como a meus irmãos Carolina e Rodrigo, sempre sabendo que a distância geográfica nunca será capaz de nos afastar.

Acima de tudo, quero expressar meu profundo agradecimento à minha esposa, Juliana Felix, que sempre com carinho me orientou ao longo das minhas hesitações e questionamentos, e à meu filho, Pedro, por sua paciência durante seus primeiros anos, quando, algumas vezes, deixei sua companhia para "subir ao andar de cima"e passar horas escrevendo.

Quase tudo o que nós fazemos está escrito na areia e apaga-se com o vento. Ainda assim, pode acontecer de termos um tablete de metal sobre o qual nós escreveremos um ou dois sinais mais duráveis.

Hendrik Casimir (1909-2000)

Este livro está escrito sobre areia. Mas a praia é bela e eu não me arrependo de ter ido passear nela.

Pierre-Gilles de Gennes (1932-2007)

Resumo

As células de plantas são altamente organizadas e muitos processos biológicos estão associados com estruturas subcelulares específicas. A localização subcelular é uma característica chave das proteínas, visto que está relacionada com a função biológica. A determinação da localização subcelular usando a predição é uma estratégia altamente desejável, principalmente porque as abordagens experimentais demandam um tempo considerável. Com o objetivo de desenvolver um método para melhorar a predição de localização subcelular, diversos algoritmos foram integrados visando à exploração ótima do potencial de cada um. O desempenho com 90% de exatidão deste novo método foi claramente superior a todos os métodos utilizados em sua criação. Usando esta estratégia foi realizada a primeira análise em larga escala da localização subcelular do proteoma de cana-de-açúcar (com 11.882 proteínas preditas), sendo encontrado que a maioria das proteínas estão localizadas em quatro compartimentos: núcleo (44%), citoplasma (19%), mitocôndria (12%), e extracelular (11%). Adicionalmente foi observado que cerca de 19% das proteínas são localizadas em múltiplos compartimentos. Outros resultados foram capazes de identificar um conjunto de proteínas de cana-de-açúcar que podem apresentar duplo direcionamento pelo uso de variações na extremidade amino. Utilizando expressão transiente em células da epiderme de cebola, foi investigada a localização subcelular de 96 proteínas de cana com fusão a proteína GFP. As construções contendo fusão amino- e carboxi-terminal dos genes foram expressas, e a localização das proteínas de fusão foi detectada por microscopia de fluorescência. É relatado também a caracterização do gene *ScBAK1*, um receptor do tipo quinase com repetições ricas em leucina, que apresenta similaridade de seqüência com o gene *brassinosteroid insensitive1-associated receptor kinase1*. Foi mostrado que transcritos desse gene se acumulam em níveis muito mais altos nas células da bainha do feixe vascular do que nas células do mesófilo, e que a fusão ScBAK1-GFP é localizada na membrana plasmática. Essa distribuição espacial e esse padrão de expressão indicam que a ScBAK1 pode estar potencialmente envolvida em cascatas de sinalização celular intermediadas por altos níveis de açúcar na folha. Ainda considerando estudos de localização subcelular, é conhecido que seqüências de nucleotídeos que flanqueiam o códon de início da tradução afetam a eficiência traducional dos mRNA, e podem indicar a presença de sítios de inicio de tradução (TIS) alternativos. O multi-direcionamento pode ser um reflexo da variabilidade traducional destas outras formas da proteína. Neste estudo foi desenvolvido um método computacional para investigar o uso de TISs alternativos na síntese de novas variantes protéicas que podem apresentar localização subcelular diferente. Visando contribuir para o nosso entendimento da complexidade do genoma da cana-de-açúcar, foi empregada uma análise em larga escala dos TIS nesta espécie. Também é demonstrado que os transcritos com expressão induzida apresentam um forte TIS quando comparados com os reprimidos, e que os transcritos constitutivos possuem uma alta freqüência de TIS alternativos. O mesmo ocorre para os genes com altas taxas evolutivas, e transcritos específicos de folhas e entrenós, levantando a hipótese de que esses genes possam codificar diferentes polipeptídeos.

Abstract

Plant cells are highly organized and many biological processes are associated with specialized subcellular structures. Subcellular localization is a key feature of proteins, since it is related to biological function. The determination of subcellular localization using computational prediction is a highly desirable strategy because experimental approaches are time-consuming. In order to develop a method for the enhanced prediction of subcellular localization, the outputs of some prediction tools were integrated so as to optimally exploit the potential of each one. The prediction performance (with 90% of accuracy) of this new method was clearly superior to all the methods used to create the predictor. Using this method, the first *in silico* genome-wide subcellular localization analysis was performed for sugarcane (with 11,882 predicted proteins). It was found that most of the proteins are localized to four compartments: nucleus (44%), cytosol (19%), mitochondria (12%), and secretory destination (11%). It is also shown that about 19% of the proteins are localized to multiple compartments, and that a potential set of sugarcane proteins can show dual targeting by use of N-truncated form of proteins. The subcellular localization of 96 sugarcane proteins fused with GFP were evaluated using transient expression in onion epidermal cells. Constructs containing the N- and C-terminal fusion of genes encoding both endogen and GFP proteins were transiently expressed, and the localization of the fusion proteins were detected by fluorescent microscopy. It was reported the characterization of *ScBAK1*, a sugarcane leucine-rich repeat receptor-like kinase, with sequence similarity to *brassinosteroid insensitive1-associated receptor kinase1*. We have found that *ScBAK1* transcripts accumulated at higher levels in bundle-sheath than in mesophyll cells. ScBAK1-GFP fusions were localized to the plasma membrane. This spatial distribution and expression pattern indicates that *ScBAK1* might be potentially involved in cellular signaling cascades mediated by high levels of sugar in this organ. The nucleotide sequence flanking the translation initiation codon affects the translational efficiency of eukaryotic mRNAs, and may indicate the presence of an alternative translation initiation site (TIS) to produce proteins with different properties. Multi-targeting may reflect the translational variability of these other protein forms. In this study it was also developed a computational method to investigate the usage of alternative TISs for the synthesis of new protein variants that might have different subcellular localization. To contribute to our understanding of the genome complexity of sugarcane, we undertook a genome wide TIS analysis in sugarcane data. It is demonstrated that up-regulated transcripts show a stronger TIS when compared with the down-regulated, and that ubiquitous transcripts have a high frequency of alternative TIS in the next downstream AUG codon. The same occurs for fast-evolving genes, and leaf and internodes specific transcripts, that may encode different polypeptides by N-terminal polymorphism.

Lista de Figuras

Lista de Tabelas

Sumário

There are no applied sciences; there are only applications of science. The study of the application of science is very easy to anyone who is master of the theory.

Louis Pasteur (1822-1895)

1 Introdução

A cana-de-açúcar é uma das mais importantes espécies vegetais cultivadas em todo o mundo, com uma área maior que 20 milhões de hectares em 101 países (FAO 2005; http://apps.fao.org), sendo que o Brasil é o maior produtor mundial, participando com 25% da produção. A cana-de-açúcar (*Saccharum* spp.) faz parte de um complexo poliplóide pertencente à tribo Andropogoneae, da família Gramineae (Poaceae). As variedades atualmente cultivadas de cana-de-açúcar são resultado de hibridação interespecífica envolvendo as espécies *S. officinarum*, a qual contribui com o alto teor de açúcar, e *S. spontaneum*, responsável pelo vigor vegetativo e resistência a estresses bióticos e abióticos (Ming *et al.*, 2001). Devido a sua capacidade única de estocar sacarose em seus entrenós, a cana-de-açúcar tornou-se uma importante fonte de informação no que se diz respeito à síntese de sacarose, seu transporte e acúmulo.

Devido à sua origem interespecífica, a cana-de-açúcar possui um dos mais complexos genomas vegetais, apresentando um variável nível de ploidia, dificultando a aplicação de técnicas convencionais no melhoramento genético da espécie. A identificação de genes responsáveis por qualidades agronomicamente desejáveis e sua posterior manipulação por meio de técnicas de biologia molecular podem proporcionar a obtenção de variedades bem sucedidas, reduzindo drasticamente as perdas na agricultura, além de permitir o aproveitamento de solos até então não utilizáveis. Além disso, devido aos altos níveis de similaridade genética entre as gramíneas (Guimaraes *et al.*, 1997, Paterson *et al.*, 2000, Paterson *et al.*, 2004), os estudos realizados em cana-de-açúcar podem ser extrapolados para outras gramíneas cultivadas.

Organelas em células de plantas

Durante a evolução, as células eucarióticas desenvolveram um complexo sistema de endomembranas que deu origem a uma grande variedade de compartimentos subcelulares. A constituição destes microambientes subcelulares, denominados organelas, é que as tornaram capazes de desempenhar processos fisiológicos e bioquímicos específicos (Figura 1.1).

As células das plantas constituem um interessante sistema no qual a informação genética encontra-se localizada em 3 diferentes compartimentos intracelulares: núcleo, plastídeos e mitocôndrias. No entanto, a capacidade codificante dos genomas de mitocôndrias e dos plastídeos é muito limitada, uma vez que seus genes foram perdidos ou transferidos para o núcleo durante a evolução. No caso das mitocôndrias, a origem está relacionada a um evento de endossimbiose do ancestral de uma α-proteobactéria. Já um segundo evento endossimbiótico, no caso uma cianobactéria, originou a organela plastidial.

O perfeito funcionamento de uma organela depende de seu conteúdo, que é definido de acordo com as suas funções. Sendo assim, um sistema de importação seletivo e eficiente é fundamental para a manutenção da identidade funcional e estrutural dos diferentes compartimentos subcelulares. Apesar de cloroplastos e mitocôndrias possuírem seus próprios genomas, a maioria das proteínas que eles contêm são codificadas por genes nucleares e precisam ser importadas do citoplasma, seu local de tradução (Adams *et al.*, 2000, Bauer *et al.*, 2001, Duby and Boutry, 2002). Como exemplo temos os cloroplastos, que ainda retêm uma reduzida capacidade de codificar entre 100 a 200 proteínas de um total de cerca de 3200 presentes em cianobactérias, seus ancestrais mais prováveis (Bruce, 2000). Assim, cerca de 90% das proteínas cloroplásticas são codificadas por genes nucleares e sintetizadas nos ribossomos citosólicos sob a forma de grandes precursores (denominadas preproteínas).

Importação e direcionamento de proteínas em células de plantas

As proteínas codificadas por genes nucleares são transportadas para os compartimentos subcelulares graças a uma complexa maquinaria (Figura 1.2). Schatz & Dobberstein (1996) cunharam os termos importação e exportação para descrever dois diferentes tipos de

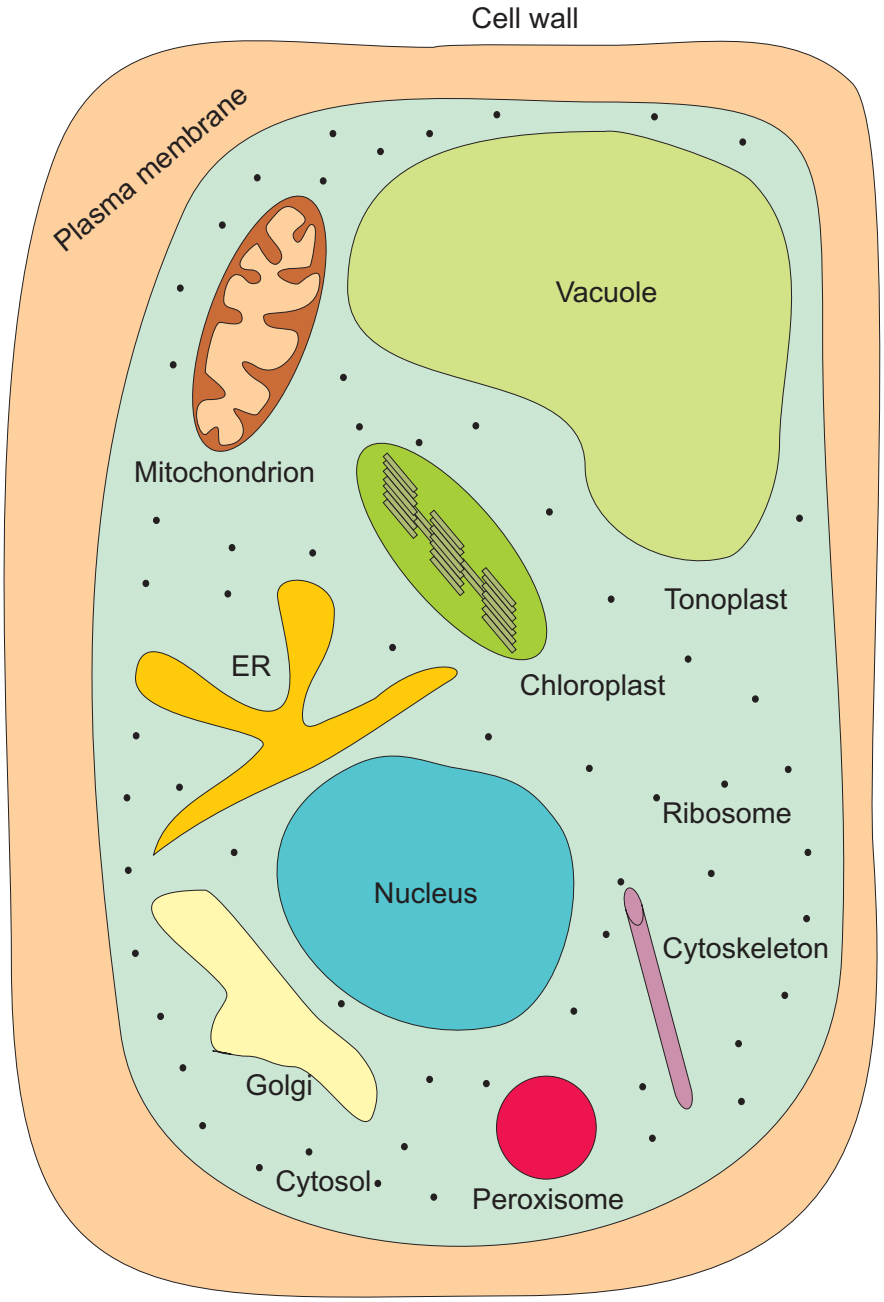


Figura 1.1: **Representação esquemática de uma célula de planta, demonstrando as principais organelas do sistema de endomembranas, assim como a parede celular.**

translocação de proteínas através das membranas celulares. De acordo com essa classificação, o processo de importação é aquele que transporta proteínas para o interior de organelas que se apresentam envoltas por membranas. Por outro lado, o processo de exportação transporta proteínas para o espaço extracelular.

A eficiência do processo de importação de uma proteína reside nas interações que ocorrem entre a sua pré-seqüência e os aparatos de translocação presentes nas membranas da organela e no citosol. O endereçamento das proteínas, porém, é mais complexo do que a simples presença de seqüências de direcionamento. Como no caso da mitocôndria, ele envolve vários subcompartimentos. As proteínas direcionadas para cada um desses subcompartimentos requerem informações de direcionamento específicas e vias de direcionamento que envolve fatores comuns e distintos a cada etapa (Duby and Boutry, 2002). Alem disto, fatores ambientais ou modificações pós-traducionais podem interferir na localização final de uma proteína na célula (Rodriguez-Concepcion *et al.*, 1999, Anandatheerthavarada *et al.*, 1999, Kircher *et al.*, 2002, Silva-Filho, 2003).

Geralmente, o processo de importação de proteínas é específico para cada organela (Boutry *et al.*, 1987, Schmitz and Lonsdale, 1989, Chaumont *et al.*, 1990, Silva *et al.*, 1996, Barata *et al.*, 2000). No entanto, casos de direcionamento inespecífico podem ser encontrados na literatura (Hurt *et al.*, 1986, Pfaller *et al.*, 1989, Franzen *et al.*, 1990, Huang *et al.*, 1990, Brinks *et al.*, 1994, Chow *et al.*, 1997, Silva-Filho *et al.*, 1997), onde a maioria é causada por seqüências de direcionamento atípicas.

Sinais de direcionamento e sítios de clivagem

A maioria das proteínas em células eucarióticas é codificada no genoma nuclear e sintetizada no citosol, e muitas necessitam posterior direcionamento a um ou outro compartimento subcelular. Quando o destino final é o cloroplasto, a mitocôndria ou as vias secretoras, o endereçamento geralmente requer a presença de uma seqüência etiquetada na região amino-terminal que é reconhecida pela maquinaria de translocação (Emanuelsson *et al.*, 2000). Na maioria dos casos, a seqüência etiquetada é proteoliticamente removida durante ou após sua entrada na organela. Para futuros endereçamentos dentro da organela, informações adicionais podem estar localizadas em seqüências etiquetadas secundárias, que podem estar adjacentes às seqüências sinal originais ou em outras regiões da proteína.

As proteínas direcionadas para cada compartimento subcelular possuem seqüên-

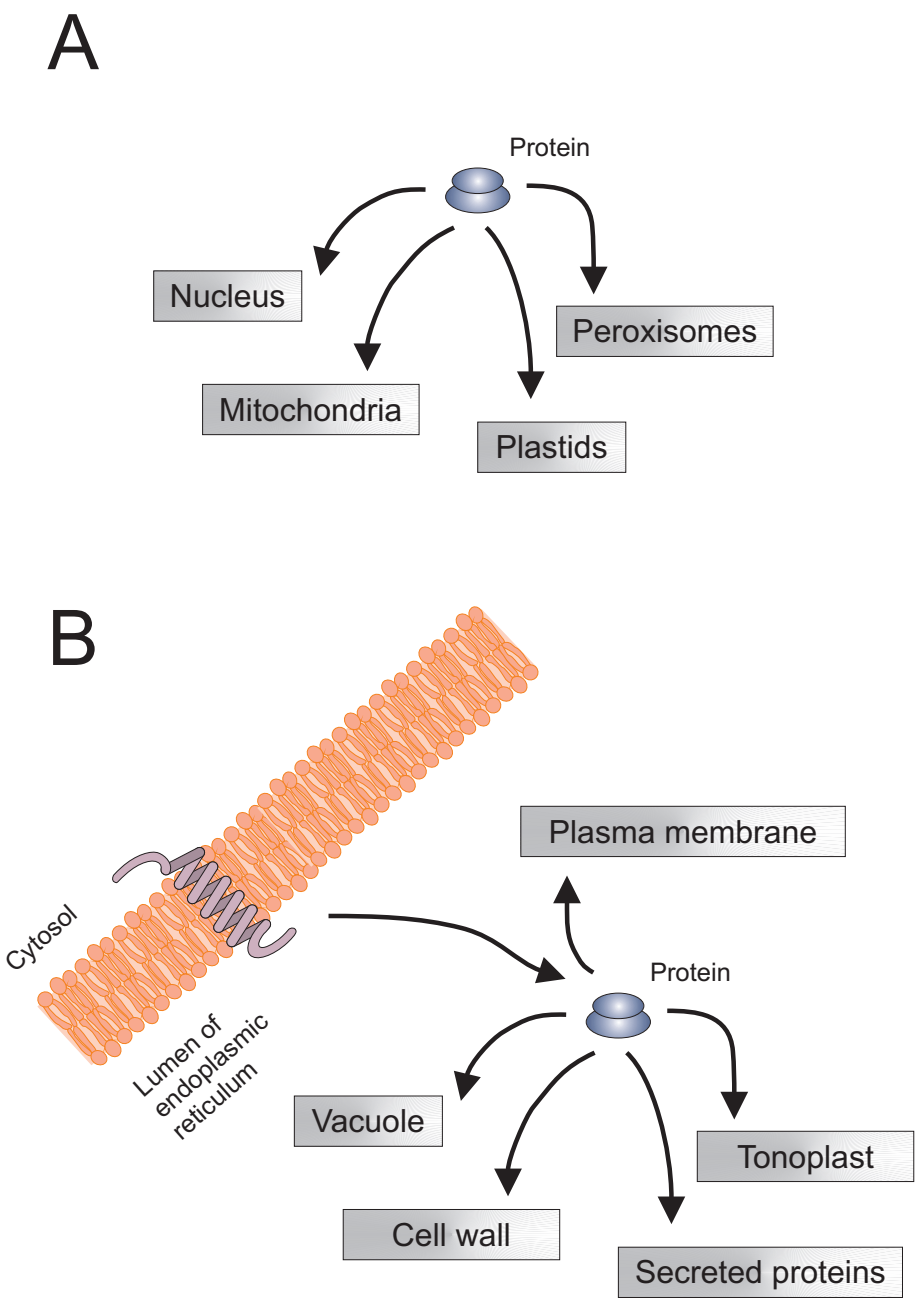


Figura 1.2: **Visão esquemática do mecanismo de direcionamento protéico.** (A) Proteínas sintetizadas nos ribossomos livres permanecem no citosol ou são direcionadas para o núcleo, mitocôndria, plastídeos ou peroxissomos. (B) Já as proteínas sintetizadas no ribossomos ligados a membrana do retículo endoplasmático são primeiro translocadas para o lumem e depois transportadas para o complexo de Golgi.

cias características de cada compartimento:

- Cloroplastos: apresentam um peptídeo de trânsito, também amino-terminal, com estrutura secundária do tipo "espiral ao acaso"(von Heijne *et al.*, 1989, Claros and Vincens, 1996, Soll and Tien, 1998);

- Mitocôndrias: as proteínas apresentam na extremidade amino terminal uma seqüência de direcionamento em α-hélice anfifílica, composta por aminoácidos hidrofóbicos de um lado, e do outro, aminoácidos carregados positivamente (Glaser *et al.*, 1998);

- Núcleo: apresentam pequenos sinais de 4 a 8 aminoácidos básicos localizados em diferentes posições da cadeia polipeptídica (BarPeled *et al.*, 1996);

- Peroxissomos: na maioria dos casos é caracterizado pela seqüência SKL (serina, lisina e leucina) carboxi-terminal (Subramani, 1996);

- Via secretora (incluindo retículo endoplasmático, complexo de Golgi, membrana plasmática, vesículas de transporte, vacúolos e meio extracelular): duas situações são possíveis: clivagem do peptídeo sinal na cadeia nascente e as âncoras sinais, que não são clivadas liberando as proteínas na membrana do retículo endoplasmático, existindo também as partículas de reconhecimento do sinal (SRPs) (BarPeled *et al.*, 1996, Corsi and Schekman, 1996);

Os cloroplastos são organelas complexas, apresentando seis diferentes subcompartimentos (membrana externa, membrana interna, espaço intramembranas, estroma, membrana do tilacóide e lúmen). A seqüência de direcionamento para os cloroplastos denomina-se peptídeo de trânsito e apesar de apresentar um enriquecimento de Ser, Thr e aminoácidos positivos (notadamente Arg), não apresenta uma estrutura secundária característica (von Heijne *et al.*, 1989). O tamanho varia de 20 a 120 resíduos e apresentam basicamente três regiões distintas (Bruce, 2000): uma região amino-terminal não carregada de aproximadamente 10 resíduos, geralmente começando com metionina seguida de alanina e terminando com glicina e prolina; um domínio central, onde faltam resíduos ácidos, mas enriquecido de serina e treonina e uma região carboxi-terminal rica em argininas e que potencialmente forma uma estrutura β-pregueada anfifílica, com aproximadamente 10 resíduos.

Assim como no caso dos cloroplastos, as mitocôndrias apresentam localizações distintas para suas proteínas: membrana externa, espaço intermembranas, membrana interna

e matriz mitocondrial. Muitas proteínas precursoras mitocondriais sintetizadas no citoplasma são reconhecidas e/ou mantidas em uma forma pouco estruturada, via ação das chaperonas moleculares, e importadas por complexos de translocação presentes nas membranas externa e interna. No peptídeo alvo para a mitocôndria, Arg, Ala e Ser estão sobre-representadas, enquanto que resíduos de aminoácidos de carga negativa (Asp e Glu) são raros. Apenas uma fraca seqüência conservada tem sido encontrada, uma Arg nas posições -2 ou -3 relativas ao sitio de clivagem pela peptidase de processamento mitocondrial. Algumas proteínas de matriz são clivadas uma segunda vez pela peptidase intermediária mitocondrial que remove adicionalmente de oito a nove resíduos da proteína madura. Variações raras destes mecanismos também podem ocorrer (Emanuelsson *et al.*, 2000).

O intercambio entre núcleo e citoplasma ocorre através de poros presentes no envelope nuclear, denominado complexo de poros. Este atua como uma peneira seletiva e serve de caminho para o transporte bidirecional de macromoléculas. No caso de proteínas com massa molecular superior a 40.000 este transporte é dependente de sinais de localização nuclear (NLS, *nuclear localization signals*). Estes sinais são reconhecidos por receptores que facilitam a passagem da proteína através do canal central do complexo (Kutay and Muhlhausser, 2006). Os NLSs em geral possuem resíduos de carga positiva em abundância, no entanto existem motivos ricos em glicina com pouca carga positiva. Eles podem ser descritos como possuidores de motivos únicos ou duplos. Os motivos únicos são caracterizados por agrupamentos de aminoácidos básicos precedidos por um aminoácido prolina ou glicina. Já os motivos duplos consistem de dois agrupamentos de aminoácidos básicos separados por 9-12 aminoácidos (Cokol *et al.*, 2000).

Os peroxissomos adquirem suas proteínas do citoplasma por dois caminhos distintos, cada um utilizando um sinal conservado de direcionamento para o peroxissomo (PTS, *peroxisomal targeting signals*). O PTS1 é o mais comum, consistindo de apenas três aminoácidos na extremidade carboxi-terminal da proteína, notadamente SKL. Já o PTS2 é composto por um sinal com duas partes, onde o consenso é a seqüência [RK]-[LVI]-x5-[HQ]-[LA]. A maquinaria de importação é realizada por diversas proteínas denominadas peroxinas. O PST1 interage com as repetições de tetratricopeptídeo (TPRs, *tetratricopeptide*) do receptor Pex5p. Já as proteínas carregando o sinal PST2 se ligam aos motivos WD40 das proteínas Pex7p. Após esta etapa de ligação aos receptores, os dois mecanismos PST parecem convergir para um mesmo caminho de direcionamento (Emanuelsson *et al.*, 2003).

Os peptídeos sinais são os responsáveis pelas proteínas que serão enviadas ao ER

para transporte subseqüente através das vias secretoras, consistindo geralmente em três regiões: uma região amino-terminal de carga positiva, uma central hidrofóbica, e uma carboxi-terminal polar que carrega o sinal para o sítio de clivagem pela peptidase. O motivo mais conservado é a presença de um aminoácido neutro na posição -3 e -1 em relação ao sitio de clivagem (Emanuelsson *et al.*, 2000).

Duplo direcionamento de proteínas

Certos processos metabólicos são realizados em mais de um local dentro da célula como, por exemplo, a síntese de DNA e RNA no núcleo, na mitocôndria e no cloroplasto e a β-oxidação de ácidos graxos nos peroxissomos e mitocôndrias de células animais (Danpure, 1995). Para estes casos, a célula eucariótica tem desenvolvido uma série de mecanismos de direcionamento de proteínas com funções semelhantes para as diferentes localizações intracelulares.

Evidências têm sido fornecidas (Chabregas, 2001, Small *et al.*, 1998, Silva-Filho, 2003) ao que tange o múltiplo direcionamento de isoformas organelares codificadas por um único transcrito nuclear. Experimentos *in vitro* indicaram a ocorrência de dois produtos da tradução a partir de diferentes códons de iniciação em fase de leitura. Por exemplo, o endereçamento de uma dada proteína para o cloroplasto é realizado pelo primeiro AUG, sendo que o segundo AUG é responsável pelo endereçamento para a mitocôndria. Outras observações permitem também a hipótese de que o peptídeo de trânsito cloroplástico seja capaz de direcionar proteínas a mitocôndria e cloroplasto simultaneamente. Temos como exemplo o direcionamento simultâneo a estas organelas da enzima glutationa redutase, envolvida no sistema de resposta ao estresse oxidativo (Creissen *et al.*, 1995).

Dentre os mecanismos possíveis de duplo direcionamento, destacam-se:

Iniciação alternativa da transcrição

Um único gene apresenta diferentes sítios de início da transcrição e conseqüentemente vários transcritos podem ser diferenciados devido à presença ou ausência de uma seqüência de direcionamento amino-terminal. Este é, por exemplo, o caso da Hsp70 plastidial e glioxissomal de cotilédones de melancia (Wimmer *et al.*, 1997).

Pré-mRNA variável

Um mRNA pode ser processado de várias formas, resultando em diferentes transcritos que possuem informações de direcionamento distintas localizadas em qualquer região das proteínas.

Um único peptídeo ambíguo

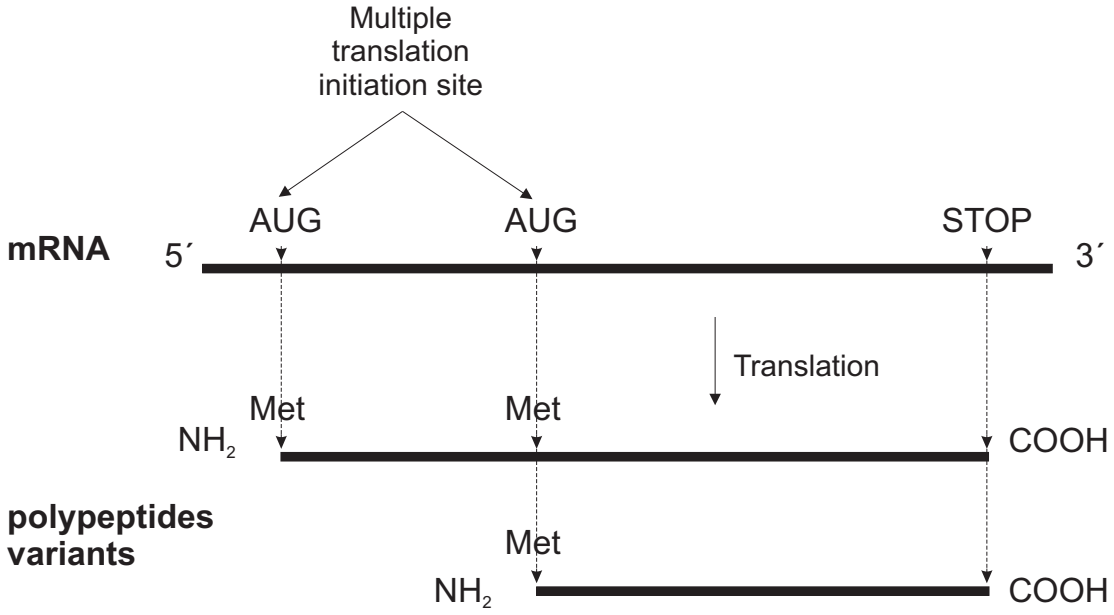
Existem ainda casos em que uma única seqüência de direcionamento é reconhecida pela maquinaria de importação de duas organelas diferentes. É o que acontece com as formas mitocondriais e cloroplásticas da glutationa redutase de ervilha, demonstradas *in vivo* (Creissen *et al.*, 1995).

Sítios de inicio de tradução polimórficos

A tradução por ribossomos citoplasmáticos geralmente corre a partir do primeiro AUG presente no transcrito. No entanto, em mRNAs eucarióticos, a eficiência de reconhecimento do códon AUG como sítio de inicio de tradução (TIS, *translation initiation site*) depende de diversos fatores, como a seqüência de nucleotídeos ao redor deste sítio (Kozak, 1991, Kozak, 2002). Existem evidências de que o contexto ao redor do códon de iniciação contribui para o controle do inicio da tradução (Kawaguchi and Bailey-Serres, 2005). A seqüência contexto do primeiro códon AUG, em particular a parte localizada na região não traduzida, pode modular a eficiência na qual ele é reconhecido como códon de inicio de tradução (Mignone *et al.*, 2002). Se o primeiro códon de iniciação apresentar um contexto apropriado, a síntese protéica será iniciada a partir dele. Quando o contexto se apresenta menos favorável, a maior parte da síntese protéica será iniciada pelo próximo códon AUG *downstream* (Nadershahi *et al.*, 2004). Alem disto, outras características no mRNA são consideradas importantes para a eficiência do inicio da tradução a partir de um códon AUG específico, sendo elas: a proximidade do AUG em relação a extremidade 5´, a estrutura secundaria *upstream* e *downstream* do códon AUG, o tamanho da seqüência *leader*, e a presença de múltiplos códons AUG *upstream* (Kozak, 1991, Wang and Rothnagel, 2004).

Neste cenário (Figura 1.3), múltiplas formas protéicas podem ser obtidas de um mesmo mRNA (Mignone *et al.*, 2002). Desta forma, AUGs com localização *downstream* podem ter papel na geração de diversidade protéica (Kozak, 2002). O uso de um códon AUG

como TIS alternativo pode resultar em uma proteína truncada em sua extremidade amino, podendo assim apresentar a mesma função e ser direcionada para uma localização diferente da proteína completa (Watanabe *et al.*, 2001, Kochetov and Sarai, 2004, Kochetov, 2005). A determinação *in silico* da localização subcelular de proteínas é dependente da correta identificação do primeiro AUG e de suas potênciais formas polimórficas na extremidade amino. Este tipo de polimorfismo de tradução pode servir como uma importante fonte de diversidade nos proteomas citoplasmáticos e organelares (Kochetov and Sarai, 2004, Kochetov, 2005).



Adaptado de Danpure, C.J. Trends in Cell Biology 1995, 5:230-238

Figura 1.3: **Início alternativo da tradução.** Um único gene com apenas um tipo de trascrito, contendo dois sítios de inicio da tradução *in frame*. A tradução apartir do primeiro sítio produz um polipeptídeo contendo a seqüência de direcionamento N-terminal, enquanto a tradução a partir do segundo sítio produz um polipeptídeo sem essa seqüência de direcionamento.

Predição das seqüências codificantes completas e da localização subcelular baseada em seqüência protéica

A predição gênica tem recebido considerável atenção nos últimos anos (Burge and Karlin, 1997, Claverie, 1998, Hiller *et al.*, 2006, Saeys *et al.*, 2007). Esses esforços têm trazido muitos progressos, mas o problema está longe de uma solução satisfatória. Os métodos atuais podem ser agrupados em duas categorias principais, *ab initio* e baseados em homologia. O primeiro método reconhece sinais ou características de composição em uma seqüência por reconhecimento de padrões, probabilidades ou métodos estatísticos, utilizado, por exemplo, no software GENSCAN (Burge and Karlin, 1997). Os métodos baseados em homologia usam informações externas para a comparação de seqüências, e na maioria das vezes apenas indicam aproximadamente as localizações de exons codificantes, não delimitando suas extremidades (Rogic *et al.*, 2001).

Um passo importante na análise de informações genômicas é decifrar a seqüência codificante completa (CDS, *complete coding sequence*) de cada gene. O CDS é a seqüência de nucleotídeo que corresponde à seqüência de aminoácido da proteína, começando tipicamente com o ATG e terminando com o *stop codon*. Existem muitos fatores que complicam a predição de CDS a partir de seqüências de cDNA, particularmente quando esta predição é realizada em larga escala. Em primeiro lugar, a predição de sítios de início no mRNAs de eucariotos é consideravelmente mais complicada do que em procariotos. Em segundo lugar, é difícil de distinguir mRNAs que não codificam, de mRNA que codificam para proteínas muito pequenas. E por último, a qualidade da seqüência pode interferir significativamente na predição do CDS.

Já no que tangem os métodos de predição da localização subcelular, estes podem ser classificados em duas classes: uma baseada em sinais de endereçamento amino-terminal e a outra baseada na composição de aminoácidos. Uma das vantagens do primeiro método é a clara implicação biológica. Contudo em projetos de análise de genes em larga escala, nem sempre a região 5' das seqüências está disponível (Reinhardt and Hubbard, 1998) ou estão apenas parcialmente acessíveis, o que pode causar problemas na predição dependendo do algoritmo utilizado (Cui *et al.*, 2004).

A habilidade dos vários programas em predizer corretamente a localização subcelular tem se demonstrado bem abaixo do esperado. Com isso tem-se buscado a sobreposição de predição por dois ou mais programas distintos, aumentando-se assim a confiabilidade da

predição (Heazlewood *et al.*, 2004).

Determinação da localização subcelular em larga escala com uso da proteína GFP

A proteína fluorescente verde (GFP, *green-fluorescent protein*) de *Aequorea victoria*, assim como suas variantes que fluorescem em outras faixas de emissão de luz, vem sendo utilizada como um repórter universal em ampla gama de células e organismos heterólogos (Chiu *et al.*, 1996). As novas tecnologias de clonagem permitem gerar rapidamente cDNAs fusionados com o gene *gfp* tanto na extremidade amino quanto na carboxi-terminal. Essa metodologia possibilita o exame da localização subcelular em larga escala, sendo que em cerca de 80% das construções gênicas é possível detectar claramente a localização intracelular (Simpson *et al.*, 2000).

Um problema freqüente na utilização da GFP refere-se à extremidade da molécula de interesse a qual ela está fusionada. Contudo, atualmente os métodos de clonagem permitem gerar tanto fusões amino quanto carboxi-terminal em uma única reação, possibilitando assim identificar imediatamente qualquer efeito que a GFP pode estar causando no sinal de direcionamento da proteína. Exemplos demonstram que proteínas com direcionamento especificamente mitocondrial são detectadas no núcleo e no citosol caso apresentem a proteína GFP fusionada em sua extremidade amino, demonstrando claramente que a perturbação no sinal de direcionamento mitocondrial acaba com a habilidade de estas proteínas serem localizadas corretamente (Simpson *et al.*, 2000).

Diversos grupos vêm utilizando com sucesso estes novos sistemas de clonagem, obtendo centenas e até mesmo milhares de construções viáveis (Simpson *et al.*, 2000, Reboul *et al.*, 2003, Palmer and Freeman, 2004, Rual *et al.*, 2004). Os resultados sugerem que apenas uma minoria dos eventos de expressão transiente sofre interferência destes sistemas de recombinação e ressaltam que a determinação da localização subcelular de novas proteínas pode se valer destas estratégias com alto grau de confiabilidade (Simpson *et al.*, 2000).

O transcritoma da cana-de-açúcar

O projeto EST da cana-de-açúcar (SUCEST) disponibilizou um conjunto básico e fundamental de dados para um maior entendimento dos processos fisiológicos e bioquímicos da cana-de-açúcar. Foram gerados 43.141 prováveis transcritos de cana-de-açúcar (SASs, *Sugarcane Assembled Sequences*), dos quais atualmente cerca de 29,7% não apresentam nenhuma homologia com seqüências gênicas ou protéicas previamente identificadas em qualquer outro organismo, o que pode significar tanto mRNAs não traduzidos quanto novos genes específicos de cana-de-açúcar (Vettore *et al.*, 2003).

Objetivos

Objetivo geral

Caracterizar e avaliar proteínas de cana-de-açúcar (*Saccharum* spp.) no que se refere à localização subcelular, assim como a correlação desta com dados de expressão gênica. As hipóteses formuladas para tais problemas são as de que a função de uma proteína esta intimamente relacionada com sua localização subcelular; e a necessidade de determinado nível de expressão, em uma dada condição, acaba por refletir na especificidade da localização protéica.

Objetivos específicos

1. O estudo da localização subcelular das proteínas associadas com o acúmulo de sacarose da cana-de-açúcar, proteínas com padrões de expressão particulares, proteínas com função desconhecida, ou até mesmo sem identidade com outras proteínas já identificadas;

2. A elaboração de um método de bioinformática capaz de auxiliar na tomada de decisão em larga escala, para uma correta determinação da localização subcelular. Este objetivo está centrado na hipótese de que a combinação de diversos métodos de predição de localização é capaz de contornar a freqüente discrepância entre eles.

Apresentação da Tese

Esta tese está baseada em estudos realizados experimentalmente e computacionalmente visando à determinação da localização subcelular de proteínas de cana-de-açúcar. De forma concreta, foram realizados estudos experimentais visando identificar o perfil de expressão de genes de interesse, assim como ensaios de determinação da localização subcelular das proteínas codificadas por tais genes. Computacionalmente foram desenvolvidos métodos de predição, sendo que tais métodos foram aplicados aos dados existentes para cana-de-açúcar. Os capítulos 2, 3 e 9 foram publicados em revistas indexadas internacionais, e os capítulos 4, 5 e 7 estão submetidos à publicação em revistas do mesmo nível. Já os capítulos 6 e 10 estão em fase de revisão para submissão. Por último são apresentadas as conclusões gerais deste trabalho de tese.

Capítulos 2, 3 e 4

Diversos processos metabólicos envolvidos no crescimento, desenvolvimento e adaptação a variações ambientais são regulados por vias de transdução de sinal. Nestes três capítulos são apresentados os resultados das pesquisas visando identificar os perfis de expressão gênica das vias de transdução de sinal em diferentes tecidos de cana-de-açúcar, assim como em resposta a variações ambientais e em indivíduos contrastantes para teor de sacarose. Estas pesquisas forneceram diversos genes candidatos a terem sua localização subcelular predita computacionalmente (Capítulo 6) e avaliadas experimentalmente (Capítulo 7 e 8).

Capítulos 5 e 6

A predição da localização subcelular de proteínas de eucariotos é uma abordagem que vem sido utilizada com freqüência para explorar e dar maior significância a grande quantidade de dados genômicos existentes. No entanto a habilidade em predizer corretamente a localização subcelular tem se demonstrado abaixo do esperado nos métodos existentes. No capítulo 5 é apresentado um novo método computacional que mostra significativas melhoras na predição com base em predições previamente realizadas por dois ou mais métodos distintos. Já no capítulo 6 é mostrado o resultado da aplicação deste método nos dados de

cana-de-açúcar, permitindo assim, pela primeira vez, a elaboração de uma visão da compartimentalização do proteoma de uma célula de cana-de-açúcar.

Capítulos 7 e 8

A determinação da localização subcelular de uma proteína é capaz de fornecer valiosos detalhes da possível função desta. Estes dois capítulos apresentam os esforços realizados visando determinar experimentalmente a localização subcelular de uma vasta gama de proteínas de cana-de-açúcar, em sua maioria identificadas por estudos de expressão gênica (Capítulos 2, 3 e 4). Especificamente, no Capítulo 7 é apresentada a caracterização do gene *ScBAK1* de cana-de-açúcar, um receptor do tipo quinase rico em leucina, que é predominantemente expresso nas células da bainha do feixe vascular em folhas maduras, e que se apresenta potencialmente envolvido na cascata de sinalização celular intermediada por altos níveis de açúcar neste órgão. Já no capítulo 8 são apresentados os resultados dos esforços visando ampliar o estudo funcional de genes de cana-de-açúcar através do estabelecimento de uma rotina para a determinação da localização subcelular de suas respectivas proteínas.

Capítulos 9 e 10

Durante o desenvolvimento deste projeto de tese, surgiu a possibilidade de investigar um interessante fenômeno biológico ainda pouco estudado computacionalmente. Este fenômeno pode ser descrito como o uso alternativo de sítios para início da tradução protéica. No capítulo 9 é apresentado um método que baseado na teoria da informação, busca fornecer indícios de possíveis sítios alternativos para o início da tradução. Finalmente no capítulo 10 é apresentada uma caracterização dos padrões genômicos existentes em cana-de-açúcar no que se refere a esses sítios de início da tradução.

Se dermos aos fracos e deformados a capacidade de viver e de propagar sua espécie, enfrentaremos a perspectiva de um crepúsculo genético. Mas se os deixarmos morrer ou sofrer, mesmo podendo salvá-los ou ajudá-los, enfrentaremos a certeza de um crepúsculo moral.

Theodosius Dobzhansky (1900-1975)

2

Transcription profiling of signal transduction-related genes in sugarcane tissues†

Flávia Stal Papini-Terzi[1,*], Flávia Riso Rocha[1,*], Ricardo Zorzetto Nicoliello Vêncio[2], Kátia Cristina Oliveira[1], Juliana de Maria Felix[3,4], Renato Vicentini[4], Cristiane de Souza Rocha[4], Ana Carolina Quirino Simões[1], Eugênio César Ulian[5], Sônia Marli Zingaretti di Mauro[6], Aline Maria Da Silva[1], Carlos Alberto de Bragança Pereira[2], Marcelo Menossi[3,4] and Gláucia Mendes Souza[1]

†*DNA Research* 2005, **12**:27-38 (doi:10.1093/dnares/12.1.27)

*The first two authors contributed equally to this work.

§Accession numbers: The gene expression data was deposited in The Gene Expression Omnibus (GEO) Database under access numbers GSE1702 (series), GPL1375 and GPL1376 (platforms) and GSM29453 to GSM29506 (samples).

[1]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, Av. Prof. Lineu Prestes, 748 05508-900, São Paulo, Brasil

[2]Departamento de Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão, 1010 05508-090, São Paulo, Brasil

[3]Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, CP 6010, 13083-970, Campinas, Brasil

[4]Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, CP 6010, 13083-970, Campinas, Brasil

[5]Centro de Tecnologia Copersucar, CP 162, 13400-970, Piracicaba, São Paulo, Brasil

[6]Departamento de Tecnologia, Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal, Universidade Estadual Paulista, Centro de Estocagem Clones (BCCCenter) Via de Acesso Professor Paulo Donato Castellane, S/N 14870-000, Jaboticabal, São Paulo, Brasil

Abstract

A collection of 237,954 sugarcane ESTs was examined in search of signal transduction genes. Over 3,500 components involved in several aspects of signal transduction, transcription, development, cell cycle, stress responses and pathogen interaction were compiled into the Sugarcane Signal Transduction (SUCAST) Catalogue. Sequence comparisons and protein domain analysis revealed 477 receptors, 510 protein kinases, 107 protein phosphatases, 75 small GTPases, 17 G-proteins, 114 calcium and inositol metabolism proteins, and over 600 transcription factors. The elements were distributed into 29 main categories subdivided into 409 sub-categories. Genes with no matches in the public databases and of unknown function were also catalogued. A cDNA microarray was constructed to profile individual variation of plants cultivated in the field and transcript abundance in six plant organs (flowers, roots, leaves, lateral buds, and 1^{st} and 4^{th} internodes). From 1280 distinct elements analyzed, 217 (17%) presented differential expression in two biological samples of at least one of the tissues tested. A total of 153 genes (12%) presented highly similar expression levels in all tissues. A virtual profile matrix was constructed and the expression profiles were validated by real-time PCR. The expression data presented can aid in assigning function for the sugarcane genes and be useful for promoter characterization of this and other economically important grasses.

Introduction

The unraveling of signal transduction pathways is of strategic importance to the understanding of fundamental processes such as growth and development as well as cellular responses triggered by biotic and abiotic stresses. In recent years, the wealth of information related to signal transduction generated by several genome sequencing projects, coupled with the global transcription profiling of a diversity of organisms, has brought many aspects of signaling under scrutiny. Protein superfamilies, such as protein kinases and transcription factors, have been systematically classified and analyzed following their identification by the sequencing projects (Manning *et al.*, 2002*a*, Manning *et al.*, 2002*b*, Chen *et al.*, 2002, Gribskov *et al.*, 2001) and comparative studies of complete genomes are defining the conserved signaling modules and revealing their inherent differences (Rives and Galitski, 2003, Lin *et al.*, 2002, McCarty and Chory, 2000).

The tropical crop sugarcane (*Saccharum* sp.) is of great economical interest,

contributing to about two thirds of the world's raw sugar production. In some countries, part of the crop is also destined to the production of ethanol, a less polluting fuel alternative. Traditional breeding programs that select for varieties showing high productivity and resistance to stresses and diseases are slow. Therefore, it could be greatly advantageous to have genes associated with desirable traits targeted for directed improvement of sugarcane varieties. With the aim of expediting sugarcane genomics, the SUCEST consortium (http://sucest.lad.ic.unicamp.br/public) sequenced and annotated 237,000 expressed sequence tags (ESTs) derived from 26 cDNA libraries (Vettore *et al.*, 2003). The sequences were assembled into 43,141 contigs or sugarcane assembled sequences (SASs) covering an estimated 90% of the expressed genome. As for all other cDNA and genomic sequences released, the challenge now is to attribute relevant biological information to the extracted data. Several studies have described particular features of sugarcane's general metabolism, growth and development based on the analysis of the data from the SUCEST project (Arruda, 2001). Notwithstanding, given the enormous amount of data generated by a project of this magnitude, many topics remain to be investigated. The SUCAST Project (Sugarcane Signal Transduction) (Souza *et al.*, 2001) is an ongoing effort that aims to identify the sugarcane signaling components and define their role in grasses. In this study, we present the SUCAST Catalogue and its categories, and investigate gene expression patterns using cDNA microarrays.

Materials and Methods

Annotation

The Sugarcane cDNA sequences can be found at the SUCEST database (http://sucest.lbi.ic.unicamp.br/public/) and GenBank under Accession Numbers CA064599-CA301538. Members of the SUCAST catalogue were identified using the BLAST algorithm (Altschul *et al.*, 1997) with conserved protein sequences as drivers. Conserved protein family domains were identified by searches at the Pfam (Sonnhammer *et al.*, 1998) and SMART (Schultz *et al.*, 1998) databases using default parameters.

PCR amplification and array printing

Sugarcane cDNA plasmid clones of 1,632 ESTs obtained from the SUCEST collection were re-arranged and amplified in 100-μl PCR reactions (40 cycles, annealing at 51 °C), directly from bacterial clones in culture, using T7 and SP6 primers. Ninety percent of the clones had their identity validated by re-sequencing. PCR products were purified by filtration using 96-well filter plates (Millipore Multiscreen^{TM}MAFBN0B50). Samples were visualized on 1% agarose gels to inspect PCR amplification quality and quantity. Purified PCR products (in 10 mM Tris-HCl solution at pH 8.0) were mixed with an equal volume of DMSO in 384 well V-bottom plates. Microarrays were constructed by arraying cDNA fragments on DMSO optimized, metal-coated glass slides (type 7, Amersham Biosciences) using the Generation III Microarray Spotter (Molecular Dynamics/Amersham Pharmacia Biotech). Each cDNA fragment was spotted on the slides at least four times (i.e., technical replicates). Following printing, the slides were allowed to dry and the spotted DNA was bound to the slides by UV cross-linking (50 mJ).

Sugarcane tissue samples

Two different samples (i.e., biological samples) were collected for cDNA microarray tissue profiling from leaves (LV), flowers (FL), lateral buds (LB), roots (RT), first internode (IN1), and fourth internode (IN4) of distinct plants. Five leaf samples, each from a single field grown plant, were collected and tested for field variability (LV-1, LV-2, LV-a, LV-b, LV-c). Culms of the commercial variety SP80-3280 were planted in May 2001 and May 2002 at the Copersucar Experimental Station. The first leaf with a visible dewlap (leaf+1) was collected from a 12-month-old plant for the LV-1 sample, from a 14-month-old plant for the LV-2 sample (both planted in 2001) and from 12-month-old plants for LV-a, LV-b and LV-c (planted in 2002). Two flower samples were collected from immature inflorescences (variety SP87-342) with 5 to 30 cm (FL-1) or 50 cm (FL-2) in length. Lateral bud and root samples derived from single-eyed seed setts were collected from 12- to 14-month-old field grown plants (variety SP80-3280). For the LB-1 and RT- 1 samples, seed setts were treated with Benlate (Benomyl) 0.6 g/l and Decis (Deltamethrin) 5 ml/l, and germinated in the dark on wet paper towels for 10 days at 25 °C. For the LB-2 and RT-2 samples, seed setts were planted in 200-ml plastic cups containing moist white sand and tissues and were collected after 12 days. The internode samples were collected from field grown plants of the commer-

cial variety SP80-3280. For the IN1-1 and IN1-2 samples, the leaves were removed and the first and second internodes visible below the apical meristem were used. For the IN4-1 and IN4-2 samples, the fourth internode was collected. Also, an independent collection of leaves, flowers, lateral buds, roots, first internode and fourth internode was performed, which were used in realtime PCR assays. Tissues were sectioned, frozen in liquid nitrogen, and stored at -80°C.

RNA extraction

Frozen tissues were ground using a homogenizer. Tissue samples of 2-2.5 g were weighed and ground to a fine powder in liquid nitrogen using a pre-cooled mortar and pestle. The pulverized tissue was transferred to a 50 ml tube and homogenized with 5 ml of Trizol™(Life Technologies) per gram of tissue, according to the manufacturer´s instructions. RNA pellets were resuspended in 20 μl of warm diethyl pyrocarbonate-treated water, vortexing gently for about 15 min. RNA samples were quantified in a spectrophotometer and loaded on 1.0% agarose/formaldehyde gels for quality inspection. An equimolar pool of RNA samples of five sugarcane tissues (flower, leaf, stem, root, bud) was prepared for use as a common reference in all hybridizations.

Probe preparation and hybridization

Ten micrograms of total RNA were reversetranscribed, labeled, and hybridized using the reagents provided with the *CyScribe Post-Labeling kit* (Amersham Biosciences), according to the manufacturer's instructions. The products of the labeling reactions were purified in Millipore Multiscreen™filtering plates to remove unincorporated labeled nucleotides. Microarrays were co-hybridized with the fluorescently labeled probes. Hybridizations were performed overnight at 42 °C in humid chambers. The slides were then washed in 1xSSC and 0.2% SDS (10 min, 55 °C), twice in 0.1xSSC and 0.2% SDS (10 min, 55 °C), and in 0.1xSSC (1 min, at room temperature). Slides were rinsed briefly in filtered milli-Q water and dried under a nitrogen stream. Each experimental step was carefully monitored to ensure high quality of the slides and extracted data. The hybridizations were performed as displayed in Table 2.1.

Tabela 2.1: **cDNA microarray hybridizations performed with sugarcane tissue samples.** All hybridizations were performed against a reference sample (pool of tissues composed of an equimolar mixture of flower, leaf, stem, root and bud RNA). The table indicates which CyDye was used to label each sample in each different hybridization.

flower			**lateral bud**			**leaf**			**root**			**1st internode**			**4th internode**		
Cy3		**Cy5**	**Cy3**		**Cy5**	**Cy3**		**Cy5**	**Cy3**		**Cy5**	**Cy3**		**Cy5**	**Cy3**		**Cy5**
Pool	vs.	FL-1	Pool	vs.	LB-1	Pool	vs.	LV-1	Pool	vs.	RT-1	Pool	vs.	IN1-1	Pool	vs.	IN4-1
Pool	vs.	FL-1	Pool	vs.	LB-1	Pool	vs.	LV-1	Pool	vs.	RT-1	Pool	vs.	IN1-1	Pool	vs.	IN4-2
FL-1	vs.	Pool	LB-1	vs.	Pool	Pool	vs.	LV-2	RT-1	vs.	Pool	IN1-1	vs.	Pool	IN4-1	vs.	Pool
FL-2	vs.	Pool	LB-2	vs.	Pool	LV-1	vs.	Pool	RT-2	vs.	Pool	IN1-2	vs.	Pool			
						LV-2	vs.	Pool									
						LV-a	vs.	Pool									
						LV-b	vs.	Pool									
						LV-c	vs.	Pool									

Data acquisition, processing and statistical analysis

Slides were scanned using the *Generation III ScannerTM* (Molecular Dynamics) adjusting the photomultiplier tube (PMT) to 700 for both channels. Images were processed and data collected using the *ArrayVision* (Imaging Research Inc.) software. Local median background was subtracted from the median-based trimmed mean (MTM) density for each spot. Data from clones that generated poor quality PCR fragments (no amplification or unspecific bands) or poor quality spots (visually inspected) were excluded. The data were stored and managed by the BioArray Software environment (Saal *et al.*, 2002) free web-based database.

A set of custom programs based on R language were developed for data processing based on methods described previously (Koide *et al.*, 2004) (available at http://verjo19.iq.usp.br/xylella/microarray/). Pearson correlation values among the leaf samples were calculated using normalized expression ratios obtained from leaf versus pool hybridizations for 1,280 genes (Table S-1, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t1.html). We used homotypic or 'self-self' hybridizations of the reference pool sample to define intensity dependent cutoff levels that would indicate differentially expressed genes. Based on these results, eight intervals were set integrating the probability density function to 99.5% for different signal intensity levels, which were used to define differentially expressed genes in the inspected tissues. Figure S1 (http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_f1.html) shows the data from four 'self-self' hybridizations of the reference pool sample computed to establish the limits of the random variations in the SUCAST microarray experiments. The fluorescence ratios were normalized to account for systematic errors using the LOWESS fitting (Yang *et al.*, 2002) and used to calculate the expression ratios for all genes between the tissue sample and the reference sample (Tables S-1, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t1.html and S-2, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t2.html). For every gene, the percentage of replicates within or outside the cutoff limits was calculated in each tissue sample (Table S-3, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t3.html). Genes with at least 70% of the replicate points above or below the cutoff limits were considered differentially expressed in that particular sample, while genes with 55% of the points within the cutoff were considered ubiquitous among samples.

For the clustering analysis and the visualization of a profile matrix, a single in-

tensity value for each gene was obtained by calculating the median of all replicate points representing the same clone. Data were clustered hierarchically using the unweighted pair-group method average (UPGMA) algorithm with the euclidian distance as a measure. Further details are available at the supplementary web site (http://www.sucestfun.org/pub/SUCAST).

Validation of microarray results by real-time PCR (RT-PCR)

Five micrograms of total RNA were treated with DNAse (Promega) according to the manufacturer's instructions and an aliquot of 7.5 μl of the treated RNA was reverse-transcribed using the *SuperScript First-Strand Synthesis System for RT-PCR* (Invitrogen). The 20 μl reverse transcription reactions contained the RNA template; 2 μl 10X RT buffer; 0.5 mM each dATP, dGTP, dCTP and dTTP; 50 ng random hexamers; 0.25 μg oligo(dT); 5 mM $MgCl_2$; 10 mM DTT (dithiothreitol); 40 U Rnase OUT; and 50 U *SuperScript II Reverse Transcriptase*. RNA, random hexamers, dNTPs, and oligo(dT) were mixed first, incubated at 70 °C for 5 min and placed on ice. Subsequently, the remaining components, except the *SuperScript II Reverse Transcriptase*, were added to the reaction and the mixture was heated to 25 °C for 10 min and then incubated at 42 °C for 2 min. The *SuperScript II Reverse Transcriptase* was added to each tube and the reaction was incubated at 42 °C for 1.5 hr, 72 °C for 10 min, and chilled on ice. An identical reaction without the reverse transcriptase was performed as a control to confirm the absence of genomic DNA. The cDNA product was treated with 2 U of RNAseH (Invitrogen) for 30 min at 37 °C and for 10 min at 72 °C. Realtime PCR reactions were performed using *SYBR Green PCR Master Mix* (Applied Biosystems) in a *GeneAmp 5,700 Sequence Detection System* (Applied Biosystems). Primers were designed using the *Primer Express 2.0* Software (Applied Biosystems). BLAST searches against the SUCEST database were conducted to ensure the specificity of the selected primers. The primer sequences designed are listed in Table S-4 (http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t4.html). Each reaction was performed in duplicate and contained 2 μl of a 1:10 dilution of the synthesized cDNA, primers to a final concentration of 600 nM each, 12.5 μl of the *SYBR Green PCR Master Mix* and PCR-grade water to a total volume of 25 μl. The parameters for the PCR reaction were 50 °C for 2 min, 95 °C for 10 min, 40 cycles of 95 °C for 15 sec and 60 °C for 1 min. The specificity of the amplified products was evaluated by the analysis of the dissociation curves generated by the equipment. Negative controls were also prepared in order to confirm the

absence of any contamination. The ratio between the relative amounts of the target gene and the endogenous control gene in the RT-PCR reactions was determined based on the $2^{-\triangle\triangle C_t}$ method (Livak and Schmittgen, 2001) with modifications. The normalized expression level was calculated as $L = 2^{-\triangle C_t}$ and $\triangle C_t = C_{t,target} - C_{t,reference}$, for each tissue. To classify a gene's distribution of expression levels among the different tissues, ranging from ubiquitous to tissue specific, we used the entropy measure:

$$H_{gene} = \sum p_t * log_6(\frac{1}{p_t}) \quad (2.1)$$

where $p_t = L_t / \sum L_t$, L_t is the expression level of the gene in the t-th tissue, and the sums are taken over the six tissues.

Results and Discussion

The SUCAST catalogue

Plant responses to developmental and environmental signals rely on the activity of different cellular components, which detect these signals and transduce them through the cytoplasm and nucleus to trigger the appropriate metabolic answer. These signaling pathways coordinate growth and development, as well as responses to stress and pathogens. With the aim of creating a signal transduction catalogue for sugarcane we undertook a detailed survey of 43 thousand transcripts identified by the SUCEST project (Vettore *et al.*, 2003). This EST project sequenced the 5' and 3' end of clones from 26 libraries prepared from 11 different sugarcane tissues and plants submitted to three stress treatments. The large sampling of many tissues allowed possibly 90% of the sugarcane expressed genes to be tagged.

We used BLAST (Altschul *et al.*, 1997) searches, Pfam (Sonnhammer *et al.*, 1998) and SMART (Schultz *et al.*, 1998) domain analyses to identify conserved signal transduction components such as receptors, adapters, G-proteins, small GTPases, members of the two-component relay system, nucleotide cyclases, protein kinases, protein phosphatases and elements of the ubiquitination machinery and infer their putative functions. Around 2000 SASs encoding signal transduction related proteins and also 611 transcription factors were indexed in the SUCAST catalogue, which is organized into 29 categories and 409 subcategories (Table S-5, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_

t5.html). These elements represent 5% of the total SASs from the current SUCEST dataset. In addition, 717 SASs that might be involved in processes triggered by stress and pathogens or that may play a role in growth and development were also catalogued. Table 2.2 summarizes the SUCAST categories. The combined analysis of the sugarcane EST data bank, by means of an in depth annotation and gene architecture analysis, generated a catalogue with 3,563 members, which covers several aspects of signaling and transcription. It includes around 100 SASs for hormone biosynthesis (Souza *et al.*, 2001) and also 548 SASs with no similarities to known proteins, which were selected due to our interest in associating function to new genes.

On the basis of sequence analysis, it has been inferred that 13% of the Arabidopsis genes are involved in transcription or signal transduction (Initiative, 2000). The automated categorization of the SUCEST data indicated that 13.6% of the tagged genes belong to these categories (Vettore *et al.*, 2003). With a genome size expected to be similar to the rice genome, the sugarcane genome might have around 3 thousand genes encoding putative signal transduction components.

The SUCAST cDNA microarray

To evaluate the expression profile of the SUCAST components in different sugarcane tissues we constructed glass slide cDNA microarrays with PCR products derived from 1632 cDNA clones. For 21% of the clones we could not obtain satisfactory PCR fragments and the corresponding data were removed from the analysis. Therefore, data of transcript abundance for 1,280 SASs are presented as indicated in the categories of Table 2.2. As a reference sample in all microarray hybridizations we used an equimolar pool of total RNA extracted from flowers, lateral buds, leaves, stems and roots.

Assessing individual variability in the field

Since sugarcane is propagated vegetatively, the genetic variability among the individuals should be low. Expression patterns obtained assaying few individuals of the same variety should be representative of a population in the field, provided that growth conditions are similar. To minimize individual differences and differences attributable to local field variations, RNA samples were typically obtained from more than one plant in our experiments. Even so, we reasoned that it would be important to evaluate whether the in-

Tabela 2.2: **The SUCAST Catalogue.** The number of SASs in the catalogue and the number selected for the cDNA microarray analysis for each category are shown. For a list of all SASs refer to the Supplementary Material, Table S-5 (http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t5.html).

SUCAST classification	**# SAS**	**# SAS in the array**
Protein categories:		
Receptors	477	181
Adapters	12	9
G proteins	17	10
Small GTPases	75	36
Two component relay	19	8
Cyclase	1	0
Calcium metabolism	68	31
Inositol metabolism	46	21
Protein Phosphatases	107	36
Protein kinases	510	65
Ubiquitination	106	41
Transcription factors	611	175
Hormone biosynthesis	75	30
Hormone related	22	13
Functional categories:		
Development	30	13
Cell cycle	34	10
Stress	305	119
Pathogenicity	382	104
'No matches' and unknown proteins	548	294
Others	118	84
TOTAL	3563	1280

dividual variability was as low as expected. With that purpose, we collected leaves from five different sugarcane individual stools and extracted the RNA separately. Leaves were collected in May 2002 (LV-1), July 2002 (LV-2) and May 2003 (LV-a, LV-b, LV-c). The RNA samples were labeled and hybridized to the microarrays against the reference sample (Table S-1, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t1.html). Pairwise Pearson correlation calculations show a high correlation between leaves of the three individuals collected at the same time or within a short interval of time (p=0.84 to 0.88), and a lower correlation between individuals collected in different years (p=0.61 to 0.64) (Fig. S2, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_f2.html). The results imply sufficiently low individual variation within each sampling event, and even between close events, and indicate that pooling a large number of plants to represent a subpopulation is not necessary.

Differentially and evenly distributed genes

Total RNA samples extracted from six different sugarcane tissues were labeled and hybridized to the microarrays against the reference sample. Two different biological samples of each tissue were analyzed, and the results of at least two technical replicates were computed. Median ratio values for each gene in each sample can be found in Table S-2 (http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t2.html). Cutoff limits for differential expression were calculated based on 'self-self' hybridizations (see Methods). To estimate replicate data consistency, the expression ratio versus signal intensity data of the replicates of a given gene in different 'tissue vs. pool' hybridizations (Table 2.1) were studied (Fig. 1). Genes with at least 70% of the replicate points outside the cutoff limits (above or below) in both biological samples of one or more tissues were considered differentially expressed, whereas genes with more than 55% of the replicates within the cutoff limits were considered ubiquitous. Figure 2.1 shows four cases of gene expression distribution: (a) over 70% of the data points are above the cutoff limits, indicating that this gene was more expressed in the tissue being tested than in the reference; (b) all the replicates are within the cutoff limits, indicating that there was no differential expression; (c) a variable pattern among the replicates was observed, showing low reproducibility for the expression levels of the gene; (d) over 70% of the data points are below the cutoff limits, indicating that this gene is less expressed in the tissue being tested than in the reference. A graphical representation of the

global distribution of the data in the M-S space, taking into account the reproducibility of the technical replicates for each gene in each hybridization is seen in Fig. 1e and summarized in Table 2.3. The number of genes analyzed for each tissue varied from 1045 to 1235, due to differences in the quality of some spots in different slides. The majority of the genes analyzed (avg. 76%) showed expression levels in each particular tissue similar to those of the pool of tissues. This result is in agreement with the observations of Obayashi and colleagues (Obayashi *et al.*, 2004), who have identified a large cluster of ubiquitously expressed genes after global macroarray analysis of the Arabidopsis transcriptome. It is important to note, however, that only part of the sugarcane transcriptome is represented in our array. Nevertheless, this could be an indication that the majority of the signaling elements in sugarcane are not differentially expressed in the different tissues analyzed. The highest percentages of preferential expression in one tissue were found in leaves (8.84% and 7.36%), in the RT-2 root sample (11.78%), and in the internode sample IN4-2 (8.26%). Likewise, a high proportion of the genes in these samples showed reduced expression in one particular tissue. The flower samples exhibited a high number of underrepresented or non-expressed genes (8.33% and 6.87%, respectively, for each of the two different biological samples). An average of 11% of the genes showed a variable pattern of expression, with high variation among the technical replicates.

For the majority of the genes present in our microarray there is little information in the literature regarding tissue distribution of transcripts. Cho and colleagues (Cho *et al.*, 2002) performed microarray hybridizations using samples from seven different organs of maize. This approach allowed the elucidation of organ relationships and the detection of organ-specific gene expression. Recently, a comprehensive study of organ-specific gene expression has been reported for *Arabidopsis* (Obayashi *et al.*, 2004). Other reports focus on a few genes or specific metabolic routes, involving families of closely related genes such as the MADS transcription factors (Alvarez-Buylla *et al.*, 2000), genes involved in particular pathways such as the acyl lipid metabolism (Beisson *et al.*, 2003), and often rely solely on EST *in silico* data. Watson and collaborators (Watson *et al.*, 2003) described the mapping of the proteome of the model legume barrel medic (*Medicago truncatula*). Spatial mapping of the transcriptome and proteome of diverse plant species can shed light on the regulation of many developmental pathways.

Our results indicated 217 genes that presented differential expression in both biological samples in at least one of the tissues analyzed. These elements were clustered ac-

Tabela 2.3: **Distribution of differentially expressed and ubiquitous genes among sugarcane tissue samples.** Genes with at least 70% of replicate points outside the cutoff limits in both biological samples of one or more tissues were considered differentially expressed, whereas genes with more than 55% of replicates within the cutoff limits in all samples were considered ubiquitous. The percentage of genes with expression above, within or below the ratio cutoff limits are indicated for each sample.

tissue	sample	% of genes above	% of genes within	% of genes below	% of genes variable	Total # of SAS analyzed
flowers	FL-1	3.19%	78.59%	8.33%	9.89%	1224
	FL-2	5.00%	80.15%	6.87%	7.97%	1179
lateral buds	LB-1	1.18%	90.10%	3.47%	5.25%	1182
	LB-2	4.61%	77.30%	5.36%	12.73%	1194
leaves	LV-1	8.84%	72.50%	9.98%	8.67%	1222
	LV-2	7.36%	76.05%	9.13%	7.46%	1073
roots	RT-1	5.52%	80.46%	7.31%	6.71%	1177
	RT-2	11.78%	61.33%	10.46%	16.43%	1205
internodes	IN1-1	3.91%	86.34%	4.58%	5.16%	1201
	IN1-2	6.20%	69.31%	5.80%	18.69%	1225
	IN4-1	4.40%	88.33%	1.91%	5.36%	1045
	IN4-2	8.26%	54.57%	10.77%	26.40%	1235
Average		5.87%	76.02%	7.05%	11.05%	1180

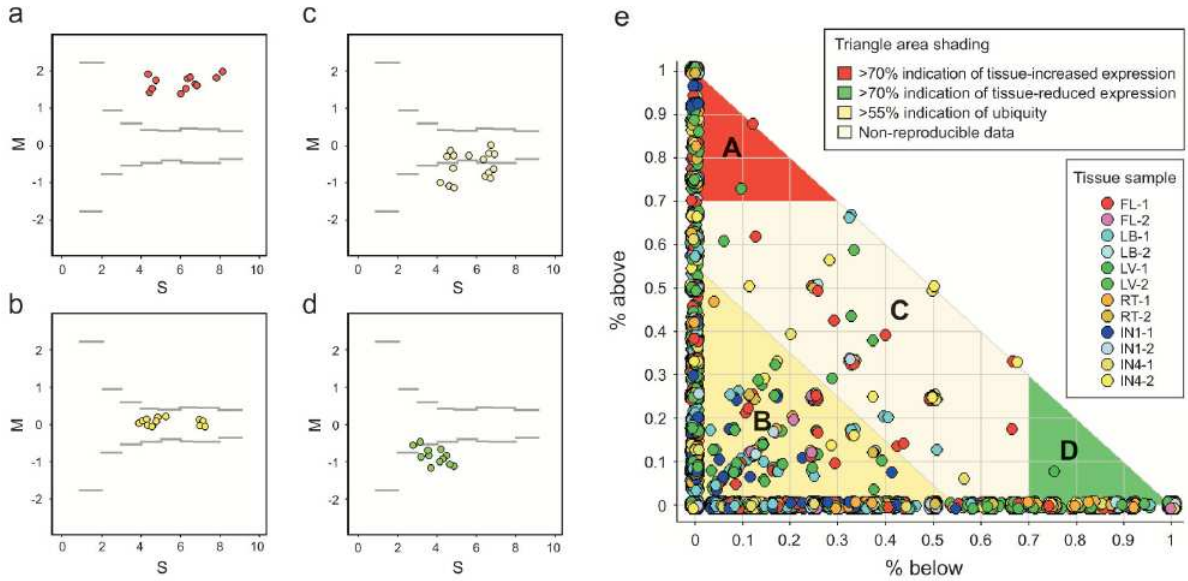


Figura 2.1: **Differential expression consistency.** a through d - Normalized log ratios against log average intensity plots, where $M=\log_2(Cy5 / Cy3)$, and $S=\log_2(0.5 * (Cy5 + Cy3))$. The graphics correspond to examples of the data distribution of the technical replicates for a gene we considered to have tissue-increased expression (a), ubiquity or no differential expression (b), low reproducibility (c), or tissue-reduced expression (d). The bars indicate the intensity-dependent cutoffs. E - Global distribution of SUCAST microarray data. The position of each dot in the triangle relates to the percentage of reproducible replicates in each hybridization. The areas A, B, C and D contain data as exemplified in the corresponding graphics a, b, c and d.

cording to their expression patterns, evidencing groups of genes with marked expression in leaves, roots, or internodes (Fig. S3, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_f3.html). Smaller groups of genes with prominent expression in flowers and lateral buds were also uncovered.

Several differentially expressed genes encode transcription factors. Among these, we detected ten genes highly expressed in flowers: a GARP transfactor, an AP2, four zinc-finger and four MADS-box domain-containing proteins. MADS and the zinc finger YABBY transcription factor play important roles during organ development and together with AP2 and zinc finger C3H proteins were shown to have enhanced expression levels in flowers (Hennig *et al.*, 2004).

A discrete group of eight receptor genes was found to be preferentially expressed in leaves, most of which are members of the receptor-like kinase (RLK) family. We also found a sugarcane receptor possibly involved in signaling pathways in the sugarcane reproductive tissues. This SASs is very similar to MSP1 from rice and EXS/EMS1 from *Arabidopsis*. MSP1 and EXS/EMS1 are genes expressed in reproductive tissues and, among other functions, control the fate of germinative cells (Zhao *et al.*, 2002, Canales *et al.*, 2002, Nonomura *et al.*, 2003) These observations indicate that this SAS possibly represents an ortholog of the MSP1 and EXS/EMS1 genes. Three SASs were found that code for putative receptors containing a protein kinase domain and a Ubox as predicted by the SMART database (Schultz *et al.*, 1998). All of them showed a homogeneous transcript distribution among the tissues analyzed. One of them, SAS SCQSRT2031C08.g, possesses a complex structure, comprised by TPRs (tetratricopeptide repeats), low complexity regions, a pkinase domain and a Ubox domain followed by a ZnF NFX domain (a presumed zinc binding domain) near the C-terminus. The U-box is believed to have a role in ubiquitination (Aravind and Koonin, 2000). Protein kinases containing the U-box domain have already been reported for *Arabidopsis* (Azevedo *et al.*, 2001). However, the function of these plant proteins remains to be determined. Moreover, 13 receptors of unknown function were found to have a differential expression pattern. Six of these presented predominant expression and seven showed weak expression in at least one of the tissues. The elucidation of the expression profiles of new receptors is of great interest, since it can help in assigning putative functions to these proteins.

It is remarkable that several genes related to the ubiquitination system have been found to be more expressed in the internodes than in the other tissues examined. The ubiquitin/26S proteasome pathway (Vierstra, 2003) is implicated in selected protein breakdown,

used to control the level and activity of proteins in a diverse range of metabolic routes. In sugarcane, an intense protein degradation activity in the internodes could be related to their specialization in sucrose storage.

A group of hormone-related elements, including four nitrilases and three lipoxygenases, showed prevalent expression in root tissues. The nitrilases are homologous to the *Arabidopsis nit4* gene, which was characterized as being predominantly expressed in roots (Bartel and Fink, 1994). The lipoxygenases (LOX) are a functionally diverse class of dioxygenases implicated in physiological processes such as growth, senescence, and stress responses in plants, that show different organ-specific expression in different plants (Kolomiets *et al.*, 2001,Porta and Rocha-Sosa, 2002) Another group of hormone-related genes composed by members of biosynthetic pathways of salicylic acid (phenylalanine ammonia-lyase), abscisic acid (zeaxanthin epoxidase), and ethylene (ACC oxidase, ACC synthase) biosynthetic pathways was mainly expressed in leaves

We detected a caffeic acid 3-*O*-methyltransferase (COMT) gene expressed primarily in the fourth internode. This enzyme is involved in lignin biosynthesis and, in association with other enzymes like the CCOMT (caffeoyl CoA 3-*O*-methyltransferase), keeps in check the content and the composition of lignin in cells. A correlation between the lignin content of alfalfa internodes of progressive maturity and the activity of COMT and CCOMT has been demonstrated (Inoue *et al.*, 1998). A sugarcane COMT has been cloned and exhibited a peak of expression in culms (Selman-Housein *et al.*, 1999) The SUCEST database indicates the presence of five complete sequences for this enzyme, that may represent a promising target for sugarcane genetic engineering with the aim of modifying the content and/or composition of sugarcane bagasse, allowing it to become a useful and low cost raw material for paper production and animal feed.

Among the 43,141 SASs in the SUCEST database, 35% did not show similarity to known proteins (no matches) and are therefore new genes of unknown function. We found nine no-matches to be predominantly expressed in internodes, seven in leaves, two in roots and two in flowers. The latter are non-coding transcripts. The involvement of non-coding RNAs in floral development has been described (Schmid *et al.*, 2003, Aukerman and Sakai, 2003, Chen, 2004) and thus it is possible that these sugarcane elements are involved in gene regulation during floral development. Among all the SUCEST no-matches, 2010 SASs correspond to sugarcane-specific non-coding sequences that could also contain regulatory elements.

In addition to the differentially expressed genes, we also investigated genes that showed similar expression levels in all tissues. The identification of "housekeeping"genes is of great interest in expression studies, since they are valuable experimental controls and indicate promoters useful for plant biotechnology. Among the analyzed genes, 153 presented over 55% of the replicate data points within the cutoff limits for all 12 samples analyzed (Table S-6, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t6.html). A total of 35 no-matches were found among them, an indication that these sugarcane-specific genes may have a central role in this plant's physiology.

Validation of microarray data by real-time PCR

To validate the present work, 25 SASs (Table S-4, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t4.html) were selected and analyzed by realtime PCR. To normalize the expression data, several ESTs were tested in search of a gene that showed strong and ubiquitous expression in the sugarcane tissues. An ideal reference gene has the same level of expression in all conditions under study. The commonly used tubulin gene did not show an adequate pattern, being expressed in varying levels in the tissues analyzed (Fig. 2a). The same was observed for an actin gene (not shown). Based on the number of ESTs sequenced in the SUCEST project and on the expression profile obtained from the microarray data we selected two SASs as references: SCCCLR1048F12. g (a 14-3-3 gene) and SCCCST2001G02.g (a polyubiquitin gene). These genes were found to be homogeneously expressed and adequate for normalization purposes, showing equivalent transcript levels in the tissue samples, except for a slight variation in flower tissues (for the polyubiquitin gene) and roots (for the 14-3-3 gene) (Fig. 2b,c). Therefore, all our real-time PCR data was normalized to both the 14-3-3 gene and the polyubiquitin gene. When the normalization was done using the 14-3-3 gene, expression data for the root tissue was disregarded. The same was done for the flower sample when normalizing data with the polyubiquitin gene. Although none of the reference genes tested presents absolutely the same expression levels in all tissues, the use of two different reference genes increases the reliability of the results. We considered that a gene had its expression profile validated when both results (using the 14-3-3 gene and the polyubiquitin gene as references) were consistent with the microarray data. Eighteen differentially expressed and seven ubiquitous SUCAST genes were assayed. To further confirm the expression profile obtained, the RNA samples used for the validation experiments were

different from those used in the microarray hybridizations. Figure 2.2d-k shows the relative levels of transcripts for 8 SASs normalized to the polyubiquitin levels. The results using the 14-3-3 gene as a reference yielded essentially the same patterns (not shown).

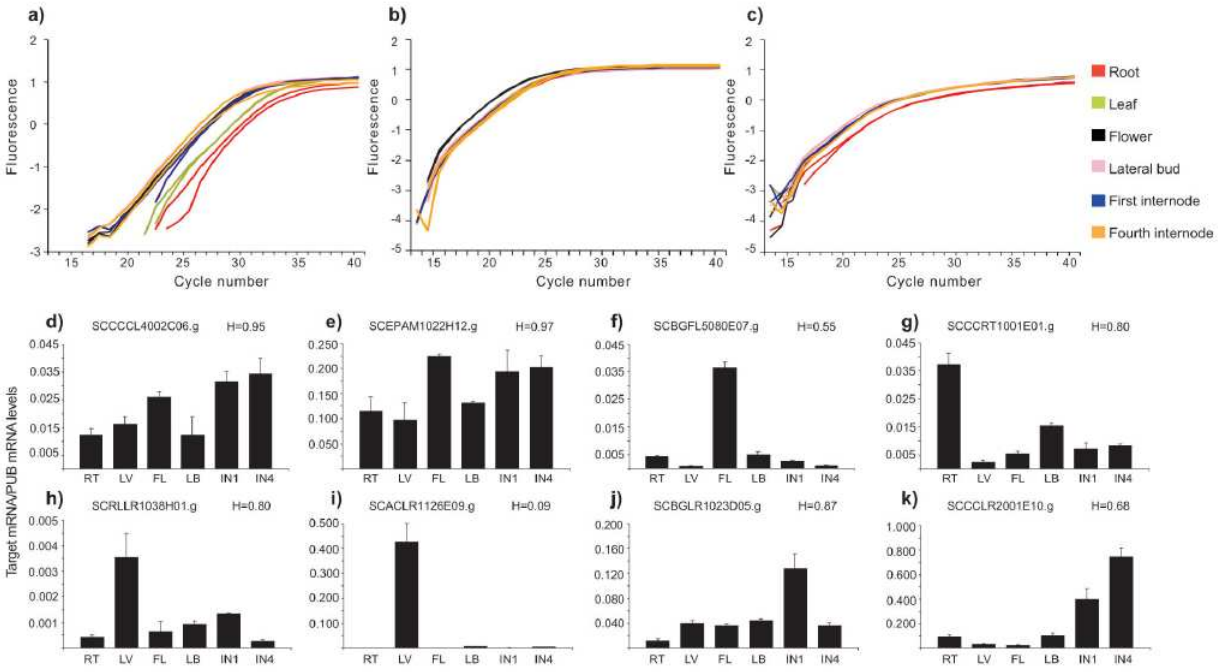


Figura 2.2: **Validation of microarray results using real-time PCR.** a through c - Raw values of the log10 fluorescence were plotted against the cycle number for a tubulin gene (a), a polyubiquitin gene (b), and a 14-3-3 gene (c). Each tissue analyzed is represented by a different color. All the reactions were carried out in parallel and each reaction was performed in duplicate. d through k - Real-time PCR results for a Phosphatase - PP6 / catalytic subunit (d), a Cytochrome P450 - CYP71A (e), an EXS receptor kinase (f), a lipoxygenase (g), a zeaxanthin epoxidase (h), SCACLR1126E09.g (no match) (i), LSD1 gene (j) and SCCCLR2001E10.g (no match) (k). The bars show target mRNA levels relative to the polyubiquitin mRNA. RT = root, LV = leaf, FL = flower, LB = lateral bud, IN1 = first internode, IN4 = fourth internode. The measured entropy (H) for each distribution obtained is indicated. Error bars were calculated as described by Livak and Schmittgen (Livak and Schmittgen, 2001)

To rank the differential expression results obtained in real-time PCR analysis, we measured the entropy (H) of the distribution of expression levels among the tissues. The entropy is widely used in information theory to measure how distant the observed distribution is from a uniform distribution. Ideally, the ubiquitous genes should have a uniform distribution of expression levels in all considered tissues. According to the entropy defini-

tion, this property is mathematically translated to H closer to 1. In contrast, tissue-specific genes should have relevant expression in just one of the considered tissues and the entropy of this expression level distribution is translated to H closer to zero (note, however, that this is not a linear scale). We observed that 6 out of 7 genes expected to present a ubiquitous expression profile in fact presented an H value equal to or higher than 0.9 (Table S-4, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t4.html), indicating that they can represent real "housekeeping"genes. Four differentially expressed genes seem to be highly specific to one tissue, with H values below 0.6. Eight differentially expressed genes showed enrichment in a particular tissue, as pointed out by the microarray data, but were also expressed at high levels in other tissues. In these cases, the H values were higher than the ones obtained for genes expressed in a single tissue, as expected, but were always lower than 0.9.

In summary, 18 out of the 25 genes tested (72%) had a profile in real-time PCR assays consistent with the differential or ubiquitous expression observed in the microarray experiments. It is important to stress that the RNA samples used in the real-time PCR experiments derived from a third biological sample, further suggesting that the data set generated in our microarray experiments is robust in indentifying ubiquitous and differentially expressed genes. The criteria used to select the differentially expressed and ubiquitous genes, although arbitrary, proved to be effective. The selection of data with at least 70% of the replicates in agreement with the cutoff for differential expression and 55% for ubiquitous expression was adequate, as shown by the high validation rate obtained. The less stringent value for ubiquity proved to be as effective probably because the genes selected had similar expression in all biological samples of all tissues.

The SUCAST expression matrix

As pointed out previously, all hybridizations were made against a common reference, consisting of a pool of tissues. When there are several samples to be compared, this strategy requires fewer hybridizations than a direct pair-wise setup, and is useful when there is no natural control (like a non-treated sample) as in treatment versus control studies. Additionally, the pool of transcripts theoretically represents the transcripts of all tissues, minimizing the occurrence of spots without a hybridization signal, for which it is not possible to calculate the expression ratio.

Although this approach allowed us to identify ubiquitous and differentially expressed genes among the sugarcane tissues, it generated relative - not absolute - information on the expression profiles. This means it evidences, for example, that a certain gene is more highly expressed in leaves than in the average of the tissues, but it does not tell us whether this same gene is more highly expressed in flowers than in roots. To get access to this type of information, we calculated "virtual ratios"between pairs of tissues using the reference values of the common sample (pool) as the common denominator.

This approach provided us with the expression patterns of each individual SAS in the SUCAST microarray among all tissues analyzed (Table S-7, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t7.html and Fig. S4, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_f4.html). Additionally, the clustering of the patterns allows a spatial comparative picture of transcript abundance, which can complement the information provided by the ratio cutoff analysis (that uses 70% replication stringency levels). Using the matrix, we could note expression patterns not evident in the previous analysis. As an example, all four of the MADS transcription factors that were identified as differentially expressed due to lower expression in roots or other tissues than in the reference, are indicated by the matrix to be primarily expressed in flowers (Table S-7, http://www.dna-res.kazusa.or.jp/12/1/03/supplement/supplement_t7.html), although only SAS SCSGSB1007G01.g had been classified as such.

Conclusions

The success of the sugarcane culture has relied for decades on traditional breeding of varieties resistant to plagues and diseases, with increased sucrose content, and more adaptable to different soils and environmental conditions, a slow and uncertain approach. Therefore, genomic data that could assist traditional breeding in the improvement of sugarcane varieties are awaited. There are very few molecular studies on sugarcane signaling response to environmental changes, and none on the distribution of these components in the different plant tissues. The comparison of the transcript complement found in six tissues using microarrays provided a spatial picture of the transcriptome of this grass, which can greatly contribute to the assignment of function to new genes. The present work focused on the identification of genes that may participate in tissue-specific activities and ubiquitous genes. The cloning of strong ubiquitous promoters or tissue-specific promoters can incre-

ase the availability of tools for sugarcane transformation and study. The identification of genes highly expressed in stems or leaves could also help in the understanding of metabolic pathways involved in sugar production and accumulation, and could constitute targets for crop improvement. The described signaling elements are currently being studied in search for candidates that might regulate hormone responses, the accumulation of sucrose in the stalk, and the response to several biotic and abiotic stresses allowing us to step forward in the efficient manipulation of sugarcane varieties. The knowledge accumulated on the role for signal transduction processes in the regulation of stress and pathogenesis responses brings the SUCAST components to center stage in the search for genes that might be modified to obtain plants with desired traits.

Acknowledgements

This work was funded by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo). We are indebted to Adriana Y. Matsukuma and Denise Yamamoto for technical assistance on microarray printing and scanning performed in the laboratory of the Cooperation for Analysis of Gene Expression (CAGE) inter-departmental Project, Dr. Sergio Verjovski-Almeida and Dr. Hugo Aguirre Armelin for their support throughout the development of this work, Apuã C. M. Paquola for valuable help providing bioinformatic tools and Dr. Jesus Ferro for coordinating the SUCEST cDNA clone collection in the beginning of this work. AMdS was partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico, (CNPq). FSPT, FRR, RZNV, KCO and ACQS were supported by FAPESP fellowships.

Então, só posso desejar uma coisa para vocês - a boa sorte de estarem em um lugar onde sejam livres para manter o tipo de integridade que eu descrevi e onde não se sintam forçados a perder sua integridade por uma necessidade de manter sua posição na organização, ou o apoio financeiro, ou coisa assim. Que vocês possam ter essa liberdade.

Richard Feynman (1918-1988)

3

Signal transduction-related responses to phytohormones and environmental challenges in sugarcane†

Flávia R Rocha[1], Flávia S Papini-Terzi[1], Milton Y Nishiyama Jr[1], Ricardo ZN Vêncio[2], Renato Vicentini[3], Rodrigo DC Duarte[3], Vicente E de Rosa Jr[3], Fabiano Vinagre[4], Carla Barsalobres[5], Ane H Medeiros[5], Fabiana A Rodrigues[7], Eugênio C Ulian[6], Sônia M Zingaretti[7], João A Galbiatti[7], Raul S Almeida[8], Antonio VO Figueira[8], Adriana S Hemerly[4], Marcio C Silva-Filho[5], Marcelo Menossi[3] and Gláucia M Souza[1]

Abstract

Background: Sugarcane is an increasingly economically and environmentally important C4 grass, used for the production of sugar and bioethanol, a low-carbon emission fuel. Sugar-

†*BMC Genomics* 2007, **8**:71 (doi:10.1186/1471-2164-8-71)

[1]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, SP, Brazil
[2]BIOINFO-USP Núcleo de Pesquisas em Bioinformática, Universidade de São Paulo, São Paulo, SP, Brazil
[3]Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Campinas, SP, Brazil
[4]Instituto de Bioquímica Médica, Universidade Federal do Rio de Janeiro, UFRJ, Rio de Janeiro, RJ, Brazil
[5]Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz, ESALQ, Universidade de São Paulo, Piracicaba, SP, Brazil
[6]Centro de Tecnologia Canavieira, Piracicaba, São Paulo, SP, Brazil
[7]Departamento de Tecnologia, Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal, Universidade Estadual Paulista, Jaboticabal, SP, Brazil
[8]Centro de Energia Nuclear na Agricultura (CENA), Universidade de São Paulo, Piracicaba, SP, Brazil

cane originated from crosses of *Saccharum* species and is noted for its unique capacity to accumulate high amounts of sucrose in its stems. Environmental stresses limit enormously sugarcane productivity worldwide. To investigate transcriptome changes in response to environmental inputs that alter yield we used cDNA microarrays to profile expression of 1,545 genes in plants submitted to drought, phosphate starvation, herbivory and N_2-fixing endophytic bacteria. We also investigated the response to phytohormones (abscisic acid and methyl jasmonate). The arrayed elements correspond mostly to genes involved in signal transduction, hormone biosynthesis, transcription factors, novel genes and genes corresponding to unknown proteins.

Results: Adopting an outliers searching method 179 genes with strikingly different expression levels were identified as differentially expressed in at least one of the treatments analysed. Self Organizing Maps were used to cluster the expression profiles of 695 genes that showed a highly correlated expression pattern among replicates. The expression data for 22 genes was evaluated for 36 experimental data points by quantitative RT-PCR indicating a validation rate of 80.5% using three biological experimental replicates. The SUCAST Database was created that provides public access to the data described in this work, linked to tissue expression profiling and the SUCAST gene category and sequence analysis. The SUCAST database also includes a categorization of the sugarcane kinome based on a phylogenetic grouping that included 182 undefined kinases.

Conclusion: An extensive study on the sugarcane transcriptome was performed. Sugarcane genes responsive to phytohormones and to challenges sugarcane commonly deals with in the field were identified. Additionally, the protein kinases were annotated based on a phylogenetic approach. The experimental design and statistical analysis applied proved robust to unravel genes associated with a diverse array of conditions attributing novel functions to previously unknown or undefined genes. The data consolidated in the SUCAST database resource can guide further studies and be useful for the development of improved sugarcane varieties.

Background

Sugarcane is an increasingly economically attractive crop, used for the production of approximately 60% of the world's sugar and also of ethanol, a low-carbon emission fuel. Sugarcane varieties with improved tolerance to adverse environmental conditions are

highly desirable. Unfavorable environmental factors are the major culprits of losses in agriculture and can reduce average productivity by 65% to 87% depending on the crop (Bray *et al.*, 2000). Crops better fit to withstand biotic and abiotic stresses have been selected by traditional genetic breeding programs but the slow pace in obtaining plants with the desirable traits limits the development of improved crop varieties. In this scenario, the use of molecular tools that enable gene-targeted modifications to achieve a phenotype of interest is highly promising.

Plants react to changes in the environment through an array of cellular responses that are activated by stress stimuli, leading to plant defense and/or adjustment to adverse conditions. Physiological changes elicited by external signals can be modulated by transcriptional regulation leading to the induction or repression of target genes. Many high throughput studies have been conducted to define gene expression changes in plants submitted to stress (Schenk *et al.*, 2000, Kawasaki *et al.*, 2001, Oono *et al.*, 2003, Oono *et al.*, 2006). Such studies showed that signal transduction gene expression is altered in response to stress possibly leading to changes in growth and development and adjustment to environmental conditions. Few studies have been conducted to unravel sugarcane's responses to biotic and abiotic stresses or the role of phytohormones in these processes. Examples of these are those that evaluated changes in the sugarcane transcriptome induced by cold and methyl jasmonate treatment (Nogueira *et al.*, 2003, Rosa *et al.*, 2005). The aim of this work was to profile sugarcane gene expression under conditions that affect crop yield: drought, phosphate starvation, herbivory and endophytic bacteria interaction. Drought is a condition of special interest, not only for sugarcane, but also for other crops, since increasing water scarcity has been observed throughout the world. Plant irrigation currently accounts for approximately 65% of global freshwater use indicating that the development of plant varieties resistant to drought will be a necessity in the near future (Kates and Parris, 2003, Riera *et al.*, 2005). Plant responses to drought are complex, partially dependent on ABA signaling and dependent on the intensity and duration of the stimulus. The main responses include changes in ion fluxes, stomatal closing, production of osmoprotectants and alteration in plant growth patterns (Riera *et al.*, 2005).

A significant portion of the arable land in tropical areas presents either limiting concentrations of essential nutrients or toxicity. Phosphorous (P), an essential macronutrient, is one of the most limiting nutrients for plant growth because of its low solubility and high sorption capacity in soil (Kochian, 2000). Plant roots acquire P as inorganic phosphate (Pi),

although the concentration of Pi in the soil solution is often low (2 to 10 mM) (Raghothama, 1999). The low availability of Pi in the acid soils of tropical and subtropical regions is a major limiting factor for crop production (Raghothama, 2000). P constitutes around 0.2% of plants dry weight (Schachtman *et al.*, 1998) and plays important roles in several biological processes, such as nucleic acid and phospholipid biosynthesis, energy metabolism, signal transduction and enzyme activity regulation.

Insect pests frequently challenge sugarcane productivity. The sugarcane borer *Diatraea saccharalis* is the major sugarcane pest in Brazil causing plant death due to apical bud death (dead heart) in plants of up to four months of age and damage to lateral bud development, aerial rooting, weight loss and stalk breakage in older plants. The attack also allows for infection by opportunistic fungi, which results in production loss for both the sugar and alcohol industries ((Braga *et al.*, 2003) and references herein).

The sugarcane culture is highly benefited by the association with N_2-fixing endophytic bacteria (*Herbaspirillum seropedicae/Herbaspirillum rubrisubalbicans* and *Gluconacetobacter diazotrophicus*). Unlike rhizobium/leguminosae symbiosis, where bacteria are restricted to nodules, *Herbaspirillum* spp. and *G. diazotrophicus* are endophytic, and colonize intercellular spaces and vascular tissues of most plant organs without causing damage to the host (James and Olivares, 1998, Reinhold-Hurek and Hurek, 1998). These bacteria promote plant growth possibly by nitrogen fixation and also by the production of plant hormones (Sevilla *et al.*, 2001). Despite the non-pathogenic aspects of this interaction, plants should limit bacterial growth inside their tissues, or the association can result in disease (Olivares *et al.*, 1997). Little is known about the signaling mechanisms that are involved in the establishment of a beneficial association with the plant.

A study on the response of sugarcane plants to methyl jasmonate (MeJA) and abscisic acid (ABA) treatments is needed since the role of these phytohormones in biotic and abiotic stress responses is well characterized and could point us to the regulatory mechanisms behind the stress treatments of interest. Several evidences point to a complex signaling network triggered by the action of ABA, including cross-talk with other hormone response pathways (Himmelbach *et al.*, 2003). Moreover, several genes that are induced by ABA also have their expression induced by drought and cold stress (Seki *et al.*, 2002). Protein kinases (Petersen *et al.*, 2000) and transcription factors (van der Fits and Memelink, 2000) have been shown to mediate the signal transduction network of MeJA action. All MeJA actions seem to need a functional COI protein, involved in ubiquitin-mediated proteolysis (Devoto *et al.*,

2002). cDNA arrays have been used to evaluate changes in gene expression in sugarcane leaves treated with MeJA (Rosa *et al.*, 2005). Two transcriptional factors encoding a putative zinc finger protein, a heat shock factor protein, protein kinases, proteins with a role in secondary metabolism, protein synthesis, stress response and photosynthesis were found to be differentially expressed.

The genes studied in this work were identified by the SUCEST (Sugarcane EST) Project. The SUCEST Project sequenced over 238,000 ESTs, which were grouped into over 43,000 SAS (Sugarcane Assembled Sequences) (Vettore *et al.*, 2003). The SUCAST Project (Sugarcane Signal Transduction) (Souza *et al.*, 2001, Papini-Terzi *et al.*, 2005) used BLAST searches, Pfam and SMART domain analysis to identify conserved signal transduction components such as receptors, adapters, G-proteins, small GTPases, members of the two-component relay system, nucleotide cyclases, protein kinases, protein phosphatases, elements of the ubiquitination machinery and transcription factors. In addition, SAS that might be involved in processes triggered by stress and pathogens or play a role in growth and development were also catalogued. The combined analysis of the sugarcane EST data bank, by means of an in depth annotation and gene architecture analysis, generated the SUCAST catalogue with over 3,500 members including around 100 SAS for hormone biosynthesis and around 600 SAS with no similarities to known proteins, which were selected due to our interest in associating function to new genes. These elements represent 5% of the total SAS from the current SUCEST dataset.

To define the expression pattern of these genes in the various sugarcane tissues cDNA microarrays with 1,280 distinct elements were constructed. A total of 217 genes were found to be differentially expressed when leaf, inflorescence, root, internode and lateral bud tissues were compared (Papini-Terzi *et al.*, 2005). For this work, a new array was designed with 1,228 elements in common with the array used in the previous study (Papini-Terzi *et al.*, 2005) plus an additional 317 elements including 229 representatives of the sugarcane kinome. Overall, 50% of the SAS catalogued in each SUCAST category are represented in the array that contains a total of 1,545 genes.

In the context of plant signal transduction, the role of protein kinases is remarkable. These proteins are responsible for the post-translational control of target proteins, acting as critical regulators of many signaling cascades. Moreover, many plant protein kinases act as receptors (named RLKs, from Receptor-Like Kinases) and participate in processes like disease resistance, growth, development, hormone perception and stress responses (Shiu and

Bleecker, 2001*b*). Many protein kinases remain uncharacterized, especially those corresponding to RLKs. Of 1,031 protein kinases previously catalogued by the SUCAST Project 39% could not be assigned to known categories based on BLAST searches and were annotated as undefined kinases. This work also reports the categorization of sugarcane protein kinases based on neighbor-joining (NJ) trees constructed from the alignment of the predicted catalytic domain. The association of an expression pattern to the categories generated by the phylogenetic analysis is useful in guiding studies on sugarcane kinases and other genes responsive to environmental and hormonal stimuli.

Results

Gene Expression Changes in Response to Biotic and Abiotic Stimuli

To identify genes regulated at the expression level by biotic and abiotic factors sugarcane plants were exposed to a variety of conditions that affect yield negatively (drought, phosphate deficiency, herbivory) or positively (endophytic bacteria interaction). Since a role for ABA and jasmonates has been observed in the regulation of plant stress responses in other plant systems (Himmelbach *et al.*, 2003, Berger, 2002, Howe and Schilmiller, 2002, Devoto and Turner, 2003, Verslues and Zhu, 2005), plants were also exposed to these phytohormones.

To obtain gene expression patterns and identify differentially expressed genes cDNA microarrays representing 1,545 genes were co-hybridized to fluorescently labeled probes generated from control and treated plants. The great majority of the genes were selected from the SUCAST Catalogue (Souza *et al.*, 2001, Papini-Terzi *et al.*, 2005). Some correspond to sugarcane metabolism genes indexed in the SUCAMET (Sugarcane Metabolism) Catalogue. The hybridizations were performed as shown in Table 3.1. Cultivar SP80-3280 was used for the ABA, MeJA, phosphate deficiency and herbivory experiments. Cultivar SP90-1638 was used for drought experiments and SP70-1143 for the endophytic bacteria interaction experiments. To define differential expression we used the outliers searching method (Koide *et al.*, 2006).

A total of 179 genes were identified as differentially expressed in both biological replicates in at least one of the treatments. Of these, twenty-nine were found differentially expressed in two or more treatments. Most of these (18) were responsive both to drought and phytohormones in agreement with the known role of ABA and MeJA in drought responses as

Tabela 3.1: **cDNA microarray hybridizations.** The table indicates which CyDye was used to label each sample and the experimental design. Two biological replicates were sampled for each treatment (E1 and E2) or control (C1 and C2) experiments. The table also indicates the cultivar used in each experiment.

MeJA			**ABA**			**Herbivory**			**Phosphate starvation**			***Gluconacetobacter***			***Herbaspirillum***			**Drought**		
Cy3		Cy5	Cy3		Cy5	Cy3		Cy5	Cy3		Cy5	Cy3		Cy5	Cy3		Cy5	Cy3		Cy5
0h (C1)	vs.	1h (E1)	0h (C1)	vs.	30min (E1)	30min (E1)	vs.	30min (C1)	6h (E1)	vs.	6h (C1)	E1	vs.	C1	E1	vs.	C1	24h (C1)	vs.	24 h (E1)
0h (C1)	vs.	6h (E1)	0h (C1)	vs.	1h (E1)	24h (E1)	vs.	24h (C1)	12h (E1)	vs.	12h (C1)	C2	vs.	E2	C2	vs.	E2	72h (C1)	vs.	72 h (E1)
0h (C1)	vs.	12h (E1)	0h (C1)	vs.	6h (E1)	30min (C2)	vs.	30min (E2)	24h (E1)	vs.	24h (C1)							120h (C1)	vs.	120 h (E1)
1h (E2)	vs.	0h (C2)	0h (C1)	vs.	12h (E1)	24h (C2)	vs.	24h (E2)	48h (E1)	vs.	48h (C1)							24 h (E2)	vs.	24h (C2)
6h (E2)	vs.	0h (C2)	30min (E2)	vs.	0h (C2)				6h (C2)	vs.	6h (E2)							72 h (E2)	vs.	72h (C2)
12h (E2)	vs.	0h (C2)	1h (E2)	vs.	0h (C2)				12h (C2)	vs.	12h (E2)							120 h (E2)	vs.	120h (C2)
			6h (E2)	vs.	0h (C2)				24h (C2)	vs.	24h (E2)									
			12h (E2)	vs.	0h (C2)				48h (C2)	vs.	48h (E2)									
SP80-3280												**SP70-1143**						**SP90-1638**		

discussed in the next section.

Additional file 1: Table 1 lists the differential expression (induction or repression) observed for each SAS in each treatment as well as the corresponding SUCAST categories. For reference, the table also includes the tissue expression profile for these genes in flowers, lateral buds, leaves, roots, immature and mature internodes (1^st and 4^th internodes, respectively) as published previously (Papini-Terzi *et al.*, 2005). The $\log_2$ ratio (M) values for the valid elements represented in our array for all experiments are shown in additional file 2: Table 2.

Drought elicited changes were most apparent in the late experimental data points (72 h and 120 h) as opposed to the first data point (24 h): 88% of drought-responsive genes were detected as differentially expressed exclusively after 72 h and/or 120 h of water deprivation. Conversely, the majority (78%) of the genes regulated by phosphate deficiency were detected as differentially expressed in the early data point (6 h). For the phytohormone treatments, differential expression was found throughout the experimental time-course.

In addition to the analysis of differential expression using the outliers searching method, the SOM algorithm (Tamayo *et al.*, 1999) was used to cluster the expression data for phytohormone treatments, phosphate starvation and drought. Gene expression profiles were compared between the two biological replicates. Profiles with a correlation coefficient of 0.7 or higher were identified for 158 genes in response to ABA treatment, 68 in response to MeJA treatment, 146 for phosphate deficiency and 485 for drought. The clusters obtained are partially shown in Figure 3.1 and additional files 1 and 3. The components of the SOM groups are available as additional files (see additional file 4: Table 4, additional file 5: Table 5, additional file 6: Table 6 and additional file 7: Table 7). Many of the genes included in a SOM group showing evident induction or repression patterns were not detected as being differentially expressed according to the outliers searching method. While the outliers searching method is based on criteria that take into account the intensity-dependent effect on the ratio values and data reproducibility, the clustering analysis allows for the visualization of the expression pattern along the entire time course. For this reason, both analysis were taken into account when defining sugarcane genes responsive to these treatments.

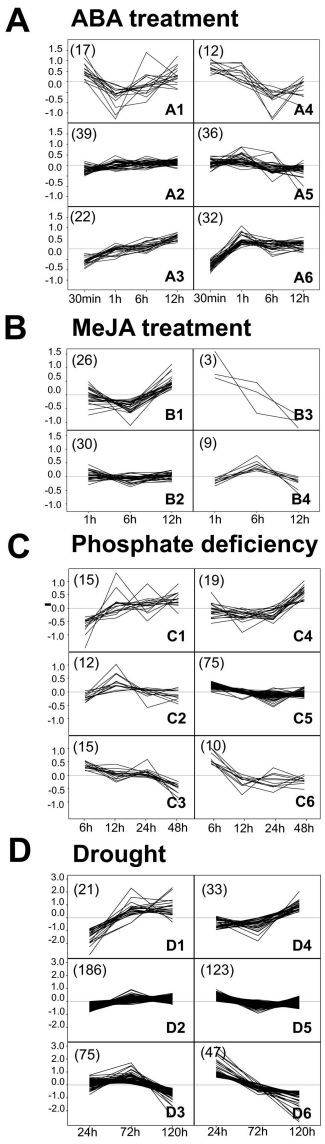


Figura 3.1: **SOM analysis for (A) ABA and (B) MeJA treatments, (C) phosphate deficiency and (D) drought.** Genes were selected based on a correlation coefficient of 0.7 or higher in the expression pattern obtained for the two biological replicates. The values of the median intensity ratios for each biological replicate were mean-centered and the average values were used as input for the SOM clustering. The geometry was chosen based on a PCA Analysis. The graphs present the average of the normalized $\log_2$ ratio (M) value between the replicates (y axis) plotted against the time course (x axis). The components of the SOM groups obtained are available in their totality as additional files (additional file 2: Table 4, additional file 3: Table 5, additional file 4: Table 6 and additional file 5: Table 7). The number in brackets indicates the number of SAS in each group.

The Sugarcane Kinome

Among the differentially expressed genes defined by the outliers search method and the SOM groups we found 185 SAS belonging to the sugarcane kinome (additional file 3: Table 3). Since 39% of SUCAST protein kinases could not be classified based on BLAST similarities and domain analysis we used a phylogenetic approach based on the analysis of *Arabidopsis thaliana* protein kinases (Shiu and Bleecker, 2001*b*) to annotate the sugarcane kinome.

Sugarcane protein kinases (277), RLKs (250) containing a putative pkinase domain and protein kinases from other organisms (156) were aligned using the neighbor-joining algorithm. The term RLCK (Receptor-like Cytoplasmic Kinase) was defined by (Shiu and Bleecker, 2001*b*) and refers to protein kinases that, in spite of having a catalytic domain very similar to the ones found for RLKs, apparently constitute cytoplasmic kinases. As it is known that RLKs/RLCKs form a monophyletic gene family with respect to other eukaryotic kinase families (Shiu and Bleecker, 2001*b*) we opted to first construct a NJ tree for sugarcane protein kinases, including only some representatives of the RLKs/RLCKs category, and then to obtain a NJ tree for RLKs/RLCKs members. A summarized view of the NJ trees obtained is depicted in Figure 3.2. A complete view of the neighbor-joining (NJ) trees is shown in additional file 8: Figure 1 and additional file 9: Figure 2. Six major groups were defined for protein kinases (KA, KB, KC, KD, KE and KF) with group KA comprising the RLKs and RLCKs representatives. Four groups were obtained (RA, RB, RC and RD) for the RLKs/RLCKs.

We observed that some SAS with BLAST best hits similar to undefined protein kinases and with no predicted transmembrane regions grouped with receptors and RLCKs in the phylogenetic analysis. For this reason, we classified these sequences as putative RLCKs instead of undefined protein kinases. In fact, some of these sequences may represent novel types of RLCKs not yet characterized. On the other hand, some of them may also be receptor-like kinases with incomplete cDNAs, lacking the extracellular domain and transmembrane region that would indicate they are receptors.

All sugarcane protein kinases were classified in the SUCAST database with the prefix cane followed by its annotation and a continuous numeration. Each undefined protein kinase or RLK received a new classification, based on the phylogenetic group and family to which it belongs. All families constituted entirely by sugarcane undefined protein kinases, RLKs or RLCKs, and supported by a bootstrap value superior to 50% received a specific

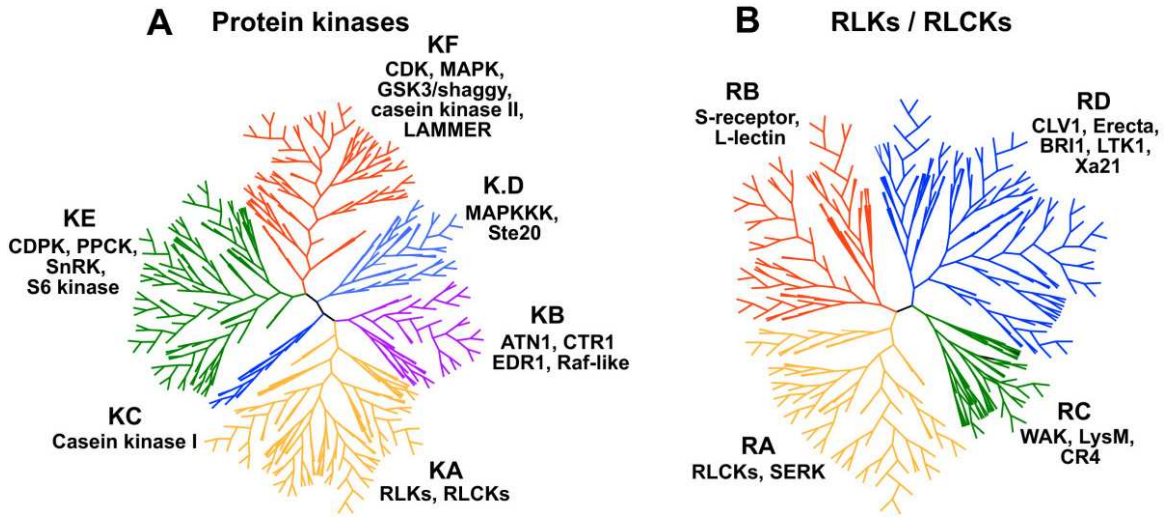


Figura 3.2: **Phylogenetic analysis of sugarcane protein kinases (A) and RLKs/RLCKs (B).** The predicted pkinase domains were aligned and used to construct a distance tree with the NJ algorithm. Only some of the main representatives of the RLK/RLCK category were included in the tree constructed for protein kinases (A). Driver sequences from other organisms were also included in the analysis. The main components of each group are: KA: RLKs and RLCKs; KB: ATN1, CTR1, EDR1, Raf-like; KC: casein kinase I; KD: MAPKKK, Ste20; KE: CDPK, PPCK, SnRK, S6 kinase; KF: CDK, MAPK, GSK3/shaggy, casein kinase II, LAMMER; RA: RLCKs, SERK; RB: S-receptor, L-lectin; RC: WAK, LysM, CR4; RD: CLV1, Erecta, BRI1, LTK1, Xa21. The complete trees indicating all SAS are available as additional files (additional file 6: Figure 1 and additional file 7: Figure 2).

nomenclature, as well as the SAS included within these families (additional file 8: Figure 1 and additional file 9: Figure 2). With this criterion, it was possible to define 6 families constituted entirely by undefined protein kinases and 33 families constituted by undefined RLKs/RLCKs.

The phylogenetic analysis allowed for the classification of 32 undefined protein kinases (additional file 8: Figure 1). Group KB contains 16 of them. Thirty-four undefined RLCKs and 117 undefined RLKs were included within the RLKs/RLCKs tree (additional file 9: Figure 2). Group RA contains most of the undefined RLCKs (76%) and group RD, the majority of the undefined RLKs (57%).

Among the 1,031 sugarcane protein kinases catalogued, 475 were represented in our array. Additional file 3: Table 3 shows expression data for all sugarcane protein kinases that were found as differentially expressed based on the outliers searching method (29) or SOM analysis (174).

Validation of microarray data by real-time PCR

Twenty-two genes were selected to have their expression data validated by quantitative real-time PCR. The primers designed for these genes and the statistical analysis (probability Pr(sample>reference) and Pr(sample < reference) for up- and down-regulated genes, respectively) of the data are shown in additional file 10: Table 8. The expression profile along the whole time-course was analysed by real-time PCR for phytohormone treatment samples. For other treatments, reactions were carried out only for the experimental point(s) in which the gene was detected as being differentially expressed.

As reference genes for normalization we used a polyubiquitin gene for the ABA treatment and drought data, a GAPDH gene for the MeJA treatment and herbivory data, a 25S rRNA gene for endophytic inoculation and a 14-3-3 gene for the phosphate starvation data. The different references were selected for their unaltered expression in each of the treatments. Curves (log fluorescence x cycles) obtained at different experimental points and the respective SAS expression profiles in the M x S space for each particular experiment were evaluated. The polyubiquitin and 14-3-3 genes were previously used as a reference for the validation of expression levels in different sugarcane tissues (Papini-Terzi *et al.*, 2005). The GAPDH and 25S rRNA genes were described as good references for sugarcane tissues and genotypes (Iskandar *et al.*, 2004). The 25S rRNA primers used were 25S rRNA1F and 25S

rRNA1R (Iskandar *et al.*, 2004).

A total of 36 results of differential expression were evaluated (additional file 10: Table 8). Of these, 80.5% had a profile in real-time PCR assays consistent with the one observed in the microarray experiments (probability value of 0.99 or higher). Validated real-time PCR results are depicted in Figure 3.3. It is important to emphasize that the RNA samples used in the real-time PCR experiments derived from a third biological sample and that the principles of real-time PCR techniques are different from the ones applied in microarray experiments. The conflicting results may correspond to biological variations in the third biological replica or even to technical limitations of the microarray method. Nevertheless, our analysis and statistical methods were efficient in evaluating differentially expressed genes, yielding only a minor percentage of unconfirmed expression data.

The SUCAST Database

A database containing all the SAS catalogued in the SUCAST Project and their respective expression data was built and is available at the SUCEST-FUN Database web site. The SUCEST-FUN database includes the expression data associated to stress responses and environmental stimuli and the expression profile of SUCAST SAS in six different sugarcane tissues (Papini-Terzi *et al.*, 2005). It also includes the SUCAMET categories of sugarcane metabolism genes.

The SUCAST databank integrates the sequence data and analysis from the SUCEST Project (Vettore *et al.*, 2003), the categorization and tissue gene expression of signal transduction genes (Papini-Terzi *et al.*, 2005) with the kinome analysis and gene expression data in response to different treatments as pointed out by the outliers searching method, SOM and quantitative PCR analysis (this work).

The SUCAST system consists of a client web interface and a server back end. The database was constructed using the MySQL database server. The scripts were written in Perl and R statistical language. Through the web interface, the SUCAST database can be easily queried to find each SAS and its associated information. For each SAS it is possible to retrieve the consensus sequence of the SAS and the alignment of its corresponding reads, according to the clusterization of SUCEST reads (Vettore *et al.*, 2003).

Besides expression data, the SUCAST database also provides, for each SAS queried, information regarding annotation, results of the blasts against the GenBank nr and the

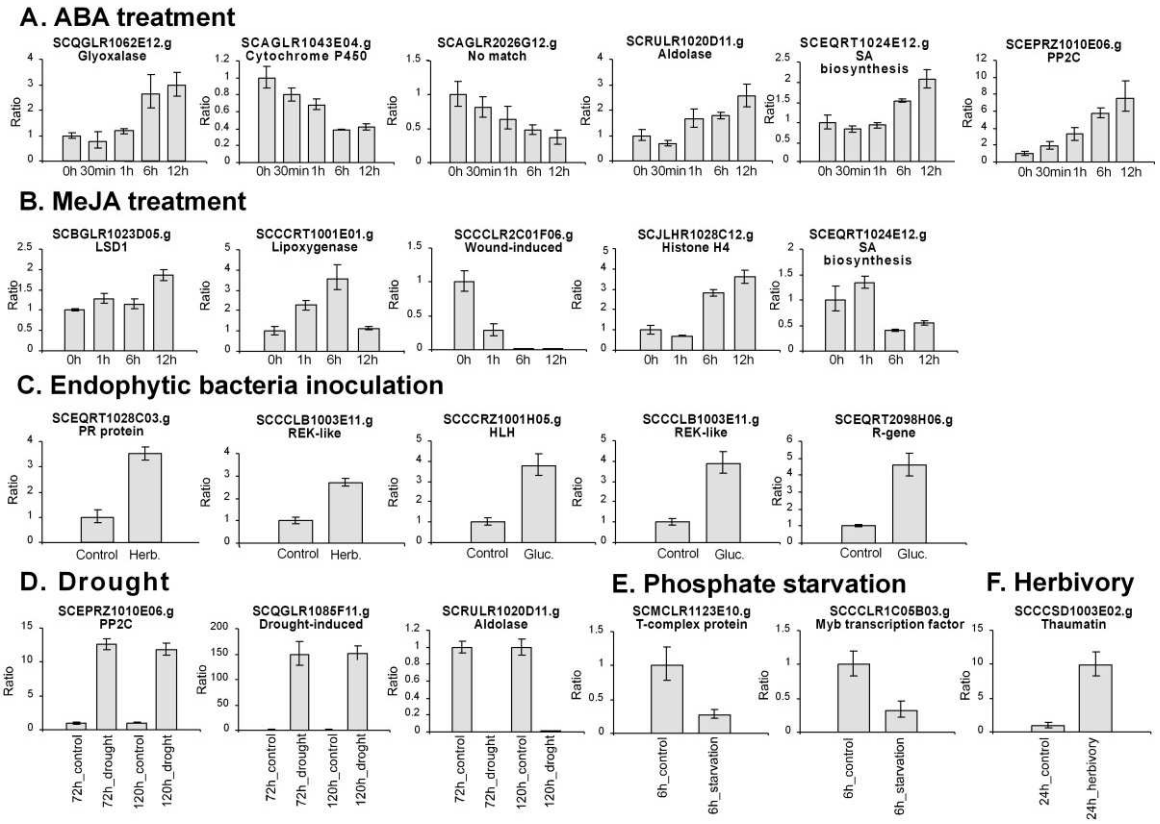


Figura 3.3: **Validation of microarray results by quantitative PCR analysis.** The y axis refers to the relative expression ratio between treated samples versus the control (untreated sample). (A) and (B) Real-time PCR results for ABA and MeJA treatments, respectively. The ratios were calculated in relation to the sample from untreated plants (0h). Transcript levels of the selected genes were profiled throughout the treatment time course. Also shown are the results for plant-endophytic bacteria association (C), drought (D), phosphate starvation (E) and herbivory (F). For these treatments, the real-time PCR reactions were carried out exclusively for the experimental point(s) in which the gene was considered differentially expressed. Only validated results are shown here. The RNA samples used in the real-time PCR experiments are from a third biological sample. All reactions were carried out in parallel and each reaction was performed in triplicates. Error bars were calculated as described previously (Livak and Schmittgen, 2001). Herb. = sample from plants inoculated with *Herbaspirillum seropedicae* and *Herbaspirillum rubrisubalbicans*; Gluc. = sample from plants inoculated with *Gluconacetobacter diazotrophicus*. The transcript levels for the reference genes were verified to not vary in response to the treatments. The reference genes used encode a polyubiquitin gene (SCCCST2001G02.g [CA179923]) for the ABA and drought dataset, a GAPDH for the MeJA and herbivory dataset (retrieved from (Iskandar *et al.*, 2004)), a 25S rRNA for the endophytic inoculation (retrieved from (Iskandar *et al.*, 2004)) and a 14-3-3 gene (SCCCLR1048F12.g [CA119519]) for phosphate starvation.

NCBI GEO databases and comparisons with the Gene Ontology database. As of the time of submission 1,083 SAS showed similarity against sequences in NCBI GEO database (using BLASTN tool, cutoff Evalue of 10^{-5}). Comparison of these SAS with the Gene Ontology database (using BLASTX tool, cutoff Evalue of $e^{-}5$) revealed significant matching with proteins for 3,030 SAS.

Over 3,600 SAS are compiled in the SUCAST database. It is possible to retrieve the predicted protein sequence encoded by the consensus sequence as well as information on conserved protein domains predicted by the Pfam (Sonnhammer *et al.*, 1998) databank. The SUCAST elements are distributed into 46 main categories subdivided into 1,678 subcategories. The database also includes 405 SUCAST SAS that present no matches with existing protein sequences in the GenBank nr database (using BLASTX tool, cutoff E value of $10^{-}5$).

Discussion

Identification of sugarcane genes responsive to biotic and abiotic stimuli

In an effort to associate gene expression changes to environmental factors that may affect sugarcane yield we profiled the expression of 1,545 sugarcane genes in response to drought, phosphate deficiency, herbivory and endophytic bacteria. We also analysed the responses of plants to ABA and MeJA treatments. The 1,545 genes selected are representative of all categories found in the SUCAST Catalogue. The outliers searching method has proved to be a robust way of differential expression analysis that considers the overall dispersion of signal intensity and provides reliable gene expression differences. A complementary approach useful to highlight the patterns observed and to include data points where differential expression was not striking was the selection of patterns highly correlated among biological replicates and their grouping by the SOM algorithm. This approach increased the number of genes analysed and confirmed the differences observed through the use of the outliers searching method. For instance, there were three histone genes (two H4 histone and one H2B histone) in SOM group B1 of genes up-regulated by MeJA. Only one of the H4 histone genes was detected as being up-regulated by MeJA exposure (after 12 h) according to the outliers searching method. This exemplifies the usefulness of the SOM analysis in addition to the outliers analysis of differentially expressed genes, revealing two additional histones induced by MeJA. Probably, the values of $\log_2$ ratio (M) for these SAS were not enough to surpass the

intensity-dependent cutoff levels that indicated differential expression, but the SOM analysis revealed that these SAS apparently were responsive to the MeJA treatment.

From the 179 genes detected as differentially expressed, 113 were also included in at least one of the SOM clustering analysis performed (additional file 1: Table 1). Some had their expression patterns evaluated by real-time PCR along the entire time course of phytohormone treatment (SCBGLR1023D05.g [CA117725], SCJLHR1028C12.g [CA106117], SCEQRT1024E12.g [CA132523], SCQGLR1062E12.g [CA124203] and SCRULR1020D11.g [CA125940]) and presented consistent results.

Analysis of sugarcane genes responsive to phytohormones and environmental challenges

Tissue expression data was obtained for most genes of the array in a previous study where samples extracted from six different sugarcane tissues were hybridized against a common reference sample (Papini-Terzi *et al.*, 2005). These studies indicated 217 genes differentially expressed in at least one of the tissues tested and 153 genes evenly expressed in all tissues. Among the 179 differentially expressed genes detected in the present study, 70 correspond to tissue-enriched genes. Most of them (19) are leaf-enriched, but there were also genes that were expressed preferentially in sugarcane roots (12) and internodes (11). While these are interesting observations, it is important to emphasize that they should be considered with caution, since the experiments reported here used plantlets cultivated under growth chamber or greenhouse conditions while in our previous studies most of the samples were collected from 12- or 14-month-old plants cultivated in the field.

Water deprivation was the condition that elicited the majority of gene expression changes. Fifty-two percent of the 179 differentially expressed genes were responsive to drought. Many studies have reported the identification of genes regulated by drought and altered expression of transcription factors. Specific recognition sequences for some categories of transcription factors were detected in the promoter of drought-responsive genes, as is the case of MYC and MYB recognition sequences and the W-box for WRKY transcription factors (Abe *et al.*, 1997, Narusaka *et al.*, 2004). We observed the induction of one MYB and two WRKY transcription factors in response to drought (group D4).

Many of the genes responsive to drought are similar to genes that in other systems have been shown to transduce additional stress signals including cold. The cold and drought

signaling pathways present a high degree of overlap and many of the responses are mediated by ABA (Chinnusamy *et al.*, 2004). Four SAS encoding low temperature induced (LTI) proteins were up-regulated in response to lack of water. Two of these SAS (SCUTST3084F06.g [CA186860] and SCACCL6008H06.g [CA096029], group D1, Figure 3.1) are grouped with genes induced by water deprivation. Many genes co-regulated in response to drought and cold in *Arabidopsis* (Seki *et al.*, 2001) have a DRE (Dehydration-Responsive Element) motif or DRE-related motifs in their promoters. A transcription factor (SCBGLR1002A09.g [CA117666]) from the AP2 family and homologous to rice DREB2 (DRE binding factor 2) [AAP70033] was induced after 72 h of watering suppression and may represent an important transcription factor for the regulation of sugarcane drought responsive genes. The overexpression of a constitutive active form of DREB2A [O82132] in *Arabidopsis thaliana* led to the development of transgenic plants more tolerant to drought (Sakuma *et al.*, 2006). Hence, the manipulation of DREB2 levels in sugarcane may represent a way of obtaining new varieties with increased resistance to water deficit. Genes induced by ABA and drought include two delta-12 oleate desaturase (SCCCLR1C03G01.g [CA189695] and SCVPST1061G05.g [CA179715]), one S-adenosylmethionine decarboxylase (SCCCLR1C05G07.g [CA189868]) and one PP2C-like protein phosphatase (SCEPRZ1010E06.g [CA147516]) homologous to *Arabidopsis* protein phosphatases ABI1 and ABI2. The protein phosphatases ABI1 and ABI2 are responsive to ABA and regulate a range of physiological responses, including stomatal closure, which minimizes the transpirational water loss (Merlot *et al.*, 2001, Tahtiharju and Palva, 2001). S-adenosylmethionine decarboxylases participate in the polyamine biosynthetic pathway, which is modulated in response to abiotic stresses (Galston and Sawhney, 1990). SCCCLR1C05G07.g [CA189868] is homologous to the S-adenosylmethionine decarboxylase *SAMDC1* [AF067194], from rice, known to accumulate in response to salinity and drought, probably through ABA-dependent pathways and with an expression positively correlated with salt tolerance (Li and Chen, 2000). The regulation of fatty acid desaturases (FAD2) may be related to changes in the degree of fatty acid desaturation in response to environmental stresses. FAD3, FAD7 and FAD8 desaturases expression is directly related to drought tolerance. A role of an omega-3 fatty acid desaturase in drought tolerance was reported in tobacco through overexpression (Zhang *et al.*, 2005) and gene silencing (Im *et al.*, 2002) studies. Additionally, it was demonstrated that the reduction in trienoic fatty acid levels by the antisense expression of the *fad7* [D26019] gene seems to affect the ABF (ABA responsive elements Binding Factor)-dependent gene expression showing a relationship between

desaturases levels and ABA-signaling pathways (Im *et al.*, 2002). Thus, the regulation of the sugarcane delta-12 oleate desaturases by ABA and drought may indicate that these genes are induced by drought and that this induction alters ABA signaling pathways.

A total of 31 differentially expressed genes were found in plants treated with ABA. Fifty-eight percent of these are exclusively regulated after 12 h of treatment with this hormone, indicating the existence of early and late response genes in the time course of our experiment. Among the ABA-responsive genes, we could observe the induction of a gene (SCCCLR1C07B07.g [CA189990]) encoding a glycine-rich protein with a predicted RNA recognition motif. The function of this class of proteins is not clear (Sachetto-Martins *et al.*, 2000, Mousavi and Hotta, 2005) but it is known that some of them play an important role in RNA turnover (van Nocker and Vierstra, 1993). A *Sorghum bicolor* gene [AF310215] similar to SCCCLR1C07B07.g [CA189990] was induced by ABA treatment, light and salinity (Aneeta Sanan-Mishra *et al.*, 2002). The maize MA16 protein is also an example of an RNA-binding protein induced by ABA (Himmelbach *et al.*, 2003, Freire and Pages, 1995).

Several regulators of ABA signaling pathways are described and characterized (Himmelbach *et al.*, 2003). Among these, the ROP10 small GTPase [NP_566897] from the Rab family in *Arabidopsis* was implicated in the down-regulation of the ABA signal transduction pathway (Zheng *et al.*, 2002). Our analysis revealed two GTPases similar to Rab11 (SCACCL6006D08.g [CA095849] and SCJFRT1059D05.g [CA134244]) induced after exposure to this hormone. The OsRab7 GTPase from rice [AAO67728] (Nahm *et al.*, 2003) and the Rab2 GTPase [AAD30658] from *Sporobolus stapfianus* (O'Mahony and Oliver, 1999) were also described as regulated by ABA. These observations point to an involvement of different Rab GTPases in the cellular responses activated by this hormone.

Among the genes up-regulated by MeJA treatment there were two H4 histones and one H2B histone genes in SOM group B1. The H4 histone SCCCLR2002G09.g [CA127138] was also identified as up-regulated after 12 h of MeJA exposure by the outliers searching method. It has been reported that jasmonates may regulate gene expression by interfering with histone acetylation and deacetylation since COI1 [O04197], an F-box protein required for jasmonates responses was able to target an *Arabidopsis* histone deacetylase to proteolysis (Devoto *et al.*, 2002). Furthermore, Kim and colleagues (Kim *et al.*, 1998) observed the induction of histones CaH2B [AF038386] and CaH4 [AF038387] from *Capsicum annuum* by MeJA. The regulation of histone transcript levels in sugarcane points towards chromatin remodeling as a possible event activated by jasmonates which may represent an

important mechanism through which jasmonates regulate the expression of target genes.

A comparison of sugarcane and rice (Lin *et al.*, 2003) ABA responsive genes indicates that a fructose-bisphosphate aldolase (SCRULR1020D11.g [CA125940]), a glyoxalase (SCQGLR1062E12.g [CA124203]) and a RUBISCO gene (SCCCLR1001E04.g [CA116155]) are induced by ABA in both grasses. The sugarcane fructose-bisphosphate aldolase induced by ABA is similar to the *NpAldP1* [AB027001] gene expressed in *Nicotiana paniculata* leaves and repressed in response to saline stress (Yamada *et al.*, 2000). The sugarcane aldolase gene was also regulated by drought and MeJA treatments (additional file 1: Table 1, Figure 1 groups A3 and D6) and according to our previous work (Papini-Terzi *et al.*, 2005) this SAS is enriched in sugarcane leaves. This suggests a potential role of this gene in signaling pathways specific of this organ. Another SAS encoding an aldolase seems to be slightly induced by ABA treatment (Figure 3.1, group A2). The analysis of transcripts levels for genes of the glycolytic and fermentation pathways in rice roots and shoots indicated the induction of an aldolase gene in response to saline stress (Minhas and Grover, 1999). In another work, an *Arabidopsis* aldolase gene was repressed by ABA (Seki *et al.*, 2002). A wealth of evidence has accumulated throughout the years revealing important interactions between sugar- and phytohormone pathways (Gibson, 2004). Additionally, ABA is implicated in the regulation of sugar transport and metabolism. The regulation of glycolytic enzymes by stressful conditions and the consequences of this regulation are particularly interesting for sugarcane, since these findings may indicate a relationship between sucrose accumulation and responses to stresses.

The identification of differential expression for genes with no homologs in the public databases ("no matches") as well as of SAS coding for unknown proteins is particularly valuable for the identification of their putative roles. We obtained 28 "no matches"differentially expressed in at least one of the experiments analysed. Of these, 13 are regulated by drought, 7 by inoculation with *Herbaspirillum*, 5 by ABA treatment, 4 by MeJA treatment, 3 by herbivory and 3 by phosphate starvation. Four of these are leaf-enriched genes. One of them was induced by both inoculation with N_2-fixing endophytic bacteria and exposure to insect attack. This may indicate a possible role for this gene in general mechanisms of defense against biotic stimuli, including endophytic recognition and the activation of defense responses until the establishment of an efficient association. It is also interesting to point out that five of these 28 no matches genes do not present a predicted coding region and may represent non-coding transcripts. Recent studies have established important

roles for some plant microRNAs in the regulation of processes like development, response to pathogens and hormone signaling (Zhang *et al.*, 2006). Among the SUCEST sequences, 239 non-coding no matches were identified. Through the present studies we see an indication that five of them may have a role in stress responses since drought regulated three of them, methyl jasmonate treatment regulated one and inoculation with *Herbaspirillum* spp regulated another.

Phosphate starvation altered the expression of 14 genes. The majority of them (11) showed decreased levels after 6 hours of starvation. Expression data indicates that during this early phase of the stress response an alteration in protein N-glycosylation may occur, as can be inferred from the repression of a gene coding for an N-acetylglucosamine-1-phosphate transferase (SCRUFL1112F04.b [CA249652]). The decreased expression of two genes coding for thioredoxins (SCCCLR2001H09.g [CA127047] and SCJFLR1073B06.g [CA122039]) indicates there are changes in the redox state of sugarcane roots in response to phosphate starvation, since these enzymes are important regulators of the intracellular redox status (Gelhaye *et al.*, 2004). A gene similar to MYB transcription factors had transcript levels reduced. Several members of this gene family show distinct expression profiles in response to phosphate starvation in *Arabidopsis*, some of them being up-regulated and some down-regulated (Wu *et al.*, 2003). It is worth to note that the promoter region of the oat homolog of this gene (*MybHv1*) [X70879] has been characterized and shown to be active only in the root apex (Wissenbach *et al.*, 1993). Alterations in the expression of root apex enriched genes could be related to the morphological changes observed in the root system in response to low levels of P.

Even though the plant-endophytic association is advantageous for both organisms, it is believed that sugarcane plants recognize these microorganisms and activate defense responses until the establishment of an efficient association (Vinagre *et al.*, 2006). In agreement with this, four R-genes were found among the genes responsive to the endophytic association. Plant disease resistance (R) genes mediate specific recognition of pathogens via perception of avirulence (avr) gene products (Ellis *et al.*, 2000). Two of them were induced by both associations under study (sugarcane-*Herbaspirillum* spp and sugarcane-*Gluconacetobacter diazotrophicus*). The inoculation with *Gluconacetobacter* also led to the induction of a salicylic acid biosynthesis gene. The phytohormone salicylic acid accumulates in plant tissues in response to pathogen attack and is essential for the induction of systemic acquired resistance and for some responses mediated by resistance genes (Gaffney *et al.*,

1993, Delaney *et al.*, 1994, Mur *et al.*, 1997).

A PP2C (SCJLRZ3077G10.g [CA160745]) was up-regulated in plants inoculated with *Gluconacetobacter*. Also, expression of five transcription factors was altered when the plants were cultivated in association with endophytic bacteria. Among these, there were two zinc-finger transcription factors (SCEQRT1033F01.g [CA133313] and SCEZST3147A10.g [CA182656]), one of which was up-regulated by inoculation with either *Gluconacetobacter* or *Herbaspirillum*. In agreement with our data, a possible role for phosphatases and zinc-finger transcription factors in response to endophytic bacteria has also been pointed out by the *in silico* analysis of the SUCEST Project libraries, that identified SAS corresponding to these categories exclusively or preferentially expressed in the SUCEST cDNA libraries constructed from plants inoculated with *Gluconacetobacter* and *Herbaspirillum* (Vargas *et al.*, 2003).

Expression data for the herbivory experiment points to the strong induction of a pathogenesis-related protein similar to a thaumatin after 24 h of the onset of this stress. This is not surprising, since it is known that proteins from this category are important for plant defense mechanisms and may present antifungal action, endo-β1,3-glucanase activity and trypsin or a-amylase inhibitory activity (Grenier *et al.*, 1999, Franco *et al.*, 2002). Further characterization of this sugarcane thaumatin-like protein should be carried out in order to define its activity and the defense mechanism that this protein may confer against the sugarcane stalk borer.

Sugarcane-responsive genes related to hormone biosynthesis and signaling

Phytohormone signaling pathways exhibit a wide degree of cross-talk among their components creating a complex network of overlapping signaling (Moller and Chua, 1999, Gazzarrini and McCourt, 2003). Interactions among phytohormone signaling pathways are highly complex and the features of these interactions are time and space dependent. Although we are only beginning to outline signaling cross-talks in sugarcane, the analysis of the expression profiles of the differentially expressed genes obtained (additional file 1: Table 1) as well as the groups obtained in the clustering analysis (Figure 3.1 and additional files 4, 5, 6, 7) uncover some aspects of these interactions. Auxin signal transduction pathways appear to be activated in response to several of the treatments studied. ABA treatment elicited an antagonistic response between the ABA and auxin pathways. A gene

(SCCCCL3002B05.b [CA093260]) coding for a protein similar to the auxin responsive protein GH3 (Hagen *et al.*, 1984) was found repressed by ABA (group A4). Furthermore, a gene (SCCCLR2002F08.g [CA127125]) coding for a protein with a predicted auxin repressed domain found in dormancy-associated- and auxin-repressed proteins (Stafstrom *et al.*, 1998) was up-regulated by this hormone (group A6). It has been shown that ABA and auxin interact antagonistically to regulate stomatal aperture (Eckert and Kaldenhoff, 2000) and the interaction between auxin and ABA signaling pathways has been demonstrated by the dual specificity of the ABI3 transcription factor, which is able to bind sequences upstream of ABA and auxin responsive genes. In the presence of ABA, ABI3 binds to the GH3 like promoter sequences and inhibits the auxin-mediated induction (Nag *et al.*, 2005).

Auxins have been implicated in several aspects of the drought response including proline accumulation (Sadiqov *et al.*, 2002), rhizogenesis (Vartanian *et al.*, 1994) and indole 3-butyric acid increases (Ludwig-Muller *et al.*, 1995). Our data also indicate auxin signaling in response to drought. We observed the induction of genes encoding auxin biosynthesis enzymes (nitrilases) after 72 h of water deprivation and the down-regulation of transcription factors from the Aux/IAA category after 120 h of drought. These transcription factors work by inhibiting auxin signaling and are rapidly induced by auxin exposure (Abel and Theologis, 1996, Woodward and Bartel, 2005). However, Aux/IAA accumulation is subject to negative feedback, since auxins target Aux/IAA for degradation by the 26S proteasome (Woodward and Bartel, 2005). Groups D1 and D2 contain two nitrilases and one auxin-binding protein (ABP). Additionally, one auxin response factor is induced by drought (group D4) and four AUX/IAA transcription factors are modulated (groups D2, D3 and D6).

Phosphate starvation leads to alterations in root architecture, resulting in increased soil exploration and phosphate acquisition. In this process of morphological adaptation, auxins and other phytohormones play important roles in root elongation and lateral root development (Hammond *et al.*, 2004). López-Bucio and colleagues (Lopez-Bucio *et al.*, 2002) showed that phosphate deprivation increases auxin sensitivity in *Arabidopsis*, what may explain the increased number of lateral roots observed when the plant is under nutritional stress. In agreement with this observation, the auxin-repressed protein found in group A6 (SCCCLR2002F08.g [CA127125]) was repressed after 6 h of phosphate starvation. This same SAS was also down-regulated by MeJA treatment what may indicate a possible synergism among MeJA and auxin pathways. Even though the majority of interactions between MeJA and auxins are antagonistics, there is evidence that these hormones may act synergistically

at the post-transcriptional level (Swarup *et al.*, 2002). It is hypothesized that, since COI1 and TIR1, components of SCF (SKP1, CDC53p, CUL1, F-box protein) complexes associated to jasmonate- and auxin-responses, respectively, are highly similar, these two signaling pathways may converge to the degradation of common target regulatory proteins (Swarup *et al.*, 2002). Phosphate starvation also causes a reduction in the expression of a gene (SCEZLB1009A09.g [CA113117]) similar to BLE1 from rice. Rice plants where the gene *OsBle1* [AB072977] was knocked-out showed reduced growth rates (Yang and Komatsu, 2004). The repression of the sugarcane homolog of *OsBle1* in the early phase of phosphate starvation could be an effort to restrain metabolism as occurs in *Arabidopsis* in response to low levels of the nutrient (Wu *et al.*, 2003). Another hormone signaling pathway that seems to be altered in response to phosphate starvation is the ethylene response pathway, since a gene for the EIL transcription factor (SCBGFL4052C11.g [CA221542]) is down-regulated after 6 h of starvation. The rice homolog of this protein, OsEIL1 [AAZ78349], acts as a positive regulator of the ethylene response and transgenic rice plants overexpressing OsEIL1 exhibit short root, coiled primary root, slightly short shoot phenotype and elevated response to exogenous ethylene (Mao *et al.*, 2006). The down-regulation of this transcription factor could be related to the changes in root architecture that occur in response to phosphate starvation.

The Sugarcane Protein Kinases and RLKs

Detailed descriptions of yeast, *Drosophila*, *C.elegans* and human kinomes are available (Manning *et al.*, 2002*a*, Manning *et al.*, 2002*b*) as well as studies on plant kinases (Shiu and Bleecker, 2001*a*, Shiu and Bleecker, 2001*b*, Shiu *et al.*, 2004). Since plant protein kinases and RLKs act as critical regulators of many signaling pathways, they represent important targets to modify pathways of interest.

Eight protein kinases were differentially expressed in response to plant inoculation with N_2-fixing bacteria. Three of these are similar to proteins involved in calcium signaling, a calcium-dependent protein kinase (caneCDPK-9) and two calcineurin B-like interacting protein kinases (caneCIPK-18 and caneCIPK-22) of the SnRK3 subgroup of plant kinases (Shi *et al.*, 1999, Hrabak *et al.*, 2000). Some reports have shown the role of calcium-dependent pathways in the processes of symbiosis and nodulation (Weaver and Roberts, 1992, Webb *et al.*, 2000, White, 2001, Levy *et al.*, 2004). CDPKs may participate in pathogen defense signaling pathways, as seen in the tomato defense responses against the fungi

Cladosporium fulvum (Romeis *et al.*, 2000). We also observed the induction of a sugarcane gene similar to a GSK3/shaggy protein (caneGSK3-5) kinase by endophytic bacteria association. In plants, these kinases are associated to floral development, brassinosteroid signaling pathways and responses to stresses such as wounding and salinity (Jonak and Hirt, 2002). Moreover, one PBS1-like protein kinase (canePBS1-4) had its transcripts increased in response to the association. The *Arabidopsis* PBS1 [NP_196820] protein recognizes avirulence factors from *Pseudomonas syringae* (Swiderski and Innes, 2001). The transcriptional regulation of this gene in sugarcane inoculated with N_2-fixing bacteria suggests a possible role for this protein in the recognition of these microorganisms.

The role of some receptors in the regulation of symbiosis has been described (Endre *et al.*, 2002, Madsen *et al.*, 2003, Capoen *et al.*, 2005). Our data indicates the induction of a putative receptor with predicted leucin-rich repeats (caneURLK-13) in plants inoculated with *Herbaspirillum* that may be regulating such an interaction. Recently, a sugarcane receptor (SHR5) [AAY67902] was shown to be repressed in plants associated with endophytic bacteria and the degree of this repression was directly related to the success of the sugarcane-endophytic bacteria association, indicating a participation of this receptor in signal transduction pathways involved in the establishment of plant-endophytic bacteria interaction (Vinagre *et al.*, 2006).

We identified ten protein kinases differentially expressed in response to drought. Seven of them are similar to the SnRK family of proteins. Four were induced by this stress (caneOsmotic stress-activated protein kinase-2, caneCIPK-8, caneCIPK-13 and caneCIPK-14). Some SnRKs are recognized players in stress responses. SRK2C leads to improved drought tolerance when overexpressed in *Arabidopsis thaliana* (Umezawa *et al.*, 2004). Mutagenesis studies on OST1 [NP_567945], a kinase whose activity is induced by drought, led to guard-cell specific effects and ABA insensitivity (Riera *et al.*, 2005, Mustilli *et al.*, 2002). Other works describing the function of this family of protein kinases in drought responses include a role in stomatal closure (Li and Assmann, 1996, Li *et al.*, 2000, Yoshida *et al.*, 2002). Furthermore, among the drought responses, the Ca^{2+}-dependent SOS signaling pathway (which involves SOS2, a SnRK similar to the CIPKs) has an important role in regulating ion homeostasis (Xiong *et al.*, 2002). Since *sos2* [NM_122932] mutants are hypersensitive to saline stress (Liu *et al.*, 2000) it will be interesting to complement our studies with a phenotypic evaluation of plants altered for caneCIPK-8, caneCIPK-13 or caneCIPK-14 genes to confirm a role for these kinases in drought responses.

Among the differentially expressed genes, six undefined kinases/RLKs (caneRLCK-AVI2, caneRLCK-DII3, canePK-BIII3, caneRLK-AX1, caneRLK-AX2 and caneRLK-C5) were regulated by drought, inoculation with *Herbaspirillum* spp. and/or phytohormone treatments and 64 undefined kinases/RLKs were selected for the SOM clustering analysis. Six protein kinases that grouped within the same phylogenetic family in group KE have very similar catalytic domains, with an insertion of around 80 aminoacids between subdomains VII and VIII, as defined by Hanks *et al*. (Hanks and Hunter, 1995,Hanks and Quinn, 1991,Hanks *et al.*, 1988). These proteins are similar to the protein kinase G11A from rice [AAA33905] (Lawton *et al.*, 1989). One of these SAS (caneG11A kinase-2) was included in SOM group A2 and two of them (caneG11A kinase-3 and caneG11A kinase-5), in group D2, with an apparent expression profile of induction by ABA or drought, respectively. It may be of interest to evaluate potential targets of these uncharacterized protein kinases and also to investigate if the sequence between subdomains VII and VIII plays a role in substrate recognition or catalytic reaction.

It is important to emphasize that our phylogenetic analysis has limitations imposed by the fact we are dealing with an EST databank. For example, many putative sugarcane protein kinases were excluded from our analysis since the available sequences do not present most of the *pkinase* subdomains. Even though, the groups obtained are in good agreement with the classes, groups and families of plant protein kinases previously defined by the PlantsP database, despite differences in their classification methodology (Gribskov *et al.*, 2001).

Conclusion

In this work, the expression of 1,545 sugarcane SAS (mostly related to signal transduction components) was evaluated by cDNA microarrays in plants submitted to a variety of challenges: drought, phosphate starvation, herbivory by *Diatraea saccharalis* and endophytic bacteria inoculation (*Herbaspirillum seropedicae/Herbaspirillum rubrisubalbicans* and *Gluconacetobacter diazotrophicus*). Additionally, plants were treated with the phytohormones ABA and MeJA, important players in the responses to biotic and abiotic stresses.

To our knowledge, this is the first broad evaluation of sugarcane gene expression in response to biotic, abiotic and hormone inputs. Since ABA and MeJA play main roles in plant responses to a plethora of stimuli the analysis of genes regulated by these hormones is

helpful in deciphering sugarcane defense pathways activated in response to stresses. Many of the differentially expressed genes belong to protein families described in the literature as associated to some of the processes studied, indicating that sugarcane responses are similar to those of other well-known plants, such as rice, maize and *Arabidopsis*. Additionally, functions were associated to genes poorly studied or novel genes such as genes with no hits in the public databases, genes encoding unknown proteins and undefined kinases/RLKs. The information generated by the protein kinase categorization using a phylogenetic approach, associated to the expression data obtained from microarray experiments, represents a useful tool in guiding the future characterization of these proteins.

Understanding the molecular mechanisms behind sugarcane stress responses will be useful for the improvement of sugarcane yield by genetic manipulation. This knowledge, allied to the use of genetic engineering, will potentially enable the development of sugarcane varieties tolerant to adverse conditions, such as drought and nutritional deficiency. Furthermore, the genes may be explored as molecular markers in traditional breeding programs or have their promoters cloned to accomplish transgene expression activated solely by a specific stimulus. It is important to emphasize the limitations intrinsic to the nature of the data presented. First, changes in mRNA levels do not always correlate to protein levels. Second, in the field, plants are exposed to a diversity of stressful conditions and the responses achieved by this combination of stimuli probably are not the same as the ones triggered by each individual stimulus. Nonetheless, the data generated in controlled experiments certainly represent an important step in the exploration of specific responses. The data also points candidates for gene silencing or overexpression experiments that may corroborate the hypothesis raised. With this in mind an expression panel is currently being constructed for several additional sugarcane cultivars tolerant or more susceptible to the stimuli that certainly will be valuable in guiding the selection of target genes. It will be important to expand the present studies to additional genotypes also if one wishes to compare the responses elicited by the different stimuli. The data obtained may reflect cultivar specific responses in the case of drought and endophytic bacteria interaction, since different cultivars were used in these experiments. The extent of genotypic variation among commercial cultivars is currently unknown. Evaluation of sugarcane responses in additional genotypes is underway to further validate commonly regulated pathways.

A databank was created that provides public access to the data described in this work, associated to tissue expression profiling and the SUCAST gene categories. As the SU-

CAST Project is an ongoing effort that aims to identify sugarcane signaling components and define their role in grasses, the database is expected to be updated each time new expression data from experiments with the SUCAST arrays are available. We expect the SUCAST database to become a useful tool for sugarcane transcriptome data mining and in guiding the selection of target genes to be modified in sugarcane and other grasses.

Methods

Plant material and cultivation

The cultivar SP90-1638 (Internal Technical Report, CTC, 2002), sensitive to drought, was used for the water deprivation experiments. The cultivar SP80-3280 sensitive to herbivory by *Diatraea saccharalis* (Barsalobres, 2004) was adopted for the herbivory experiments. The same cultivar was used for phosphate deficiency and phytohormone treatment experiments. The cultivar SP70-1143 with high inputs of nitrogen obtained from BNF and with efficient association with endophytic bacteria (Urquiaga *et al.*, 1992) was used for inoculation with *Gluconacetobacter* and *Herbaspirillum*.

Sugarcane plantlets obtained from one-eyed seed sets were used for methyljasmonate treatment, water stress and herbivory experiments. For methyljasmonate treatments, one-eyed seed sets were planted in 200 ml plastic cups containing a commercial planting mix (Plantmax, Eucatex) for 20 days under greenhouse conditions and subsequently transferred to a growth chamber at 26°C on a 16 h/8 h light/dark cycle with a photon flux density of 70 μE.m^{-2}.s^{-1} to acclimate for 48 h. For the insect attack assays, one-eyed seed sets were planted in 200 ml plastic cups as described above and maintained in the greenhouse for 60 days when they were transferred to a growth chamber at 28°C on a 14 h/8 h light/dark cycle with a photon flux density of 70 μE.m^{-2}.s^{-1}. Sugarcane one-eyed seed sets were cultivated on moist sand for 15 days prior to drought experiments. For ABA treatment, plants derived from shoot apex of 2-month-old sugarcane plants were axenically *in vitro* cultivated for approximately three months in a growth chamber at 26°C on a 16 h/8 h light/dark cycle with a photon flux density of 70 μE.m^{-2}.s^{-1}. Sugarcane rooted plantlets obtained by sterile *in vitro* meristem culture and micropropagated according to the method of Hendre *et al*. (Hendre *et al.*, 1983) were used for nutritional deficiency and inoculation with endophytic bacteria experiments.

Plant treatments

Plant treatments are described below. Three biological replicates were performed for each of these treatments. Two of the replicates were used for microarray experiments and one for real-time PCR reactions.

MeJA treatment

Plantlets were sprayed with a 100 μmol.L^{-1} MeJA solution (Bedoukian Research Inc., Danbury, CT), whereas control plantlets were treated with distilled water. Leaves were collected 0, 1, 6 and 12 h after exposure to MeJA and immediately frozen in liquid nitrogen. Six plantlets were sampled for each time point.

ABA treatment

ABA (Sigma Chem. Co) was added to the culture medium to a final concentration of a 100 μmol.L^{-1} whereas control plants were treated with distilled water. Leaves were collected 0, 0.5, 1, 6 and 12 h after exposure to ABA and immediately frozen in liquid nitrogen. Six plantlets were sampled for each time point.

Phosphate deficiency

Rooted plantlets were greenhouse acclimatized by initial cultivation on 1/20th strength Hoagland and Arnon (1950) nutrient solution. Nutrient solutions were replaced every 7 days increasing nutrient concentration to 1/4 strength in 3 weeks. Subsequently, plants were transferred to 2.8 L pots filled with fresh 1/4 strength nutrient solution. After one week, half of the plants were transferred to fresh solution containing 250 μM Pi, while the other half was transferred to nutrient solution deprived of phosphate (P_i), with H_2PO_4 being replaced by H_2SO_4 (Muchhal *et al.*, 1996). Roots from each treatment (0 and 250 μM Pi) were harvested 6, 12, 24 and 48 h after the onset of phosphate starvation and immediately frozen in liquid nitrogen. For each time point, root samples of two plants were pooled.

Plant-N_2-fixing endophytic bacteria association

Plantlets were inoculated as described (James *et al.*, 1994) with 0.1 ml of a 10^6 10^7 cells/mL bacterial suspension. Controls were inoculated with medium only. The endophytic diazotrophic bacteria used were *Gluconacetobacter diazotrophicus* (PAL5 strain) or a mixture of *Herbaspirillum seropedicae* (HRC54 strain) and *H. rubrisubalbicans* (HCC103 strain). All plants were maintained at 30°C with a photon flux density of 60 μE.m^{-2}.s^{-1} for 12 h d^{-1}. One day after the inoculation, plant tissues were examined for bacterial colonization by the Most Probable Number (MPN) estimation (Reis *et al.*, 1994) and plantlets were collected and immediately frozen in liquid nitrogen. Five plantlets were pooled for each treatment.

Herbivory by Diatraea saccharalis

Sugarcane stalk borer larvae were grown on an artificial diet (King and Hartley, 1985) and maintained at 25°C and 60 $\pm$ 10% relative humidity with a 14 h/10 h light/dark cycle. Second instar larvae were maintained under fasting conditions for 18 h prior to transfer. After transferring to plantlets, larvae were observed for a period of two hours to ensure complete boring into the sugarcane stalk. Control plantlets were kept unattacked. After 0.5 and 24 h of exposure to herbivory, plantlets were cut at the stalk/root zone and immediately frozen in liquid nitrogen. For each treatment, two plantlets were used for each time point.

Drought

The plants were transferred to pots containing moist sand, irrigated with Hoagland's solution (Hoagland and Arnon, 1950) and maintained under greenhouse conditions. Regular watering was controlled using a Livingstone atmometer (Broner and Law, 1991) and maintained for 90 days, being suppressed after this period for the experimental group. To control for water loss, soil samples were collected and the humid weight of each soil sample was compared with its dried weight, in order to verify the hydric loss in experimental plants. Aerial parts of the plants were collected 24, 72 and 120 h after the onset of drought for the control and experimental groups. Samples were collected and immediately frozen in liquid nitrogen. For each treatment, aerial parts of six plants were used for each time point.

RNA extraction

Frozen tissues were grinded using a homogenizer. Tissue samples of 2-2.5 g were weighted and grinded to a fine powder, in liquid nitrogen, using a pre-cooled mortar and pestle. The pulverized tissue was transferred to a 50 ml tube and homogenized with 5 ml Trizol (Invitrogen) per gram of tissue according to the manufacturer's instructions. RNA pellets were resuspended in 20 μl of warm diethyl pyrocarbonate-treated water, vortexing gently for about 15 min. RNA samples were quantified in a spectrophotometer and loaded on 1% agarose/formaldehyde gels for quality inspection.

PCR amplification and array printing

cDNA microarray experiments were conducted essentially as reported previously (Papini-Terzi *et al.*, 2005). Sugarcane cDNA plasmid clones obtained from the SUCEST collection were re-arranged and amplified in 100 μl PCR reactions (40 cycles, annealing at 51°C), directly from bacterial clones in culture, using T7 and SP6 primers. PCR products were purified by filtration using 96 well filter plates (Millipore Multiscreen MAFBN0B50). Samples were visualized on 1% agarose gels to inspect PCR-amplification quality and quantity. Purified PCR products (in 10 mM Tris-HCl solution at pH 8.0) were mixed with an equal volume of DMSO in 384 well V-bottom plates. Microarrays were constructed by arraying cDNA fragments on DMSO optimized metal-coated glass slides (type 7, GE Healthcare) using the Generation III Microarray Spotter (Molecular Dynamics). Each cDNA fragment was spotted on the slides at least twice (i.e., technical replicates). Following printing, the slides were allowed to dry and the spotted DNA was bound to the slides by UV-cross linking (50 mJ).

Hybridization and selection of differentially expressed genes

Ten to fifteen micrograms of total RNA were reverse transcribed, labeled, and hybridized using the reagents provided with the CyScribe Post-Labeling kit (GE Healthcare) according to the manufacturer's instructions. The products of the labeling reactions were purified in Millipore Multiscreen filtering plates to remove unincorporated labeled nucleotides. Microarrays were co-hybridized with the fluorescently labeled probes. Hybridizations were performed overnight at 42°C in humid chambers. The slides were then washed in 1x SSC

and 0.2% SDS (10 min, 55°C), twice in 0.1x SSC and 0.2% SDS (10 min, 55°C), and in 0.1 x SSC (1 min, RT). Slides were rinsed briefly in filtered milli-Q water and dried with a nitrogen stream. Each experimental step was carefully monitored to ensure high quality of the slides and extracted data. Slides were scanned using the Generation III Scanner (Molecular Dynamics) adjusting the photomultiplier tube (PMT) to 700 for both channels.

The microarray designed was composed of 1,830 genes selected from the SUCAST Catalogue. Hybridizations were carried out as depicted in Table 3.1. Two biological replicates were used for each microarray experiment. The 1,830 unique genes represented yielded 1,545 good-quality PCR fragments.

Images were processed and data collected using the ArrayVision (Imaging Research Inc.) software. Local median background was subtracted from the MTM (median-based trimmed mean) density for each spot (Papini-Terzi *et al.*, 2005). Data from clones that generated poor-quality PCR fragments (no amplification or unspecific bands) or relative to saturated, low-intensity or poor-quality spots (visually inspected) were excluded.

The fluorescence ratios were visualized and normalized in the MxS space, where M is the base 2 logarithm of the intensities ratio and S is the base 2 logarithm of the average intensity of each spot. The M values were normalized to account for systematic errors using the LOWESS fitting (Yang *et al.*, 2002). The raw and normalized data are publicly available according to the MIAME guidelines at the GEO database under the accession numbers GSE4966 to GSE4971. Differentially expressed genes were defined as the extreme outliers in each experiment, using an intensity-dependent strategy modified from the HT-self method (Vencio and Koide, 2005) and described in (Koide *et al.*, 2006). This method defines an intensity-dependent cutoff curve using the data from each hybridization, detecting non-parametrically genes with the greatest log-ratio changes (outliers) regardless of the absolute value of the log-ratio measurement. We defined as differentially expressed a gene that has at least 60% of their replicate-points above or below the cutoff curve in the two hybridizations of the biological *vs*. controls samples, indicating a reproducible result between the biological replicates. The number of technical replicates ranges from 2 to 16 since genes are spotted several times in the same array. The credibility level used to define outliers was 0.8.

Clustering of expression data using Self Organizing Maps (SOM)

Biologically reproducible expression profiles were clustered with the Self Organizing Maps (SOM) method (Tamayo *et al.*, 1999) using the Spotfire DecisionSite for Functional Genomics software (Spotfire, Somerville, Massachusetts) with default advanced parameters. For each experimental point, the median of the normalized M values among all technical replicates was calculated for each gene represented in the SUCAST microarray. The median values of M in each biological replicate were mean-centered in order to emphasize similarities in the deviations from the mean value by subtracting the average expression level of each gene along the time-course from the experimental measurement obtained in each experimental point. We considered an expression profile as biologically reproducible when the correlation coefficient was ≥ 0.7 between the mean-centered values from pair-wise biological replicate comparisons. SAS with at least one invalid M value (saturated, low-intensity or poor quality signals) were excluded from this analysis. The mean-centered values were averaged between the biological replicates and a Principal Component Analysis (PCA) was performed to estimate the number of groups to be generated by the SOM algorithm (Quackenbush, 2001). The results obtained are shown in additional file 11: Table 9, which demonstrate that the establishment of four to six groups for these data was enough to represent the main sources of variability among the selected patterns. The 2 x 2 and 2 x 3 geometries were tested when generating the SOM results for each of these treatments. We concluded that the 2 x 2 geometry in the case of MeJA treatment and a 2 x 3 geometry in the case of ABA treatment, phosphate starvation and drought resulted in groups with little internal variation.

Validation of microarray results by real-time PCR (RT-PCR)

Two to five micrograms of total RNA (from a third biological replicate for each treatment) were treated with DNase (Invitrogen) according to the manufacturer's instructions and an aliquot of 7.5 μl of the treated RNA was reverse-transcribed using the SuperScript First-Strand Synthesis System for RT-PCR (Invitrogen). The 20 μl reverse transcription reactions contained the RNA template, 2 μl 10x RT buffer, 0.5 mM each dATP, dGTP, dCTP and dTTP, 50 ng random hexamers, 0.25 μg oligo(dT), 5 mM $MgCl_2$, 10 mM DTT (dithiothreitol), 40 U RNase OUT and 50 U *SuperScript II Reverse Transcriptase*. RNA, random hexamers, dNTPs, and oligo(dT) were mixed first, incubated at 70°C for 5 min and placed on ice. The remaining components, except the *SuperScript II Reverse Transcriptase*, were

added to the reaction, the mixture was heated to 25°C for 10 min and then incubated at 42°C for 2 min. The *SuperScript II Reverse Transcriptase* was added to each tube and the reaction was incubated at 42°C for 1.5 h, 72°C for 10 min, and chilled on ice. An identical reaction without the reverse transcriptase was performed as a control (no amplification control, NAC) to confirm the absence of genomic DNA. The cDNA product was treated with 2 U of RNaseH (Invitrogen) for 30 min at 37°C and for 10 min at 72°C. Real-time PCR reactions were performed using *SYBR Green PCR Master Mix* (Applied Biosystems) in a *GeneAmp 5700 Sequence Detection System* (Applied Biosystems). Primers were designed using the *Primer Express 2.0* Software (Applied Biosystems). Real-time PCR reactions were performed in triplicates and contained 2 μl of a 1:10 dilution of the synthesized cDNA, primers to a final concentration of 600 nM each, 12.5 μl of the SYBR Green PCR Master Mix and PCR-grade water to a total volume of 25 μl. In the case of reactions using primers for 25S rRNA, a dilution of 1:1,000 of the synthesized cDNA was used. The parameters for the PCR reaction were 50°C for 2 min, 95°C for 10 min, 40 cycles of 95°C for 15 s and 60°C for 1 min. The specificity of the amplified products was evaluated by the analysis of the dissociation curves generated. No template controls (NTC) and no amplification controls (NAC) were run in order to confirm the absence of genomic DNA or reagent contamination. The relative expression ratio (experimental/control) was determined based on the $2^{-\triangle\triangle Ct}$ method (Livak and Schmittgen, 2001). To access the statistical significance of expression ratios, we assumed a log-normal model and calculated the probability Pr(sample>reference) and Pr(sample < reference) for up- and down-regulated genes, respectively. The expression profile was considered validated when P $\geq$ 0.99. The primers used are shown in additional file 10: Table 8.

Phylogenetic grouping of kinases

Sugarcane sequences containing a putative *pkinase* (catalytic domain of protein kinases) domain, as defined by the Pfam algorithm (Sonnhammer *et al.*, 1998) were selected for the phylogenetic analysis of RLKs and other protein kinases. Protein kinase sequences from other organisms were retrieved in their majority from the PlantsP database (Gribskov *et al.*, 2001) and used as drivers in the phylogenetic analysis. The sequences were aligned using ClustalW (Thompson *et al.*, 1994) with default parameters. The *pkinase* alignment was manually adjusted to preserve the conserved subdomains defined by Hanks and colleagues (Hanks and Hunter, 1995, Hanks and Quinn, 1991, Hanks *et al.*, 1988) for eukaryotic kinases

using the Se-Al Sequence Alignment Editor. The alignment was trimmed to remove gaps in most of the sequences. Sequences spanning less than 50% of this partial alignment were discarded. The alignment was analysed with PAUP using the neighbor-joining (NJ) algorithm and distance as the criterion. Bootstrap analysis was conducted with 500 replicates using NJ/UPGMA and distance as the criterion.

Additional files

Additional file 1: SUCAST SAS showing differential expression in response to the applied treatments. The table lists all SAS whose expression was enriched or decreased in both biological samples for each experimental point. The plus sign indicates that the SAS expression is up-regulated, the minus sign indicates that the gene expression is down-regulated. The asterisk indicates that the SAS identity was not confirmed by re-sequencing. The table also shows the expression profile of these genes in six sugarcane tissues (**revised from (Papini-Terzi *et al.*, 2005)). The last four columns indicate the SOM groups in which these SAS were included (numbered according to Figure 3.1). The expression data for sugarcane tissues were inferred from microarray hybridizations of tissue samples against a common reference constituted by an equimolar mixture of the tissues sampled (Papini-Terzi *et al.*, 2005). FL = flowers, LB = lateral buds, LV = leaves, RT = roots, IN1 = first internodes, IN4 = fourth internodes, PD = phosphate deficiency, H = inoculation with *Herbaspirillum* spp., GD = inoculation with *Gluconacetobacter diazotrophicus*.

Additional file 2: Log_2 ratios of microarray signals between the Cy3 and Cy5 channels. For each clone, the median intensity value of the technical replicates was used to calculate the ratio between the experimental samples and the reference sample. The numbers 1 and 2 denote the different biological samples used for each experiment. The asterisk indicates that the SAS identity was not confirmed by re-sequencing. Only valid values are shown, excluding data from saturated, low-intensity and low-quality spots. Data from clones for which PCR reactions produced low yields or multiple bands were also removed. The table contains 1,555 elements and not 1,545. This is due to the fact that some SAS are represented more than once in the Table. This occurs when the same SAS is represented in more than one position in the 384-well plates and some of these positions did not have their identity validated by re-sequencing. In these cases, we opted to analyse these data separately.

Additional file 3: SUCAST protein kinases showing differential expression in

response to the applied conditions. For SAS included in the SOM clustering analysis, the group to which the SAS belongs is indicated. The plus sign indicates that the SAS expression is up-regulated, the minus sign indicates that the gene expression is down-regulated. The asterisk indicates that the SAS identity was not confirmed by re-sequencing. The last four columns indicate the SOM groups in which these SAS were included (numbered according to Figure 3.1). PD = phosphate deficiency, H = inoculation with *Herbaspirillum* spp., GD = inoculation with *Gluconacetobacter diazotrophicus*.

Additional file 4: Components of the SOM groups generated for ABA treatment. The table indexes all genes that were included in the clustering analysis (correlation coefficient ≥ 0.7 between the expression profiles obtained for the two biological replicates of a particular experiment) and their respective groups. The asterisk indicates that the SAS identity was not confirmed by re-sequencing.

Additional file 5: Components of the SOM groups generated for MeJA treatment. The table indexes all genes that were included in the clustering analysis (correlation coefficient ≥ 0.7 between the expression profiles obtained for the two biological replicates of a particular experiment) and their respective groups. The asterisk indicates that the SAS identity was not confirmed by re-sequencing.

Additional file 6: Components of the SOM groups generated for phosphate deficiency treatment. The table indexes all genes that were included in the clustering analysis (correlation coefficient ≥ 0.7 between the expression profiles obtained for the two biological replicates of a particular experiment) and their respective groups. The asterisk indicates that the SAS identity was not confirmed by re-sequencing.

Additional file 7: Components of the SOM groups generated for drought treatment. The table indexes all genes that were included in the clustering analysis (correlation coefficient ≥ 0.7 between the expression profiles obtained for the two biological replicates of a particular experiment) and their respective groups. The asterisk indicates that the SAS identity was not confirmed by re-sequencing. ** The two different results obtained for this SAS are related to different positions in the 384-well plates. In both positions the SAS identity was not validated by re-sequencing. Because of this, these positions were treated separately.

Additional file 8: Neighbor-joining (NJ) tree of kinase domains from sugarcane protein kinases. Selected kinase domains were aligned to construct a distance tree with the NJ algorithm. Bootstrap values greater than 50% (500 replicates) are shown for nodes in the tree. The RLKs/RLCKs group is highlighted in gray and only some representatives of this

group are included. The undefined kinases are highlighted in red. Drivers are in italic. The SAS name in the figure is preceded by a prefix based on the BLAST similarity searches as indicated in the Annotation Box. The branch color indicates the presence of additional Pfam (Sonnhammer *et al.*, 1998) domain besides the pkinase domain as indicated in the Domains Box.

Additional file 9: Neighbor-joining (NJ) tree of kinase domains of sugarcane RLKs and RLCKs. Selected kinase domains from sugarcane RLKs and RLCKs and from other plants (drivers) were aligned to construct a distance tree with the NJ algorithm. Bootstrap values greater than 50% (500 replicates) are shown for nodes in the tree. Drivers are in italic. Undefined RLKs and RLCKs are highlighted in red. The SAS name in the figure is preceded by a prefix based on the BLAST similarity searches as indicated in the Annotation Box. The branch color indicates the presence of additional Pfam (Sonnhammer *et al.*, 1998) and SMART (Schultz *et al.*, 1998) domains besides the pkinase domain as indicated in the Domains Box.

Additional file 10. Sugarcane genes and primers selected for real-time PCR experiments. Primers were designed using the Primer Express 2.0 Software (Applied Biosystems) and BLAST searches against the SUCEST database were conducted to ensure the specificity of the selected primers. The column Experimental Datapoints lists the samples analysed. Primer sequences for endogenous references were retrieved from (Papini-Terzi *et al.*, 2005) and (Iskandar *et al.*, 2004). The *P value* is the probability Pr(sample>reference) and Pr(sample < reference) for up- and down-regulated genes, respectively. The expression profile was considered validated when $P \geq 0.99$.

Additional file 11: PCA analysis for defining the number of groups in SOM. The difference between the magnitude of principal component eigenvalues for each treatment (columns) estimates how many groups are minimally necessary in each SOM grouping analysis. When the difference is near zero, there is no substantial information added by increasing the number of seed groups in SOM.

Author's contributions

GMS as the General Coordinator of the SUCEST Project, conceived, advised and coordinated most aspects of the study. GMS and FRR wrote the manuscript. FRR and FSPT produced the arrays, designed the experiments, performed the hybridizations and analysed the

data. FRR also performed the real-time PCR experiments and the SOM clustering analysis. MYN and RZNV worked on the data statistical analysis and differential gene expression determination. FRR, FSPT and GSM annotated and catalogued the genes. GSM conceived the SUCEST-FUN, SUCAST and SUCAMET Databases. RV constructed the SUCEST-FUN database and MYN helped with computational issues. The following researchers were responsible for conducting plant experimentation as follows: phosphate starvation (RDCD, RSA, AVOF); sugarcane inoculation with endophytic bacteria (FV, ASH); herbivory attack (CB, AHM, MCSF); drought experiments (FAR, JAG, SMZ); phytohormone treatments (VER, MM). ECS was responsible for sugarcane cultivation at CTC, Piracicaba. All authors contributed, read, corrected and approved the final manuscript. The authors declare no conflict of interest in this work.

Acknowledgements

This work was funded by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), Centro de Tecnologia Canavieira and Central de Álcool Lucélia Ltda. We are indebted to Sarah D. B. Cavalcanti, Erica Bandeira, Adriana Y. Matsukuma and Denise Yamamoto for technical assistance on microarray experiments performed in the laboratory of the Cooperation for Analysis of Gene Expression (CAGE) inter-departmental Project and Dr. Enrico Arrigoni for providing us with the *D. saccharalis* larvae.

FRR, FSPT, JMF, RSA were supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) fellowships. RV was supported by a fellowship from the UNIEMP Institute. FV and ASH were supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) fellowships. RDD was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). AVOF, MCSF and MM are recipients of research productivity fellowships from CNPq.

O conhecimento humano havia se tornado incontrolavelmente vasto;
cada ciência havia gerado dezenas de outras mais,
a filosofia em si ... encontrou sua tarefa de coordenação
excessivamente estupenda para a sua coragem, fugiu de todas
essas frentes de batalha pela verdade ...O conhecimento humano
havia se tornado excessivamente grande para a mente humana.

William Durant (1885-1981)

4

Expression profile of signal transduction components in a sugarcane population segregated for sugar content‡

Juliana de Maria Felix[1,2,*], Flávia Stal Papini-Terzi[3,*], Flávia Riso Rocha[3], Ricardo Zorzetto Nicoliello Vêncio[3], Renato Vicentini[1,2], Milton Nishiyama-Jr[3], Eugênio César Ulian[4], Gláucia Mendes Souza[3], Marcelo Menossi[1,2]

Abstract

Sucrose is the major product of photosynthesis in many higher plants. It is transported from the source tissue (mature leaves) through the phloem to various sink tissues to support plant growth, development and reproduction. Knowledge on the signal transduction pathways involved in sucrose synthesis in mature leaves is limited. The tropical crop sugarcane (*Saccharum* spp.) contributes to about two thirds of the world's raw sugar production. Using a microarray approach, we analyzed the expression profiles of 1920 sugarcane genes encoding signal

‡submitted

*The first two authors contributed equally to this work.

[1]Functional Genomics Laboratory, Center for Molecular Biology and Genetic Engineering, State University of Campinas, P.O.Box 6010, 13083-875, Campinas, SP, Brazil

[2]Department of Genetics and Evolution, Institute of Biology, State University of Campinas, P.O.Box 6010, 13083-875, Campinas, SP, Brazil

[3]Department of Biochemistry, Institute of Chemistry, University of São Paulo, 05508-900, São Paulo, SP, Brazil

[4]Cane Technology Center - CTC, 13400-970, Piracicaba, SP, Brazil

transduction elements, transcription factors and stress-related proteins. We used individuals from a population segregating for sugar content and gene expression profiles were obtained from seven individuals with highest and the seven with lowest sugar content. Twenty-four genes were differentially expressed. Five of them were more expressed in the bulk of plants with highest sugar content and 19 in the bulk of plants with lowest sugar content. RNA-blot hybridization was used to validate the expression profiles obtained in individuals of each bulk. These genes had differential expression along the growing season and in different tissues. Our data supports a view that sugar levels modulate a complex signal transduction network that seems to envolve responses that are related to stress.

Introduction

Plants synthesize carbohydrates in leaves by photosynthetically fixing atmospheric CO_2. In C_4 plants, like sugarcane, maize and sorghum, the CO_2 fixation occurs in two photosynthetic cell types: mesophyll cells and bundle sheath cells. Mesophyll cells carry out the initial steps of CO_2 fixation via the enzyme phosphenolpyruvate (PEP) carboxylase to produce the four-carbon organic acid oxaloacetate. In the bundle sheath cells, the C_4 acid is decarboxylated and Rubisco refixes the resulting CO_2 in the photosynthetic carbon reduction (PCR) cycle (reviewed by Lunn & Furbank, 1999).

Sucrose is the major form in which carbohydrate is translocated from leaves to the rest of the plant, to supply carbon and energy for growth and the accumulation of storage reserves. After synthesized, it can be either stored temporarily in the vacuole or transported over long distance in solution in the phloem sap. Photosynthetically tissues, like mature leaves, are net exporters of sugars and are known as 'carbon sources' or source tissues. Heterotrophic cells in roots, reproductive structures, storage and developing organs rely on a supply of sugars for their nutrition; these are known as 'carbon sinks' or sink tissues (net importers). Sucrose itself is the major storage reserve in some plants, for example in sugarcane (*Saccharum* spp.) stems, sugarbeet (*Beta vulgaris*) roots and the fruits of many species.

There is a growing interest in the tropical crop sugarcane because of ethanol and biomass are important renewable biofuel sources. Moreover it is of great economic interest, contributing to about two thirds of the world's raw sugar production (Pessoa-Jr *et al.*, 2005). Due to its unique capacity of storing sucrose in the stems, sugarcane is an interesting model for studies on sugar synthesis, transport and accumulation. Sucrose metabolism components

and regulators are likely to be key players in determining sugarcane sucrose yield (Lunn and Furbank, 1999, Moore, 2005).

Sugarcane is a complex polyploid grass with commercial varieties derived from conventional breeding. Recent yield data indicate that such technology may be reaching its limit with respect to increases in sugar productivity. It would be highly advantageous to have genes associated with desirable traits targeted for directed improvement of the varieties. A useful strategy for target-gene identification has been denominated "genetical genomics". First introduced by Jansen and Nap (2001), this method aims to apply large-scale analysis of gene expression to a segregated population. The use of cDNA microarrays to evaluate an F_1 sugarcane population that segregates for a certain trait may provide more insight into plant signaling and gene function than classical mutagenesis studies (Meyers *et al.*, 2004). Recently, Casu *et al.* (2005) used this strategy to identify genes associated with high sucrose accumulation in sugarcane stem. The genomics approach has been the method of choice in the search for coarse regulatory mechanisms of sugarcane sucrose accumulation (Carson and Botha, 2002,Carson *et al.*, 2002,Casu *et al.*, 2003,Casu *et al.*, 2004,Casu *et al.*, 2005). However, the studies on the sugarcane transcriptome have focused primarily on the sugarcane stem during vegetative growth, i.e., on internodes actively accumulating sucrose. In sugarcane there are no studies indicating genes potentially involved in the sugar signaling that regulates sucrose synthesis in the leaves.

A comprehensive sugarcane EST (Expressed Sequence Tags) data collection was made available by the SUCEST Consortium in 2003 (Vettore *et al.*, 2003) (http://sucest.lad.ic.unicamp.br/public) and functional characterization of molecular components is underway (http://www.sucest-fun.org). A total of 43,141 Sugarcane Assembled Sequences (SAS) representing the putative transcripts from sugarcane have been found. A subset of 902 transcripts related to elements of signaling cascades, transcription factors and stress-related transcripts, in particular, plus 378 transcripts encoding proteins with unknown function are the focus of this work and have been described previously (Souza *et al.*, 2001, Papini-Terzi *et al.*, 2005, Rocha *et al.*, 2007).

In addition to being an important carbon reserve in different organs, such as stems, tubers and fruits, sucrose also helps to protect plants from environmental stresses as, for example, cold and drought (Smeekens, 2000). The accumulation of sucrose and other low- molecular-mass compounds under stress conditions is often regarded as an adaptive mechanism to maintain cell turgor and to protect the structure and function of proteins and

membranes.

Moreover, it has been recognized that sucrose also acts as a signal compound, affecting a variety of physiological processes, such as photosynthesis, source and sink metabolism and defense responses (Smeekens, 2000, Rolland *et al.*, 2002, Koch, 2004, Gibson, 2005, Osuna *et al.*, 2007). Metabolism control involves the coordinated regulation of genes and enzymes at the level of transcription, translation, post-translational modification and protein turnover. The carbon metabolite signaling pathways cross-talk with other pathways, including hormonal responses, cell cycle control and nitrogen response systems, amongst others (Halford and Paul, 2003). Whereas the effect of sugars on gene regulation is well established, the nature of the signals and the molecular mechanisms involved in sugar perception and intracellular signal transmission are largely unknown. Therefore, understanding sucrose synthesis in sugarcane at the transcriptional level, and finding the genes coding for proteins associated with sugar accumulation would be of great value for the long-term success of varietal improvement.

In this study we used "genetical genomics"for the identification of genes whose differential expression levels correlated with high or low sugar contents in a segregating sugarcane population. Microarrays containing 1920 signal transduction-related ESTs as well as transcription factors and stress-related elements were used to measure relative gene expression. A total of 24 SAS (Sugarcane Assembled Sequences) were defined as differentially expressed. Five of them were found to be differentially expressed in a high sugar content pool of plants and the other 19 in a low sugar content pool.

Materials and methods

SUCAST Catalogue Annotation

The Sugarcane Assembled Sequences (SAS) can be found in the SUCEST database (http://sucest.lad.dcc.unicamp.br/public) and the sugarcane ESTs sequences can be found in the GenBank under accession numbers CA064599-CA301538. The subset of the SAS discussed in this work and the corresponding ESTs can be found in Table S-II. Members of the SUCAST catalogue were identified using the BLAST algorithm (Altschul *et al.*, 1997) with conserved protein sequences as drivers, as described previously (Papini-Terzi *et al.*, 2005, Rocha *et al.*, 2007).

cDNA Microarrays

Microarrays were constructed by arraying 1920 PCR-amplified cDNA fragments on derivative glass slides as described by Papini-Terzi *et al.* (2005). Four replicates of each cDNA fragment were distributed across each array. Fragments for which the amplification reactions were not satisfactory or hybridization signals were low were removed from the analysis. High quality data was obtained for a total of 1280 SAS, all of which had their identity confirmed by re-sequencing.

Sugarcane Tissue Samples

Sugarcane F_1 plants were obtained from a cross between pre-commercial cultivars (SP80-180 X SP80-4966). The population is comprised of 498 individuals that segregated for stem sugar content in a normal manner and was previously described by Garcia *et al.* (Garcia *et al.*, 2006). The seven plants presenting extreme values for high sugar (HS) and low sugar (LS) were selected. Mature leaves (Leaf +1), according to Van Dillewijn (1952), were collected from the selected plants 6, 7, 9, 11 and 13 months after planting. For the microarray analyses, RNA extracted from leaves collected at the 9 months time point from the seven individuals of each group were pooled. The expression profiles observed in the microarrays were further validated by RNA blot using RNA from three HS and three LS individuals collected at the 9 months time point. Pooled RNA from the seven HS and LS individuals collected at all five time points were also used in RNA blots to detect the expression profiles along the growing season. The expression profile of selected genes was also evaluated for six different tissues collected from 12 month old plants: mature leaf, immature leaf, immature internode, root, lateral bud and a mixture of flowers in different developmental stages, using the same commercial sugarcane varieties used in the SUCEST project (SP87-432 for flowers and SP80-3280 for other tissues). All tissue samples were stored at -80°C.

RNA extraction

Leaf tissue (2 - 2.5g) was ground to a fine powder in liquid nitrogen, using pre-cooled mortar and pestle. RNA was isolated using the Trizol^{TM}reagent (Invitrogen, USA), following the recommended procedure. The RNA samples were quantified in a spectrophotometer and loaded onto 1.0% agarose/formaldehyde gels for a quality inspection. RNA sam-

ples of five sugarcane tissues (flower, leaf, stem, root and bud) were also prepared and equimolarly mixed, to be used in homotypic (self-self) hybridizations. The Trizol™ manufacturer's recommendations for high polysaccharide content tissues were followed for the mature internode samples.

Probe Preparation and Hybridization

Two microarray hybridizations (Lv1 and Lv2) were performed comparing one pooled sample from seven plants with high sugar content (HS) to another pool from seven plants with low sugar content (LS), in a dye-swap layout. RNA samples for Lv1 and Lv2 hybridizations derived from independent extractions from the same pools of plants.

To this end, ten micrograms of total RNA were reverse transcribed using oligo dT primers and labeled using the CyScribe Post-Labeling kit (Amersham Biosciences), according to the manufacturer's instructions. The products of the labeling reactions were purified in filtering plates (Millipore Multiscreen MAFBN0B50) to remove unincorporated labeled nucleotides. The microarrays were co-hybridized with the fluorescently labeled probes. Hybridizations were performed overnight at 42°C in moist chambers. The slides were then washed in 1x SSC and 0.2% SDS (10 min, 55°C), twice in 0.1x SSC and 0.2% SDS (10 min, 55°C) and finally in 0.1 x SSC (1 min, RT). The slides were rinsed briefly in filtered milli-Q water and dried in a nitrogen stream. Each experimental step was carefully monitored to ensure high quality of the slides and the extracted data.

Data extraction and processing

The slides were scanned using a Generation III Scanner™ (Molecular Dynamics, USA) and processed using the ArrayVision (Imaging Research Inc., Canada) software. Low-quality spots were filtered. Signal intensities were calculated for each valid spot subtracting the local median background from the MTM (median-based trimmed mean) density.

The raw fluorescence intensity values were then processed using custom programs on R language, available at http://verjo19.iq.usp.br/xylella/microarray. Firstly, intensity ratios (HS/LS) were calculated for each spot. Then, each slide dataset was normalized using the Lowess fitting (Yang *et al.*, 2002), in order to correct for systematic experimental errors such as labeling-bias and intensity dependent variation. To be able to classify a gene as

differentially expressed, a set of experimental and computational steps was established, using a local implementation of the HTself method (Vencio and Koide, 2005), as follows:

1. Homotypic or "self-self"hybridizations were performed using a tissue-pool sample in both channels (Cy3 and Cy5) to assess experimental "noise", i.e., the intrinsic technical variation of the experimental pipeline;

2. The fluctuation of the normalized ratios obtained from these homotypic hybridizations was computed in an intensity-dependent manner, integrating the probability density function to 98% for eight different signal intensity intervals. Thus, a ratio cut-off curve that determines the limits of the random variation for our data could be outlined;

3. The replicate ratio values obtained for each gene were independently compared to the cut-off limits and classified as up (above the cut-off limit), down (below the cut-off) or inside (no differential expression). Genes with at least 75% of the replicate points above or below (up or down) the cut-off limits were considered differentially expressed.

The ratios obtained for each transcript in our chip can be found in the supplemental material (Table S-I). Descriptions followed the MIAME guidelines and the data was deposited on Gene Expression Omnibus database (GEO - http://www.ncbi.nlm.nih.gov/geo/) under the accession numbers GSE4233 (series), GPL1376 (platform), GSM95526, GSM95546, GSM95547 and GSM95548 (samples).

Validation of microarray results by RNA blot

Electrophoresis of total RNA samples (10μg) was carried out on 1.5% formaldehyde-containing agarose gels by standard procedures and transferred to a nylon filter (Hybond-N^+, Amershan Biosciences). For each gene tested, the longest EST clone of each SUCEST SAS was selected as a probe for RNA blot hybridization. Inserts were labeled with the Read-To-Go kit (Amershan Biosciences) according to the protocol recommended by the manufacturer. Hybridized filters were exposed to imaging plates for 24 h and the digitized images of RNA blot hybridization signals detected using the FLA3000-G screen system (Fuji Photo Film, Japan) and quantified using the Image Gauge software v. 3.12 (Fuji Photo Film, Japan).

Results

Sugar content in a field-grown F_1 segregant progeny

In order to assess differences in gene expression associated with sugar content, individuals from sugarcane progeny contrasting for sucrose content were chosen for the analysis. The plant material used was a field-grown F_1 progeny selected from a cross between the sugarcane varieties SP 80-180 and SP 80-4966. The parental are divergent for sucrose content and differ by 3.07 points in their Brix content (data not shown). From a total of 498 individuals, seven plants with the highest (7HS) and seven with the lowest (7LS) sugar contents were picked out. Figure 4.1 shows the average values and standard deviations for the soluble solids content (Brix) of the most mature internode of these group of plants measured throughout the growing season (6, 7, 9, 11 and 13 months after planting).

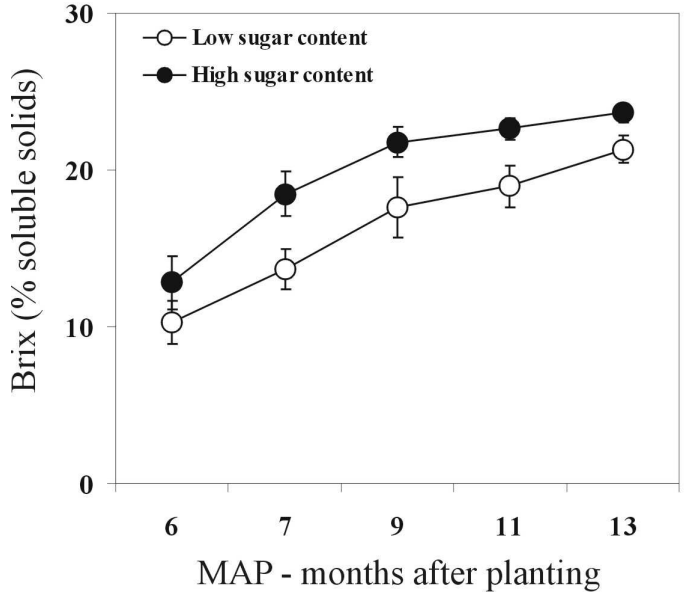


Figura 4.1: **Sugar content throughout the growing season in the extreme individuals of a sugarcane segregated population.** The Brix (soluble solids) values of the most mature internodes of each sugarcane segregant plant were measured throughout the growing season. Average Brix values and standard deviations for the seven individuals with the highest or lowest sugar contents are shown for the times indicated.

Differential gene expression in mature sugarcane leaves

A sugarcane cDNA microarray previously described by Papini-Terzi *et al.* (2005) was used to assess the gene expression of the mature leaves. 1920 ESTs involved in sugarcane signal transduction were represented on the glass slides. Target RNA was isolated from the mature leaf tissue of the 7HS and 7LS pools, reverse-transcribed into complementary DNA, fluorescently labeled, and competitively hybridized onto replicate slides. Leaves were collected nine months after planting since at this age the great difference in sugar content was observed between the two segregant samples (Figure 4.1). Twenty-four ESTs were differentially expressed in the two groups. The putative biological functions associated with these ESTs are shown in Table 4.1, and the accession numbers of the SAS (Sugarcane Assembled Sequences) as well as the cDNA sequences that compose them are available in the supplemental material (Table S-II). Five transcripts were enriched in the mature leaves of the higher sugar content plants. These encoded an omega-3 fatty acid desaturase (FAD8), two sequences with no hits in the public databases ('no match'), a putative receptor-like serine/threonine kinase (RLK1) and a Myb domain transcription factor LHY/CCA1. Nineteen transcripts were enriched in the mature leaves of the lower sugar content plants. These encoded three 14-3-3 like proteins, two proteins of the inositol metabolism (O-methyltransferase and 1,4,5-trisphosphate phosphatase), a SNF1-related protein (SnRK1), a putative protein with an unknown function, eight stress-related proteins, two transcription factors, a F-box TIR-1 and one putative protein with no match in the GenBank database. It is interesting to note that these genes encoded cellular components of various functional categories, including signaling (RLK1, SnRK, 14-3-3), transcription (tubby, DP transcription factor) and stress responses (drought and cold response, wound induced protein, dehydrin, tonoplast intrinsic protein). This indicates that the modulation of sucrose content relies on several metabolic processes, including the perception of stress signals and the regulation of gene expression.

Gene expression validation by RNA-blot and analysis in plants throughout the growing season

Three genes with greater expression in the higher sugar content plants (*SCSBAM1085B06.g*, *SCACLR1126E09.g* and *SCCCLR1C08G10.g*) and three with increased expression in the lower sugar content plants (*SCSBST3096H04.g*, *SCEQLB2019B08.g* and *SCQ-*

Tabela 4.1: **Sugarcane genes showing differential expression between high and low sugar content populations.** [a] Sugarcane Assembled Sequence; [b] The description indicates the putative function of the gene products expected from the similarity sequences by searches using the BlastX algorithm and the corresponding SUCAST category; [c] The accession number of the homologue in the NCBI public database; [d] E value; [e,f] Fold increase in expression observed for these ESTs in a high ([e]) or low ([f]) sugar content plants. Asterisks represent ESTs that were validated by RNA-blots.

Category	SAS[a]	Description of homologue[b]	Accession number of homologue[c]	E value[d]	High[e]	Low[f]
Enriched expression in the high sugar content population						
Hormone biosynthesis	SCSBAM1085B06.g	Omega-3 fatty acid desaturase-FAD8	T01696	1e-104	1.88 *	
No matches	SCACLR1126E09.g	No matches			1.89 *	
No matches	SCBFSD1035H11.g	No matches			1.93	
Receptors	SCEQRZ3020C02.g	RLK undefined with LRR- unclassified	CAB51480	1e-113	1.64	
Transcription	SCCCLR1C08G10.g	LHY/CAA1	XP_480189.1	9e-69	1.79 *	
Enriched expression in the low sugar content population						
Adapters	SCCCRZ1001D02.g	14-3-3 proteins	AAP48904	7e-140		1.81
Adapters	SCEQRT1025D06.g	14-3-3 proteins	BAB11739	2e-80		2.26
Adapters	SCEQRT1031D02.g	14-3-3 proteins	AAP48904	1e-117		1.90
Inositol	SCRFLR1012F12.g	Caffeic acid 3-O-methyltransferase	AAQ67347	0.0		1.61
Inositol	SCSBST3096H04.g	Inositol-1,4,5-trisphosphate phosphatase	XP_475767	8e-62		3.07 *
Protein kinases	SCEQLB2019B08.g	SNF1-related	CAA73067	2e-73		2.28 *
Putative protein	SCCCLR2002H11.g	Putative CGI-94 protein	BAD68235	8e-96		1.69
Stress	SCJFLR1074E09.g	Dehydrin	AAA33480	6e-48		1.81 *
Stress	SCEPRZ3087C08.g	Low temperature induced (LTI)	AAT37942	6e-24		2.40
Stress	SCUTST3084F06.g	Low temperature induced (LTI)	AAV88601	7e-18		2.65
Stress	SCBFFL5074C09.g	Reversibly glycosylated polypeptide	XP_479089	0.0		1.63
Stress	SCCCLR1024C03.g	Tonoplast intrinsic protein	AAC09245	6e-102		1.90
Stress	SCQGLR1085F11.g	Dehydrin	AAB05927	7e-20		2.57
Stress	SCJLRT1016G06.g	Ribonuclease	AAS01727	1e-106		2.38
Stress	SCCCLR2C01F06.g	Wound-induced	CAA42537	2e-17		2.03
Transcription	SCCCLB1002B01.g	DP transcription factor	AAO72671	5e-109		1.86
Transcription	SCCCCL4001D08.g	Tubby-like protein 7	AAM18187	1e-71		1.94
Ubiquitination	SCCCCL6003D08.g	F-box containing protein TIR1-like	XP_467902	1e-173		1.98
Unknown protein	SCSFFL8048D12.g	Unknown protein	XP_467976	9e-36		3.62

GLR1085F11.g) were analyzed by RNA-blots. Total RNA from each of three sugarcane individuals was used to provide replication for the gene expression profiles observed in the microarray hybridization. Figure 4.2 shows that the microarray data was confirmed in at least two of the three different sugarcane plants collected nine months after planting, indicating high consistency between the two data sets.

To identify the gene expression trends throughout the growing season, the mRNA levels for the same six genes were determined in the 7HS and 7LS pools collected 6, 7, 9, 11 and 13 months after planting (Figure 4.3). The inset graph represents the expression profile of each gene plotted for each group. The three genes found to be enriched in the higher sugar content plants were consistently differentially expressed throughout the growing season (Figure 4.3 a-c). These genes are possibly involved in the control of sucrose synthesis, accounting for the higher sugar content in these segregant plants. The genes with more transcripts in the lower sugar content plants showed a less consistent pattern (Figure 4.3 d-f). All of them were differentially expressed in the plants at nine months after planting, confirming the expression observed by microarrays, but only the one encoding dehydrin, a stress-related protein (Figure 4.3 f) had a more consistent pattern throughout the growing season.

Finally, the spatial profile of these ESTs was analyzed, comparing their expression in the source (mature leaf) and sink (immature leaf, immature internode, root, lateral bud and flower) tissues of a commercial sugarcane variety (Figure 4.4). *SCSBAM1085B06.g* mRNA accumulated at high levels in immature leaves and immature internodes, at a lower level in mature leaves and at very low levels in roots, lateral buds and flowers. *SCACLR1126E09.g* and *SCSBST3096H04.g* presented similar patterns, with preferential expressions in mature leaves and no expression, or a very weak signal, in the other tissues analyzed. *SCCCLR1C08G10.g* was expressed in all tissues analyzed, but it's mRNA accumulated at high levels in mature and immature leaves, lateral buds and flowers. *SCEQLB2019B08.g* and *SCQGLR1085F11.g* showed similar behavior, their expression levels being highest in immature leaves, immature internodes, lateral buds and flowers.

Discussion

Gene regulation is based on sensing different signals or stimuli, which are transmitted through a signaling pathway, finally leading to an increase or decrease in transcription. In sugar signaling, the first step is to sense the nature and level of the specific sugar. While

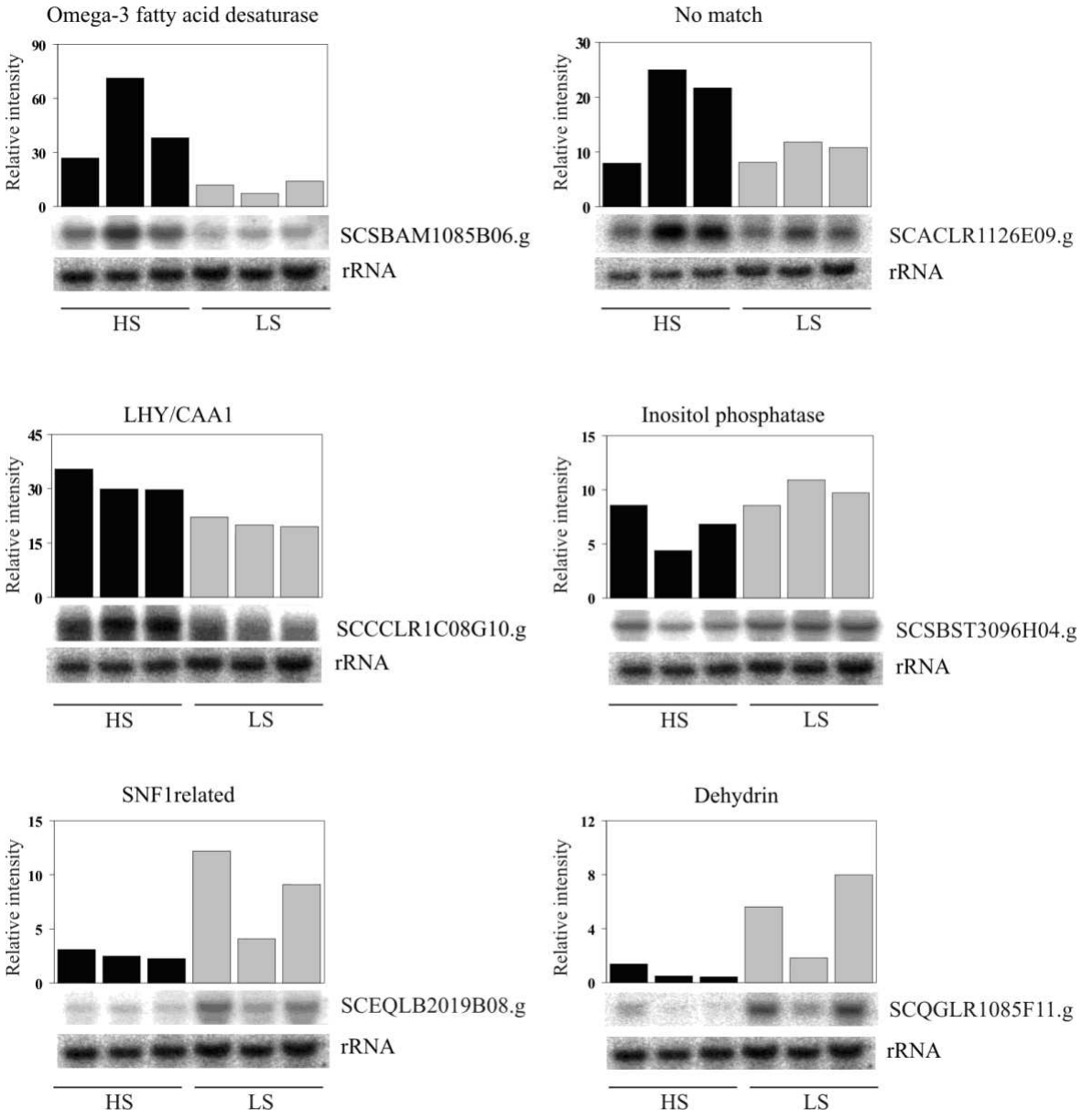


Figura 4.2: **Expression levels of differentially expressed genes in sugarcane individuals.** RNA blots were prepared using 10 μg of total RNA isolated from mature leaves of three individual clones of each segregated plants (HS - high and LS - low sugar contents). The time point evaluated in the blots corresponds to the same one used in the cDNA microarray experiments (9 months after planting). Blots were hybridized with the gene-specific radioactive probes indicated. An rDNA fragment was used as the control.

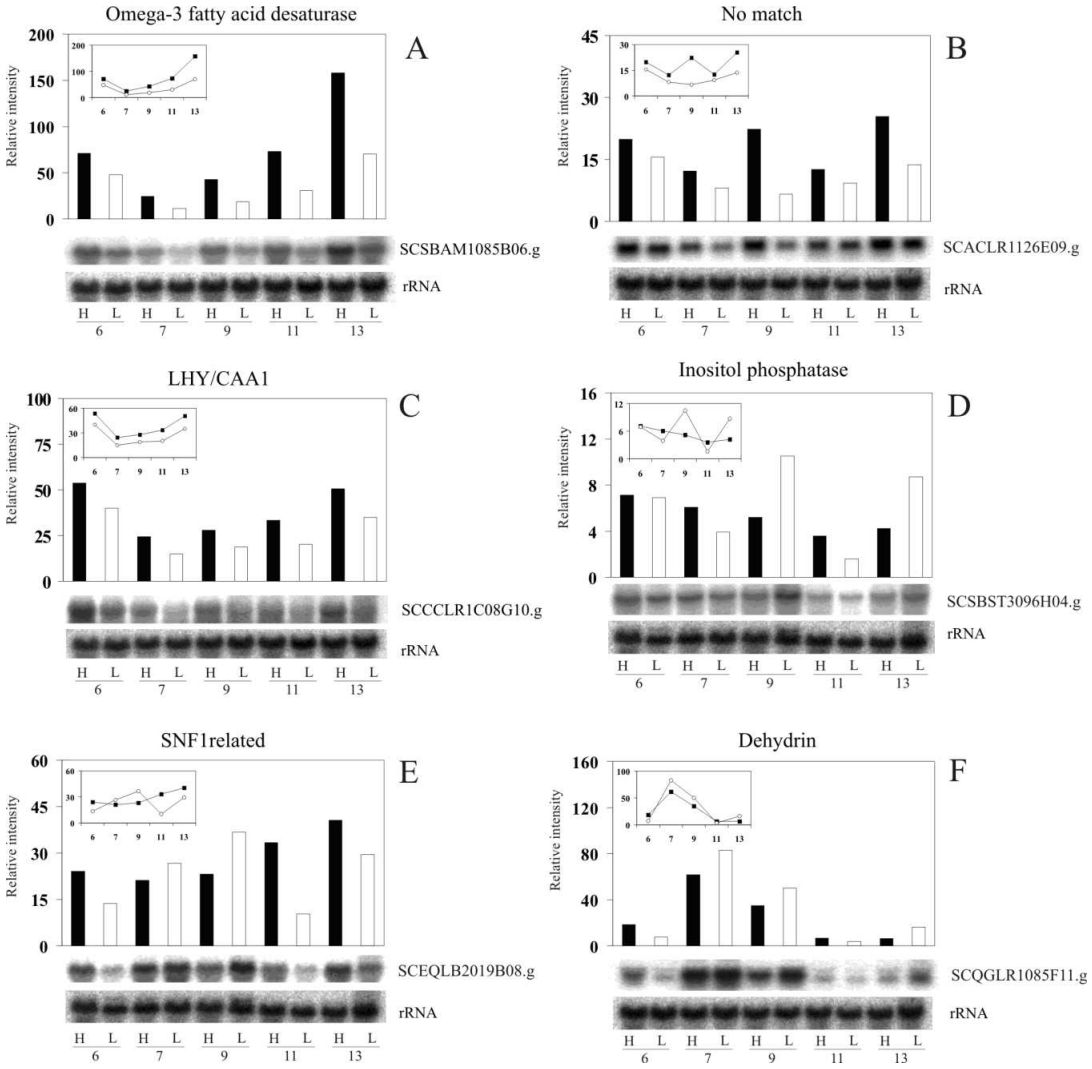


Figura 4.3: **Expression profiles of differentially expressed genes throughout the growing season.** RNA-blots were prepared from total leaf-RNA from a pool of 7 individuals with high (HS) and low (LS) sugar contents collected throughout the growing season (6, 7, 9, 11 and 13 months after planting). The inset graphs show the expression levels observed for the high (black circles) and low (white circles) sugar content plants. An rDNA fragment was used as the control.

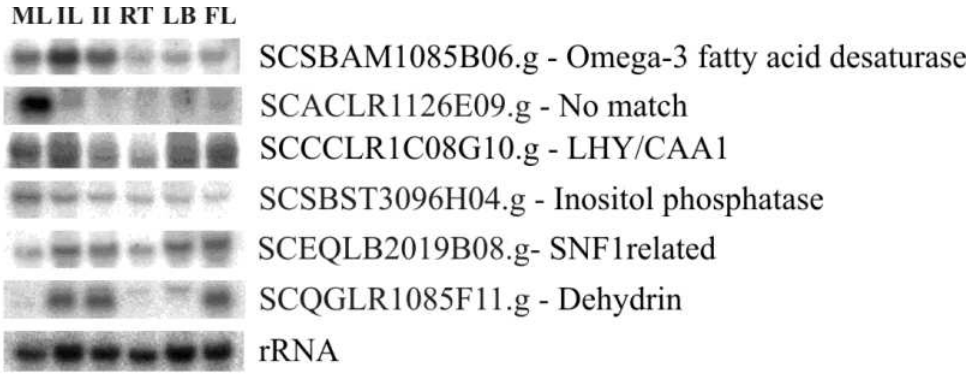


Figura 4.4: **Gene expression analysis in different tissues.** For the RNA gel blot preparation, each lane was loaded with 10 μg of total RNA isolated from one of six tissues from sugarcane. ML - mature leaves; IL - immature leaves; II - immature internode; RT - root; LB - lateral bud; FL - flowers. The same blot was hybridized to the indicated cDNA probes. An rDNA fragment was used as the control.

elevated cellular levels of sugar up-regulate genes involved in the synthesis of polysaccharides, storage proteins and pigments, as well as in genes associated with defense responses and respiration, sugar deprivation enhances the expression of genes involved in photosynthesis and resource remobilization, such as the degradation of starch, lipid, and protein (Koch, 1996, Yu, 1999, Ho *et al.*, 2001). Although the regulatory effect of sugars on photosynthetic activity and plant metabolism has long been recognized, the concept of sugars as central signaling molecules is relatively new (reviewed by Rolland *et al.* 2006).

Genome-wide expression analysis using cDNA microarray has been applied to the discovery of new insights into the mechanisms by which sugar-response pathways interact with other pathways. Price *et al.* (2004) used this approach to determine the effect of glucose and inorganic nitrogen on gene expression on a global scale in *Arabidopsis thaliana*. Glucose regulated a broad range of genes, including genes associated with carbohydrate metabolism, signal transduction and metabolite transport. In addition, a large number of stress responsive genes were also induced by glucose, indicating a role for sugar in environmental responses. Similar results were obtained using rice (*Oryza sativa*) cell cultures, where the transcription rate and mRNA stability were shown to be affected by sugars (Ho *et al.*, 2001), illustrating a diverse role of sugar in gene regulation. In a microarray study measuring the effects of sucrose and light on 8000 unique *Arabidopsis* targets revealed that genes associated with metabolism, protein synthesis/modification and energy were over represented when

compared to genes unaffected by the treatments (Thum *et al.*, 2004). In a recent study using ATH1 arrays, Osuna *et al.* (2007) identified many genes related to signal transduction like receptor kinases, soluble protein kinases and phosphatases, MAP kinase pathway components, calcium binding proteins and G-proteins that presented alterations of their transcript levels in C-starved seedlings, which are reversed by sucrose addition.

Here we report a microarray analysis of 1920 sugarcane genes encoding signal transduction elements, transcription factors and stress-related proteins. The expression profile of these genes in mature leaves of a sugarcane population segregating for sugar content was analyzed and putative targets for molecular-assisted varietal improvement identified. Possible roles of these genes in sugar signal transduction and sugar stress (low sugar) tolerance as well as sugar metabolism are discussed.

Cross-talk between hormone biosynthesis and sugar signaling

One of the five ESTs up-regulated in high sugar content (HS) mature leaves coded for an omega-3 fatty acid desaturase-*FAD8* (*SCSBAM1085B06.g*) (Table 4.1). In higher plants, the membrane lipids contain a high proportion of trienoic fatty acids (TAs). It has been suggested that these fatty acids, especially linolenic acid, are precursors of a defense-related signal molecule, jasmonate (JA). In *Arabidopsis*, three genes encoding omega-3 fatty acid desaturase, namely *FAD3*, *FAD7* and *FAD8*, are responsible for the production of TAs. Environmental stimuli, such as wounding, salt stress and pathogen invasion, which lead to a rapid increase in JA production, significantly induce the expression of the *FAD7* and *FAD8* genes (Nishiuchi and Iba, 1998). The sugarcane gene *FAD8* was enriched in the high sugar content individuals nine months after planting (Figure 4.2), and was always more expressed in these plants throughout the growing season (Figure 4.3). This EST was more expressed in immature leaves and also had high levels in immature internodes, which are considered as sink tissue, but also presented a weak expression in mature leaves (Figure 4.4). The data point to a role of JA synthesis in sucrose metabolism. This is the first report of the involvement of this gene in sucrose synthesis, and opens a wide array of possibilities to elucidate jasmonate role.

The putative involvement of a receptor-like protein kinase in sugar sensing

Recent evidence suggested that plants have many different types of receptor-like protein kinases (RLKs) that may transduce extracellular information into the cell (Becraft, 2002, Morillo and Tax, 2006). One sugarcane EST encoding for a leucine-rich repeat receptor-like kinase (*SCEQRZ3020C02.g*) was enriched in the high sugar content individuals nine months after planting (Table 4.1). RLKs have been identified from a number of plants and have been categorized into classes based on the different structural motifs found in their extra cellular domains. The physiological function of most RLKs is unknown, but some of them are involved in disease resistance and plant development (Becraft, 2002). To study the metabolic interactions between the mesophyll and bundle-sheath cells in sorghum, a C4 plant like sugarcane, Annen and Stockhaus (1999) cloned and characterized protein kinases that could be involved in the regulatory processes and signal transduction of this metabolic pathway. One of these was an *RLK1* that accumulated at much higher levels in the mesophyll cells than in the bundle sheath cells, and was almost undetectable in the roots. *In situ* analyses showed that transcripts of the sugarcane *RLK* were preferentially expressed in the bundle-sheath cells and it was expressed only in mature leaves (Vicentini *et al.*, manuscript submitted). It is known that receptor-like protein kinases are responsible for the activation of other factors that are translocated to the nucleus to induce gene expression changes (Hardie, 1999). This suggests that the sugarcane *RLK*, a bundle-sheath cell specific protein kinase, could be involved in cellular signaling cascades mediated by abundant sugar levels in sugarcane mature leaves.

Putative activation of Sucrose Phosphate Synthase (SPS) by the circadian clock

An EST homologous to a single Myb-repeat transcription factor, named CIRCADIAN CLOCK ASSOCIATED (CCA1) or LATE ELONGATED HYPOCOTYL (LHY) (*SCCCLR1C08G10.g*), was up-regulated in HS mature leaves (Table 4.1). CCA1/LHY and the TIMING OF CAB EXPRESSION 1 (TOC1) are thought to participate in a negative feedback loop, which is part of a model for the central oscillator in the *Arabidopsis* circadian clock. In higher plants the circadian clock controls hypocotyl elongation, daily leaf movements,

flowering time and the rhythm of CO_2 fixation (McClung, 2001, Blasing *et al.*, 2005). The sugarcane LHY/CCA1 gene was more expressed in HS individuals throughout the growing season (Figures 4.2 and 4.3). In tomato, Jones *et al.* (1997) demonstrated that the circadian rhythm controls the timing of sucrose-phosphate synthase phosphatase activity, which, in turn determines the activation of sucrose phosphate synthase (SPS). SPS catalyses the conversion of UDP-glucose and fructose-6-phosphate to sucrose-6-phosphate, the second to last step in sucrose biosynthesis (Huber and Huber, 1996). Pathre *et al.* (2004) demonstrated that the diurnal variation observed in the activity of SPS was not due to any intrinsic rhythm, but due to the transient changes in environmental conditions, like irradiance and temperature. When the circadian clock was correctly tuned with the environment, *Arabidopsis* plants presented increased photosynthesis and growth (Dodd *et al.*, 2005). The sugarcane EST was mainly expressed in mature and immature leaves, lateral buds and flowers, but also presented a weak expression in immature internodes and roots (Figure 4.4). We suggest that the expression profile of LHY/CCA1 transcripts in HS plants could be related to a photosynthetic advantage and, consequently, an enhanced carbon fixation and sucrose synthesis.

Putative model for sugar starvation regulation of SPS

Three ESTs coding for 14-3-3 proteins (*SCCCRZ1001D02.g*, *SCEQRT1025D06.g* and *SCEQRT1031D02.g*) were found to be more expressed in mature leaves from the LS population (Table 4.1). Recent reports have pointed out the importance of these adapter proteins in plant metabolic pathways (Ferl, 2004, Huber *et al.*, 2002). It was suggested that the members of this family affect nitrate fixation by regulating nitrate reductase (NR) and carbohydrate metabolism by binding to SPS. This enzyme has several putative phosphorylation sites that regulate its activity by 14-3-3 -dependent and -independent mechanisms. Non-14-3-3 events include phosphorylation of SPS on Ser-424 and Ser-158 that is thought to be responsible for light/dark modulation and osmotic stress activation of the enzyme (McMichael *et al.*, 1993, Toroser and Huber, 1997). However, there is a site-specific regulatory interaction between the 14-3-3 proteins and Ser-229 of spinach SPS, which inhibits SPS activity (Toroser *et al.*, 1998). This regulatory node is likely to be the same one occurring in the NR regulation. In its unphosphorylated state, SPS is active. Phosphorylation by a kinase, such as SNF1, does not inactivate SPS, but tags the enzyme for 14-3-3 binding which completes the signal-induced transition towards inactivation (Bachmann *et al.*, 1996, Moorhead *et al.*,

1999). Phosphorylated SPS, bound by 14-3-3s, may be inactivated directly in a reversible manner or may be destabilized and subjected to proteolysis (Sehnke *et al.*, 2002, Comparot *et al.*, 2003). It has been reported that during sugar starvation, targets for 14-3-3 proteins are degraded by proteases. The function of this is not clear, but it has been suggested that it represents a safety valve for metabolic regulation (Cotelle *et al.*, 2000). Various research groups have reported the impact of 14-3-3 proteins on metabolism. The over-expression of 14-3-3 proteins in potato induced an increase in catecholamine and soluble sugar contents in the leaves, whilst a 14-3-3 anti-sense experiment increased the tuber starch content, NR activity and amino acid composition (Prescha *et al.*, 2001, Swiedrych *et al.*, 2002). In addition, Zuk *et al.* (2003) observed a significant increase in potato SPS and NR activities when all of the six 14-3-3 isoforms were repressed. In line with these data, the upregulation of three ESTs coding for 14-3-3 proteins in LS mature leaves could reflect the inactivation state of SPS and consequently the low sugar content in these plants.

One of the ESTs that was up-regulated in LS plants codes for an SNF1-related protein (*SCEQLB2019B08.g*) (Table 4.1). SnRK1 (SNF1-Related Protein Kinase-1) is a plant protein kinase with a catalytic domain similar to that of SNF1 (Sucrose Non-fermenting-1) of yeast and AMPK (AMPactivated protein kinase) of animals (Halford and Paul, 2003). Carraro *et al.* (2001) identified at least 22 sugarcane expressed sequence tag (EST) contigs encoding putative SnRKs in the SUCEST database. Studies led to the hypothesis that once SnRK1 is activated in response to high intracellular sucrose and/or low intracellular glucose levels, SnRK1 can phosphorylate plant enzymes and activate starch synthesis in potato tubers (Halford and Paul, 2003, Rolland *et al.*, 2006). The first plant protein to be identified as a substrate for SnRK1 was a HMG-CoA reductase in *A. thaliana* (Dale *et al.*, 1995). Subsequently, two other important enzymes, SPS and NR were shown to be substrates for SnRK1 phosphorylation in Ser-binding sites. In both cases, phosphorylation results in inactivation of the enzyme, although the inactivation of NR and SPS also requires the binding of a 14-3-3 protein to the phosphorylation site, discussed before (Bachmann *et al.*, 1996, Moorhead *et al.*, 1999). The sugarcane *SnRK1* transcript was up-regulated in low sugar content mature leaves, nine months after planting (Figure 4.2). However it can be observed in Figure 4.3 that, at times, this transcript had the opposite expression profile. For example, its levels were lower in the low sugar content leaves, 6, 11 and 13 months after planting.

The sugarcane *SnRK1* transcript was higher expressed in sink tissues, such as immature leaves, internodes, lateral buds and flowers. A weak expression could be seen in

mature leaves and roots (Figure 4.4). There is evidence that this class of protein kinases is also involved in the regulation of C_4 photosynthesis Annen and Stockhaus (1998), but the presence of sugarcane transcripts in sink tissues also suggests the envolviment of this kinase in sugar translocation. This makes it unlikely that this kinase is involved solely in the regulation of developmental processes related to photosynthesis and leaf development. Under the conditions of low sugar content, SPS activity decreases because of an increase in the phosphorylation state of the enzyme (Huber *et al.*, 1989, Paul and Foyer, 2001). As stated above, the fact that three sugarcane 14-3-3 and a SnRK1 were more expressed in low sugar content individuals 9 months after planting could reflect their role in keeping SPS in an inactivated state that would account for the lower sucrose levels in these plants.However, the expression profile along the growing season observed for SnRK1 suggests that a complex regulation might be involved in the signaling pathway modulated by these genes. Future work with transgenic sugarcane plants would be helpful to discover the function of these genes.

Lignin biosynthesis and secondary wall synthesis in low sugar content sugarcane plants

Lignin is a complex polymer, which provides structural integrity in plants. In sugarcane bagasse it makes 23,1% by weight of biomass. Lignin remains as residual material after the sugars in the biomass have been converted to ethanol. It contains a lot of energy and can be burned to produce steam and electricity for the biomass-to-ethanol process. Three enzymes are involved in the biosynthetic pathway of lignin: cinnamoyl-coenzyme A reductase (CCR), cinnamyl alcohol dehydrogenase (CAD) and caffeic acid 3-O-methyltransferase (COMT). An EST coding for a COMT (*SCRFLR1012F12.g*) was found to be enriched in the low sugar content population (Table 4.1). Lignins are found in the secondary cell walls of vascular plants and they play an important role by reducing the permeability of the cell wall to water, providing mechanical strength and defense against wounding and infection (Lewis and Yamamoto, 1990). There are many evidences that some genes that encode lignin biosynthetic enzymes are regulated at the transcriptional level (Anterola and Lewis, 2002). One of these studies suggests that at least three different stimuli may modulate this regulation: light, circadian clock and carbohydrate availability (Rogers *et al.*, 2005). All of them are linked through light that determines the quantity of carbohydrate that are synthesized. The timing and localization of some of these transcripts show a strong correlation with the deposition

of lignin, like in mature sugarcane stems (Casu *et al.*, 2004). The storage parenchyma of the maturing sugarcane stem internodes is extensively lignified and Jacobsen *et al.* (1992) proposed that this process parallels with the increase in sucrose content observed in mature internodes. We observed that the LS plants had more lignified leaves and stem barks than HS plants (our unpublished data), in agreement with the higher levels of *COMT* transcripts in these plants.. Interestingly, in transgenic alfafa plants with reduced levels of COMT the cell walls were more amenable to enzymes (Chen and Dixon, 2007). It is important to note that, to our knowledge, an association of COMT enzymes with sugar levels in sugarcane was not related before.

Signals can be perceived and amplified at the cell membrane by receptors coupled to a variety of signaling pathways, including the inositol 1,4,5-trisphosphate (IP3) pathway. This second messenger is produced from the hydrolysis of phosphatidylinositol 4,5 bisphosphate and raises the Ca^{2+} levels in the cytosol (Berridge, 1993). In potato, Ohto *et al.* (1995) demonstrated that the sugar-inducible expression of genes for sporamin and β-amylase involves a Ca^{2+}-mediated signaling. Inositol-polyphosphate 5-phosphatases (5Ptases) comprise a large group of enzymes that can hydrolyze 5-phosphates from a variety of inositol phosphates such as IP3 (Majerus *et al.*, 1999). An EST coding for a 5PTase (*SCSBST3096H04.g*) was up-regulated in the low sugar content plants. This sugarcane *5PTase* was differentially expressed in mature LS leaves collected nine months after planting (Table 4.1). RNA-blot analysis confirmed the differential expression in three individual samples (Figure 4.2). However, as shown in Figure 4.3, the same expression profile was not observed throughout the growing season. *5PTase* was over-expressed in LS plants only at nine and thirteen months after planting. For the other time-points a slight difference in the expression profile between high and low sugar content plants could be observed. It was also more expressed in mature leaves in comparison to the other tissues and a weak expression could be seen in sink tissues (Figure 4.4). 5PTases may be involved in the modulation of Ca^{2+} levels and there is a wide array of evidence for a role of Ca^{2+} in sugar signaling (Furuichi *et al.*, 2001, Rolland *et al.*, 2002). However, Zhong *et al.* (2004) demonstrated that the *FRAGILE FIBER3* (*FRA3*) gene of *A. thaliana*, which encodes a type II 5PTase, plays an essential role in the secondary wall synthesis in fiber cells and xylem vessels. The authors showed that *fra3* mutations caused a dramatic reduction in secondary wall thickness and a concomitant decrease in stem strength. In agreement of this expression profile in LS plants, we propose that sugarcane *5PTase* is involved in secondary wall synthesis in these plants, although the putative role of 5PTase in a Ca^{2+} signaling cas-

cade triggered by sugars cannot be discarded. Further analysis is necessary to elucidate these mechanisms.

Expression of stress-related proteins in low sugar content plants

Sugar-signaling pathways do not operate in isolation but are part of cellular regulatory networks. Recent results clearly show cross talk between different signaling systems, especially those of sugars, phytohormones and light. It is interesting to note that seven stress-related genes were up-regulated in the LS plants. Most of them are cold and drought-induced and they also include a ribonuclease and a wound-induced protein. Three sugarcane stress-related ESTs (*SCUTST3084F06.g*, *SCACCL6008H06.g*, *SCEPRZ3087C08.g*) belong to a class of low-molecular-weight hydrophobic proteins involved in maintaining the integrity of the plasma membrane during cold, dehydration and salt stress conditions. These genes are activated by environmental factors, such as dehydration and salinity, and by chemical signals such as abscisic acid (Morsy *et al.*, 2005). Another differentially expressed stress-related gene encodes a plasma membrane intrinsic protein (*SCCCRZ1002E08.g*). These proteins facilitate water flux across cell membranes and play important roles in plant growth and development. One EST coding for a dehydrin (*SCQGLR1085F11.g*) was also up-regulated in the LS plants (Figures 4.2 and 4.3). These proteins are supposed to stabilize macromolecules and/or protect membranes against chilling damage (Pearce, 1999). Two putative sugarcane dehydrin-like proteins were identified by Nogueira *et al.* (2003) in a sugarcane cold-response datamining. They proposed that these putative sugarcane antifreeze proteins could confer cellular membrane protection, reducing chilling injury. The S-like RNAse (*SCJLRT1016G06.g*) is a protein present in higher plants that controls self-incompatibility. In a self-incompatible *Antirrhinum*, S-RNAse transcription was induced during leaf senescence and phosphate (Pi) starvation but not by wounding, indicating that this gene plays a role in remobilizing Pi and other nutrients (Liang *et al.*, 2002). Finally, a stress-related EST differentially expressed in low sugar content plants, is a protein described as being wound-induced (*SCCCLR2C01F06.g*).

All of these stress-related ESTs were differentially expressed in at least the nine-month samples of low sugar content sugarcane mature leaves (Table 4.1). This is of interest because, in general, a low sugar status enhances the expression of genes involved in photosynthesis, reserve mobilization and export, whereas the abundant presence of sugars pro-

motes the expression of genes involved in growth and carbohydrate storage (Koch, 1996, Rolland *et al.*, 2002, Rolland *et al.*, 2006). Nevertheless our results are in agreement with those observed in rice cell cultures, where the low sugar status also up-regulated several stress-related genes (Ho *et al.*, 2001). These authors suggested that sucrose starvation was a type of nutritional stress and that the expression of these stress-related genes might play roles in protecting cells against it. The LS plants may be under some kind of stress caused by the lower levels of sucrose triggering the differential expression of stress-related genes in this population.

The transcription and ubiquitination category

Two transcription factors and a protein involved in ubiquitination were up-regulated in low sugar content plants (Table 4.1). They code for a protein from the E2F/DP family (*SCCCLB1002B01.g*), a Tubby-like protein (*SCCCCL4001D08.g*) and a TIR1-like F-box containing protein (*SCCCCL6003D08.g*). In animals, the E2F family plays an important role in cell cycle control by regulating the transcription of genes involved in the progression from the G1 (G0) to the S phase. In *Arabidopsis*, De Veylder *et al.* (2002) showed that E2Fa-DPa was an important factor that determined the proliferative status of plant cells. Recently, Lai *et al.* (2004) identified a TUBBY-like gene family with 11 members in *Arabidopsis*, and all of them contained a conserved F-box domain. Recent studies on the auxin response demonstrated that a F-box protein TIR1 is part of an ubiquitin protein ligase required for degradation of Aux/IAA proteins (Yang *et al.*, 2004). To our knowledge, this is the first report about the expression of these categories in plants that segregate for sugar content. Further characterization of these sugarcane proteins should be carried out in order to define its activity and the involvement in sugar level.

Feedback regulation of photosynthesis

There are many evidences that sink tissues exert an influence on the photosynthetic rates and carbohydrate levels of source organs (Paul and Foyer, 2001). Recently, McCormick *et al.* (2007) demonstrated a relationship between source and sink tissues in sugarcane, where demand for carbon from sinks affects source leaf photosynthetic activity, metabolite levels and gene expression.

In fact, the activity of photosynthesis-related enzymes, and expression of associated gene transcripts in the source leaf are modified by the local levels of sugar and hexoses that will be transported to sink (Rolland *et al.*, 2002). As observed in sugarcane, a decreased hexose levels in leaf may act a signal for increased sink demand, reducing a negative feedback regulation of photosynthesis (McCormick *et al.*, 2006). Recent studies of the same group showed that hexoses, rather than sucrose, could be involved in this regulation (McCormick *et al.*, 2007). Hexoses have been involved in the regulation of source metabolism via signal transduction pathways involving protein phosphorylation via MAPK activities (Ehness *et al.*, 1997). However it is difficult to address the specific role of this enzyme in the source to sink regulation. There are a wide variety of signaling pathways that are associated with these kinases, and the same occur to many other components related to signal transduction.

As stated above, the expression of 14-3-3 and a SnRK1 give insights in a feedback regulation of photosynthesis, keeping the enzyme SPS in an inactivated state that would account for the lower sucrose levels in these plants. Our results also suggests that sugar levels seem to modulate gene expression at the transcriptional level through a complex signal transduction network that may involve common responses related to stress. The data provide an insight into the role of sugar levels in signal transduction pathways. Some expression trends of low sugar levels such as up regulation of 14-3-3 proteins, a SnRK1 and stress-related proteins were substantiated by the present data at the transcript level. These genes are interesting targets for further research using other approaches, such as overexpression or gene silencing. An in-depth analysis of these components should improve our knowledge on how signal transduction can regulate sucrose synthesis in sugarcane plants.

Supplementary Material

The following supplementary material is avaiable for this article online:

Table SI. Expression patterns of the SUCAST chip components. The table indexes the ratios of the microarray signal between the two channels (Cy3 and Cy5). For each clone, the median intensity value of the technical replicates was used to calculate the ratio between the experimental samples (high and low sugar content). Ratio of the microarray signal intensity. Lv=leaf. The numbers 1 and 2 denote the technical replicates. * indicates the SAS identity that was not confirmed by re-sequencing.

Table SII. Accession number of cDNA sequences that compose the differentially

expressed SAS.

Acknowledgements

This work was funded by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) and CNPq. We are indebted to Adriana Y. Matsukuma and Denise Yamamoto for their technical assistance in microarray printing and scanning performed in the laboratory of the Cooperation for Analysis of Gene Expression (CAGE) and Dr. Sonia di Mauro for coordinating the distribution of the SUCEST clones.

Centenas de vezes, todos os dias, eu me recordo de que minha vida interior e exterior dependem do trabalho de outros homens, vivos e mortos, portanto, devo esforçar-me para retribuir na mesma medida aquilo que recebi e que continuo recebendo.

Albert Einstein (1879-1955)

5

Improving the prediction of protein subcellular localization using predicted profiles‡

Renato Vicentini[1] and Marcelo Menossi[1]

Abstract

Background: Subcellular localization is a key feature of proteins, since it is related to biological function. The determination of subcellular localization using computational prediction is a highly desirable strategy because experimental approaches are time-consuming. Many tools are available for predicting protein localization in the cell. Each tool performs well on certain data sets, but their predictions often disagree for a given protein. The vast majority of existing methods for predicting protein subcellular localization is limited to a single location site, despite the knowledge that the protein may exist simultaneously at two or more different subcellular localizations.

Results: In order to develop a method for the enhanced prediction of subcellular localization, the outputs of seven localization prediction tools were integrated so as to optimally exploit the potential of each one, based on prediction profiles, and thus be able to deal with the case of multiple location sites. The performance was tested with two subcellular localization protein

‡submitted

[1]Department of Genetics and Evolution, Functional Genomics Laboratory, Institute of Biology, CP 6009, State University of Campinas - UNICAMP, 13083-970, Campinas, SP, Brazil

data sets. Based on the Matthews correlation coefficient measurement, the prediction performance of this new method, named PWMSubLoc, was clearly superior to all the methods used to create the predictor. The overall prediction accuracy was 90% for the GS data set, and 86% for the DBSubLoc data set.

Conclusions: This novel method provided superior prediction performance as compared to existing algorithms, and can be easily complemented with any other existing method. An implementation of the method described is freely available to both academic and commercial users as a web server at http://ipe.ib.unicamp.br/pub/PWMSubLoc.

Background

The eukaryotic cell is highly organized, and various biological processes are associated with specialized subcellular structures, or confined to particular compartments. It is important to know the subcellular localization of proteins to gain deeper understanding of the functions of organelles and the compartmentalization of cellular metabolism. As a result, knowledge of the subcellular localization of proteins provides clues to their function as well as to the interconnectivity of biological processes (Shen and Burger, 2007).

While most proteins in an eukaryotic cell are encoded in the nuclear genome and synthesized in the cytosol, many of them need to be further sorted out into one or other subcellular compartment (Emanuelsson *et al.*, 2000). Signals for protein sorting exist either in the form of primary sequences, usually N-terminal targeting sequences (Rusch and Kendall, 1995, Emanuelsson *et al.*, 2000) or internal sequence motifs (Cokol *et al.*, 2000). Proteins localized in the same organelle have been reported to show a similar overall amino acid composition and are thought to have evolved to function optimally in that specific environment (Andrade *et al.*, 1998).

Proteins may simultaneously exist at, or move between, two or more different subcellular localizations. For instance, 15% of the human protein entries that have experimentally observed subcellular localization annotations in the Swiss-Prot database (version 50.7) have multiple localization sites (Shen and Chou, 2007). Proteins with multiple locations or a dynamic feature of this type are particularly interesting, because they may have special biological functions. For example, most mitochondrial or chloroplastic proteins are encoded by the nuclear genome, translated on cytosolic ribosomes, and directed to the appropriate organelle by following a specific pathway guided by signals in their amino acid

sequence (Pujol *et al.*, 2007). The dual targeting of proteins to both mitochondria and chloroplasts was originally expected to be rare, but the number of proteins that have been shown to have dual targets has greatly increased (Millar *et al.*, 2006). Recent studies have shown an unexpectedly high frequency of dual-targeting proteins, and novel routes of protein trafficking. Such findings makes it more difficult to predict which proteins are really targeted to organelles (Millar *et al.*, 2006).

The prediction of protein subcellular localization focuses on determining the localization sites of unknown proteins in a cell. Despite recent technical advances, the experimental determination of protein subcellular localization remains time-consuming and labor-intensive. Given the size and complexity of the genomic data, prediction systems are interesting approaches to identify and screen possible candidates for further analysis. Compared with experimental methods, computational prediction methods that can provide fast and accurate assignments of protein subcellular localization are very desirable (Xie *et al.*, 2005).

Several efforts have been made to predict the protein subcellular localization. Up to now two main categories of prediction methods have been proposed. One was based on the existence of sorting signals in the N-terminal sequences, including signal peptides, mitochondrial targeting peptides and chloroplastic transit peptides (Emanuelsson *et al.*, 2000, Nielsen *et al.*, 1997, Nielsen *et al.*, 1999). The other category was based on the amino acid composition of the protein sequences in different subcellular localizations (Nakashima and Nishikawa, 1994, Reinhardt and Hubbard, 1998, Guo *et al.*, 2004*a*). Many different approaches for predicting protein subcellular localization have been proposed.

Each of the localization prediction tools available shows different strengths, and no tool is clearly and globally optimal (Shen and Burger, 2007). Any given tool will perform well on certain data but poorly on others, and often predictions by different tools disagree. This is not surprising, because the tools employ different algorithms, sequence features and training data (Shen and Burger, 2007). However, by using the proper exploitation of the combined strengths of these prediction programs, it may be possible to construct predictors whose performances surpass those of all existing prediction programs (Liu *et al.*, 2007).

One of the critical challenges in predicting protein subcellular localization is to find a way to deal with the case of multiple location sites. For most existing predictors, the multi site proteins are either excluded from consideration or even assumed not to exist (Liu *et al.*, 2007).

In this paper, a computation method called PWMSubLoc was developed to pre-

dict the subcellular localization of a protein previously deployed to predict profiles to improve the knowledge of the subcellular localization of proteins. The present study was initiated in an attempt to develop a method to predict the subcellular localization of proteins including those with multiple localizations. In the current on-line version, it is possible to use seven predicted profiles created from iPSORT (Bannai *et al.*, 2002), MitoProtII (Claros and Vincens, 1996), Predotar (Small *et al.*, 2004), PSORT (Nakai and Horton, 1999), SubLoc (Hua and Sun, 2001), TargetP (Emanuelsson *et al.*, 2000), and PredictNLS (Cokol *et al.*, 2000). Following the information theory, the present approach combines the complementary strengths of existing prediction methods. Using the example of two data sets, heterogeneous localization predictors were integrated, their performance tested with known data and the most efficient way of integration selected. The method is readily applicable to proteins with single or multi subcellular localizations. All scripts and interfaces were written in the Perl and R languages. This version of the program is available at the PWMSubLoc web server.

Implementation

Description of the method and the input and output parameters

The PWMSubLoc method can be described as the use of previous prediction profiles to better exploit their potential and improve the prediction of protein subcellular localization. Currently, the PWMSubLoc method is comprised of five position weight matrix (PWM) models, one for each subcellular localization, composed by seven predictors in each. The PWMSubLoc modularity allows for relatively simple additions of new localizations or predictors. Increasing the coverage of predictors and localizations is envisioned in future updates. PWMSubLoc and all PWM models, are freely available to both academic and commercial users as a user-friendly web interface available at http://ipe.ib.unicamp.br/pub/PWMSubLoc.

The input data are made up of a simple selection of the previous predictions made by the seven softwares adopted in this version. The background processing subsystem computes and shows the total information content (measured in bits) using the predicted profile for all available PWM models.

Prediction by individual tools

Seven prediction tools were selected for subcellular localization: iPSORT (Bannai *et al.*, 2002), MitoProtII (Claros and Vincens, 1996), Predotar (Small *et al.*, 2004), PSORT (Nakai and Horton, 1999), SubLoc (Hua and Sun, 2001), TargetP (Emanuelsson *et al.*, 2000) and PredictNLS (Cokol *et al.*, 2000). The selection was based on the diversity of the algorithms and the sequence features they employed. These tools were used as base-level classifiers, whose prediction results were combined to build the new classifier. The prediction of mitochondrial protein can be performed by Predotar, MitoProtII, PSORT, TargetP, SubLoc and iPSORT. The same tools were used for chloroplastic proteins, with the exception of SubLoc. PredictNLS, PSORT, and SubLoc were used for the prediction of nuclear localization. The proteins from the secretory pathway were predicted by PSORT, TargetP, SubLoc, and iPSORT. Finally, the PSORT and TargetP were used for the cytoplasmic proteins.

Data sets

Two data sets were used for training and testing the new model for subcellular prediction. The first one was the DBSubLoc (Guo *et al.*, 2004*b*), which, in the eukaryotic non-redundant version (retrieved August, 2007), contained 8100 eukaryotic protein sequences belonging to five used location categories: nuclear, cytoplasmic, mitochondrial, chloroplastic and secretory pathways (4074 nuclear, 930 cytoplasmic, 1659 mitochondrial, 400 chloroplastic, and 1037 secretory pathway proteins). In this data set, no pair of sequences had $\geq 60\%$ identity. The second data set was generated by Guo et al. (Guo *et al.*, 2004*b*) and contained 7844 eukaryotic protein sequences belonging to 5 used location categories: 1019 chloroplastic, 2387 cytoplasmic, 595 secretory pathway, 644 mitochondrial, 3199 nuclear. For convenience, these two data sets are referred to as the DBSubLoc and GS data sets.

Analysis of the information content at each subcellular localization based on the predicted profiles

The method begins by calculating a weight matrix from the frequencies of each prediction (localization) for each predictor (software) for all the sequences that have the same known subcellular localization. This procedure is applied to all the subcellular localizations. These matrixes are then applied to the prediction profile to determine the more probable

subcellular localization of each individual protein.

A PWM model for each subcellular localization, called $R_{iw}(l, s)$, is created by using a training set consisting of previously predicted sequences from the databases described before. The PWM is computed using the widely accepted information theoretical approach with modifications (Schneider, 1997, Vicentini and Menossi, 2007).

The information content of predicted subcellular localization profile ($R_i(s)$), given by manipulation of the $R_i(j)$ (Schneider, 1997), is the product between the prediction and the weight matrix:

$$R_i(s) = \sum_{l=1}^{5} S(l, s)\, R_{iw}(l, s) \qquad \text{(bits per predictor)} \tag{5.1}$$

And the total information content of the subcellular localization is the R_i:

$$R_i = \sum_{l} R_i(s) \qquad \text{(bits per subcellular localization)} \tag{5.2}$$

Performance assessment

Preparation of the training and testing data. Each of the previous described data sets was used for training and testing the PWMSubLoc. The total sequences were divided into two equal sets, training and testing. The training set was cross-validated by testing with the testing data set.

Confusion matrix method. Using the confusion matrix, various measurements of quality such as accuracy, false discovery rate, specificity, sensitivity and the Matthews correlation coefficient (MCC) (Matthews, 1975) were determined. In the following equation (5.3-5.7), TP refers to true positives, TN to true negatives, FN to false negatives and FP to false positives.

accuracy (AC): proportion of correct predictions amongst the total predictions.

$$AC = \frac{TP + TN}{TP + TN + FN + FP} \tag{5.3}$$

false discovery rate (FDR): proportion of false predictions amongst the total positive predictions.

$$FDR = \frac{FP}{TP + FP} \tag{5.4}$$

specificity (SP): ratio of true negatives to the total negatives.

$$SP = \frac{TN}{TN + FP} \tag{5.5}$$

sensitivity (SN): ratio of true positives to the total positives.

$$SN = \frac{TP}{FN + TP} \tag{5.6}$$

MCC: This is regarded as a more rigorous measurement to evaluate the performance of class prediction methods. MCC equals 1 for perfect predictions (Matthews, 1975).

$$MCC = \frac{(TP * TN) - (FP * FN)}{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)} \tag{5.7}$$

Accuracy, FDR, specificity and sensitivity tests. Specificity and sensitivity are two competing quality measurements for any two-classifier method. Prediction can be performed at different levels of specificity and sensitivity by defining various thresholds for the final score. At each threshold, the numbers for the TPs, TNs, FPs and FNs are calculated, and, based on these, the value parameters such as MCC, accuracy, false discovery rate, specificity and sensitivity, are determined using equations (5.3-5.7). Table 5.1 (for the GS data set) and Table 5.2 (from the DBSubLoc data set) show these parameters of the PWMSubLoc at the different threshold levels. For example, for mitochondrial predictions performed in the GS data set, it was observed that at the highest specificity (0.99), the sensitivity was at 0.33, and at the highest sensitivity level (0.88), the specificity was at 0.66. At a mid-range sensitivity level of 0.51 (with a threshold score of ≥ 50), 51% of all the positives could be predicted with only 1% FPs, and the best correlation was obtained (MCC = 0.61). The same procedure was applied to all the five subcellular localizations in both data sets, and the ranges of threshold scores with the best correlation (MCC) are indicated in Tables 5.1 and 5.2. Figure 5.1 shows the relationship between the true positive rate plotted against the false positive rate, using a receiver operating characteristic (ROC) plot. This figure demonstrates the prediction performance of PWMSubLoc for the five subcellular compartments (Chloroplastic, Mitochondrial, Nuclear, Secretory pathway, Cytoplasmic), and is able to show that PWMSubLoc give a high

true positive discovery rate with a low false positive rate.

Comparison of the PWMSubLoc performance

The performance of the PWMSubLoc was compared with that of the seven programs used to compute the PWM models. The iPSORT, Predotar, and MitoProtII programs were installed locally and trained with the same data used for training PWMSubLoc. PSORT, SubLoc, TargetP, and PredictNLS predictions were carried out from the respective web servers.

Multi subcellular localization

A new class of plant organellar proteins was found recently, corresponding to dual targeted proteins to both mitochondria and chloroplasts (Pujol *et al.*, 2007). As mentioned at the beginning, several of the previous prediction studies were confined to within the scope of a single location prediction. To demonstrate the applicability of the PWMSubLoc to predict proteins with multiple subcellular locations, predictions were performed in a data set composed by 35 *Arabidopsis thaliana* dual targeted proteins (Pujol *et al.*, 2007).

Results and discussion

The performance of the PWMSubLoc was compared against all the programs used in the construction of PWM (Tables 5.3 and 5.4). The MCC value at each location was also shown in order to show a more comprehensive evaluation of the performance of the new predictor. Since MCC considers not only the number of true positives but also the number of false positives, false negatives and true negatives, it is more reliable and more comprehensive than accuracy statistics. PWMSubLoc attained the best overall accuracy of 90.2% and 86% for the GS and DBSubLoc data sets, respectively. In both sets, PWMSubLoc achieved improvements of 2.5% at 7.5% in the overall accuracy as compared to the other approaches in each data set. With respect to MCC, the PWMSubLoc performed better than the other approaches for all the subcellular localizations tested, with the exception of the cytoplasmic one, where SubLoc showed a similar score. In the case of DBSubLoc cytoplasmic proteins, both the predictors used to create the model showed a poor performance, with higher FDR values,

Tabela 5.1: **Confusion matrix values, obtained with the GS data set, and the dependent parameters at each threshold value.**

Score threshold	Positives tested TP	FN	Negatives tested TN	FP	Accuracy	SN	SP	FDR	MCC
Chloroplastic									
≥70	309	710	7195	90	0.904	0.303	0.988	0.226	0.446*
≥60	409	610	5171	2114	0.672	0.401	0.710	0.838	0.079
Mitochondrial									
≥70	213	431	7631	29	0.945	0.331	0.996	0.120	0.520
≥60	288	356	7607	53	0.951	0.447	0.993	0.155	0.593
≥50	331	313	7569	91	0.951	0.514	0.988	0.216	0.611*
≥40	390	254	7464	196	0.946	0.606	0.974	0.334	0.606
≥30	465	179	7197	463	0.923	0.722	0.940	0.499	0.562
≥20	559	85	5296	2364	0.705	0.868	0.691	0.809	0.313
≥10	571	73	5083	2577	0.681	0.887	0.664	0.819	0.303
Nuclear									
≥70	966	2233	5089	16	0.729	0.302	0.997	0.016	0.450
≥60	1570	1629	4969	136	0.787	0.491	0.973	0.080	0.559
≥50	1987	1212	4956	149	0.836	0.621	0.971	0.070	0.659
≥30	2845	354	4476	629	0.882	0.889	0.877	0.181	0.756*
≥10	2928	271	4110	995	0.848	0.915	0.805	0.254	0.702
>0	2972	227	4092	1013	0.851	0.929	0.802	0.254	0.712
Secretory pathway									
≥70	180	415	7678	31	0.946	0.303	0.996	0.147	0.489
≥60	254	341	7668	41	0.954	0.427	0.995	0.139	0.587
≥50	340	255	7535	174	0.948	0.571	0.977	0.339	0.587
≥40	395	200	7514	195	0.952	0.664	0.975	0.331	0.641
≥30	422	173	7485	224	0.952	0.709	0.971	0.347	0.655*
≥20	549	46	7123	586	0.924	0.923	0.924	0.516	0.636
>0	560	35	6716	993	0.876	0.941	0.871	0.639	0.537
Cytoplasmic									
≥70	918	1469	5718	199	0.799	0.385	0.966	0.178	0.466
≥60	1876	511	4976	941	0.825	0.786	0.841	0.334	0.599*
≥20	2104	283	4099	1818	0.747	0.881	0.693	0.464	0.521

(*) shows the range of threshold scores at the best correlation (MCC) obtained.

Tabela 5.2: **Confusion matrix values, obtained with the DBSubLoc data set, and the dependent parameters at each threshold value.**

Score threshold	Positives tested		Negatives tested		Accuracy	SN	SP	FDR	MCC
	TP	FN	TN	FP					
Chloroplastic									
≥70	122	278	7663	37	0.961	0.305	0.995	0.233	0.469*
≥60	164	236	7549	151	0.952	0.410	0.980	0.479	0.438
≥50	201	199	7050	650	0.895	0.503	0.916	0.764	0.295
≥40	243	157	5726	1974	0.737	0.608	0.744	0.890	0.171
≥30	259	141	5192	2508	0.673	0.648	0.674	0.906	0.147
Mitochondrial									
≥70	547	1112	6363	78	0.853	0.330	0.988	0.125	0.480
≥60	676	983	6303	138	0.862	0.407	0.979	0.170	0.518
≥50	831	828	6156	285	0.863	0.501	0.956	0.255	0.535*
≥40	1001	658	5697	744	0.827	0.603	0.884	0.426	0.479
≥30	1290	369	4397	2044	0.702	0.778	0.683	0.613	0.377
≥20	1312	347	4236	2205	0.685	0.791	0.658	0.627	0.365
Nuclear									
≥70	1559	2515	3925	101	0.677	0.383	0.975	0.061	0.443
≥60	1802	2272	3893	133	0.703	0.442	0.967	0.069	0.480
≥50	3121	953	3086	940	0.766	0.766	0.767	0.231	0.533
≥20	3341	733	2928	1098	0.774	0.820	0.727	0.247	0.550*
≥10	3390	684	2871	1155	0.773	0.832	0.713	0.254	0.549
Secretory pathway									
≥70	469	568	6863	200	0.905	0.452	0.972	0.299	0.515
≥50	525	512	6834	229	0.909	0.506	0.968	0.304	0.545
≥40	655	382	6744	319	0.913	0.632	0.955	0.328	0.602
≥30	743	294	6608	455	0.908	0.716	0.936	0.380	0.614*
≥20	833	204	6072	991	0.852	0.803	0.860	0.543	0.530
≥10	945	92	5705	1358	0.821	0.911	0.808	0.590	0.533
Cytoplasmic									
≥70	484	446	5973	1197	0.797	0.520	0.833	0.712	0.278*
≥40	672	258	4947	2223	0.694	0.723	0.690	0.768	0.274

(*) shows the range of threshold scores at the best correlation (MCC) obtained.

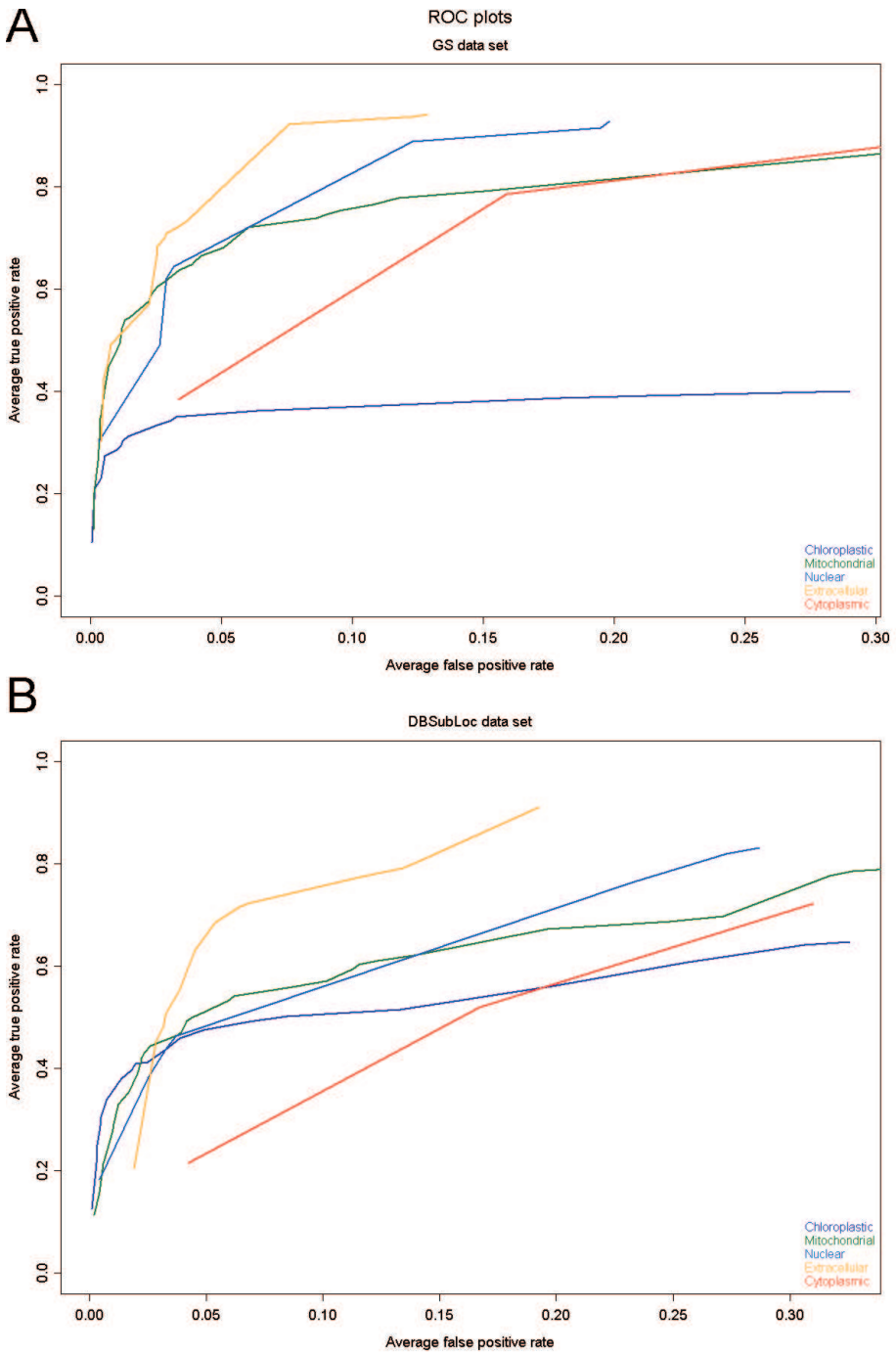


Figura 5.1: **PWMSubLoc prediction performance for the five subcellular compartments (Chloroplastic, Mitochondrial, Nuclear, Secretory pathway, Cytoplasmic) using ROC plots.** The data points used in the ROC plots corresponded to a full range of discrete score thresholds. (A) GS data set, and (B) DBSubLoc data set.

and the new method depends on accurate predictions made by these. The benchmarking results clearly showed that the PWMSubLoc was an effective method for improving subcellular localization prediction, with higher overall prediction accuracy than other popular methods.

Protein import machineries have been studied extensively in the organelles of many organisms. In contrast to the obvious differences in import machineries, the mitochondrial and chloroplast N-terminal targeting sequences have a similar over all composition (Pujol *et al.*, 2007). The difficulties in predicting dual targeted proteins were illustrated by Pujol *et al.* (2007).

The current method was applied to real data sets from different *A. thaliana* organellar proteins, to demonstrate their use in discriminating multi subcellular localization proteins. A total of 35 *A. thaliana* proteins with clearly established dual targeting (mitochondrial and chloroplastic) were used in the analysis. As a discrimination criterion, a threshold of 40% was set. In this analysis, the PWMSubLoc achieved 20% (7 out of 35) accuracy. The other two programs that are able to predict dual localization, Predotar and PSORT, showed 11% and 28% of accuracy respectively (Figure 5.2). Besides the low accuracy of all three methods, which had a minimal accuracy of 80% when considering mono localization for these two organelles, each was able to predict multi subcellular localization proteins that were not predicted by the other two (mainly in the analysis performed by PWMSubLoc and PSORT, Figure 5.2). This indicates that for multi subcellular localization prediction, the use of complementary methods could increase the prediction success rate. Also, the data indicated that future work to develop new algorithms to increase the accuracy for dual targeted proteins would be highly desirable.

Experimental approaches can also be used for subcellular determinations and frequently use cell fractionation, purification of organelles and mass spectrometry to identify peptides. Such approaches have been extensively undertaken, with a number of studies producing significant protein sets from major locations such as the plastids, the nucleus, the plasma membrane and the mitochondrion. For example, many hundreds of *Arabidopsis* proteins had their subcellular localization determined by visualization of their expression with fluorescent protein fusion (Heazlewood *et al.*, 2007). The same occurred for rice, where are available many reference maps based on two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) of proteins from subcellular compartments (Komatsu *et al.*, 2004).

The analyse described in this work are based on databases composed of protein sequences that do not a necessarily have a reliable location annotation obtained from

Tabela 5.3: **Performance comparison for the different subcellular localization prediction methods (PWMSubLoc, iPSORT, MitoProtII, Predotar, PSORT, TargetP, PredictNLS) for the GS data set.**

Score threshold	Positives tested		Negatives tested		Accuracy	SN	SP	FDR	MCC
	TP	FN	TN	FP					
Chloroplastic									
PWMSubLoc	309	710	7195	90	0.904	0.303	0.988	0.226	0.446*
iPSORT	240	779	6541	744	0.817	0.236	0.898	0.756	0.135
MitoProtII	298	721	6910	375	0.868	0.292	0.949	0.557	0.290
Predotar	341	678	6599	686	0.836	0.335	0.906	0.668	0.240
PSORT	207	812	6941	344	0.861	0.203	0.953	0.624	0.206
TargetP	365	654	6596	689	0.838	0.358	0.905	0.654	0.260
Mitochondrial									
PWMSubLoc	331	313	7569	91	0.951	0.514	0.988	0.216	0.611*
iPSORT	402	242	6644	1016	0.849	0.624	0.867	0.717	0.349
MitoProtII	274	370	7296	364	0.912	0.425	0.952	0.571	0.380
Predotar	420	224	7033	627	0.898	0.652	0.918	0.599	0.460
PSORT	281	363	7278	382	0.910	0.436	0.950	0.576	0.381
SubLoc	395	249	6820	840	0.869	0.613	0.890	0.680	0.379
TargetP	395	249	6904	756	0.879	0.613	0.901	0.657	0.398
Nuclear									
PWMSubLoc	2845	354	4476	629	0.882	0.889	0.877	0.181	0.756*
PredictNLS	1503	1696	5044	61	0.788	0.470	0.988	0.039	0.570
PSORT	1729	1470	4589	516	0.761	0.540	0.899	0.230	0.481
SubLoc	2769	430	4490	615	0.874	0.866	0.880	0.182	0.738
Secretory pathway									
PWMSubLoc	422	173	7485	224	0.952	0.709	0.971	0.347	0.655*
iPSORT	287	308	7438	271	0.930	0.482	0.965	0.486	0.461
PSORT	394	201	7205	504	0.915	0.662	0.935	0.561	0.496
SubLoc	478	117	7319	390	0.939	0.803	0.949	0.449	0.635
TargetP	357	238	7391	318	0.933	0.600	0.959	0.471	0.527
Cytoplasmic									
PWMSubLoc	1876	511	4976	941	0.825	0.786	0.841	0.334	0.599*
PSORT	1146	1241	4841	1076	0.721	0.480	0.818	0.484	0.305
SubLoc	1876	511	4976	941	0.825	0.786	0.841	0.334	0.599*

(*) shows the method(s) at the best correlation (MCC) obtained.

Tabela 5.4: **Performance comparison for the different subcellular localization prediction methods (PWMSubLoc, iPSORT, MitoProtII, Predotar, PSORT, TargetP, PredictNLS) for the DBSubLoc data set.**

Score threshold	Positives tested		Negatives tested		Accuracy	SN	SP	FDR	MCC
	TP	FN	TN	FP					
Chloroplastic									
PWMSubLoc	122	278	7663	37	0.961	0.305	0.995	0.233	0.469*
iPSORT	155	245	6910	790	0.872	0.388	0.897	0.836	0.192
MitoProtII	160	240	7030	670	0.888	0.400	0.913	0.807	0.224
Predotar	186	214	6839	861	0.867	0.465	0.888	0.822	0.228
PSORT	100	300	7331	369	0.917	0.250	0.952	0.787	0.187
TargetP	196	204	6947	753	0.882	0.490	0.902	0.793	0.264
Mitochondrial									
PWMSubLoc	831	828 6	156	285	0.863	0.501	0.956	0.255	0.535*
iPSORT	904	755	5527	914	0.794	0.545	0.858	0.503	0.390
MitoProtII	606	1053	6067	374	0.824	0.365	0.942	0.382	0.380
Predotar	916	743	5543	898	0.797	0.552	0.861	0.495	0.400
PSORT	598	1061	6041	400	0.820	0.360	0.938	0.401	0.366
SubLoc	759	900	5899	542	0.822	0.458	0.916	0.417	0.410
TargetP	897	762	5754	687	0.821	0.541	0.893	0.434	0.442
Nuclear									
PWMSubLoc	3341	733	2928	1098	0.774	0.820	0.727	0.247	0.550*
PredictNLS	1126	2948	3905	121	0.621	0.276	0.970	0.097	0.341
PSORT	1870	2204	3751	275	0.694	0.459	0.932	0.128	0.443
SubLoc	3030	1044	3102	924	0.757	0.744	0.770	0.234	0.514
Secretory pathway									
PWMSubLoc	743	294	6608	455	0.908	0.716	0.936	0.380	0.614*
iPSORT	654	383	6685	378	0.906	0.631	0.946	0.366	0.578
PSORT	684	353	6376	687	0.872	0.660	0.903	0.501	0.501
SubLoc	478	559	6377	686	0.846	0.461	0.903	0.589	0.347
TargetP	675	362	6558	505	0.893	0.651	0.929	0.428	0.549
Cytoplasmic									
PWMSubLoc	484	446	5973	1197	0.797	0.520	0.833	0.712	0.278*
PSORT	389	541	5842	1328	0.769	0.418	0.815	0.773	0.182
SubLoc	484	446	5973	1197	0.797	0.520	0.833	0.712	0.278*

(*) shows the method(s) at the best correlation (MCC) obtained.

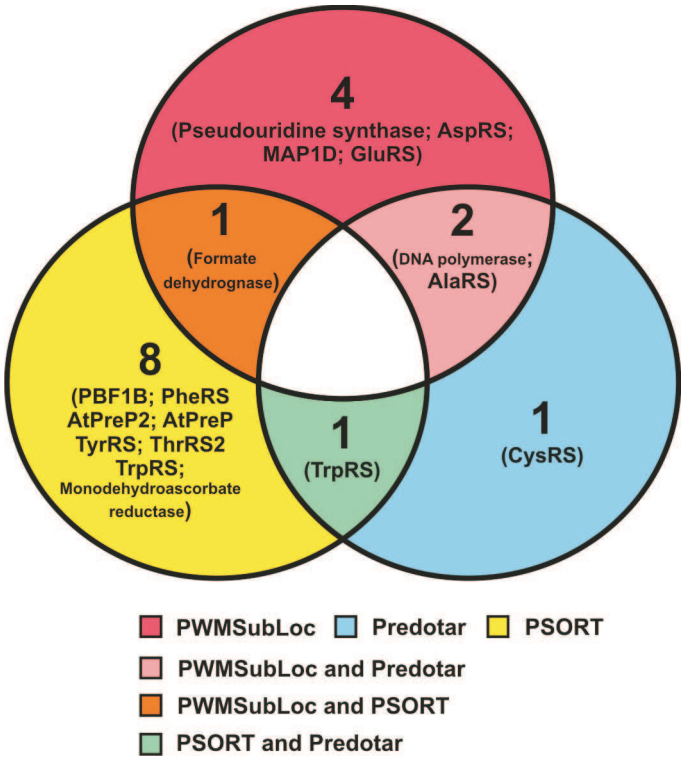


Figura 5.2: **Venn diagram illustrating the low overlap in the PWMSubLoc, Predotar and PSORT prediction for the 35 dual targeted *Arabidopsis* proteins.** The numbers inside the Venn diagram represent the protein set in each group, the protein name being indicated between brackets. The blank group indicates a class without prediction.

experimental methods. Additional new PWMs models derivable from predicted sequences confirmed by subcellular proteomic studies will be included in future versions. Future work will also include the ability to predict other subcellular localizations by incorporating new predictors.

Conclusions

Several prediction programs have emerged showing different strengths due to the different types of data used. When more than one of these is used on the same data, they may produce conflicting predictions (Liu *et al.*, 2007). This question may be resolved if a predictor can be developed with predicting performance exceeding that of many individual element predictors. The purpose of this study was to enhance prediction accuracy by integrating the available subcellular localization prediction tools. In this work, a novel method for predicting eukaryotic protein subcellular localization exclusively from its previous prediction is presented. This new approach provides superior prediction performance compared with the other popular methods. A PWMSubLoc online service has been developed for predicting protein subcellular localizations based on the PWMs described in this work.

Availability and requirements

- Project name: PWMSubLoc
- Project home page: http://ipe.cbmeg.unicamp.br/pub/PWMSubLoc
- Operating systems(s): Platform independent
- Programming language: Perl and R
- License: under the GNU General Public License

Authors' contributions

RV conceived the study, conducted the work and drafted the manuscript. MM participated in coordinating the study and helped draft the manuscript. All the authors read and approved the manuscript.

Acknowledgements

RV was supported by a fellowship from UNIEMP Institute and MM received a research fellowship from CNPq. This work was partially supported by grant 05/58104-0 from FAPESP to MM.

Era idílico, sim. Naquele tempo, lembro-me, você podia ser um cientista sem sentir-se culpado; podia ainda acreditar que estava trabalhando para maior glória de Deus. Hoje em dia não lhe permitem ao menos o conforto de enganar-se a si mesmo ... Nem por um só momento lhe consentem que esqueça os seus reais objetivos. Ad majorem Dei gloriam? Não seja idiota! Ad majorem hominis degradationem - é para o que o você trabalha.

Aldous Huxley (1894-1975)

6

The predicted subcellular localization of the sugarcane proteome

Renato Vicentini[1] and Marcelo Menossi[1]

Abstract

Plant cells are highly organized, and many biological processes are associated with specialized subcellular structures. Subcellular localization is a key feature of proteins, since it is related to biological function. The subcellular localization of these proteins can be predicted, providing information that is particularly relevant to those on proteins with unknown or putative function. We performed the first *in silico* genome-wide subcellular localization analysis for the sugarcane transcriptome (with 11,882 predicted proteins) and found that most of the proteins are localized to four compartments: nucleus (44%), cytosol (19%), mitochondria (12%), and secretory destination (11%). We also show that about 19% of the proteins are localized to multiple compartments. Other results were able to identify a potential set of sugarcane proteins that can show dual targeting by use of N-truncated forms that start from the nearest downstream in-frame AUG codons. This study is a first step to increase the knowledge of the subcellular localization of sugarcane proteome.

[1]Departamento de Genética e Evolução, Laboratório de Genoma Funcional, Instituto de Biologia, CP 6109, Universidade Estadual de Campinas - UNICAMP, 13083-970, Campinas, SP, Brazil

Introduction

The eukaryotic cell is highly organized, and various biological processes are associated with specialized subcellular structures, or confined to particular compartments (Shen and Burger, 2007). Membrane biology has to be adapted to specific requirements, such as in the chloroplasts, to organize the phosynthetic pathways, the formation of organelles for specific storage, to permit changes in vacuolar content and in membrane composition in response to stress (Moreau *et al.*, 2007). Assigning a subcellular location to a protein is highly desirable to help elucidate how proteins are spatially organized according to their function (Xie *et al.*, 2005), to refine our knowledge on cellular processes by certain activities to specific organelles (Lilley and Dupree, 2007). As a result, the knowledge of subcellular localization of proteins provides clues to their function as well as the interconnectivity of biological process (Shen and Burger, 2007). Changes in localization may result from cell signaling events, environmental changes and progressing through the cell cycle (O'Rourke *et al.*, 2005).

Finally, protein localization has important implications regarding other proteins with which it interacts. Many studies reveal a very large numbers of protein-protein interactions, but the experimental evidence has shown that many of the interactions turn out to be false positives (O'Rourke *et al.*, 2005). One way to eliminate false positives is to determine whether the two proteins reside in the same cellular structures (O'Rourke *et al.*, 2005), increasing the confidence in the interactions in those cases where the proteins co-localize to the same organelle.

Signals for protein sorting exist either in the form of primary sequences, usually N-terminal targeting sequences (Rusch and Kendall, 1995, Emanuelsson *et al.*, 2000), or in internal sequence motifs (Cokol *et al.*, 2000). Proteins localized in the same organelle have been reported to show a similar overall amino acid composition and are thought to have evolved to function optimally in that specific environment (Andrade *et al.*, 1998). When the final destination is the mitochondria, the chloroplast, or the secretory pathway, sorting usually relies on the presence of an N-terminal targeting sequence (von Heijne *et al.*, 1989, Hiller *et al.*, 2004). These signal peptides are responsible for targeting proteins to the ER for subsequent transport through the secretory pathway. There are known cases of variation in the use of alternative signal peptides, and in the majority of cases this is due to the exclusion of the signal peptide from protein products of the same gene.

Proteins may simultaneously exist at, or move between, two or more different

subcellular localizations. Proteins with multiple locations are particularly interesting, because they may have special biological functions. The dual targeting of proteins to both mitochondria and chloroplast was originally expected to be rare, but the number of proteins that have been shown to have dual targets has greatly increased (Millar *et al.*, 2006). Recent studies shows an unexpectedly frequency of dual-targeting proteins, and novel routes of protein trafficking. Such finding makes it more difficult to predict which proteins really are targeted to organelles (Millar *et al.*, 2006).

Experimental approaches were be used for subcellular determinations. For example, many hundreds of *Arabidopsis* proteins had their subcellular localization determined by visualization of their expression with fluorescent proteins fusion (Heazlewood *et al.*, 2007). The same occurred in rice, for which are available many reference maps based on two-dimensional polyacrylamide gel electrophoresis of proteins from subcellular compartments (Komatsu *et al.*, 2004).

The prediction of protein subcellular localization focuses on determining localization sites of unknown proteins in a cell. Given the size and complexity of the genomic data, prediction systems are interesting approaches to identify and screen possible candidates for further analysis. Compared with experimental methods, computational prediction methods that can provide fast and accurate assignment of protein subcellular localization are very desirable (Xie *et al.*, 2005).

The tropical sugarcane (*Saccharum* spp.) is an important industrial crop and is cultivated on close to 20 million hectares in more than 90 countries (FAO; http://apps.fao.org). Increasing interest in sugarcane biology has been boosted due to the growing interest in ethanol, a green fuel (Pessoa-Jr *et al.*, 2005). Sugarcane belongs to the grass family (*Poaceae*), an economically important seed plant family that includes cereals such as maize, wheat, rice, and sorghum as well as many forage crops. The main product of sugarcane is sucrose, which accumulates in the stalk internodes, contributing to about two thirds of the world's raw sugar production. With the aim of expediting sugarcane genomics, several sugarcane ESTs collections have been developed (Carson and Botha, 2002, Vettore *et al.*, 2003, Ma *et al.*, 2004, Casu *et al.*, 2004, Bower *et al.*, 2005), the challenge now is to attribute relevant biological information to these data.

To contribute to our understanding of the distribution of proteome in sugarcane, we undertook an *in silico* genome wide subcellular localization analysis. In this paper, we describe a general set of predicted sugarcane protein and show the subcellular localization

predicted for these proteins. We also performed an analysis of putative dual targeting proteins, obtained by the N-truncated forms, and an attempt to create sugarcane predicted protein interaction map, showing the location of all proteins.

Materials and methods

Determining sugarcane complete coding sequences

The sugarcane (*Saccharum* spp.) EST sequences were from the SUCEST project, which has been described previously (Vettore *et al.*, 2003). These sequences represents 43.141 SASs (sugarcane assembled sequences), which were estimated to represent over 30,000 unique genes. The first step of the analysis was to determine the putative ORFs (open reading frame) for each SAS. To predict the ORFs we used three prediction programs: 1. GeneMark.SPL (Borodovsky and McIninch, 1993), a variant of GeneMark.hmm used to analyze ESTs and cDNAs data with improved prediction of gene boundaries; 2. GENSCAN (Burge and Karlin, 1997), which captures the general and specific compositional properties of the distinct functional units of eukaryotic genes and 3. ESTScan (Iseli *et al.*, 1999), particularly useful to solve problems caused by frameshift or stop codon errors. The next step of the analysis was to estimate the set of ORFs that represents a sugarcane complete CDSs (coding sequences). Those that start with ATG and end with stop codon were selected for the SCCDS set 1 (a total of 11,882 amino acid sequences, see additional data file 1). When exist more of one ORF for the same SAS, it was selected that with best match against NCBI protein database (for known proteins) or that represent the longest protein sequence (for unknown proteins). Finally, to obtain a more accurate sugarcane protein data set, that can be very useful in experimental analysis, we performed a final comparison against NCBI protein database. These results were able to indicate proteins that had evidences that the first predicted methione encoded the initial amino acid of protein, and shows the same protein hit in the SAS blastx result (SCCDS set 2, with 1,519 amino acid sequences).

Predicting protein subcelular localization

We recently developed a new method, called PWMSubLoc, for subcellular localization protein prediction (manuscript submitted). Following the information theory, this

approach combines the complementary strengths of many existing prediction methods. The PWMSubLoc method can be described as the use of previous prediction profiles to better exploit their potential and improve the prediction of protein subcellular localization. The benchmarking results showed that the PWMSubLoc provides superior prediction performance compared with the other popular methods. The overall prediction accuracy was 90% for the GS data set (Guo *et al.*, 2004*a*), and 86% for the DBSubLoc data set (Guo *et al.*, 2004*b*).

In this study, seven prediction tools were selected for subcellular localization: iPSORT (Bannai *et al.*, 2002), MitoProtII (Claros and Vincens, 1996), Predotar (Small *et al.*, 2004), PSORT (Nakai and Horton, 1999), SubLoc (Hua and Sun, 2001), TargetP (Emanuelsson *et al.*, 2000) and PredictNLS (Cokol *et al.*, 2000). The selection was based on the diversity of the algorithms and the sequence features they employed. These tools were used as base-level classifiers, whose prediction results were combined to build the new classifier.

We performed the prediction of subcellular localization of sugarcane proteome using this method. Currently, the PWMSubLoc method is comprised of five position weight matrix (PWM) models (one for each subcellular localization: mitochondria, plastid, secretory pathway, nucleus, and cytoplasm). The method begins by calculating a weight matrix from the frequencies of each prediction (localization) for each predictor (software) for all the sequences that have the same known subcellular localization in training set. This procedure is applied to all the subcellular localizations. These matrixes are then applied to the prediction profile to determine the more probable subcellular localization of each individual protein. The PWM is computed using the widely accepted information theoretical approach with modifications (Schneider, 1997, Vicentini and Menossi, 2007).

We classify the sugarcane proteins in eight subcellular localizations (plastid, mitochondria, nucleus, cytoplasm, secreted proteins, vacuole, endoplasmic reticulum, and Golgi complex). For protein prediction to vacuole, endoplasmic reticulum, and Golgi complex, we used the PSORT prediction program (Nakai and Horton, 1999), since this is only one able to perform these predictions. For subcellular localization protein prediction with PWMSubLoc, we use followings information content cutoff values: 3.4 bits for plastid proteins, 6.7 bits for mitochondria, 2.1 bits for nuclear proteins, 2.9 bits for proteins to secretory pathway, and 2.3 bits for cytoplasmic proteins. In the text below when we refer to "Secretory pathway"prediction, it means that the prediction was one of these: vacuole, endoplasmic reticulum, Golgi complex, or secretory destination. The same occurs for "Others"prediction, which means that the prediction was not to mitochondria, plastid or "Secretory pathway".

Subcellular localization of full and N-truncated proteins

The prediction of subcellular localization of full and N-truncated proteins was evaluated by the TargetP prediction program (Emanuelsson *et al.*, 2000). A second set comprising 4.887 sugarcane CDSs, that start with the nearest AUG codon in the same frame of the first AUG was also used in this analysis.

Analysis of overrepresented biological processes of sugarcane proteins

Significantly overrepresented biological processes of proteins localized to different subcellular location were evaluated using the GeneMerge 1.2 software (Castillo-Davis and Hartl, 2003), and ontology file from the GO Consortium (Ashburner *et al.*, 2000), and annotations from Gene Ontology Annotation (GOA) Database (Camon *et al.*, 2004). A Bonferroni corrected *P* value (e-score) of 0.01 was used as threshold for measuring significance.

Predicting protein-protein interaction

The interactions map of subcellular localization sugarcane proteins was generated using the STRING tool (von Mering *et al.*, 2007) and the Medusa Network Viewer software (Hooper and Bork, 2005). The data set was created by selecting all sugarcane protein pairs that are involved in protein-protein interactions annotated in the *Arabidopsis* dataset of the STRING database (von Mering *et al.*, 2007). The nodes were colored according to their subcellular localization predictions, and the confidence limits were set to medium confidence (50%).

Results

Sugarcane predicted protein sequences

The strategy adopted to predict complete CDSs from sugarcane generated two data sets (Figure 6.1). The SCCDS set 1 is composed by 11,882 amino acid sequences (see additional data file 1) most of which without homology with known proteins in public databases. Another data set, SCCDS set 2, contains 1,519 complete amino acid sequences and is a more accurate sugarcane protein set that can be useful in experimental analysis. All protein

in this set shows full coverage, with various similarities thresholds, with a known protein in public dataset, but not necessarily with known function.

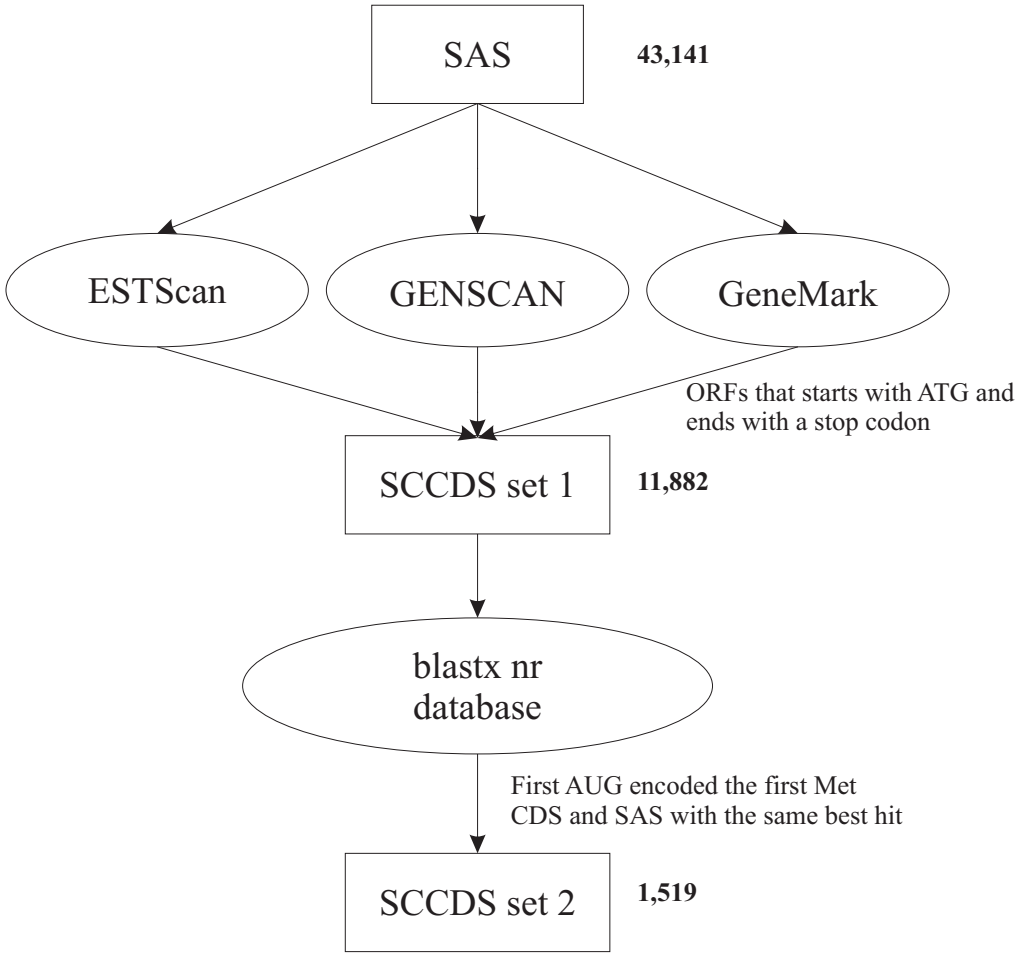


Figura 6.1: **Scheme for determination of the complete coding sequences for sugarcane transcriptome.**

To evaluate if SCCDS set 1 is not composed by truncated protein sequences, we compared the polypeptides length of this set with protein data sets of others three plant species, and with the SCCDS set 2. The *Arabidopsis* protein set was downloaded from The Arabidopsis Information Resource (TAIR) web site (http://www.arabidopsis.org), the rice set was downloaded from Rice Annotation Project (RAP) web site (http://rapdb.dna.affrc.go.jp, second version), and maize protein set was downloaded from PlantGDB (http://www.plantgdb.org). The density plot of the proteins length shows clearly that the SCCDS set 1 do not show a strong pattern for shortest proteins (Figure 6.2). In fact, the rice set and the two

set from sugarcane show a similar profile for protein length. The maize and *Arabidopsis* sets contains less short proteins (< 300 amino acids).

Sugarcane predicted subcellular localization proteome

To date, most proteomic studies in plants have focused on *Arabidopsis thaliana*. Proteomic and fluorescent protein-labeling studies have revealed the locations of over 4,000 proteins in *Arabidopsis thaliana* (Lunn, 2007), but the whole proteome of this species is estimated to contain another 12,000-24,000 proteins whose intracellular locations are unknown (Heazlewood *et al.*, 2004). To lackey this huge problem, various computational approaches have been developed to predict intracellular location from the primary protein sequence, and recent studies had showed that the combining results from several prediction programs improve specificity of prediction. Each of the localization prediction tools available shows different strengths, and no tool is clearly and globally optimal (Shen and Burger, 2007). Any given tool will perform well on some data but poorly in others, and often predictions disagree between different tools (Millar, 2004).

We perform the prediction of subcellular localization in the two sugarcane data sets (SCCDS set 1 and 2). Initially we utilized the seven distinct programs described above, and how expected, their predictions often disagree for a given protein (Figures 6.3A, 6.4A and 6.5A, for SCCDS set 1). Another limitations about this strategy, is that the vast majority of existing methods for predicting protein subcellular localization is limited to a single location site, despite the knowledge that the protein may exist simultaneously at two or more different subcellular localizations. To enhance the accuracy of our analysis we utilized the PWMSubLoc for predict subcellular localization for proteins where exist more of one site prediction (mitochondria, plastid, cytoplasm, nucleus, and secretory pathway) by the different programs. The result, for SCCDS set1, is shows in Figures 6.3B, 6.4B and 6.5B. These results represents a more realistic view of the sugarcane protein distribution in the cell, mainly because with the proper exploitation of the combined strengths of these prediction programs, it is possible to obtain a prediction whose performances surpass those of all existing prediction programs (Liu *et al.*, 2007).

In the Figure 6.6 we shows a cell schematic diagram of the putative subcellular localizations of sugarcane proteins. This analyze showed that 44% of the sugarcane proteins were localized in nucleus, 19% in cytosol, 12% in mitochondria, 18% in secretory pathway,

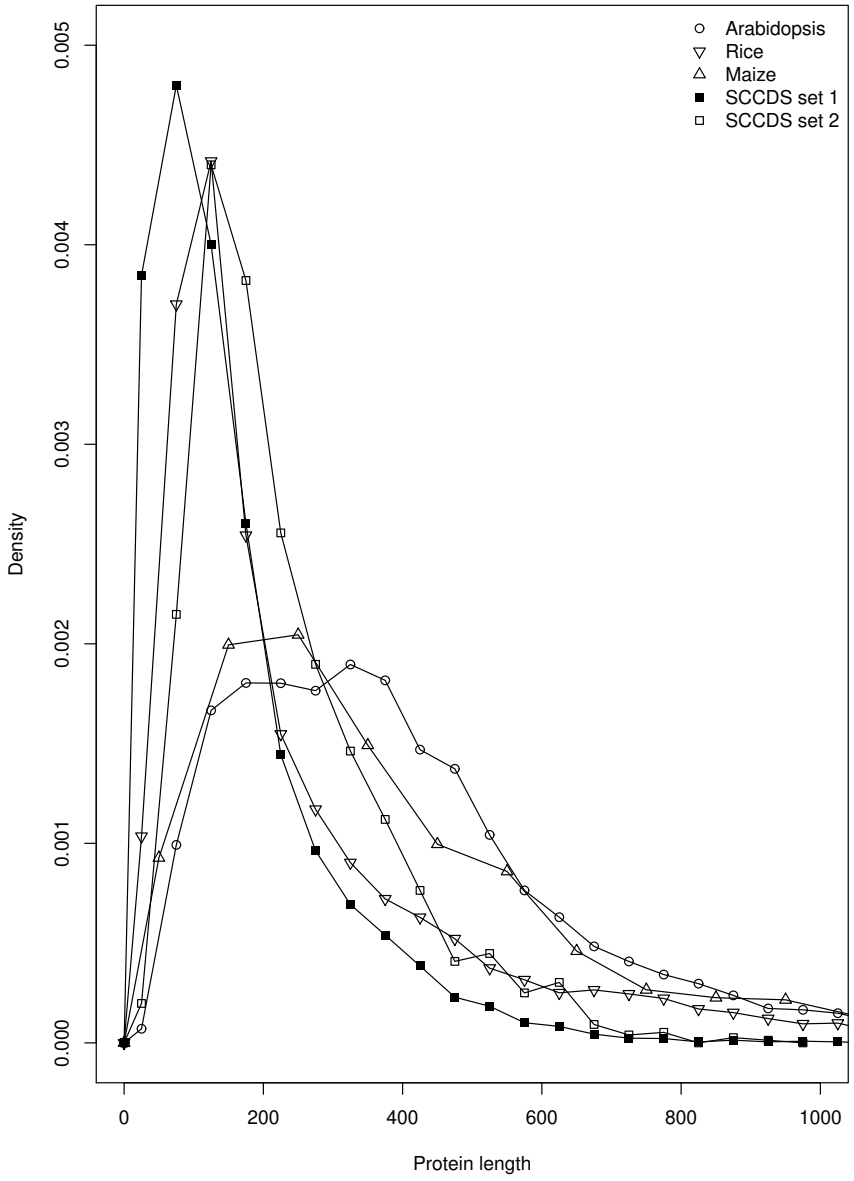


Figura 6.2: **Density plot of the protein length in four plant species data sets.** The plot shows the density estimates of the protein length for *Arabidopsis*, rice, maize, and for the two sugarcane data sets. *Arabidopsis* n = 32.009; Rice n = 22.006, Maize n = 2.782, SCCDS set 1 n = 11.881, and SCCDS set 2 n = 1.519.

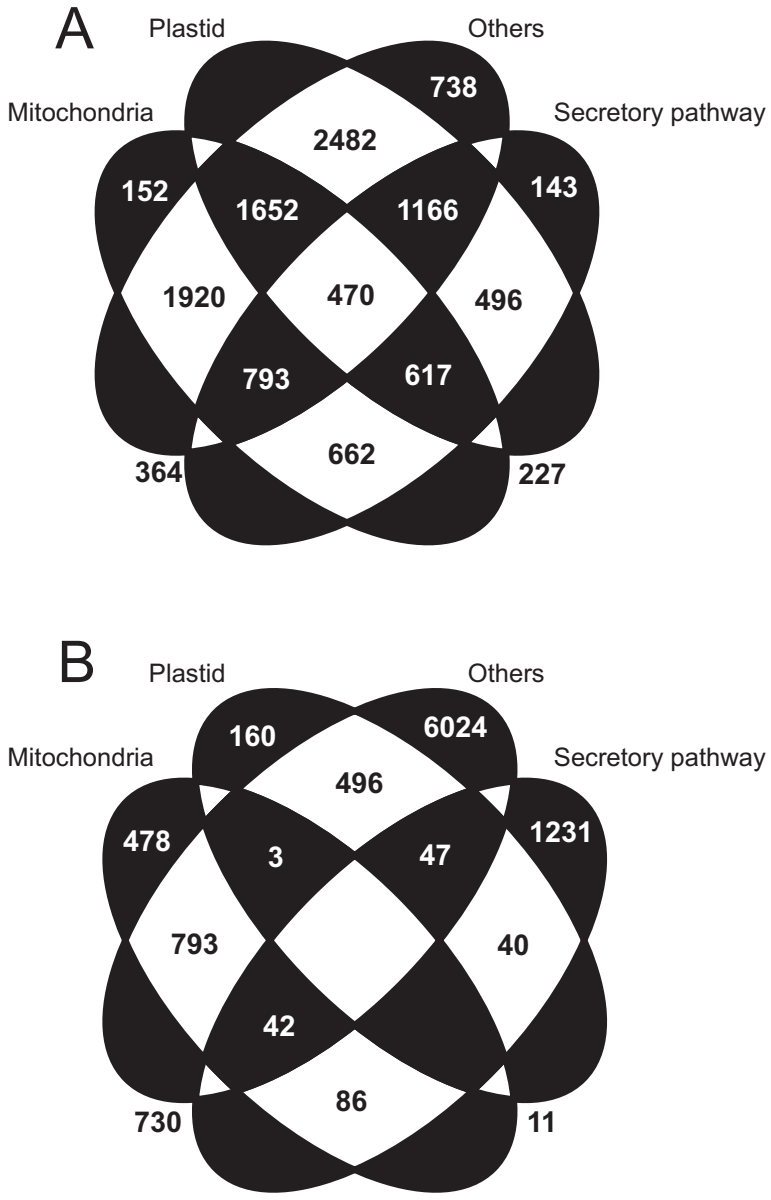


Figura 6.3: **Venn diagram illustrating the overlap of Mitochondria, Plastid, "Secretory Pathway"and "Others"predictions of the sugarcane proteins.** These diagrams clearly show a putative set of multi-targeted sugarcane proteins. (A) All predictions and (B) selected localization by PWMSubLoc prediction method. Groups without numbers indicate classes without prediction.

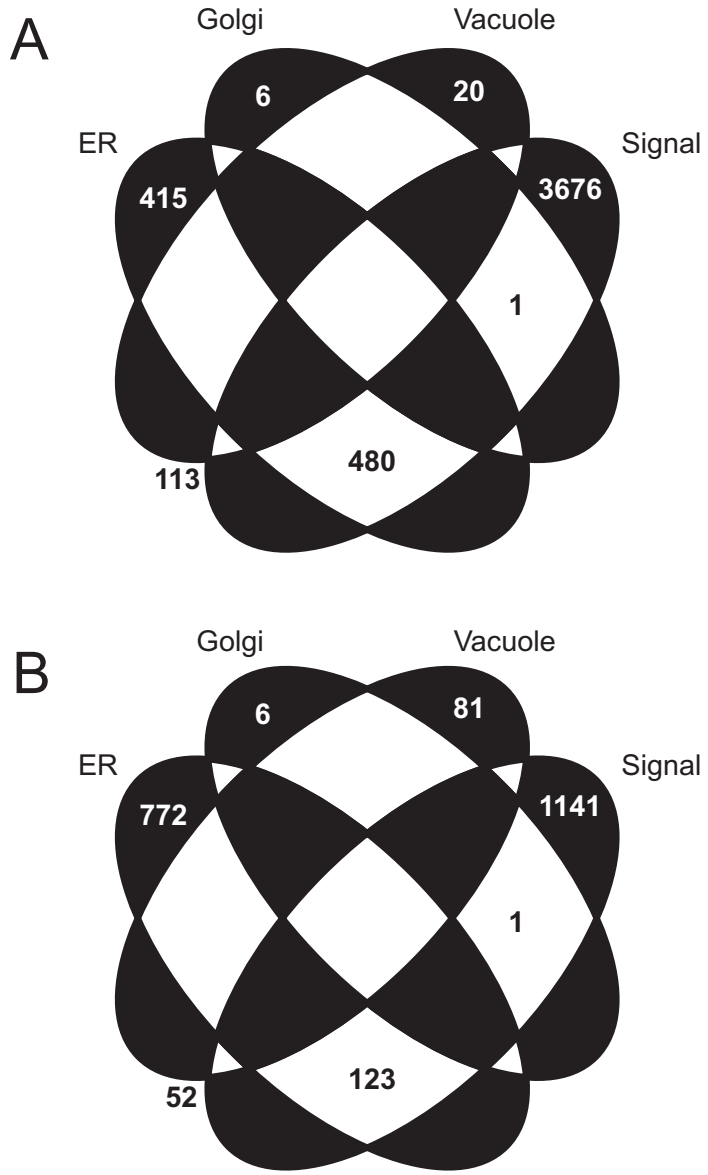


Figura 6.4: **Venn diagram illustrating the prediction of subcellular localization of expanded groups "Secretory Pathway"(Golgi complex, Endoplasmic Reticulum, Vacuole and Signal).** (A) All predictions and (B) selected localization by PWMSubLoc prediction method. Groups without numbers indicate classes without prediction.

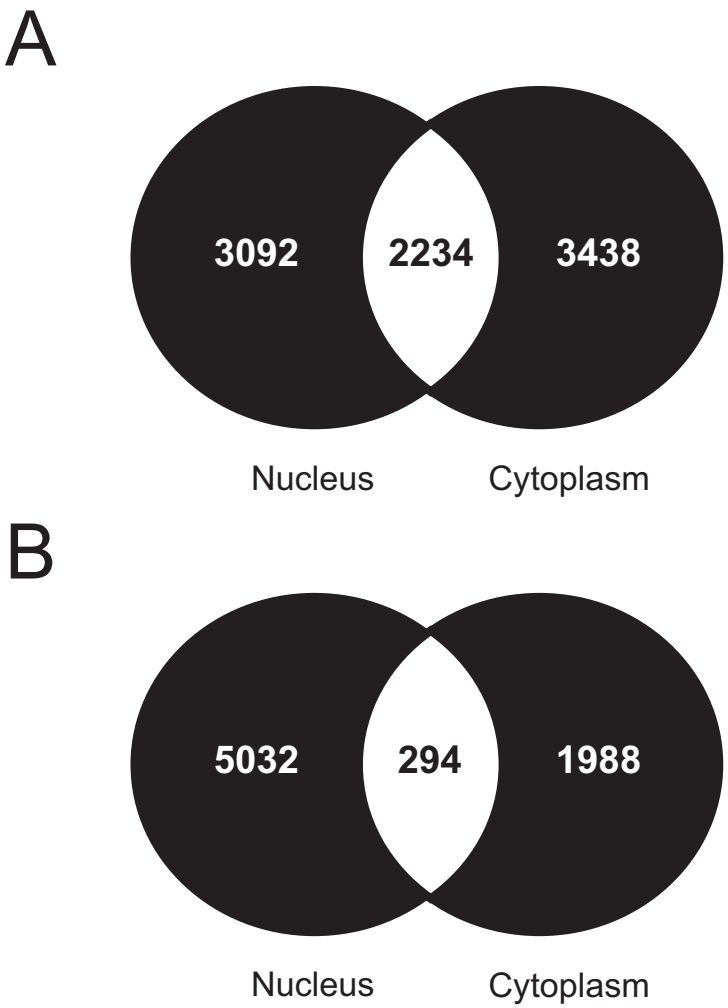


Figura 6.5: **Venn diagram illustrating the prediction of subcellular localization of expanded groups "Others"(Nucleus, and Cytoplasm).** (A) All predictions and (B) selected localization by PWMSubLoc prediction method. Groups without numbers indicate classes without prediction.

and 6% in plastid. Of these secretory pathway proteins, we identify that 60% were secreted proteins, 41% were endoplasmatic reticulum, and 6% were vacuole proteins. We found that a significant number of sugarcane proteins (19%) localize to more than one compartment, similar to the results from the *Arabidopsis* proteome (Millar *et al.*, 2006). The most commonly occurring paired compartments are nucleus and cytoplasm (13% of the proteins), that was the same highly paired compartments identified in *Arabidopsis* proteome (Millar *et al.*, 2006). Which suggests that there is a significant amount of interaction between these organelles. These proteins may be involved in process such as electron transport or energy pathways, and response to stress (Geisler-Lee *et al.*, 2007). Figure 6.7 shows the overrepresented biological processes of sugarcane plastidial (Figura 6.7A), and mitochondrial (Figura 6.7B) proteins. Note that much ontology classifies proteins of both organelles, and some specify biological process that is attributed to proteins that have specific localization in one of these organelles (e.g. photosynthesis and response to oxidative stress).

Subcellular targeting of putative N-truncated sugarcane proteins

A possible source of new protein forms is translational polymorphism where several AUG codons within mRNAs may serve as alternative translation start sites (TISs) to produce proteins with overlapping sequences and displaying different properties (Kochetov and Sarai, 2004, Vicentini and Menossi, 2007, Christensen *et al.*, 2005). We compared predicted subcellular localizations of complete sugarcane proteins and their potential N-terminally truncated forms started from the nearest downstream in frame AUG codon. The results of prediction are shown in Table 6.1. One can see that N-truncated forms of many mitochondrial, plastid or secreted proteins lose their targets (18%). It was expected since N-truncated polypeptides could lose their signal peptides. Fifteen percent of N-truncated proteins retained their localizations. It was found that localizations of full and N-truncated proteins differ in many cases: 10.3% of N-truncated proteins acquired sorting signals *de novo* and 6.5% changed their predicted subcellular localization. Very similar results were obtained by Kochetov and Sarai (2004) for *Arabidopsis thaliana*. A complete list of all sequences used in this study, with description of subcellular prediction for full and N-truncated form, is provided in Additional data file 1.

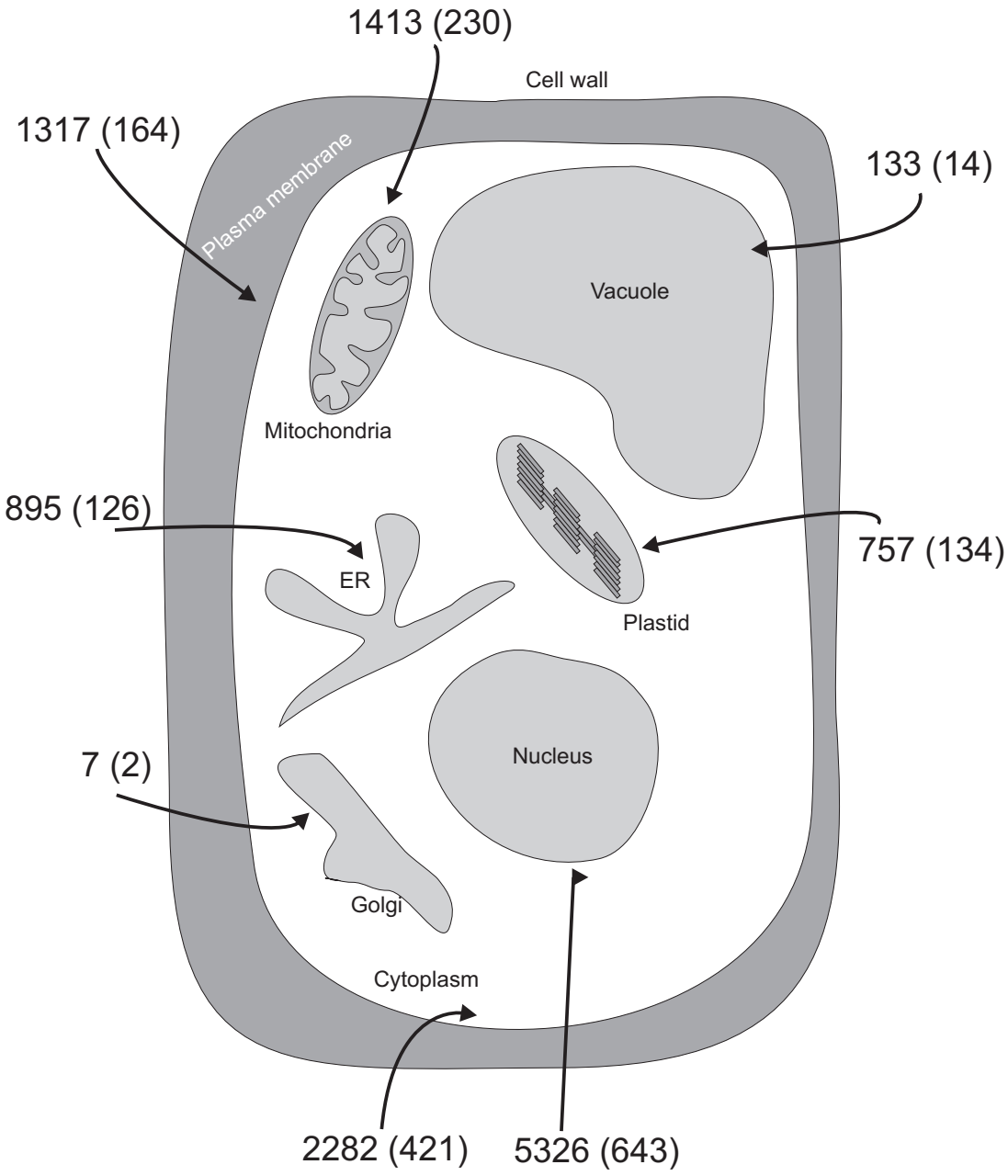


Figura 6.6: **Schematic diagram showing the putative subcellular localizations of sugarcane proteins.** Numbers indicate the total of predict proteins in each location (SCCDS set 1). The numbers between brackets represents the total of predict proteins for the SCCDS set 2.

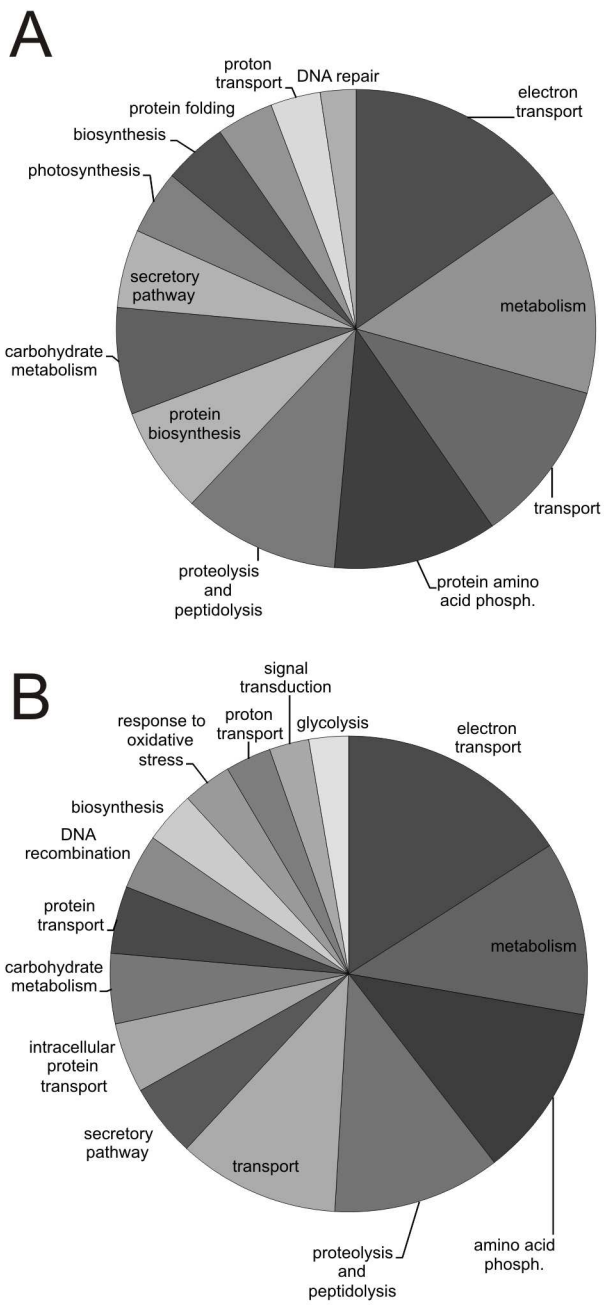


Figura 6.7: **Overrepresented ontologies of biological process for the sugarcane proteins that were predicted for theses two subcellular localizations.** (A) Plastid, and (B) Mitochondria.

Tabela 6.1: **Subcellular localization of full and putative N-truncated sugarcane proteins (%) predicted with TargetP program.**

	Full	N-truncated			
Location	**Size of fraction**	**mTP**	**cTP**	**SP**	**Others**
mTP	16,1	6,4	0,9	1,2	7,6
cTP	9,0	1,7	2,8	0,2	4,2
SP	15,0	1,5	1,0	5,9	6,6
Others	59,9	4,7	2,2	3,4	49,5
Total	100,0	14,3	7,0	10,8	68,0

Predicted protein-protein interaction in sugarcane

In a cell, interacting proteins are significantly more likely to be found within the same subcellular location, although some proteins will interact across adjacent subcellular locations (Geisler-Lee *et al.*, 2007). Some of the know Golgi/ER and Golgi/vacuole protein interaction can be attributed to real interactions between members of complexes involved in the endomembrane trafficking pathway (Geisler-Lee *et al.*, 2007). To investigate this topic, we performed an *in silico* prediction analysis for sugarcane protein-protein interaction, using *Arabidopsis* data to established the link between these interaction. Unfortunately the result of the analysis was not be able to show a clearly pattern for interaction between proteins in the same subcellular location (with exception for nucleus) or in adjacent subcellular locations. However, the protein-protein interaction map obtained (Figure 6.8) shows a preference for interaction between nuclear proteins (blue squares in the figure), and higher number of interaction between cytoplasmatic proteins (yellow squares) and organellar proteins.

Discussion

Experimental approaches were be used for subcellular determinations. Such approaches have been extensively undertaken, with a number of studies producing significant protein sets from major locations such as the plastids, the nucleus, the plasma membrane and the mitochondrion. Many proteins change their localization in the cell depending on their state of activation (O'Rourke *et al.*, 2005). Translocation of proteins from the cytosol to the plasma membrane and other internal membranes is also critical for many signaling events

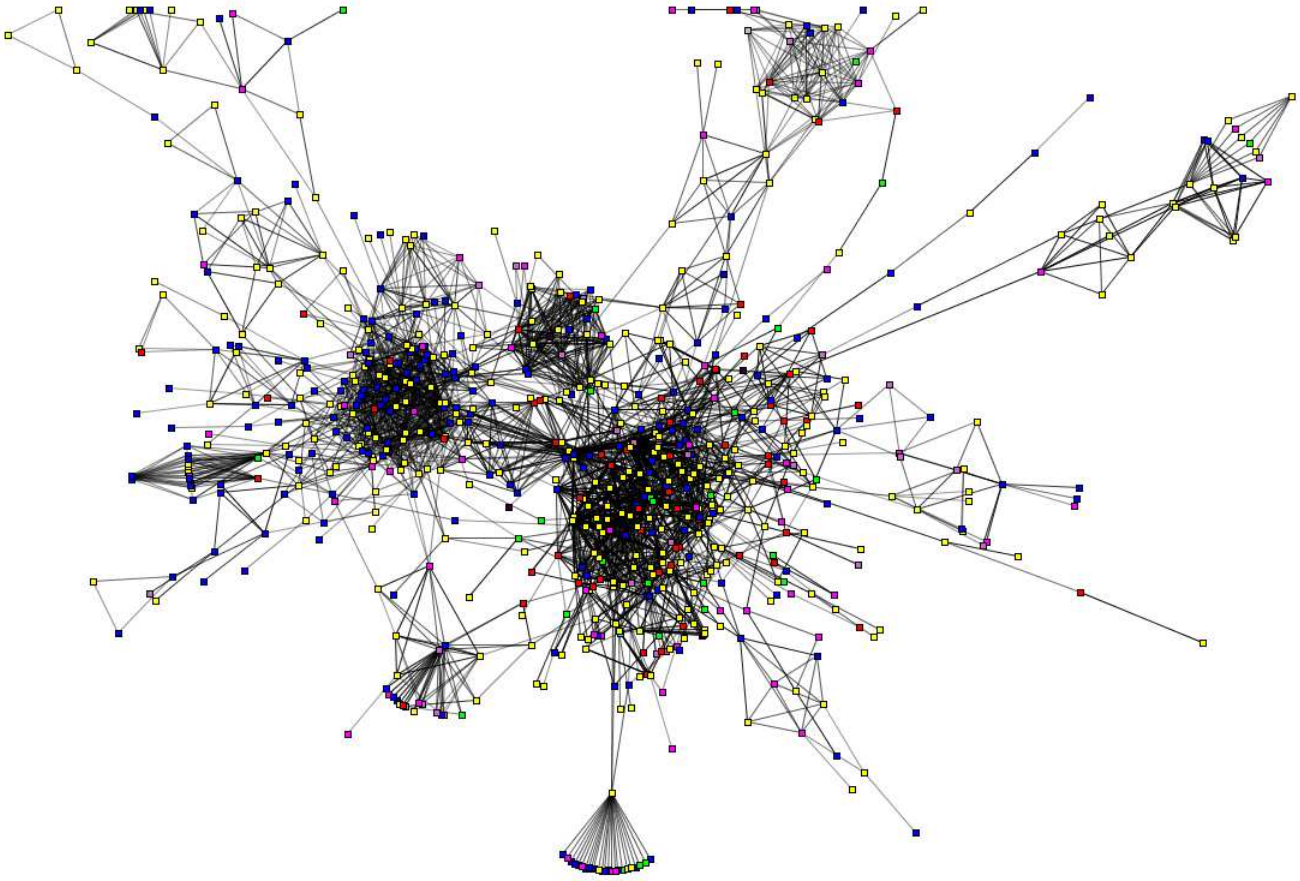


Figura 6.8: **Predicted protein-protein interaction map for subcellular localization sugarcane proteins.** Proteins are depicted as squares colored according to their predicted compartments localization (Nucleus in blue, Plastid in green, Mitochondria in red, Cytoplasmic in yellow, Secretory pathway in purple, and Endoplasmic Reticulum in pink). Interactions are shown as lines between proteins. No interactions were found for Golgi complex and Vacuolar proteins.

(O'Rourke *et al.*, 2005).

Targeting prediction programs are mostly used to determine the likelihood that a specific protein sequence is targeted to a particular subcellular structure. Usually only one, or two programs are used one at a time (Heazlewood *et al.*, 2007). Comparisons of the prediction sets obtained from each predictor program indicate that the consensus between programs are much smaller (3% of all *Arabidopsis* proteins), indicating there are relatively large non-overlapping sets of positives within these predicted sets. However, as PWMSubLoc utilize predicted profiles, the set of sugarcane proteins predicted to a specific location by multiple targeting prediction programs can be rapidly compared.

The general distribution of protein subcellular localization in sugarcane is look like to the distribution of the *Arabidopsis* proteome that has been experimentally verified. In *Arabidopsis*, the 1300 experimentally verified proteins were distributed among many different compartments, with most of the proteins localized to four compartments: mitochondria, nucleus, plastid, and cytosol. About 19% of the proteins are found in multiple compartments, in which a high proportion is localized to both cytosol and nucleus (Millar *et al.*, 2006).

In recent studies numerous genes have been identified which contain two or more in frame ATGs at the 5' end, any of which might correspond to the initiation codon used to initiate translation of catalytically active protein (Small *et al.*, 1998, Christensen *et al.*, 2005). In the present study we identify that 41% of putative sugarcane proteins had an in frame ATGs at the 5' end. If the sequence between two AUGs encodes an organelle targeting sequence (many of which are N-terminal), then the protein initiated from the upstream sequence will contain this targeting information whereas the protein initiated from the downstream one will not. This will naturally lead to differential targeting of the two proteins. It is likely that translational polymorphism may be considered as an important source of subcellular proteomes and should be taken into account in further investigations (Kochetov and Sarai, 2004, Kochetov, 2005). In this study four different variants were considered: (1) N-truncated protein may lose its sorting signal (with 902 sugarcane proteins); (2) full and N-truncated proteins may be targeted to the same compartment and have the same function (with 739 sugarcane proteins); (3) full and N-truncated proteins of the same function may be target to different locations (with 321 sugarcane proteins); (4) full protein may have no working targeting signal, whereas its N-truncated form may have an active targeting signal (with 504 sugarcane proteins). These results demonstrate that this phenomenon in sugarcane deserves attention, and further investigation may be shows a high source of translational polymorphism

in sugarcane, which a change of organelle proteome content.

The aim of this study was give the first step to increase the scientific knowledge about the subcellular localization of sugarcane proteome. This is a preliminary view of this plant molecular cell biology question, and further experimental investigations are necessary to attribute another relevant biological information to these data.

Acknowledgements

RV was supported by a fellowship from UNIEMP Institute and MM received a research fellowship from CNPq. This work was partially supported by grant 05/58104-0 from FAPESP to MM.

Supplementary material

Additional data file 1. Listed are the all protein sequences, with subcellular localization description of full and N-truncated form, in FASTA format.

MÁ DOBA PRIJDE.
Gregor Mendel (1822-1884)

7

Characterization of a sugarcane (*Saccharum* spp.) gene homolog to brassinosteroid insensitive1-associated receptor kinase 1 that is associated to sugar content‡

Renato Vicentini[1,*], Juliana de Maria Felix[1,*], Marcelo Carnier Dornelas[2] and Marcelo Menossi[1]

Abstract

The present paper reports on the characterization of *ScBAK1*, a leucine-rich repeat receptor-like kinase from sugarcane (*Saccharum* spp.), expressed predominantly in bundle-sheath cells of the mature leaf and potentially involved in cellular signaling cascades mediated by high levels of sugar in this organ. In this report it was shown that the *ScBAK1* sequence was similar to the brassinosteroid insensitive1-associated receptor kinase1. The putative cytoplasmatic domain of ScBAK1 contains all the amino acids characteristic of protein kinases, and the

‡submitted

*The first two authors contributed equally to this work.

[1]Departamento de Genética e Evolução, Laboratório de Genoma Funcional, Instituto de Biologia, CP 6109, Universidade Estadual de Campinas - UNICAMP, 13083-970, Campinas, SP, Brazil

[2]Departamento de Fisiologia Vegetal, Instituto de Biologia, CP 6109, Universidade Estadual de Campinas - UNICAMP, 13083-970, Campinas, SP, Brazil

extracellular domain contains five leucine-rich repeats and a putative leucine zipper. Transcripts of *ScBAK1* were almost undetectable in sugarcane roots or in any other sink tissue, but accumulated abundantly in the mature leaves. The *ScBAK1* expression was higher in the higher sugar content individuals from a population segregating for sugar content throughout the growing season. *In situ* hybridization in sugarcane leaves showed that the *ScBAK1* mRNA accumulated at much higher levels in bundle-sheath cells than in mesophyll cells. In addition, using biolistic bombardment of onion epidermal cells, it was shown that ScBAK1-GFP fusions were localized in the plasma membrane as predicted for a receptor kinase. Taken all together, the present data indicate that ScBAK1 might be a receptor involved in the regulation of specific processes in bundle-sheath cells and in sucrose synthesis in mature sugarcane leaves.

Introduction

Sucrose is the major form in which carbohydrates are translocated from the leaves to the rest of the plant to supply carbon and energy for growth and the accumulation of storage reserves. Moreover, it has been recognized that sucrose also acts as a signal compound, affecting a variety of physiological processes such as photosynthesis, source and sink metabolism and defense responses (Gibson, 2005, Koch, 2004, Osuna *et al.*, 2007, Rolland *et al.*, 2002, Smeekens, 2000). The carbon metabolite signaling pathways cross talk with other pathways, including hormonal responses, cell cycle control and nitrogen response systems, amongst others (Halford and Paul, 2003). Due to its unique capacity for storing sucrose in the stems, sugarcane is an interesting model for studies on sugar synthesis, transport and accumulation. The photosynthetic C4 cycle is based on metabolic interactions between two types of leaf cell, the mesophyll cells and the bundle-sheath cells. Proper functioning of C4 photosynthesis is dependent on the correct compartmentation of the enzymes involved. C4 plants such as sugarcane show higher rates of photosynthesis at high light intensities and high temperatures, due to the increased efficiency of the PCR (photosynthetic carbon reduction) cycle (Furbank and Taylor, 1995). In favorable environments, C4 plants outperform C3 plants, making them the most productive crops. However, little is known about the regulatory factors and signal transduction chains that control the differentiation of the mesophyll cells and bundle-sheath cells, and the expression of the genes involved in C4 photosynthesis. Plants use the coordinated action of several small-molecule hormones to grow

and develop optimally in response to a changing environment. Plant polyhydroxylated steroid hormones called brassinosteroids (BRs) are involved in diverse processes such as stem and root elongation, vascular differentiation, male fertility, timing of senescence and flowering, leaf development, and resistance to biotic and abiotic stress (Nemhauser and Chory, 2004). Interestingly, a significant number of the identified BR-responsive genes encode putative transcriptional factors, implying that plant steroids use a complex regulatory mechanism to control gene expression (Li and Jin, 2006). BRs binding to a protein complex that includes the leucine-rich repeat receptor-like protein kinase (LRR-RLK) brassinosteroid-insensitive 1 (BRI1), activates a phosphorylation-mediated signaling cascade that changes the amount, subcellular localization, and/or DNA-binding activity of a family of novel transcription factors (Belkhadir and Chory, 2006,Li and Jin, 2006,Vert *et al.*, 2005). The structure of BRI1 and its localization in the plasma membrane support the hypothesis that BRI1 interacts with an extracellular ligand. A second plasma membrane-localized LRR-RLK called BAK1 (BRI1-associated receptor kinase 1), also known as SERK3 (somatic embryogenesis receptor kinase 3), is a signaling partner of BRI1. BAK1 has a shorter extracellular domain, with only five LRRs. Genetic analysis demonstrated a role for BAK1 in BR signaling, and a direct physical interaction between BRI1 and BAK1 was found both in yeast cells and in *Arabidopsis* plants by co-immuno-precipitation experiments (Li *et al.*, 2002). The current model suggests that BAK1 is a co-receptor and/or downstream target of BRI1 (Morillo and Tax, 2006, Vert *et al.*, 2005). A recent study has suggested other possible functions for BAK1 in BR signaling. BAK1 over-expression dramatically stimulates the endocytosis of BRI1 in a protoplast system, indicating that BAK1 might be crucial for BR signaling by bringing the cell surface receptor into the cytosol (Belkhadir and Chory, 2006, Li and Jin, 2006, Russinova *et al.*, 2004). This study showed that in plant cells the interaction of BRI1 and BAK1 occurred in restricted areas of the membrane, and recycled on internalization by endocytosis (Russinova *et al.*, 2004). Both BRI1 and BAK1 contain a leucine-zipper motif and at least one cysteine pair in their extracellular domains, which are known to mediate protein-protein interactions and could thus be directly involved in the BRI1/BAK1 interaction. Brassinosteroids are also linked to sugar-mediated pathways and the BR response is related to several hormones that participate in sugar signaling (Rognoni *et al.*, 2007). There are several examples of interactions between signaling pathways and sugar (Koch, 1996), and a role for brassinosteroids has been demonstrated in the process of sugar uptake in tomato seedlings (Goetz *et al.*, 2000). In plants, the control of enzymatic activity by sugars and sugar metabolites

has been investigated in detail and the regulation of metabolic pathways highlighted (Smeekens, 1998). The picture that emerges is that of a sugar-responsive regulatory network in which endogenous developmental programs and external stimuli are integrated, resulting in a coordinated metabolic response. The sugar-sensing and signal transduction systems interact closely with pathways responsive to other stimuli like phytohormones and light. Proposed functions include the control of the activity of sugar transporters located in the membrane by the calcium-dependent protein kinases (Ohto and Nakamura, 1995), and the sugar-mediated repression of a gene involved in brassinolide biosynthesis (Smeekens, 1998, Szekeres *et al.*, 1996). In this study, a sugarcane cDNA designated *ScBAK1*, a leucine-rich repeat receptor-like kinase from sugarcane (*Saccharum* spp.) with sequence similarity to the brassinosteroid insensitive1-associated receptor kinase, was characterized. Transcripts of *ScBAK1* accumulated predominantly in the mature leaves, more specifically in the bundle sheath cells, and the protein was located in the plasma membrane, according to the subcellular location of GFP fusions in onion epidermal cells. Interestingly, gene expression analyses in individuals from a population segregating for sugar content, showed that the *ScBAK1* transcript levels were higher in those individuals with higher sugar contents.

Material and methods

Sugarcane plants

Sugarcane F1 plants were obtained from a cross between the pre-commercial cultivars (SP80-180 X SP80-4966) previously described by Garcia *et al.* (2006) and Felix *et al.* (submitted). The population was comprised of 498 individuals that segregated for sugar content according to a normal distribution of frequencies. The seven plants presenting extreme values for high sugar (HS) and low sugar levels (LS) were selected. Mature leaves (Leaf +1), according to Van Dillewijn (1952), were collected from the selected plants 6, 7, 9, 11 and 13 months after planting. The expression profiles were evaluated by the RNA blot method, using RNA from three HS and three LS individuals collected at the 9 months time point. Pooled RNA from the seven HS and LS individuals collected at all five time points were also used in the RNA blots, to detect the expression profiles throughout the growing season. The expression profile of ScBAK1 was also evaluated for six different tissues collected from 12 month old plants: mature leaf, immature leaf, immature internode, root, lateral bud and a mixture

of flowers in different developmental stages, using the same commercial sugarcane varieties used in the SUCEST project (SP87-432 for flowers and SP80-3280 for other tissues, (Vettore *et al.*, 2003)). All plants were field-grown at the CTC (Centro de Tecnologia Canavieira, Piracicaba, São Paulo, Brazil).

cDNA microarray analyses

A cDNA microarray containing 1920 sugarcane genes was used to identify gene expression patterns associated with the sucrose levels in high and low sugar content plants. cDNA array hybridization and the data analysis were performed as described by Papini-Terzi *et al.* (2005) and Rocha *et al.* (2007), and the complete set of genes identified as differentially expressed will be published elsewhere (Felix *et al.*, submitted).

RNA extraction

Sugarcane tissues (2 - 2.5g) were ground to a fine powder in liquid nitrogen, using a pre-cooled mortar and pestle. RNA was isolated using the Trizol reagent (Invitrogen, USA), following the recommended procedure. The RNA samples were quantified in a spectrophotometer and loaded onto 1.5% agarose/formaldehyde gels for a quality inspection. The Trizol manufacturer's recommendations for high polysaccharide content tissues were followed for the internode samples.

RNA blot

Electrophoresis of the total RNA samples (10μg) was carried out on 1.5% formaldehyde-containing agarose gels by standard procedures, and transferred to a nylon filter (Hybond-N+, GE Healthcare, USA). The RZ3020C02 (accession CA156919) EST clone was selected as a probe for RNA blot hybridization. The insert was labeled with the Ready-To-Go kit (Amershan Biosciences, USA) according to the protocol recommended by the manufacturer. Hybridized filters were exposed to imaging plates for 24 h and the digitalized images of the RNA blot hybridization signals detected using the FLA3000-G screen system (Fuji Photo Film, Japan) and quantified using the Image Gauge software v. 3.12 (Fuji Photo Film, Japan).

Phylogenetic analyses

Database searches were performed using the ScBAK1 protein sequence as the query sequence at the GenBank protein database, using the BLAST algorithm (Altschul *et al.*, 1997). Phylogenetic analyses of the full-length amino acid sequences of AtSERK1, 2, 3, 4 and 5 from *Arabidopsis* (Hecht *et al.*, 2001); OsSERK1 (Hu *et al.*, 2005), OsBAK1 (Messing, 2005) and OsBISERK1 (Song *et al.*, 2007) from rice; ZmSERK1 and 2 from maize (Baudino *et al.*, 2001); MtSERK1 from *Medicago truncatula* (Nolan *et al.*, 2003); DcSERK from carrot (Schmidt *et al.*, 1997); SbRLK from sorghum (Annen and Stockhaus, 1999); and ScBAK1 from sugarcane (this study); were performed using the ClustalW (Thompson *et al.*, 1994) program and PHYLIP package (Felsenstein, 1989). The trees were drawn using Tree View (Page, 1996). To build the bootstrapped parsimony tree for these dataset, the SEQBOOT program for bootstrap analysis with 1000 replicates was used, the PROTPARS program used to generate 100 parsimonious trees, and the CONSENSE program to reduce the 100 trees to a single consensus tree and indicate the bootstrap values as numbers on the branches. The ERECTA protein sequence (GenBank protein accession number NP_180201) from *Arabidopsis* (Torii *et al.*, 1996) was used as an outgroup.

In situ hybridization

Digoxigenin labeling of the RNA probes, tissue preparation and the hybridization conditions were performed as described before by Dornelas *et al.* (1999). Samples of sugarcane leaf +1 were collected, cut transversely, and then cut into less than 1cm square pieces. Sample material was fixed under vacuum for 2 h at room temperature in FAA (50% Ethanol, 10% formaldehyde, 5% acetic acid). The template for *ScBAK1* digoxigenin-labeled riboprobes was an 1857 pb fragment from clone RZ3020C02, cloned in the pSPORT vector (Invitrogen, USA). The hybridized sections were viewed after overnight staining and photographed under a Zeiss Axioscope microscope.

Generation of plant expression vectors for the fluorescence analysis

The complete coding sequence of *ScBAK1* (GenBank accession number EU189960) was PCR amplified from the RZ3020C02 (accession CA156919) cDNA clone with the primers ScBAK1-5' (5' CACC**ATG**ATGATATATTCAGAAATAATGAATCT 3') containing a

site necessary for directional cloning (underlined sequence), ScBAK1-3' (5' TCTGCCAGTTGACAACTCAA 3'), and ScBAK1stop-3' (5' TCATCTGCCAGTTGACAACTCAA 3') containing a stop codon (underlined sequence). The PCR products were cloned into the pENTR/D-TOPO vector using the TOPO cloning procedure (Invitrogen, USA). The cloned cDNA were subsequently recombined into the binary destination vectors p2FGW7 (for carboxi-terminal GFP fusion) and p2GWF7 (for amino-terminal GFP fusion) (Karimi *et al.*, 2002) using the Gateway technology (Invitrogen, USA). All constructs were verified by sequencing.

Transient expression of the ScBAK1-GFP protein fusions in onion epidermal cells

Transient expression of the GFP-fusion proteins in onion epidermal cells was performed using 8 mg of each plasmid DNA precipitated onto tungsten particles, and delivered into onion epidermal cells using the biolistic system. The tungsten particles (1.0 mm, Bio-Rad) were coated with DNA according to the manufacturers instructions. The particles were bombarded into the onion epidermal cells using a Biolistic Particle Delivery System (Bio-Rad). After bombardment, the onion peels were incubated on solid MS medium for 24 h at 28 °C in the absence of light. The peels were screened for GFP fluorescence using a Leica DMI4000 fluorescence microscope (Leica, Germany). Excitation of the GFP was performed with a 488-nm laser line, and GFP fluorescence detected by a band-pass 505- to 550-nm filter. The positive control used for the secretory pathway and plasma membrane localization was the pSGFP5/psecGFP construct (Di Sansebastiano *et al.*, 1998). To visualize GFP fluorescence in the plasma membrane, the bombarded peels were plasmolyzed with 1 M mannitol. A 10x objective (numerical aperture 1.3) was used for scanning. Digital images were captured using the Leica IM50 software (Leica, Germany).

Results

ScBAK1 expression profile

Microarrays containing 1920 ESTs encoding signal transduction-related components, transcription factors and stress-related elements (Papini-Terzi *et al.*, 2005) were used in a "genetical genomics"approach (Jansen and Nap, 2001) to identify genes differentially

expressed in a sugarcane population segregating for sugar content. In the present work, the *ScBAK1* gene differentially expressed in a high sugar content pool of these plants, was described.

The gene expression trends in a segregating sugarcane population were determined in the seven high sugar content (H) and seven low sugar content (L) pools, collected 6, 7, 9, 11 and 13 months after planting (Figure 7.1A). The inset graph represents the expression profiles of *ScBAK1* plotted for each group (H x L). This gene was shown to be enriched in the higher sugar content plants, throughout the growing season. To provide a replication of the gene expression profile observed throughout the growing season, the total RNA from each of three sugarcane individuals collected 9 months after planting was extracted and used in the RNA blot (Figure 7.1B).

The spatial profile of this gene was also analyzed, comparing its expression in the source (mature leaf) and sink (immature leaf, immature internode, root, lateral bud and flower) sugarcane tissues (Figure 7.2). The mRNA accumulated at high levels in the mature leaves, but no expression, or a very weak signal, was found in the other tissues analyzed.

In situ hybridization was used to localize *ScBAK1* transcripts in the leaves +1 tissue, using an antisense RNA probe corresponding to the *ScBAK1* cDNA. A strong expression of this gene was observed in the bundle sheath cells (Figure 7.3B and 7.3C), and also in the vascular parenchyma cells that separate the xylem from the phloem. There was no expression in the phloem, xylem vessel or mesophyll cells. In control experiments with a sense RNA probe, only weak background color developed (Figure 7.3A).

The *ScBAK1* gene codes for a receptor kinase expressed preferentially in bundle-sheath cells

The deduced gene product was named ScBAK1 (brassinosteroid insensitive1-associated receptor kinase 1 from *Saccharum* spp.) based on its sequence similarity to members of the LRR-RLKs subfamily. These proteins were characterized as containing 5 LRRs in their extracellular domains, and displaying good similarity with the family of somatic embryogenesis receptor kinases (SERKs), mainly with the SERK3/BAK1 proteins. The sequence of the longest cDNA clone was determined by sequencing. The complete ScBAK1 coding sequence was 1,860 bp in length, and encoded a protein consisting of 619 amino acids with an estimated molecular mass of 69 kDa and a predicted pI of 5.73. The overall structure

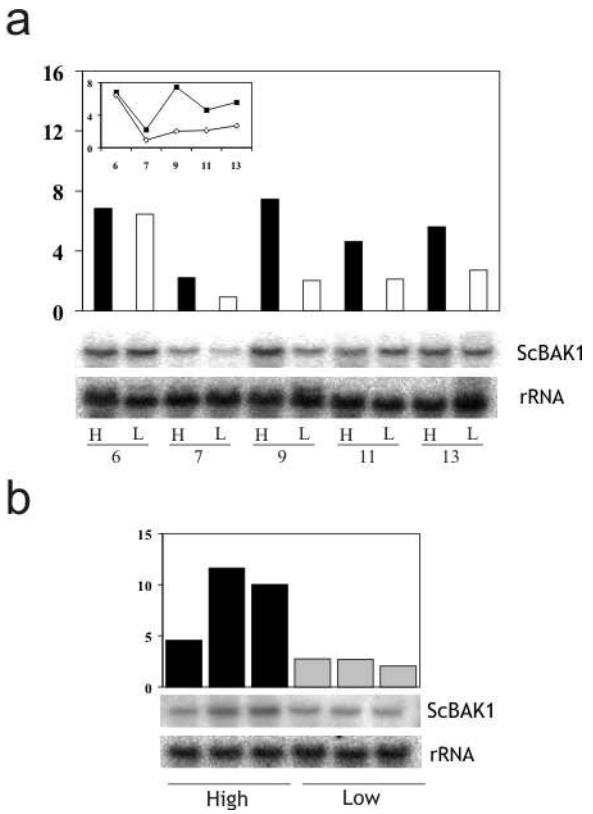


Figura 7.1: **Expression profile of *ScBAK1* throughout the growing season (A) and individual clones of segregated plants (B).** RNA blots were prepared using 10 μg of total RNA isolated from mature leaves of (A) a pool of 7 individuals with high (H) and low (L) sugar contents collected throughout the growing season (6, 7, 9, 11 and 13 months after planting), and (B) three individual clones of each segregated plant. In A, the inset graphs show the expression levels observed for the high (black circles) and low (white circles) sugar content plants, and B, the time point evaluated in the blots corresponds to 9 months after planting. An rDNA fragment was used as the control.

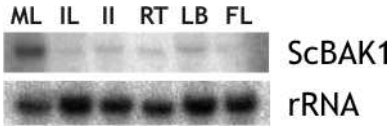


Figura 7.2: **Gene expression analysis in different tissues.** For the RNA gel blot preparation, each lane was loaded with 10 μg of total RNA isolated from one of six tissues from sugarcane. ML - mature leaves; IL - immature leaves; II - immature internode; RT - root; LB - lateral bud; FL - flowers. An rDNA fragment was used as the control.

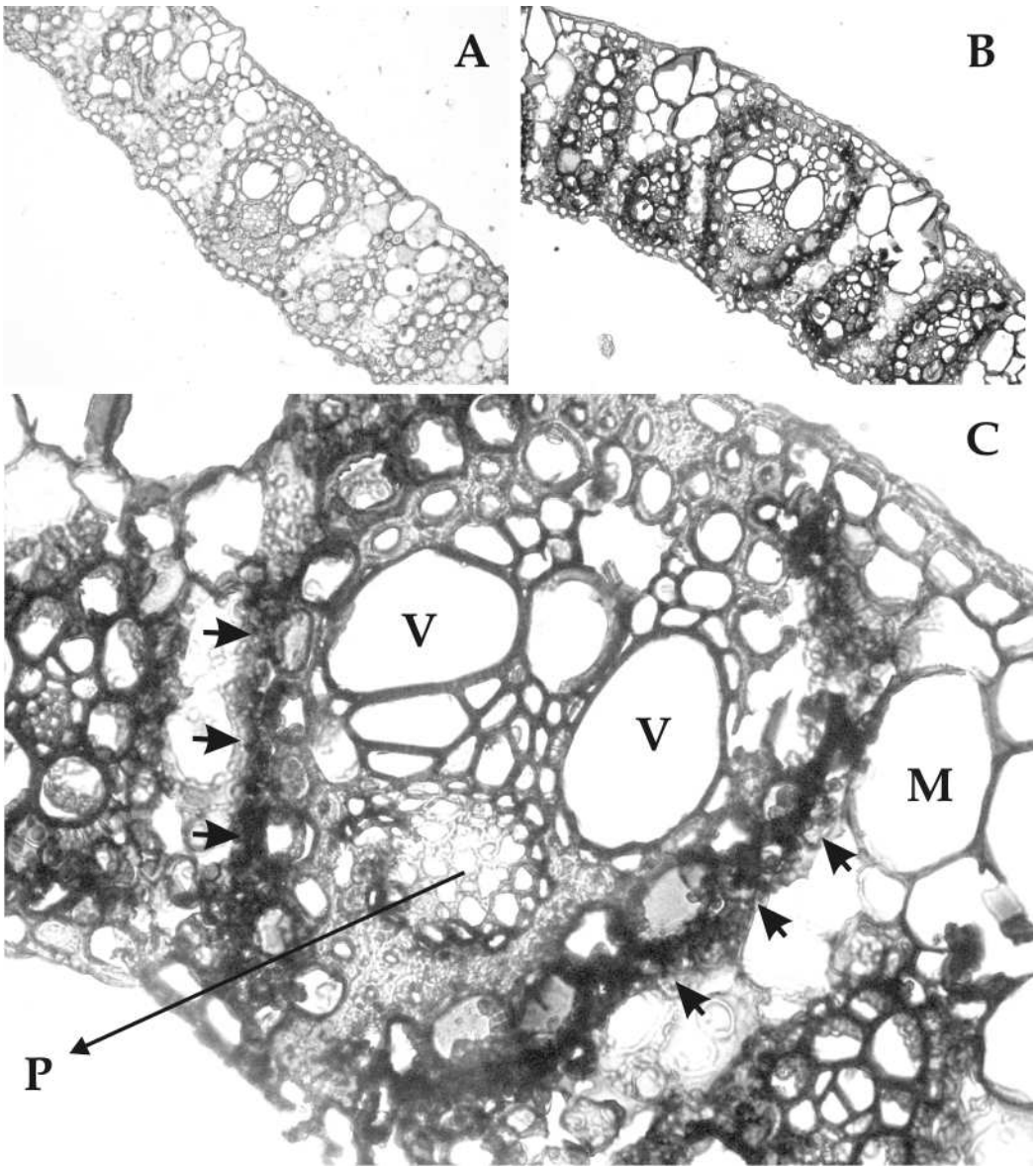


Figura 7.3: **Detection of sugarcane *ScBAK1* transcripts by *in situ* hybridization of sections of mature sugarcane leaf.** Transversal sections of mature leaf hybridized to the *ScBAK1* sense (A) and antisense (B) probes. (C) Detail of B showing that the majority of the transcripts were present in the bundle-sheath cells (indicated by arrows). V- xylem vessel; P - phloem; M - mesophyll cells.

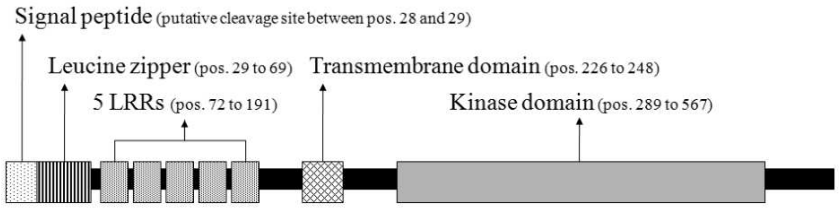
of the predicted polypeptide (Figure 7.4) shared all the characteristic features of the SERK proteins, including the five LRRs, the hydrophobic transmembrane domain and the kinase domain. The N-terminus of the ScBAK1 protein was characterized by a stretch of hydrophobic amino acids that could function as a signal peptide sequence for the rough endoplasmic reticulum. A possible processing site was located between the amino acids alanine and isoleucine at positions 28 and 29, respectively. The N-terminal domain was separated from the C-terminal domain by a putative transmembrane domain (amino acids at positions 226 to 248). The C-terminal domain showed a high sequence similarity with the catalytic domain of the protein kinases and contained the characteristic sub-domains and all the amino acids conserved in the eukaryotic protein kinases.

Evolutionary relationship of ScBAK1 with other SERK proteins of higher plants

A database search using the BLAST program revealed that the ScBAK1 protein shared the greatest similarity with proteins from the SERK kinase family. It most closely identified with SbRLK (94%) and OsBAK1 (73%), whilst showing a considerable percentage of identification with other SERK proteins (Figure 7.5). With the exception of the RLK identified in *Daucus carota* (Schmidt *et al.*, 1997), all the other thirteen SERK proteins had a signal peptide and a leucine zipper. The proline-rich box N-terminal to the transmembrane domain of the SERKs, was missing in ScBAK1, SbRLK1 and OsBAK1. Figure 7.5 shows an alignment of fourteen LRR-containing receptor-like protein kinases of higher plants that were classified in the SERK family.

A phylogenetic tree based on the whole sequence of these plant SERKs revealed their evolutionary relationships (Figure 7.6). The ScBAK1, SbRLK and OsBAK1 proteins are closely related and were clustered together (Monocotyledons Subfamily II group in Figure 7.6). These last two proteins are SERK3-type proteins. Other groups of related SERKs were those formed by the SERK3, 4, and 5 proteins of *Arabidopsis thaliana* (Eudicotyledons Subfamily II in the same figure), the one formed by the SERK1 and 2 proteins of *Oryza sativa* and *Zea mays* (Monocotyledons Subfamily I), and finally the clade containing SERK1 and 2 proteins of *Arabidopsis thaliana*, *Daucus carota*, and *Medicago truncatula* (Eudicotyledons Subfamily I). The amino acid sequence alignment, the protein structure, and the bootstrapped parsimony tree were consistent with ScBAK1 being an ortholog of SERK3-type proteins.

a



b

```
MMIYSEIMNLLAFVLLLWGCQQLSLSLA                              28   Signal Peptide

IEFQVEALVEMKMQLVDNRGVLSDWKDNQMSPCYWEYVNCQDN              71   3 Leucine Zippers

KVTTITLSSSGLTGTLSPSIAKLT                                  95   5 LRRs
TLQQLKLDNNITGGIPLEFGNLS                                   119
SLTILNLGRNNLNGSIPDSLGQLS                                  143
KLQILDLSHNHLSGNIPSSFSNPP                                  167
SLNNINLAHNNISGEIPQHLLQAA                                  191

HYNFTGNHLNCGQNLFPCEGGSTRTGGSKNSKLKVVIGSIAGAVTLFVTV       241  Transmembrane
VLVLLWW                                                   248

QRMRYRPEIFIDVSGQNDHMLEFGQIKRFSWRELQIATNYFSEQNVLGKG       298  Cytoplasmic
GFGKVYKGVLPRPDSIKIAVKPLFNVESREGEMAFLREVELISIAVHKNI       248  Kinase
LRLIRFCTTTTERLLVYPFMENLNVASRLRDIKLNEPALDWSTRMRIAPG       398  Domain
AARGLEYPHEHCNPKIIHSDVKAANVLLDGNFEAVVGDFGLAKMMDIGRN       448
TVTTGLRGTMGHIAPEYIKTGRPSVKTDIFGYGVMLLEIVTGDRAIAFHP       498
DRIEEAGEIMLIDQVKLWMEEGRLLDLVDHNLGGVYNLEELEKVTQIALL       548
CTHMEPNQRPTMSEVVQMLEGEIVPAERWEEWQLAELQRRQQHEMRQQGK       598
LFDFSEESLNIQEAIELSTGR                                     619
```

Figura 7.4: **The structure and the deduced amino acid sequence of the ScBAK1.** (A) Diagrammatic structure of the ScBAK1 protein. (B) The numbers show the position of the amino acid residues. The extracellular domain of the protein contains a signal peptide, three putative leucine zippers and five LRRs (conserved residues are in bold). The extracellular domain is separated from the cytoplasmic kinase domain by a hydrophobic transmembrane domain.

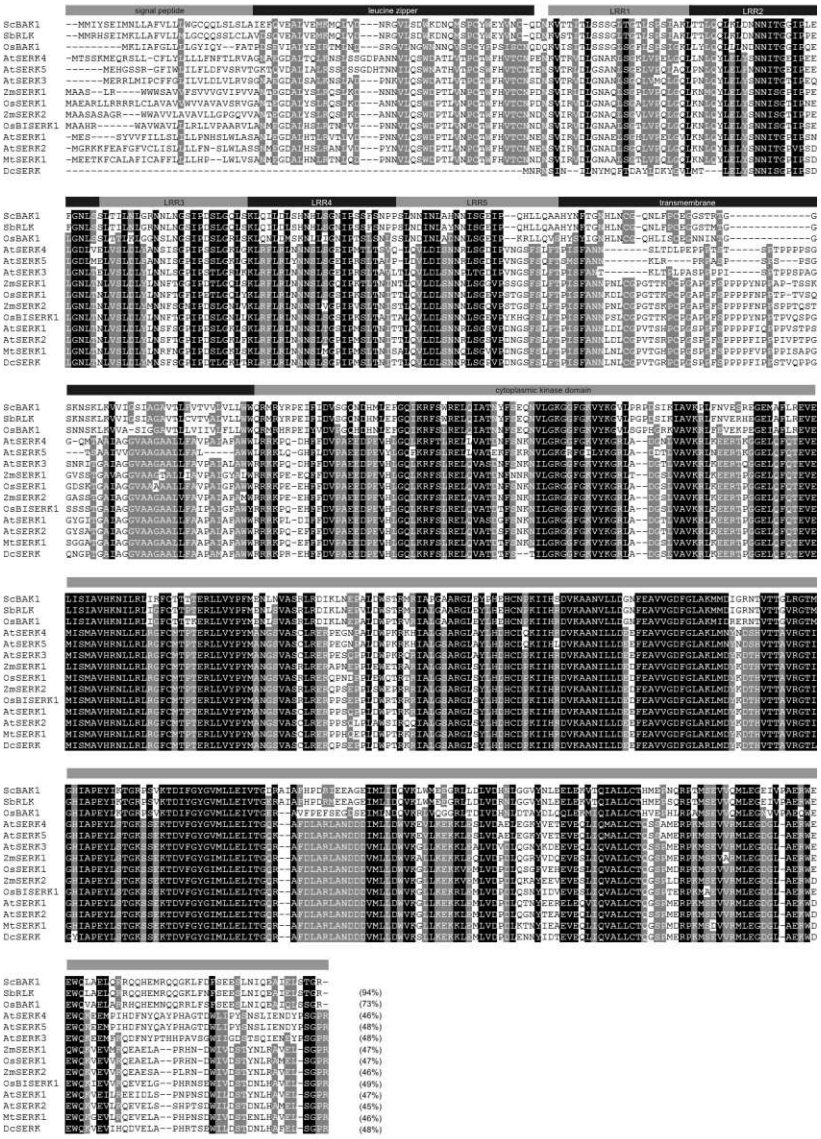


Figura 7.5: **Alignment of the amino acid sequence of fourteen LRR-containing receptor-like protein kinases of higher plants that were classified in the SERK family.** The amino acid residues conserved in all the proteins listed, are shaded in black. The percentages indicate amino acid identity as compared to ScBAK1. AtSERK1, 2, 3, 4 and 5 from *Arabidopsis*; OsSERK1, OsBAK1 and OsBISERK1 from rice; ZmSERK1 and 2 from maize; MtSERK1 from *Medicago truncatula*; DcSERK from carrot; SbRLK from sorghum; and ScBAK1 from sugarcane. GenBank protein accession numbers: AtSERK1, NP_177328; AtSERK2, AAK68073; AtSERK3, AAK68074; AtSERK4, NP_178999; AtSERK5, NP_179000; OsSERK1, AAU88198; OsBAK1, ABA94685; OsBISERK1, AAR26543; ZmSERK1, CAC37638; ZmSERK2, CAC37639; MtSERK1, AAN64293; DcSERK, AAB61708; SbRLK, CAB51480; ScBAK1, EU189960.

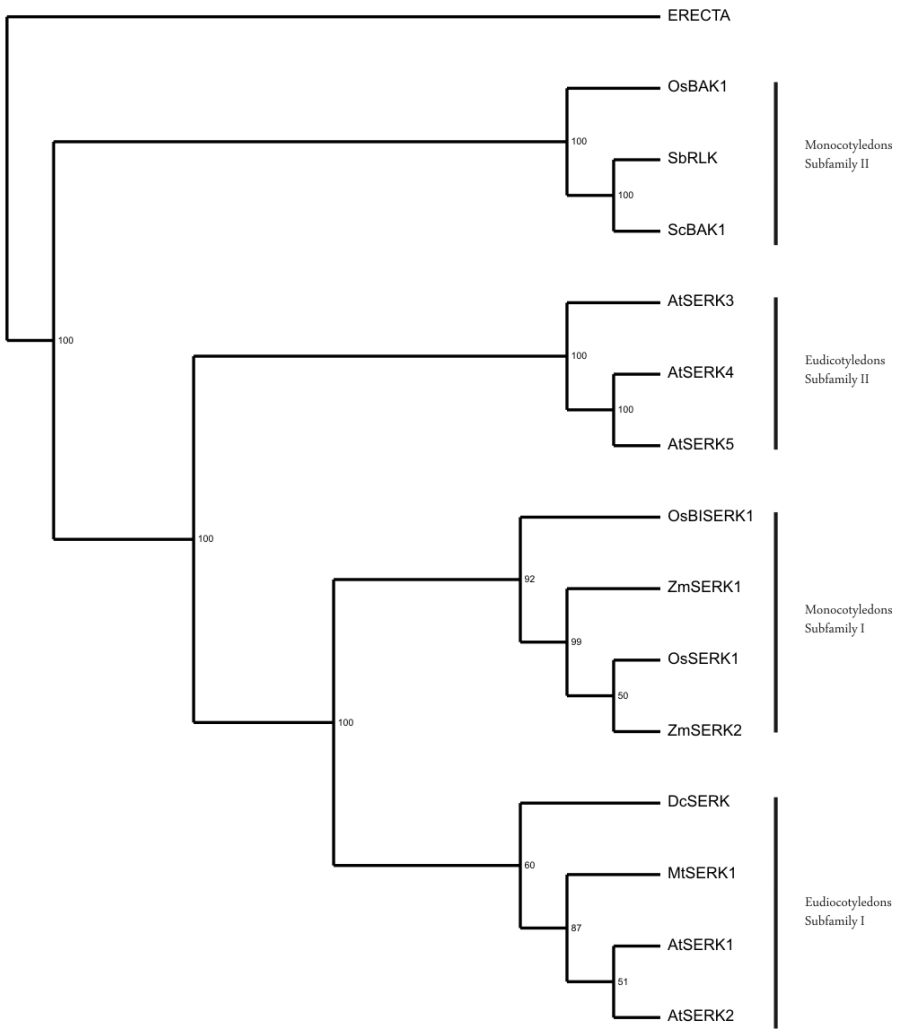


Figura 7.6: **Bootstrapped parsimony tree for ScBAK1 with representative plant SERK proteins, constructed based on the alignment of the amino acid sequences.** ScBAK1 is clustered together with thirteen LRR-containing receptor-like protein kinases of higher plants that were classified in the SERK family. The tree represents a consensus tree generated by 1000 bootstrap replicates. The numbers represent the bootstrap support (in percentages).

ScBAK1 is plasma membrane localized

To determine the subcellular localization of the ScBAK1 protein in the plant cells, *in vivo* targeting experiments were performed. ScBAK1 protein was tagged at the C- and N-terminus with GFP and then transiently expressed in onion epidermal cells under the control of the constitutive 35S promoter of the Cauliflower mosaic virus. In onion cells, ScBAK1-GFP is localized on the plasma membrane (Figure 7.7), which is the same subcellular localization reported previously for AtSERK3/BAK1 (Li *et al.*, 2002) and for BRI1 (Friedrichsen *et al.*, 2000).

The onion epidermal cells transiently expressing the non-fused GFP control, exhibited the nuclear and cytoplasmic green fluorescence characteristic of the GFP localization (Figure 7.7A), whereas the ScBAK1/GFP and GFP/ScBAK1 fusion proteins re-directed fluorescence throughout the periphery of the transformed cells, consistent with the proposed localization of ScBAK1 in the plasma membrane (Figure 7.7B, C and E). In order to differentiate between the plasma membrane and the cell wall, ScBAK1/GFP transformed cells were treated with 1 M mannitol, which induces plasmolysis resulting in the internalization of the plasma membrane along with cellular organelles, while the cell wall remains unchanged (Friedrichsen *et al.*, 2000). Figures 7.7C and D illustrate fluorescence and transmission images of a plasmolysed cell, in which the ScBAK1/GFP fluorescence at the plasma membrane was internalized or pulled away from the cell wall, as indicated by the bold face arrows. As a positive control for the secretory pathway and plasma membrane localization, the pSGFP5/psecGFP (Di Sansebastiano *et al.*, 1998) was also bombarded into onion epidermal cells. The pSGFP5/psecGFP fluorescence was localized along the secretory pathway, in the periphery of the cell and in the extracellular space (Figure 7.7F).

Discussion

Previous studies of the DcSERK from carrot and the AtSERK1 from *Arabidopsis*, suggested that the SERK genes were expressed in the embryonic cells and provided embryonic competence to the cells (Hecht *et al.*, 2001, Schmidt *et al.*, 1997). In contrast, analyses of other SERK genes suggested that they may have broader roles rather than being specific for embryogenesis (Baudino *et al.*, 2001, Li *et al.*, 2002, Nam and Li, 2002, Nolan *et al.*, 2003). AtSERK3/BAK1 is known to be a component of a brassinosteroid receptor complex

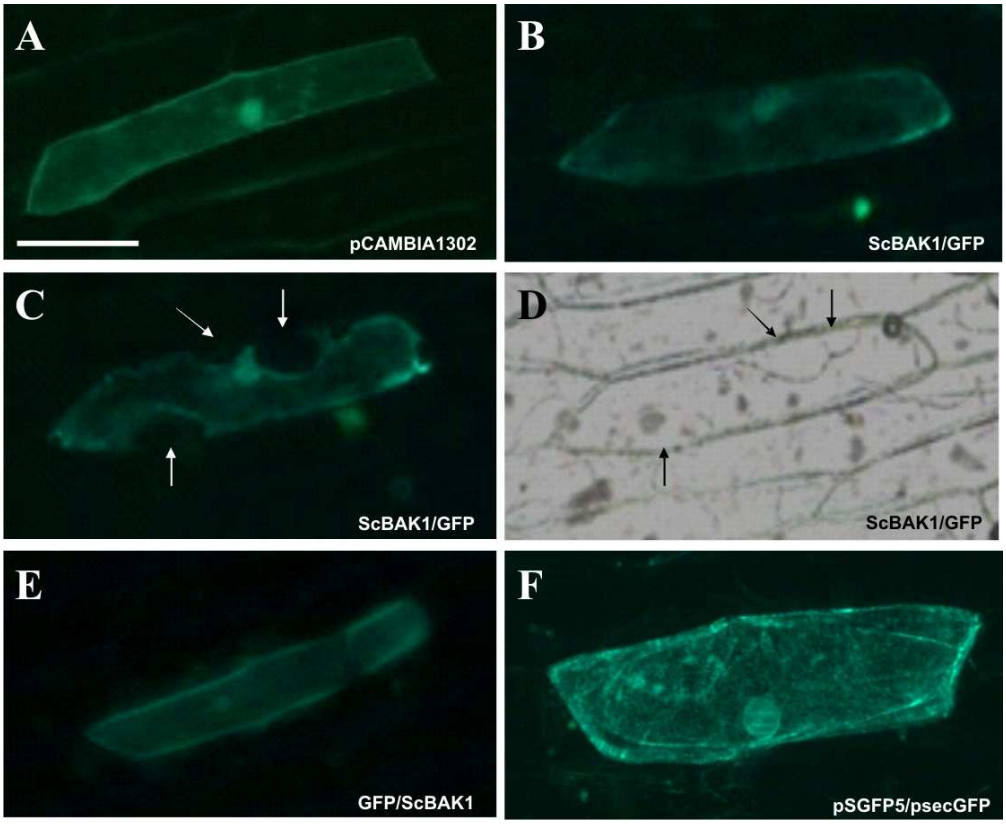


Figura 7.7: **The transient expression of GFP-tagged ScBAK1 in onion epidermal cells visualized by epifluorescence microscopy.** GFP protein with no signal (A, GFP control), secretory signal (F, pSGFP5/psecGFP control), and C- and N-terminal fusion protein between GFP and the ScBAK1 (B and E, respectively). Note that the cells expressing only GFP show cytoplasmatic fluorescence (A), while the fluorescence of ScBAK1-GFP and GFP-ScBAK1 is localized predominantly in the plasma membrane (B and E). This plasma membrane localization of the ScBAK1-GFP and GFP-ScBAK1 was further confirmed by plasmolysing the cells with 1 M mannitol, where the fluorescence signal was localized only at the retracting plasma membrane and not at the cell wall (C and D). Bold face arrows indicate where the plasma membrane has detached from the cell wall as a result of plasmolysis. The scale bar is equivalent to 120 μm.

and involved in brassinosteroid signaling in *Arabidopsis* (Li *et al.*, 2002, Nam and Li, 2002). These results suggest that the functions of the members of the SERK family are not limited to embryogenesis, but may play divergent roles. In the SERK family, conservation is very extensive within a subgroup encompassing ZmSERK1-2, DcSERK, AtSERK1-2, MtSERK1, OsSERK1 and OsBISERK1. In contrast, conservation is only partial for AtSERK3-5, OsBAK1, SbRLK and ScBAK1, mainly because pro-rich domains with SPP are missing in these proteins, and the number of leucine residues in the putative ZIP domain is reduced from four to three. The Pro-rich domain with the SPP motif, located between the LRRs and the transmembrane region, has been suggested to act like a hinge, providing flexibility to the extracellular part of the receptor or as a region for interaction with the cell wall (Hecht *et al.*, 2001). This finding reflects the phylogenetic tree of the kinase domains, and may be an indication of the common specialized functions of each subgroup member. The ScBAK1 protein has a number of features that are characteristic of the LRR receptor protein kinases classified in the SERK family. At the N-terminus this protein contains a typical signal sequence that could be involved in targeting to the rough endoplasmic reticulum. The putative extracellular N-terminal domain contains five LRRs that have been proposed to be involved in protein-protein or protein-ligand interactions (Kobe and Deisenhofer, 1994). This N-terminal domain is separated from the cytoplasmic protein-kinase domain by a transmembrane domain. The putative protein-kinase domain contains all the diagnostic amino acids that have been described for the catalytic domain of protein kinases (Hanks *et al.*, 1988). Due to the considerable sequence similarity between ScBAK1 and SbRLK and OsBAK1, it was suggested that the sugarcane protein was a component of a brassinosteroid receptor complex, and might play a role in brassinosteroid signaling. BAK1 and BRI1 are both localized in the plasma membrane, supporting the idea that BAK1 and BRI1 may interact with one other directly in a physiological setting (Li *et al.*, 2002). Recent studies detected that both the BRI1 and AtSERK3/BAK1 proteins were present in small vesicle-like compartments in the cytoplasm, close to the plasma membrane. The results of several experiments led to the conclusion that these vesicles represented endosomes (Russinova *et al.*, 2004, Ueda *et al.*, 2001, Vida and Emr, 1995). The turnover of the BRI1 and AtSERK3 proteins when expressed individually was relatively slow, but when co-expressed, the complex first co-localized at the plasma membrane and was then rapidly internalized into the endosomes. Most of the internalized BRI1 and AtSERK3 proteins were not recycled back to the membrane and were possibly targeted for degradation, resulting in a nearly complete depletion of the BRI1 and At-

SERK3 fluorescent proteins from the plasma membrane (Russinova *et al.*, 2004, Ueda *et al.*, 2001, Vida and Emr, 1995). These data raise the possibility that ScBAK1 might play a role in bringing the cell surface BR receptor into the cytosol. The localization of the ScBAK1-GFP fusion protein at the plasma membrane supports the prediction that it is a plasma membrane protein, similar to that observed in Arabidopsis for BAK1 (Li *et al.*, 2002, Nam and Li, 2002) and BRI1 (Friedrichsen *et al.*, 2000). Gene regulation is based on sensing different signals or stimuli, leading to an increase or decrease in transcription. In sugar signaling, the first step is to sense the nature and level of the specific sugar. The expression of genes involved in photosynthesis and in starch, lipid and protein remobilization, is enhancing by sugar deprivation. However when cellular sugar levels are elevated, the genes involved in the synthesis of polysaccharides, storage proteins, defense responses and respiration are up-regulated (Ho *et al.*, 2001, Koch, 1996, Yu, 1999). Although the regulatory effect of sugars on photosynthetic activity and plant metabolism has long been recognized, the concept of sugars as central signaling molecules is relatively new (Rolland *et al.*, 2006). The expression profile of the signal transduction components in a sugarcane population segregating for sugar content (Felix *et al.* submitted) was evaluated, and in the present work *ScBAK1* showed greater expression in the pool of plants with higher sugar contents throughout the growing season. In order to study the metabolic interactions between the mesophyll cells and bundle-sheath cells in sorghum, a C4 plant, Annen and Stockhaus (1999) cloned and characterized protein kinases that could be involved in the regulatory processes and signal transduction of this metabolic pathway. One of these was SbRLK, which accumulated at much higher levels in the mesophyll cells than in the bundle sheath cells, and was almost undetectable in the roots. The sugarcane *ScBAK1* transcripts were almost undetectable in the roots or in any other sink tissue, but accumulated differentially in mature leaves, which have photosynthetic active cell types. Contrary to that previously reported for the sorghum ortholog, *in situ* analyses showed that transcripts encoding ScBAK1 were preferentially expressed in the bundle-sheath cells of mature leaves. It is known that receptor-like protein kinases are responsible for the activation of other factors that are translocated to the nucleus to induce gene expression changes (Hardie, 1999). Taking together with the fact that this gene was found to be enriched in the high sugar content individuals throughout the growing season, this suggests that the sugarcane ScBAK1, a bundle-sheath cell predominant protein kinase, could be involved in cellular signaling cascades mediated by abundant sugar levels in sugarcane mature leaves. The authors are currently characterizing transgenic events in sugarcane for this gene, to increase knowledge about the

role of this protein in sucrose synthesis.

Acknowledgements

We are grateful to G.P. Di Sansebastiano (Università di Lecce, Italy) for providing DNA of pSGFP5/psecGFP. JMF was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) fellowships, RV was supported by a fellowship from the UNIEMP Institute, and MM received a research fellowship from CNPq. This work was partially supported by FAPESP grant 05/58104-0 awarded to MM.

tout pour le mieux dand le meilleur des mondes possibles.

Voltaire (1694-1778)

8

Análise sistemática da localização subcelular de proteínas de cana-de-açúcar

Renato Vicentini[1] e Marcelo Menossi[1]

Resumo

Ao término do projeto EST da cana-de-açúcar (SUCEST), apresentou-se um enorme desafio à comunidade científica envolvida: a caracterização funcional do grande número de genes descoberto. A identificação da localização subcelular das proteínas de cana-de-açúcar é capaz de contribuir nesse sentido, pois a localização determina potenciais padrões de interação metabólica. O principal objetivo deste trabalho foi ampliar o estudo funcional de genes de cana-de-açúcar através do estabelecimento de uma rotina para a determinação da localização subcelular de suas respectivas proteínas. Neste capítulo apresentamos os progressos obtidos no que se refere clonagem dos genes de interesse nos vetores de expressão, obtenção e avaliação dos controles de localização subcelular, assim como a avaliação da localização subcelular das construções obtidas. São relatadas também as dificuldades que foram enfrentadas na detecção do sinal das proteínas fusionadas a proteína GFP, dificuldades estas que inviabilizaram até o momento a determinação da localização subcelular de várias proteínas selecionadas para este fim. Com isto damos início a busca por correlações entre a localização subcelular e dados funcionais gerados por pesquisas atuais, visando ampliar o

[1]Departamento de Genética e Evolução, Laboratório de Genoma Funcional, Instituto de Biologia, CP 6009, Universidade Estadual de Campinas - UNICAMP, 13083-970, Campinas, SP, Brasil

conhecimento sobre importantes processos da cana-de-açúcar.

Introdução

A cana-de-açúcar (*Saccharum* spp.) é uma das mais importantes espécies vegetais cultivadas em todo o mundo, principalmente nas regiões tropicais e subtropicais, com uma área maior que 20 milhões de hectares em 101 países (FAO 2004; http://apps.fao.org). O Brasil é o maior produtor, participando com 25% da produção, seguido pela Índia, China, Tailândia e Paquistão. Sua importância econômica provém do crescente consumo de açúcar e álcool (etanol), sendo o estado de São Paulo responsável por metade da produção brasileira (Pessoa-Jr *et al.*, 2005).

Durante a evolução, as células eucarióticas desenvolveram um complexo sistema de endomembranas que deu origem a uma grande variedade de compartimentos subcelulares. A constituição destes microambientes subcelulares, denominados organelas, é que as tornaram capazes de desempenhar processos fisiológicos e bioquímicos específicos. O perfeito funcionamento de uma organela depende de seu conteúdo, que é definido de acordo com as suas funções. Sendo assim, um sistema de importação seletivo e eficiente é fundamental para a manutenção da identidade funcional e estrutural dos diferentes compartimentos subcelulares.

A proteína fluorescente verde (GFP, *green-fluorescent protein*) de *Aequorea victoria*, assim como suas variantes que fluorescem em outras faixas de emissão de luz, vêem sendo utilizada como um gene repórter universal em ampla gama de células e organismos heterólogos (Chiu *et al.*, 1996). As novas tecnologias de clonagem permitem gerar rapidamente fusões com o gene gfp tanto na extremidade amino quanto na carboxi-terminal. Essa metodologia possibilita o exame da localização subcelular em larga escala, sendo que em cerca de 80% das construções gênicas é possível detectar claramente a localização intracelular (Simpson *et al.*, 2000).

Em qualquer estratégia de análise funcional de proteínas em larga escala é evidente a necessidade da realização das atividades em paralelo, principalmente no que se refere ao sistema de clonagem e expressão. A clonagem dos CDS precisa ser rápida, eficiente, direcional e compatível com a vasta esfera de vetores de expressão existente. Muitas das limitações podem ser solucionadas com a utilização de métodos de clonagem recombinacional, como o sistema *Gateway* (Invitrogen, Brasil) que pode ser aplicado *in vitro* e apresenta

compatibilidade com vários vetores de clonagem, incluindo os utilizados em sistemas de expressão transiente. Esses sistemas mimetizam os eventos de recombinação sitio específico utilizado pelo bacteriófago *Lambda* para se integrar e se remover do genoma de *Escherichia coli*, permitindo assim a clonagem unidirecional do gene de interesse (Rual *et al.*, 2004).

Um problema freqüente na utilização da GFP refere-se à extremidade da molécula de interesse a qual ela está fusionada. Contudo, atualmente os métodos de clonagem permitem gerar tanto fusões amino quanto carboxi-terminal em uma única reação, possibilitando assim identificar imediatamente qualquer efeito que a GFP pode estar causando no sinal de direcionamento da proteína. Exemplos demonstram que proteínas com direcionamento especificamente mitocondrial são detectadas no núcleo e no citosol caso apresentem a proteína GFP fusionada em sua extremidade amino, demonstrando claramente que a perturbação no sinal de direcionamento mitocondrial acaba com a habilidade de estas proteínas serem localizadas corretamente (Simpson *et al.*, 2000).

Diversos grupos vêm utilizando com sucesso estes sistemas de clonagem, obtendo centenas e até mesmo milhares de construções viáveis (Simpson *et al.*, 2000, Reboul *et al.*, 2003, Palmer and Freeman, 2004, Rual *et al.*, 2004). Os resultados sugerem que apenas uma minoria dos eventos de expressão transiente sofre interferência destes sistemas de recombinação e ressaltam que a determinação da localização subcelular de novas proteínas pode se valer destas estratégias com alto grau de confiabilidade (Simpson *et al.*, 2000).

Neste estudo buscamos a determinação da localização subcelular das novas seqüências protéicas identificadas em cana-de-açúcar, visando assim o fornecimento de valiosos detalhes das possíveis funções destas proteínas, uma vez que a localização determina potenciais padrões de interação metabólica. No caso de proteínas já conhecidas, a informação sobre a localização subcelular pode fornecer indícios sobre vias enzimáticas específicas. Se junta a isto, o fato de que as proteínas menos caracterizadas e específicas de plantas apresentam tanto órgão especificidade quanto direcionamento específico para organelas celulares (Gutierrez *et al.*, 2004).

Materiais e métodos

Predição *in silico* da localização subcelular das prováveis proteínas de cana-de-açúcar

Os CDS foram obtidos utilizando os *softwares* GeneMark.SPL (Borodovsky and McIninch, 1993), GENSCAN (Burge and Karlin, 1997) e ESTScan (Iseli *et al.*, 1999), que se caracterizam por serem métodos *ab initio*, em conjunto com uma estratégia de escolha dos CDS válidos baseada em homologia de seqüências protéicas (Capítulo 6). Já a predição da localização subcelular destas prováveis proteínas foi realizada por sete *softwares* distintos: iPSORT (Bannai *et al.*, 2002), MitoProtII (Claros and Vincens, 1996) e PSORT (Nakai and Horton, 1999), que detectam peptídeos sinais; Predotar (Small *et al.*, 2004), SubLoc (Hua and Sun, 2001) e TargetP (Emanuelsson *et al.*, 2000), que utilizam técnicas de *machine-learning*; e o PredictNLS sendo específico na detecção de sinais de localização subcelular (NLS, *nuclear localization signals*), também aliado a uma estratégia de tomada de decisão (Capítulos 5 e 6).

Seleção dos genes de interesse e confirmação dos clones desejados

Os genes selecionados foram identificados a partir dos clones de cDNA obtidos no projeto EST da cana-de-açúcar (SUCEST). A seleção valeu-se da busca por transcritos de cana-de-açúcar (SASs, *Sugarcane Assembled Sequences*) no banco de dados do projeto SUCEST, assim como em resultados do estudo dos níveis de expressão gênica. Foram selecionados 96 genes de cana-de-açúcar que apresentam potencial biotecnológico e/ou predição de localização subcelular única. Em seguida, foi selecionado o clone do projeto SUCEST mais representativo de cada gene, e estes tiveram suas identidades confirmadas por seqüenciamento de DNA. A representatividade de um clone é determinada pela maior proximidade com a região 5' do gene homologo. Já nos casos onde o SAS não possui homologia, é selecionado o clone que contribui para o início da seqüência consenso.

Amplificação e clonagens dos genes de interesse visando fusão com o gene *gfp*

Uma vez obtido os DNAs dos genes de interesse, foi realizada a amplificação visando à obtenção dos CDSs a serem utilizados nas clonagens. Foram desenhados três oligonucleotídeos para a amplificação de cada CDS (totalizando assim 288 oligonucleotídeos, Tabela 8.1), os oligonucleotídeos *forward* foram desenhados para permitir a amplificação a partir do ATG que codifica o aminoácido metionina do início da síntese protéica, já os oligonucleotídeos *reverse* foram desenhados com a remoção do sítio *stop* dos CDS para a subseqüente fusão na extremidade N-terminal da proteína GFP (*primers* reverso). Oligos utilizados como *primers reverse* apresentando o sítio *stop* também foram desenhados para permitir a fusão na extremidade C-terminal da proteína GFP. Outra característica dos oligos *forward* é a presença dos nucleotídeos 5' CACC 3' na extremidade 5', necessários para a realização da clonagem direcional no vetor pENTR/D-TOPO. As reações de PCR foram realizadas em placa de 96 wells utilizando um gradiente de temperatura de 60 $\pm$ 5°C. Obtido os CDSs, foram realizadas as clonagens no vetor de entrada pENTR/D-TOPO (Invitrogen, Brasil) compatíveis com o sistema *Gateway* (Invitrogen, Brasil). Estas construções foram então recombinadas com os vetores *p2FGW7* e *p2GWF7* (Karimi *et al.*, 2002, Joubes *et al.*, 2004, Karimi *et al.*, 2005), obtendo-se assim construções que permitem a fusão da proteína GFP na extremidade amino e carboxi das proteínas de interesse, respectivamente (Figura 8.1). Com o uso destas duas construções distintas somos capazes de identificar qualquer efeito que a GFP possa estar causando no sinal de direcionamento da proteína devido a extremidade de fusão. Estes vetores, baseados no promotor 35S, foram adquiridos da *Functional Genomics Division of the Department Plant Systems Biology* (VIB-Ghent University).

O protocolo padronizado para a clonagem no vetor de entrada pENTR/D-TOPO foi estabelecido como uma reação contento 1 μl do produto de PCR, 1 μl da solução de sais, 3 μl de H_2O estéril e 1 μl do vetor TOPO, sendo iniciada a transformação das bactérias competentes após 10 minutos em temperatura ambiente. Já para a recombinação com os vetores de destino foi estabelecida a reação da seguinte forma: 1 μl do DNA da clonagem no TOPO, 1,5 μl do DNA do vetor de destino, 5,5 μl de T.E. (ph 8.0) e 2 μl da enzima clonase. Incuba-se esta reação a 25°C por duas horas e meia, sendo que após este período adiciona-se 1 μl de Proteinase K e incuba-se a 37°C por 10 minutos antes de iniciar a transformação das bactérias competentes.

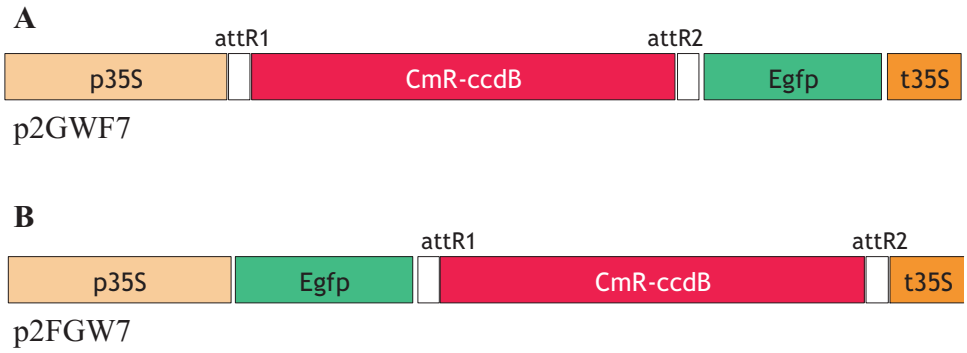


Figura 8.1: **Construções apresentadas pelos vetores a serem utilizados na expressão transiente.** A e B fusão carboxi-terminal e amino-terminal, respectivamente. p35S (DeLoose *et al.*, 1995); Egfp (*Enhanced fluorescence protein*, Clontech, EUA); attR1 e attR2, sítios de recombinação, e CmR-ccdB, gene marcador de resistência, (Invitrogen, Brasil); t35S (Goddijn *et al.*, 1993).

A confirmação das clonagens foi realizada por PCR utilizando os *primers* específicos, por digestão enzimática com as enzimas *MluI* (confirmando a clonagem no pENTR/D-TOPO) e *PstI* (confirmando a clonagem no *p2FGW7* e *p2GWF7*), assim como por seqüenciamento, através do qual temos a possibilidade de verificação da seqüência que foi clonada.

Tabela 8.1: ***Primers forward* e *reverse* desenhados para a amplificação dos CDS dos clones do SUCEST.**

CDS	Direção		Primer		TM	Tamanho
CDS5956	forward	5'	CACCATGGCGGCCTCCGGT	3'	62.5	19
	reverse	5'	CAGATCATTTTTTGGTTCATCC	3'	57.3	22
CDS1092	forward	5'	CACCATGGACCCGGACGCTGT	3'	62.60	21
	reverse	5'	GCCATAGTTCAGGCGGAA	3'	59.76	18
CDS10984	forward	5'	CACCATGAGGACGCATGAACGAG	3'	59.80	23
	reverse	5'	GAGGTCATCAAGTCCGTTTAGC	3'	60.14	22
CDS1038	forward	5'	CACCATGGGCTCCTTCCTCTCGA	3'	62.23	23
	reverse	5'	GGATGAGGAAGATCGCGTG	3'	61.75	19
CDS1024	forward	5'	CACCATGGTGGTGGAGGAGATCACTG	3'	63.58	26
	reverse	5'	AGCCTGCTCTTCACCATCTTC	3'	60.92	21
CDS6524	forward	5'	CACCATGGAGGCAGCAGTAGCC	3'	58.94	22
	reverse	5'	GTTGGTCACGAAATGGTACCTG	3'	61.56	22
CDS890	forward	5'	CACCATGGAGCCAGATTCGGC	3'	59.69	21
	reverse	5'	GAACAGCACCATGCCAAC	3'	58.56	18
CDS2273	forward	5'	CACCATGGCCGGCGCCGAGG	3'	55.72	20
	reverse	5'	ACCTTCGGTGCTTGATGACTTGAC	3'	55.40	24
CDS10295	forward	5'	CACCATGGCGTCAGAAAAGAAGCA	3'	60.91	24

continua na próxima página

Tabela 8.1 – continuação da página anterior

CDS	Direção		Primer		TM	Tamanho
	reverse	5’	CGAGGTGTTTGCTTGGG	3’	58.71	17
CDS125	forward	5’	CACCATGGACCCGTGCCCGT	3’	63.53	20
	reverse	5’	CCTGAGTAGGTCGGCCCG	3’	64.10	18
CDS1286	forward	5’	CACCATGGTAGGGGTGCTGTCGAATCG	3’	56.99	27
	reverse	5’	CGGTGCTGCGGGGGATATAATG	3’	56.86	22
CDS2152	forward	5’	CACCATGATCATCTCCAAGAAGAACCG	3’	61.68	27
	reverse	5’	CTCCATGGAAGATCCACCG	3’	61.44	19
CDS5183	forward	5’	CACCATGGGCTCCTTCCTCTCG	3’	60.30	22
	reverse	5’	GGATGAGGAAGATGGCGTG	3’	61.60	19
CDS3725	forward	5’	CACCATGACGAACGGCTACCTG	3’	57.17	22
	reverse	5’	GTGGTTGGCAATGATCTGTT	3’	58.42	20
CDS4650	forward	5’	CACCATGGCGGACTGGGGC	3’	61.71	19
	reverse	5’	GCCCGCGTAGACGTGGAT	3’	64.52	18
CDS2474	forward	5’	CACCATGGCCTCCGGTGGC	3’	60.3	19
	reverse	5’	CGACTCTCGCACGTAGC	3’	55.2	17
CDS4583	forward	5’	CACCATGGAGAGGGCGGTGC	3’	60.78	20
	reverse	5’	GATCAGAGCGATGCTGCTCTC	3’	62.70	21
CDS4123	forward	5’	CACCATGTCGGCCCAGCAGAT	3’	60.18	21
	reverse	5’	GTAGAGATCTTGGACTTTCTGGG	3’	58.41	23
CDS1967	forward	5’	CACCATGGAAGGGAAGGAGGAGGA	3’	62.22	24
	reverse	5’	AGACCTGCTCTTGAATGGGATG	3’	62.77	22
CDS3242	forward	5’	CACCATGGCGGGTCTGGGAA	3’	61.04	20
	reverse	5’	TTGGGCTGGATTTGGG	3’	58.86	16
CDS11487	forward	5’	CACCATGGCAGAGGTTGTTGTGG	3’	59.53	23
	reverse	5’	TCGCTTCCCAGTGTGTCC	3’	60.85	18
CDS6243	forward	5’	CACCATGCCGATCAGCAGGAT	3’	57.52	21
	reverse	5’	GTAGTCGGTGGTGGGGAG	3’	58.88	18
CDS8136	forward	5’	CACCATGGATTTGGATCTGTGGATC	3’	58.66	25
	reverse	5’	GAGGAAATATGGGACTGCA	3’	56.02	19
CDS10829	forward	5’	CACCATGGCGAAGGACATCGAA	3’	60.16	22
	reverse	5’	TGTACTGGAGGGATCGGAGT	3’	59.53	20
CDS11691	forward	5’	CACCATGGACGGTGGTTCCTGTG	3’	61.84	23
	reverse	5’	GGAAGCTTCAGCAATCGAGA	3’	60.62	20
CDS9840	forward	5’	CACCATGGCTCATCGCGTTCTC	3’	59.91	22
	reverse	5’	CTCCTTCGAGGTGTGGACTG	3’	60.85	20
CDS8278	forward	5’	CACCATGGAGTGCGAGACCGG	3’	60.82	21
	reverse	5’	GCAGAGCGGCAGGGG	3’	62.87	15
CDS2210	forward	5’	CACCATGGCGGACTGGGGC	3’	61.71	19
	reverse	5’	GCCCGCGTAGACGTGGAT	3’	64.52	18
CDS3121	forward	5’	CACCATGGCGATGATGGTGGAT	3’	59.68	22
	reverse	5’	TGACATGCTTATTGCTGTTACAA	3’	58.44	23
CDS10592	forward	5’	CACCATGGGGCCGTCCATG	3’	59.76	19
	reverse	5’	GGCTGGCTGGGCTAGAACT	3’	61.84	19
CDS9723	forward	5’	CACCATGATCATCCCCAAGAAGAAC	3’	57.90	25
	reverse	5’	CTCCATGGAAGACGAGGTC	3’	58.14	19
CDS7778	forward	5’	CACCATGAGGCAGCCCACCAG	3’	60.82	21
	reverse	5’	TGGCCACCCAACCGAC	3’	62.71	16

continua na próxima página

Tabela 8.1 – continuação da página anterior

CDS	Direção		Primer		TM	Tamanho
CDS6252	forward	5’	CACCATGTGTGGCGGTGAGATCC	3’	61.98	23
	reverse	5’	AAACAACAGGTTGTCCTCGAC	3’	59.10	21
CDS7965	forward	5’	CACCATGGCGGTGTCGATCG	3’	58.8	20
	reverse	5’	GAGTGTCTCCCTGATGACCA	3’	57.3	20
CDS11693	forward	5’	CACCATGATTACGGCGACGGACTTCTAC	3’	55.40	28
	reverse	5’	GAGATGCCCCGCCGGC	3’	55.72	16
CDS517	forward	5’	CACCATGAGTACAACCAAGGTGAAGAGA	3’	59.21	28
	reverse	5’	GGTTCCCATATGAATCAAGCA	3’	59.78	21
CDS11518	forward	5’	CACCATGGCGAGGTCTTCCAAGATG	3’	63.68	25
	reverse	5’	GGGCTTGTTGGCGTGCT	3’	63.44	17
CDS11117	forward	5’	CACCATGGTCGCTGCTAGCCT	3’	55.7	21
	reverse	5’	GCATGCATCATGCATGG	3’	57.6	17
CDS6228	forward	5’	CACCATGCCTTCCAGGGCCATGAACC	3’	56.86	26
	reverse	5’	GGCGCGGGCCGACAG	3’	55.44	15
CDS4119	forward	5’	CACCATGGCGGGGGCCGGG	3’	55.44	19
	reverse	5’	CATGGAACTGCAGCACTTGCTCT	3’	55.21	23
CDS9529	forward	5’	CACCATGGCTTTCCTCAACTTGGA	3’	59.67	24
	reverse	5’	CGCCTCCACGGCAAT	3’	60.32	15
CDS6392	forward	5’	CACCATGGTCTGCGTCGCGT	3’	59.94	20
	reverse	5’	TTCCTCTTTCTTATCTTCTGCCG	3’	61.15	23
CDS4642	forward	5’	CACCATGTCGACGCCGGCG	3’	63.0	19
	reverse	5’	GTCGGCTGTCCAGCTCT	3’	56.1	17
CDS2447	forward	5’	CACCATGGCGGACTGGGGC	3’	61.71	19
	reverse	5’	GCCCGCGTAGACGTGGAT	3’	64.52	18
CDS7976	forward	5’	CACCATGGTGGGTCTCGTCGG	3’	60.50	21
	reverse	5’	CTGATCAATCGGCGATGTAG	3’	59.25	20
CDS2561	forward	5’	CACCATGGCGGCACAGTTTG	3’	57.50	20
	reverse	5’	TTTAGCGTTTGCATGAAGC	3’	57.68	19
CDS10970	forward	5’	CACCATGGCTATGATTGATGAACCTCTGT	3’	62.21	29
	reverse	5’	GCTAGATATCACCATTTGGTCACACG	3’	64.82	26
CDS3321	forward	5’	CACCATGGCCGTGATGGAGAAG	3’	59.59	22
	reverse	5’	CTTTTTCTTGCCGGTGACA	3’	59.83	19
CDS11056	forward	5’	CACCATGGGTAAGGAGAAGTCCCACA	3’	61.98	26
	reverse	5’	TTTCTTCTTGGCAGCAGCCTT	3’	63.58	21
CDS1959	forward	5’	CACCATGGCAGGCAAGGCCTC	3’	62.36	21
	reverse	5’	GAGGCATTTGAATCCGGTG	3’	61.41	19
CDS1074	forward	5’	CACCATGATGGAGAGCCAGCC	3’	57.02	21
	reverse	5’	GTCTTCCGCGGTGCC	3’	59.84	15
CDS157	forward	5’	CACCATGGAGTGGACGACGGTG	3’	60.55	22
	reverse	5’	TTTTGGCAAGGAATTAATGAATTTTA	3’	61.09	26
CDS5587	forward	5’	CACCATGGCAGGCAAGGCCTC	3’	62.36	21
	reverse	5’	GAGGCATTTGAATCCGGTG	3’	61.41	19
CDS4197	forward	5’	CACCATGGGGCTCACCTTCACG	3’	62.09	22
	reverse	5’	AGTCTCAGGGTCAAAATGACCA	3’	60.91	22
CDS11026	forward	5’	CACCATGGCCTCCGCCAACAG	3’	63.27	21
	reverse	5’	GTTGTAGGCGTCGGGGTTG	3’	63.74	19
CDS1864	forward	5’	CACCATGGCGGCGGGGAATA	3’	63.37	20

continua na próxima página

Tabela 8.1 – continuação da página anterior

CDS	Direção		Primer		TM	Tamanho
	reverse	5’	AGCATAGTACTTGACCATGCTTGTG	3’	61.75	25
CDS4157	forward	5’	CACCATGGGCGTCGTCTCGG	3’	61.90	20
	reverse	5’	GACCCTCTTCTTCCGGGCA	3’	64.88	19
CDS8179	forward	5’	CACCATGTCGTCGGTGTTCAGCG	3’	63.32	23
	reverse	5’	ATCGGCCCGGGATTG	3’	61.44	15
CDS2906	forward	5’	CACCATGGACGAGGCGGCG	3’	61.8	19
	reverse	5’	AATTCGGGTAACACTTATTTTTATC	3’	55.4	25
CDS1232	forward	5’	CACCATGGAGAAGATGCTCCACGC	3’	62.67	24
	reverse	5’	GAGCTCCTTCTGCACGTCCT	3’	62.06	20
CDS2645	forward	5’	CACCATGGTCTGCGTCGCGT	3’	59.94	20
	reverse	5’	TTCCTCTTTCTTATCTTCTGCCG	3’	61.15	23
CDS492	forward	5’	CACCATGACGACGGCGGCG	3’	63.0	19
	reverse	5’	CTTCATGTACCTCTGAAGAAATTTA	3’	55.1	25
CDS7565	forward	5’	CACCATGTCGACGCCGGCG	3’	63.0	19
	reverse	5’	GTCGGCTGTCCAGCTCT	3’	56.1	17
CDS8242	forward	5’	CACCATGAGTCTCTTCGGGCTTG	3’	57.95	23
	reverse	5’	GTAAGGGACGATGATGGATTC	3’	58.32	21
CDS3751	forward	5’	CACCATGCCGATCAGCAGGAT	3’	57.52	21
	reverse	5’	GATAACGCCAGCGATGC	3’	58.85	17
CDS2602	forward	5’	CACCATGGCCACCATCCTGG	3’	57.3	20
	reverse	5’	GTGCGCCATGGCGTG	3’	62.0	15
CDS2008	forward	5’	CACCATGGCGACCTCCTTCCA	3’	60.15	21
	reverse	5’	ATACTCTTCTTCATCCTCCAGTGA	3’	58.42	24
CDS10413	forward	5’	CACCATGGACGGAGCTCCGG	3’	59.4	20
	reverse	5’	ATATCGGAAATTGGGGTTC	3’	55.1	19
CDS9265	forward	5’	CACCATGCGCGTTGCCGCGACTC	3’	56.38	23
	reverse	5’	CTGGCTGCTCGACTCATCATGC	3’	56.86	22
CDS3352	forward	5’	CACCATGGGCTTCCCCGTG	3’	58.82	19
	reverse	5’	CTGGAAGCCGCCGAG	3’	60.17	15
CDS3310	forward	5’	CACCATGCATCCAACGCGTT	3’	55.5	20
	reverse	5’	ATACTCATGTTCATCCTCCTCC	3’	56.0	22
CDS11014	forward	5’	CACCATGGCAGAGGGCAACAACT	3’	60.67	23
	reverse	5’	CTTGAGGAAGCCCTGGG	3’	59.27	17
CDS1584	forward	5’	CACCATGGTGAAGATCTGCTGCATCG	3’	64.77	26
	reverse	5’	ATGGCCGGGGGTCTCC	3’	64.38	16
CDS9564	forward	5’	CACCATGGTGAACGGCGAGC	3’	59.25	20
	reverse	5’	GAACCCAGATAGAGGGAACG	3’	58.59	20
CDS658	forward	5’	CACCATGAGGCGGTGGGTCTT	3’	58.98	21
	reverse	5’	CCAAGGCTTCAGGTCGC	3’	60.50	17
CDS2626	forward	5’	CACCATGGGTATACTGGTGATTACGG	3’	57.77	26
	reverse	5’	CATGACAGAGCAGGAGTTCTCA	3’	60.59	22
CDS8068	forward	5’	CACCATGAAGAGATGGGGCAGCAG	3’	62.65	24
	reverse	5’	ATAGGGAATCCAATCCTCGGTA	3’	60.86	22
CDS11178	forward	5’	CACCATGGCGAAGAGCGAGGGT	3’	63.25	22
	reverse	5’	ATCGACCTCCTCGATCTTAGGAC	3’	62.13	23
CDS8892	forward	5’	CACCATGGAGAGGCCGGCG	3’	60.7	19
	reverse	5’	AACCAGCGCAATGCTG	3’	56.5	16

continua na próxima página

Tabela 8.1 – continuação da página anterior

CDS	Direção		Primer		TM	Tamanho
CDS9420	forward	5'	CACCATGGGGAAGTACACGCG	3'	57.97	21
	reverse	5'	GGTGAATGACTGTACCACGG	3'	58.85	20
CDS3470	forward	5'	CACCATGGACGGCGCCCGA	3'	64.8	19
	reverse	5'	ACCAAACATATTATGCGGGG	3'	58.5	20
CDS7050	forward	5'	CACCATGGCGGACCAGCTCAC	3'	60.83	21
	reverse	5'	CTTGGCCATCATAACCTTGA	3'	58.57	20
CDS6535	forward	5'	CACCATGGGGAGGGGGCGG	3'	64.1	19
	reverse	5'	GGGTAGCCATGTCGGC	3'	57.1	16
CDS2160	forward	5'	CACCATGGCATCCCCGGCC	3'	64.06	19
	reverse	5'	CGGGCAGAACGTCACCTT	3'	62.26	18
CDS1385	forward	5'	CACCATGCAGATCTTTGTCAAGACCCTTACG	3'	55.90	31
	reverse	5'	GGCGGTGGGGGGCTTG	3'	55.72	16
CDS10100	forward	5'	CACCATGGCGATCCGCAGG	3'	60.28	19
	reverse	5'	GAGCTCATCATGCTTCTCATCA	3'	60.51	22
CDS8814	forward	5'	CACCATGGCTTCAGCTGGTGTAG	3'	56.40	23
	reverse	5'	CAATCCGACATAGGAGCATT	3'	57.64	20
CDS11007	forward	5'	CACCATGTGGTGCGCATCGTG	3'	61.39	21
	reverse	5'	GGAGAAATCACGGTCTGGGAAG	3'	64.29	22
CDS2676	forward	5'	CACCATGTCAACTGCCACCGC	3'	59.15	21
	reverse	5'	GTACAGCTCCTCCTCCTTGTG	3'	58.95	21
CDS300	forward	5'	CACCATGGACCTGGTGCCGCACC	3'	56.38	23
	reverse	5'	AGCCGCTGCGCCGGG	3'	55.44	15
CDS2972	forward	5'	CACCATGGCCAAGGACGACG	3'	59.07	20
	reverse	5'	CGCGTTGCTCCTGAAGG	3'	61.72	17
CDS8709	forward	5'	CACCATGGTGCGGGGCAAG	3'	60.15	19
	reverse	5'	GCCTGACCTGATCGCTACTG	3'	60.97	20
CDS10968	forward	5'	CACCATGAGTTCACTAAATAAAGTTGTTTCA	3'	56.92	31
	reverse	5'	GTACACCATCACTCGGAGCA	3'	59.71	20
CDS8133	forward	5'	CACCATGGCGGAGGAGCTGGT	3'	61.81	21
	reverse	5'	TTTCTTCAGCAATTGTTTCAACAC	3'	60.54	24
CDS2643	forward	5'	CACCATGTCTCATCTGAGCATTCTCTC	3'	56.5	27
	reverse	5'	CGGCTTCCTCCGCGG	3'	64.0	15
CDS10795	forward	5'	CACCATGGCCAAGGGGGAG	3'	57.68	19
	reverse	5'	GTCGACCTCCTCGATCTTG	3'	58.31	19

Validação dos vetores de fusão e dos controles para avaliação da localização subcelular

Construções controles para diferentes localizações subcelulares foram selecionadas na literatura especializada e solicitadas a seus respectivos autores (Tabela 8.2). A partir de cada construção recebida, foram feitas colônias permanentes e extração dos respectivos DNAs.

Todas as construções utilizadas como controles de localização subcelular, assim

Tabela 8.2: **Construções selecionadas como controles de localização subcelular.**

Construção	Localização subcelular	Referência
pCAMBIA-1302	citoplasma	Hajdukiewicz et al. 1994
RecA-GFP	cloroplasto	Köhler et al. 1997a
ERD2-GFP e STtmd-GFP	Golgi	Boevink et al. 1998
ARR1-GFP e ARR6-GFP	núcleo	Hwang and Sheen 2001
β-GFP	mitocondria	Duby et al. 2001
SYCO-GFP e SYCO-RFP	mitocondria e cloroplasto	Peeters et al. 2000
GFP-PST1	peroxissomo	Mano et al. 2002
pSGFP5K/pGFPKDEL	retículo endoplasmático	
pSGFP5/psecGFP	via de secreção	Di Sansebastiano el al. 1998
pSGFP5T/pGFPChi	vacúolo	
pAleuGFP	vacúolo lítico	Di Sansebastiano el al. 2001

como as duas construções utilizadas na obtenção da fusão com a proteína codificada pelo gene *gfp*, tiveram suas identidades validadas por digestão enzimática. Ensaios de expressão transiente em epitélio de cebola foram realizados visando à validação do perfil de expressão no que tange a localização subcelular. A confirmação da identidade dos controles por digestão enzimática baseou-se nos mapas de restrição fornecidos pelos autores das mesmas. Já a confirmação da identidade dos dois vetores de expressão da fusão foi realizada por PCR (utilizando os primers *M13*) e digestão enzimática com a enzima *NcoI*, capaz de diferenciar as duas construções entre si.

Material vegetal, cultura de células, cultura de calos e obtenção de protoplastos

Os calos foram fornecidos pelo CTC (Centro de Tecnologia Canavieira, Piracicaba, SP) e mantidos por subcultivo em nosso laboratório em meio MS (Murashige and Skoog, 1962) suplementado com ácido nicotínico (0,5 mg/L), piridoxina.HCl (0,5 mg/L), tiamina.HCl (0,1 mg/L), sacarose (30 g/L), ácido 2,4-diclorofenoxiacético (1,5 mg/L), caseína hidrolizada (500 mg/L) e água de coco (5%), de acordo com protocolo padronizado por Falco (1994). Já as suspensões celulares de cana-de-açúcar (linhagem SUG2847) foram fornecidas pelo laboratório do prof. Marcio de Castro (ESALQ/USP, Piracicaba, SP), sendo estas

oriundas de calos da variedade SP83-2847. O meio de cultivo destas culturas em suspensão possui os mesmos componentes do meio semi-sólido descrito acima, com exceção do ácido 2,4-diclorofenoxiacético que é utilizado na concentração de 3 mg/L, sendo que o subcultivo é realizado a cada 5 dias.

Os ensaios para obtenção de protoplastos foram realizados a partir da cultura de calos assim como das de células em suspensão de acordo com protocolo padronizado para cana-de-açúcar (Falco, 1994).

Transformação de células de cana-de-açúcar e epitélio de cebola

Os ensaios de transformação de protoplastos de cana-de-açúcar para a posterior visualização da localização das proteínas fluorescentes foram realizados pelo emprego de polietilenoglicol (PEG), que permite a passagem passiva do DNA para dentro da célula vegetal. O procedimento adotado foi baseado no protocolo descrito por Chen *et al.* (1987), onde em 1 ml do meio de cultura com protoplasto de cana-de-açúcar (2,0 x 10^6 células) é adicionado 30 mg do DNA plasmidial e 50 mg de DNA de esperma de salmão, assim como 0,5 ml de PEG 6000 (40%). Após incubação de 30 minutos a 25°C, a solução é diluída por adições graduais da solução F (Krens *et al.*, 1982) que contém 111 mM glicose, pH 5,6 (5 x 2,0 ml a cada 25 minutos).

Nas transformações de epitélio de cebola utilizamos o sistema de transformação gênica por biobalística de hélio (Bio-Rad, EUA), amplamente usado para a determinação da localização subcelular (Scott *et al.*, 1999). O tecido a ser transformado foi depositado sobre meio MS sólido (Murashige and Skoog, 1962). Cinco miligramas do DNA plasmidial purificado em coluna (Qiagen, EUA) e dializado em água foram precipitadas sobre partículas de ouro (Bio-Raid, EUA) usando $CaCl_2$ e espermidina (Sigma, EUA) tal como descrito por Menossi *et al.* (1997). As partículas revestidas com DNA foram então lavadas em etanol absoluto, ressuspendida em etanol, e usadas para o bombardeamento do tecido em 1200 psi.

Avaliação da localização subcelular das proteínas de cana-de-açúcar fusionadas ao gene gfp

As avaliações da localização subcelular foram realizadas em microscópio de fluorescência modelo Leica DMI4000 (Leica, Alemanha), visando assim a detecção (com filtro de 505 a 550 nm) da emissão da GFP (excitada em 488 nm). A captura digital das imagens foi

realizada utilizando-se o software Leica IM50 (Leica, Alemanha). Os padrões de localização subcelular foram comparados aos obtidos utilizando-se os controles testados.

Resultados e discussão

Determinação das prováveis proteínas de cana-de-açúcar

Neste trabalho todos os CDS selecionados começam com o códon ATG e terminam com o códon de parada, fornecendo assim um conjunto com 11.882 seqüências. Deste conjunto buscamos os CDS que, por homologia com seqüências protéicas, forneciam maiores indícios de que a estrutura de sua ORF estava correta (principalmente pela presença do primeiro aminoácido metionina da proteína como sendo codificado pela primeira trinca do CDS). Ao utilizar esta premissa, nosso conjunto foi reduzido para 1.519 CDS.

Com o objetivo de caracterizar e avaliar funcionalmente proteínas de cana-de-açúcar, a partir desse grupo foi selecionado 65 CDS devido ao potencial biotecnológico evidenciado em estudos de expressão gênica (Tabela 8.3). Por um lado temos genes relacionados às vias de transdução de sinal cuja expressão diferencial relacionada à síntese de açúcar foi observada em folhas de populações segregantes para teor de açúcar (Felix *et al.*; submetido) ou associada ao acúmulo de açúcar em entrenós de populações segregantes para teor de açúcar ao longo do desenvolvimento e da maturação. Outros genes, identificados por experimentos de microarrays, que apresentam expressão constitutiva ou tecido preferencial (Papini-Terzi *et al.*, 2005), assim como expressão diferencial em experimentos com fitohormônios e de resposta ambiental (Rocha *et al.*, 2007) foram selecionados, e finalmente SASs representando genes altamente expressos que codificam proteínas desconhecidas (provavelmente as mais abundantes) no projeto SUCEST. Foram selecionados outros 31 CDS (Tabela 8.3) para complementar os 65 CDSs selecionados previamente, finalizando 96 CDS depositados em uma placa de microtitulação. A seleção destes novos CDS foi baseada em localizações subcelulares pouco representadas no conjunto inicial.

Tabela 8.3: **CDS selecionados para a determinação da localização subcelulares de suas respectivas proteínas.** (A) CDSs codificados por cDNAs que apresentam expressão diferencial em folha de plantas de populações segregantes para teor de açúcar (Felix *at al.*; submetido). (B) CDSs codificados por cDNAs que apresentam expressão diferencial em entrenós de plantas de populações segregantes para teor de açúcar ao longo do desenvolvimento e da maturação. Os sinais - e + indicam repressão ou indução respectivamente, em plantas com alto teor de sacarose, em relação a plantas com baixo teor. (C) CDSs codificados por cDNAs que apresentam padrão de expressão sem alteração, não apresentam expressão preferencial em nenhum dos tecidos de cana (Papini-Terzi *et al.*, 2005). (D) CDSs codificados por cDNAs que apresentam expressão diferencial em diversos tecidos de cana-de-açúcar (Papini-Terzi *et al.*, 2005). Os sinais - e + indicam, respectivamente, menor ou maior expressão em relação à amostra de referência (pool de tecidos). (E) CDSs codificados por SASs que apresentam identidade com proteínas hipotéticas. (F) CDSs que apresentam localizações subcelulares pouco representadas no conjunto inicial. (G) Nome codificado do gene. (H) localização subcelular predita pelo nosso método. (I) localização subcelular de possíveis genes homólogos presentes no banco de dados de referência DBSubLoc (Guo *et al.*, 2004*b*). (J) Perfil de expressão gênica determinado por experimentos de microarrays, quando disponível. (K) Expressão diferencial em experimentos com fitohormônios e de resposta ambiental (Rocha *et al.*, 2007). Legenda: C, cloroplasto; G, complexo de Golgi; M, mitocôndria; N, núcleo; O, outras localizações; P, peroxissomo; R, retículo endoplasmático; S, sinal para secreção; V, vacúolo; X, citoplasma; n.d., sem homologia; MAP, meses após plantio.

	CDS[G]	Predição[H]	DBSubLoc[I]	Perfil de Expressão Gênica[J]						Perfil de Expressão Gênica[K]
						Alto Brix				
A	SCCDS1967	OS	membrane			-				
	SCCDS517	MX	nucleus			-				Drought 72h (+)
	SCCDS2643	M	n.d.			-				
					7 MAP			11 MAP		
				Colmo 1	Colmo 5	Colmo 9	Colmo 1	Colmo 5	Colmo 9	
	SCCDS11178	NX	ER	+	-	+	+	+	+	Drought 24h (+)
	SCCDS10795	NX	ER	+	+	+				
	SCCDS3751	S	membrane	-	-	+	-	-	-	
	SCCDS4197	M	Golgi complex	+	+	+				
	SCCDS6243	S	membrane	-	+	+	-	-	-	
	SCCDS1967	OS	membrane	+	+	+	-	-	-	
	SCCDS10295	NX	protoplasm	-	+	+				
	SCCDS11691	N	n.d.				-	-	-	
	SCCDS10413	NX	nucleus	-	-	+				
B	SCCDS3310	NX	n.d.				+	+	-	Herbivory 30min (+)
	SCCDS2160	S	cell wall	+	-	+				
	SCCDS2972	OS	membrane	+	+	+	-	-	-	
	SCCDS2008	NX	n.d.	+	+	-	-	-	-	
	SCCDS8278	X	cytoplasm	+	+	-				Drought 72h e 120h (-), ABA 12h (+), MeJA 6h e 12h (-)
	SCCDS11026	C	n.d.	+	-	+				
	SCCDS10829	OS	membrane	+	-	+				
	SCCDS11056	NX	cytoplasm	+	+	+				
	SCCDS9564	X	nucleus				-	-	-	Drought 120h (-)
	SCCDS4642	MNX	n.d.	+	+	+				
	SCCDS5183	CP	chloroplast	+	+	-				
	SCCDS1385	N	nucleus	+	+	+				
	SCCDS2626	OS	n.d.	+	+	+	+	+	+	Drought 72h (-), Drought 120h (-)
	SCCDS8242	PX	n.d.	+	+	+	-	-	+	
	SCCDS6252	NX	nucleus							
	SCCDS8136	X	n.d.							
	SCCDS2906	NX	n.d.							
	SCCDS7050	NX	n.d.							
	SCCDS11691	N	n.d.							

Continua na próxima página

Tabela 8.3 – continuação da página anterior

	CDS[G]	Predição[H]	DBSubLoc[I]	Perfil de Expressão Gênica[J]						Perfil de Expressão Gênica[K]
	SCCDS5956	P	nucleus							
	SCCDS9265	MNSX	n.d.							
	SCCDS10970	X	n.d.							
	SCCDS1232	X	n.d.							
	SCCDS10592	N	nucleus							
	SCCDS1038	C	chloroplast							
				Flor	Gema lateral	Folha	Raiz	Colmo 1	Colmo 4	
	SCCDS8814	NX	nucleus				-			
	SCCDS3751	S	membrane				+			
	SCCDS8709	N	nucleus				-			
	SCCDS4119	X	nucleus		+					
	SCCDS4197	M	Golgi complex						+	
	SCCDS6243	S	membrane	-						
	SCCDS1967	OS	membrane			-	+			
	SCCDS3121	NX	nucleus			+				
	SCCDS10295	NX	protoplasm			-				Drought 72h (-)
	SCCDS7565	MNX	n.d.						+	
	SCCDS3310	NX	n.d.			-				
	SCCDS2160	S	cell wall		-			-		Herbivory 24h (+)
D	SCCDS2972	OS	membrane			-				MeJA 6h (-)
	SCCDS10100	V	ER			-		-	-	
	SCCDS2008	NX	n.d.			-				
	SCCDS8278	X	cytoplasm			+	-	-		
	SCCDS11026	C	n.d.	-	-	-		-	-	*Herbaspirillum spp.* (+)
	SCCDS10829	OS	membrane			-	+			
	SCCDS9840	S	ER					-		
	SCCDS2561	S	membrane				+			
	SCCDS4642	MNX	n.d.						+	
	SCCDS5183	CP	chloroplast	-						Drought 120h (-)
	SCCDS11117	M	chloroplast			+	-			Drought 72h (-), Drought 120h (-), ABA 12h (+), MeJA 6h (-)
	SCCDS6535	N	nucleus				-			
	SCCDS6228	CMN	n.d.			+				
	SCCDS2626	OS	n.d.						+	
	SCCDS4123	C	n.d.	+						
	SCCDS8133	R	n.d.							
	SCCDS890	S	n.d.							
	SCCDS7976	X	n.d.							
	SCCDS1074	OS	n.d.							
	SCCDS3242	NX	mitochondrion							
	SCCDS11014	C	n.d.							
	SCCDS2273	X	n.d.							
	SCCDS4583	X	n.d.							
E	SCCDS1286	X	n.d.							
	SCCDS300	N	n.d.							
	SCCDS10968	C	n.d.							
	SCCDS1024	O	n.d.							
	SCCDS6524	C	n.d.							
	SCCDS2602	NP	n.d.							
	SCCDS125	NX	n.d.							
	SCCDS492	N	n.d.							
	SCCDS658	S	n.d.							
	SCCDS8892	X	n.d.							
	SCCDS3470	R	n.d.							
	SCCDS2152	P	n.d.							
	SCCDS8068	P	n.d.							
	SCCDS3321	G	n.d.							
	SCCDS10984	R	n.d.							
	SCCDS1864	P	n.d.							
	SCCDS3725	P	n.d.							
	SCCDS1584	R	n.d.							
	SCCDS8179	R	membrane							
	SCCDS1092	P	protoplasm							
	SCCDS9529	R	n.d.							
	SCCDS6392	V	n.d.							
	SCCDS2447	R	n.d.							
	SCCDS9723	P	n.d.							
	SCCDS1959	V	membrane							
F	SCCDS5587	V	membrane							
	SCCDS2645	V	n.d.							
	SCCDS4650	R	n.d.							
	SCCDS7778	P	n.d.							
	SCCDS2210	R	n.d.							
	SCCDS7965	R	n.d.							
	SCCDS11518	V	n.d.							
	SCCDS11487	V	n.d.							
	SCCDS2474	R	n.d.							
	SCCDS9420	R	n.d.							

continua na próxima página

Tabela 8.3 – continuação da página anterior

CDS[G]	Predição[H]	DBSubLoc[I]	Perfil de Expressão Gênica[J]	Perfil de Expressão Gênica[K]
SCCDS11007	R	n.d.		
SCCDS3352	R	n.d.		
SCCDS11693	R	n.d.		
SCCDS4157	R	n.d.		
SCCDS157	PX	n.d.		
SCCDS2676	C	chloroplast		

Amplificação dos genes de interesse visando fusão com o gene *gfp*

Uma vez tendo obtido os DNAs necessários, foi realizada a amplificação visando à obtenção dos CDSs a serem utilizados nas clonagens. Com as amplificações em larga escala, foram obtidos cerca de 70% de produtos específicos. A Figura 8.2 ilustra os CDS que foram clonados com sucesso no vetor *p2FGW7* e que desta foram tiveram suas clonagens realizadas no vetor *p2GWF7*.

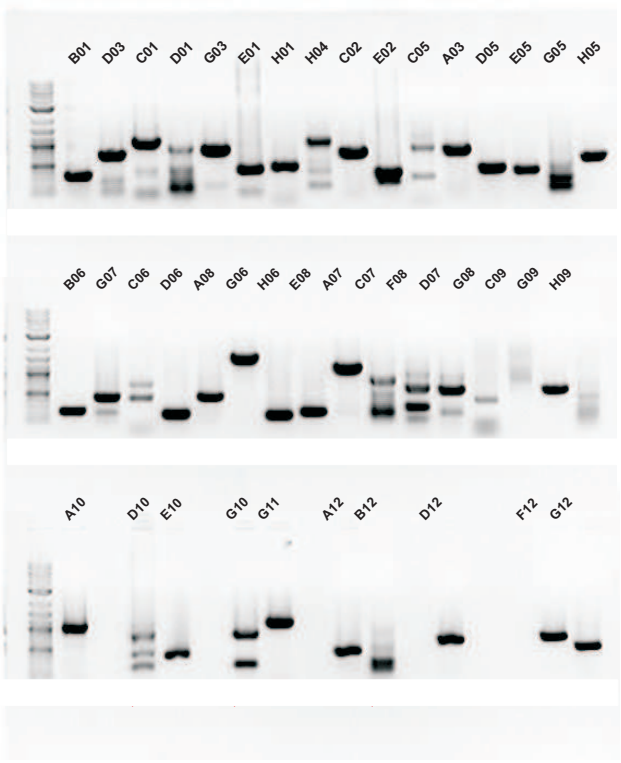


Figura 8.2: **PCR dos 42 CDS clonados com sucesso no vetor *p2FGW7*.** Gel de agarose 1% contendo 10 μL de cada reação de PCR.

Clonagens dos genes de interesse visando fusão com a proteína codificada pelo gene *gfp*

Com a realização das clonagens visando a obtenção de construções com fusão amino-terminal, à proteína GFP, dos os 96 CDS previamente selecionados (utilizando-se o vetor *p2FGW7*) obtivemos o seguinte resultado: 44% (42 CDS) das clonagens neste vetor de destino foram realizadas com sucesso, confirmação esta fornecida pelo seqüenciamento dos respectivos DNAs. Com isto, todos estes 42 CDS foram clonados no vetor *p2GWF7*, visando a obtenção de construções com a fusão carboxi-terminal.

Na Figura 8.3 é apresentada a ontologia atribuída a estas 42 proteínas codificadas pelos genes já clonados (componente celular, Figura 8.3A; e processo biológico, Figura 8.3B). Já a Figura 8.4 apresenta a distribuição subcelular predita das proteínas codificadas por estes genes, onde estes são indicados por sua classificação funcional. Estas figuras mostram que este conjunto de 42 CDS, apesar de ser uma parcela do conjunto inicial, é muito representativo tanto no que se refere a distribuição dentro da célula quanto a processos moleculares de extrema relevância para a cana-de-açúcar.

Validação dos vetores de fusão à proteína codificada pelo gene *gfp*

Uma vez recebidas as construções a serem utilizadas como vetores de expressão da fusão, passamos para o processo de confirmação da identidade e da expressão em células vegetais. A Figura 8.5 apresenta a confirmação das identidades das construções *p2GWF7* e *p2FGW7*. Com a identidade confirmada dos clones, foi realizado o ensaio de expressão transiente visando a confirmação da expressão do gene *gfp*. Este ensaio apresentou os resultados esperados; emissão de fluorescência em níveis detectáveis e localização dispersa no citoplasma e no núcleo (Figura 8.6).

Obtenção de protoplastos

A realização de ensaios para obtenção de protoplastos a partir de cultura de calos não forneceu resultados positivos. Imediatamente após o período estabelecido para a digestão da parede celular não foi possível verificar a existência de protoplastos viáveis a partir do uso de azul de metileno como corante positivo (dados não mostrados). Esta observação repetiu-se após as lavagens da solução de células. Procedimento similar foi realizado para a cultura de

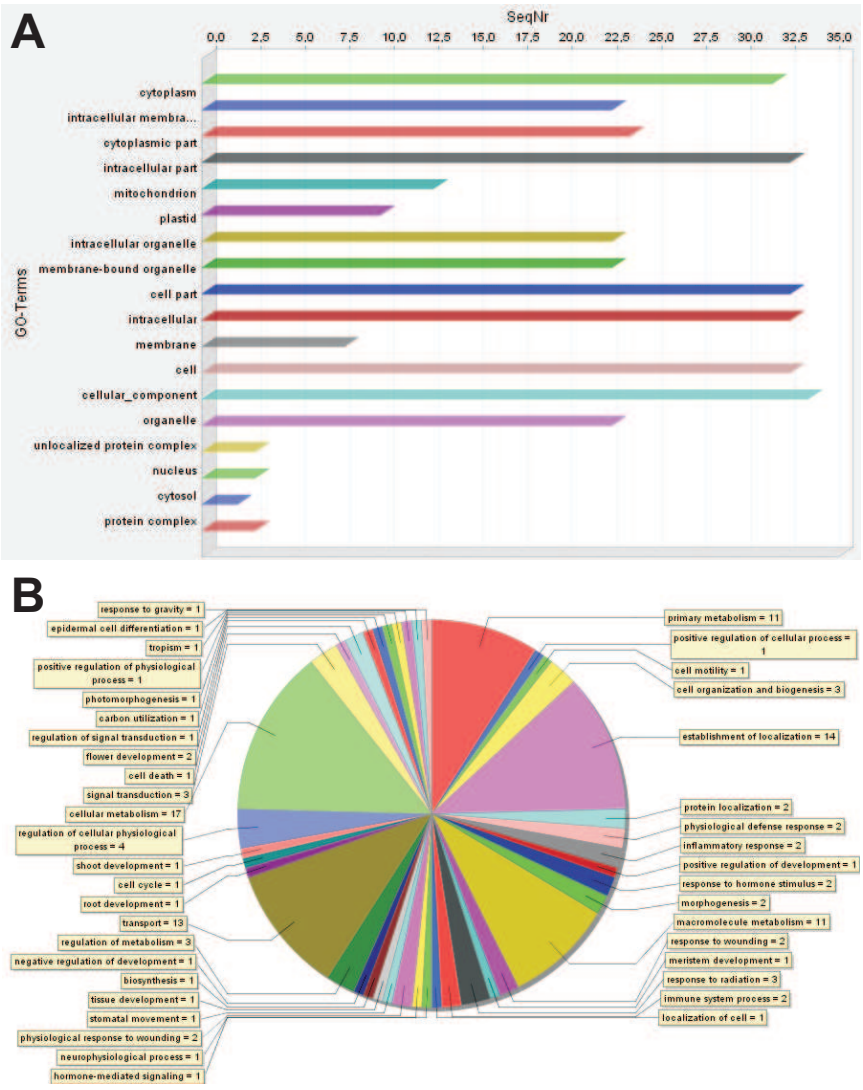


Figura 8.3: **Classificação por ontologias dos genes já clonados.** Em A é apresentada a distribuição destes genes em relação aos diversos componentes celulares; e B em relação aos processos biológicos.

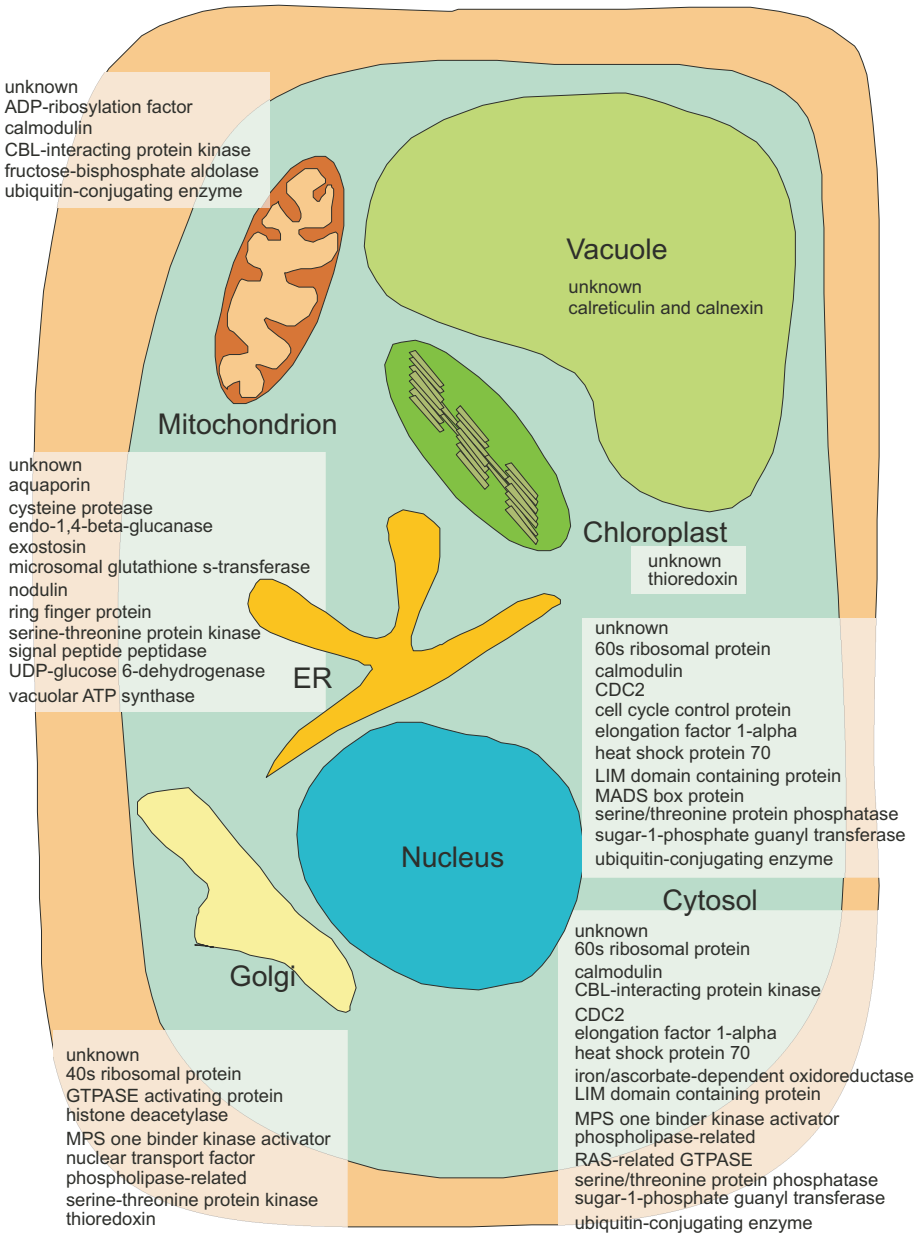


Figura 8.4: **Distribuição celular das proteínas codificadas pelos genes já clonados.**

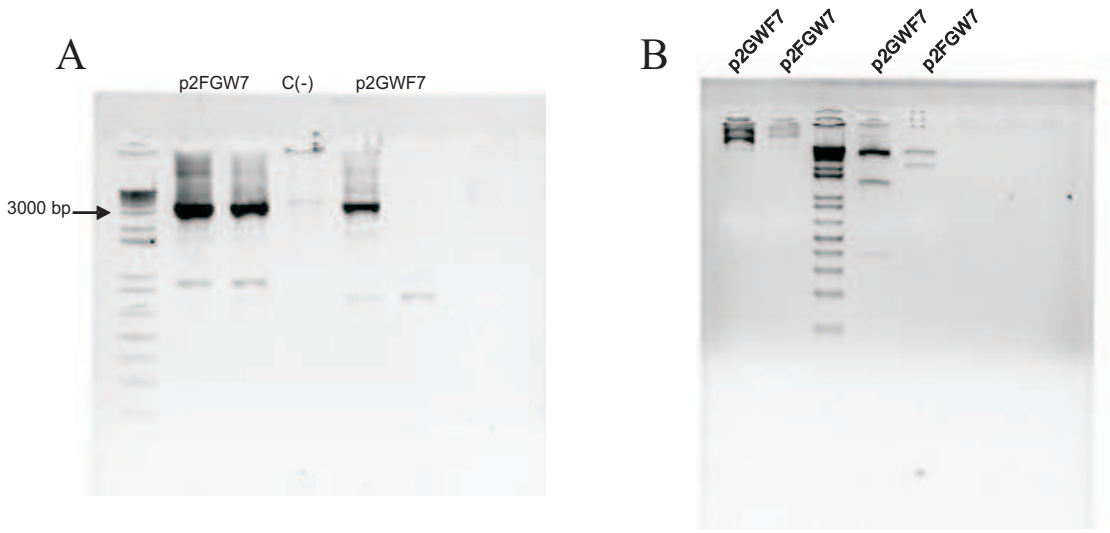


Figura 8.5: **Obtenção e confirmação da identidade dos clones com as construções a serem utilizadas na expressão da fusão.** Em A é apresentado o resultado da PCR de duas colônias de cada construção utilizando-se os *primers* universais M13, com resultados positivos para três colônias; e em B é apresentada a eletroforese feita com o DNA extraído para cada construção (colunas 1 e 2), assim como confirmação da identidade por digestão enzimática com a enzima *NcoI* (colunas 4 e 5).

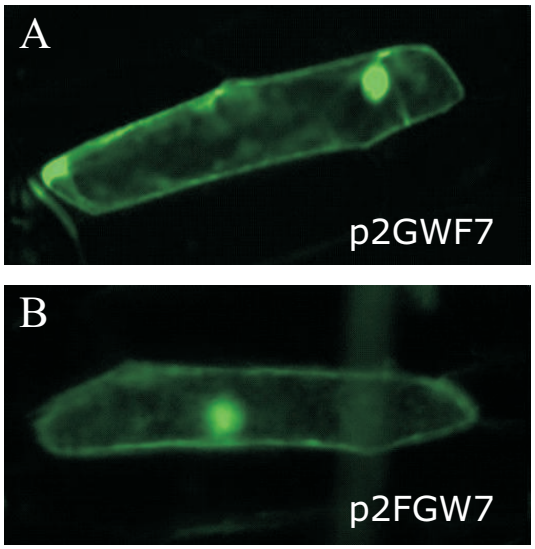


Figura 8.6: **Ensaio de expressão transiente dos vetores de expressão.** Epitélio de cebola bombardeado com os vetores *p2GWF7* (A) e *p2FGW7* (B) expressando a proteína GFP sob ação do promotor 35S. Aumento de 10X.

células em suspensão (Figura 8.7), onde neste caso foram obtido protoplastos viáveis, mais que não resultaram em eventos positivos de transformação, via protocolo estabelecido por PEG.

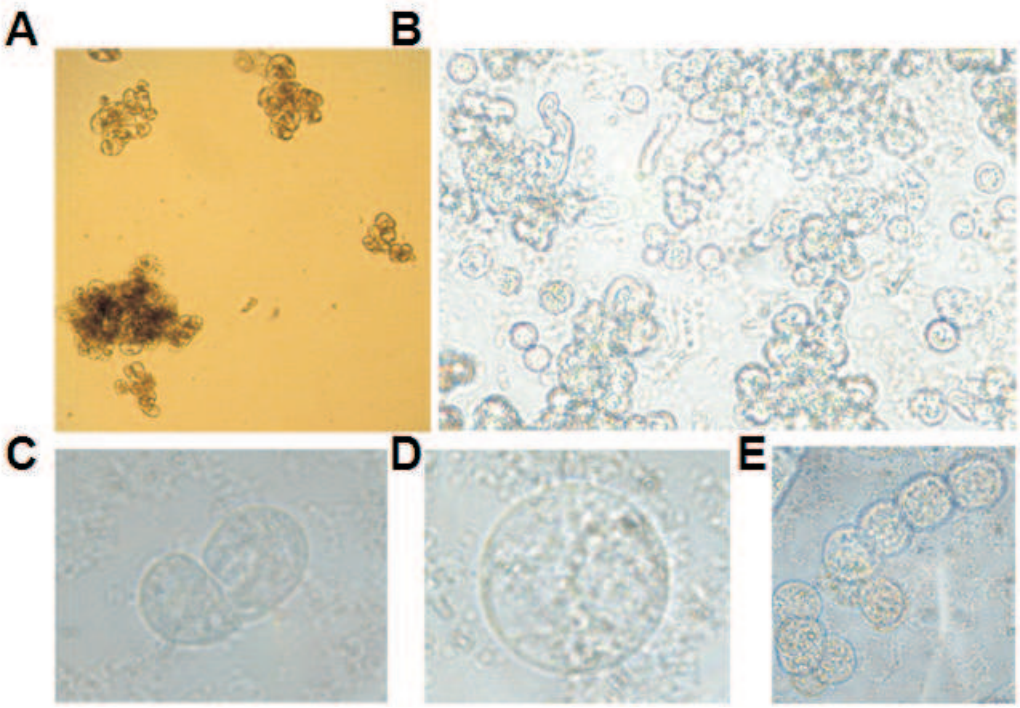


Figura 8.7: **Obtenção de protoplastos de cana-de-açúcar.** (A) Suspensão celular, (B-E) exemplos protoplastos isolados.

Validação do perfil de localização de proteínas controles para avaliação da localização subcelular

Foram realizados ensaios de expressão transiente em epitélio de cebola visando a validação do perfil de expressão no que tange a localização subcelular das construções a serem utilizadas como controle. Os resultados destes ensaios são apresentados na Figura 8.8.

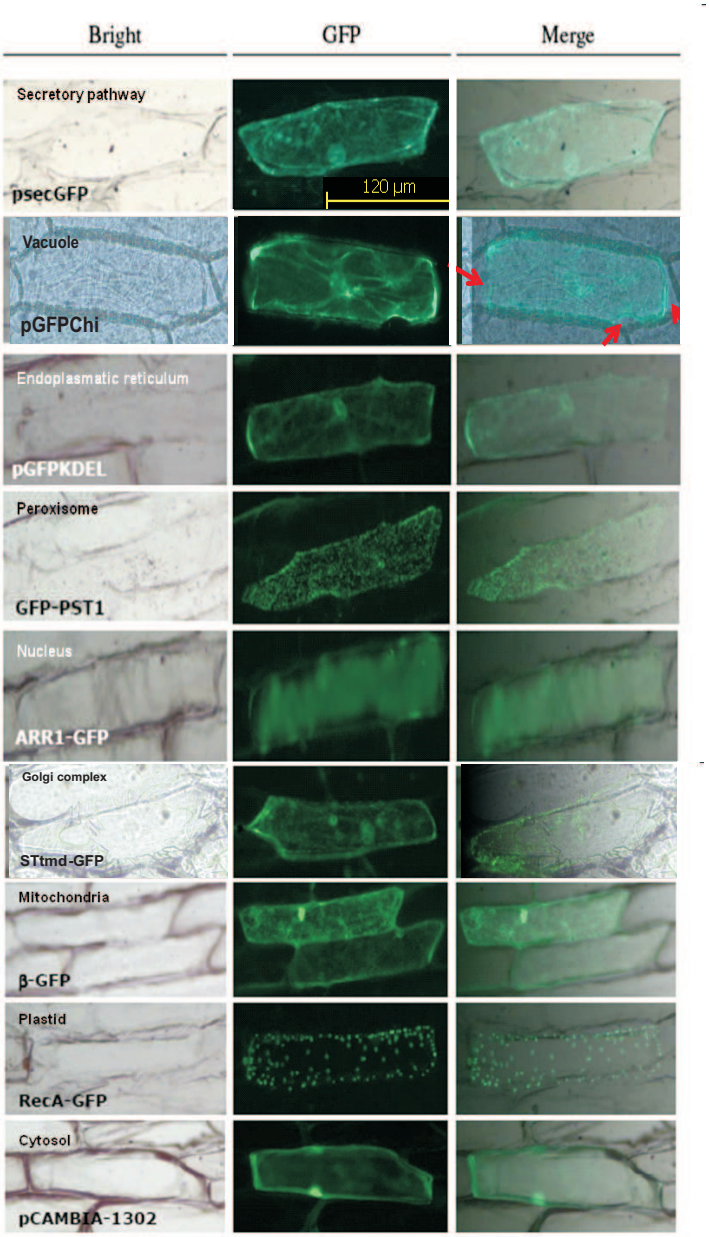


Figura 8.8: **Ensaios de expressão transiente dos controles para avaliação da localização subcelular.** Epitélio de cebola expressando as construções controles sob incidência de luz clara (coluna 1), luz ultravioleta sobre o mesmo campo (coluna 2) e a imagem resultante da junção das duas anteriores (coluna 3). Aumento de 10X.

Avaliação da localização subcelular das proteínas codificadas pelos genes clonados

Infelizmente e, principalmente, inesperadamente, as construções obtidas durante o estudo (com fusão tanto na extremidade amino quanto na carboxi) não forneceram nenhum resultado conclusivo, como visto em exemplos selecionados e apresentados na Figura 8.9.

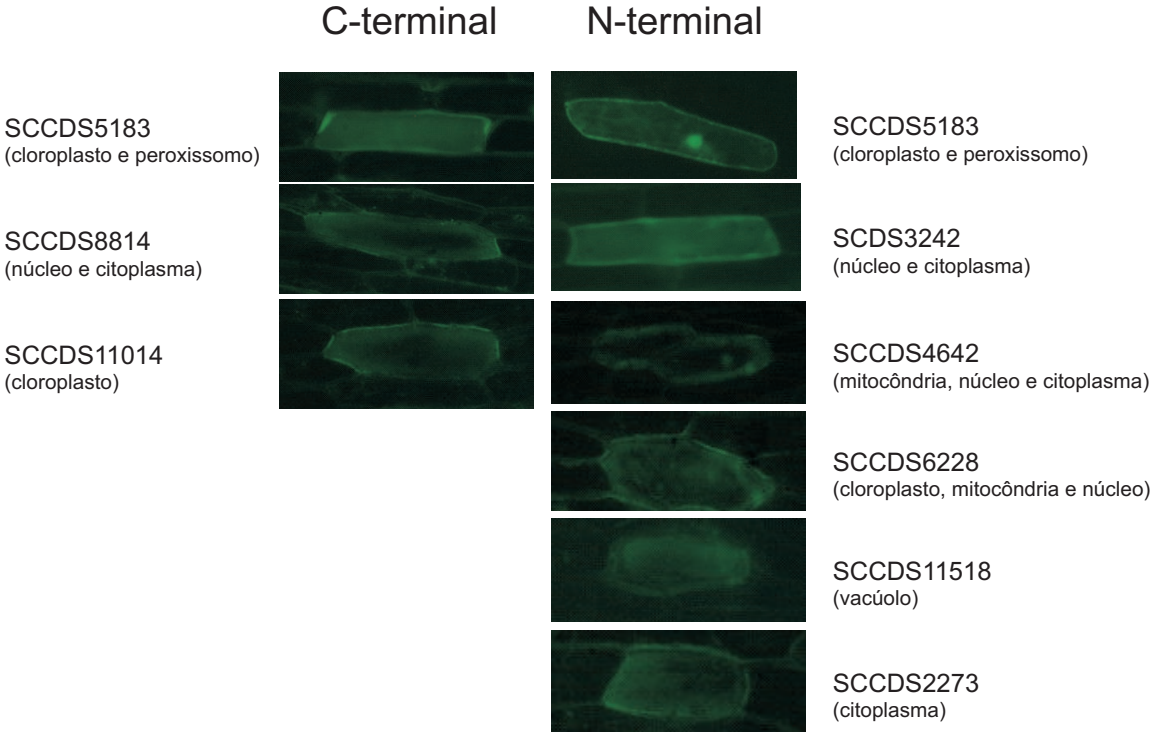


Figura 8.9: **Exemplos dos resultados dos ensaios de expressão transiente das construções obtidas.** Epitélio de cebola bombardeado com as construções com fusão carboxi- e amino-terminal. Entre parênteses é apresentada a localização subcellular predita. Aumento de 10X.

Diante deste resultado inesperado, algumas possibilidades podem ser levantadas. Uma vez que todas as 42 construções foram seqüenciadas e confirmadas, o principal fator que poderia gerar resultados inconclusivos fica por conta dos próprios vetores de destino (*p2FGW7* e *p2GWF7*). Como visto, estes vetores apresentam boa expressão da proteína GFP quando esta não esta fusionada a nenhuma outra proteína (Figura 8.6), e o mesmo ocorre, mas em menor intensidade, quando com a fusão ao gene *ScBAK1* (Capítulo 7). Foi observado também que em diversos experimentos ocorreu uma rápida queda na emissão da fluorescência assim que o tecido é colocado sobre luz ultravioleta. Uma vez que os experimentos com a

expressão do gene *ScBAK1* foram repetidos em um número muito maior do que os com as 42 construções, fica a possibilidade da necessidade de condições ainda não padronizadas para a observação da expressão destas construções.

Conclusão

Neste capítulo são apresentados os resultados obtidos durante o estudo que teve como objetivo caracterizar proteínas de cana-de-açúcar (*Saccharum* spp.) no que se refere à localização subcelular. Infelizmente estes resultados não foram conclusivos, possivelmente por problemas metodológicos. Nosso grupo está buscando alternativas que possam contornar esse problema, dentre elas a utilização de microscopia confocal e a transformação de protoplastos de milho, embora ainda insistamos com a biobalística em cebola e transformação de protoplastos de cana.

Agradecimentos

Este trabalho foi financiado pela FAPESP, através do auxilio 05/58104-0 concedido a Marcelo Menossi.

Neste dia perfeito em que tudo amadurece e não somente a uva começa a tomar uma cor escura, um raio de sol cai sobre minha vida: olhei para trás e olhei para frente e nunca havia visto de uma só vez tantas e tão boas coisas.

Friedrich Nietzsche (1844-1900)

9

TISs-ST: a web server to evaluate polymorphic translation initiation sites and their reflections on the secretory targets†

Renato Vicentini[1,2] and Marcelo Menossi[1,2]

Abstract

Background: The nucleotide sequence flanking the translation initiation codon (start codon context) affects the translational efficiency of eukaryotic mRNAs, and may indicate the presence of an alternative translation initiation site (TIS) to produce proteins with different properties. Multi-targeting may reflect the translational variability of these other protein forms. In this paper we present a web server that performs computations to investigate the usage of alternative translation initiation sites for the synthesis of new protein variants that might have different functions.

Results: An efficient web-based tool entitled TISs-ST (Translation Initiation Sites and Secretory Targets) evaluates putative translation initiation sites and indicates the prediction of a signal peptide of the protein encoded from this site. The TISs-ST web server is freely available

†*BMC Bioinformatics* 2007, **8**:160 (doi:10.1186/1471-2105-8-160)

[1]Functional Genomics Laboratory, Center for Molecular Biology and Genetic Engineering, State University of Campinas, P.O.Box 6010, 13083-875, Campinas, SP, Brazil

[2]Department of Genetics and Evolution, Institute of Biology, State University of Campinas, Campinas, SP, Brazil

to both academic and commercial users and can be accessed at http://ipe.cbmeg.unicamp.br/pub/TISs-ST.

Conclusions: The program can be used to evaluate alternative translation initiation site consensus with user-specified sequences, based on their composition or on many position weight matrix models. TISs-ST provides analytical and visualization tools for evaluating the periodic frequency, the consensus pattern and the total information content of a sequence data set. A search option allows for the identification of signal peptides from predicted proteins using the PrediSi software.

Background

Translation by cytosolic ribosomes generally occurs at the first AUG in the transcript. However, in eukaryotic mRNAs, efficient recognition of an AUG codon as a translation initiation site (TIS) depends on several factors, such as the nucleotide sequence that flanks the site (Kozak, 1991, Kozak, 2002, Kozak, 2005). There is evidence that the context surrounding the initiation codon contributes to the control of translational initiation (Kawaguchi and Bailey-Serres, 2005). The sequence context of the first AUG codon, in particular that part located in the untranslated region, may modulate the efficiency with which it is recognized as a translation initiation codon (Mignone *et al.*, 2002). If the first initiation codon lies in a suitable context, protein synthesis will be started. When the context is less than favorable, most of the protein synthesis will start at the next downstream AUG codon (Nadershahi *et al.*, 2004). Moreover, other structural features of the mRNA are considered important for the efficiency of the translation initiation at a specific AUG codon, such as: the proximity of AUG to the 5' end, the secondary structure upstream and downstream from the AUG codon, the leader sequence length and the multiple upstream AUG codons (Kozak, 1991, Wang and Rothnagel, 2004, Kozak, 2005). Recent studies indicate that start codons of a large proportion of the human and mouse mRNAs reside in evolutionary conserved local loop structures, and some of these structures may be common in mammals and important for the efficient initiation of translation (Shabalina *et al.*, 2006, Kozak, 2002). The frequency of the nucleotides surrounding the initiation AUG (context) has been extensively analysed in sequences available in public databases (Cavener, 1987). The importance of a particular position in a sequence is more clearly and consistently given by the information required to describe the pattern. The information in the sequence patterns allows one to investigate how the information is

distributed across the sites and to compare one site to another (Schneider *et al.*, 1986).

Statistical analyses of the AUG initiation codon context in many organisms identified a preferential nucleotide frequency in some positions around the AUG. Recent analyses have revealed variations in the initiation context between different groups of eukaryotes. Distinct inter-taxon variations in the AUG context sequences are repeatedly observed when invertebrates, higher plants and protozoa are considered separately (Lukaszewicz *et al.*, 2000). For instance, in vertebrates, C(A/G)CCAUGG was observed to be a consensus sequence (Kozak, 1987). For plant genes, a consensus context was deduced as c(A/G)(C/A)CAUGGC for monocots and A(A/C)aAUGGC for eudicots (Joshi *et al.*, 1997, Lukaszewicz *et al.*, 2000).

Upstream out-frame AUG may severely affect the translation of a gene, even if surrounded by a poor context (Luehrsen and Walbot, 1994), suggesting that upstream AUGs may have a role in keeping the basal translation level of a gene low (Mignone *et al.*, 2002). Recently it was demonstrated that downstream AUG codons are utilized as alternative TISs even in mRNAs with multiple strong upstream AUGs (Wang and Rothnagel, 2004). Their occurrence must correlate with the start codon context: sub-optimal context should be accompanied by a higher frequency of downstream AUGs (Kochetov and Sarai, 2004). With this mechanism, called 'leaky scanning', multiple different proteins can be obtained from the same mRNA (Mignone *et al.*, 2002). In this sense AUGs located downstream of the major coding sequences (CDS), may play a role in generating protein diversity (Kozak, 2002). The usage of a closely located downstream in-frame AUG codon as an alternative TIS can result in full and N-truncated proteins that may have the same function and be targeted at the different compartments (Watanabe *et al.*, 2001, Kochetov and Sarai, 2004, Kochetov, 2005). Since eukaryotic mRNAs frequently contain TISs in a sub-optimal context (Rogozin *et al.*, 2001), the problems of polypeptide N-end heterogeneity and finding of the genuine TIS are very topical.

In silico determination of the sub-cellular localization of the proteins can provide information on their function, and is dependent on the correct identification of the first AUG and their potential N-terminally polymorphic forms. This translational polymorphism may serve as an important source of diversity in both cytoplasmic and organelle proteomes (Kochetov and Sarai, 2004, Kochetov, 2005).

Proteins must be localized correctly at the sub-cellular level to have normal biological functions (Xie *et al.*, 2005). When the final destination is the mitochondria, the chloroplast, or the secretory pathway, sorting usually relies on the presence of an N-terminal

targeting sequence (von Heijne *et al.*, 1989). In the secretory pathway, proteins designated for export from the cell are labelled by an N-terminal signal sequence (Hiller *et al.*, 2004). These signal peptides are responsible for targeting proteins to the ER for subsequent transport through the secretory pathway, and the prediction of signal peptides has become an important application of genomic and proteomic investigations. There are known cases of variation in the use of alternative signal peptides, and in the majority of cases this is due to the exclusion of the signal peptide from one or more protein products of the same gene. However in other cases, this variation involves the replacement of one signal peptide by another signal. For example, a single gene encoded 48 isoforms of protocadherin using 34 different signal peptides, each encoded by its own initial exon (Davis *et al.*, 2006). Alternative initial exon usage is the most common mechanism for replacing one signal peptide with another.

In this paper, we present the TISs-ST (Translation Initiation Sites and Secretory Targets) web server that investigated the usage of alternative TISs for synthesis of new protein variants possibly possessing different functions. This server deployed previously annotated complete CDSs retrieved from the NCBI UniGene database (Wheeler *et al.*, 2003) and the PrediSi prediction program (Hiller *et al.*, 2004) for inspection and evaluation of alternative TISs and also target prediction of the proteins encoded by this variable site. It can be useful for finding proteins that have signal peptides in the polymorphic form, for assisting research by evaluating alternative coding potentials for eukaryotic mRNAs, and in designing synthetically created genes, especially for maximizing the translational level of an interesting protein. TISs-ST uses user-specified sequences and an optional position weight matrix (PWM) model, derived computationally from a subset of the NCBI UniGene data set, to infer the consensus around the AUG sites. This subset only includes sequences previously annotated as complete CDS. The program determines the consensus sequence and the total information content around the AUGs in five situations: (i) first transcript AUG, (ii) second in-frame downstream AUG, (iii) second out-frame downstream AUG, (iv) all other in-frame downstream AUGs, and (v) all other out-frame downstream AUGs. This information is provided with the probability of alternative TIS based on the frequency of the AUG codons. The use of these five alternatives in the analyses can provide some advantages, mainly in the prediction of TIS originating in genes with alternative splicing. All scripts and interfaces were written in Perl and R languages. This version of program is available at TISs-ST web server.

Implementation

Description of the web server

We developed a web server named TISs-ST. Basically it can be divided into two subsystems: (i) the web interface system, which is written in the Perl and HTML languages and (ii) the background process system, which is written in the Perl and R statistical languages. The web interface subsystem mainly deals with the task of receiving information from the user and checking the validity of the data submitted. The background processing subsystem computes all the analytical and prediction tasks: extracts features from the sequences, computes and displays the consensus sequences and total information content, and predicts the signal peptides of the user sequences. We used PrediSi for the prediction of signal peptides in our implementation of TISs-ST.

The web interface allows for the easy evaluation of sequences (provided in FASTA format), for the presence of putative translation initiation sites and for the prediction of signal peptides. The TISs-ST interface is designed in such a way that the user specifies all necessary parameters in the initial page, and provides many possibilities to work with the sequence files. The content of a file can be pasted in the input window, or taken from a directory on a local computer. The user has the option of setting several parameters manually.

Input parameters

TISs-ST takes the following input data:

(i) DNA sequence data. This contains one or several sequences in FASTA format. The sequences can be in a nucleotide format, where the first ATG codon after the first 15 base pairs will be considered to be the one that encodes the initial Methionine. Since the analysis calculates the nucleotide frequencies in a context position, if the user does not choose to run the analysis with a pre-defined PWM, the amount of sequences must be bigger in order to obtain a significant result. A representative sample data is available on the website.

(ii) An optional predefined PWM model for species or species group data sets. Currently, 32 species data sets are available on the web server (Table 9.1) and these data can also be grouped into 20 data sets (12 phylogenetic Class, and 8 Phylum or Division).

(iii) A filter for the AUG sites.

(iv) An option for signal peptide prediction in the protein deduced from the submitted sequence(s). This option requires the selection of the genetic code used to translate the DNA sequences.

Output

After submitting the data to the server, the TISs-ST program searches the consensus sequences according to the parameters selected. While the analyses are running, the web server shows a checklist of the steps finalized, from which the user can estimate the total time required for the analysis. An example of the output of the TISs-ST program is shown in Figure 9.1. For each site selected, the final result is a summary table with detailed information on the analysis of the data set generated. Every AUG flanking site analysed by the program is shown in a separate row of the result table, including the name of the site analysed, the number of sequences submitted, the consensus sequence found, and the total information content with or without correction for bias (measured in bits). The last column shows hyperlinks to files that include the amino acid sequences and the signal peptide prediction of the data submitted. More detailed information about a consensus sequence and the probability of localizing alternative translation initiation sites is provided by hyperlinks at the top of the graphic results.

In addition to the tabulated output, the result page shows a graphic representation of the consensus and the information content found at each site analysed. There are also hyperlinks for each sequence pattern from the user sequences, the sequence header providing information about the probability of alternative TIS based on the frequency of AUG codons, and hyperlinks to frequency tables of this pattern generated in each analysis (Figure 9.1).

Graphical representation

A typical consensus graphical representation concentrates the following information into a single graph: the general consensus of the sequences; the predominance order of the nucleotides at every position (the most frequent nucleotides, in a same position, are showed before the less frequent); the relative frequencies of every nucleotide at every posi-

Tabela 9.1: **Description of the data set available in TISs-ST using a non-redundant set of genes.**

Group type I (Phylum[a] or Division[b])	**Group type II (Class)**	**Species**	**Number of sequences**
Arthropoda[a]	Insecta	*Anopheles gambiae*	599
		Bombyx mori	274
		Drosophila melanogaster	8821
Ascomycota[b]	Sordariomycetes	*Magnaporthe grisea*	152
Bryophyta[b]	Bryopsida	*Physcomitrella patens*	154
Chordata[a]	Aves	*Gallus gallus*	3135
	Actinopterygii	*Danio rerio*	7074
		Oncorhynchus mykiss	367
		Oryzias latipes	171
	Amphibia	*Xenopus laevis*	7440
		Xenopus tropicalis	3336
	Mammalia	*Bos taurus*	2541
		Canis familiaris	322
		Homo sapiens	12387
		Mus musculus	11872
		Ovis aries	180
		Rattus norvegicus	8594
		Sus scrofa	603
Echinodermata[a]	Echinoidea	*Strongylocentrotus purpuratus*	134
Magnoliophyta[b]	Magnoliopsida	*Arabidopsis thaliana*	14525
		Brassica napus	182
		Glycine max	373
		Lycopersicon esculentum	511
		Malus x domestica	133
		Solanum tuberosum	281
	Liliopsida	*Hordeum vulgare*	334
		Oryza sativa	1090
		Triticum aestivum	315
		Zea mays	518
Nemata[a]	Secernentea	*Caenorhabditis elegans*	3051
Platyhelminthes[a]	Trematoda	*Schistosoma japonicum*	1093
		Schistosoma mansoni	135

The number of sequences represents the total of unique genes. The sequences were taken from a subset of the NCBI UniGene data set (Wheeler *et al.*, 2003) (retrieved August, 2005). The taxonomic classification was based on the Integrated Taxonomic Information System on-line database. For the group classification, the phylum or division, and the class were used for the taxonomic level.

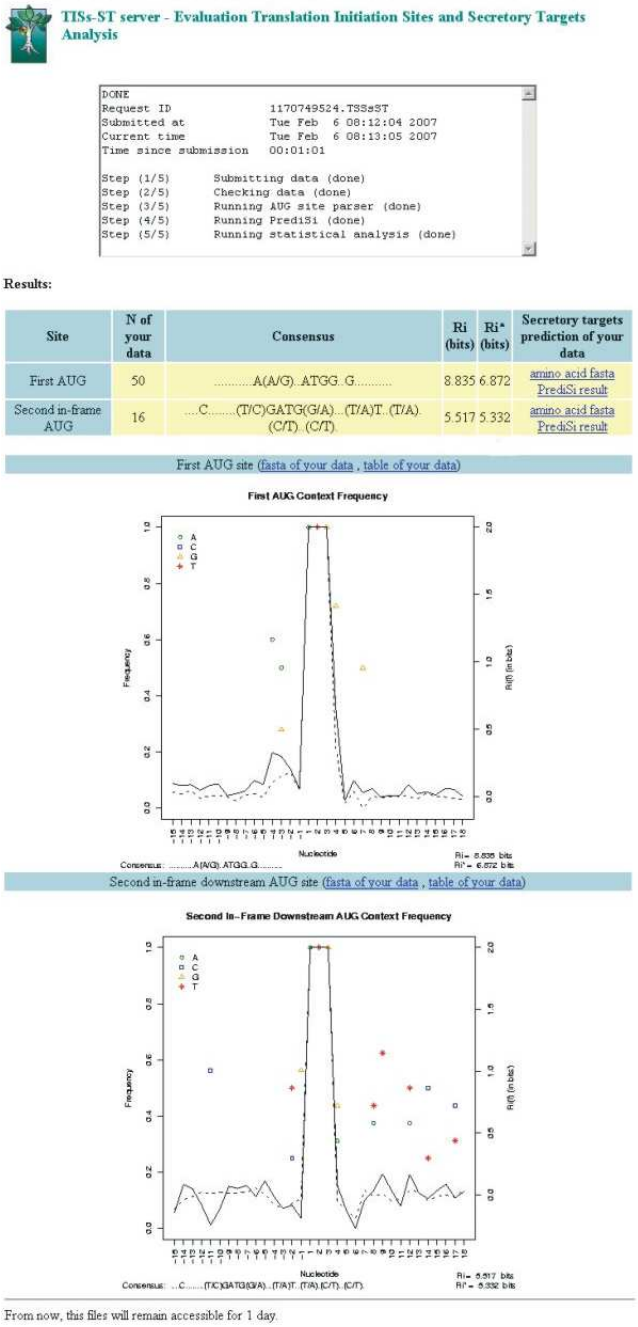


Figura 9.1: **The output interface of the TISs-ST web server.** In the TISs-ST visual output result page, every site that showed a consensus sequence had a graphic representation for the consensus and information content found. This example identified consensus for proteins encoded by first and second in-frame downstream AUG. The total information contents were 8.8 and 7.0 bits for these sites, respectively.

tion; and the amount of information present at every position in the sequence (Schneider and Stephens, 1990).

TISs-ST uses two different ways to display the graphical representation of consensus sequences: (i) the information content required to describe the pattern, which is the total information content (measured in bits) at every position in a site, and (ii) a useful graphical representation for displaying the global patterns in a set of aligned sequences, to focus on the periodic frequencies of the nucleotides in the consensus pattern.

Resources and CDS database

Non-redundant data sets of nucleotide sequences were compiled from the NCBI UniGene (retrieved August, 2005). Following the removal of sequences not annotated as 'complete CDS' (not identified previously as putatively and complete coding sequence), only sequences that had termination codon and had 15 bp before the first ATG annotated, remained in the data set. For this set, sequences were grouped according to species, based on their taxonomic classification in the Integrated Taxonomic Information System on-line database. Species were also grouped according to their taxonomic level (phylum or division, and class).

Classifying sequences into AUG site types and consensus determination

Data sets of fragments flanking the AUG were created from the -15 to +15 nucleotides of each AUG in every sequence. All fragments were grouped and the consensus sequences were determined separately for each AUG (first AUG, second in-frame downstream AUG, second out-frame downstream AUG, all other in-frame downstream AUGs, all other out-frame downstream AUGs) for each species and group of species using the 50/75 consensus rule described by Cavener (Cavener, 1987). The reading frame determination is based on the first AUG from the complete CDS annotated.

The consensus at a position is computed according to the following rules, with decreasing order of priority: (i) if a nucleotide at that position has a relative frequency greater than 50% and greater than twice the relative frequency of the second most frequent nucleotide, the nucleotide is given consensus status that is indicated in uppercase; (ii) if the sum of the relative frequencies of a pair of nucleotides exceeds 75%, these two nucleotides are given co-consensus status, indicated in uppercase; (iii) if there is a single most frequent nucleotide,

it is given dominant status, indicated in lowercase; (iv) if two bases have the same highest frequency, they are given co-dominant status that is indicated in lowercase.

Analysis of the information content at each position around each AUG site

The method begins by calculating a weight matrix from the frequencies of each nucleotide at each position of the aligned sequences. This matrix is then applied to the sequences to determine the sequence conservation of each individual site. Additionally we considered the nucleotide bias in genomes by using a linear noise correction (Schreiber and Brown, 2002).

A PWM model of first AUG site, called $R_{iw}(b,l)$, is created by using an aligned training set consisting of sequences from the nucleotides databases described before. The PWM is computed using the widely accepted information theoretical approach with some modifications (Schneider, 1997, Reents *et al.*, 2006). In TISs-ST, nucleotide biases can be corrected by using the nucleotide composition observed 15 bases upstream of the start codon annotated, and the triplet noise can be corrected by using the observed frequency of each nucleotide at each position of every codon (Schreiber and Brown, 2002). If a particular base does not appear in the data set used to create the frequency matrix, then we apply a penalty function that depends on the sample size n as follows: $R_{iw}(b,l) = \frac{1}{n+2}$ (Schneider, 1997). Since this weight matrix is created from many sequences, it can give statistically significant evaluations of individual sites, including those used to create the matrix itself (Schneider, 1997).

In a set of submitted sequences, we represent the jth sequence by a matrix $S_j(b,l)$ that contains only 0's and 1's. The individual information content of a base ($R_i(l)$), given by some manipulation of the $R_i(j)$ (Schneider, 1997), is the product between the base and the weight matrix:

$$R_i(l) = \frac{\sum_{j=1}^{n}\sum_{b=A}^{T} S_j(b,l)\, R_{iw}(b,l)}{n} \qquad \text{(bits per base)} \tag{9.1}$$

And the total information content of the sites is the R_i:

$$R_i = \sum_{l} R_i(l) \qquad \text{(bits per site)} \tag{9.2}$$

Prediction of secretory targets

The prediction of signal peptides was evaluated by the PrediSi prediction program (Hiller *et al.*, 2004), which allows an accurate and fast prediction of signal peptides.

Testing the TISs-ST

Preparation of training and testing data. The maize data set (n = 518) and *Arabidopsis* data set (n = 14525) were used. Each of these data sets was used for training and testing the program. The total sequences were divided into two equal sets, training and testing. The training set was cross-validated by testing with the testing data set.

Confusion matrix method. Using the confusion matrix, various measurements of quality such as accuracy, specificity, sensitivity and the Matthews correlation coefficient (MCC) (Matthews, 1975) were determined. In the following equation (9.3-9.6), TP refers to true positives (correctly predicted TIS), TN to true negatives (correctly predicted non-TIS), FN to false negatives (incorrectly predicted TIS) and FP to false positives (incorrectly predicted non-TIS).

accuracy (AC): proportion of correct predictions of the total predictions.

$$AC = \frac{TP + TN}{TP + TN + FN + FP} \tag{9.3}$$

specificity (SP): proportion of true negatives to the total negatives.

$$SP = \frac{TN}{TN + FP} \tag{9.4}$$

sensitivity (SN): proportion of true positives to the total positives.

$$SN = \frac{TP}{FN + TP} \tag{9.5}$$

MCC: This is regarded as a more rigorous measurement to evaluate the performance of class prediction methods. MCC equals 1 for perfect predictions, whilst it is zero for completely random predictions (Matthews, 1975).

$$MCC = \frac{(TP * TN) - (FP * FN)}{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)} \tag{9.6}$$

Accuracy test. Prediction can be performed at different levels of specificity and sensitivity by defining various thresholds for the final score. At each threshold, the numbers for the TPs, TNs, FPs and FNs are calculated, and based on these values parameters such as accuracy, specificity and sensitivity are determined using equations (9.3-9.5). As seen in Table 9.2, the prediction accuracy ranges from 0.64 to 0.95 at the different thresholds.

Tabela 9.2: **Confusion matrix values and dependent parameters at each threshold value.**

Score	**Positives tested (259)**		**Negatives tested (5692)**		**Accuracy**	**SN**	**SP**	**MCC**
threshold	**TP**	**FN**	**TN**	**FP**				
≥70	79	180	5570	122	0.949	0.305	0.978	0.320
≥60	105	154	5456	236	0.934	0.405	0.958	0.319
≥50	131	128	5309	383	0.914	0.505	0.932	0.318
≥40	157	102	5159	533	0.893	0.606	0.906	0.326
≥30	183	76	4801	891	0.837	0.706	0.843	0.291
≥20	209	50	4414	1278	0.776	0.806	0.775	0.274
≥10	235	24	3600	2092	0.644	0.907	0.632	0.225

Values in parentheses indicate the number of proteins in each set.

Specificity and sensitivity test. Specificity and sensitivity are two competing quality measurements for any two-classifier method. Table 9.2 shows the specificity and sensitivity of the TISs-ST at the different threshold levels, as well as the MCC values. At the highest specificity (0.98), the sensitivity is at 0.30, and at the highest sensitivity level (0.91), the specificity is reasonable, at 0.63. At a mid-range sensitivity level of 0.61 (with a threshold score of ≥ 40), 61% of all the positives can be predicted with only 9% FPs, and the best correlation is obtained (MCC = 0.33).

Results and discussion

In the literature, purines are usually claimed to be important at position -3. This was the case for maize and tobacco suspension cells (Lukaszewicz *et al.*, 2000). Although the -3 position is most conserved upstream of the start codon, experimental evidence in eudicots, showed that changes at the -2 and -1 positions affected translation efficiency at least as much as changes at the -3 position. For monocots, this effect seems to be even more pronounced

because changing the C at the -1 and/or -2 position resulted in an approximately 50% reduction in translation efficiency (Lukaszewicz *et al.*, 2000). In Figure 9.2, our analysis shows this topic for the chicken data set, where purines are most conserved upstream of the start codon (Figure 9.2A). The same did not occur for the second in-frame downstream AUG (Figure 9.2B).

We also performed analyses on the rice and tomato data sets, and the results of a TISs-ST search of the context surrounding the first translated AUG are shown in Figure 9.3. Different consensus patterns for the first AUG were found in the two data sets, and corroborate the known consensus obtained from the monocot and eudicot species (Joshi *et al.*, 1997, Lukaszewicz *et al.*, 2000).

The current method was applied to a real data set of different mRNA variants produced by differential splicing in the human phosphodiesterase 9A gene. The PDE9A gene encodes a cGMP-specific high-affinity phosphodiesterase, and the physiological implication of the isoforms produced by this gene is not yet known (Rentero *et al.*, 2003, Puigdomènech, 2006). At least 21 different human PDE9A mRNA transcripts have so far been identified, which are produced as a result of alternative splicing of the 5' exons. The different PDE9A splice variants could present different translation start codons to produce the functional protein, allowing for the synthesis of a variety of polypeptides that differ in their N-terminal regions and also show differential subcellular localization (Rentero *et al.*, 2003, Puigdomènech, 2006).

We performed analyses on the data set composed of splice variants that used the first start codon in exon 1 (7 isoforms) and the possible start codon present in exon 8 (5 isoforms). The results of the analyses of these sites are shown in Figure 9.4. The isoforms possibly encoded by the start codon in exon 8 (Figure 9.4B) showed an information content (Ri = 8.3 bits) as high as that encoded by the first start codon in exon 1. This prediction corroborates the hypothesis that this AUG codon may be an alternative TIS in the splice variants of the human phosphodiesterase 9A gene (Rentero *et al.*, 2003).

The observation that sub-optimal AUG contexts are present in many genes suggests the hypothesis that this context might be involved in modulation of gene expression. This might be the case for transcripts encoding two proteins that differ at their N-terminal end (Lukaszewicz *et al.*, 2000), which would reflect in multi-targeting of the protein. Another instance where modulation of the expression by the AUG context might be important, concerns transcripts that contain a small open reading frame upstream from the main open reading

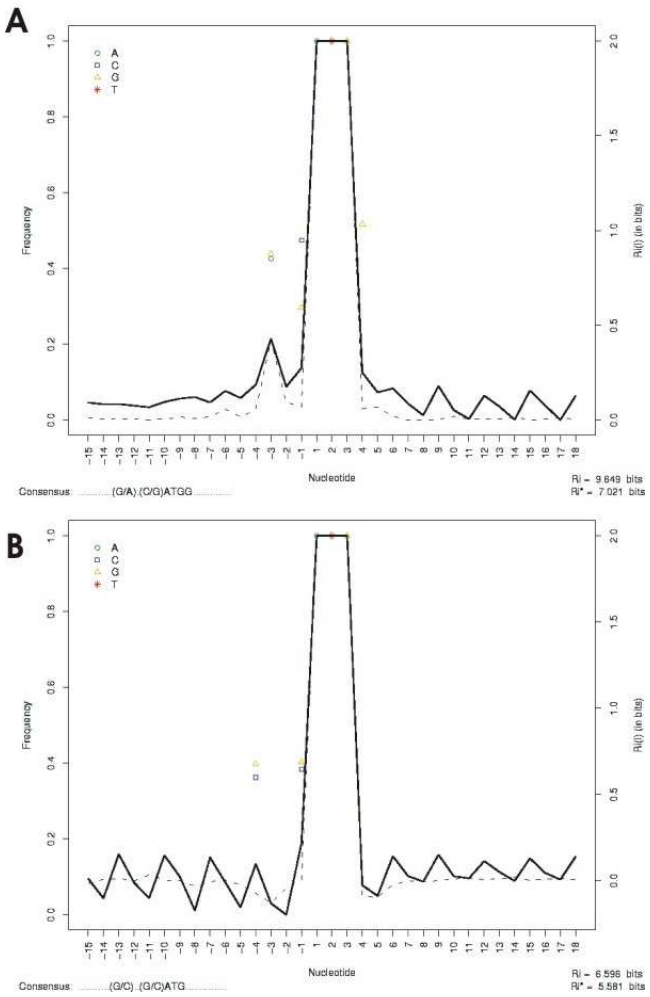


Figura 9.2: **Information content and differences in the context between two sites of** *Gallus gallus.* The nucleotide sequences were selected for the determination of nucleotide frequency at positions between -15 and +15 (codon AUG corresponds to position +1 to +3). The application of two different ways of displaying the consensus sequences allows one to display the nucleotide periodic frequencies in addition to the site information content. In the analysis performed a consensus context was deduced as a ".....(G/A).(C/G)ATGG....."and ".....(G/C)..(G/C)ATG....."for first and second in-frame downstream AUG, respectively. The total information content consisted of 9.6 and 6.6 bits for these sites. (A) First AUG site (n = 3135). (B) Second in-frame downstream AUG site (n = 1190). The frequency of each consensus base is indicated on the left Y axis, according to the 50/75 consensus rule. On the right Y axis, the lines represent the degree of site conservation measured in bits of information according to the equation given in the methods section. The continuous line is the information content without correction, and the broken line is the same information corrected for bias.

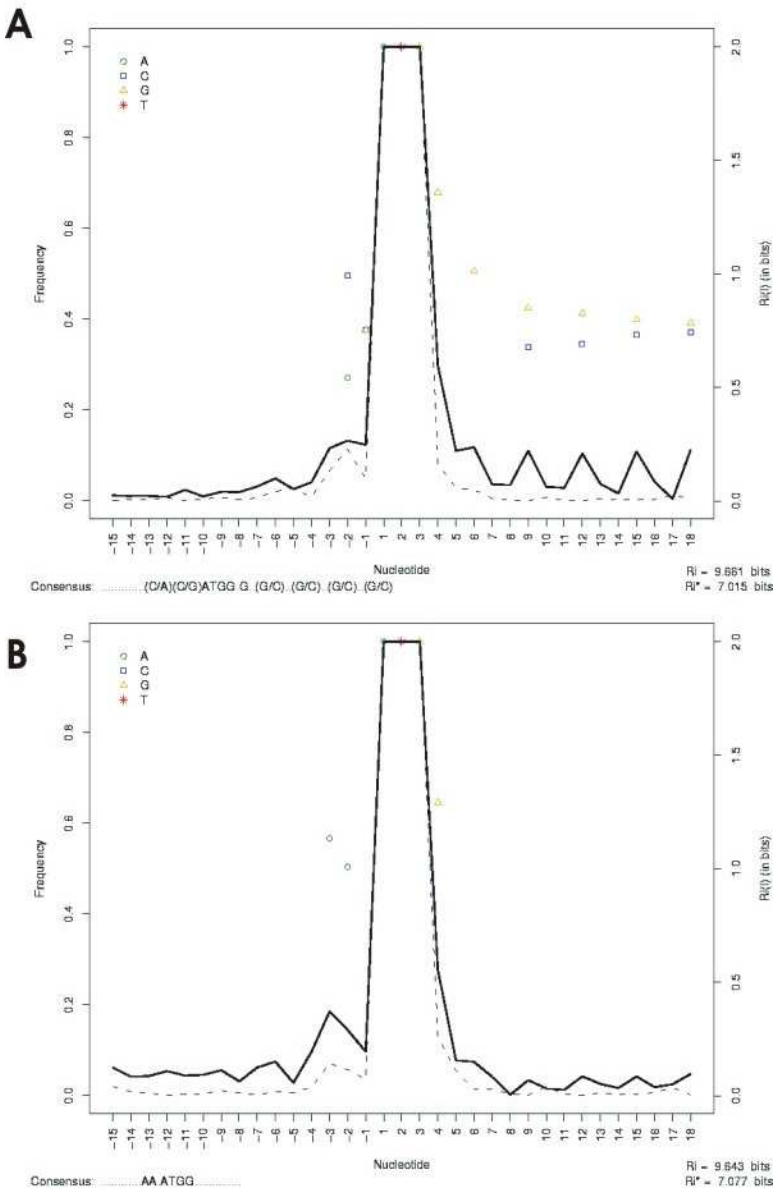


Figura 9.3: **An example of an evaluation frequency of nucleotides surrounding the initiation codon.** In the analysis performed a consensus context was deduced as a ".....(C/A)(C/G)ATGG.G..(G/C)..(G/C)..(G/C)..(G/C)"and ".....AA.ATGG....."for rice and tomato, respectively. (A) Sequence data set from *Oryza sativa* (n = 1090). (B) Sequence data set from *Lycopersicon esculentum* (n = 511).

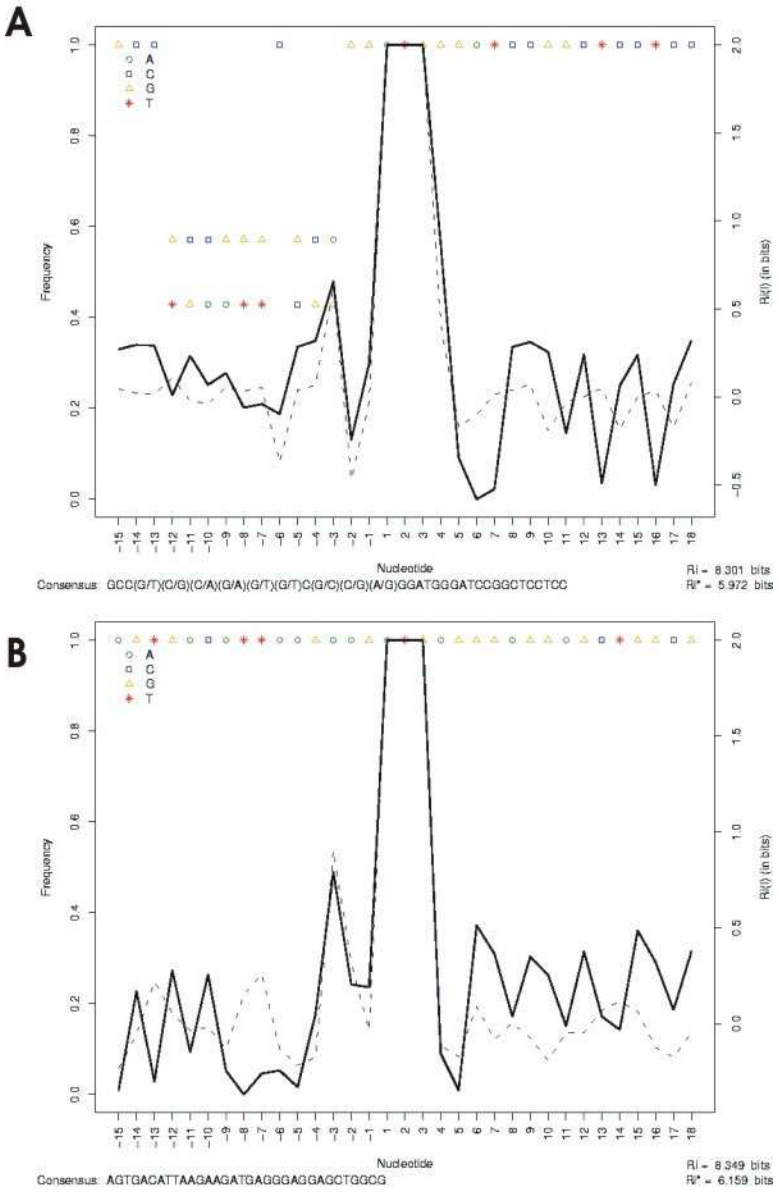


Figura 9.4: **High information content in alternative translation initiation sites of human PDE9A splice forms.** (A) The data set from the 7 isoforms that use the first start codon in exon 1. (B) The data set from the 5 isoforms that use the possible start codon present in exon 8.

frame.

The main limitation of TISs-ST is that target prediction is limited to signal peptides, mainly because there are only a few free stand-alone tools available for protein subcellular localization prediction (Petsalakis *et al.*, 2005). These free available tools do not permit one to create derivative works based on their software, or their applications are time consuming for a web access. Future work includes making the ability to predict the subcellular localization of proteins from the N-terminal amino acid sequence available. The analysis described above is based on the NCBI UniGene data set and the taxonomy classification of Integrated Taxonomic Information System on-line database. The sequences in the NCBI UniGene data set are mostly represented by gene-specific EST clusters, and many of them are often annotated as complete CDS. This makes the determination of reading frames and the search for the 15 bp upstream of CDS, an easy step. Additional new taxonomic groups will be included in future versions. The TISs-ST local data set will be updated twice a year to incorporate future UniGene updates.

Conclusions

Several molecular mechanisms might provide for efficient translation of the mRNAs containing upstream AUGs including leaky scanning and reinitiation or internal initiation of the translation. The contributions of these mechanisms remain uncertain but several recent studies suggest that the impact of at least some of them might be substantial (Kozak, 2001, Vadim, 2001, Pestova *et al.*, 2001, Kozak, 2002, Wang and Rothnagel, 2004, Kozak, 2005).

Many databases in this research area are available on the web. Examples include Transterm, an information resource devoted to contexts of translation initiation and termination sites (Jacobs *et al.*, 2006), and UTRdb, a curate database of 5' and 3' untranslated sequences of eukaryotic mRNAs (Mignone *et al.*, 2005). With these databases, it is possible to obtain the structure and detect the presence of known regulatory elements in UTR sequences, and the annotation of experimentally defined functional motifs from sequence contexts of annotated translation initiation and termination codons. But currently no computational tools are available for the accurate prediction of alternative TIS, and investigations in this field could contribute to a better understanding of the complexity of mechanisms used by the cell to expand the diversity of proteins encoded by the genome. In this work we have

presented a new online web server to evaluate translational variability reflected by alternative TISs. Comprehensive comparisons of contexts that surround the alternative TISs are very topical in eukaryotic mRNAs, and in addition, such translational polymorphism is a source of variability in cytoplasmic and organellar proteomes (Kochetov and Sarai, 2004). TISs-ST provides a collection of pre-analysed data sets extracted from the NCBI UniGene database, and focuses on the sequences flanking the various AUG along the complete CDSs.

Availability and requirements

- Project name: TISs-ST
- Project home page: http://ipe.cbmeg.unicamp.br/pub/TISs-ST
- Operating systems(s): Platform independent
- Programming language: Perl and R
- Other requirements: to build a local version of the web-service it is necessary to have a web server that allows CGI and Perl.
- License: under the GNU General Public License

Authors' contributions

RV conceived the study, conducted the work and drafted the manuscript. MM participated in the design and coordination of the study and helped draft the manuscript. All the authors read and approved the manuscript.

Acknowledgements

RV was supported by a fellowship from the UNIEMP Institute and MM received a research fellowship from the "Conselho Nacional de Desenvolvimento Científico e Tecnológico". This work was partially supported by grants from the "Fundação de Amparo à Pesquisa do Estado de São Paulo"(02/01167-1, 03/07244-0 and 05/58104-0).

As Ciências são uma espécie de grande edifício em que muitas pessoas trabalham em comum acordo. Uns extraem a pedra da pedreira com o suor do seu corpo, outros a arrastam com esforço até a base da construção, outros a erguem com a força de braços e maquinas, mas aquele que a coloca no lugar e a faz servir tem o mérito da construção.

Jean le Rond d'Alembert (1717-1783)

10

Genomic patterns of translation initiation sites in sugarcane

Renato Vicentini[1] and Marcelo Menossi[1]

Abstract

The nucleotide sequence flanking the translation initiation codon and the features of the 5' untranslated regions affects the translational efficiency of eukaryotic mRNAs, and may indicate the presence of an alternative translation initiation site (TIS) to produce proteins with different properties. To contribute to our understanding of the genome complexity of sugarcane, we undertook an *in silico* genome wide TIS analysis. We describe specific periodic pattern of nucleotide base pairing in mRNA coding regions. We also performed a comparative analysis between TIS in sugarcane genes with different expression profiles, and classified in different evolutionary groups. We show that up-regulated transcripts shows a more strong TIS when compared with down-regulated, and that ubiquitous transcripts have a high frequency of alternative TIS in the next downstream AUG codon. The same occurs for fast-evolving genes, and leaf and internodes specific transcripts, that may be encodes different polypeptide by N-terminal polymorphism. Our data show that different classes of sugarcane genes can often be utilized as additional start sites to increase translation rate of mRNAs and many proteins can be characterized by N-terminal heterogeneity.

[1]Departamento de Genética e Evolução, Laboratório de Genoma Funcional, Instituto de Biologia, CP 6109, Universidade Estadual de Campinas - UNICAMP, 13083-970, Campinas, SP, Brazil

Introduction

The tropical sugarcane (*Saccharum* spp.) is an important industrial crop and is cultivated on close to 20 million hectares in more than 90 countries (FAO; http://apps.fao.org). Sugarcane belongs to the grass family (Poaceae), an economically important seed plant family that includes cereals such as maize, wheat, rice, and sorghum as well as many forage crops. The *Saccharum* complex includes the agronomically and industrially important sugarcane genotypes obtained from *S. officinarum*, *S. spontaneum* and *S. robustum* crosses (Grivet and Arruda, 2002). The main product of sugarcane is sucrose, which accumulates in the stalk internodes, contributing to about two thirds of the world's raw sugar production.

The translation start site plays an important role in the control of translation efficiency of eukaryotic mRNAs. Translation by cytosolic ribosomes generally occurs at the first AUG in the transcript. However, in eukaryotic mRNAs, efficient recognition of an AUG codon as a translation initiation site (TIS) depends on several factors, such as the nucleotide sequence that flanks the site (Kozak, 1991, Kozak, 2002, Kozak, 2005). There is evidence that the context surrounding the initiation codon contributes to the control of translational initiation (Kawaguchi and Bailey-Serres, 2005). The sequence context of the first AUG codon, in particular that part located in the untranslated region, may modulate the efficiency with which it is recognized as a translation initiation codon (Mignone *et al.*, 2002). If the first initiation codon lies in a suitable context, protein synthesis will be started. When the context is less than favorable, most of the protein synthesis will start at the next downstream AUG codon (Nadershahi *et al.*, 2004). Moreover, other structural features of the mRNA are considered important for the efficiency of the translation initiation at a specific AUG codon, such as: the proximity of AUG to the 5' end, the secondary structure upstream and downstream from the AUG codon, the leader sequence length and the multiple upstream AUG codons (Kozak, 1991, Wang and Rothnagel, 2004, Kozak, 2005).

Statistical analyses of the AUG initiation codon context in many organisms identified a preferential nucleotide frequency in some positions around the AUG. Recent analyses have revealed variations in the initiation context between different groups of eukaryotes. Distinct inter-taxon variations in the AUG context sequences are repeatedly observed when invertebrates, higher plants and protozoa are considered separately (Lukaszewicz *et al.*, 2000). For plant genes, a consensus context was deduced as c(A/G)(C/A)CAUGGC for monocots and A(A/C)aAUGGC for eudicots (Joshi *et al.*, 1997, Lukaszewicz *et al.*, 2000).

Upstream out-frame AUG may severely affect the translation of a gene, even if surrounded by a poor context (Luehrsen and Walbot, 1994), suggesting that upstream AUGs may have a role in keeping the basal translation level of a gene (Mignone *et al.*, 2002). Recently it was demonstrated that downstream AUG codons are utilized as alternative TISs even in mRNAs with multiple strong upstream AUGs (Wang and Rothnagel, 2004). mRNAs with a suboptimal context of start AUG codon are relatively abundant. It is likely that at least some mRNAs with suboptimal start codon context contain the other signals providing additional information for efficient AUG recognition (Kochetov and Sarai, 2004). Their occurrence must correlate with the start codon context: sub-optimal context should be accompanied by a higher frequency of downstream AUGs (Kochetov and Sarai, 2004). With this mechanism, called 'leaky scanning', multiple different proteins can be obtained from the same mRNA (Mignone *et al.*, 2002). In this sense AUGs located downstream of the major coding sequences (CDS), may play a role in generating protein diversity (Kozak, 2002). The usage of a closely located downstream in-frame AUG codon as an alternative TIS can result in full and N-truncated proteins that may have the same function and be targeted at the different compartments (Watanabe *et al.*, 2001, Kochetov and Sarai, 2004, Kochetov, 2005). Since eukaryotic mRNAs frequently contain TISs in a sub-optimal context (Rogozin *et al.*, 2001), the problems of polypeptide N-end heterogeneity and finding of the genuine TIS are very topical.

In 1973 (Ball, 1973) was proposed the hypotheses that selection pressure for specific RNA secondary structure could affect the choice of nucleotide at both synonymous and non-synonymous positions. Since then, several lines of evidence have supported the idea that the redundancy of the genetic code allows preservation of mRNA folding. Periodical patterns complementary to the proof-reading site in the ribosome and presumably involved in the translation frame monitoring mechanism have been found in many transcripts (Shabalina *et al.*, 2006).

Proteins must be localized correctly at the sub-cellular level to have normal biological functions (Xie *et al.*, 2005). When the final destination is the mitochondria, the chloroplast, or the secretory pathway, sorting usually relies on the presence of an N-terminal targeting sequence (von Heijne *et al.*, 1989). There are known cases of variation in the use of alternative signal peptides, and in the majority of cases this is due to the exclusion of the signal peptide from one or more protein products of the same gene. If the sequence between two AUGs encodes an organelle targeting sequence, then the protein initiated from the ups-

tream sequence will contain this targeting information whereas the protein initiated from the downstream one will not.

To contribute to our understanding of the genome complexity of sugarcane, we undertook a genome wide TIS analysis. In this paper, we report the results of this *in silico* analysis in the sugarcane ESTs data. We describe specific periodic pattern of nucleotide base pairing in mRNA coding regions. We also performed a comparative analysis between TIS in sugarcane genes with different expression profiles, and classified in different evolutionary groups.

Results and Discussion

General structure of sugarcane coding region

To study the sequence conservation related with mRNA stability, associated with secondary structure, we evaluated sequence nucleotide pattern in sugarcane data. A total of 11,436 complete CDS (with first and stop codon) and 4,699 sequences with second downstream in-frame AUG were analyzed (Table 10.1). Profiles of nucleotide base conservation frequency in the 5'-UTR, CDS and 3'-UTR are shown on Figure 10.1. A well-defined periodic pattern of conservation is observed in the coding region. This pattern emerges at nucleotide +1 (Figure 10.1A) and terminates at the stop codon (Figure 10.1C). The nucleotide in the first codon site is the most conserved, while those in the second positions are the least conserved. Nucleotide G preferentially participated in the secondary structures at codon site 1, nucleotides C and A at codon site 2, G at codon site 3. When we considers the nearest downstream AUG codon (Figure 10.1B) this pattern are: nucleotide G at codon site 1, nucleotides A, C and U at codon site 2, GC at codon site 3. This pattern of nucleotide base pairing was related with the mRNA self-folding. Recent study demonstrate that the obvious reasons for preferential realization of pairing at third codon sites are the elevated GC content and the near equivalent frequencies of base pairing nucleotides at third codon site (Shabalina *et al.*, 2006).

The 5' UTR-CDS and CDS-3' UTR boundaries are characterized with conserved secondary structures (Shabalina *et al.*, 2006). Figure 10.1 show that sequence conservation profiles around the start codon shows a degree of symmetry around this site. Shabalina *et al.* (2006) describe that this region of mRNA are subject to a stronger purifying selection than

those in the rest of CDS and the 5' UTR. This pattern of sequence conservation may be important for efficient initiation and termination of translations, and a strong mRNA secondary structure formed due to gene-specific codon usage may have been implicated in discontinuous translation and pauses in synthesis.

Tabela 10.1: Data sets of sugarcane genes used in this study.

Groups	**Genes**	**Genes in second AUG data set**	**Reference**
complete CDSs	11436	4699	this study
Expression groups			
Up-regulated	49	25	
Down-regulated	57	25	
Ubiquitous expression	48	22	(Papini-Terzi *et al.*, 2005)
Internode specific	32	15	
Leaf specific	44	21	
Evolutionary groups			
Monocot fast evolving	31	12	(Vincentz *et al.*, 2004)
Sugarcane specific	1333	424	

Translation initiation site of sugarcane genes classified by expression level

The initiation efficiency often depends upon both a consensus sequence context and the secondary structure surrounding the codon (Kozak, 1990). Despite some variations in the consensus sequences, the presence of a purine in positions -3 and +4 seems to be determinant for an efficient initiation in both animals and plants. The observation that sub-optimal AUG contexts are present in many genes suggests the hypothesis that this context might be involved in modulating gene expression (Lukaszewicz *et al.*, 2000).

In recent studies numerous genes have been identified which contain two or more in frame AUGs at the 5' end, any of which might correspond to the initiation codon used to initiate translation of catalytically active protein (Small *et al.*, 1998). In the present study we identify that 41% of putative sugarcane proteins had an in frame AUG codon at the 5' end. We performed a comparison of consensus sequence and information content of sugarcane genes that are classified according to their evolutionary or expression classes. Table 10.2 presents

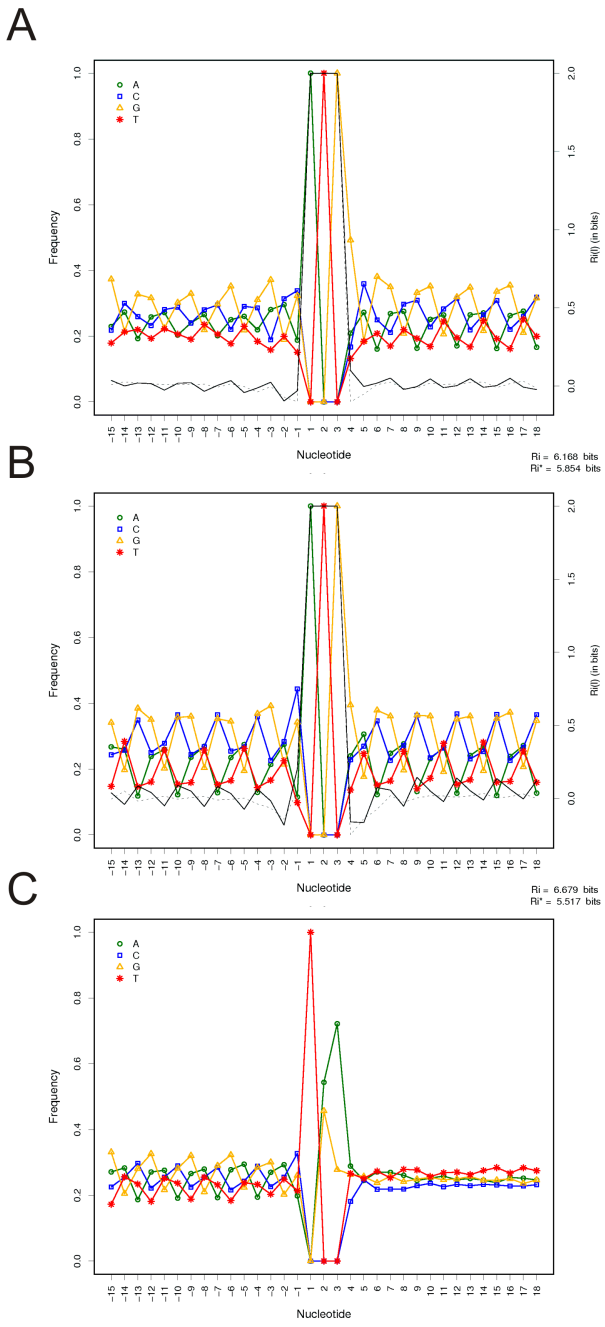


Figura 10.1: **Profiles of nucleotide base pairing around the start codon (A), the downstream AUG codon (B) and stop codon (C) for 11436 sugarcane mRNAs.** The total information content consisted of 6.1 and 6.6 bits for start codon and downstream AUG codon, respectively. The frequency of each base is indicated on the left Y axis. On the right Y axis (A and B), the lines represent the degree of site conservation measured in bits of information. The continuous line is the information content without correction, and the broken line is the same information corrected for bias. Orange, guanine; Blue, cytosine; Green, adenosine; Red, uridine/thymine.

these results, where is shows that second in-frame AUG codon had a higher degree of purines conservation in the important nucleotides positions (-4, -3, and +6). How expected, the first AUG codon shows preference for a guanine in +4 positions (Table 10.2).

Translation of mRNAs can vary in efficiency, so that the amount of protein produced is modulated. This is an important level of gene regulation; indeed, a correlation between mRNA and protein abundance is seen only for secreted proteins, whereas for intracellular proteins the differing rates of translation of different mRNAs removes this correlation. To investigate the correlation between gene expression and mRNA translation, we analyze the total information content showed by proteins encoded by genes that shows organ specific expression. In Figure 10.2 are showed the total information content and consensus base determination for start codon of up-regulated (Figure 10.2A) and down-regulated (Figure 10.2B) genes in at least one sugarcane tissue. Despite of a more evident consensus in up-regulated data set, the both sets shows highly information content (7.2 and 6.8 bits, respectively) when compared to complete CDSs, our control set. These results suggest that genes with organ specific expression can be more effectively translated.

Surprising, when we analyzed the TISs of sugarcane genes that presented similar expression levels in all tissues (Figure 10.3), we indentified a highly information content (9.3 bits) present in the first in-frame downstream AUG codon. In this case the putative first AUG codon, shows a similar measure of the control data set (6.1 bits). This strong frequency of putative alternative TISs in genes with similar expression in the whole plant represent that these genes may be encodes different polypeptide by N-terminal polymorphism.

Highly level of alternative TIS in the main sinks tissues of sugarcane

In sugarcane, growing young leaves and internodes are the main sink tissues. Light and sugars regulate growth activities by a coordinated modulation of gene expression and enzyme activities in both, carbohydrate-exporting (source) and carbohydrate-importing (sink) tissues. Gene regulation is based on sensing different signals or stimuli, which then is transmitted through a signaling pathway that in the end leads to an increase or decrease of transcription. In a similar way that the ubiquitous genes, those genes that shows a specific expression in leaf or internodes presents highly total information content for the first in-frame downstream AUG codon (Figure 10.4). Our data show that this both specialized classes of sugarcane genes can often be utilized as additional start sites to increase translation rate of

Tabela 10.2: Comparison of consensus sequence and information content of sugarcane genes, with strongest or poor context, that are classified according to their evolutionary or expression classes.

Groups	-3	+4	Ri in first AUG data set (bits)	Consensu in first AUG data set	-4	-3	+6	Ri in second AUG data set (bits)	Consensu in second AUG data set
complete CDSs			6,168	.....ATG.....				6,679	.....(C/G)ATG.....
Up-regulated		G	7,29	.....ATGG....(C/G)..(G/C)..(G/C)...			(G/C)	6,948	..(G/C)(C/A)....(C/G).....(C/G)ATG..(G/C).....(C/G)..(C/G)..C
Down-regulated		G	6,87	G.....ATGG.....			(C/G)	6,579	..(G/C)..(C/G)..C.....CATG..(C/G)..(C/G)..(C/G)..(C/G)..(C/G)
Ubiquitous expression		G	6,163	.....ATGG.....	(C/G)	G	(G/C)	9,305	..(C/G)..(C/G)..G..(C/G)G.(C/G)ATG..(G/C)..(G/C)..C(G/A).C(G/A).(G/C)
Internode specific		G	6,722	G.....ATGG.....(A/G).....			G	7,377	..(G/C)..C..(C/G).....CATG.(G/C)G..(C/A)..C..(C/G)..(C/G)
Leaf specific			6,331	.....ATG.....			(G/C)	6,944	..(G/C)C.(G/C)..C.....CATG..(G/C).(A/C)...(C/G)..(C/G)..(C/G)
Monocot fast evolving	G		6,233	.....(A/C)....G..ATG.....	(G/C)	(G/A)		8,391	.....(G/C).....(G/C)(G/A).(G/C)ATG...G.(C/G).....(G/C)..G
Sugarcane specific			3,229	.....ATG.....				2,628	.....ATG.....

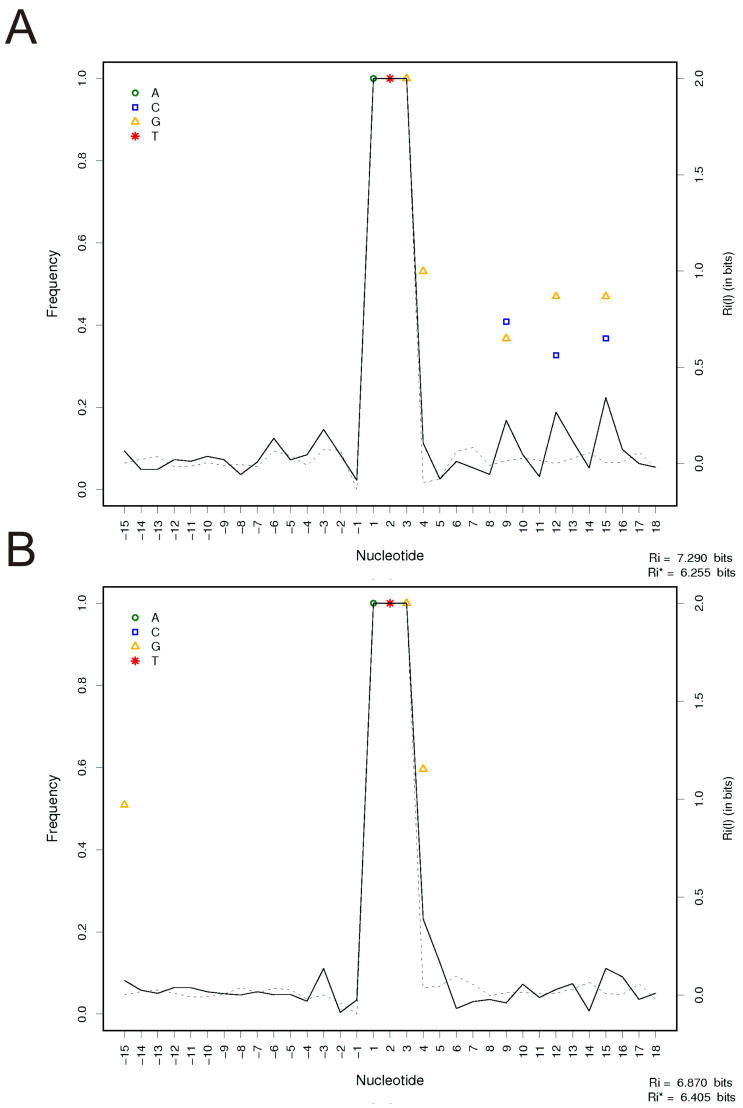


Figura 10.2: **Total information content and consensus base determination for start codon of up-regulated (A) and down-regulated (B) genes at least one sugarcane tissue.** The frequency of each consensus base is indicated on the left Y axis, according to the 50/75 consensus rule. The total information content consisted of 7.2 and 6.8 bits for start codon of up-regulated and down-regulated genes, respectively.

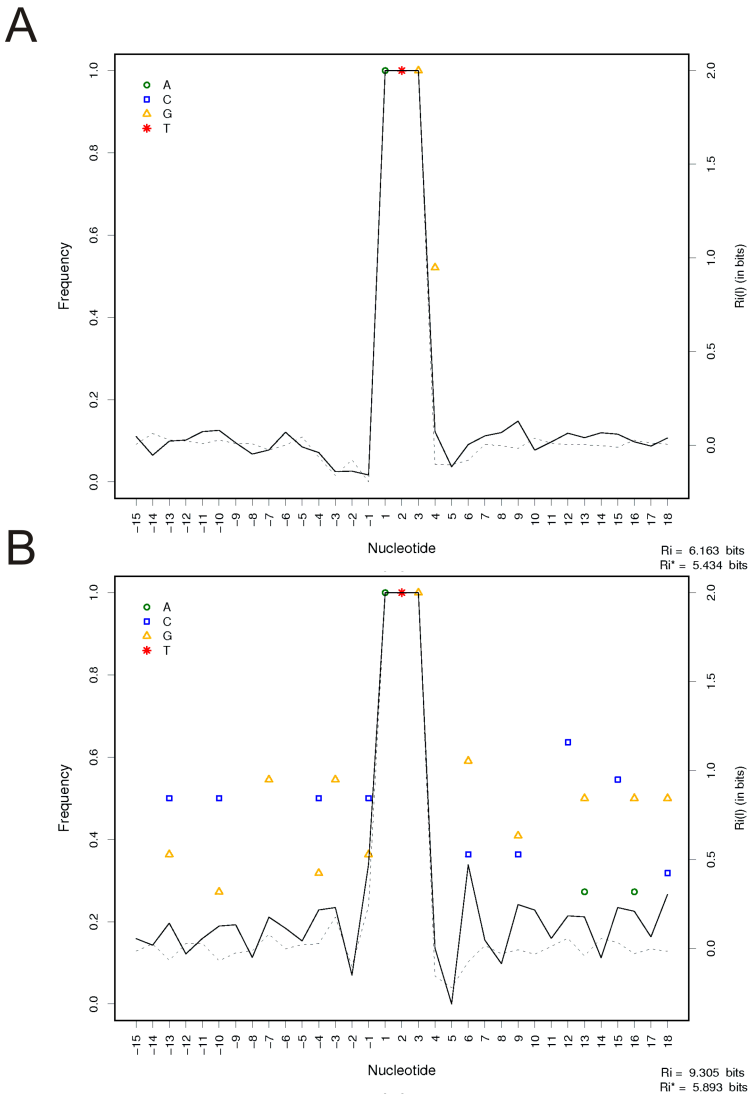


Figura 10.3: **Total information content and consensus base determination for start codon (A) and first in-frame downstream AUG codon (B) of sugarcane genes that presented highly similar expression levels in all tissues.** The total information content consisted of 6.1 and 9.3 bits for start codon and downstream AUG codon, respectively.

mRNAs and many proteins can be characterized by N-terminal heterogeneity.

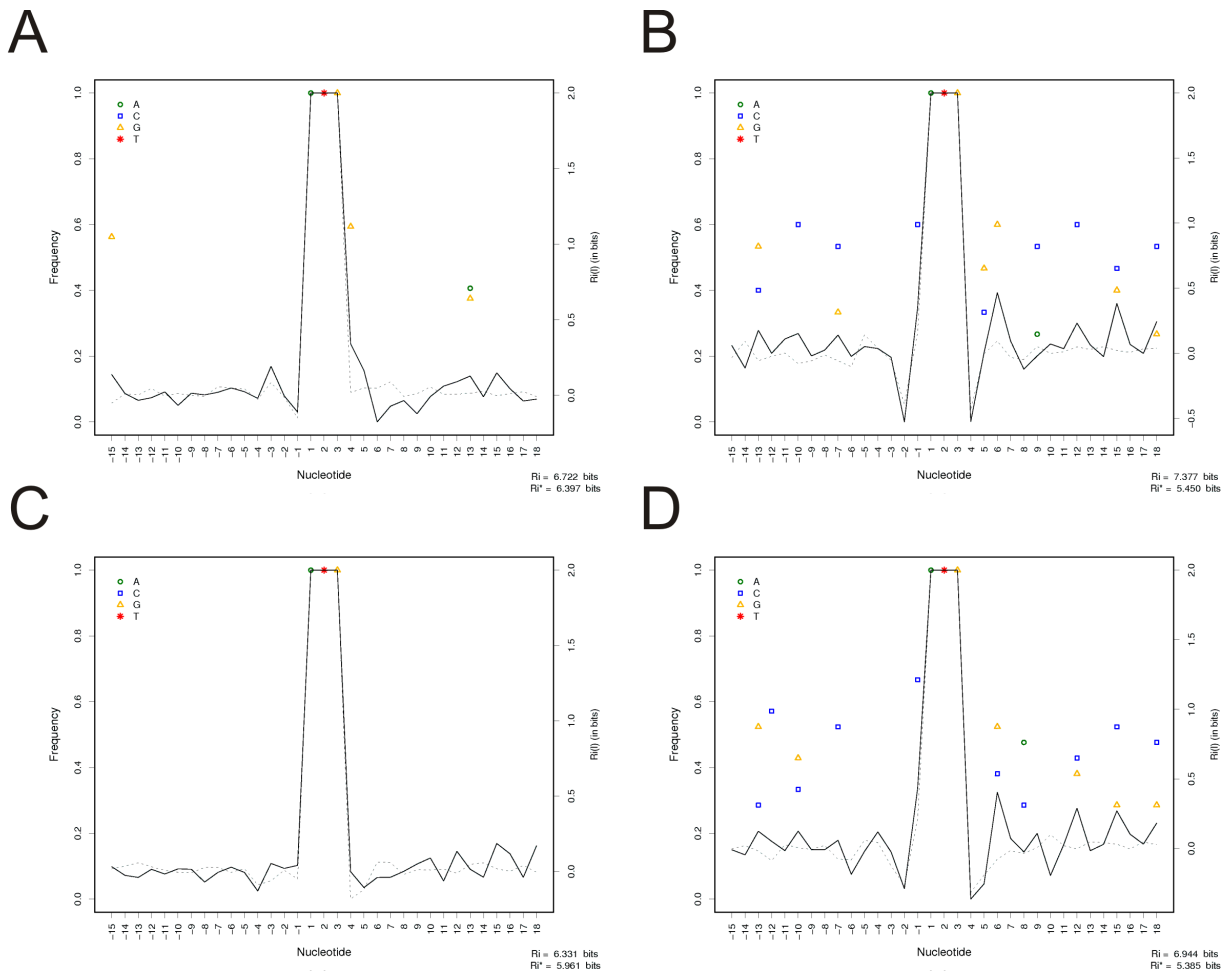


Figura 10.4: **Total information content and consensus base determination for start codon (A and C) and first in-frame downstream AUG codon (B and D) of internodes (A and B) and leaf (C and D) specific sugarcane genes.**

Information content of sugarcane specific genes and fast evolving genes

The two main groups of flowering plants, the monocots and dicots, diverged 200 million years ago, but despite this long period of independent evolution, plant genes display significant conservation. The accelerated evolution of specific sequences of conserved eudicot-monocot protein families is an important aspect of angiosperm evolution (Vincentz *et al.*, 2004) and the comparative genomics provides a starting point for understanding the

genetic basis of the biological diversity among plant species. We analyze two data sets of sugarcane proteins that are classified in evolutionary classes: (1) sugarcane specific proteins; and (2) monocot fast-evolving proteins. Figure 10.5 shows the analysis of the total information content and consensus base determination for start codon and first in-frame downstream AUG codon of these both sets of genes.

This analysis shows that the sugarcane specific proteins set presents lower information content for the both AUG sites analyzed (3.2 and 2.6 bits). This group correspond to highly variable sequences that either diverged significantly (evolving at high rates) from their homologs in other monocots or that are specific for sugarcane. Since the method used in determination of the information content was created based on a PWM for Liliopsida taxonomic class, it not can be able to give a statistically significant evaluation for proteins in this evolutionary class. The same not occurs for the monocot fast-evolving proteins, where the PWM is adapted for the analysis. In this case, the results (Figure 10.5C and 10.5D) are similar to the obtained for the ubiquitous genes, and leaf and internodes specific transcripts, where the first in-frame downstream AUG codon shows a strong information content and can be utilized how a TIS.

Characterization of 5' UTR in sugarcane

In recent study (Mignone *et al.*, 2002) was showed that the comparison of the various completed and partial genome sequences reveals some conserved aspects of the structure of UTRs sequences. The average length of 5' UTRs is roughly constant over diverse taxonomic classes and ranges between 100 and 200 nucleotides. Structural features of the 5' UTR have a major role in the control of mRNA translation. Messenger RNAs encoding proteins involved in developmental processes, which need to be strongly and finely regulated, often have 5' UTRs that are longer than average (Kozak, 1987).

It is noteworthy that a large fraction of 5' UTRs contain upstream AUGs, suggesting that the 'first AUG rule' predicted by the scanning model of ribosome start-site selection is disobeyed in a large number of cases. This implies that the 40S ribosomal subunit can sometimes bypass the most upstream AUG codon, possibly because its sequence context makes it a poor initiation codon, to initiate translation at a more distal AUG (Xiong *et al.*, 2001). Moreover, it has been calculated that the presence of an upstream AUG correlates with a long 5' UTR and with a weak start codon context of the AUG that is usually used, whereas

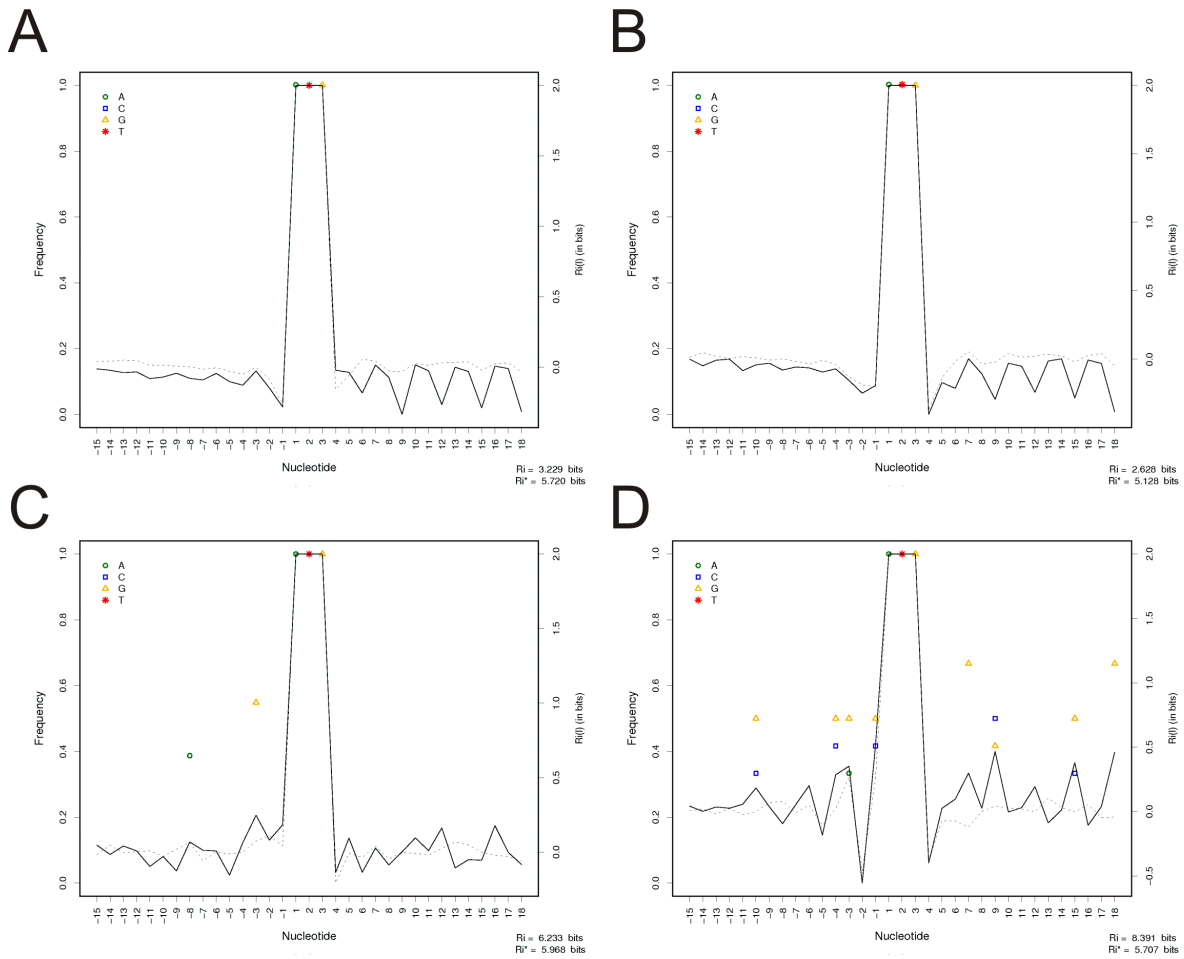


Figura 10.5: **Total information content and consensus base determination for start codon (A and C) and first in-frame downstream AUG codon (B and D) of sugarcane specific (A and B) and fast-evolving (C and D) genes.**

transcripts with an optimal start-codon context have short 5' UTRs without upstream AUGs (Rogozin *et al.*, 2001), suggesting that upstream AUGs may have a role in keeping the basal translational level of a gene.

We analyze the frequency of AUG codon in 5' UTR sequences, and the length average of these UTR sequences in the eight data sets of this study (Table 10.3). These analyses were able to show that exist a negative correlation between the information content of the putative start codon and the number of upstream AUGs in the 5' UTR, the same occurs for the correlation between the information content and the 5' UTR length. The two sets of sequences that have high frequency of AUG codon in UTR region, when compared with the control group (complete CDSs), were the 'Ubiquitous' and 'Sugarcane specific'. Those are the only two groups with low information content in comparison with control group. The opposite occurs for the sequence sets with high information content. This is the case for the genes that shows sequences with short 5' UTR and small number of AUGs codon ('Down-regulated', 'Internode specific', and 'Monocot fast evolving'). The other groups of genes that show high information content ('Up-regulated' and 'Leaf specific) are composed by sequences with longest 5' UTR region. This supports the hypothesis that the genes with organ specific expression, main in sugarcane leaf, are involved in biological processes that need to be finely regulated.

Tabela 10.3: Features of sugarcane 5' UTR of genes that are classified according to their evolutionary or expression classes.

Groups	**Average frequency of AUG codon in 5' UTR**	**Average 5' UTR length (bp)**
complete CDSs	5,55	288,99
Up-regulated	5,27	318,82
Down-regulated	4,82	269,95
Ubiquitous expression	8,75	437,73
Internode specific	4,81	277,34
Leaf specific	5,56	315,02
Monocot fast evolving	4,45	205,21
Sugarcane specific	5,93	251,70

Subcellular targeting of putative N-truncated sugarcane proteins

We compared predicted subcellular localizations of sugarcane proteins and their potential N-terminally truncated forms started from the nearest downstream in frame AUG codon. *In silico* determination of the subcellular localization of the proteins can provide information on their function, and is dependent on the correct identification of the first AUG and their potential N-terminally polymorphic forms. This translational polymorphism may serve as an important source of diversity in both cytoplasmic and organelle proteomes (Kochetov and Sarai, 2004, Kochetov, 2005). The results of prediction are shown in Table 10.4 for each data set analyzed.

In this study two distinct groups were identified: (1) full and N-truncated proteins may be target to the same locations, the two evolutionary groups presented this characteristic; and (2) full protein may have no working targeting signal, whereas its N-truncated form may have an active targeting signal. This last case was detected for all expression groups of genes, where an increase of N-truncated form of proteins with target to an organelle was identified. Interestingly, this analysis shows increase of plastid proteins in all data sets when a N-truncated form of protein acquired sorting signals *de novo*. These results demonstrate that this phenomenon in sugarcane deserves attention, and further investigation may be shows a high source of translational polymorphism in sugarcane, which a change of organelle proteome content.

Tabela 10.4: Subcellular localization of full and putative N-truncated sugarcane proteins (%) predicted with *TargetP* program.

Groups	**Full**				**N-truncated**			
	mTP	**cTP**	**SP**	**Others**	**mTP**	**cTP**	**SP**	**Others**
complete CDSs	6,1	3,0	4,6	29,1	6,9	3,9	6,4	25,6
Up-regulated	1,6	0,8	0,0	18,3	4,0	2,4	3,2	11,1
Down-regulated	2,3	0,0	0,0	16,7	2,3	3,0	1,5	12,1
Ubiquitous expression	2,0	0,0	2,6	11,1	3,9	1,3	2,6	7,8
Internode specific	1,4	1,4	0,0	19,2	6,8	5,5	0,0	9,6
Leaf specific	1,8	0,0	0,0	17,5	2,6	3,5	1,8	11,4
Monocot fast evolving	1,8	0,6	0,0	4,9	1,2	0,6	0,6	4,9
Sugarcane specific	1,0	0,4	0,6	5,3	1,2	0,4	0,9	4,8

Conclusion

During the 200 million years since separation of the monocots and eudicots, approximately two-thirds of their genes have been kept very similar, whereas the remaining one-third consists of fast-evolving sequences or specific genes that could account for the differences between these two types of plants (Vincentz *et al.*, 2004). A detailed analysis of these sequences in sugarcane indicated that a significant proportion corresponded to fast-evolving sequences found in members of conserved angiosperm gene families (Vincentz *et al.*, 2004). A high rate of evolution can be related to the production of new protein functions that may be involved in the differentiation of a specific evolutionary lineage. The characterization of sugarcane genes that may participate in tissue-specific or ubiquitous activities can increase the availability of tools for sugarcane biotechnology and our knowledge about molecular biology of this plant (Papini-Terzi *et al.*, 2005). The identification of genes highly expressed in stems or leaves could also help in the understanding of metabolic pathways involved in sugar production and accumulation, and could constitute targets for crop improvement.

Comparisons of contexts that surround the alternative TISs are very topical in eukaryotic mRNAs. A comprehensive genome analysis using the consensus sequences assembled from sugarcane ESTs has revealed sets of putative proteins that show a strong TIS and/or strong alternative TIS in the next downstream AUG codon. These results suggest that genes with organ specific expression can be more effectively translated, and that genes with similar expression in the whole plant present a strong frequency of alternative TIS. The same occurs for fast-evolving genes, and leaf and internodes specific transcripts, that may be encodes different polypeptide by N-terminal polymorphism represented by strong alternative TIS. Our data show that different classes of sugarcane genes can often be utilized as additional start sites to increase translation rate of mRNAs and many proteins can be characterized by N-terminal heterogeneity.

In conclusion, modulation of protein translation in sugarcane has different roles for different classes of gene, remarkably in organ specific or ubiquitous genes, and sugarcane specific or fast-evolving genes. The aim of this study was give a view of the regulation of protein translation in sugarcane, increasing the knowledge about this very topical phenomenon in plants. This is a preliminary study of this question, and further experimental investigations are necessary to attribute others relevant biological information to these data.

Material and Methods

Sequence datasets

The sugarcane (*Saccharum* spp.) EST sequences were from the SUCEST project, which has been described previously (Vettore *et al.*, 2003). These sequences represents 43,141 SASs (Sugarcane Assembled Sequences), which were estimated to represent over 30,000 unique genes. The first step was to determine the putative ORFs (open reading frame) for each SAS. To predict the ORFs we used three prediction programs: 1. GeneMark.SPL (Borodovsky and McIninch, 1993), a variant of GeneMark.hmm used to analyze ESTs and cDNAs data with improved prediction of gene boundaries; 2. GENSCAN (Burge and Karlin, 1997), which captures the general and specific compositional properties of the distinct functional units of eukaryotic genes and 3. ESTScan (Iseli *et al.*, 1999), particularly useful to solve problems caused by frameshift or stop codon errors. The next step was to identify the set of ORFs that represents complete CDSs (coding sequences). Those that start with ATG and end with stop codon were selected for the sugarcane complete CDS data set (a total of 11,882 amino acid sequences). Only sequences with complete CDSs possessing the 5'- and 3'-UTRs (15 nt or longer) were analyzed in this study.

With objective of characterize putative and alternative TIS in groups of genes classified by their expression profiles, we generate datasets based on previously expression study using a cDNA microarray (Papini-Terzi *et al.*, 2005). These microarrays were constructed to profile individual variation of sugarcane plants cultivated in the field and transcript abundance in six plant organs (flowers, roots, leaves, lateral buds, and 1st and 4th internodes), and contains 1280 distinct elements. This study showed that 217 sugarcane genes presented differential expression in two biological samples of at least one of the tissues tested, and a total of 153 genes (12%) presented highly similar expression levels in all tissues. In the present study these genes were separated in Up-regulated, Down-regulated and Ubiquitous expression classes (Table 10.1, expression groups). We also created groups of genes that showed differential expression in each organ analyzed (flowers, roots, leaves, lateral buds, and internodes).

Approximately two-thirds of the sugarcane transcriptome have similar sequences in *Arabidopsis*. The remaining sequences represent putative monocot-specific genetic material, one-half of which were found only in sugarcane (Vincentz *et al.*, 2004). To further

evaluate the participation of different evolutionary processes in the production of sugarcane- or monocot-specific sequences, an evaluation of information content present in putative and alternative TIS was performed in two evolutionary categories of sugarcane genes (Vincentz *et al.*, 2004): sugarcane-specific sequences, and fast-evolving sequences. (Table 10.1, evolutionary groups). AUG frequencies were calculated in the 5' UTR regions of annotated coding sequences.

Identification of the consensus sequence

Data sets of fragments flanking the AUG were created from the -15 to +15 nucleotides of each AUG in every sequence of the eight groups of sugarcane genes. All fragments were grouped and the consensus sequences were determined separately for each AUG (first AUG or second in-frame downstream AUG) using the 50/75 consensus rule described by Cavener (Cavener, 1987). The consensus at a position is computed according to the following rules, with decreasing order of priority: (i) if a nucleotide at that position has a relative frequency greater than 50% and greater than twice the relative frequency of the second most frequent nucleotide, the nucleotide is given consensus status that is indicated in uppercase; and (ii) if the sum of the relative frequencies of a pair of nucleotides exceeds 75%, these two nucleotides are given co-consensus status, indicated in uppercase.

Identification of the information content

We recently developed a method, called TISs-ST (Vicentini and Menossi, 2007), for investigate the usage of alternative TISs for synthesis of new protein variants possibly possessing different functions. TISs-ST provides a collection of pre-analyzed data sets extracted from the NCBI UniGene database, and focuses on the sequences making the various AUG along the complete CDSs. In the present study, we utilize this method and the position weight matrix (PWM) model created specifically for the Liliopsida (monocotyledons) taxonomic group (Vicentini and Menossi, 2007). All sequences in each data set were aligned and this PWM was then applied to the sequences to determine the sequence conservation of each individual site. Additionally we considered the nucleotide bias in coding region by using a linear noise correction (Schreiber and Brown, 2002). The individual information content of a site (R_i(l)), given by some manipulation of the R_i(j) (Schneider, 1997), is the product between the base and the weight matrix; and the total information content of the sequence

sites is the R_i.

Subcellular localization of full and N-truncated proteins

The prediction of subcellular localization of full and N-truncated proteins was evaluated by the TargetP prediction program (Emanuelsson *et al.*, 2000). Two samples were generated for each data sets of Table 10.1, either started from first AUG or from the nearest downstream in frame AUG codon.

Acknowledgements

This work was partially supported by grant 05/58104-0 from the FAPESP.

Não importa quanto aprendamos. O que resta, por menor que possa parecer, é tão infinitamente complexo quanto o todo era no princípio.

Isaac Asimov (1920-1992)

11 Conclusão

I Nossos resultados permitiram a elaboração de um cenário referente à distribuição e compartimentalização do proteoma de uma célula de cana-de-açúcar. Neste cenário 44% das proteínas estão localizadas no núcleo, 19% no citoplasma, 18% na via secretora, 12% na mitocôndria, e 6% nos plastídios. Dentre as proteínas da via secretora 60% foram identificadas como proteínas a serem secretadas da célula, 41% pertencentes ao reticulo endoplasmático e 6% como sendo proteínas de vacúolo.

II A cana-de-açúcar apresenta diversas proteínas possuidoras de múltiplas localizações subcelulares, em concordância com resultados recentes em outros organismos. De fato, nossos resultados indicam que cerca de 19% das proteínas de cana-de-açúcar são localizadas em mais de um compartimento celular. A ocorrência mais comum é a de proteínas com localização citoplasmática e nuclear, que perfazem 13% dos casos.

III O uso de isoformas protéicas com a extremidade amino truncada também tem sido relatado por estudos recentes. Em nossas análises com cana-de-açúcar identificamos que em 18% dos casos ocorre a perda do sinal de direcionamento nas proteínas truncadas, já em 15% das proteínas truncadas ocorre a manutenção desse sinal. Nossos resultados mostraram também que ocorre uma mudança nos sinais em diversos casos: 10% das proteínas truncadas adquirem um novo sinal de direcionamento, e cerca de 6% acabam alterando sua localização subcelular devido à mudança do sinal presente na isoforma truncada.

IV O uso de métodos distintos de predição da localização subcelular possibilita a otimização do potencial de predição, auxiliando inclusive nos casos de múltiplas localizações.

Com isto foi possível a elaboração de um método com exatidão superior aos atualmente existentes.

V O gene *ScBAK1* provavelmente está envolvido em cascatas de sinalização celular intermediadas por altos níveis de açúcar na folha de cana-de-açúcar.

VI A comparação detalhada dos contextos ao redor dos sítios de início da tradução é uma abordagem capaz de identificar sítios alternativos, podendo inclusive indicar polimorfismos de tradução como fonte de variabilidade nos proteomas citoplasmáticos e organelares.

VII Há uma clara tendência por fortes contextos ao redor de sítios de inicio de tradução em transcritos de cana-de-açúcar que apresentam expressão em um ou mais tecido da própria planta. Esses resultados indicam que pode não haver uma correlação linear entre quantidade de mRNA e quantidade de proteína para estes genes, mas sim que os genes mais expressos acabam por fornecerem material para uma síntese protéica ainda em maior escala.

VIII Os genes de cana-de-açúcar com expressão constitutiva apresentam uma tendência por possuírem fortes sítios de início da tradução alternativos, localizados no próximo códon AUG *downstream* ao sítio tradicional. Resultados como estes podem indicar uma alta demanda por quantidade de proteínas codificadas por genes que se expressam de forma similar na planta como um todo.

IX Resultados similares ao acima foram encontrados para os genes com expressão específica em folhas e entrenós da cana-de-açúcar, com o detalhe de que nestes dois casos além de um forte sítio de início da tradução alternativo foi também identificado que o sítio tradicional possui um conteúdo de informação mais evidente do que o apresentado por genes específicos de outros tecidos.

Referências Bibliográficas

Abe,H., Yamaguchi-Shinozaki,K., Urao,T., Iwasaki,T., Hosokawa,D. and Shinozaki,K. (1997) Role of Arabidopsis MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. *Plant Cell,* **9**, 1859–1868.

Abel,S. and Theologis,A. (1996) Early genes and auxin action. *Plant Physiology,* **111**, 9–17.

Adams,K., Daley,D., Qiu,Y., Whelan,J. and Palmer,J. (2000) Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. *Nature,* **408** (6810), 354–357.

Altschul,S., Madden,T., Schaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* **25** (17), 3389–3402.

Alvarez-Buylla,E.R., Liljegren,S.J., Pelaz,S., Gold,S.E., Burgeff,C., Ditta,G.S., Vergara-Silva,F. and Yanofsky,M.F. (2000) MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes. *The Plant Journal,* **24** (4), 457–466.

Anandatheerthavarada,H., Biswas,G., Mullick,J., Sepuri,N., Otvos,L., Pain,D. and Avadhani,N. (1999) Dual targeting of cytochrome P4502B1 to endoplasmic reticulum and mitochondria involves a novel signal activation by cyclic AMP-dependent phosphorylation at Ser128. *EMBO Journal,* **18** (20), 5494–5504.

Andrade,M., O'Donoghue,S. and Rost,B. (1998) Adaptation of protein surfaces to subcellular location. *Journal of Molecular Biology,* **276** (2), 517–525.

Aneeta Sanan-Mishra,N., Tuteja,N. and Kumar Sopory,S. (2002) Salinity- and ABA-induced up regulation and light-mediated modulation of mRNA encoding glycine-rich RNA-binding protein from Sorghum bicolor. *Biochemical and Biophysical Research Communications,* **296**, 1063–1068.

Annen,F. and Stockhaus,J. (1999) SbRLK1, a receptor-like protein kinase of Sorghum bicolor (L.) Moench that is expressed in mesophyll cells. *Planta,* **208** (3), 420–425.

Anterola,A. and Lewis,N. (2002) Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. *Phytochemistry,* **61** (3), 221–294.

Aravind,L. and Koonin,E. (2000) The U box is a modified RING finger - a common domain in ubiquitination. *Current Biology,* **10** (4), R132–R134.

Arruda,P. (2001) Sugarcane transcriptome. A landmark in plant genomics in the tropics. *Genetics and Molecular Biology,* **24**, 1.

Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,J., Davis,A., Dolinski,K., Dwight,S., Eppig,J., Harris,M., Hill,D., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J., Richardson,J., Ringwald,M., Rubin,G. and Sherlock,G. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics,* **25** (1), 25–29.

Aukerman,M.J. and Sakai,H. (2003) Regulation of Flowering Time and Floral Organ Identity by a MicroRNA and Its APETALA2-Like Target Genes. *Plant Cell,* **15** (11), 2730–2741.

Azevedo,C., Santos-Rosa,M. and Shirasu,K. (2001) The U-box protein family in plants. *Trends in Plant Science,* **6** (8), 354–358.

Bachmann,M., Huber,J., Athwal,G., Wu,K., Ferl,R. and Huber,S. (1996) 14-3-3 proteins associate with the regulatory phosphorylation site of spinach leaf nitrate reductase in an isoform-specific manner and reduce dephosphorylation of Ser-543 by endogenous protein phosphatases. *FEBS Letters,* **398** (1), 26–30.

Ball,L. (1973) Secondary structure and coding potential of the coat protein gene of bacteriophage MS2. *Nature - New Biology,* **242** (115), 44–45.

Bannai,H., Tamada,Y., Maruyama,O., Nakai,K. and Miyano,S. (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics,* **18** (2), 298–305.

Barata,R., Chaparro,A., Chabregas,S., Gonzalez,R., Labate,C., Azevedo,R., Sarath,G., Lea,P. and Silva,M. (2000) Targeting of the soybean leghemoglobin to tobacco chloroplasts: effects on aerobic metabolism in transgenic plants. *Plant Science,* **155** (2), 193–202.

BarPeled,M., Bassham,D. and Raikhel,N. (1996) Transport of proteins in eukaryotic cells: More questions ahead. *Plant Molecular Biology,* **32** (1-2), 223–249.

Barsalobres,C. (2004). *Analise da expressao genica induzida por Diatraea saccharalis em cana-de-acucar via macroarranjos de colonias bacterianas*. PhD thesis, Universidade de São Paulo.

Bartel,B. and Fink,G. (1994) Differential Regulation of an Auxin-Producing Nitrilase Gene Family in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences,* **91** (14), 6649–6653.

Baudino,S., Hansen,S., Brettschneider,R., Hecht,V.F., Dresselhaus,T., Lorz,H., Dumas,C. and Rogowsky,P.M. (2001) Molecular characterisation of two novel maize LRR receptor-like kinases, which belong to the SERK gene family. *Planta,* **213** (1), 1–10.

Bauer,J., Hiltbrunner,A. and Kessler,F. (2001) Molecular biology of chloroplast biogenesis: gene expression, protein import and intraorganellar sorting. *Cellular and Molecular Life Sciences,* **58** (3), 420–433.

Becraft,P.W. (2002) Receptor kinase signaling in plant development. *Annual Review of Cell and Developmental Biology,* **18** (1), 163–192.

Beisson,F., Koo,A.J., Ruuska,S., Schwender,J., Pollard,M., Thelen,J.J., Paddock,T., Salas,J.J., Savage,L., Milcamps,A., Mhaske,V.B., Cho,Y. and Ohlrogge,J.B. (2003) Arabidopsis Genes Involved in Acyl Lipid Metabolism. A 2003 Census of the Candidates, a Study of the Distribution of Expressed Sequence Tags in Organs, and a Web-Based Database. *Plant Physiology,* **132** (2), 681–697.

Belkhadir,Y. and Chory,J. (2006) Brassinosteroid signaling: a paradigm for steroid hormone signaling from the cell surface. *Science,* **314** (5804), 1410–1411.

Berger,S. (2002) Jasmonate-related mutants of Arabidopsis as tools for studying stress signaling. *Planta,* **214**, 497–504.

Berridge,M.J. (1993) Inositol trisphosphate and calcium signalling. *Nature,* **361** (6410), 315–325.

Blasing,O.E., Gibon,Y., Gunther,M., Hohne,M., Morcuende,R., Osuna,D., Thimm,O., Usadel,B., Scheible,W.R. and Stitt,M. (2005) Sugars and Circadian Regulation Make Major Contributions to the Global Regulation of Diurnal Gene Expression in Arabidopsis. *Plant Cell,* **17** (12), 3257–3281.

Borodovsky,M. and McIninch,J. (1993) GENMARK: Parallel gene recognition for both DNA strands. *Computers & Chemistry,* **17** (2), 123–133.

Boutry,M., Nagy,F., Poulsen,C., Aoyagi,K. and Chua,N. (1987) Targeting Of Bacterial Chloramphenicol Acetyltransferase To Mitochondria In Transgenic Plants. *Nature,* **328** (6128), 340–342.

Bower,N., Casu,R., Maclean,D., Reverter,A., Chapman,S. and Manners,J. (2005) Transcriptional response of sugarcane roots to methyl jasmonate. *Plant Science,* **168** (3), 761–772.

Braga,D., Arrigoni,E., Silva-Filho,M. and Ulian,E. (2003) Expression of the Cry1Ab protein in genetically modified sugarcane for the control of Diatraea saccharalis (Lepidoptera: Crambidae). *Journal of New Seeds,* **5**, 209–222.

Bray,E., Bailey-Serres,J. and Weretilnyk,E. (2000) *Biochemistry and Molecular Biology of Plants*. Rockyville, Maryland: American Society of Plant Physiologists pp. 1158–1203.

Brinks,S., Flugge,U., Chaumont,F., Boutry,M., Emmermann,M., Schmitz,U., Becker,K. and Pfanner,N. (1994) Preproteins of Chloroplast Envelope Inner Membrane Contain Targeting Information For Receptor-Dependent Import Into Fungal Mitochondria. *Journal of Biological Chemistry,* **269** (23), 16478–16485.

Broner,I. and Law,R. (1991) Evaluation of a modified atmometer for estimating reference ET. *Irrigation Science,* **12**, 21–26.

Bruce,B.D. (2000) Chloroplast transit peptides: structure, function and evolution. *Trends in Cell Biology,* **10** (10), 440–447.

Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology,* **268** (1), 78–94.

Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research,* **32**, D262–D266.

Canales,C., Bhatt,A., Scott,R. and Dickinson,H. (2002) EXS, a putative LRR receptor kinase, regulates male germline cell number and tapetal identity and promotes seed development in Arabidopsis. *Current Biology,* **12** (20), 1718–1727.

Capoen,W., Goormachtig,S., De Rycke,R., Schroeyers,K. and Holsters,M. (2005) SrSymRK, a plant receptor essential for symbiosome formation. *Proceedings of the National Academy of Sciences,* **102**, 10369–10374.

Carson,D. and Botha,F. (2002) Genes expressed in sugarcane maturing internodal tissue. *Plant Cell Reports,* **20** (11), 1075–1081.

Carson,D., Huckett,B. and Botha,F. (2002) Sugarcane ESTs differentially expressed in immature and maturing internodal tissue. *Plant Science,* **162**, 289–300.

Castillo-Davis,C.I. and Hartl,D.L. (2003) GeneMerge – post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics,* **19** (7), 891–892.

Casu,R., Dimmock,C., Chapman,S., Grof,C., McIntyre,C., Bonnett,G. and Manners,J. (2004) Identification of differentially expressed transcripts from maturing stem of sugarcane by in silico analysis of stem expressed sequence tags and gene expression profiling. *Plant Molecular Biology,* **54** (4), 503–517.

Casu,R., Manners,J., Bonnett,G., Jackson,P., McIntyre,C., Dunne,R., Chapman,S., Rae,A. and Grof,C. (2005) Genomics approaches for the identification of genes determining important traits in sugarcane. *Field crops research,* **92**, 137–147.

Casu,R.E., Grof,C.P., Rae,A.L., McIntyre,C.L., Dimmock,C.M. and Manners,J.M. (2003) Identification of a novel sugar transporter homologue strongly expressed in maturing stem vascular tissues of sugarcane by expressed sequence tag and microarray analysis. *Plant Molecular Biology,* **52** (2), 371–386.

Cavener,D.R. (1987) Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. *Nucleic Acids Research,* **15**, 1353–1361.

Chabregas,S.M. (2001). *Caracterização da localização subcelular da proteína THI1 de Arabidopsis thaliana*. PhD thesis, Universidade de São Paulo.

Chaumont,F., Oriordoan,V. and Boutry,M. (1990) Protein-Transport Into Mitochondria Is Conserved Between Plant And Yeast Species. *Journal of Biological Chemistry,* **265** (28), 16856–16862.

Chen,F. and Dixon,R.A. (2007) Lignin modification improves fermentable sugar yields for biofuel production. *Nature Biotechnology,* **25** (7), 759–761.

Chen,W., Provart,N.J., Glazebrook,J., Katagiri,F., Chang,H.S., Eulgem,T., Mauch,F., Luan,S., Zou,G., Whitham,S.A., Budworth,P.R., Tao,Y., Xie,Z., Chen,X., Lam,S., Kreps,J.A., Harper,J.F., Si-Ammour,A., Mauch-Mani,B., Heinlein,M., Kobayashi,K., Hohn,T., Dangl,J.L., Wang,X. and Zhu,T. (2002) Expression Profile Matrix of Arabidopsis Transcription Factor Genes Suggests Their Putative Functions in Response to Environmental Stresses. *Plant Cell,* **14** (3), 559–574.

Chen,X. (2004) A MicroRNA as a Translational Repressor of APETALA2 in Arabidopsis Flower Development. *Science,* **303** (5666), 2022–2025.

Chinnusamy,V., Schumaker,K. and Zhu,J. (2004) Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants. *Journal of Experimental Botany,* **55**, 225–236.

Chiu,W., Niwa,Y., Zeng,W., Hirano,T., Kobayashi,H. and Sheen,J. (1996) Engineered GFP as a vital reporter in plants. *Current Biology,* **6** (3), 325–330.

Cho,Y., Fernandes,J., Kim,S.H. and Walbot,V. (2002) Gene-expression profile comparisons distinguish seven organs of maize. *Genome Biology,* **3** (9), research0045–research0045.

Chow,K., Singh,D., Roper,J. and Smith,A. (1997) A single precursor protein for ferrochelatase-I from Arabidopsis is imported in vitro into both chloroplasts and mitochondria. *Journal of Biological Chemistry,* **272** (44), 27565–27571.

Christensen,A.C., Lyznik,A., Mohammed,S., Elowsky,C.G., Elo,A., Yule,R. and Mackenzie,S.A. (2005) Dual-Domain, Dual-Targeting Organellar Protein Presequences in Arabidopsis Can Use Non-AUG Start Codons. *Plant Cell,* **17** (10), 2805–2816.

Claros,M. and Vincens,P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *European Journal of Biochemistry,* **241** (3), 779–786.

Claverie,J. (1998) Computational methods for exon detection. *Molecular Biotechnology,* **10** (1), 27–48.

Cokol,M., Nair,R. and Rost,B. (2000) Finding nuclear localization signals. *EMBO Reports,* **1** (5), 411–415.

Comparot,S., Lingiah,G. and Martin,T. (2003) Function and specificity of 14-3-3 proteins in the regulation of carbohydrate and nitrogen metabolism. *Journal of Experimental Botany,* **54** (382), 595–604.

Corsi,A. and Schekman,R. (1996) Mechanism of polypeptide translocation into the endoplasmic reticulum. *Journal of Biological Chemistry,* **271** (48), 30299–30302.

Cotelle,V., Meek,S., Provan,F., Milne,F., Morrice,N. and MacKintosh,C. (2000) 14-3-3s regulate global cleavage of their diverse binding partners in sugar-starved Arabidopsis cells. *EMBO Journal,* **19** (12), 2869–2876.

Creissen,G., Reynolds,H., Xue,Y. and Mullineaux,P. (1995) Simultaneous Targeting of Pea Glutathione-Reductase and of a Bacterial Fusion Protein to Chloroplasts and Mitochondria in Transgenic Tobacco. *The Plant Journal,* **8** (2), 167–175.

Cui,Q., Jiang,T., Liu,B. and Ma,S. (2004) Esub8: A novel tool to predict protein subcellular localizations in eukaryotic organisms. *BMC Bioinformatics,* **5** (1), 66.

Dale,S., Arro,M., Becerra,B., Morrice,N.G., Boronat,A., Hardie,D.G. and Ferrer,A. (1995) Bacterial Expression of the Catalytic Domain of 3-Hydroxy-3-Methylglutaryl-CoA Reductase (Isoform HMGR1) from Arabidopsis thaliana, and Its Inactivation by Phosphorylation at Ser577 by Brassica oleracea 3-Hydroxy-3-Methylglutaryl-CoA Reductase Kinase. *European Journal of Biochemistry,* **233** (2), 506–513.

Danpure,C. (1995) How Can The Products of a Single-Gene be Localized to more than one Intracellular Compartment. *Trends in Cell Biology,* **5** (6), 230–238.

Davis,M.J., Hanson,K.A., Clark,F., Fink,J.L., Zhang,F., Kasukawa,T., Kai,C., Kawai,J., Carninci,P., Hayashizaki,Y. and Teasdale,R.D. (2006) Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units. *PLoS Genetics,* **2**, e46.

Delaney,T., Uknes,S., Vernooij,B., Friedrich,L., Weymann,K., Negrotto,D., Gaffney,T., Gut-Rella,M., Kessmann,H., Ward,E. and Ryals,J. (1994) A central role of salicylic acid in plant disease resistance. *Science,* **266**, 1247–1250.

DeLoose,M., Danthinne,X., van Bockstaele,E., van Montagu,M. and Depicker,A. (1995) Different 5'-leader sequences modulate beta-glucuronidase accumulation levels in transgenic nicotiana-tabacum plants. *Euphytica,* **85** (1-3), 209–216.

Devoto,A., Nieto-Rostro,M., Xie,D., Ellis,C., Harmston,R., Patrick,E., Davis,J., Sherratt,L., Coleman,M. and Turner,J. (2002) COI1 links jasmonate signalling and fertility to the SCF ubiquitin-ligase complex in Arabidopsis. *The Plant Journal,* **32**, 457–466.

Devoto,A. and Turner,J. (2003) Regulation of jasmonate-mediated plant responses in Arabidopsis. *Annals of Botany,* **92**, 329–337.

Di Sansebastiano,G.P., Paris,N., Marc-Martin,S. and Neuhaus,J.M. (1998) Specific accumulation of GFP in a non-acidic vacuolar compartment via a C-terminal propeptide-mediated sorting pathway. *The Plant Journal,* **15** (4), 449–457.

Dodd,A.N., Salathia,N., Hall,A., Kevei,E., Toth,R., Nagy,F., Hibberd,J.M., Millar,A.J. and Webb,A.A.R. (2005) Plant Circadian Clocks Increase Photosynthesis, Growth, Survival, and Competitive Advantage. *Science,* **309** (5734), 630–633.

Duby,G. and Boutry,M. (2002) Mitochondrial protein import machinery and targeting information. *Plant Science,* **162** (4), 477–490.

Eckert,M. and Kaldenhoff,R. (2000) Light-induced stomatal movement of selected Arabidopsis thaliana mutants. *Journal of Experimental Botany,* **51**, 1435–1442.

Ehness,R., Ecker,M., Godt,D.E. and Roitsch,T. (1997) Glucose and Stress Independently Regulate Source and Sink Metabolism and Defense Mechanisms via Signal Transduction Pathways Involving Protein Phosphorylation. *Plant Cell,* **9** (10), 1825–1841.

Ellis,J., Dodds,P. and Pryor,T. (2000) Structure, function and evolution of plant disease resistance genes. *Current Opinion in Plant Biology,* **3**, 278–284.

Emanuelsson,O., Elofsson,A., von Heijne,G. and Cristobal,S. (2003) In Silico Prediction of the Peroxisomal Proteome in Fungi, Plants and Animals. *Journal of Molecular Biology,* **330** (2), 443–456.

Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence,. *Journal of Molecular Biology,* **300** (4), 1005–1016.

Endre,G., Kereszt,A., Kevei,Z., Mihacea,S., Kalo,P. and Kiss,G. (2002) A receptor kinase gene regulating symbiotic nodule development. *Nature,* **417**, 962–966.

Falco,M. (1994). *Cultura de Células em Suspensão e protoplastos em cana-de-açúcar*. PhD thesis, Universidade de São Paulo.

Felsenstein,J. (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics,* **5**, 164–166.

Ferl,R.J. (2004) 14-3-3 proteins: regulation of signal-induced events. *Physiologia Plantarum,* **120** (2), 173–178.

Franco,O., Rigden,D., Melo,F. and Grossi-De-Sa,M. (2002) Plant alpha-amylase inhibitors and their interaction with insect alpha-amylases. *European Journal of Biochemistry,* **269**, 397–412.

Franzen,L., Rochaix,J. and von Heijne,G. (1990) Chloroplast Transit Peptides From The Green-Alga Chlamydomonas-Reinhardtii Share Features With Both Mitochondrial And Higher-Plant Chloroplast Presequences. *FEBS Letters,* **260** (2), 165–168.

Freire,M. and Pages,M. (1995) Functional characteristics of the maize RNA-binding protein MA16. *Plant Molecular Biology,* **29**, 797–807.

Friedrichsen,D., Joazeiro,C., Li,J., Hunter,T. and Chory,J. (2000) Brassinosteroid-insensitive-1 is a ubiquitously expressed leucine-rich repeat receptor serine/threonine kinase. *Plant Physiology,* **123** (4), 1247–1256.

Furbank,R.T. and Taylor,W.C. (1995) Regulation of Photosynthesis in C3 and C4 Plants: A Molecular Approach. *Plant Cell,* **7** (7), 797–807.

Furuichi,T., Mori,I.C., Takahashi,K. and Muto,S. (2001) Sugar-Induced Increase in Cytosolic Ca2+ in Arabidopsis thaliana Whole Plants. *Plant Cell Physiology,* **42** (10), 1149–1155.

Gaffney,T., Friedrich,L., Vernooij,B., Negrotto,D., Nye,G., Uknes,S., Ward,E., Kessmann,H. and Ryals,J. (1993) Requirement of salicylic acid for the induction of systemic acquired resistance. *Science,* **261**, 754–756.

Galston,A. and Sawhney,R. (1990) Polyamines in plant physiology. *Plant Physiology,* **94**, 406–410.

Garcia,A., Kido,E., Meza,A., Souza,H., Pinto,L., Pastina,M., Leite,C., Silva,J., Ulian,E., Figueira,A. and Souza,A. (2006) Development of an integrated genetic map of a sugarcane (Saccharum spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. *Theoretical and Applied Genetics,* **112** (2), 298–314.

Gazzarrini,S. and McCourt,P. (2003) Cross-talk in plant hormone signalling: what Arabidopsis mutants are telling us. *Annals of Botany,* **91**, 605–612.

Geisler-Lee,J., O'Toole,N., Ammar,R., Provart,N.J., Millar,A. and Geisler,M. (2007) A Predicted Interactome for Arabidopsis. *Plant Physiology,* **145** (2), 317–329.

Gelhaye,E., Rouhier,N. and Jacquot,J. (2004) The thioredoxin H system of higher plants. *Plant Physiology and Biochemistry,* **42**, 265–271.

Gibson,S. (2004) Sugar and phytohormone response pathways: navigating a signalling network. *Journal of Experimental Botany,* **55**, 253–264.

Gibson,S. (2005) Control of plant development and gene expression by sugar signaling. *Current Opinion in Plant Biology,* **8** (1), 93–102.

Glaser,E., Sjoling,S., Tanudji,M. and Whelan,J. (1998) Mitochondrial protein import in plants - Signals, Sorting targeting, processing and regulation. *Plant Molecular Biology,* **38** (1-2), 311–338.

Goddijn,O.J., Lindsey,K., Lee,F.M., Klap,J.C. and Sijmons,P.C. (1993) Differential gene expression in nematode-induced feeding structures of transgenic plants harbouring promoter-gusA fusion constructs. *The Plant Journal,* **4** (5), 863–873.

Goetz,M., Godt,D. and Roitsch,T. (2000) Tissue-specific induction of the mRNA for an extracellular invertase isoenzyme of tomato by brassinosteroids suggests a role for steroid hormones in assimilate partitioning. *The Plant Journal,* **22** (6), 515–522.

Grenier,J., Potvin,C., Trudel,J. and Asselin,A. (1999) Some thaumatin-like proteins hydrolyse polymeric beta-1,3-glucans. *The Plant Journal,* **19**, 473–480.

Gribskov,M., Fana,F., Harper,J., Hope,D.A., Harmon,A.C., Smith,D.W., Tax,F.E. and Zhang,G. (2001) PlantsP: a functional genomics database for plant phosphorylation. *Nucleic Acids Research,* **29** (1), 111–113.

Grivet,L. and Arruda,P. (2002) Sugarcane genomics: depicting the complex genome of an important tropical crop. *Current Opinion in Plant Biology,* **5** (2), 122–127.

Guimaraes,C.T., Sills,G.R. and Sobral,B.W. (1997) Comparative mapping of Andropogoneae: Saccharum L. (sugarcane) and its relation to sorghum and maize. *Proceedings of the National Academy of Sciences,* **94** (26), 14261–14266.

Guo,J., Lin,Y. and Sun,Z. (2004*a*) A novel method for protein subcellular localization based on boosting and probabilistic neural network. *Proceedings of the second conference on Asia-Pacific bioinformatics,* **29**, 21–27.

Guo,T., Hua,S., Ji,X. and Sun,Z. (2004*b*) DBSubLoc: database of protein subcellular localization. *Nucleic Acids Research,* **32** (90001), D122–D124.

Gutierrez,R., Green,P., Keegstra,K. and Ohlrogge,J. (2004) Phylogenetic profiling of the Arabidopsis thaliana proteome: what proteins distinguish plants from other organisms? *Genome Biology,* **5** (8), R53.

Hagen,G., Kleinschmidt,A. and Guilfoyle,T. (1984) Auxin-regulated gene expression in intact soybean hypocotyl and excised hypocotyl sections. *Planta,* **162**, 147–153.

Halford,N. and Paul,M. (2003) Carbon metabolite sensing and signalling. *Plant Biotechnol Journal,* **1** (6), 381–398.

Hammond,J., Broadley,M. and White,P. (2004) Genetic responses to phosphorus deficiency. *Annals of Botany,* **94**, 323–332.

Hanks,S. and Hunter,T. (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain strucuture and classification. *FASEB Journal,* **9**, 576–596.

Hanks,S. and Quinn,A. (1991) Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. *Methods in Enzymology,* **200**, 38–62.

Hanks,S., Quinn,A. and Hunter,T. (1988) The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science,* **241** (4861), 42–52.

Hardie,D. (1999) PLANT PROTEIN SERINE/THREONINE KINASES: Classification and Functions. *Annual Review of Plant Physiology and Plant Molecular Biology,* **50** (1), 97–131.

Heazlewood,J.L., Tonti-Filippini,J.S., Gout,A.M., Day,D.A., Whelan,J. and Millar,A. (2004) Experimental Analysis of the Arabidopsis Mitochondrial Proteome Highlights Signaling and Regulatory Components, Provides Assessment of Targeting Prediction Programs, and Indicates Plant-Specific Mitochondrial Proteins. *Plant Cell,* **16** (1), 241–256.

Heazlewood,J.L., Verboom,R.E., Tonti-Filippini,J., Small,I. and Millar,A. (2007) SUBA: the Arabidopsis Subcellular Database. *Nucleic Acids Research,* **35**, D213–D218.

Hecht,V., Vielle-Calzada,J.P., Hartog,M.V., Schmidt,E.D., Boutilier,K., Grossniklaus,U. and de Vries,S.C. (2001) The Arabidopsis Somatic Embryogenesis Receptor Kinase 1 Gene Is Expressed in Developing Ovules and Embryos and Enhances Embryogenic Competence in Culture. *Plant Physiology,* **127** (3), 803–816.

Hendre,K., Iyer,R., Kotwain,M., Kluspe,S. and Mascarenhas,A. (1983) Rapid multiplication of sugarcane by tissue culture. *Sugarcane,* **1**, 5–8.

Hennig,L., Gruissem,W., Grossniklaus,U. and Kohler,C. (2004) Transcriptional Programs of Early Reproductive Stages in Arabidopsis. *Plant Physiology,* **135** (3), 1765–1775.

Hiller,K., Grote,A., Scheer,M., Munch,R. and Jahn,D. (2004) PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Research,* **32**, W375–W379.

Hiller,M., Pudimat,R., Busch,A. and Backofen,R. (2006) Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Research,* **34** (17), e117.

Himmelbach,A., Yang,Y. and Grill,E. (2003) Relay and control of abscisic acid signaling. *Current Opinion in Plant Biology,* **6**, 470–479.

Ho,S., Chao,Y., Tong,W. and Yu,S. (2001) Sugar coordinately and differentially regulates growth- and stress-related gene expression via a complex signal transduction network and multiple control mechanisms. *Plant Physiology,* **125** (2), 877–890.

Hoagland,D. and Arnon,D. (1950) *Califórnia Agricultural Experiment Station*. Berkely: California Agr. Expt. Sta. Cir p. 347.

Hooper,S.D. and Bork,P. (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics,* **21** (24), 4432–4433.

Howe,G. and Schilmiller,A. (2002) Oxylipin metabolism in response to stress. *Current Opinion in Plant Biology,* **5**, 230–236.

Hrabak,E., Chan,C., Gribskov,M., Harper,J., Choi,J., Halford,N., Kudla,J., Luan,S., Nimmo,H., Sussman,M., Thomas,M., Walker-Simmons,K., Zhu,J. and Harmon,A. (2000) The Arabidopsis CDPK-SnRK superfamily of protein kinases. *Plant Physiology,* **132**, 666–680.

Hu,H., Xiong,L. and Yang,Y. (2005) Rice SERK1 gene positively regulates somatic embryogenesis of cultured cell and host defense response against fungal infection. *Planta,* **222** (1), 107–117.

Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics,* **17** (8), 721–728.

Huang,J., Hack,E., Thornburg,R. and Myers,A. (1990) A Yeast Mitochondrial Leader Peptide Functions in vivo as a Dual Targeting Signal for Both Chloroplasts and Mitochondria. *Plant Cell,* **2** (12), 1249–1260.

Huber,J., Huber,S. and Nielsen,T. (1989) Protein phosphorylation as a mechanism for regulation of spinach leaf sucrose-phosphate synthase activity. *Archives of Biochemistry and Biophysics,* **270** (2), 681–690.

Huber,S.C. and Huber,J.L. (1996) Role and regulation of sucrose-phosphate synthase in higher plants. *Annual Review of Plant Physiology and Plant Molecular Biology,* **47** (1), 431–444.

Huber,S.C., MacKintosh,C. and Kaiser,W.M. (2002) Metabolic enzymes as targets for 14-3-3 proteins. *Plant Molecular Biology,* **50** (6), 1053–1063.

Hurt,E., Soltanifar,N., Goldschmidtclermont,M., Rochaix,J. and Schatz,G. (1986) The Cleavable Pre-Sequence of an Imported Chloroplast Protein Directs Attached Polypeptides into Yeast Mitochondria. *EMBO Journal,* **5** (6), 1343–1350.

Im,Y., Han,O., Chung,G. and Cho,B. (2002) Antisense expression of an Arabidopsis omega-3 fatty acid desaturase gene reduces salt/drought tolerance in transgenic tobacco plants. *Molecular Cell,* **13**, 264–271.

Initiative,A.G. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature,* **408** (6814), 796–815.

Inoue,K., Sewalt,V.J., Murray Ballance,G., Ni,W., Sturzer,C. and Dixon,R.A. (1998) Developmental Expression and Substrate Specificities of Alfalfa Caffeic Acid 3-O-Methyltransferase and Caffeoyl Coenzyme A 3-O-Methyltransferase in Relation to Lignification. *Plant Physiology,* **117** (3), 761–770.

Iseli,C., Jongeneel,C. and Bucher,P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proceedings of International Conference on Intelligent Systems for Molecular Biology,* **1**, 138–148.

Iskandar,H., Simpson,R., Casu,R., Bonnett,G., MacLean,D. and Manners,J. (2004) Comparison of Reference Genes for Quantitative Real-Time Polymerase Chain Reaction Analysis of Gene Expression in Sugarcane. *Plant Molecular Biology Reporter,* **22**, 325–337.

Jacobs,G.H., Stockwell,P.A., Tate,W.P. and Brown,C.M. (2006) Transterm - extended search facilities and improved integration with other databases. *Nucleic Acids Research,* **34**, D37–D40.

James,E. and Olivares,F. (1998) Infection and Colonization of Sugar Cane and Other Graminaceous Plants by Endophytic Diazotrophs. *Critical Reviews in Plant Sciences,* **17**, 77–119.

James,E., Reis,V., Olivares,F., Baldani,J. and Dobereiner,J. (1994) Infection of sugar cane by the nitrogen-fixing bacterium Acetobacter diazotrophicus. *Journal of Experimental Botany,* **45**, 757–766.

Jansen,R. and Nap,J. (2001) Genetical genomics: the added value from segregation. *Trends in Genetics,* **17** (7), 388–391.

Jonak,C. and Hirt,H. (2002) Glycogen synthase kinase 3/SHAGGY-like kinases in plants: an emerging family with novel functions. *Trends in Plant Science,* **7**, 457–461.

Joshi,C.P., Zhou,H., Huang,X. and Chiang,V.L. (1997) Context sequences of translation initiation codon in plants. *Plant Mol Biology,* **35**, 993–1001.

Joubes,J., De Schutter,K., Verkest,A., Inze,D. and De Veylder,L. (2004) Conditional, recombinase-mediated expression of genes in plant cell cultures. *The Plant Journal,* **37** (6), 889–896.

Karimi,M., De Meyer,B. and Hilson,P. (2005) Modular cloning in plant cells. *Trends in Plant Science,* **10** (3), 103–105.

Karimi,M., Inze,D. and Depicker,A. (2002) GATEWAY vectors for Agrobacterium-mediated plant transformation. *Trends in Plant Science,* **7** (5), 193–195.

Kates,R. and Parris,T. (2003) Long-term trends and a sustainability transition. *The Proceedings of the National Academy of Sciences,* **100**, 8062–8067.

Kawaguchi,R. and Bailey-Serres,J. (2005) mRNA sequence features that contribute to translational regulation in Arabidopsis. *Nucleic Acids Research,* **33**, 955–965.

Kawasaki,S., Borchert,C., Deyholos,M., Wang,H., Brazille,S., Kawai,K., Galbraith,D. and Bohnert,H. (2001) Gene expression profiles during the initial phase of salt stress in rice. *Plant Cell,* **13**, 889–905.

Kim,S., Kwak,H., Park,M. and Kim,S. (1998) Induction of reproductive organ-preferential histone genes by wounding or methyl jasmonate. *Molecular Cell,* **8**, 669–677.

King,E. and Hartley,G. (1985) Diatraea saccharalis. *Handbook of insect rearing,* **2**, 265–270.

Kircher,S., Gil,P., Kozma-Bognar,L., Fejes,E., Speth,V., Husselstein-Muller,T., Bauer,D., Adam,E., Schafer,E. and Nagy,F. (2002) Nucleocytoplasmic partitioning of the plant photoreceptors phytochrome A, B, C, D, and E is regulated differentially by light and exhibits a diurnal rhythm. *Plant Cell,* **14** (7), 1541–1555.

Kobe,B. and Deisenhofer,J. (1994) The leucine-rich repeat: a versatile binding motif. *Trends in Biochemical Sciences,* **19** (10), 415–421.

Koch,K. (1996) Carbohydrate-modulated gene expression in plants. *Annual Review of Plant Physiology and Plant Molecular Biology,* **47** (1), 509–540.

Koch,K. (2004) Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. *Current Opinion in Plant Biology,* **7** (3), 235–246.

Kochetov,A.V. (2005) AUG codons at the beginning of protein coding sequences are frequent in eukaryotic mRNAs with a suboptimal start codon context. *Bioinformatics,* **21** (7), 837–840.

Kochetov,A.V. and Sarai,A. (2004) Translational polymorphism as a potential source of plant proteins variety in *Arabidopsis thaliana*. *Bioinformatics,* **20**, 445–447.

Kochian,L. (2000) *Biochemistry and Molecular Biology of plants*. Rockyville, Maryland: American Society of Plant Physiologists pp. 1204–1249.

Koide,T., Salem-Izaac,S., Gomes,S. and Vencio,R. (2006) SpotWhatR: a user-friendly microarray data analysis system. *Genetics and Molecular Research,* **5**, 93–107.

Koide,T., Zaini,P.A., Moreira,L.M., Vencio,R.Z., Matsukuma,A.Y., Durham,A.M., Teixeira,D.C., El-Dorry,H., Monteiro,P.B., da Silva,A.C., Verjovski-Almeida,S., da Silva,A.M. and Gomes,S.L. (2004) DNA Microarray-Based Genome Comparison of a Pathogenic and a Nonpathogenic Strain of Xylella fastidiosa Delineates Genes Important for Bacterial Virulence. *The Journal of Bacteriology,* **186** (16), 5442–5449.

Kolomiets,M.V., Hannapel,D.J., Chen,H., Tymeson,M. and Gladon,R.J. (2001) Lipoxygenase Is Involved in the Control of Potato Tuber Development. *Plant Cell,* **13** (3), 613–626.

Komatsu,S., Kojima,K., Suzuki,K., Ozaki,K. and Higo,K. (2004) Rice Proteome Database based on two-dimensional polyacrylamide gel electrophoresis: its status in 2003. *Nucleic Acids Research,* **32** (90001), D388–D392.

Kozak,M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research,* **15**, 8125–8148.

Kozak,M. (1990) Downstream Secondary Structure Facilitates Recognition of Initiator Codons by Eukaryotic Ribosomes. *Proceedings of the National Academy of Sciences,* **87** (21), 8301–8305.

Kozak,M. (1991) An analysis of vertebrate mRNA sequences: intimations of translational control. *The Journal of Cell Biology,* **115**, 887–903.

Kozak,M. (2001) New ways of initiating translation in eukaryotes? *Molecular and Cellular Biology,* **21**, 1899–1907.

Kozak,M. (2002) Pushing the limits of the scanning mechanism for initiation of translation. *Gene,* **299**, 1–34.

Kozak,M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene,* **361**, 13–37.

Krens,F., Molendijk,L., Wullems,G. and Schilperoort,R. (1982) In vitro transformation of plant protoplasts with Ti-plasmid DNA. *Nature,* **296** (5852), 72–74.

Kutay,U. and Muhlhausser,P. (2006) Cell biology: taking a turn into the nucleus. *Nature,* **442** (7106), 991–992.

Lawton,M., Yamamoto,R., Hanks,S. and Lamb,C. (1989) Molecular cloning of plant transcripts encoding protein kinase homologs. *The Proceedings of the National Academy of Sciences,* **86**, 3140–3144.

Levy,J., Bres,C., Geurts,R., Chalhoub,B., Kulikova,O., Duc,G., Journet,E., Ane,J., Lauber,E., Bisseling,T., Denarie,J., Rosenberg,C. and Debelle,F. (2004) A putative Ca2+ and calmodulin-dependent protein kinase required for bacterial and fungal symbioses. *Science,* **303**, 1361–1364.

Lewis,N.G. and Yamamoto,E. (1990) Lignin: Occurrence, Biogenesis and Biodegradation. *Annual Review of Plant Physiology and Plant Molecular Biology,* **41** (1), 455–496.

Li,J. and Assmann,S. (1996) An abscisic acid activated and calcium-independent protein kinase from guard cells of fava bean. *Plant Cell,* **8**, 2359–2368.

Li,J. and Jin,H. (2006) Regulation of brassinosteroid signaling. *Trends in Plant Science,* **12** (1), 37–41.

Li,J., Wang,X., Watson,M. and Assmann,S. (2000) Regulation of abscisic acid-induced stomatal closure and anion channels by guard cell AAPK kinase. *Science,* **287**, 300–303.

Li,J., Wen,J., Lease,K., Doke,J., Tax,F. and Walker,J. (2002) BAK1, an Arabidopsis LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signaling. *Cell,* **110** (2), 213–222.

Li,Z. and Chen,S. (2000) Differential accumulation of the S-adenosylmethionine decarboxylase transcript in rice seedlings in response to salt and drought stresses. *Theoretical and Applied Genetics,* **100**, 782–788.

Liang,L., Lai,Z., Ma,W., Zhang,Y. and Xue,Y. (2002) AhSL28, a senescence- and phosphate starvation-induced S-like RNase gene in Antirrhinum. *Biochimica et Biophysica Acta,* **1579** (1), 64–71.

Lilley,K. and Dupree,P. (2007) Plant organelle proteomics. *Current Opinion in Plant Biology,* **10** (6), 594–599.

Lin,F., Xu,S., Ni,W., Chu,Z., Xu,Z. and Xue,H. (2003) Identification of ABA-responsive genes in rice shoots via cDNA macroarray. *Cell Research,* **13**, 59–68.

Lin,J., Qian,J., Greenbaum,D., Bertone,P., Das,R., Echols,N., Senes,A., Stenger,B. and Gerstein,M. (2002) GeneCensus: genome comparisons in terms of metabolic pathway activity and protein family sharing. *Nucleic Acids Research,* **30** (20), 4574–4582.

Liu,J., Ishitani,M., Halfter,U., Kim,C. and Zhu,J. (2000) The Arabidopsis thaliana SOS2 gene encodes a protein kinase that is required for salt tolerance. *The Proceedings of the National Academy of Sciences,* **97**, 3730–3734.

Liu,J., Kang,S., Tang,C., Ellis,L.B. and Li,T. (2007) Meta-prediction of protein subcellular localization with reduced voting. *Nucleic Acids Research,* **35** (15), e96.

Livak,K. and Schmittgen,T. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods,* **25** (4), 402–408.

Lopez-Bucio,J., Hernandez-Abreu,E., Sanchez-Calderon,L., Nieto-Jacobo,M., Simpson,J. and Herrera-Estrella,L. (2002) Phosphate availability alters architecture and causes changes in hormone sensitivity in the Arabidopsis root system. *Plant Physiology,* **129**, 244–256.

Ludwig-Muller,J., Schubert,B. and Pieper,K. (1995) Regulation of IBA synthetase from maize (Zea mays L.) by drought stress and ABA. *Journal of Experimental Botany,* **46**, 423–432.

Luehrsen,K.R. and Walbot,V. (1994) The impact of AUG start codon context on maize gene expression in vivo. *Plant Cell Reports,* **13**, 454–458.

Lukaszewicz,M., Feuermann,M., Jerouville,B., Stas,A. and Boutry,M. (2000) In vivo evaluation of the context sequence of the translation initiation codon in plants. *Plant Science,* **154**, 89–98.

Lunn,J. and Furbank,R. (1999) Sucrose biosynthesis in C4 plants. *New phytologist,* **143**, 221–237.

Lunn,J.E. (2007) Compartmentation in plant metabolism. *Journal of Experimental Botany,* **58** (1), 35–47.

Ma,H., Schulze,S., Lee,S., Yang,M., Mirkov,E., Irvine,J., Moore,P. and Paterson,A. (2004) An EST survey of the sugarcane transcriptome. *Theoretical and Applied Genetics,* **108** (5), 851–863.

Madsen,E., Madsen,L., Radutoiu,S., Olbryt,M., Rakwalska,M., Szczyglowski,K., Sato,S., Kaneko,T., Tabata,S., Sandal,N. and Stougaard,J. (2003) A receptor kinase gene of the LysM type is involved in legume perception of rhizobial signals. *Nature,* **425**, 637–640.

Majerus,P.W., Kisseleva,M.V. and Norris,F.A. (1999) The Role of Phosphatases in Inositol Signaling Reactions. *Journal of Biological Chemistry,* **274** (16), 10669–10672.

Manning,G., Plowman,G., Hunter,T. and Sudarsanam,S. (2002*a*) Evolution of protein kinase signaling from yeast to man. *Trends in Biochemical Sciences,* **27** (10), 514–520.

Manning,G., Whyte,D., Martinez,R., Hunter,T. and Sudarsanam,S. (2002*b*) The Protein Kinase Complement of the Human Genome. *Science,* **298** (5600), 1912–1934.

Mao,C., Wang,S., Jia,Q. and Wu,P. (2006) OsEIL1, a rice homolog of the Arabidopsis EIN3 regulates the ethylene response as a positive component. *Plant Molecular Biology,* **61**, 141–152.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta,* **405**, 442–451.

McCarty,D. and Chory,J. (2000) Conservation and innovation in plant signaling pathways. *Cell,* **103** (2), 201–209.

McClung,C.R. (2001) Circadian Rhythms in Plants. *Annual Review of Plant Physiology and Plant Molecular Biology,* **52** (1), 139–162.

McCormick,A.J., Cramer,M.D. and Watt,D.A. (2006) Sink strength regulates photosynthesis in sugarcane. *New Phytologist,* **171** (4), 759–770.

McCormick,A.J., Cramer,M.D. and Watt,D.A. (2007) Changes in Photosynthetic Rates and Gene Expression of Leaves during a Source Sink Perturbation in Sugarcane. *Annals of Botany,* **101** (1), 89–102.

McMichael,R.W.,J., Klein,R., Salvucci,M. and Huber,S. (1993) Identification of the major regulatory phosphorylation site in sucrose-phosphate synthase. *Archives of Biochemistry and Biophysics,* **307** (2), 248–252.

Merlot,S., Gosti,F., Guerrier,D., Vavasseur,A. and Giraudat,J. (2001) The ABI1 and ABI2 protein phosphatases 2C act in a negative feedback regulatory loop of the abscisic acid signalling pathway. *The Plant Journal,* **25**, 295–303.

Messing,J. (2005) The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications. *BMC Biology,* **3** (1), 20.

Meyers,B.C., Galbraith,D.W., Nelson,T. and Agrawal,V. (2004) Methods for Transcriptional Profiling in Plants. Be Fruitful and Replicate. *Plant Physiology,* **135** (2), 637–652.

Mignone,F., Gissi,C., Liuni,S. and Pesole,G. (2002) Untranslated regions of mRNAs. *Genome Biology,* **3**, reviews0004.

Mignone,F., Grillo,G., Licciulli,F., Iacono,M., Liuni,S., Kersey,P.J., Duarte,J., Saccone,C. and Pesole,G. (2005) UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Research,* **33**, D141–D146.

Millar,A. (2004) Location, location, location: surveying the intracellular real estate through proteomics in plants. *Functional Plant Biology,* **31** (6), 563–582.

Millar,A., Whelan,J. and Small,I. (2006) Recent surprises in protein targeting to mitochondria and plastids. *Current Opinion in Plant Biology,* **9** (6), 610–615.

Ming,R., Liu,S.C., Moore,P.H., Irvine,J.E. and Paterson,A.H. (2001) QTL Analysis in a Complex Autopolyploid: Genetic Control of Sugar Content in Sugarcane. *Genome Research,* **11** (12), 2075–2084.

Minhas,D. and Grover,A. (1999) Transcript level of genes encoding various glycolytic and fermentation enzymes change in response to abiotic stresses. *Plant Science,* **146**, 41–51.

Moller,S. and Chua,N. (1999) Interactions and intersections of plant signaling pathways. *Journal of Molecular Biology,* **293**, 219–234.

Moore,P. (2005) Integration of sucrose accumulation processes across hierarchical scales: towards developing an understanding of the gene-to-crop continuum. *Field Crops Research,* **92**, 119–135.

Moorhead,G., Douglas,P., Cotelle,V., Harthill,J., Morrice,N., Meek,S., Deiting,U., Stitt,M., Scarabel,M., Aitken,A. and MacKintosh,C. (1999) Phosphorylation-dependent interactions between enzymes of plant metabolism and 14-3-3 proteins. *The Plant Journal,* **18** (1), 1–12.

Moreau,P., Brandizzi,F., Hanton,S., Chatre,L., Melser,S., Hawes,C. and Satiat-Jeunemaitre,B. (2007) The plant ER-Golgi interface: a highly structured and dynamic membrane complex. *Journal of Experimental Botany,* **58** (1), 49–64.

Morillo,S. and Tax,F. (2006) Functional analysis of receptor-like kinases in monocots and dicots. *Current Opinion in Plant Biology,* **9** (5), 460–469.

Morsy,M., Almutairi,A., Gibbons,J., Yun,S. and de Los Reyes,B. (2005) The OsLti6 genes encoding low-molecular-weight membrane proteins are differentially expressed in rice cultivars with contrasting sensitivity to low temperature. *Gene,* **344**, 171–180.

Mousavi,A. and Hotta,Y. (2005) Glycine-rich proteins: a class of novel proteins. *Applied Biochemistry Biotechnology,* **120**, 169–174.

Muchhal,U., Pardo,J. and Raghothama,K. (1996) Phosphate transporters from the higher plant Arabidopsis thaliana. *Proceedings of the National Academy of Sciences,* **93**, 10519–10523.

Mur,L., Bi,Y., Darby,R., Firek,S. and Draper,J. (1997) Compromising early salicylic acid accumulation delays the hypersensitive response and increases viral dispersion during lesion establishment in TMV-infected tobacco. *The Plant Journal,* **12**, 1113–1126.

Murashige,T. and Skoog,F. (1962) A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. *Physiologia Plantarum,* **15** (3), 473–497.

Mustilli,A., Merlot,S., Vavasseur,A., Fenzi,F. and Giraudat,J. (2002) Arabidopsis OST1 protein kinase mediates the regulation of stomatal aperture by abscisic acid and acts upstream of reactive oxygen species production. *Plant Cell,* **14**, 3089–3099.

Nadershahi,A., Fahrenkrug,S.C. and Ellis,L.B.M. (2004) Comparison of computational methods for identifying translation initiation sites in EST data. *BMC Bioinformatics,* **5**, 14.

Nag,R., Maity,M. and Dasgupta,M. (2005) Dual DNA binding property of ABA insensitive 3 like factors targeted to promoters responsive to ABA and auxin. *Plant Molecular Biology,* **59**, 821–838.

Nahm,M., Kim,S., Yun,D., Lee,S., Cho,M. and Bahk,J. (2003) Molecular and biochemical analyses of OsRab7, a rice Rab7 homolog. *Plant Cell Physiology,* **44**, 1341–1349.

Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Sciences,* **24** (1), 34–36.

Nakashima,H. and Nishikawa,K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *Journal of Molecular Biology,* **238**, 54–61.

Nam,K. and Li,J. (2002) BRI1/BAK1, a receptor kinase pair mediating brassinosteroid signaling. *Cell,* **110** (2), 203–212.

Narusaka,Y., Narusaka,M., Seki,M., Umezawa,T., Ishida,J., Nakajima,M., Enju,A. and Shinozaki,K. (2004) Crosstalk in the responses to abiotic and biotic stresses in Arabidopsis : analysis of gene expression in cytochrome P450 gene superfamily by cDNA microarray. *Plant Molecular Biology,* **55**, 327–342.

Nemhauser,J. and Chory,J. (2004) BRing it on: new insights into the mechanism of brassinosteroid action. *Journal of Experimental Botany,* **55** (395), 265–270.

Nielsen,H., Brunak,S. and von Heijne,G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Engineering,* **12**, 3–9.

Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering,* **10**, 1–6.

Nishiuchi,T. and Iba,K. (1998) Roles of plastid e-3 fatty acid desaturases in defense response of higher plants. *Journal of Plant Research,* **111** (4), 481–486.

Nogueira,F., De Rosa,V., Menossi,M., Ulian,E. and Arruda,P. (2003) RNA expression profiles and data mining of sugarcane response to low temperature. *Plant Physiology,* **132**, 1811–1824.

Nolan,K.E., Irwanto,R.R. and Rose,R.J. (2003) Auxin Up-Regulates MtSERK1 Expression in Both Medicago truncatula Root-Forming and Embryogenic Cultures. *Plant Physiology,* **133** (1), 218–230.

Nonomura,K.I., Miyoshi,K., Eiguchi,M., Suzuki,T., Miyao,A., Hirochika,H. and Kurata,N. (2003) The MSP1 Gene Is Necessary to Restrict the Number of Cells Entering into Male and Female Sporogenesis and to Initiate Anther Wall Formation in Rice. *Plant Cell,* **15** (8), 1728–1739.

Obayashi,T., Okegawa,T., Sasaki-Sekimoto,Y., Shimada,H., Masuda,T., Asamizu,E., Nakamura,Y., Shibata,D., Tabata,S., Takamiya,K.i. and Ohta,H. (2004) Distinctive Features of Plant Organs Characterized by Global Analysis of Gene Expression in Arabidopsis. *DNA Research,* **11** (1), 11–25.

Ohto,M. and Nakamura,K. (1995) Sugar-induced increase of calcium-dependent protein kinases associated with the plasma membrane in leaf tissues of tobacco. *Plant Physiology,* **109** (3), 973–981.

Olivares,F., James,E., Baldani,J. and Dobereiner,J. (1997) Infection of mottled stripe disease-susceptible and resistant sugarcane varieties by the endophytic diazotrophs Herbaspirillum. *New Phytology,* **135**, 723–727.

O'Mahony,P. and Oliver,M. (1999) Characterization of a desiccation-responsive small GTP-binding protein (Rab2) from the desiccation-tolerant grass Sporobolus stapfianus. *Plant Molecular Biology,* **39**, 809–821.

Oono,Y., Seki,M., Nanjo,T., Narusaka,M., Fujita,M., Satoh,R., Satou,M., Sakurai,T., Ishida,J., Akiyama,K., Iida,K., Maruyama,K., Satoh,S., Yamaguchi-Shinozaki,K. and Shinozaki,K. (2003) Monitoring expression profiles of Arabidopsis gene expression during rehydration process after dehydration using ca 7,000 full-length cDNA microarray. *The Plant Journal,* **34**, 868–887.

Oono,Y., Seki,M., Satou,M., Iida,K., Akiyama,K., Sakurai,T., Fujita,M., Yamaguchi-Shinozaki,K. and Shinozaki,K. (2006) Monitoring expression profiles of Arabidopsis genes during cold acclimation and deacclimation using DNA microarrays. *Functional & Integrative Genomics,* **4**, 1–23.

O'Rourke,N.A., Meyer,T. and Chandy,G. (2005) Protein localization studies in the age of 'Omics'. *Current Opinion in Chemical Biology,* **9** (1), 82–87.

Osuna,D., Usadel,B., Morcuende,R., Gibon,Y., Blasing,O., Hohne,M., Gunter,M., Kamlage,B., Trethewey,R., Scheible,W. and Stitt,M. (2007) Temporal responses of transcripts, enzyme activities and metabolites after adding sucrose to carbon-deprived Arabidopsis seedlings. *The Plant Journal,* **49** (3), 463–491.

Page,R. (1996) TreeView: an application to display phylogenetic trees on personal computers. *Computer Applications In The Biosciences,* **12** (4), 357–358.

Palmer,E. and Freeman,T. (2004) Investigation into the use of C- and N-terminal GFP fusion proteins for subcellular localization studies using reverse transfection microarrays. *Comparative and functional genomics,* **5**, 342–353.

Papini-Terzi,F., Rocha,F., Vencio,R., Oliveira,K., Felix,J., Vicentini,R., Rocha,C.S., Simoes,A., Ulian,E., Di Mauro,S., da Silva,A., Pereira,C., Menossi,M. and Souza,G. (2005) Transcription profiling of signal transduction-related genes in sugarcane tissues. *DNA Research,* **12** (1), 27–38.

Paterson,A.H., Bowers,J.E., Burow,M.D., Draye,X., Elsik,C.G., Jiang,C.X., Katsar,C.S., Lan,T.H., Lin,Y.R., Ming,R. and Wright,R.J. (2000) Comparative Genomics of Plant Chromosomes. *Plant Cell,* **12** (9), 1523–1540.

Paterson,A.H., Bowers,J.E. and Chapman,B.A. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences,* **101** (26), 9903–9908.

Paul,M.J. and Foyer,C.H. (2001) Sink regulation of photosynthesis. *Journal of Experimental Botany,* **52** (360), 1383–1400.

Pearce,R.S. (1999) Molecular analysis of acclimation to cold. *Plant Growth Regulation,* **29** (1), 47–76.

Pessoa-Jr,A., Roberto,I., Menossi,M., dos Santos,R., Filho,S. and Penna,T. (2005) Perspectives on bioenergy and biotechnology in Brazil. *Applied Biochemistry Biotechnology,* **121-124**, 59–70.

Pestova,T.V., Kolupaeva,V.G., Lomakin,I.B., Pilipenko,E.V., Shatsky,I.N., Agol,V.I. and Hellen,C.U.T. (2001) Molecular mechanisms of translation initiation in eukaryotes. *The Proceedings of the National Academy of Sciences,* **98**, 7029–7036.

Petersen,M., Brodersen,P., Naested,H., Andreasson,E., Lindhart,U., Johansen,B., Nielsen,H., Lacy,M., Austin,M., Parker,J., Sharma,S., Klessig,D., Martienssen,R., Mattsson,O., Jensen,A. and Mundy,J. (2000) Arabidopsis map kinase 4 negatively regulates systemic acquired resistance. *Cell,* **103**, 1111–1120.

Petsalakis,E.I., Bagos,P.G., Litou,Z.I. and Hamodrakas,S.J. (2005) N-terminal sequence-based prediction of subcellular location. *BMC Bioinformatics,* **6**, S11.

Pfaller,R., Pfanner,N. and Neupert,W. (1989) Mitochondrial Protein Import - Bypass of Proteinaceous Surface-Receptors can occur with Low Specificity and Efficiency. *Journal of Biological Chemistry,* **264** (1), 34–39.

Porta,H. and Rocha-Sosa,M. (2002) Plant Lipoxygenases. Physiological and Molecular Features. *Plant Physiology,* **130** (1), 15–21.

Prescha,A., Swiedrych,A., Biernat,J. and Szopa,J. (2001) Increase in Lipid Content in Potato Tubers Modified by 14-3-3 Gene Overexpression. *Journal of Agriculture and Food Chemistry,* **49** (8), 3638–3643.

Puigdomènech,C.R.P. (2006) Specific use of start codons and cellular localization of splice variants of human phosphodiesterase 9A gene. *BMC Molecular Biology,* **7**, 39.

Pujol,C., Maréchal-Drouard,L. and Duchêne,A. (2007) How can organellar protein N-terminal sequences be dual targeting signals? In silico analysis and mutagenesis approach. *Journal of Molecular Biology,* **369**, 356–367.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nature Reviews Genetics,* **2**, 418–427.

Raghothama,K. (1999) Phosphate acquisition. *Annual Review of Plant Physiology and Plant Molecular Biology,* **50**, 665–693.

Raghothama,K. (2000) Phosphate transport and signaling. *Current Opinion in Plant Biology,* **3**, 182–187.

Reboul,J., Vaglio,P., Rual,J., Lamesch,P., Martinez,M., Armstrong,C., Li,S., Jacotot,L., Bertin,N., Janky,R., Moore,T., Hudson,J.R.,J., Hartley,J., Brasch,M., Vandenhaute,J., Boulton,S., Endress,G., Jenna,S., Chevet,E., Papasotiropoulos,V., Tolias,P., Ptacek,J., Snyder,M., Huang,R., Chance,M., Lee,H., Doucette-Stamm,L., Hill,D. and Vidal,M. (2003) C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genetics,* **34** (1), 35–41.

Reents,H., Münch,R., Dammeyer,T., Jahn,D. and Härtig,E. (2006) The Fnr regulon of *Bacillus subtilis*. *The Journal of Bacteriology,* **188**, 1103–1112.

Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Research,* **26** (9), 2230–2236.

Reinhold-Hurek,B. and Hurek,T. (1998) Life in grasses: diazotrophic endophytes. *Trends in Microbiology,* **6**, 139–144.

Reis,V., Olivares,F. and Dobereiner,J. (1994) Improved methodology for isolation of Acetobacter diazotrophicus and confirmation of its endophythic habitat. *World Journal of Microbiology and Biotechnology,* **10**, 101–104.

Rentero,C., Monfort,A. and Puigdomènech,P. (2003) Identification and distribution of different mRNA variants produced by differential splicing in the human phosphodiesterase 9A gene. *Biochemical and Biophysical Research Communications,* **301**, 686–692.

Riera,M., Valon,C., Fenzi,F., Giraudat,J. and Leung,J. (2005) The genetics of adaptive responses to drought stress: abscisic acid-dependent and abscisic acid-independent signalling components. *Physiologia Plantarum,* **123**, 111–119.

Rives,A.W. and Galitski,T. (2003) Modular organization of cellular networks. *Proceedings of the National Academy of Sciences,* **100** (3), 1128–1133.

Rocha,F., Papini-Terzi,F., Nishiyama,M., Vencio,R., Vicentini,R., Duarte,R., de Rosa,V., Vinagre,F., Barsalobres,C., Medeiros,A., Rodrigues,F., Ulian,E., Zingaretti,S., Galbiatti,J., Almeida,R., Figueira,A., Hemerly,A., Silva-Filho,M., Menossi,M. and Souza,G. (2007) Signal transduction-related responses to phytohormones and environmental challenges in sugarcane. *BMC Genomics,* **8** (1), 71.

Rodriguez-Concepcion,M., Yalovsky,S., Zik,M., Fromm,H. and Gruissem,W. (1999) The prenylation status of a novel plant calmodulin directs plasma membrane or nuclear localization of the protein. *EMBO Journal,* **18** (7), 1996–2007.

Rogers,L.A., Dubos,C., Cullis,I.F., Surman,C., Poole,M., Willment,J., Mansfield,S.D. and Campbell,M.M. (2005) Light, the circadian clock, and sugar perception in the control of lignin biosynthesis. *Journal of Experimental Botany,* **56** (416), 1651–1663.

Rogic,S., Mackworth,A.K. and Ouellette,F.B. (2001) Evaluation of Gene-Finding Programs on Mammalian Sequences. *Genome Research,* **11** (5), 817–832.

Rognoni,S., Teng,S., Arru,L., Smeekens,S. and Perata,P. (2007) Sugar effects on early seedling development in Arabidopsis. *Plant Growth Regulation,* **52** (3), 217–228.

Rogozin,I.B., Kochetov,A.V., Kondrashov,F.A., Koonin,E.V. and Milanesi,L. (2001) Presence of ATG triplet in 5' untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. *Bioinformatics,* **17**, 890–900.

Rolland,F., Baena-Gonzalez,E. and Sheen,J. (2006) Sugar sensing and signaling in plants: conserved and novel mechanisms. *Annual Revision in Plant Biology,* **57**, 675–709.

Rolland,F., Moore,B. and Sheen,J. (2002) Sugar sensing and signaling in plants. *Plant Cell,* **14**, S185–S205.

Romeis,T., Piedras,P. and Jones,J. (2000) Resistance gene-dependent activation of a calcium-dependent protein kinase in the plant defense response. *Plant Cell,* **12**, 803–816.

Rosa,V., Nogueira,F., Menossi,M., Ulian,E. and Arruda,P. (2005) Identification of methyl jasmonate-responsive genes in sugarcane using cDNA arrays. *Brazilian Journal of Plant Physiology,* **17**, 173–180.

Rual,J., Hill,D. and Vidal,M. (2004) ORFeome projects: gateway between genomics and omics. *Current Opinion in Chemical Biology,* **8**, 20–25.

Rusch,S. and Kendall,D. (1995) Protein transport via amino-terminal targeting sequences: common themes in diverse systems. *Molecular Membrane Biology,* **12** (4), 295–307.

Russinova,E., Borst,J., Kwaaitaal,M., Cano-Delgado,A., Yin,Y., Chory,J. and de Vries,S. (2004) Heterodimerization and endocytosis of Arabidopsis brassinosteroid receptors BRI1 and AtSERK3 (BAK1). *Plant Cell,* **16** (12), 3216–3229.

Saal,L., Troein,C., Vallon-Christersson,J., Gruvberger,S., Borg,A. and Peterson,C. (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biology,* **3** (8), software0003–software0003.

Sachetto-Martins,G., Franco,L. and de Oliveira,D. (2000) Plant glycine-rich proteins: a family or just proteins with a common motif? *Biochimica et Biophysica Acta,* **1492**, 1–14.

Sadiqov,S., Akbulut,M. and Ehmedov,V. (2002) Role of Ca2+ in drought stress signaling in wheat seedlings. *Biochemistry,* **67**, 491–497.

Saeys,Y., Rouze,P. and Van de Peer,Y. (2007) In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists. *Bioinformatics,* **23** (4), 414–420.

Sakuma,Y., Maruyama,K., Osakabe,Y., Qin,F., Seki,M., Shinozaki,K. and Yamaguchi-Shinozaki,K. (2006) Functional analysis of an Arabidopsis transcription factor, DREB2A, involved in drought-responsive gene expression. *Plant Cell,* **18**, 1292–1309.

Schachtman,D., Reid,R. and Ayling,S. (1998) Phosphorus Uptake by Plants: From soil to Cell. *Plant Physiology,* **116**, 447–453.

Schenk,P., Kazan,K., Wilson,I., Anderson,J., Richmond,T., Somerville,S. and Manners,J. (2000) Coordinated plant defense responses in Arabidopsis revealed by microarray analysis. *The Proceedings of the National Academy of Sciences,* **97**, 11655–11660.

Schmid,M., Uhlenhaut,N., Godard,F., Demar,M., Bressan,R., Weigel,D. and Lohmann,J.U. (2003) Dissection of floral induction pathways using global expression analysis. *Development,* **130** (24), 6001–6012.

Schmidt,E., Guzzo,F., Toonen,M. and de Vries,S. (1997) A leucine-rich repeat containing receptor-like kinase marks somatic plant cells competent to form embryos. *Development,* **124** (10), 2049–2062.

Schmitz,U. and Lonsdale,D. (1989) A Yeast Mitochondrial Presequence Functions as a Signal for Targeting to Plant-Mitochondria in vivo. *Plant Cell,* **1** (8), 783–791.

Schneider,T.D. (1997) Information content of individual genetic sequences. *Journal of Theoretical Biology,* **189**, 427–441.

Schneider,T.D. and Stephens,R.M. (1990) Sequence Logos: a new way to display consensus sequences. *Nucleic Acids Research,* **18**, 6097–6100.

Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology,* **188**, 415–431.

Schreiber,M. and Brown,C. (2002) Compensation for nucleotide bias in a genome by representation as a discrete channel with noise. *Bioinformatics,* **18**, 507–512.

Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences,* **95** (11), 5857–5864.

Scott,A., Wyatt,S., Tsou,P., Robertson,D. and Allen,N. (1999) Model system for plant cell biology: GFP imaging in living onion epidermal cells. *Biotechniques,* **26** (6), 1125–1132.

Sehnke,P.C., DeLille,J.M. and Ferl,R.J. (2002) Consummating Signal Transduction: The Role of 14-3-3 Proteins in the Completion of Signal-Induced Transitions in Protein Activity. *Plant Cell,* **14** (90001), S339–354.

Seki,M., Ishida,J., Narusaka,M., Fujita,M., Nanjo,T., Umezawa,T., Kamiya,A., Nakajima,M., Enju,A., Sakurai,T., Satou,M., Akiyama,K., Yamaguchi-Shinozaki,K., Carninci,P., Kawai,J., Hayashizaki,Y. and Shinozaki,K. (2002) Monitoring the expression pattern of around 7,000 Arabidopsis genes under ABA treatments using a full-length cDNA microarray. *Functional & Integrative Genomics,* **2**, 282–291.

Seki,M., Narusaka,M., Abe,H., Kasuga,M., Yamaguchi-Shinozaki,K., Carninci,P., Hayashizaki,Y. and Shinozaki,K. (2001) Monitoring the expression pattern of 1,300 Arabidopsis genes under drought and cold stresses by using a full-length cDNA microarray. *Plant Cell,* **13**, 61–72.

Selman-Housein,G., Lopez,M., Hernandez,D., Civardi,L., Miranda,F., Rigau,J. and Puigdomenech,P. (1999) Molecular cloning of cDNAs coding for three sugarcane enzymes involved in lignification. *Plant Science,* **143**, 163–171.

Sevilla,M., Burris,R., Gunapala,N. and Kennedy,C. (2001) Comparison of benefit to Sugar Cane plant growth an 15N2 incorporation following inoculation of sterile plants with Acetobacter diazotrophicus wild-type and Nif- mutant strains. *Molecular Plant-Microbe Interactions,* **14**, 358–366.

Shabalina,S.A., Ogurtsov,A.Y. and Spiridonov,N.A. (2006) A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Research,* **34**, 2428–2437.

Shen,H. and Chou,K. (2007) Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochemical and Biophysical Research Communications,* **355**, 1006–1011.

Shen,Y. and Burger,G. (2007) 'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools. *BMC Bioinformatics,* **8** (1), 420.

Shi,J., Kim,K., Ritz,O., Albrecht,V., Gupta,R., Harter,K., Luan,S. and Kudla,J. (1999) Novel protein kinases associated with calcineurin B-like calcium sensors in Arabidopsis. *Plant Cell,* **11**, 2393–2405.

Shiu,S. and Bleecker,A. (2001*a*) Plant receptor-like kinase gene family: diversity, function, and signaling. *Science Signaling,* **2001** (113), RE22.

Shiu,S. and Bleecker,A. (2001*b*) Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. *The Proceedings of the National Academy of Sciences,* **98**, 10763–10768.

Shiu,S., Karlowski,W., Pan,R., Tzeng,Y., Mayer,K. and Li,W. (2004) Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. *Plant Cell,* **16**, 1220–1234.

Silva,M., Chaumont,F., Leterme,S. and Boutry,M. (1996) Mitochondrial and chloroplast targeting sequences in tandem modify protein import specificity in plant organelles. *Plant Molecular Biology,* **30** (4), 769–780.

Silva-Filho,M. (2003) One ticket for multiple destinations: dual targeting of proteins to distinct subcellular locations. *Current Opinion in Plant Biology,* **6** (6), 589–595.

Silva-Filho,M., Wieers,M., Flugge,U., Chaumont,F. and Boutry,M. (1997) Different in vitro and in vivo targeting properties of the transit peptide of a chloroplast envelope inner membrane protein. *Journal of Biological Chemistry,* **272** (24), 15264–15269.

Simpson,J., Wellenreuther,R., Poustka,A., Pepperkok,R. and Wiemann,S. (2000) Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Reports,* **1** (3), 287–292.

Small,I., Peeters,N., Legeai,F. and Lurin,C. (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics,* **4** (6), 1581–1590.

Small,I., Wintz,H., Akashi,K. and Mireau,H. (1998) Two birds with one stone: genes that encode products targeted to two or more compartments. *Plant Molecular Biology,* **38** (1-2), 265–277.

Smeekens,S. (1998) Sugar regulation of gene expression in plants. *Current Opinion in Plant Biology,* **1** (3), 230–234.

Smeekens,S. (2000) Sugar-induced signal transduction in plants. *Annual Review of Plant Physiology and Plant Molecular Biology,* **51**, 49–81.

Soll,J. and Tien,R. (1998) Protein translocation into and across the chloroplastic envelope membranes. *Plant Molecular Biology,* **38** (1 - 2), 191–207.

Song,D., Li,G., Song,F. and Zheng,Z. (2007) Molecular characterization and expression analysis of OsBISERK1 , a gene encoding a leucine-rich repeat receptor-like kinase, during disease resistance responses in rice. *Molecular Biology Reports,* **May**.

Sonnhammer,E., Eddy,S., Birney,E., Bateman,A. and Durbin,R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research,* **26** (1), 320–322.

Souza,G., Simoes,A., Oliveira,K., Garay,H., Fiorini,L., Gomes,F., Nishiyama-Junior,M. and da Silva,A. (2001) SUCAST: prospecting signal transduction in sugarcane. *Genetics and Molecular Biology,* **24**, 25–34.

Stafstrom,J., Ripley,B., Devitt,M. and Drake,B. (1998) Dormancy-associated gene expression in pea axillary buds. Cloning and expression of PsDRM1 and PsDRM2. *Planta,* **205**, 547–552.

Subramani,S. (1996) Protein translocation into peroxisomes. *Journal of Biological Chemistry,* **271** (51), 32483–32486.

Swarup,R., Parry,G., Graham,N., Allen,T. and Bennett,M. (2002) Auxin cross-talk: integration of signalling pathways to control plant development. *Plant Molecular Biology,* **49**, 411–426.

Swiderski,M. and Innes,R. (2001) The Arabidopsis PBS1 resistance gene encodes a member of a novel protein kinase subfamily. *The Plant Journal,* **26**, 101–112.

Swiedrych,A., Prescha,A., Matysiak-Kata,I., Biernat,J. and Szopa,J. (2002) Repression of the 14-3-3 Gene Affects the Amino Acid and Mineral Composition of Potato Tubers. *Journal of Agriculture and Food Chemistry,* **50** (7), 2137–2141.

Szekeres,M., Nemeth,K., Koncz-Kalman,Z., Mathur,J., Kauschmann,A., Altmann,T., Redei,G., Nagy,F., Schell,J. and Koncz,C. (1996) Brassinosteroids rescue the deficiency of CYP90, a cytochrome P450, controlling cell elongation and de-etiolation in Arabidopsis. *Cell,* **85** (2), 171–182.

Tahtiharju,S. and Palva,T. (2001) Antisense inhibition of protein phosphatase 2C accelerates cold acclimation in Arabidopsis thaliana. *The Plant Journal,* **26**, 461–470.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E. and Golub,T. (1999) Interpreting patterns of gene expression with self-organizing maps. *The Proceedings of the National Academy of Sciences,* **96**, 2907–2912.

Thompson,J., Higgins,D. and Gibson,T. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research,* **22** (22), 4673–4680.

Thum,K., Shin,M., Palenchar,P., Kouranov,A. and Coruzzi,G. (2004) Genome-wide investigation of light and carbon signaling interactions in Arabidopsis. *Genome Biology,* **5** (2), R10.

Torii,K., Mitsukawa,N., Oosumi,T., Matsuura,Y., Yokoyama,R., Whittier,R. and Komeda,Y. (1996) The Arabidopsis ERECTA Gene Encodes a Putative Receptor Protein Kinase with Extracellular Leucine-Rich Repeats. *Plant Cell,* **8** (4), 735–746.

Toroser,D., Athwal,G.S. and Huber,S.C. (1998) Site-specific regulatory interaction between spinach leaf sucrose-phosphate synthase and 14-3-3 proteins. *FEBS Letter,* **435** (1), 110–114.

Toroser,D. and Huber,S.C. (1997) Protein Phosphorylation as a Mechanism for Osmotic-Stress Activation of Sucrose-Phosphate Synthase in Spinach Leaves. *Plant Physiology,* **114** (3), 947–955.

Ueda,T., Yamaguchi,M., Uchimiya,H. and Nakano,A. (2001) Ara6, a plant-unique novel type Rab GTPase, functions in the endocytic pathway of Arabidopsis thaliana. *EMBO Journal,* **20** (17), 4730–4741.

Umezawa,T., Yoshida,R., Maruyama,K., Yamaguchi-Shinozaki,K. and Shinozaki,K. (2004) SRK2C, a SNF1-related protein kinase 2, improves drought tolerance by controlling stress-responsive gene expression in Arabidopsis thaliana. *The Proceedings of the National Academy of Sciences,* **101**, 17306–17311.

Urquiaga,S., Cruz,K. and Boddey,R. (1992) Contribution of nitrogen fixation to sugar cane: Nitrogen-15 and nitrogen-balance estimates. *Soil Science Society of America Journal,* **56**, 105–114.

Vadim,I. (2001) New ways of initiating translation in eukaryotes? *Molecular and Cellular Biology,* **21**, 8238–8246.

van der Fits,L. and Memelink,J. (2000) ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. *Science,* **289**, 295–297.

van Nocker,S. and Vierstra,R. (1993) Two cDNAs from Arabidopsis thaliana encode putative RNA binding proteins containing glycine-rich domains. *Plant Molecular Biology,* **21**, 695–699.

Vargas,C., de Padua,V., Nogueira,E., Vinagre,F., Masuda,H., da Silva,F., Baldani,J., Ferreira,P. and Hemerly,A. (2003) Signaling pathways mediating the association between sugarcane and endophytic diazotrophic bacteria: a genomic approach. *Symbiosis,* **35**, 159–180.

Vartanian,N., Marcotte,L. and Giraudat,J. (1994) Drought Rhizogenesis in Arabidopsis thaliana (Differential Responses of Hormonal Mutants). *Plant Physiology,* **104**, 761–767.

Vencio,R. and Koide,T. (2005) HTself: Self-Self Based Statistical Test for Low Replication Microarray Studies. *DNA Research,* **12**, 211–214.

Verslues,P. and Zhu,J. (2005) Before and beyond ABA: upstream sensing and internal signals that determine ABA accumulation and response under abiotic stress. *Biochemical Society Transactions,* **33**, 375–379.

Vert,G., Nemhauser,J., Geldner,N., Hong,F. and Chory,J. (2005) Molecular mechanisms of steroid hormone signaling in plants. *Annual Review of Cell and Developmental Biology,* **21**, 177–201.

Vettore,A.L., da Silva,F.R., Kemper,E.L., Souza,G.M., da Silva,A.M., Ferro,M.I., Henrique-Silva,F., Giglioti,E.A., Lemos,M.V., Coutinho,L.L., Nobrega,M.P., Carrer,H., Franca,S.C., Bacci,Mauricio,J., Goldman,M.H., Gomes,S.L., Nunes,L.R., Camargo,L.E., Siqueira,W.J., Van Sluys,M.A., Thiemann,O.H., Kuramae,E.E., Santelli,R.V., Marino,C.L., Targon,M.L., Ferro,J.A., Silveira,H.C., Marini,D.C., Lemos,E.G., Monteiro-Vitorello,C.B., Tambor,J.H., Carraro,D.M., Roberto,P.G., Martins,V.G., Goldman,G.H., de Oliveira,R.C., Truffi,D., Colombo,C.A., Rossi,M., de Araujo,P.G., Sculaccio,S.A., Angella,A., Lima,M.M., de Rosa,Vicente E.,J., Siviero,F., Coscrato,V.E., Machado,M.A., Grivet,L., Di Mauro,S.M., Nobrega,F.G., Menck,C.F., Braga,M.D., Telles,G.P., Cara,F.A., Pedrosa,G., Meidanis,J. and Arruda,P. (2003) Analysis and Functional Annotation of an Expressed Sequence Tag Collection for Tropical Crop Sugarcane. *Genome Research,* **13** (12), 2725–2735.

Vicentini,R. and Menossi,M. (2007) TISs-ST: a web server to evaluate polymorphic translation initiation sites and their reflections on the secretory targets. *BMC Bioinformatics,* **8** (1), 160.

Vida,T. and Emr,S. (1995) A new vital stain for visualizing vacuolar membrane dynamics and endocytosis in yeast. *J.Cell Biol.,* **128** (5), 779–792.

Vierstra,R. (2003) The ubiquitin/26S proteasome pathway, the complex last chapter in the life of many plant proteins. *Trends in Plant Science,* **8** (3), 135–142.

Vinagre,F., Vargas,C., Schwarcz,K., Cavalcante,J., Nogueira,E., Baldani,J., Ferreira,P. and Hemerly,A. (2006) SHR5: a novel plant receptor kinase involved in plant-N2-fixing endophytic bacteria association. *Journal of Experimental Botany,* **57**, 559–569.

Vincentz,M., Cara,F., Okura,V., da Silva,F., Pedrosa,G., Hemerly,A., Capella,A., Marins,M., Ferreira,P., Franca,S., Grivet,L., Vettore,A., Kemper,E., Burnquist,W., Targon,M., Siqueira,W., Kuramae,E., Marino,C., Camargo,L., Carrer,H., Coutinho,L., Furlan,L., Lemos,M., Nunes,L., Gomes,S., Santelli,R., Goldman,M., Bacci,M.,J., Giglioti,E., Thiemann,O., Silva,F., Van Sluys,M., Nobrega,F., Arruda,P. and Menck,C. (2004) Evaluation

of monocot and eudicot divergence using the sugarcane transcriptome. *Plant Physiology,* **134** (3), 951–959.

von Heijne,G., Steppuhn,J. and Herrmann,R. (1989) Domain structure of mitochondrial and chloroplast targeting peptides. *European Journal of Biochemistry,* **180** (3), 535–545.

von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Kruger,B., Snel,B. and Bork,P. (2007) STRING 7 – recent developments in the integration and prediction of protein interactions. *Nucleic Acids Research,* **35**, D358–D362.

Wang,X.Q. and Rothnagel,J.A. (2004) 5'-Untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. *Nucleic Acids Research,* **32**, 1382–1391.

Watanabe,N., Che,F.S., Iwano,M., Takayama,S., Yoshida,S. and Isogai,A. (2001) Dual targeting of spinach protoporphyrinogen oxidase II to mitochondria and chloroplasts by alternative use of two in-frame initiation codons. *Journal of Biological Chemistry,* **276**, 20474–20481.

Watson,B.S., Asirvatham,V.S., Wang,L. and Sumner,L.W. (2003) Mapping the Proteome of Barrel Medic (Medicago truncatula). *Plant Physiology,* **131** (3), 1104–1123.

Weaver,C. and Roberts,D. (1992) Determination of the site of phosphorylation of nodulin 26 by the calcium-dependent protein kinase from soybean nodules. *Biochemistry,* **31**, 8954–8959.

Webb,K., Skot,L., Nicholson,M., Jorgensen,B. and Mizen,S. (2000) Mesorhizobium loti increases root-specific expression of a calcium-binding protein homologue identified by promoter tagging in Lotus japonicus. *Molecular Plant-Microbe Interactions,* **13**, 606–616.

Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database Resources of the National Center for Biotechnology. *Nucleic Acids Research,* **31**, 28–33.

White,P. (2001) A Nod and a wave: calcium signals during nodulation. *Trends in Plant Science,* **6**, 141.

Wimmer,B., Lottspeich,F., vanderKlei,I., Veenhuis,M. and Gietl,C. (1997) The glyoxysomal and plastid molecular chaperones (70-kDa heat shock protein) of watermelon cotyledons are encoded by a single gene. *Proceedings of the National Academy of Sciences,* **94** (25), 13624–13629.

Wissenbach,M., Uberlacker,B., Vogt,F., Becker,D., Salamini,F. and Rohde,W. (1993) Myb genes from Hordeum vulgare: tissue-specific expression of chimeric Myb promoter/Gus genes in transgenic tobacco. *The Plant Journal,* **4**, 411–422.

Woodward,A. and Bartel,B. (2005) Auxin: regulation, action, and interaction. *Annals of Botany,* **95**, 707–735.

Wu,P., Ma,L., Hou,X., Wang,M., Wu,Y., Liu,F. and Deng,X. (2003) Phosphate Starvation Triggers Distinct Alterations of Genome Expression in Arabidopsis Roots and Leaves. *Plant Physiology,* **132**, 1260–1271.

Xie,D., Li,A., Wang,M., Fan,Z. and Feng,H. (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Research,* **33**, W105–W110.

Xiong,L., Schumaker,K. and Zhu,J. (2002) Cell signaling during cold, drought, and salt stress. *Plant Cell,* **14**, s165–s183.

Xiong,W., Hsieh,C.C., Kurtz,A.J., Rabek,J.P. and Papaconstantinou,J. (2001) Regulation of CCAAT/enhancer-binding protein-beta isoform synthesis by alternative translational initiation at multiple AUG start sites. *Nucleic Acids Research,* **29** (14), 3087–3098.

Yamada,S., Komori,T., Hashimoto,A., Kuwata,S., Imaseki,H. and Kubo,T. (2000) Differential expression of plastidic aldolase genes in Nicotiana plants under salt stress. *Plant Science,* **154**, 61–69.

Yang,G. and Komatsu,S. (2004) Molecular cloning and characterization of a novel brassinolide enhanced gene OsBLE1 in Oryza sativa seedlings. *Plant Physiology and Biochemistry,* **42**, 1–6.

Yang,X., Lee,S., So,J.h., Dharmasiri,S., Dharmasiri,N., Ge,L., Jensen,C., Hangarter,R., Hobbie,L. and Estelle,M. (2004) The IAA1 protein is encoded by AXR5 and is a substrate of SCFTIR1. *The Plant Journal,* **40** (5), 772–782.

Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research,* **30** (4), e15.

Yoshida,R., Hobo,T., Ichimura,K., Mizoguchi,T., Takahashi,F., Aronso,J., Ecker,J. and Shinozaki,K. (2002) ABA-activated SnRK2 protein kinase is required for dehydration stress signaling in Arabidopsis. *Plant Cell Physiology,* **43**, 1473–1483.

Yu,S. (1999) Cellular and genetic responses of plants to sugar starvation. *Plant Physiology,* **121** (3), 687–693.

Zhang,B., Pan,X., Cobb,G. and Anderson,T. (2006) Plant microRNA: a small regulatory molecule with big impact. *Developmental Biology,* **289**, 3–16.

Zhang,M., Barg,R., Yin,M., Gueta-Dahan,Y., Leikin-Frenkel,A., Salts,Y., Shabtai,S. and Ben-Hayyim,G. (2005) Modulated fatty acid desaturation via overexpression of two distinct omega-3 desaturases differentially alters tolerance to various abiotic stresses in transgenic tobacco cells and plants. *The Plant Journal,* **44**, 361–371.

Zhao,D.Z., Wang,G.F., Speal,B. and Ma,H. (2002) The EXCESS MICROSPOROCYTES1 gene encodes a putative leucine-rich repeat receptor protein kinase that controls somatic and reproductive cell fates in the Arabidopsis anther. *Genes & Development,* **16** (15), 2021–2031.

Zheng,Z., Nafisi,M., Tam,A., Li,H., Crowell,D., Chary,S., Schroeder,J., Shen,J. and Yang,Z. (2002) Plasma membrane-associated ROP10 small GTPase is a specific negative regulator of abscisic acid responses in Arabidopsis. *Plant Cell,* **14**, 2787–2797.

Índice de Autores