UNIVERSIDADE DE SÃO PAULO

Systems Integration Tool: uma ferramenta para integração
e visualização de dados em larga escala e sua aplicação
em cana-de-açúcar

Amanda Rusiska Piovezani

Tese Apresentada ao
Programa Interunidades de Pós-Graduação em Bioinformática da
Universidade de São Paulo para
Obtenção do Título de Doutora em Ciências
Área de Concentração: Bioinformática

Orientador: Prof. Dr. Marcos Silveira Buckeridge (LAFIECO – IB/USP)
Co-orientador: Prof. Dr. Fabrício Martins Lopes (Universidade Tecnológica
Federal do Paraná – UTFPR)

Durante a elaboração deste trabalho, a autora recebeu auxílio financeiro da
FAPESP-Microsoft Research

– São Paulo, SP – 2017 –

Systems Integration Tool: uma ferramenta para integração e visualização de dados em larga escala e sua aplicação em cana-de-açúcar

Amanda Rusiska Piovezani

Esta é a versão original e corrigida da tese elaborada pela candidata Amanda Rusiska Piovezani, tal como submetida à Comissão Julgadora como requisito final para a obtenção do título de Doutora em Ciências.

Área de Concentração: Bioinformática.

Orientador: Prof. Dr. Marcos Silveira Buckeridge

DEDICATÓRIA

À mulher da minha vida Gesiele, pelo amor incondicional e todos os conselhos e momentos de questionamentos sobre a vida.

À nossa filha Fryda, que chegou para nos ensinar um jeito diferente e intenso de amar.

À minha mãe Luisa, pelo simples e completo fato de ser quem você é, me ensinando dignamente a importância da família, o caminho da honestidade e persistência.

AGRADECIMENTOS

Quando iniciei o doutorado, não imaginava no que ele me transformaria. E não digo isso pelo título, mas pela transformação da alma que me causou. Parte do cérebro só é aberta com o caminho percorrido, não há como pular etapas. Isso foi possível e eu sou muito grata por isso, primeiramente a Deus, pela saúde física e mental, mas também graças a muitas pessoas.

Às minhas amadas esposa Gesiele e filha Fryda, que me apoiaram de forma muito próxima e intensa nesta busca. Ge você sabe que foi minha mentora nessa tese. Agradeço a minha mãe Luisa, irmão Alberto por serem parte de tudo o que me tornei, obrigada por cada palavra e ato de amor e incentivo. Agradeço ainda à todos os outros familiares que sempre me apoiaram nessa busca.

Em agradecimento à USP eu gostaria de começar agradecendo ao Prof. Marcos Buckeridge, orientador de fato, que com sua maneira holística de pensar, me ensinou o que é a ciência, a compreender que o sistema é muito mais complexo do que conseguimos ver, mas devemos considerar isso e buscar entender como as partes formam o todo e para todos os sistemas. Sou muito grata por cada longa conversa e café. Agradeço também ao meu co-orientador Fabrício Lopes pelas conversas e direcionamentos, e a comissão julgadora que compõe a banca, por contribuírem no aprimoramento desta tese.

É claro que não apenas o orientador contribuiu para o que chamamos de final da tese, agradeço aos amigos do laboratório Lafieco - IB-USP, em especial à Adriana Grandis, Eveline Tavares, Amanda P. de Souza, Vinícius Carvalho, Arthur Cambler, e todos outros, além da Viviane Costa e Eglee Igarashi pelo suporte técnico em tudo, sem vocês as coisas não andariam.

Há ainda a equipe da Vision do Instituto de Matemática e Estatística da USP, sou imensamente grata à estrutura computacional disponibilizada para os pesados processamentos de dados. Agradeço à Fapesp (Processo nº 2013/24205-0), à Microsoft Research e ao Instituto Nacional de Ciência e Tecnologia do Bioetanol (INCT do Bioetanol) pelo financiamento do estudo. À secretaria da pós-graduação da Bioinformática da USP, por serem sempre tão solícitos.

Agora é isso aí, a outra jornada começa agora. Sou grata a tudo e todos que passaram por mim até hoje.

EPÍGRAFE

“A tarefa não é tanto ver aquilo que ninguém viu,
mas pensar o que ninguém ainda pensou
sobre aquilo que todo mundo vê.”
(Arthur Schopenhauer)

Piovezani, Amanda Rusiska. *Systems Integration Tool*: uma ferramenta para integração e visualização de dados em larga escala e sua aplicação em cana-de-açúcar. 2017. Tese (Doutorado em Bioinformática) – Universidade de São Paulo, São Paulo. Brasil.

RESUMO

As respostas das plantas ao ambiente são orquestradas por fatores genéticos, bem como sua flexibilidade metabólica, uma vez que essas são sésseis. Respostas ao ambiente são regidas por fatores genéticos, bem como sua flexibilidade metabólica, uma vez que essas são sésseis. A forma com que os padrões gênicos e metabólicos redundam entre as células, refletem nos diferentes níveis organizacionais (célula, tecido, órgão e até o organismo como um todo). Por isso, para entendermos as respostas das plantas em determinados estágios de desenvolvimento ou condições é importante explorarmos ao máximo os diferentes níveis de regulação. Neste sentido, tem crescido a quantidade de dados biológicos obtidos através de métodos que produzem dados em larga escala, visando um estudo de forma sistêmica. Embora existam várias ferramentas para a integração de dados biológicos, elas estão desenvolvidas para organismos modelos, inviabilizando análises para outros, como a cana-de-açúcar, que possui vários dados biológicos disponíveis, mas com genoma complexo e incompleto. Tendo em vista a importância econômica da cana-de-açúcar e o interesse em entendermos o processo de degradação da parede celular, desenvolvemos a ferramenta SIT (*Systems Integration Tool*), para integração dos dados disponíveis (transcritoma, proteoma e atividade enzimática). A implementação da ferramenta foi realizada utilizando as linguagens de programação Perl e Java. SIT possui uma interface gráfica, podendo ser executada localmente, a qual possibilita a integração de até seis diferentes conjuntos de dados. A visualização do resultado é obtida na forma de redes complexas, permitindo ao usuário a visualização e edição dinâmica da integração. Uma aplicação de SIT utilizando transcritômica, proteômica e fisiologia de regiões da raiz de cana-de-açúcar que compreende o desenvolvimento do aerênquima, foi executada e permitiu entre outros, a identificação de elementos chave na degradação da parede celular, presentes nos diferentes conjuntos de dados explorados. Apontando portanto, potenciais alvos de estudos experimentais. SIT pode ser aplicada à diferentes conjuntos de dados, a qual poderá auxiliar em estudos futuros em várias áreas do conhecimento.

Palavras-chave: biologia de sistemas, integração de dados, cana-de-açúcar, aerênquima, parede celular, redes complexas.

Piovezani, Amanda Rusiska. *Systems Integration Tool: an integration and visualization tool for big data and their application on sugarcane*. 2017. These (Philosophy Doctor: Bioinformatics) – University of São Paulo, São Paulo. Brazil.

ABSTRACT

Plant are sessile organisms, and their responses to environmental stimuli are orchestrated by genetic factors, as well as by their metabolic flexibility. Inside the cell, there are genetic and metabolic patterns responsible for cell redundancy, and that reflects on different organizational levels (cell, tissue, organ, until a whole organism). Thus, to understand plant responses to certain conditions, it is important to understanding different regulatory levels. Recently, there was a large increase in availability of biological data. This happened due to the advance in next-generation sequencing techniques, which now enables more profound system biology studies. Despite the availability of several integration tools for analysis of biological data, these were developed for organism modeling. However, such tools are partially effective for sugarcane, for which there are large amounts of data, but has incomplete genome data. Due to the economic importance of sugarcane and aiming at understanding cell wall degradation process, we develop the software Systems Integration Tool (SIT). The tool integrates available data (transcriptomics, proteomics, and enzymatic activity). The implementation was performed in Perl and Java. SIT has a graphical interface, standalone execution, enabling integration until six layers of data. Integration results are generated as complex networks, allowing the users to visualize and dynamically edit the networks. The present study allowed the identification of key cell wall regulatory elements present on different data sets pointing out to potential targets for experimental validation. SIT can be applied to various data sets being capable of helping future studies in different areas of knowledge.

Keywords: system biology, big data integration, sugarcane, aerenchyma, cell wall, biological networks.

Sumário

1 Introdução.....	1
2 Revisão Bibliográfica.....	3
2.1 Cana-de-açúcar e sua importância econômica.....	3
2.1.1 Parede celular de cana-de-açúcar.....	8
2.1.1 Aerênquima em raízes de cana-de-açúcar.....	13
2.2 Níveis de regulação biológica.....	17
2.2.1 Transcritômica.....	17
2.2.2 Proteômica.....	21
2.2.3 Atividade Enzimática.....	23
2.3 Redes gênicas.....	26
2.3.1 Centralidades das redes de co-expressão.....	33
2.4 Biologia de sistemas e a integração de dados.....	36
3 Objetivos.....	43
3.1 Objetivo geral.....	43
3.2 Objetivos específicos.....	43
4 Metodologia.....	44
4.1 Desenvolvimento da ferramenta SIT.....	44
4.2 Preparação dos dados biológicos.....	46
4.2.1 Processamento do transcrito de segmentos de raízes de cana-de-açúcar.....	46
4.2.1.1 Preparação das reads.....	47
4.2.1.2 Estratégias para a quantificação dos transcritos.....	48
4.2.1.3 Análise de expressão diferencial.....	49
4.2.2 Anotação de transcritos de biossíntese e degradação da parede celular.....	50
4.3 Aplicação da ferramenta SIT em dados de aerênquima em raízes de cana-de-açúcar.....	53
4.4 Microtomografia de segmentos da raiz de cana-de-açúcar.....	56
4.5 Desenvolvimento de um banco de dados.....	56
5 Resultados.....	57
5.1 Ferramenta SIT para integração e visualização de dados.....	57
5.2 Transcritoma de segmentos de raízes de cana-de-açúcar.....	66
5.2.1 Processamento das reads do RNA-Seq.....	66
5.2.2 RSEM para a quantificação da expressão gênica.....	68
5.2.3 Características gerais observadas no RNA-Seq.....	69
5.2.4 Transcritos de parede celular observados no transcrito.....	70
5.2.5 Domínios de biossíntese, degradação e expansinas da parede celular em cana-de-açúcar.....	73
5.2.6 Domínios funcionais de transcritos diferencialmente expressos.....	75
5.3 Aplicação da ferramenta SIT em dados de cana-de-açúcar.....	77
5.3.1 Quantificação geral da integração de dados.....	77
5.3.2 Propriedade das redes de co-expressão integradas.....	79
5.3.3 Análise de enzimas de degradação da parede celular.....	87
5.3.4 Perfil de potenciais elementos efetores no desenvolvimento do aerênquima em raízes de cana-de-açúcar.....	89
5.4 Visualização microtomográfica de segmentos da raiz de cana-de-açúcar.....	97
6 Discussão.....	98

6.1 Biologia de Sistemas.....	98
6.2 <i>Systems Integration Tool</i> , uma nova ferramenta de integração de dados.....	99
6.3 A aplicação do SIT no processo de formação do aerênquima de raiz de cana-de-açúcar.....	102
6.4 Aplicabilidade de SIT.....	111
7 Conclusões e perspectivas futuras.....	114
Referências.....	116
Apêndice.....	134

1 Introdução

Por serem sésseis, as plantas não têm capacidade de buscar áreas com melhores condições para sua adaptação. Por este motivo, elas respondem ao ambiente através de diversos mecanismos que foram selecionados ao longo de sua evolução para manter a homeostase do indivíduo, mesmo em condições adversas. Tais mecanismos são dependentes de respostas sistêmicas que conectam o nível de expressão gênica às respostas fisiológicas.

O funcionamento sistêmico das plantas é muito complexo, com diversas rotas metabólicas altamente controladas e que atuam conectadas. Neste sentido, entender a dinâmica de integração dos vários níveis de regulação biológica traz luz ao entendimento sobre o funcionamento genético que envolve cada processo fisiológico.

Para organismos modelo, já existem alguns recursos computacionais envolvendo a abordagem sistêmica, que avaliam o comportamento da planta como um todo e que essencialmente correspondem às plataformas de exploração de dados biológicos disponíveis em bancos de dados públicos.

Embora possam contribuir para um estudo sistêmico, tais recursos estão restritos apenas à aplicação em organismos de genoma completo e apenas para dados disponíveis em plataformas públicas, ou seja, não disponíveis para dados locais, como é o caso do conjunto de dados utilizados neste estudo, que são provenientes de cana-de-açúcar, um organismo de genoma complexo e incompleto. Além disso, não apresentam um modelo capaz de integrar dados de vários níveis de regulação biológica e que tenha capacidade para gerar novas interações.

Por esse motivo o enfoque principal desse estudo foi desenvolver uma ferramenta que permita a visão integrada dos diferentes níveis de regulação da

planta. Para tanto, utilizou-se a raiz de cana-de-açúcar como alvo, tendo o processo de formação do aerênquima que ocorre em células do seu córtex, como sistema modelo. A abordagem técnica desse sistema em estudo, utiliza conjuntos de dados referentes à transcritômica, proteômica e fisiologia obtidos em estudos anteriores e que possuem a necessidade de serem explorados de forma integrada para um entendimento sistêmico do processo.

2 Revisão Bibliográfica

2.1 Cana-de-açúcar e sua importância econômica

A cana-de-açúcar (Figura 1-a) é uma gramínea que pertence à ordem *Poales*, família *Poaceae* e gênero *Saccharum* (TZVELEV, 1989). Este gênero é composto por seis espécies, *S. officinarum*, *S. spontaneum*, *S. robustum*, *S. barberi*, *S. sinense* e *S. edule*.

As variedades comercializadas atualmente são conhecidas como *Saccharum spp.* e se caracterizam por serem resultados de gerações avançadas de híbridos de duas das espécies do gênero, *S. officinarum* e *S. spontaneum*, obtidas por hibridização e sucessivos retrocruzamentos (DILLON et al., 2007).

Uma importante característica do genoma da cana-de-açúcar é sua poliploidia (SREENIVASAN; AHLOOWALIA; HEINZ, 1987), que é a presença na célula, de mais de dois conjuntos de cromossomos homólogos, ou seja, há múltiplas cópias do genoma. É comum nas variedades modernas a ocorrência de aneuploidia (variação no número de cromossomos, podendo ocorrer redução ou aumento da quantidade) (GRIVET L.; ARRUDA, 2002). Nessas variedades, a maioria dos cromossomos vem de *S. officinarum*, sendo aproximadamente entre 10 à 25% herdado de *S. spontaneum* (D'HONT et al., 1996).

No final da década de 1990, um grupo de pesquisadores produziu, pela primeira vez, um grande número de sequências de DNA complementares aos RNAs mensageiros expressos em vários tecidos da cana-de-açúcar. Este projeto, denominado **SUCST**, sequenciou etiquetas de sequência expressas do genoma (do inglês, *Expressed Sequence Tags* - ESTs) de cana-de-açúcar a partir de 26

bibliotecas (vários tecidos e combinações experimentais) e 13 diferentes cultivares, onde ao todo foram identificados 43.141 agrupamentos (do inglês, *clusters*) de ESTs de cana-de-açúcar (VETTORE et al., 2001). Posteriormente, nosso grupo utilizou o banco de dados do SUCEST para analisar os padrões de expressão de genes de parede celular em todas as bibliotecas de tecidos da cana-de-açúcar (LIMA et al., 2001), encontrando um total de 459 agrupamentos (potenciais genes) relacionados à síntese e degradação de polímeros da parede celular.

Figura 1: Aspecto geral (a) e anatômico (b e c) de uma planta de cana-de-açúcar. (a) Indicação de uma raiz, órgão utilizado para os cortes anatômicos e visualização da região do córtex. b) destaque de uma célula viva no córtex delimitada pela parede celular, com destaque ao córtex sem aerênquima. c) presença de espaços de ar (aerênquima cortical) nessa região apontada pela seta.

A cana-de-açúcar figura entre as principais culturas da economia brasileira, sendo o Brasil o maior produtor do mundo. Dos produtos gerados a partir de sua

matéria prima, o país ocupa atualmente a primeira posição mundial na produção de açúcar e compete com os Estados Unidos da América pela primeira posição na produção de etanol, embora os Estados Unidos extraiam esse produto do milho e não da cana-de-açúcar como no Brasil (OECD/FOOD; OF THE UNITED NATIONS, 2015).

Especialmente na geração de combustíveis, a previsão é de que haverá um aumento na demanda energética para os próximos anos e este chegará até mais de 50% até 2025. A principal causa seria a demanda emergente gerada por nações de rápido desenvolvimento (RAGAUSKAS et al., 2006).

Neste sentido, o etanol proveniente da cana-de-açúcar (bioetanol/bicombustível) tem se mostrado uma ótima alternativa de combustível comparado às fontes petroquímicas, por ser uma fonte renovável, de fácil acesso e com liberação de baixo teor de carbono (DE SOUZA et al., 2014).

A produção de combustíveis a partir de carboidratos vegetais depende da energia solar armazenada na biomassa na forma de açúcares solúveis, amido e polissacarídeos estruturais, por meio da fotossíntese. A forma mais utilizada atualmente para extração de bioetanol da cana-de-açúcar é por meio do caldo. O caldo é processado para extração de açúcares solúveis através da sua fermentação. Contudo, apenas um terço da energia total é extraída dessa forma, o restante do conteúdo energético está presente na forma de carboidratos, armazenados como polissacarídeos estruturais e localizados na parede celular (Figura 1-b) da biomassa resultante da produção de etanol de primeira geração (1G). Essa biomassa é chamada de palhiço e bagaço e é composta principalmente por parte das folhas e colmo de cana-de-açúcar, respectivamente (SOCCOL et al., 2010).

Apesar do potencial energético da biomassa restante da cana-de-açúcar, o principal destino deste material ainda é como descarte agrícola ou queima para fornecimento de energia para usinas de açúcar e etanol, uma alternativa poluente e

ineficiente na questão de transformar a energia disponível a partir da biomassa (DE CARVALHO PIRES, 2009; DIAS et al., 2012).

A principal questão na produção do bioetanol a partir da biomassa é o acesso aos componentes da parede celular, que pode ser obtido por exemplo, por meio de um fracionamento químico e sua conversão em açúcares fermentáveis. Esse processo é essencial para que a matéria prima renovável da cana-de-açúcar possa ser utilizada na produção de bioetanol, conhecido também como etanol de segunda geração (2G) (BUCKERIDGE; DOS SANTOS; DE SOUZA, 2010; HIMMEL M.E et al., 2007; MOSIER et al., 2005; SOCCOL et al., 2010). A sua produção tem sido possível através de algumas etapas: o pré-tratamento, a hidrólise enzimática e a fermentação.

Existem diversos pré-tratamentos e com diferentes rendimentos. De forma geral podem ser classificados em métodos físicos (ex. processo de moagem), químicos (ex. tratamentos alcalinos e oxidativos), físico-químicos (ex. explosão a vapor) e biológicos (ex. pré-tratamento por microrganismos, especialmente fungos) (SOCCOL et al., 2010).

Para cada tipo de pré-processamento existem certas desvantagens que limitam a sua aplicação. Elas variam quanto ao tempo de aplicação, custo operacional (ex. equipamentos de pressão e reagentes químicos), produção de elementos inibidores da fermentação (ex. H_2SO_4 ou SO_2 em tratamentos físico-químicos), entre outros. Neste sentido, métodos combinados têm sido considerados promissores no aumento da eficiência da produção de açúcares, reduzindo a formação de inibidores e reduzindo o tempo de processamento (SOCCOL et al., 2010).

A biomassa da cana-de-açúcar, por se caracterizar um recurso acessível e de baixo custo para sua obtenção, tem atraído popularidade como uma fonte de açúcar fermentável para a produção de etanol 2G. Contudo, uma das etapas com maior

custo embutido neste processo é o pré-tratamento seguido pelo tratamento enzimático. Este, ainda hoje é o fator limitante para a produção em escala industrial, de bioetanol 2G (DE SOUZA et al., 2014; TAVARES; DE SOUZA; BUCKERIDGE, 2015).

Nesta perspectiva, pesquisas têm sido realizadas no sentido de melhorar a obtenção de carboidratos fermentáveis, reduzir os inibidores que são produzidos a partir da degradação do açúcar durante o processo de pré-tratamento, diminuir a utilização de produtos químicos e o consumo de energia, produzir subprodutos com maior valor agregado e reduzir o custo do processo de bioetanol.

Por outro lado, recentes avanços na área de ômicas, como a genômica funcional, metagenômica, engenharia genética e metabólica, têm mostrado avanços que levam à produção de plantas desenhadas artificialmente e especialmente voltadas à produção de bioetanol, as chamadas “plantas de etanol de 2G”. Elas podem conter ambos, teores elevados de celulose e capacidade de produção de enzimas hidrolases para a sua degradação.

O Brasil tem, pelo menos, duas empresas que já desenvolveram essa planta de etanol 2G e que estão sendo utilizadas em escala industrial (GRANBIO, 2014; RAÍZEN, 2014). Todavia, o custo final para a produção de bioetanol 2G ainda é considerado alto em relação às atuais tecnologias e não competitivo com o etanol 1G. Tal fato requer maiores investimentos em pesquisa, visando principalmente a degradação do material lignocelulósico (biomassa) (DE SOUZA et al., 2014).

2.1.1 Parede celular de cana-de-açúcar

A parede celular nas plantas é uma elaborada matriz extracelular, que

circunda cada célula da planta, fornecendo proteção e resistência (recalcitrância) à hidrólise enzimática. A parede de células vizinhas que são cimentadas juntas para formar a planta inteira, são geralmente mais espessas, mais fortes e mais rígidas do que a matriz extracelular produzida por células animais (BUCKERIDGE et al., 2015; BUCKERIDGE; CAVALARI; SILVA, 2008).

A organização e composição da parede celular depende do tipo de célula. Todas as novas células têm origem na etapa de divisão celular e em regiões chamadas meristemáticas. Em plantas, geralmente as células são pequenas comparadas ao tamanho final. Para acomodar o crescimento subsequente dessas células, cada parede celular é chamada “parede celular primária”, que é fina e extensível, embora resistente. A parede primária está presente por exemplo, em raízes de cana-de-açúcar como células parenquimáticas do córtex. Para algumas estruturas da planta, depois de cessado o seu crescimento a parede não precisa mais da sua característica extensível. As vezes a parede primária é retida sem maiores modificações (como acontece no córtex da raiz de cana-de-açúcar), mas geralmente ocorre a formação de uma rígida parede celular secundária, que é produzida pela deposição de novas camadas dentro da primeira. Essa segunda pode apresentar uma composição similar à primeira, mas pode também ser completamente diferente. O principal polímero adicional da parede celular secundária é a lignina, que pode ser entendida como uma complexa rede de compostos fenólicos, encontrados nas paredes de vasos como o xilema (ALBERTS et al., 2002).

Diferente do milho, que é uma das principais culturas utilizadas como fonte de bioenergia e cuja estrutura da parede celular já é profundamente estudada há anos (CARPITA, 1984, 1996), em gramíneas como a cana-de-açúcar, apenas recentemente, a composição e arquitetura da parede celular foi profundamente estudada e associada ao processo de degradação da mesma na produção de etanol

2G (DE SOUZA et al., 2013; LEITE et al., 2017). A lacuna no conhecimento sobre a parede celular de cana-de-açúcar ocorreu devido ao fato de os esforços dessa área terem sido realizados predominantemente na busca por enzimas para a degradação da parede celular (BUCKERIDGE et al., 2015).

A parede de cana-de-açúcar pode ser didaticamente e hipoteticamente entendida como o arranjo de polímeros de hemicelulose, celulose, pectina e lignina (BUCKERIDGE et al., 2017a). A Figura 2-A mostra a sua composição e arquitetura. Nessa estrutura, existem sete microfibrilas de celulose, que estão ligadas à hemicelulose, tendo sido descritas como moléculas de Arabinosilano esterificado com ácidos ferúlico e p-cumárico. Uma outra camada que circunda estas primeiras é ainda de hemicelulose, com moléculas de Arabinosilano, Xiloglucano e β -glucano. As hemicelulose estão embebidas em pectina e o pouco que é observado de lignina e fenilpropanóides, foram observados envolvendo os polissacarídeos. As pectinas garantem a porosidade e a flexibilidade da parede celular. Essa seria a unidade repetitiva da parede celular de cana-de-açúcar. Na Figura 2-B,C, tem-se um esquema do que seria a porção tridimensional dessa parede.

Do ponto de vista de bioenergia, o primeiro e o mais utilizado nível de organização em pesquisas é a composição de monossacarídeos da parede celular, que são principalmente a celulose, hemicelulose e pectina. Análises sobre a composição de monossacarídeos são realizadas para o material de biomassa intacta e também sobre o material que é liberado da etapa de hidrólise. Contudo, os resultados geralmente focam em padrões de ação das enzimas voltadas para a hidrólise de celulose (GOMEZ et al., 2010; WALTON; BANERJEE; CAR, 2011). Essa

Figura 2: Estrutura didática e hipotética para a parede celular de cana-de-açúcar. (A) Composição e arquitetura, mostrando unidade repetitiva da parede celular de cana-de-açúcar. Existem sete microfibrilas de celulose ligadas à hemicelulose, tendo sido descritas como moléculas de Arabinoxilano esterificado com ácidos ferúlico e p-cumárico e ainda o (Gluco)manano. Uma terceira camada que circunda estas primeiras é ainda de hemicelulose, com moléculas de Arabinoxilano, Xiloglucano e β -glucano. Até aqui, tudo está embebido em moléculas de fenilpropanóides e/ou lignina. A terceira e quarta camadas são de pectina, que circunda as anteriores e garante a porosidade da parede celular. (B,C) Esquema representativo do que seria a porção tridimensional dessa parede. Este modelo ainda não inclui proteínas da parede celular, como as ricas em hidroxiprolinas, que por sua vez ainda não são bem compreendidas em cana-de-açúcar. Figura adaptada (BUCKERIDGE et al., 2017a).

seria uma limitação do método, que por estar focado na ação de celulases, acaba analisando apenas uma parte da parede celular, já que em cana-de-açúcar, os domínios de hemicelulose (que compõe grande parte de pentoses que não são facilmente fermentáveis) e pectinas são quantitativamente relevantes (DE SOUZA et al., 2013; LEITE et al., 2017).

Ainda relacionado a essa organização, uma prática comum em pesquisas com bioenergia mas que deve ser olhada com cautela, é a inferência de polissacarídeos a partir da quantificação de monossacarídeos. Isso porque, muitos monossacarídeos são comuns de vários polissacarídeos complexos. Por exemplo nas gramíneas, a glucose é encontrada na celulose, mas também é encontrada em ligações mistas de β -glucano e xiloglucano. Tem também a arabinose, que é encontrada em arabinoxilano, mas também está presente nas pectinas.

O segundo nível de complexidade da parede celular está relacionada à montagem de polímeros na parede e é chamado de arquitetura da parede celular. Neste caso, o modelo mais aceito para a parede celular defende a presença de domínios independentes (MCCANN; ROBERTS, 1991).

Posteriormente, foi proposta a existência de tipos diferentes de parede celular, dependendo da hemicelulose e da proporção de pectinas encontrada, o tipo I sendo uma arquitetura rica em xiloglucano e pectinas e o tipo II sendo rica em arabinoxilano e com menor quantidade de pectinas (CARPITA; GIBEAUT, 1993). Seguindo essa linha, foi demonstrada a existência do tipo III de parede celular, como sendo uma estrutura rica em mananos (SILVA et al., 2011).

Recentemente, foi determinado que a parede celular de cana-de-açúcar apresenta uma arquitetura do tipo II, com relativamente pouca matriz de pectina e tendo arabinoxilano > β -glucano > xiloglucano > manano como as principais hemiceluloses (DE SOUZA et al., 2013). Isso considerando que em cana-de-açúcar a parede é do tipo primária.

O terceiro nível de organização da parede celular refere-se a combinação de domínios da parede, para os diferentes tecidos da planta. Especialmente em cana-de-açúcar, a composição dos diferentes domínios tem se mostrado uniforme entre os órgãos. As folhas, o colmo e a raiz são compostos de celulose, hemiceluloses (arabinoxilano, ligações mistas de β -glucano, xiloglucano e traços de manano) e pectinas (homogalacturonano, ramnogalacturonano I ramificado com curtas cadeias de 1,3-1,6- β -galactano e 1,5- α -arabinano) (DE SOUZA et al., 2013; LEITE et al., 2017). Já a distribuição desses domínios, está arranjada de forma diferente entre os tecidos da cana-de-açúcar.

Com foco na bioenergia, a arquitetura da parede celular de cana-de-açúcar é o ponto chave para a produção de etanol 2G, pois é justamente a composição e a forma aos quais os domínios estão arranjados, que pode interferir no processo de hidrólise e consequentemente permitir um ataque mais eficiente à parede celular. O modelo proposto (BUCKERIDGE et al., 2015), aponta três principais limitações envolvendo a etapa de hidrólise da parede celular: a porosidade da parede, o código glicômico das hemiceluloses e a desmontagem da estrutura microfibrilar. A porosidade pode ser entendida de forma geral, como a acessibilidade com que as enzimas atravessam a estrutura da matriz péctica da parede celular a fim de acessar as hemiceluloses e celulose.

O código glicômico recentemente proposto (BUCKERIDGE; DE SOUZA, 2014), está relacionado ao domínio das hemiceluloses, mostrando que a resistência da cana-de-açúcar à hidrólise, parece ser dada pela combinação de interações polímero-polímero, com os padrões de ramificação das cadeias principais de polissacarídeos. A estrutura microfibrilar mencionada se refere a ligação entre as microfibras de celulose e hemicelulose. A desmontagem dessa estrutura, bem como os dois aspectos mencionados anteriormente, afetam a recalcitrância da parede celular.

A recalcitrância da parede celular em cana-de-açúcar é uma barreira na produção de tecnologias para a produção de etanol 2G, pois dificulta o acesso das enzimas aos polímeros da parede. Reduzir a recalcitrância da parede celular, é portanto, um caminho para aumentar a produção de etanol 2G sem a necessidade de grande esforços nos pré-tratamentos químicos e físicos.

Neste sentido, uma estratégia natural e portanto, desempenhada pela própria planta, como o desenvolvimento do aerênquima, pode ser um caminho para elucidar os mecanismos e elementos envolvidos no processo de degradação da parede celular das plantas, e com isso auxiliar no avanço de tecnologias do processo de produção de etanol.

2.1.1 Aerênquima em raízes de cana-de-açúcar

Aerênquima é o termo dado a um tecido vegetal contendo um grande espaço de ar, com dimensões que excedem o tamanho normalmente encontrado em espaços intracelulares (EVANS, 2003). Ele pode se formar em folhas, pecíolos, caules e raízes e ocorrer de forma constitutiva pelo desenvolvimento normal da planta (ex. como é o caso da cana-de-açúcar (GILBERT et al., 2007) ou de forma induzida sob condições de estresse. Este último é geralmente observado em condições de hipóxia em regiões alagadas (ex. a formação do aerênquima em milho pode contribuir para a anulação da anoxia sob condições de alagamento), mas também pode ocorrer sob condições como alta temperatura, deficiência nutricional e seca. Dessa forma, o aerênquima pode ainda promover o transporte de gases entre a raiz e outras partes da planta (ARMSTRONG, 1980; EVANS, 2003; HE; MORGAN;

DREW, 1996).

O aerênquima pode ocorrer de várias formas, muitas das quais podem ser estrutural e metabolicamente complexas, apesar da sua pequena quantidade de massa de tecido (ARBER, 1920). Tem-se conhecimento sobre dois tipos de aerênquima: i) aerênquima cortical (aerênquima primário), o qual é formado nas raízes de cana-de-açúcar, arroz (*Oryza sativa*), milho (*Zea mays*), cevada (*Hordeum vulgare*) e trigo (*Triticum aestivum*) (EVANS, 2003; LEITE et al., 2017; NISHIUCHI et al., 2012); ii) e o outro é um tecido branco e esponjoso preenchido com espaços de ar que se formam no pecíolo, hipocótilo, raízes adventícias e nódulos de raízes de plantas leguminosas como a soja (*Glycine max*), soja selvagem (*Glycine soja*) e *Sesbania rostrata*, em condições de alagamento (ARIKADO, 1954; SARASWATI; MATOH; SEKIYA, 1992).

Em especial, o aerênquima cortical apresenta dois tipos de formação: o esquizógeno que é iniciado por um crescimento celular não uniforme e resulta da separação celular e o lisígeno, no qual as células crescem de forma uniforme mas diferente do anterior, o aerênquima aqui é formado pela morte de algumas células do córtex (SACHS, 1882). Em cana-de-açúcar, a formação é lisígena, tendo a morte celular programada seguida da degradação de polissacarídeos da parede celular, pela presença de enzimas hidrolíticas (GRANDIS, 2015; LEITE et al., 2017).

A Figura 1-b aponta para uma parede celular de uma célula viva do córtex da raiz de cana-de-açúcar, em região sem aerênquima. Já a degradação da parede celular que ocorre para a formação do aerênquima em raízes de cana-de-açúcar (Figura 1-c) é também observada em outros processos naturais como o amadurecimento de frutos, interação planta-microrganismo e abscisão foliar (GRANDIS et al., 2014). Tais processos levam à uma modificação celular através da hidrólise da parede. Segundo os autores, o processo envolvendo a percepção do sinal ambiental até o processo de hidrólise da parede ocorrem na sequência dos

seguintes módulos: 1) Percepção de um sinal, 2) Separação celular, 3) Expansão celular, 4) Morte celular programada, 5) Hidrólise ou afrouxamento de hemicelulose e celulose, 6) Hidrólise de celulose.

A degradação que foi descrita nas raízes de cana-de-açúcar durante o processo de desenvolvimento do aerênquima foi de β -glucano e pectinas, principalmente as associadas às lamelas médias, resultando na separação celular (LEITE et al., 2017). Já as hemiceluloses arabinoxilano e xiloglucano mostraram-se presentes nas paredes ao longo do desenvolvimento do aerênquima, embora revelando modificações na estrutura fina de suas moléculas (LEITE et al., 2017).

Embora presente, a degradação da parede celular durante a formação do aerênquima ocorre de forma parcial, o que ocorre é que parte dos polissacarídeos são modificados e rearranjados. Portanto, a formação do aerênquima em raízes de cana-de-açúcar é vista como uma sequência de eventos envolvendo a modificação da estrutura da parede celular, ao invés apenas de sua degradação por completo. Neste sentido, ainda pouco se sabe sobre tais modificações.

Em raízes de cana-de-açúcar, o aerênquima foi observado na transição do segundo para o terceiro centímetro da raiz, contados a partir do seu ápice e se apresentou completo já no quarto centímetro. Ao longo do desenvolvimento do aerênquima, apenas observou-se a degradação de células na região do córtex, nas quais o volume de aerênquima é de 0%, 6%, 19% e 32% para o primeiro, segundo, terceiro e quarto centímetro da raiz, respectivamente (Figura 3) (GRANDIS, 2015; LEITE et al., 2017). O desenho experimental para o estudo do aerênquima em raízes de cana-de-açúcar considerou cada centímetro da raiz, a começar a contagem pelo ápice, como sendo um segmento (S) da raiz.

Toda essa caracterização do aerênquima é uma resposta da integração dos diferentes níveis de regulação biológica, incluindo o transcricional e traducional.

Figura 3: Desenvolvimento do aerênquima em raiz de cana-de-açúcar, a partir do ápice até 5 cm. Os números acima das barras mostram a porcentagem média da área da secção transversal ocupada por aerênquima no córtex. Barras = erro padrão. Os números do eixo x mostram a distância em relação ao ápice. (B – F) Cortes histológicos transversais. Ápice da raiz com as células inteiras: tecidos periféricos compostos por epiderme, exoderme e cilindro esclerenquimatoso (pe), seguido pelo córtex (co), endoderme (en) e cilindro vascular (cv) (marcados em B). A degradação das paredes celulares se inicia em células mais periféricas (asteriscos em C e D) e evolui em direção ao cilindro esclerenquimatoso (setas em E e F); a partir da região entre 4 e 5cm de distância do ápice até a base da raiz se considera que o aerênquima esteja completamente desenvolvido (F); os espaços de ar são separados entre si por paredes não degradadas e raios de células íntegras (indicado por “a” em F) (barra=20 μ m). Adaptado de (LEITE et al., 2017).

2.2 Níveis de regulação biológica

O genoma é o conjunto completo de DNA de um organismo, incluindo todos os seus genes. Cada genoma contém todas as informações necessárias para construir e manter um organismo vivo e está presente igualmente em todas as células do indivíduo. O que diferencia a especialização de cada célula é o grupo de genes que nela são expressos. Assim, em uma dada célula, apenas um conjunto destes genes será expresso (i.e. transcritos) sob uma determinada condição. Entre eles, existe um gradiente do nível de expressão gênica e além disso, apenas alguns destes transcritos serão codificadores de proteínas.

2.2.1 Transcritômica

A transcritômica corresponde à identificação e quantificação em larga de escala de todos os transcritos expressos em uma amostra sob determinada condição, independente de ser ou não de um gene codificador de proteína.

No caso das plantas que são organismos eucariotos, os transcritos de genes codificadores de proteínas precisam passar por um processamento (Figura 4), até que a molécula madura de RNA seja formada. Nesta etapa, ainda no núcleo da célula, os íntrons são excisados levando à junção dos éxons (processo conhecido como *splicing*), onde o produto dessa junção é uma sequência de RNA mensageiro (mRNA), que é a sequência madura. Esse processo de *splicing* pode ocorrer de diferentes maneiras (do inglês, *alternative splicing*), gerando assim mais de uma isoforma de transcrito para um mesmo gene. Esta última pode se tornar complexa,

uma vez que cada nova isoforma determinará uma ação metabólica com características próprias (REDDYA et al., 2013; STAIGERA; BROWN, 2013). Além disso, uma dada isoforma pode estar relacionada às ações em uma célula ou tecido específico, estando envolvida em processos biológicos distintos no mesmo ou em diferentes órgãos, ou ainda estando envolvida em um mesmo processo biológico mas exercendo funções com diferenças sutis.

Uma forma de analisar o nível de expressão gênica de um determinado tecido é através de técnicas moleculares, como por exemplo a técnica de RNA-Seq (WANG; GERSTEIN; SNYDER, 2009), aqui descrita para quantificação de mRNA. Esta é uma metodologia de sequenciamento para obtenção de grande quantidade de dados de expressão gênica. Nessa técnica, as moléculas de RNAs são fragmentadas em curtos segmentos (chamados de *reads*) que são então sequenciados e posteriormente devem ser processados para a quantificação de cada transcrito.

Figura 4: Ilustração do processamento de um gene eucariótico codificador de proteína. No DNA, um gene codificador de proteína é composto por unidades de exons (unidades codificadoras) e íntrons (reguladores gênico). O gene é então transcrito em uma molécula de RNA precursora para posteriormente ser processada e gerar a sequência madura. No processamento ocorre a excisão dos íntrons e uma nova molécula de RNA, chamada de RNA mensageiro (mRNA) é formada. Fonte: cadernomedicina_o-dogma-central-da-biologia.

A quantificação pode ser obtida através de uma das três principais métricas (CONESA et al., 2016), o RPKM (do inglês, *Reads Per Kilobase Million*)

(MORTAZAVI et al., 2008), o FPKM (do inglês, *Fragments Per Kilobase Million*) (TRAPNELL et al., 2010, 2012a) e o TPM (do inglês, *Transcripts Per Kilobase Million*) (WAGNER; KIN; LYNCH, 2012). Tais métricas fazem uma normalização da profundidade do sequenciamento e do tamanho do gene.

O RPKM foi desenvolvido para execuções de RNA-Seq em *single-end*, onde cada *read* corresponde a um único fragmento que foi sequenciado. Já o FPKM foi desenvolvido para experimentos *paired-end*. Neste último, duas *reads* podem corresponder a um único fragmento, ou se uma *read* do par não é mapeada, apenas uma delas pode corresponder a um único fragmento. A única diferença entre estas duas métricas é que o FPKM considera que essas duas *reads* podem mapear para um fragmento e portanto, não conta esse fragmento duas vezes. Por fim, a terceira maneira de quantificação é o TPM, que é muito similar ao RPKM e FPKM, porém diverge na forma de calculá-la.

Para RPKM e FPKM podemos assumir que a somatória das *reads* (Σ_r) é dividida pelo fator escala (10^6). Essa etapa normaliza a profundidade do sequenciamento e gera o valor de RPM (*Reads Per Million*). Esse valor é então dividido pelo comprimento dos genes (L_g) em kilobases, gerando o valor de RPKM ou FPKM.

Similar à RPKM e FPKM, o TPM difere na ordem das equações. Onde a somatória das *reads* pelo comprimento de cada gene (L_g) em kilobase gera o valor de *Reads Per Kilobase* (RPK). Então a somatória de todos os RPKs dividido pelo fator de escala “*Per Million*” (10^6) gera o valor de TPM.

Usando o TPM a soma de todos os TPMs em cada amostra é o mesmo, facilitando a comparação da proporção de *reads* que mapeiam para um certo gene por cada amostra. Por outro lado, com RPKM e FPKM, a soma das *reads* normalizadas em cada amostra pode ser diferente, dificultando a comparação direta entre amostras.

A etapa de processamento dos dados gerados pelo RNA-Seq inclui várias sub-etapas, mas uma delas é crucial para a decisão da metodologia de análise. Esta sub-etapa é a reconstrução dos transcritos a partir das *reads* e neste caso existem duas abordagens principais: i) utilizando um genoma ou um exoma (conjunto de éxons) como referência; ii) ou assumir que não existe uma referência e fazer a montagem do início, através de uma técnica chamada *De novo*. Abaixo são descritos três exemplos de ferramentas computacionais capazes de realizar cada uma dessas etapas de processamento.

Entre os métodos de processamento que utilizam uma sequência de referência, o pipeline Tuxedo (TRAPNELL et al., 2012b) é uma metodologia bastante utilizada em trabalhos da área. Faz o mapeamento das *reads* no genoma de referência e em seguida faz a montagem dos transcritos com base nesse mapeamento. Posteriormente, realiza a quantificação do nível de expressão de cada transcrito. Uma vantagem desse método é a possibilidade de identificação de variantes de *splicing*, que está implementado no algoritmo *TopHat* (TRAPNELL; PACHTER; SALZBERG, 2009), que faz o mapeamento.

Uma segunda maneira de processar os dados do RNA-Seq é utilizando o transcrito do organismo de interesse, ao invés do genoma. A ferramenta RSEM (*RNA-Seq by Expectation-Maximization*) (LI; DEWEY, 2011) é baseada nessa metodologia e apresenta algumas particularidades em relação aos outros métodos. O fato de independer de um genoma para sua execução, é interessante em situações onde ainda não se tem o genoma do organismo de interesse, contudo existe uma referência de transcrito confiável, como é o caso da cana-de-açúcar, com o arquivo SUCEST (*Sugarcane Expressed Sequence Tags*) (VETTORE et al., 2001), onde cada transcrito é referido como SAS (*Sugarcane Assembly Sequence*). Uma vantagem dessa metodologia em relação à anterior, se o objetivo não for a identificação de novos variantes de *splicing*, é que o transcrito tem tamanho de

sequência menor comparado ao genoma e com isso o tempo de execução do algoritmo é menor. O tempo de RSEM também é menor comparado a metodologia descrita abaixo (*De novo*).

Quanto à montagem de novo, a ferramenta Trinity (GRABHERR et al., 2011) é um exemplo dessa metodologia e faz a reconstrução dos transcritos com base em algoritmos de grafos (*Bruijn graphs*). Essa estratégia independe de qualquer sequência de referência, mas, ainda assim, é capaz de fazer a identificação de novos variantes de *splicing*.

Independente da metodologia utilizada para o processamento dos dados de transcritômica, o resultado esperado são predições de listas de genes que estão sendo expressos/transcritos em cada amostra estudada, bem como seus respectivos valores de expressão. Contudo, o tamanho das listas de genes pode diferir entre metodologias.

2.2.2 Proteômica

Na seção anterior foram abordadas algumas estratégias de estudo de transcritoma em larga escala, mais precisamente sobre a quantificação de moléculas de mRNAs. Entretanto, o valor quantificado de mRNAs não reflete a quantidade de proteínas nas células. Isso ocorre porque existem eventos reguladores pós-transcricionais que atuam sobre os RNAs, podendo inclusive impedi-los de serem traduzidos.

No genoma humano existem aproximadamente 19.000 genes codificadores de proteínas (CHI, 2016; EZKURDIA et al., 2014) e 25.500 relatados para espécies diplóides de plantas (MARTINEZ, 2013; THE ARABIDOPSIS GENOME INITIATIVE,

2000).

A proteômica tem permitido estudos em larga escala de expressão proteica em diferentes tecidos, condições e momentos. Essa é uma análise complementar ao estudo do transcrito. A proteômica não é um trabalho trivial e requer o fracionamento, identificação, caracterização e quantificação das moléculas. De forma geral, pode-se dividir em separação e identificação das proteínas.

A separação pode ser através de eletroforese bidimensional, que em essência permite que as proteínas sejam separadas através de duas propriedades físico-químicas: ponto isoelétrico e massa molecular. Depois de separadas, elas são submetidas à digestão enzimática com proteases e os fragmentos (peptídeos) seguem para identificação através de espectrometria de massas.

Embora bastante utilizada, essa estratégia apresenta limitações para capturar proteínas com massa molecular muito pequena ou muito grande. Para estes casos, uma estratégia é utilizar a cromatografia líquida de alta eficiência (do inglês, *High Performance Liquid Chromatography* – HPLC) (MEYER, 2004) para a fase de separação, acoplada a uma técnica complementar de análise quantitativa, como a espectrometria de massas (do inglês, *Mass Spectrometry* – MS) (SCHULZE; USADEL, 2010) que utiliza métodos de ionização para a fase de detecção/identificação dos peptídeos. A junção destas duas metodologias é chamada de LC/MS (*Liquid Chromatography/Mass Spectrometry*).

A espectrometria de massa é uma ferramenta analítica, capaz de fornecer informações quanto às unidades elementares de uma amostra, estrutura molecular, composição quantitativa e qualitativa dos componentes, entre outros. Em princípio, as proteínas podem ter suas sequências inferidas a partir das massas de seus peptídeos, mas posteriormente faz-se necessário o uso de técnicas de análise de sequências em bioinformática, para confirmar a compatibilidade destas com outras sequências já conhecidas através das técnicas de genômica e transcritômica.

A proteômica quantitativa para cada um dos quatro primeiros centímetros (a contar pelo ápice) da raiz de cana-de-açúcar, utilizou a técnica de cromatografia líquida acoplada com espectrometria de massa em tandem (LC/MS-MS) para a análise de separação das proteínas. Para a quantificação foram usados marcadores moleculares que permitiram calcular quanto de cada proteína estava presente nos extratos usando medidas relativas. Os dados foram analisados utilizando a plataforma MaxQuant (COX; MANN, 2008; TYANOVA; TEMU; COX, 2016) no trabalho de Grandis (2015).

Ao todo foram identificadas 2.031 proteínas, sendo 186 de parede celular. A quantificação relativa destas proteínas foi determinada com base na razão entre os segmentos de raiz subsequente (1/2, 2/3, 3/4 e 4/5). Como resultado dessa etapa foram geradas listas de quantificação das proteínas para cada um dos segmentos da raiz.

A proteômica revelou respectivamente 140, 139, 141 e 150 proteínas de parede celular para os segmentos de 1 à 4 da raiz, sempre considerando como o segmento 1 o ápice da raiz e o segmento 4 o mais próximo à base. Adicionalmente, a Figura 31 (apêndice) mostra para cada segmento da raiz, as 15 subcategorias funcionais mais representadas na proteômica.

2.2.3 Atividade Enzimática

Enzimas são catalisadores biológicos, que por definição têm o papel de aumentar a taxa de reações químicas que sejam termodinamicamente possíveis, sem de fato participar delas como reagentes e que ao final podem ser recuperadas inalteradas.

Elas são também proteínas que são dobradas em formas complexas e que permitem que moléculas ainda menores sejam inseridas nessa estrutura. O local de inserção dessas moléculas substrato é chamado de sítio ativo e é extremamente específico. Dada essa especificidade, geralmente uma enzima catalisa apenas um só tipo de reação química, mediante a especificidade da enzima com o substrato. Decorrente deste fato, o tipo de enzima encontrada em uma célula determina o tipo de metabolismo que a mesma efetua. Se a forma da enzima é alterada, o seu sítio ativo pode não mais estar funcional e passa a ser considerada desnaturada. A desnaturação pode ocorrer devido às altas temperaturas ou valores extremos de pH.

As enzimas são classificadas de acordo com a reação que catalisam. As Hidrolases (p.ex. Glicosil Hidrolases (GHs), Lipases e Proteases) catalisam reações de hidrólise, as Oxirredutases (p.ex. Desidrogenase e Oxidase) agem sob reações de óxido-redução, as Transferases (p.ex. Metiltransferases), agem na transferência de grupos funcionais como o grupo metil-CH₃, as Liases (p.ex. Descarboxilase) agindo na quebra de ligações duplas como a remoção de CO₂, as Isomerases (p.ex. Isomerase cis-trans) agindo em isomerizações, e por fim as Ligases (p.ex. Sintetase), que agem formando ligações.

Essa classificação é acompanhada de uma nomenclatura internacional, criada pela International Union of Biochemistry and Molecular Biology (IUBMB), que atribui uma classificação numérica conhecida como “E.C.” (Enzyme Commission of the IUBMB) e que é composto por quatro dígitos: i) a Classe, ii) Sub-classe dentro da Classe, iii) Grupos químicos específicos que participam da reação, iv) A enzima propriamente dita. A lista completa das enzimas e sua classificação pode ser consultada em <http://www.chem.qmul.ac.uk/iubmb/enzyme/>.

Atividade enzimática é o nome dado ao processo de ação das enzimas e que depende de alguns fatores adicionais além da temperatura e pH, como por exemplo, a presença de cofatores e inibidores, a quantidade de substrato e até da própria

enzima, além do tempo de contato com o substrato. Cada enzima tem uma condição ótima, na qual a sua atividade é máxima.

Para a quantificação da atividade das enzimas de degradação de parede celular em segmentos de raiz de cana-de-açúcar, foram utilizados substratos sintéticos de 4-Nitrophenyl (PNP) e seus respectivos açúcares conjugados, que quando atacados pelas hidrolases específicas, resultaram em uma coloração que quando calculado o valor é possível estimar a atividade dessas enzimas no meio. Ao todo, foi possível medir a atividade de sete enzimas (Figura 5) (GRANDIS, 2015).

Figura 5: Atividade média das enzimas α -L-arabinofuranosidase (1 e 2), celulase (1 e 2), β -xilosidase, β -galactosidases e α -galactosidase, de degradação da parede celular, identificadas ao longo dos quatro segmentos (S) da raiz de cana-de-açúcar. Fonte dos dados: (GRANDIS, 2015).

Os resultados indicaram a presença de α -arabinofuranosidase, observada pela conjugação do substrato com o açúcar α -L-arabinopiranoside (α -

arabinofuranosidase 1), bem como do substrato com o açúcar α -L-arabinofuranoside (α -arabinofuranosidase 2). Identificou-se também a atividade de uma suposta endocelulase devido a utilização do substrato β -D-cellobioside (Celulase 1 – endo), e também de outra celulase, porém do tipo exocelulase devido à atividade com o substrato β -D-glucopiranoside (Celulase 2 – endo/exo), que tanto pode ser uma endo ou de exo-enzima. Por fim, as conjugações do substrato com os açúcares β -D-xilopiranoside (β -xilosidase), β -D-galactopiranoside (β -galactosidase), α -D-galactopiranoside (α -galactosidase). Muitas enzimas de parede celular não puderam ter sua atividade mensurada devido ao seu baixo limiar de atividade ou dificuldade na extração para o ensaio enzimático.

A Tabela 1 mostra todos os transcritos (SAS_IDs) de cana-de-açúcar, codificadores de cada enzima mencionada.

2.3 Redes gênicas

Os resultados gerados por técnicas de sequenciamento em larga escala têm fornecido uma grande quantidade de dados e para os vários níveis de regulação biológica de um organismo. Dessa forma, ao lado dos métodos experimentais de sequenciamento de genomas e de análise do perfil de expressão gênica, existem as técnicas computacionais que auxiliam na geração de hipóteses. Tais técnicas têm sido utilizadas para investigar os diferentes aspectos da biologia vegetal, incluindo o nível do transcrito, que analisa os diferentes padrões de expressão gênica, região específica de um tecido ou algumas poucas células. Estas, podem contribuir para uma predição mais acurada do comportamento regulador dos genes e com isso acelerar projetos biotecnológicos e direcionamentos experimentais em laboratório.

Uma forma de analisar os dados de expressão é calculando a correlação

Tabela 1: Transcritos codificadores (SAS_ID) de enzimas de degradação da parede celular (α -L-arabinofuranosidase, celulase, β -xilosidase, β -galactosidases e α -galactosidase). Cada anotação da enzima é ainda categorizada segundo o domínio que está relacionado à degradação: H/P para enzimas que podem agir sobre ambos hemicelulose pectina; C para enzimas que degradam celulose; H para enzimas que degradam hemicelulose.

SAS_ID	Anotação SUCEST Domínio Parede Celular	Enzima
SCEZRZ3099G01.g	α -L-arabinofuranosidase H/P_degradação	α -L-arabinofuranosidase
SCEZRZ3100F05.g	α -L-arabinofuranosidase H/P_degradação	
SCJFAD1010G05.b	α -L-arabinofuranosidase H/P_degradação	
SCJLRT1017C05.g	α -L-arabinofuranosidase H/P_degradação	
SCJLRT2049H07.g	α -L-arabinofuranosidase H/P_degradação	
SCJLST1027C06.g	α -L-arabinofuranosidase H/P_degradação	
SCMCST1055D01.g	α -L-arabinofuranosidase H/P_degradação	
SCQGLR1041C07.g	α -L-arabinofuranosidase H/P_degradação	
SCQSRT1034D03.g	α -L-arabinofuranosidase H/P_degradação	
SCSGSB1007C09.g	α -L-arabinofuranosidase H/P_degradação	
SCVPLB1020G05.g	α -L-arabinofuranosidase H/P_degradação	
SCACCL6010F02.g	Glicosil hidrolase família 1 C_degradação	Celulase
SCACLV1022D02.g	exoglucanase C_degradation	
SCAGLR1021F11.g	Glicosil hidrolase família 1 C_degradação	
SCBFAD1045D12.g	Glicosil hidrolase família 1 C_degradação	
SCBGRT1047G02.g	Glicosil hidrolase família 1 C_degradação	
SCBGSB1026A09.g	Glicosil hidrolase família 1 C_degradação	
SCCCCL3001B10.b	Glicosil hidrolase família 1 C_degradação	
SCCCCL4017A06.g	Glicosil hidrolase família 1 C_degradação	
SCCCCLB1004E10.g	Glicosil hidrolase família 1 C_degradação	
SCCCRT1C05G04.g	Glicosil hidrolase família 1 C_degradação	
SCCCRT3008H12.g	Glicosil hidrolase família 1 C_degradação	
SCCCSD1089A09.g	Glicosil hidrolase família 1 C_degradação	
SCCCST1002E03.g	Glicosil hidrolase família 1 C_degradação	
SCCCST1006H12.g	Glicosil hidrolase família 1 C_degradação	
SCEPCL6026C06.g	Glicosil hidrolase família 1 C_degradação	
SCEPSB1130A03.g	Glicosil hidrolase família 1 C_degradação	
SCEPSD1072G12.g	Glicosil hidrolase família 1 C_degradação	
SCEQHR1082B01.g	Glicosil hidrolase família 1 C_degradação	
SCEQLB1066E08.g	Glicosil hidrolase família 1 C_degradação	
SCEQRZ3023D01.g	Glicosil hidrolase família 1 C_degradação	
SCEZRT3067G10.g	Glicosil hidrolase família 1 C_degradação	
SCEZSD2080C09.g	Glicosil hidrolase família 1 C_degradação	
SCJFAD1011A07.b	Glicosil hidrolase família 1 C_degradação	
SCJFHR1C03E12.b	Glicosil hidrolase família 1 C_degradação	
SCJFLR1017E03.g	Glicosil hidrolase família 1 C_degradação	
SCJFST1009E07.b	Glicosil hidrolase família 1 C_degradação	
SCMCCL6048D12.g	Glicosil hidrolase família 1 C_degradação	
SCMCSB1108H08.g	Glicosil hidrolase família 1 C_degradação	
SCPIRT3022G07.g	Glicosil hidrolase família 1 C_degradação	

Continuação Tabela 1.

SAS_ID	Anotação SUCEST Domínio Parede Celular	Enzima
SCQGSB1066B02.g	Glicosil hidrolase família 1 C_degradação	Celulase
SCQGSD1048E08.g	Glicosil hidrolase família 1 C_degradação	
SCQSRT2032A07.g	Glicosil hidrolase família 1 C_degradação	
SCRLAD1040B02.g	Glicosil hidrolase família 1 C_degradação	
SCRLAD1043E01.g	Glicosil hidrolase família 1 C_degradação	
SCRUFL3060G01.g	Glicosil hidrolase família 1 C_degradação	
SCRULR1020A02.g	Glicosil hidrolase família 1 C_degradação	
SCSBSB1057E03.g	Glicosil hidrolase família 1 C_degradação	
SCSFRT2067G04.g	Glicosil hidrolase família 1 C_degradação	
SCSFRT2068B05.g	Glicosil hidrolase família 1 C_degradação	
SCSGAM2104E10.g	Glicosil hidrolase família 1 C_degradação	
SCUTSB1073H05.g	Glicosil hidrolase família 1 C_degradação	
SCVPRT2076H07.g	Glicosil hidrolase família 1 C_degradação	
SCVPRT3089B12.g	Glicosil hidrolase família 1 C_degradação	
SCVPRZ2042G03.g	Glicosil hidrolase família 1 C_degradação	
SCCCCL4009F05.g	β -xilosidase/ α -L-arabinofuranosidase (GH family 3) H/P_degradação	Xilosidase
SCEZFL4044G03.g	β -xilosidase/ α -L-arabinofuranosidase (GH family 3) H/P_degradação	
SCQSAM1030G04.g	β -xilosidase/ α -L-arabinofuranosidase (GH family 3) H/P_degradação	
SCAGST3137A07.g	β -galactosidase H/P_degradação	β -galactosidase
SCBGLR1044E06.g	β -galactosidase H/P_degradação	
SCCCCL3120A04.g	β -galactosidase H/P_degradação	
SCCCCL4001H11.g	β -galactosidase H/P_degradação	
SCCCCL6004H07.g	β -galactosidase H/P_degradação	
SCCCRT1002C07.g	β -galactosidase H/P_degradação	
SCCCST1006B07.g	β -galactosidase H/P_degradação	
SCEPAM1022E11.g	β -galactosidase H/P_degradação	
SCEPRT2044F09.g	β -galactosidase H/P_degradação	
SCEPSB1131E01.g	β -galactosidase H/P_degradação	
SCEPSB1133A06.g	β -galactosidase H/P_degradação	
SCEQFL5044F07.g	β -galactosidase H/P_degradação	
SCJFAD1014H06.b	β -galactosidase H/P_degradação	
SCJFAD1C12B09.b	β -galactosidase H/P_degradação	
SCJLFL3014G04.g	β -galactosidase H/P_degradação	
SCJLFL4186G07.g	β -galactosidase H/P_degradação	
SCQSFL3033B04.g	β -galactosidase H/P_degradação	
SCR FAM2127G01.g	β -galactosidase H/P_degradação	
SCSFFL3091E06.g	β -galactosidase H/P_degradação	
SCSFSB1101E12.g	β -galactosidase H/P_degradação	
SCVPAM1059F06.g	β -galactosidase H/P_degradação	
SCVPRZ2037A02.g	β -galactosidase H/P_degradação	

Continuação Tabela 1.

SAS_ID	Anotação SUCEST Domínio Parede Celular	Enzima
SCAGAM2016B10.g	α-galactosidase H_degradação	α-galactosidase
SCAGLB1071G05.g	α-galactosidase H_degradação	
SCBFRZ3005H01.g	α-galactosidase H_degradação	
SCCCRZ3004D03.g	α-galactosidase H_degradação	
SCEPSD2006B02.g	α-galactosidase H_degradação	
SCEQRT2025H02.g	α-galactosidase H_degradação	
SCJLLB2078F05.b	α-galactosidase H_degradação	
SCQSRT2034G08.g	α-galactosidase H_degradação	
SCUTLR1037C07.g	α-galactosidase H_degradação	

entre os padrões de expressão de dois genes e em seguida visualizar graficamente as correlações mais fortes. Neste caso, sistemas complexos como o conjunto de genes expressos em uma condição biológica, podem ser representados por estruturas chamadas redes.

As redes são matematicamente representadas por grafos (G) e definidas por uma coleção de vértices (V), também chamados de nós e pelas *Edges* (E), também chamadas de arestas, as quais podem ser representadas por: $G=[V,E]$.

Em se tratando de redes de expressão gênica, os nós são os genes e as arestas (ou ligações) são as interações gênicas, as quais podem ser entendidas como interações par-a-par de correlação (Figura 6) (BARABÁSI; OLTVAI, 2004). As redes de correlação são também chamadas de redes de co-expressão (do inglês: *Gene Co-expression Network* (GCN)) e têm ajudado na investigação de diferentes aspectos da biologia vegetal, principalmente ao nível de regulação da transcritômica, permitindo a identificação de padrões de expressão tecido específico (REID et al., 2006) bem como da expressão de padrões de desenvolvimento (CHÁVEZ-BÁRCENAS et al., 2000). Tais redes devem ser analisadas com informações adicionais, como o nível de expressão gênica e localização celular, além de outras informações que estejam disponíveis quanto ao processo de estudo em questão.

Elas permitem a identificação de quais genes estão sendo co-expressos e para qual condição ou estado da planta.

Tais redes permitem a identificação de co-expressão de pares de genes, mas isso não implica na causalidade de regulação entre eles, pois cada aresta indica a correlação entre os genes e não a regulação, que por outro lado pode ser obtida nas redes de regulação gênica (do inglês: *Gene Regulatory Networks* (GRNs)). Diferente da primeira, estas podem representar processos bioquímicos como reações, transformações, interações, ativações e inibições (LOPES; DE OLIVEIRA; CESAR-JR, 2009; LOPES; DE OLIVEIRA; CESAR, 2011; ROY; BHATTACHARYYA; KALITA, 2014). Dessa forma, redes de co-expressão costumam ser não direcionadas enquanto as de regulação gênica são direcionadas (Figura 6).

Figura 6: Ilustração de uma rede/grafo. A rede possui cinco nós e seis arestas, apontados nas setas. Arestas podem ser não-direcionadas como as encontradas em redes de co-expressão, enquanto as arestas direcionadas ocorrem em redes de regulação gênica.

Para a construção das redes é necessária uma medida de co-expressão/similaridade entre os possíveis pares de genes. A maioria das medidas

pode ser categorizada em: coeficiente de correlação e medidas de informação mútua.

Medidas de informação mútua (do inglês: *Mutual Information* – MI) (BASSO et al., 2005; DAUB et al., 2004; MARGOLIN et al., 2006; PRINCESS; MAIMON; BENGAL, 2007; SONG; LANGFELDER; HORVATH, 2012) podem ser utilizadas para medidas de associação não lineares. O método é bem definido para variáveis discretas e categóricas.

Em contraste, o coeficiente de correlação e outros modelos baseados em medidas de associação são ideais para capturar o relacionamento linear entre duas variáveis, por exemplo, dadas observações amostrais, verificar se o valor de expressão de um gene “A” altera conforme varia o valor de expressão do gene “B” (AOKI; OGATA; SHIBATA, 2007; EISEN MB, SPELLMAN PT, BROWN PO, 1998; LANGFELDER; HORVATH, 2008; STUART, 2003; ZHANG; HORVATH, 2005). Uma vantagem das medidas de correlação é que por capturarem o relacionamento par a par entre todas as variáveis amostradas, essa informação abre margem para uma outra vertente de análise através de propriedades da rede de co-expressão, pela análise da sua conectividade, permitindo investigar assim, as centralidades da rede. Considerando o conjunto de dados gerados por tecnologias de larga escala (do inglês; *Big Data*), como é o caso do RNA-Seq, tais métodos contribuem para a redução da dimensionalidade das informações dos dados.

Como uma das medidas que são utilizadas para determinar a similaridade por coeficiente de correlação, Pearson é frequentemente utilizada e com boa performance comparada com outros métodos sofisticados, considerando a capacidade de identificar a relação entre os pares de genes e pela capacidade em lidar com grandes conjuntos de dados (BALLOUZ; VERLEYEN; GILLIS, 2015; KUMARI et al., 2012; SONG; LANGFELDER; HORVATH, 2012).

A correlação de Pearson (PEARSON, 1920) é um método paramétrico que

assume uma distribuição normal para os dados, capaz de medir a força da relação linear entre duas variáveis aleatórias. O valor de correlação está entre -1 e +1 e valores próximos destes representam correlação linear quase perfeita. Valores positivos indicam que a variação aumenta para ambas variáveis, enquanto valores de correlação negativa mostram que a variação de uma amostra sobe enquanto da outra diminui. Coeficiente de correlação igual a zero indica ausência de relação linear entre as variáveis. O método é ainda sensível aos valores isolados de expressão gênica (do inglês: *outliers*) e seu coeficiente é normalmente denominado por r de Pearson e definido como:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_x S_y} \quad (1)$$

Onde, $\bar{X}$ e $\bar{Y}$ são as médias das amostras e S_x e S_y correspondem ao desvio padrão das amostras X e Y, respectivamente.

Sobre a hipótese nula das duas variáveis serem independentes calcula-se a distribuição t de *Student*, com $n-2$ grau de liberdade, como segue:

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}} \quad (2)$$

De forma geral, a utilização de um método para a construção de uma rede de co-expressão é muito importante, porque o método pode apontar grupos de genes verdadeiramente associados e com isso aumentar o conhecimento sobre o processo biológico. Não existe o melhor método para a construção das redes de co-expressão pois a sua eficiência pode variar de acordo com o processo biológico em questão (KUMARI et al., 2012).

2.3.1 Centralidades das redes de co-expressão

O estudo de centralidades de redes tem sido aplicado há muito tempo em redes de relacionamentos de indivíduos, organizações, países, comércios (BARRAT et al., 2004; FREEMAN; BORGATTI; WHITE, 1991; JI et al., 2016; MAKAGON; MCCOWAN; MENCH, 2012; OPSAHL; AGNEESSENS; SKVORETZ, 2010), bem como em aplicações biológicas em estudos com redes de co-expressão gênica (AZUAJE, 2014; FULLER et al., 2007; IANCU et al., 2012; JIANG et al., 2016; LIANG et al., 2014; VAN DAM et al., 2017).

A centralidade é uma propriedade da rede, capaz de ranquear os nós e que vem acompanhada de uma pergunta biológica relacionada. Em redes biológicas, é importante por exemplo, detectar elementos centrais e intermediários que afetam a topologia da rede, dependendo também do curso da pergunta biológica. Neste caso, entende-se por topologia, o padrão pelo qual os nós e as arestas estão conectados. A pergunta biológica pode ser a busca para encontrar uma molécula em uma via metabólica que não é necessariamente central mas que tem um papel biológico crucial na transdução do sinal ou em redes de interação proteína-proteína, ou ainda identificar moléculas que são cruciais para o estímulo da expressão de genes.

A interação de dados biológicos, como dos genes, das proteínas e dos metabólitos, podem ser representadas por pequenas redes, mas geralmente alcançam complexas proporções com centenas de nós e arestas, desafiando assim, os métodos de processamento de tais estruturas.

Entre os vários métodos já desenvolvidos para a análise de redes complexas, o uso das centralidades, como o grau de conectividade (do inglês: *degree*) (EGERSTEDT, 2011; GILLIS et al., 2011; GILLIS; PAVLIDIS, 2011; LIU; SLOTINE; BARABÁSI, 2011; MÜLLER; SCHUPPERT, 2011; NEWMAN, 2003) e medidas de

comprimento entre os nós de uma rede (do inglês: *Closeness* e *Betweenness*) (BAVELAS, 1950; FREEMAN, 1978) são amplamente utilizadas.

Enquanto os nós com maior valor de *degree* representam aqueles mais centrais (do inglês: *hubs*), os nós com maiores valores de *closeness* e *betweenness* de uma rede têm algumas vantagens em relação aos primeiros: i) são mais acessados, ii) podem alcançar outros nós mais rapidamente e iii) podem controlar o fluxo entre outros nós (BARRAT et al., 2004; FREEMAN, 1978; NEWMAN, 2003).

Uma vantagem do cálculo do *degree* é que o seu cálculo apenas requer informação local de seus vizinhos conectados. Mas em contrapartida, não leva em consideração a estrutura global da rede. Por exemplo, um gene com alto valor de *degree* mas que não está em uma posição que se conecte aos outros rapidamente. Para capturar essa característica a centralidade *closeness* foi definida como o caminho mais curto para acessar todos os outros nós através de um mesmo nó fonte. A principal limitação do *closeness* é a sua aplicação em redes com nós desconectados, limitando o seu uso para estes casos. Assim, um alto valor de *closeness* indica maior proximidade de todos os outros nós da rede (DRRKSRRRA, 1959; NEWMAN, 2001).

A terceira medida de centralidade é o *betweenness*, que de forma geral busca identificar a essencialidade de cada nó. Primeiro captura-se o *degree* em que um nó se encontra, considerando o caminho mais curto entre dois outros nós, para então construir um caminho quanto ao fluxo da rede, ou seja, o *betweenness* calcula para cada nó, quantas vezes ele é acessado para permitir o caminho entre nós vizinhos. Embora essa medida considere a estrutura global da rede, ela pode ser aplicada em redes com componentes desconectados. Pode ocorrer, por exemplo, de uma grande quantidade de nós de uma rede, não serem o caminho mais curto entre outros dois nós e com isso terão valor de *betweenness* igual a zero. Dessa forma, um alto valor de *betweenness* indica que o nó é altamente acessado no caminho de outros nós

vizinhos.

Estas três medidas de centralidades são aplicadas para redes ponderadas (BARRAT et al., 2004; OPSAHL; AGNEESSENS; SKVORETZ, 2010), como é o caso das redes de co-expressão.

A centralidade do *degree* (C_D) do nó (x) é calculada pelo grau de conexões do nó, dividido pelo número de nós menos 1 (valor normalizado, indicando o grau de conexão média), como segue:

$$C_D(x) = \frac{\deg(x)}{n-1} \quad (3)$$

Onde n é o número de nós da rede. O *closeness* pode ser entendido como a soma de todos os menores caminhos entre um nó e todos os outros em uma rede, geralmente referido à sua forma normalizada que indica o comprimento médio de menores caminhos ao invés da sua soma. Assim, quanto mais central for um nó, mais próximo ele estará de todos os outros. O *closeness* é definido como:

$$C_c(x) = \frac{1}{\sum_y d(x,y)} \quad (4)$$

onde $d(x,y)$ é o menor caminho entre x e y .

Já a medida de *betweenness* é calculada pelo número médio de menores caminhos que passam através de um determinado nó. A centralidade *betweenness* C_B de um nó (v) é calculada como:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (5)$$

onde σ_{st} é o número dos menores caminhos a partir de s até t e $\sigma_{st}(v)$ é o número de menores caminhos de s até t passando por v . Posteriormente, soma-se essa fração para todos pares de vértices (s,t). Na prática, se uma grande

porcentagem de menores caminhos passa por um certo nó, a remoção deste nó implicará em um efeito significativo na topologia da rede.

2.4 Biologia de sistemas e a integração de dados

Biologia de sistemas é o estudo de sistemas biológicos. Não necessariamente envolve o processamento de grandes conjuntos de dados, mas frequentemente requer a utilização de métodos computacionais complexos, capazes de integrar os diferentes sistemas biológicos.

Um sistema biológico pode ser entendido como uma rede complexa de entidades biologicamente relevantes. A organização biológica pode ocorrer de diversas maneiras. Exemplos de sistemas biológicos podem ser representados pelo sistema circulatório, respiratório e nervoso, ou também pela escala dos tecidos e órgãos nos animais, ou ainda pela grande quantidade de dados gerados pela biologia molecular (ômicas) na era do sequenciamento em larga escala (conhecido como *Next-Generation Sequencing* - NGS) (MA'AYAN et al., 2014; SALEEM et al., 2014), que são: genômica, transcritômica, proteômica, metabolômica e etc.

Neste contexto, uma visão integrada dos dados de diferentes sistemas, disponibiliza ao pesquisador uma visão global de seus experimentos. Decifrando a complexidade dos sistemas biológicos, que vão desde a identificação das redes funcionais das partes, aqui chamadas de *layers*, que formam o todo, até a identificação de propriedades emergentes, as quais são decorrentes dessa abordagem. Essa visão pode contribuir para tomadas de decisões mais rápidas a respeito de seus resultados (TREWAVAS, 2006).

Tecnicamente, soluções para a integração de dados têm sido desenvolvidas e

aplicadas pelos setores acadêmico (KUO; TIAN; TSENG, 2013; RITCHIE et al., 2015; SRIVASTAVA; KUMAR; RAMASWAMY, 2014), corporativo, bem como ocorrendo por parceria de ambos setores, como é o caso do presente estudo, em colaboração da Universidade de São Paulo, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e *Microsoft Research*.

A partir do momento em que se realiza uma pesquisa biológica, a integração de dados pode ocorrer por meio de diferentes metodologias e para diferentes níveis de dados.

Exemplos de estudos de integração de dados que permitem novas descobertas biológicas incluem: a integração de genômica, transcritômica e proteômica em humanos revelando sub-tipos invasivos de lóbulos de câncer de mama (MICHAUT et al., 2016), estudo da robustez (rigidez e flexibilidade) de redes metabólicas em *Mycobacterium tuberculosis* (PUNIYA et al., 2016), compreender o funcionamento de pêlos radiculares em *Arabidopsis thaliana*, revelando propriedades espécie-específica e comuns desse tecido (WANG; LAN; SHEN, 2016), a identificação de elementos fundamentais do metabolismo das plantas (KOLBE et al., 2006), a integração metabólica entre órgãos vegetais respondendo ao ambiente, o que possibilita a visão do organismo como um todo (DE SOUZA et al., 2015), bem como a inferência do processo de regeneração da planária através da integração de dados experimentais e a aplicação de métodos de engenharia reversa (LOBO; LEVIN, 2015).

A escolha da metodologia está diretamente relacionada aos tipos de dados biológicos a serem integrados, já que se trata de uma análise heterogênea, com diferentes formatos e tamanhos de arquivos. Os formatos costumam ser padrões para cada tipo de dado e isso deve ser respeitado e mantido, aumentando assim a comunicação entre as diferentes bases de dados biológicos existentes na literatura, diminuindo a redundância entre eles e permitindo a reprodutibilidade de suas

análises.

Além da produção do dado em si, existem os desafios do gerenciamento, armazenamento e consequente processamento das informações de cada sistema (NEKRUTENKO; TAYLOR, 2012; WANDELT et al., 2012). Para isso, o profissional de bioinformática geralmente aplica técnicas de paralelização dos algoritmos, execução baseada em nuvem e complexos fluxogramas de trabalho (do inglês: *pipelines*).

Por se tratar de uma grande demanda de dados biológicos, é essencial e urgente que existam metodologias apropriadas para o seu processamento. Neste sentido, estudos com ômicas têm se movido das técnicas dirigidas pela hipótese para as dirigidas pelos dados (MERELLI et al., 2014). Este é um caminho para melhor se conhecer as vias moleculares de doenças e também de estudos de desenvolvimento de estruturas ou tecidos vegetais, por exemplo. São análises precursoras da medicina personalizada, estudos de melhoramento genético e também de metodologias preditivas, mas que apenas são possíveis após a compreensão dos dados ou do modelo biológico em estudo (SCHNEIDER; JIMENEZ, 2012).

Existem diversos métodos para se desenvolver a integração de dados (RITCHIE et al., 2015), podendo ser essencialmente categorizadas entre: banco de dados, onde os três mundialmente principais são: [NCBI](#) (*National Center for Biotechnology Information*), [ENSEMBL](#) e [DDBJ](#) (*DNA Data Bank of Japan*), e as ferramentas computacionais, que são programas de informática específicos para análise dos dados. Estas últimas variam muito em relação as técnicas e ao tipo de origem dos dados em questão. Neste caso, as ferramentas tanto podem utilizar as bases acima, quanto podem ser adaptadas para execução de bases locais de dados.

Com foco nas ferramentas de integração de dados, elas suportam dados de

organismos biológicos cujos genomas já estão completos e podem ser acessadas essencialmente através de acesso *online* (i.e. em uma página de *internet*), como as disponíveis para humanos (HU et al., 2005, 2009; JUNKER; KLUKAS; SCHREIBER, 2006; KARNOVSKY et al., 2012; KUO; TIAN; TSENG, 2013; LEMMENS et al., 2009), existindo alguns poucos modelos envolvendo a integração em plantas como arroz (*Oriza sativa*) (CHO et al., 2008), *Arabidopsis* (*Arabidopsis thaliana*) (KOLBE et al., 2006; WESTON et al., 2011), pópulus (*Populus trichocarpa*) e soja (*Glycine max*) (WESTON et al., 2011), além de algumas outras (KATARI et al., 2010).

Contudo, o fato de estarem restritas a genomas completos deixa de permitir integração de dados de organismos muitas vezes com importância biotecnológica como é o caso da cana-de-açúcar, que enfrenta os desafios do projeto genoma, mas que já possuem bases consolidadas de seus outros conjuntos de dados, como de transcritômica e proteômica, por exemplo.

Com o enfoque nas plantas, tais ferramentas de integração são voltadas para a visualização gráfica de dados disponíveis em bancos públicos, o que é diferente de um modelo com capacidade de executar novas análises e novas interações, além das já catalogadas em plataformas públicas. Tal questão é portanto uma das principais lacunas na área de integração de dados em plantas.

Embora possam ser analisados individualmente, na célula, os diferentes níveis de regulação biológica (transcritômica, proteômica, etc.), são organizados seguindo o dogma central da biologia molecular (CRICK, 1970) e portanto, uma *layer* é dependente de uma ou algumas *layers* anteriores. Ademais, entre cada uma das camadas existe um filtro de transição, a exemplo das modificações epigenéticas, que são capazes de regular uma certa molécula mesmo em nível pós-traducional. De tal forma que mesmo traduzida, uma proteína pode ainda ser regulada e impedida de exercer a sua função. Existe portanto, um nível de regulação que ocorre dentro de cada *layer*, aqui denominada *intra-layer*, bem como existe a regulação que

ocorre entre *layers*, aqui denominada *inter-layer*.

Esta variada possibilidade de regulação biológica intra e *inter-layers* demonstra uma parte da complexidade que as ferramentas computacionais enfrentam para modelagem e integração dos dados, principalmente quando se deseja desenvolver um modelo preditor para identificação de novas relações, sem estar restrito às já existentes em bancos de dados públicos.

Neste sentido, investigar o estado funcional das *layers*, bem como da transição da informação biológica entre as várias *layers* pode dizer muito sobre o processo funcional e de desenvolvimento de um tecido ou organismo.

O estudo *intra-layer* é o que mais tem sido abordado até hoje, principalmente para a transcritômica (GENOUD; MÉTRAUX, 1999; HASHIMOTO et al., 2004; KROUK et al., 2013; LOPES et al., 2014a, 2014b; STEEN, 2012; VICENTE et al., 2012) e proteômica (HAN et al., 2004; JANSEN; GREENBAUM; GERSTEIN, 2002; SZKLARCZYK et al., 2014).

Entre as estratégias utilizadas na integração envolvendo dados de transcritômica, o uso de redes complexas e métodos de centralidades auxiliam na priorização de elementos. De tal forma que um elemento apontado anteriormente como central em uma rede, por exemplo considerando apenas uma categoria funcional, pode não ser mais central quando todo o sistema é considerado em conjunto, ou seja, considerando os genes de todas as categorias funcionais, e com isso pode emergir uma nova centralidade da rede.

O processo de integração de dados vai muito além do processamento de arquivos em si e dependendo da maneira como são compartilhados os seus resultados, podem apresentar menor ou maior impacto sobre a comunidade científica como um todo. Geralmente, plataformas de visualização gráfica de dados, contribui não apenas, mas principalmente, para a parcela de usuários sem expertise com linguagens de programação. Tais plataformas podem ser de visualização gráfica

dos arquivos integrados (GeneMANIA) (WARDE-FARLEY et al., 2010), como podem ser de algoritmos para a integração de dados, como o projeto Cytoscape (SHANNON, 2003), com vários plugins já desenvolvidos para este intuito. Assim, muitas análises ganham significado biológico após serem exploradas em plataformas gráficas.

Considerando os desafios que as ferramentas computacionais enfrentam na integração de dados podem-se mencionar: a capacidade em acessar as informações dos vários e heterogêneos conjunto de dados, a análise em si, que inclui como os vários sistemas irão se comunicar (integrar), as estratégias para a identificação de elementos emergentes, bem como a visualização dos dados.

O acesso aos dados vai depender de como estes arquivos estão armazenados. Os dados biológicos podem estar armazenados em pequenos arquivos (unidade de medida em kilobytes - KB), mas podem alcançar grandes tamanhos (unidade de medida em terabytes - TB), principalmente em se tratando de análises de ômicas (LI; CHEN, 2014), e análises que consideram vários tratamentos, tecidos e réplicas biológicas. Por isso faz-se necessário um planejamento sobre quais arquivos e análises cada ferramenta realiza, para que então seja criada uma estrutura computacional adaptada.

Existem vários formatos de arquivos para dados biológicos e o acesso vai depender do tipo de análise em execução, que indica os formatos a serem processados e conseqüentemente as vias que podem ser aplicadas para a obtenção da informação. Existem diversas linguagens de programação que podem auxiliar nessa etapa, por exemplo, [Perl](#), [Python](#) e [R](#), além das ferramentas disponibilizadas pelas próprias plataformas geradoras dos dados. Muitas vezes é necessário aplicar várias etapas de conversão dos dados, até que se obtenha o formato desejado.

Estes são alguns dos desafios encontrados no processo de integração de dados. Neste estudo, busca-se essencialmente a compreensão sistêmica da

integração dos níveis de transcritômica, proteômica e fisiológicos com atividade enzimática, todos observados sob o mesmo processo biológico, que é o estudo do desenvolvimento do aerênquima.

3 Objetivos

3.1 Objetivo geral

Desenvolver uma ferramenta computacional que seja capaz de integrar os dados de transcrito, proteoma e fisiologia e aplicá-la ao processo de desenvolvimento do aerênquima em raízes de cana-de-açúcar.

3.2 Objetivos específicos

- a) Processar o transcrito dos quatro segmentos da raiz;
- b) Predizer redes de interações gênicas;
- c) Integrar os níveis de regulação de transcrito e proteômica;
- d) Integrar os resultados obtidos nas etapas anteriores com os dados de fisiologia;
- e) Gerar visualização dos dados.

4 Metodologia

4.1 Desenvolvimento da ferramenta SIT

Neste trabalho foi desenvolvida a ferramenta computacional SIT (*Systems Integration Tool*), capaz de fazer a integração de dados heterogêneos como os de transcritômica, proteômica e dados de fisiologia, além de permitir a análise de redes complexas por meio de uma interface dinâmica para exploração de parâmetros e análise de centralidades. SIT é composta por dois principais módulos, o primeiro para a integração e criação das redes, enquanto o segundo realiza a análise e exploração das redes geradas.

O primeiro módulo foi integralmente desenvolvido em linguagem de programação [Perl](#), utilizando [Perl-Tk](#) para a construção das aplicações. Ele foi subdividido em três abas: integração, construção da rede e análise de correlação.

A primeira aba foi feita para o carregamento dos dados, visualização dos arquivos carregados, configuração de parâmetros, até finalizar com a integração de *layers* propriamente dita.

O algoritmo de integração das *layers* é baseado em expressões regulares, com processamento textual dos nomes das colunas escolhidas nas tabelas pelo usuário. SIT permite que o usuário configure a sua busca pelo nome completo ou ao invés disso, permitir que apenas uma parte do nome seja comparado, e obtido todos os nomes que compõe o que foi buscado. Na condição deste último caso, consideramos a possibilidade de inconsistências entre os nomes, já que não por acaso podem existir entradas de nomes repetidos dentro de uma mesma *layer*.

Para evitar inconsistências e falsos resultados nessas situações, SIT requer

que, para estes casos apenas, sejam adicionados um termo chave, que é um caracter ‘_’ seguido de um prefixo de identificação (p.ex.: “_p” para proteômica e “_t” para transcritômica) para essas entradas.

A cada nova execução dessa primeira aba, SIT gera um arquivo de integração, que deve agora ser configurado para a construção da rede. Essa segunda etapa de configuração é realizada na segunda aba. Nela, o usuário deverá carregar um novo arquivo com a anotação dos nós, ou seja, é uma informação que será atribuída a cada elemento integrado na etapa anterior (aba um). Opcionalmente, esse arquivo pode conter um valor para o tamanho do nó e que pode ser, por exemplo, o valor de expressão gênica. E um segundo arquivo com a descrição dos pares das redes, que podem contudo, ser criado conforme interesse do usuário. Os pares podem ser de correlação, regulação ou simplesmente de relação entre os pares de nós.

A terceira aba desse módulo é para a análise de correlação, utilizando as medidas das variáveis de interesse (p.ex. medidas de expressão gênica). Nela foi implementado o cálculo de correlação por Pearson e Spearman, com botões de configuração para o filtro dos resultados.

Já com a rede de integração de dados gerada, o arquivo é lido pelo segundo módulo de SIT, que é composto de uma visualização dinâmica das redes, além de permitir análises adicionais de propriedades (*Degree*, *Closeness* e *Betweenness*) das mesmas. O desenvolvimento dessa etapa foi realizado em JavaScript/HTML com colaboração de Raphael Branco.

Esse módulo é executado em ambiente HTML, com uso do provedor de internet de preferência. A construção de algumas funções desse módulo foi baseada em bibliotecas do [cytoscape.js](#), dos quais implementamos os diferentes *layouts*, a palheta de zoom, a edição dinâmica dos nós, a visualização/ocultação de vizinhos, opção para remover/mostrar nós individuais, a edição dos rótulos (do inglês, *labels*)

dos nós e arestas, tipos de arestas (direcionadas ou não), bem como os cálculos de centralidades (*Degree* e *Betweenness*).

As demais características desse módulo foram de programação personalizada. Entre elas tem-se duas barras de ferramentas, uma localizada na parte superior da janela e outra que é contrátil e está localizada na lateral direita da janela. Entre as funcionalidades dessas barras tem-se: um campo de busca por palavra-chave; botões para execução das bibliotecas *cytoscape.js* mencionadas acima; filtro por quantidade e/ou tipo de *layers*, incluindo a funcionalidade que disponibiliza uma palheta de cores para que o usuário pode aplicar uma configuração personalizada da sua *layer*; a exportação das redes na forma de figura com extensão “.png”; botões para filtro de valores das arestas; uma aba de navegação da imagem e por fim, um campo de texto que registra cada ação executada no programa, mas que permite edição do usuário. Assim, esse relatório que contém todas as atividades feitas durante as análises, pode ser utilizado como um registro de bordo.

4.2 Preparação dos dados biológicos

4.2.1 Processamento do transcritoma de segmentos de raízes de cana-de-açúcar

A preparação do material biológico e a construção da biblioteca de RNA foram desenvolvidas durante a tese de doutorado da aluna Eveline Queiroz de Pinho Tavares (Processo-FAPESP nº 2011/07586-5) (TAVARES, 2015) e dentro do

conjunto de dados a ser utilizado na aplicação da ferramenta de integração, o transcrito era o único que ainda não havia sido processado, ou seja, os dados estavam no formato bruto necessitando serem trabalhados antes de iniciarmos as análises.

O RNA-Seq de raízes de cana-de-açúcar foi realizado com três réplicas biológicas compostas para cada segmento da raiz, respeitando a metodologia de segmentação da raiz a cada 1 cm do ápice em direção à base da raiz. Para o RNA-Seq utilizou-se até o quarto segmento da raiz. Essa metodologia tem sido aplicada no grupo em trabalhos anteriores (GRANDIS, 2015; LEITE et al., 2017; TAVARES, 2015).

A etapa inicial do processamento de dados do transcrito foi a preparação das *reads*. Após a preparação dessas *reads*, estas foram quantificadas através de pelo menos duas metodologias, pipeline Tuxedo (TRAPNELL et al., 2012b) e ferramenta RSEM (LI; DEWEY, 2011), ambas utilizando o transcrito de cana-de-açúcar (SUCEST) como sequência de referência para a reconstrução dos transcritos, já que não temos o genoma disponível.

4.2.1.1 Preparação das reads

O sequenciamento do RNA-Seq foi realizado através da plataforma SOLID (*Sequencing by Oligo Ligation and Detection* - [Applied Biosystems](#)), cuja biblioteca foi *single-end*. A avaliação de qualidade das *reads* foi realizada por meio da aplicação do algoritmo [FASTQC](#) (ANDREWS, 2010). Essa análise foi realizada para cada uma das réplicas dos segmentos da raiz.

De acordo com os resultados obtidos, a presença de bases com baixa qualidade foi corrigida aplicando-se um filtro por bases com valores de qualidade

Phred maiores ou iguais a 20, o que corresponde a uma acurácia de 99% para cada base (EWING; GREEN, 1998). Esse filtro foi realizado com a execução da ferramenta Trimmomatic (BOLGER; LOHSE; USADEL, 2014). Também através dessa ferramenta, aplicou-se o filtro para comprimento de sequências maiores ou iguais a 40 nucleotídeos (nt), uma vez que o processo de filtragem por qualidade implica na geração de uma grande quantidade de pequenos fragmentos de sequências e que devido à sua redundância no transcrito, pode influenciar negativamente a taxa de reconstrução dos transcritos. Para confirmar a qualidade das *reads* filtradas, aplicou-se novamente a ferramenta FASTQC.

4.2.1.2 Estratégias para a quantificação dos transcritos

As *reads* já processadas para os filtros de qualidade e comprimento, foram analisadas utilizando duas diferentes abordagens: a) pipeline Tuxedo (TRAPNELL et al., 2012b), que aplica as seguintes etapas: primeiro realiza o mapeamento das *reads* sobre a sequência de referência usando a ferramenta Tophat, posteriormente utiliza os dados gerados para quantificar a expressão gênica, este último realizado por meio do pacote de programas cufflinks; b) ferramenta RSEM, que constrói um mapeamento das *reads* sobre a sequência de referência e em seguida, quantifica esses transcritos mapeados. Para ambos os casos, o SUCESST foi utilizado como sequência de referência, além disso, a unidade de quantificação dos transcritos adotada no trabalho foi a TPM.

Além das listas com os transcritos e respectivos valores de expressão, as ferramentas geram também, arquivos de alinhamento das *reads* (extensão .bam) e também de coordenadas dos transcritos (extensão .bam.bai). Todos estes arquivos podem ser úteis em análises posteriores de enriquecimento dos transcritos, incluindo

a visualização gráfica dos mesmos.

A importância dessa etapa para o trabalho está em poder analisar comparativamente o padrão de expressão gênica entre os segmentos vizinhos da raiz, os quais compõem exatamente as transições do desenvolvimento do aerênquima. Tais informações podem ser analisadas junto aos dados de fisiologia, já obtidos pelo grupo em trabalhos anteriores.

4.2.1.3 Análise de expressão diferencial

A análise de expressão diferencial entre as amostras da raiz foi realizada com o uso do algoritmo EBSeq (LENG et al., 2013), utilizando como corte de significância o p-valor menor ou igual à 0,05 ($p \leq 0,05$). Essa análise considerou apenas resultados da ferramenta RSEM. Foram comparadas as listas de transcritos expressos do segmento 1 com segmento 2 (S1xS2), segmento 2 com segmento 3 (S2xS3) e segmento 3 com segmento 4 (S3xS4) de raízes de cana-de-açúcar. Dessa forma, foram geradas três listas com a identificação dos genes que estão diferencialmente expressos (DE) entre os pares de segmentos analisados. A Figura 7 ilustra como foram geradas cada uma dessas listas, seguindo a metodologia de segmentação da raiz de cana-de-açúcar (LEITE et al., 2017).

Figura 7: Representação esquemática da metodologia empregada na análise de expressão diferencial para os segmentos da raiz de cana-de-açúcar. Segmentos (S) de 1 cm de raiz, a partir do ápice da raiz sentido à base da mesma até o quarto centímetro. As setas pretas marcam a transição entre os segmentos e a forma a qual cada análise de expressão diferencial foi aplicada. Listas de transcritos diferencialmente expressos (DE) foram geradas primeiro considerando a expressão do primeiro com o segundo segmento (S1xS2), em seguida do segundo com o terceiro segmento (S2xS3) e por fim do terceiro com o quarto segmento (S3xS4).

4.2.2 Anotação de transcritos de biossíntese e degradação da parede celular

Para a identificação funcional dos transcritos observados em cada segmento da raiz, foi utilizado o arquivo de anotação do próprio SUCEST, que inclui anotação de transcritos para, entre outras categorias funcionais (VETTORE et al., 2003), também para parede celular (LIMA et al., 2001).

Paralelamente, utilizou-se a ferramenta Blast2GO (CONESA et al., 2005) para recuperar informações de cada transcrito sobre o nome do processo biológico que está envolvido, sua função molecular e componente celular. Também para a proteína codificadora destes genes, foram recuperadas informações sobre o código de

enzima e informações gerais de sua assinatura, como os modelos preditivos da estrutura, identificação de domínios importantes e a anotação funcional em famílias.

A estrutura da parede celular é composta por uma arquitetura de várias unidades de polissacarídeos, atuando na biossíntese e degradação de domínios da parede celular. Além de elementos envolvidos em outras funções que não estas e que podem ser classificados como sendo do metabolismo de parede celular. Pode-se entender domínio da parede celular, como sendo um elemento componente da sua estrutura, são eles: hemiceluloses, celulose, pectinas, lignina e expansinas.

Buscando identificar e mapear, ao longo dos segmentos da raiz de cana-de-açúcar, os transcritos envolvidos com cada um destes domínios, aplicou-se em colaboração com os alunos Arthur Cambler e David Amorim, uma etapa de anotação dos transcritos para os respectivos domínios. Os transcritos foram obtidos a partir da análise de RNA-Seq e as nomenclaturas utilizadas foram: H para hemiceluloses, C para celulose, P para pectinas, L para lignina e E para expansinas, além de domínios compostos como H/P e H/C, que são domínios que podem agir em ambos domínios envolvidos. Foram considerados apenas aqueles possivelmente envolvidos no processo de biossíntese e degradação da parede.

A anotação manual dos domínios da parede celular foi realizada com base em duas outras anotações, uma anotação do SUCEST e uma segunda anotação gerada com a ferramenta [Blast2go](#), ambas para todos os genes referentes à parede celular.

Diante disso, para os transcritos anotados da mesma maneira em ambas anotações, uma busca na literatura foi realizada para a identificação da função da enzima correspondente e com isso, a enzima foi atribuída a um domínio.

Entretanto, em muitos casos as fontes de informação apresentavam classificações conflitantes ou insuficientes, tornando impossível a sua classificação em domínios. Nestes casos, aplicou-se duas buscas. A primeira delas levou em consideração o código da enzima (*Enzyme Commission numbers – EC numbers*),

verificando a existência na base de dados [ExplorEnz](#) para números EC completos, ou então verificando a base [Enzyme Nomenclature](#) para números parciais. O segundo tipo de busca foi realizado utilizando as sequências de aminoácidos correspondentes à cada transcrito, contra os bancos de dados referentes a proteínas, [UniProt](#) e [InterPro](#) e de domínios conservados ([NCBI Conserved Domains](#)).

Os resultados obtidos pelos dois tipos de busca foram então confrontados com as classificações do Blast2go e do SUCEST para confirmação de alguma destas classificações. Entretanto, em muitos casos, mesmo após essas duas buscas, não foi possível confirmar qual era a enzima em questão. Para mostrar se foi possível ou não classificar os transcritos e quais foram as análises feitas as análises dos transcritos foram categorizadas desta forma: *T-True*, para os transcritos classificados diretamente a partir das informações do Blast2go, do SUCEST e das informações adicionais ou com auxílio do EC number; *TW-True after Web*, para os transcritos classificados após as buscas nos sites supracitados (UniProt, NCBI-CD e InterPro) e *DW-Doubt after Web*, para os casos em que mesmo após a busca nos sites não foi possível identificar o transcrito. Feitas todas essas etapas, foi obtida a lista com a classificação dos transcritos em domínios.

A etapa automatizada ocorreu pela criação e execução de scripts personalizados utilizando Perl como linguagem de programação e auxílio da ferramenta [BLAST+](#), que permite a execução de alinhamentos de sequências em ambiente local.

4.3 Aplicação da ferramenta SIT em dados de aerênquima em raízes de cana-de-açúcar

A SIT foi aplicada para os dados gerados no estudo sobre o desenvolvimento do aerênquima em raízes de cana-de-açúcar, utilizando dados anotados na categoria de parede celular, o qual foi executada sob dois principais contextos, todos tendo sido gerados segundo o mesmo desenho experimental (GRANDIS, 2015; LEITE et al., 2017; TAVARES, 2015). O primeiro contexto foi para analisar a integração entre os dados moleculares disponíveis e para um mesmo segmento, neste caso os dados de transcritômica recém-quantificados, além da proteômica e da atividade enzimática. No segundo contexto considerou-se uma análise de integração ao longo dos segmentos, buscando avaliar a ocorrência dos elementos.

Adotou-se no trabalho que, cada elemento (p.ex. transcrito ou proteína) de uma *layer* é chamado de identificador (ID). Dessa forma, considerando a ocorrência dos identificadores nas *layers*, eles podem ocorrer em uma ou mais *layers* dentro de uma mesma amostra.

A primeira integração para os níveis de regulação molecular, foi aplicada considerando por um lado, os dados totais para transcritômica, proteômica e enzimas (Tabela 2-a). Por outro lado, os resultados da análise de expressão gênica diferencial, calculados para as transições entre os segmentos da raiz, também foram considerados (Tabela 2-b).

Embora identificados em estudos individuais, os dados destas quatro *layers* ainda não tinham sido analisados em conjunto, a fim de encontrar os principais fatores que contribuem para a formação do aerênquima. Os resultados individuais quanto às características funcionais do transcritoma são contemplados no tópico resultados desta tese.

Tabela 2: Metodologia de execução de SIT em dados de aerênquima de raiz de cana-de-açúcar. a) execução por segmento da raiz, considerando a lista de todos os transcritos expressos no RNA-Seq como entrada da layer1, todas as proteínas expressas na proteômica na layer2, todos os elementos de enzimas de degradação da parede celular já quantificados na layer3. Os valores das réplicas da expressão gênica do RNA-Seq de cada segmento foram utilizados para o cálculo da correlação. b) Seis diferentes execuções de SIT consideraram para cada uma, a lista de transcritos diferencialmente expressos entre os segmentos vizinhos como entrada da layer1, todas as proteínas do segmento em questão na layer2, todas os IDs das proteínas disponíveis na layer3. Aqui, a rede foi ponderada com base no valor de expressão apenas de transcritos diferencialmente expressos. Em a) e b) o valor da expressão gênica foi utilizado como escala do tamanho do nó da rede.

a) RNA-Seq total de parede celular				
Segmento	Proteína	Enzima	Valor de expressão gênica (tamanho do nó)	Correlação (peso da aresta)
S1	S1	S1	S1	S1
S2	S2	S2	S2	S2
S3	S3	S3	S3	S3
S4	S4	S4	S4	S4

b) Transcritos diferencialmente expressos de parede celular				
Condição	Proteína	Enzima	Valor de expressão gênica (tamanho do nó)	Correlação (peso da aresta)
S1_S1xS2	S1	S1	S1	S1
S2_S1xS2	S2	S2	S2	S2
S2_S2xS3	S2	S2	S2	S2
S3_S2xS3	S3	S3	S3	S3
S3_S3xS4	S3	S3	S3	S3
S4_S3xS4	S4	S4	S4	S4

A informação resultante da anotação de domínios para biossíntese e degradação da parede celular, com a anotação funcional do SUCEST, foram utilizados como atributos de descrição de cada nó na construção da visualização gráfica dos dados. Uma informação adicional para o nó.

A visualização da integração é feita na forma de redes complexas, sendo

assim, o resultado da integração foi utilizado como cor dos nós da rede. Já as arestas (responsáveis por ligar os nós nas redes) podem ser construídas com base nos interesses do usuário. Neste trabalho aplicamos a análise de correlação de Pearson para identificar transcritos co-expressos em cada segmento e assim construir redes ponderadas.

A análise de correlação foi realizada utilizando o módulo específico da ferramenta SIT. Foram aplicados filtros de significância para p-valor menor ou igual à 0.05 e filtro pelo valor absoluto do coeficiente de correlação, desde que maior que 0.9. O arquivo com os pares de correlação entre os transcritos foi utilizado como base para a construção das redes de co-expressão, as quais foram geradas e analisadas quanto às propriedades das redes, na ferramenta de visualização gráfica.

Após as redes geradas, aplicou-se uma análise de propriedade das redes para as condições apresentadas nos itens a e b da Tabela 2. Apenas para os dados totais por cada segmento (Tabela 2-a), aplicou-se uma nova análise de propriedades de redes, que utilizou apenas elementos de degradação da parede celular, tendo como base os resultados da etapa de anotação de domínios da parede celular.

O segundo contexto de execução de SIT considerou os resultados da primeira etapa de execução de dados moleculares, filtrando apenas as listas de identificadores que integraram nas *layers* de transcritômica, proteômica e enzimas, ou seja, que ocorrem nessas três *layers*. Essa etapa foi realizada separadamente para os itens a e b da Tabela 2. A ideia aqui foi verificar, por segmento da raiz, a intersecção entre os identificadores e assim identificar identificadores comuns e exclusivos por nível de regulação biológica.

4.4 Microtomografia de segmentos da raiz de cana-de-açúcar

Adicionalmente às análises moleculares e fisiológicas já obtidas pelo nosso laboratório, foi realizada uma análise de microtomografia para os 5 centímetros de segmentos de raiz considerados no início do estudo. A microtomografia foi realizada no aparelho escaner SkyScan 1173 de microtomografia (micro-CT), disponível no Instituto de Biociências da Universidade de São Paulo. A raiz foi escaneada em cortes de 9µm.

4.5 Desenvolvimento de um banco de dados

Para organizar todos os dados que estão sendo utilizados neste trabalho (transcritômica, proteômica e fisiologia), foi desenvolvido um banco de dados local, cuja implementação utilizou MySQL como Sistema Gerenciador de Banco de Dados (SGBD). Este banco está inicialmente disponível para o Lafieco. Contudo, pretende-se futuramente disponibilizá-lo à comunidade científica como parte da contribuição desta tese.

5 Resultados

5.1 Ferramenta SIT para integração e visualização de dados

Visando integrar os dados gerados por tecnologias em larga escala, desenvolveu-se a ferramenta SIT, que consiste em um programa com interface gráfica, criada para a integração e visualização dos dados, esta última feita com base na teoria de redes complexas. Adicionalmente, SIT traz um módulo para a análise de correlação, onde os dados para a correlação podem ou não ter natureza genética. A Figura 8 ilustra as principais etapas e ordem de execução da ferramenta SIT. Abaixo são apresentadas as características e possibilidades de execução de SIT.

Em SIT, a integração de dados propriamente dita pode ser aplicada para até seis *layers* e está disponível na primeira aba da ferramenta (do inglês: *Data Integration*) (Figura 9). Como padrão a ferramenta já mostra campos para a inclusão de duas *layers* (p ex. Transcritômica e Proteômica), mas o usuário pode incluir quantas desejar adicionando-as pelo botão *Add Layer*. O arquivo a ser carregado deve estar em formato de texto, mas apresentar os dados em colunas separadas por tabulação ou vírgula. Cada *layer* deve ser configurada com nome da *layer* no campo *Layer name*, nome do arquivo com os dados, adicionado no campo *Upload file*. Há um botão chamado *Choose integration_id*, onde o usuário pode visualizar o arquivo carregado e escolher a coluna correspondente ao identificador que será utilizado para a integração das *layers*. Por fim, existe um campo de preenchimento para o estado dessa *layers*, que configura se é fonte (do inglês: *source*) ou alvo (do inglês: *Target*). Na existência de mais de duas *layers*, podem existir alvos de 1 a 5.

Figura 8: Fluxograma da ferramenta SIT. Como entrada podem existir até seis layers de dados. A primeira etapa é de integração, realizada localmente e via interface gráfica. Ainda nesta etapa, o usuário configura a forma que deseja visualizar a rede, podendo ou não utilizar redes ponderadas. A segunda etapa é para visualização e edição dinâmica das redes geradas na etapa anterior. Isso é feito em ambiente online, apenas carregando o arquivo gerado da integração. Nessa interface o usuário pode aplicar novas análises e calcular as propriedades das redes, como análises de centralidades. Entre as diversas características de SIT, é possível alterar o layout de visualização e poder identificar agrupamentos de co-expressão gênica. Como saída da ferramenta, o usuário pode salvar as redes geradas e configuradas na forma de imagem com extensão .png, além de arquivos em formato texto que são gerados a cada nova análise.

Essa informação de dependência entre as *layers* é gravada no arquivo de integração, mas só é interpretada se o usuário configurar uma rede de regulação (com arestas direcionadas), o que é feito na segunda aba de SIT, na configuração da

rede. Caso contrário, o usuário não precisa se preocupar com essa configuração e a informação apenas segue no arquivo.

Figura 9: Tela principal da ferramenta SIT para a integração de dados de larga escala. A ferramenta permite análises de integração de até seis layers, sendo por padrão iniciada já com duas. Essa é a primeira aba de SIT, que realiza a integração das layers, gerando o arquivo de integração dos dados. A cada nova execução de integração, SIT cria um novo diretório, onde são adicionados todos os arquivos gerados. O nome do diretório pode ser personalizado pelo usuário no campo “Enter experiment name” da interface. O arquivo gerado aqui é utilizado na segunda etapa (segunda aba), onde a rede é gerada a partir dos dados de integração.

Após incluir os dados correspondentes a cada *layer*, o usuário deve incluir um nome de trabalho dessa execução no campo *Enter experiment name*, nome este que permita diferenciar cada execução da ferramenta.

A cada nova execução da ferramenta, um novo diretório é criado no local de instalação da mesma, com o prefixo “*SystemsIntegrationTool_outputs*”. Caso já

exista outra pasta com este mesmo nome, um sufixo numérico é adicionado para diferenciar, por exemplo em “*SystemsIntegrationTool_outputs_1*”.

Nesta pasta são incluídos todos os resultados de cada análise, seja ela de integração, visualização ou correlação, onde cada um ficará destacado com um nome específico. Resultados da integração são incluídos em um diretório com prefixo “*Layers_Integration_*” seguido do nome dado pelo usuário. A visualização dos dados é adicionada também dentro dessa pasta, com o sufixo “*_sit_output.txt*”. O usuário pode executar a integração de dados uma única vez, esta etapa gera um arquivo a cada nova execução, para posteriormente, utilizar este arquivo gerado para explorar quantas vezes desejar pelo carregamento na interface de visualização, descrita a seguir.

A segunda aba de SIT é para a configuração da visualização dos dados (do inglês: *Setting Data Visualization*) (Figura 10), que é feita na forma de redes complexas, formata as informações de atributos dos nós através da caixa *Node description*, onde o usuário faz o carregamento do arquivo. Para que existam conexões entre os nós (i.e. arestas), o usuário deve incluir um outro arquivo com essas informações, na caixa *Edge Weighting* e dizendo se a rede será ou não ponderada (i.e. com pesos nas arestas). E caso afirmativo o usuário pode incluir o resultado da análise de correlação, por exemplo, ou apenas fonte e alvo, em uma análise de regulação, onde existem dependências entre os nós.

Uma outra característica implementada é o direcionamento das arestas, podendo ser direcionada ou não, dependendo do interesse do usuário. Por fim, pode-se formatar também o nome do arquivo que será criado no campo “*Enter a name to create the output file*”.

A terceira aba da ferramenta (Figura 11) é a implementação da análise de correlação, que pode ser realizada para construir o arquivo de relacionamentos da rede. Embora este módulo esteja implementado na ferramenta, ele é opcional na

execução de SIT. É possível que o usuário implemente outra métrica e até mesmo em outro ambiente de programação para a identificação das medidas de similaridade, como a entropia por exemplo. Mas independente da métrica escolhida, é necessário que o resultado seja importado na aba de configuração dos dados para visualização. O formato desse arquivo de entrada deve ser tabular, com o nome dos pares da rede na primeira e segunda colunas, respectivamente, deixando o valor da relação na terceira coluna. É possível que esse valor seja zero.

Figura 10: Visualização da ferramenta SIT, com destaque à aba de configuração da visualização dos dados, o que é feito na forma de redes complexas. Nessa interface é possível carregar os arquivos correspondentes à descrição do nó na rede, bem como do arquivo de pares de elementos que construirá a rede. Esse arquivo pode ser configurado pelo usuário ou ser gerado como resultado de uma análise de correlação (arestas ponderadas e não direcionadas) ou regulação (arestas direcionadas e não necessariamente ponderadas). O usuário pode ainda escolher se deseja visualizar uma rede de arestas direcionadas ou não. Também para cada execução realizada, um novo arquivo é criado e seu nome pode ser personalizado.

Embora opcional, o módulo de correlação da SIT está contemplado para calcular o método de Pearson e Spearman com possibilidade para corte de significância (do inglês: *p-value*) para valores menores que 0,05 (significante) e 0,01 (altamente significativa). Além disso, é possível também filtrar os valores de correlação de acordo com um limiar (do inglês: *threshold*), neste caso para valores maiores ou igual ao valor da correlação em módulo e para os seguintes intervalos: 0,5, 0,9 e 0,99 ($|corr| \geq 0,5$, $|corr| \geq 0,9$, $|corr| \geq 0,99$).

A ferramenta de integração de dados foi planejada para permitir a análise de integração e também a visualização em si (Figura 12). A visualização é feita em ambiente HTML, disponível em <http://vision.ime.usp.br/~amanda/SIT/Default.html> e permite de forma interativa a aplicação de novas análises.

Após carregar o arquivo gerado pela integração, o mesmo será carregado na página inicial, sendo possível a aplicação de diferentes topologias de rede, como por exemplo a randômica (do inglês: *random*), grade (do inglês: *grid*) e circular (do inglês: *circle*). Bem como mostrar ou ocultar os rótulos (do inglês: *labels*) dos nós e arestas.

O usuário pode então ir aplicando filtros que permitam a redução por identificadores de interesse. Sugerimos o início com o layout Cose, que permite a visualização de grupos de redes.

Existe uma barra de busca na aba superior de ferramentas, que procura pelo nome do ID ou da informação contida no atributo do nó. O sistema permite a inclusão do caracter * (asterisco) como expressão regular para indicar que a busca é por tudo que contiver este texto. Por exemplo, *Degradation** retornará todos os IDs cuja a palavra *Degradation* aparece contido no nome da anotação. Para cada um, será mostrado também os seus primeiros vizinhos.

Ao clicar em um nó o sistema apresenta algumas opções de informações e edições do mesmo. O usuário pode visualizar o nome do ID, na condição de uma

rede com ponderação nos nós, como por exemplo o valor de expressão de um transcrito, este valor é mostrado no campo *Weighting*. Este valor também configura o tamanho do nós, portanto para modificar o tamanho de um nó, basta alterar o valor desse campo. São mostradas também os nomes de cada um dos primeiros vizinhos desse nó, sendo permitido ocultar ou mesmo remover completamente alguns deles ou todos. O usuário pode desejar ainda, ocultar ou remover o nó em questão, ou ainda apenas mostrar esse nó e ocultar todos os outros da rede.

Figura 11: Visualização da ferramenta SIT, com destaque à aba para análise de correlação. SIT depende de um arquivo que construa as interações das redes complexas, de forma que disponibilizamos o método de correlação de Pearson, com opções de filtros para seleção de resultados mais e menos estridentes, através do p-valor e valor do coeficiente de correlação.

Figura 12: Interface online de SIT para visualização e exploração das redes de integração geradas. A figura mostra um exemplo de visualização de uma rede integrada com layout randômico. A informação da integração de layers está atribuída ao nó enquanto as arestas correspondem à correlação de expressão gênica, com relacionamentos positivos (arestas verdes) e negativos (arestas vermelhas). As cores dos nós correspondem às layers integradas e são editáveis nos botões da ferramenta. Na parte superior e lateral direita, existem botões funcionais para a manipulação da rede. Essa interface está disponível em: <http://vision.ime.usp.br/~amanda/SIT/Default.html>

É possível filtrar os identificadores por quantidade de layers que eles ocorrem, além do filtro por identidade da layer, que inclui a edição da paleta de cor, ou seja, o usuário pode querer visualizar apenas ID que ocorrem em, pelo menos, duas layers e que seja a primeira e a segunda layer.

Além disso, outras funcionalidades foram implementadas na visualização dos dados, como a análise de propriedades de redes, a citar o *degree* e o *betweenness*. O usuário pode assim, calcular as centralidades para um identificador pontual ou em uma análise global, identificar as centralidades do sistema, permitindo a identificação de propriedades emergentes do sistema.

Há também um botão para seleção de elementos da rede, onde é possível selecionar um ou mais nós e arestas e movê-los, removê-los ou então deixar apenas

a seleção visível.

Na lateral direita da página de visualização há uma aba de ferramentas com mais algumas funcionalidades, a saber, o filtro dos valores das arestas (*Filter Edge*), permitindo a visualização apenas de correlações positivas e/ou negativas e cortes para os valores maiores ou iguais em módulo que $corr \geq |0,5|$ $corr \geq |0,6|$ $corr \geq |0,7|$ $corr \geq |0,8|$ $corr \geq |0,9|$.

Embora o usuário tenha configurado a sua rede para ser direcionada, é possível através da função *Edge directions* (também na aba lateral), que seja alterada essa informação.

Como geralmente a análise de ômicas inclui grandes conjuntos de dados, foi desenvolvido um painel de zoom para ajuste da escala de visualização (do inglês: *Navigator*), que pode ser utilizada como suporte para cada nova interação com a ferramenta.

A ferramenta possui um botão para exportação de figuras (formato png) de redes, geradas pela ferramenta e que permite ao usuário arquivar a figura produzida da análise. Neste sentido, é muito importante documentar as etapas aplicadas para obtenção da rede final. Para isso, existe na aba lateral de ferramentas, um campo interativo de texto chamado *Tracking*, que registra cada função executada pelo usuário. É permitido edições nessa caixa, bem como o usuário pode selecionar o seu conteúdo e salvar em um novo arquivo de texto.

Por fim, o usuário pode desejar reiniciar a rede, através do botão *reset*, eliminando toda e qualquer alteração realizada, como também acessar ajuda no botão *Help* e ter acesso ao Manual de utilização da ferramenta.

5.2 Transcritoma de segmentos de raízes de cana-de-açúcar

5.2.1 Processamento das reads do RNA-Seq

O relatório final da análise de qualidade das *reads*, gerado pela ferramenta FASTQC para cada segmento de raiz, mostrou trechos de baixa qualidade nas reads, como exemplificado pelos dados contidos na faixa horizontal vermelha da Figura 13, correspondente à amostra S1A. Na faixa amarela aquelas com qualidade média e na verde estão as sequências de qualidades consideradas ótimas.

Figura 13: Representação de uma qualidade ruim de reads do sequenciamento. Imagem gerada pela ferramenta FASTQC para a amostra S1A de raiz de cana-de-açúcar. Eixo x é a posição das bases nucleotídicas nas reads sequenciadas, enquanto eixo y mostra a margem de qualidade phred das bases. As barras são as leituras do sequenciamento para cada base da read. A faixa horizontal em verde representa uma ótima qualidade phred das bases, enquanto a faixa cor laranja, na região intermediária, representa qualidade média das bases. A faixa em vermelho representa um trecho de baixa qualidade das bases. A figura mostra a existência de trechos de sequência com qualidade ruim e que devem ser eliminadas antes de análises posteriores.

Contudo, depois de filtrar apenas os resultados cujas bases tinham valor *Phred* igual ou maior à 20 e após isso, recuperar apenas as sequências de comprimento maior ou igual à 40 nt, uma nova execução da ferramenta FASTQC foi realizada para cada amostra buscando verificar o novo padrão de qualidade das sequências gerado pelo filtro. Como resultado dessa análise, todas as amostras (de S1 até S4) incluindo as réplicas, tiveram a qualidade corrigida. A Figura 14 exemplifica uma boa qualidade para as *reads*. Agora as *reads* pré-processadas estão prontas para serem quantificadas com maior confiabilidade.

Figura 14: Representação de uma boa qualidade para reads do sequenciamento. Imagem gerada pela ferramenta FASTQC para a amostra S1A de raiz de cana-de-açúcar. Eixo x é a posição das bases nucleotídicas nas reads sequenciadas, enquanto eixo y mostra a margem de qualidade phred das bases. As barras são as leituras do sequenciamento para cada base da read. A faixa horizontal em verde representa uma ótima qualidade phred das bases, enquanto a faixa cor laranja, na região intermediária, representa qualidade média das bases. A faixa em vermelho representa um trecho de baixa qualidade das bases. A figura mostra que apenas reads com boa qualidade estão presentes nos arquivos.

Com a aplicação de filtros, a quantidade de *reads* iniciais e finais não são as mesmas, como mostra a Tabela 3. Os filtros retiraram aproximadamente 150 milhões

de reads de baixa qualidade, gerando um banco de dados de 503 milhões de *reads*. Qualidade ruim pode interferir negativamente na etapa de montagem das *reads*, por isso é tão importante essa etapa.

Tabela 3: Descrição da quantidade de reads do RNA-Seq por etapa de processamento. As reads do RNA-Seq foram processadas para correção de qualidade phred maior e igual à 20, contudo isso gerou fragmentos menores que 40 nucleotídeos, que por sua vez foram também posteriormente filtrados.

Nº reads	Comprimento da Read	%GC	Taxa de Filtro (%)
<u>Dados brutos</u>			
653.576.011	75	51	-
<u>Filtro de qualidade Phred >=20</u>			
563.699.858	1-75	52	14
<u>Filtro de qualidade Phred >=20, comprimento >=40</u>			
503.483.924	40-70	52	23

5.2.2 RSEM para a quantificação da expressão gênica

A quantidade de transcritos gerados por RSEM foi comparada aos do pipeline Tuxedo buscando selecionar aquele com melhor capacidade de reconstrução dos transcritos de cana-de-açúcar, sendo que a ferramenta RSEM apresentou melhor resultado (Figura 15). Dessa forma, apenas estes resultados foram submetidos à análise de expressão diferencial e utilizados em análises posteriores.

5.2.3 Características gerais observadas no RNA-Seq

A partir da análise do RNA-Seq dos segmentos da raiz, foram identificados 23.458 transcritos únicos, considerando aqueles presentes nas 3 réplicas biológicas em cada segmento. Analisando a distribuição dos transcritos totais ao longo de todos os segmentos (Figura 16), identificou-se 10.574 comuns entre os quatro segmentos. Já uma análise por segmento apontou para S1 a maior taxa de transcritos (21.025), seguido por S3 (17.631), S2 (17.410) e S4 (12.254). Desse total, 2.924 (14%), 602 (3,5%), 852 (4,8%) e 180 (1,46%), são transcritos exclusivos dos segmentos S1, S2, S3 e S4, respectivamente.

Figura 15: Comparação de resultados entre Tuxedo e RSEM. Os resultados da quantificação pela ferramenta RSEM e pipeline Tuxedo foram comparados entre si, a fim de avaliá-las e escolher a um resultado para análises posteriores. No eixo x estão os segmentos de 1 à 4 (S1, S2, S3, S4), mostrando para cada um, os resultados de RSEM e Tuxedo. No eixo y à esquerda são mostradas as quantidades de transcritos. Já o eixo z (à direita) estão descritas as condições, comuns ou exclusivos, de transcritos por metodologia utilizada (RSEM ou Tuxedo) e ainda considerando todas as categorias funcionais e apenas parede celular.

Para cada conjunto de transcritos identificados em cada segmento, foi realizada a anotação funcional (Figura 17), revelando a categoria de parede celular com a maior quantidade de transcritos anotados, seguida por fator de transcrição e metabolismo de proteínas. Seguindo o objetivo do trabalho, uma análise mais específica foi realizada para a categoria de parede celular.

5.2.4 Transcritos de parede celular observados no transcritoma

Analizando os transcritos de parede celular, foram identificados 672, 585, 622 e 418 transcritos no S1, S2, S3 e S4 das raízes, respectivamente. Quando divididos em subcategorias funcionais, estes apresentaram 86, 83, 86 e 76, respectivamente. Nessa categorização foi observado que a enzima glicosil hidrolase da família GH17 foi a que apresentou maior quantidade de transcritos expressos para os três primeiros segmentos. Esta GH17 pertence a um grupo de enzimas de degradação de hemicelulose e β -glucano. Já em S4 a maior quantidade de transcritos foi anotada como celulose sintase, um domínio de enzimas que são responsáveis pela síntese de celulose. A Figura 18 mostra a distribuição dos transcritos de parede celular ao longo dos segmentos de raízes e adicionalmente uma descrição para as duas subcategorias funcionais de parede celular, com maior ocorrência e nas Figuras 27 à 30 do apêndice mostram as 20 primeiras subcategorias que foram mais representadas para os segmentos S1 à S4, respectivamente.

Comparando os segmentos de raiz em uma ordem de desenvolvimento, de S1 para S2 foram observados 3.491 transcritos diferencialmente expressos, enquanto que esse valor reduz na transição de S2 para S3 (601 transcritos) e volta a aumentar de S3 para S4 com 1.165 transcritos diferencialmente expressos. Sendo

que dessas respectivas transições, 169, 42 e 34 dos transcritos diferenciais correspondem apenas à parede celular.

Figura 16: Diagrama de Venn do RNA-Seq para os segmentos de raiz de cana-de-açúcar. Um total de 10.574 transcritos foi apontado sendo comuns a todos os segmentos. Já a distribuição ao longo dos segmentos mostrou S1 com a maior taxa de transcritos (21.025), seguido S3 (17.631), S2 (17.410) e S4 (12.254). Desse total, 2.924, 602, 852 e 180, são transcritos exclusivos dos segmentos S1, S2, S3 e S4, respectivamente. Ferramenta utilizada: [Venn Diagrams](#).

Figura 17: Quantidade de transcritos para as 10 categorias funcionais mais representadas no transcrito para os quatro segmentos da raiz de cana-de-açúcar. No eixo x são mostradas as categorias funcionais observadas para os quatro segmentos. O eixo y mostra a quantidade de transcritos observados. Destaque para a categoria de parede celular como a mais representada no transcrito.

Figura 18: Distribuição de transcritos de parede celular e as duas subcategorias mais representadas ao longo dos quatro primeiros segmentos da raiz de cana-de-açúcar. O diagrama de venn mostra a distribuição dos transcritos ao longo dos segmentos, apontando 354 comuns de todos eles, e por outro lado, 73, 13, 38 e 12 específicos dos segmentos S1, S2, S3 e S4, respectivamente. Adicionalmente, para cada segmento são mostradas as duas subcategorias com maior quantidade de transcritos. Os parênteses mostram a quantidade de transcritos por subcategoria. Ferramenta utilizada: [Venn Diagrams](#).

5.2.5 Domínios de biossíntese, degradação e expansinas da parede celular em cana-de-açúcar

A análise de domínios da parede celular permitiu a anotação de 441, 411, 53 transcritos envolvidos na biossíntese, degradação e anotados como expansinas, respectivamente.

O domínio mais representado foi o de degradação de hemicelulose (184 transcritos), seguido pela biossíntese de hemicelulose (92) e biossíntese de lignina (90). O resultado completo da anotação de domínios para todos os transcritos de

parede celular de cana-de-açúcar está na Tabela 4.

Tabela 4: Distribuição de todos os transcritos de parede celular do SUCEST em domínios de biossíntese, degradação e expansinas. H=hemicelulose, C=celulose, P=pectina, L=lignina, H/C=hemicelulose e celulose, H/P=hemicelulose e pectina.

Domínio	Porcentagem de Transcritos (quantidade)
H_degradação	20(184)
H_biossíntese	10(92)
L_biossíntese	10(90)
P_degradação	9(83)
H/P_biossíntese	9(79)
H/C_biossíntese	8(74)
P_biossíntese	7(64)
Expansin	6(53)
C_biossíntese	5(42)
C_degradação	5(44)
H/C_degradação	4(37)
H/P_degradação	4(37)
L_degradação	3(26)

5.2.6 Domínios funcionais de transcritos diferencialmente expressos

Buscando uma análise mais específica entre os segmentos, considerou-se o conjunto de transcritos que são diferencialmente expressos entre os segmentos, mas dessa vez buscando o perfil funcional daqueles onde o valor de expressão é maior em um ou no outro segmento considerado.

Entre S1xS2 existem 3.490 transcritos, sendo 169 de parede celular. Considerando a lista de transcritos de parede celular diferencialmente expressos e com maior valor de expressão em S1 (72), observa-se uma predominância na quantidade de transcritos anotados funcionalmente como biossíntese de hemicelulose com 8 transcritos (SCBGLR1002A07.g, SCEZHR1049A07.g, SCQGLB1039H03.g, SCCCCL4002D03.g, SCCCLB1C06E01.g, SCSBRZ3121G09.g, SCVPAM1058E05.g, SCCCLR1048B04.g) e degradação de hemicelulose com 7 transcritos (SCEPAM1023F03.g, SCRFLR1055B07.g, SCRLLR1059E04.g, SCRLSB1042A05.g, SCEZRZ3015A12.g, SCBFRT1068E05.g, SCCCRT1002F03.g). O terceiro domínio mais representado é o de transcritos que podem estar envolvidos na degradação tanto de ambos hemicelulose quanto de pectina, com seis transcritos (SCCCCL6004H07.g, SCCCST1006B07.g, SCQSFL3033B04.g, SCVPAM1059F06.g, SCVPRZ2037A02.g, SCCCCL4009F05.g).

Já a lista de transcritos com maior valor de expressão em S2 (97), observou-se os maiores números de transcritos nos domínios de biossíntese de lignina com 12 transcritos (SCJFRT1061B07.g, SCJLRT1023B09.g, SCRFLR1012F12.g, SCJLRT2050C09.g, SCQSRT1034C09.g, SCCCZRZ2C01A04.g, SCEPSB1135A04.g, SCSFFL4084B02.b, SCCCZRZ2C01F05.g, SCEPRZ1011A02.g, SCSBAD1054A07.g,

SCQSRT1036E09.g), biossíntese de celulose com 11 transcritos (SCAGLR1064F05.g, SCCCLR1065F07.g, SCCCLR2C02H01.g, SCCCZR2002G07.g, SCEQLR1029D03.g, SCJFLR1017B09.g, SCJLLR1106H06.g, SCRLST3163A02.g, SCSBSB1053C05.g, SCVPCL6064E05.g, SCVPRT2075E11.g) e biossíntese de hemicelulose com 10 transcritos (SCCCLB1026G03.g, SCEPRZ1009A11.g, SCJFLR1073B10.g, SCJLLR1033F06.g, SCQGFL4080C07.g, SCEZRT2017G09.g, SCUTSD2027H09.g, SCVPHR1092E06.g, SCBFRZ2017F10.g, SCCCST3006B11.g).

Na transição S2xS3, um total de 601 transcritos estão diferencialmente expressos, sendo 42 de parede celular.

Considerando os transcritos de parede celular e mais expressos em S2 (27), observou-se um predomínio das categorias de biossíntese de celulose com quatro transcritos (SCACLR1036F04.g, SCCCCL3002B03.b, SCCCLR1066B10.g, SCCCZR2002G07.g), seguido de uma sequência de domínios com três transcritos cada, sendo eles: degradação de ambos hemicelulose e pectina (SCBFLR1083E01.g, SCSFFL3091E06.g, SCCCCL4009F05.g), biossíntese de hemicelulose (H_biossíntese) (SCUTSD2027H09.g, SCCCLR1048B04.g, SCUTSB1030B12.g) e degradação de hemicelulose (SCCCTR1002G03.g, SCBFLR1046E10.g, SCJLST1020H07.g).

Em S3 observou-se o predomínio de transcritos relacionados à biossíntese de hemicelulose com cinco transcritos (SCJLLR1033F06.g, SCEZRT2017G09.g, SCVPHR1092E06.g, SCCCLR1065A04.g, SCEZLR1009C10.g), degradação de celulose (SCCCTR1C05G04.g, SCCCSD1089A09.g, SCVPRZ2042G03.g) e a biossíntese de lignina (SCCCZR2C01A04.g, SCJLRT2050F04.g).

Já a transição S3xS4, observou-se 1.164 transcritos diferencialmente expressos, sendo 34 destes apenas de parede celular. O domínio mais representado em S3 foi a biossíntese de celulose com seis transcritos (SCCCCCL3001G09.b,

SCCCLR1066B10.g, SCCCLR2C02H01.g, SCEQLR1029D03.g, SCJLLR1106H06.g, SCSBSB1053C05.g), biossíntese de hemicelulose com cinco transcritos (SCEQLB1066G04.g, SCEZLR1009C10.g, SCEZRT2017G09.g, SCQSST1040A06.g, SCVPHR1092E06.g) e degradação de hemicelulose com dois transcritos (SCCCRT1002G03.g, SCEQRT1025C10.g).

Já para S4 apenas dois domínios estavam mais expressos, sendo biossíntese de hemicelulose com um transcrito (SCUTSD2027H09.g) e degradação de celulose também com um transcrito (SCCCSD1089A09.g).

5.3 Aplicação da ferramenta SIT em dados de cana-de-açúcar

5.3.1 Quantificação geral da integração de dados

A ferramenta SIT foi aplicada a um conjunto de dados biológicos especialmente já anotados para a categoria funcional de parede celular. Diferentes aplicações foram realizadas utilizando dados de cana-de-açúcar.

Como esperado, a *layer* de transcritômica foi para todos os segmentos, a que mais apresentou elementos. O resultado completo descrevendo a ocorrência dos identificadores SAS por *layer* analisada, está no arquivo *intersectedIds.txt*, criado para cada execução da integração de dados, bem como compilado na Tabela 5.

Tabela 5: Compilação dos resultados de SIT para transcritômica, proteômica e enzimas gerados no estudo de desenvolvimento do aerênquima de raízes de cana-de-açúcar, considerando apenas elementos da categoria funcional de parede celular. a) execução por segmento da raiz. b) Seis diferentes execuções de SIT consideraram para cada uma, a lista de transcritos diferencialmente expressos entre os segmentos vizinhos. Em a) e b) são mostradas as quantidades de transcritos que foram comuns na integração de três (3), duas (1|2, 1|3, 2|3) e apenas um layer (1, 2 e 3). Com exceção de transcritos exclusivos de uma única layer, destacamos a maior quantidade de integrações ocorridas para as layers 1|2 tanto em a) visto em S1, quanto em b), visto nas condições S1_S1xS2 e S1_S1xS2 de b).

a) RNA-Seq total de parede celular							
Segmento	3 layers	2 layers			1 layer		
		1 2	1 3	2 3	1	2	3
S1	19	101	34	3	517	16	33
S2	17	89	36	5	442	28	31
S3	17	97	38	5	469	22	29
S4	16	73	20	9	308	52	44

b) Transcritos de parede celular diferencialmente expressos							
Condição	3 layers	2 layers			1 layer		
		1 2	1 3	2 3	1	2	3
S1_S1xS2	12	41	9	10	107	76	58
S2_S1xS2	12	41	9	10	107	76	58
S2_S2xS3	3	10	4	19	25	107	63
S3_S2xS3	3	10	4	19	25	109	63
S3_S3xS4	0	6	1	22	27	113	66
S4_S3xS4	0	6	1	25	27	119	63

Novamente utilizando a ferramenta SIT, agora para as listas de elementos presentes nas *layers* de transcritômica, proteômica e atividade enzimática, verificou-se a intersecção e especificidade dos mesmos por segmento da raiz, a qual foi realizada mediante uma nova execução de SIT, agora sob o contexto de analisar a integração entre os segmentos da raiz.

Inicialmente considerando todo o RNA-Seq, observou-se 20 elementos únicos dos quais, 15 ocorrem em todos os quatro segmentos da raiz, apenas um para cada uma das integrações de S1,S2,S3; S1,S3; S1 e S4 (Tabela 6). Já considerando apenas os transcritos diferencialmente expressos tem-se três ocorrendo nas transições de S1XS2, S2XS3 e S3XS4, enquanto os nove restantes são comuns das transições de S1XS2 e S2XS3 (Tabela 7).

5.3.2 Propriedade das redes de co-expressão integradas

Uma análise de propriedades das redes foi aplicada para cada resultado de integração obtido com SIT e descritos na Tabela 5. Foram analisados os parâmetros relacionados à conectividade e centralidades (*degree betweenness*) das redes.

Embora complexa, a visualização dos dados na forma de redes permite observar a sua conectividade. Para todos os transcritos de parede celular, a conectividade das redes mostrou 6.958, 4.848, 5.905 e 2.589 conexões para os segmentos S1, S2, S3 e S4, respectivamente. Considerando a existência de 559, 493, 510 e 403 nós existentes para os mesmos quatro segmentos, respectivamente (Figura 19). O resultado da conectividade foi obtido dos respectivos arquivos gerados pela ferramenta SIT.

Uma outra característica recuperada das redes foi a formação de grupos de co-expressão gênica. Para essa análise considerando cada segmento individualmente, observou-se a formação de apenas um grande grupo de conexões, conforme mostrado na Figura 20.

Tabela 6: Ocorrência de IDs compartilhados e exclusivos, de parede celular para todo o RNA-Seq, integrados nas layers de transcritômica, proteômica e enzimas, ao longo dos quatro segmentos (S) da raiz de cana-de-açúcar.

Segmentos	SAS_IDs	Domínio Parede Celular	Anotação Funcional
1 2 3 4	SCAGLR1021F11.g	C_degradação	Glicosil hidrolase família 1
1 2 3 4	SCCCCL3001B10.b	C_degradação	Glicosil hidrolase família 1
1 2 3 4	SCCCST1002E03.g	C_degradação	Glicosil hidrolase família 1
1 2 3 4	SCEQHR1082B01.g	C_degradação	Glicosil hidrolase família 1
1 2 3 4	SCEQLB1066E08.g	C_degradação	Glicosil hidrolase família 1
1 2 3 4	SCJFLR1017E03.g	C_degradação	Glicosil hidrolase família 1
1 2 3 4	SCPIRT3022G07.g	C_degradação	Glicosil hidrolase família 1
1 2 3 4	SCVPRZ2042G03.g	C_degradação	Glicosil hidrolase família 1
1 2 3 4	SCQSRT2034G08.g	H_degradação	α -galactosidase
1 2 3 4	SCCCCL4001H11.g	H/P_degradação	β -galactosidase
1 2 3 4	SCCCCL4009F05.g	H/P_degradação	β -xilosidase/ α -L-arabinofuranosidase
1 2 3 4	SCCCCL6004H07.g	H/P_degradação	β -galactosidase
1 2 3 4	SCJLRT1017C05.g	H/P_degradação	α -L-arabinofuranosidase
1 2 3 4	SCQSRT1034D03.g	H/P_degradação	α -L-arabinofuranosidase
1 2 3 4	SCVPLB1020G05.g	H/P_degradação	α -L-arabinofuranosidase
1 2 3	SCEPRT2044F09.g	H/P_degradação	β -galactosidase
1 2	SCSFRT2067G04.g	C_degradação	Glicosil hidrolase família 1
1 3	SCSFSB1101E12.g	H/P_degradação	β -galactosidase
4	SCACLV1022D02.g	C_degradação	Exoglucanase
1	SCQSAM1030G04.g	H/P_degradação	β -xilosidase/ α -L-arabinofuranosidase (GH família 3)

Tabela 7: Relação dos IDs de parede celular que foram observados nas layers de transcritômica, proteômica e enzimas, agora mostrando a integração ao longo das listas e diferencialmente expressos para as transições de segmentos da raiz de cana-de-açúcar. As condições 1, 2 3 são para S1xS2, S2xS3 e S3xS4, respectivamente. Entre as informações mostradas para cada identificador, está a anotação de domínio funcional de parede celular, resultado de etapas anteriores.

Condição	SAS_IDs	Domínio Parede Celular	Anotação Funcional
1 2 3	SCCCCL3001B10.b	C_degradação	Glicosil hidrolase família 1
1 2 3	SCCCCL4009F05.g	H/P_degradação	β -xilosidase/ α -L-arabinofuranosidase (GH família 3)
1 2 3	SCVPRZ2042G03.g	C_degradação	Glicosil hidrolase família 1
1 2	SCAGLR1021F11.g	C_degradação	Glicosil hidrolase família 1
1 2	SCCCCL4001H11.g	H/P_degradação	β -galactosidase
1 2	SCCCCL6004H07.g	H/P_degradação	β -galactosidase
1 2	SCEQHR1082B01.g	C_degradação	Glicosil hidrolase família 1
1 2	SCEQLB1066E08.g	C_degradação	Glicosil hidrolase família 1
1 2	SCJFLR1017E03.g	C_degradação	Glicosil hidrolase família 1
1 2	SCJLRT1017C05.g	H/P_degradação	α -L-arabinofuranosidase
1 2	SCQSRT1034D03.g	H/P_degradação	α -L-arabinofuranosidase
1 2	SCVPLB1020G05.g	H/P_degradação	α -L-arabinofuranosidase

Figura 19: Conectividade das redes de co-expressão geradas por SIT para cada um dos segmentos (S1 à S4) de raiz de cana-de-açúcar. Os resultados apontaram 6958, 4848, 5905 e 2589 conexões para os segmentos S1, S2, S3 e S4, respectivamente. Considerando a existência de 559, 493, 510 e 403 nós existentes para os mesmos quatro segmentos, respectivamente. Embora com aproximadas quantidade de nós para os segmentos, observou-se um padrão de conectividade decrescente seguindo S1, S3, S2 e S4.

Uma segunda análise de conectividade foi realizada agora para os resultados de SIT, que consideram as listas de transcritos diferencialmente expressos. Para estes grupos de dados observou-se uma fragmentação das redes, observadas do ápice em direção à base da raiz, com um aumento de nós órfãos seguindo este sentido (Figura 21).

Quanto às centralidades, analisou-se o *degree* e *betweenness*, para os resultados do RNA-Seq total de parede celular (Tabela 8) e também para os transcritos diferencialmente expressos (Tabela 9). Especialmente para esta última análise, considerou-se que todos os transcritos de parede foram utilizados para o cálculo da correlação, mas apenas transcritos diferencialmente expressos foram carregados na *layer 1* de transcrito, ou seja, podem existir nós correlacionados com outros, mas que não necessariamente tem expressão diferencial entre os seg-

Figura 20: Visualização da estrutura geral de agrupamentos das redes de cada segmento da raiz de cana-de-açúcar. A visualização é obtida através do layout Cose e para os dados de transcriptoma total de parede celular, para os segmentos de 1 à 4 (S1-S4) da raiz de cana-de-açúcar. Para cada segmento, observou-se a formação de uma grande rede e alguns nós órfãos, que não se conectam a nenhum outro nó. Arestas vermelhas mostram correlações negativas e verdes as positivas. A cor dos nós mostra a transcriptômica em amarela, proteômica em azul e atividade enzimática em verde. Através do link de manipulação das redes é possível aproximar os nós e arestas, buscando identificar os detalhes de cada uma, como os nomes dos nós e peso das arestas.

mentos.

Essa estratégia buscou identificar todos os possíveis vizinhos dos transcritos diferencialmente expressos, ainda que não reportados na análise de expressão diferencial.

Figura 21: Conectividade das redes de co-expressão geradas por SIT para as condições que consideram as listas de transcritos diferencialmente expressos. Seguindo as condições de S1_S1xS2, S2_S1xS2, S2_S2xS3, S3_S2xS3, S3_S3xS4, S4_S3xS4, foram apontados 244, 243, 175, 176, 173 e 178 nós e 480, 424, 26, 28, 17 e 12 conexões, respectivamente. Apenas para as condições S1_S1xS2 e S2_S1xS2 existem mais conexões do que quantidade de nós, nas outras condições, ocorre uma fragmentação das conexões, gerando uma maior quantidade de nós órfãos e com isso a conectividade da rede é reduzida do ápice para a base da raiz, considerando as listas de expressão diferencial.

Tabela 8: Maiores centralidades das redes de co-expressão para o RNA-Seq total de parede celular para os segmentos de 1 à 4 (S1-S4) da raiz de cana-de-açúcar. Os parênteses mostram em qual layer esse identificador integrou, sendo 1 para transcritômica, 2 para proteômica e 3 enzimas.

Maiores centralidades degree e betweeness para o RNA-Seq total de parede celular	
Segmentos	Anotação das centralidades (C _D = degree, C _B = betweeness)
S1	C _D = SCACFL8022F04.G – Glicosil Transferase Família 8 P_Biossíntese (1)
	C _B = SCRLFL1004C08.G – Semelhante à Celulose Sinthase (CSL) H_Biossíntese(1)
S2	C _D = SCBFLB2091E01.G- 1-3-β-Glucano Sinthase H_Biossíntese(1)
	C _B = SCEPRT2044F09.G – β-Galactosidase H/P_Degradação(1 2 3)
S3	C _D = SCBFRZ3009B03.G - 4-Coumarate-Coa Ligase L_Biossíntese (1 2)
	C _B = SCSFFL4084B02.B - Cinnamoyl Coa Redutase (CCR) L_Biossíntese (1 2)
S4	C _{D1} = SCJLRT1023E06.G – β-1-3-Glucanase Glicosil Hidrolase Família 17 H_Degradação (1)
	C _{D2} = SCEPRT2046G07.G – Proteína Semelhante Lacase L_Degradação (1)
	C _{D3} = SCAGAM2016B10.G – α-Galactosidase H_Degradação (1 3)
	C _{D4} = SCVPRZ3027D06.G – Endo-1-4-β-Glucanase Glicosil Hidrolase Família 9 H/C_Degradação (1)
	C _{B1} = SCJLRT1023A03.G - Endo-1-4-β-Glucanase Glycosil Hidrolase Família 9 H/C_Degradação (1)
	C _{B2} = SCVPRZ2042G03.G – Glicosil Hidrolase Família 1 C_Degradação (1 2 3)

Tabela 9: Maiores centralidades das redes de co-expressão para as condições que analisam os transcritos diferencialmente expressos entre os segmentos da raiz de cana-de-açúcar, acompanhando a metodologia de análise entre as transições (S1_S1xS2, S2_S1xS2, S2_S2xS3, S3_S2xS3, S3_S3xS4, S4_S3xS4). Os parênteses mostram em qual layer esse identificador integrou, sendo 1 para transcritômica, 2 para proteômica e 3 enzimas.

Maiores centralidades degree e betweeness para as condições de transcritos diferencialmente expressos	
Condições	Anotação das centralidades (C _D = degree, C _B = betweeness) e destaque de transcritos diferencialmente expressos (*)
S1_S1xS2	C _{D1} = SCJLRT1023A03.G - Endo-1-4-β-Glucanase Glicosil Hidrolase Família 9 H/C_Degradação (1*)
	C _{D2} = SCVPRZ2042G03.G - Glicosil Hidrolase Família 1 C_Degradação (1* 2 3)
	C _B = SCRFHR1008D11.G - β-1-3-Glucanase Glicosil Hidrolase Família 17 H_Degradação (1*)
S2_S1xS2	C _{D1} = SCCCCL4001H11.G - β-Galactosidase H/P_Degradação (1* 2 3)
	C _{D2} = SCAGLR1064F05.G - Celulose Synthase (CESA) C_Biossíntese (1*)
	C _B = SCCCLR1066B10.G - Celulose Synthase (CESA) C_Biossíntese (1*)
S2_S2xS3	C _{D1} = SCCCLR1065A04.G - Glicosil Transferase Família 2 H_Biossíntese (1*)
	C _{D2} = SCSFFL3091E06.G - β-Galactosidase H/P_Degradação (1* 3)
	C _{D3} = SCVPAM1059F06.G - B-Galactosidase H/P_Degradação (1* 3)
	C _{D4} = SCJLST1020H07.G - Xiloglucano Endotransglicosilase/Hidrolase (XTH) H_Degradação (1*)
	C _{D5} = SCACLR1036F04.G - Celulose Sintase (CESA) C_Biossíntese (1)
	C _{D6} = SCEZLR1009C10.G - Glicosil Transferase Família 34 H_Biossíntese (1*)
	C _{D7} = SCEQLR1093F09.G - Glicosil Hidrolase Família 3 H/C_Degradação (1* 2)
	C _{D8} = SCEZRT2017G09.G - Semelhante à Celulose Synthase (CSL) H_Biossíntese (1*)
	C _B = SCEZRT2017G09.G - Semelhante à Celulose Synthase (CSL) H_Biossíntese (1*)
S3_S2xS3	C _{D1} = SCJLST1020H07.G - Xiloglucano Endotransglicosilase/Hidrolase (XTH) H_Degradação (1*)
	C _{D2} = SCCCLR1048B04.G - Celulose Sintase H_Biossíntese (1*)
	C _{D3} = SCMCRT2085B03.G - Proteína Semelhante à Lacase L_Degradação (1*)
	C _{D4} = SCVPAM1059F06.G - β-Galactosidase H/P_Degradação (1* 3)
	C _B = SCCCLR1048B04.G - Celulose Sintase H_Biossíntese (1*)
S3_S3xS4	C _{D1} = SCEPLR1051B04.G - Proteína Arabinogalactano Semelhante a Fasciclina P_Pectina(1* 2)
	C _{D2} = SCEQLB1066G04.G - Semelhante a Celulose Sintase (CSL) H_Biossíntese (1*)
	C _B = SCQSST1040A06.G - Glicosiltransferase Família 34 H_Biossíntese (1*)
S4_S3xS4	C _{D1} = SCCCLR1066B10.G - Celulose Sintase (CESA) C_Biossíntese (1*)
	C _{D2} = SCQSST1037B03.G - Proteína Semelhante à Lacase L_Degradação (1*)
	C _{D3} = SCMCSB1114E08.G - Proteína Arabinogalactano Semelhante a Fasciclina P_Pectina (1*)
	C _{D4} = SCCCCL3002A03.B - 4-Coumarate-Coa Ligase L_Biossíntese (1*)
	C _{D5} = SCCCCL3001G09.B - Celulose Sintase (CESA) C_Biossíntese (1*)
	C _{D6} = SCCCSD1089A09.G - Glicosil Hidrolase Família 1 C_Degradação (1* 3)
	C _{D7} = SCEQLB1066G04.G - Semelhante a Celulose Sintase (CSL) H_Biossíntese (1*)
	C _B = SCEQLB1066G04.G - Semelhante a Celulose Sintase (CSL) H_Biossíntese (1*)

5.3.3 Análise de enzimas de degradação da parede celular

Nesta etapa, realizou-se uma análise focada apenas em elementos de degradação da parede celular, tendo sido aplicada considerando todos os dados de ômicas para cada segmento individualmente.

Buscando capturar a estrutura geral das redes de co-expressão, avaliou-se a conectividade das redes e perfil funcional das altas centralidades (Figura 22), bem como as centralidades dessas novas redes (Tabela 10).

Figura 22: Conectividade das redes de co-expressão geradas por SIT considerando apenas elementos de degradação de parede celular, para cada segmento que analisa o desenvolvimento do aerênquima (S1 à S4). Seguindo os segmentos de desenvolvimento do aerênquima foram apontados 251, 234, 250 e 199 nós e 1171, 901, 1153, 432 conexões, respectivamente. Embora com aproximadas quantidade de nós para os segmentos, observou-se um padrão de conectividade decrescente seguindo S1, S3, S2 e S4.

Tabela 10: Maiores centralidades das redes de co-expressão para elementos anotados apenas na degradação de parede celular para os segmentos de 1 à 4 (S1-S4) da raiz de cana-de-açúcar. Os parênteses mostram em qual layer esse identificador integrou, sendo 1 para transcritômica, 2 para proteômica e 3 enzimas.

Segmentos	Anotação das centralidades	
	(C _D = degree, C _B = betweenness)	
S1	C _{D1}	SCVPRZ2042G03.G - Glicosil Hidrolase Família 1 C_Degradação (1 2 3)
	C _{D2}	SCEZHR1088G11.G - β -1-3-Glucanase Glicosil Hidrolase Família 17 H_Degradação (1 2)
	C _B	SCCCLR1076C07.G - β -1-3-Glucanase Glicosil Hidrolase Família 17 H_Degradação (1)
S2	C _D	SCJLLR1102A08.G - Glicosil-Hidrolase Família 28 (poligalacturonase) P_degradação (1)
	C _B	SCCCCL4001H11.G - β -Galactosidase H/P_Degradação
S3	C _{D1}	SCRLSB1042A05.G - Glicosil Hidrolase Família 10 (Xilanase) H_Degradação (1)
	C _{D2}	SCRURT2006B10.G - β -1-3-Glucanase Glicosil Hidrolase Família 17 H_Degradação (1 2)
	C _{D3}	SCJFRZ2029F03.G - Glicosil Hidrolase Família 31 (α -Glucosidase/ α -Xilosidase) H_Degradação (1 2)
	C _{D4}	SCCCLR1069D11.G - Endo-1-4- β -1-3-Glucanase Glicosil Hidrolase Família 9 H/C_Degradação (1)
	C _B	SCCCLR1069D11.G - Endo-1-4- β -1-3-Glucanase Glicosil Hidrolase Família 9 H/C_Degradação (1)
S4	C _{D1}	SCVPRZ3027D06.G – Endo-1-4-Glucanase Glicosil Hidrolase Família 9 H/C_Degradação (1)
	C _{D2}	SCEPRT2046G07.G - Proteína Semelhante à Lacase L_Degradação (1)
	C _{D3}	SCAGAM2016B10.G – α -Galactosidase H_Degradação (1 3)
	C _{D4}	SCJLRT1023E06.G - β -1-3-Glucanase Glicosil Hidrolase Família 17 H_Degradação (1)
	C _B	SCJFLR1017E03.G - Glicosil Hidrolase Família 1 C_Degradação (1 2 3)

De forma geral, para a análise de degradação, os resultados dos maiores *betweenness* apontam a predominância em elementos de degradação de hemicelulose. O que ocorre é uma intercalação de domínios, observados ao longo do desenvolvimento do aerênquima.

As redes de co-expressão foram carregadas na interface online de SIT e visualizadas utilizando o *layout* Cose da plataforma. Assim foi possível observar para este conjunto de dados, a ocorrência de mais de um grupo de co-expressão para os segmentos S1 e S4, enquanto S2 e S3 apresentaram um único grupo de transcritos co-expressos (Figuras anexas 32 à 35).

Os resultados gerados pela ferramenta SIT apontam para centralidades das

redes de co-expressão, analisadas para os segmentos que correspondem ao desenvolvimento do aerênquima. Uma vantagem de SIT é podermos partir das altas centralidades, como um guia, para então investigar pontualmente as condições de relação (positiva ou negativa) com seus vizinhos. Adicionalmente, os resultados de integração permitem aumentar as evidências de ação dos elementos apontados, uma vez que nesta aplicação, utilizou dados de regulação molecular e que portanto, são dependentes uma das outras.

5.3.4 Perfil de potenciais elementos efetores no desenvolvimento do aerênquima em raízes de cana-de-açúcar

Essa análise objetivou a identificação de elementos efetores do processo de desenvolvimento do aerênquima em raízes de cana-de-açúcar, sob a perspectiva da integração da transcritômica, proteômica e enzimas, todos de parede celular e para cada segmento da raiz de cana-de-açúcar.

Considerando que a abertura de espaços de ar do aerênquima inicia-se na transição de S2 para S3 foram recuperados os 97 transcritos com maior valor de expressão em S2, apontados na análise de expressão diferencial da condição S1xS2.

Uma vez que a presença de uma proteína pode indicar a sua atividade no segmento, a lista desses transcritos foi submetida à uma análise de enriquecimento para: a) obtenção do perfil de expressão gênica ao longo dos segmentos de S1 a S4, incluindo possíveis diferenças estatísticas entre eles (S2xS3 e S3xS4); b) anotação funcional de cada transcrito; c) obter as informações sobre a integração de layers, dos resultados gerados anteriormente por SIT para cada segmento. O

resultado completo da análise de enriquecimento para os 97 transcritos é mostrado na Tabela 12 (anexa).

Tais resultados mostram elementos envolvidos na biossíntese e degradação da parede celular. Com foco na degradação da parede celular, observou-se um predomínio decrescente dos domínios de hemicelulose > Celulose > Pectina > Lignina.

Com base nos resultados de SIT sobre a integração de *layers*, um novo filtro foi aplicado para recuperar transcritos cuja ocorrência era em pelo menos duas *layers*, sendo transcrito e proteoma. Foram recuperados 95 elementos. Então deste total, filtramos aqueles cujo valor de expressão aumenta em S3, que é o segmento onde o aerênquima apresenta a primeira ocorrência de maior volume de área do córtex na raiz (LEITE et al., 2017).

Então foi gerada uma nova lista de 60 elementos que estão diferencialmente expressos na condição S1xS2, ocorrem nas *layers* de pelo menos, transcrito e proteoma mas principalmente, mostram um perfil de expressão aumentado em S3, na comparação com S2.

Entre os 60 (inseridos na Tabela 12 anexa) potenciais alvos efetores do processo biológico, destacamos pelo menos nove transcritos codificadores de enzimas de degradação da parede celular, cuja participação no processo biológico já foi apontada em outros estudos sobre o mesmo processo fisiológico. Foram dois transcritos de α -L-arabinofuranosidases (SCJLRT1017C05.g, SCVPLB1020G05.g), um de β -galactosidase (SCCCCL4001H11.g), quatro para glicosil hidrolase família 1 (SCCCCL4017A06.g, SCAGLR1021F11.g, SCEQHR1082B01.g e SCVPRZ2042G03.g) e dois para α -galactosidase (SCEQRT2025H02.g, SCUTLR1037C07.g) (Índices de 1 à 9 da Tabela 11).

Adicionalmente, a identificação recorrente de uma proteína arabinogalactana semelhante à fasciclina (SCEPLR1051B04.g) (Índice 10 da Tabela 11), chamou a

atenção neste estudo sobre o desenvolvimento do aerênquima em raízes de cana-de-açúcar. Esta é uma proteína do domínio das pectinas e que para melhor entender a sua interação com as enzimas acima mencionadas e também com os outros transcritos do segmento, aplicamos a ferramenta SIT para recuperar a sua rede de co-expressão, obtendo os primeiros vizinhos que integraram em pelo menos duas *layers* e para cada segmento da raiz (Figuras 23 a 26).

Analisando o perfil de domínios de biossíntese e degradação dos vizinhos mais conectados à proteína arabinogalactana semelhante à fasciclina, observou-se em S2 (dois SAS_IDs), a maior quantidade de transcritos cujas enzimas correspondentes foram igualmente indicadas na *layer* de enzimas. Nenhum vizinho integrado nas três *layers* foi apontado em S1, enquanto S3 e S4 apontou um SAS_ID cada.

Em S1 estão presentes quatro vizinhos de degradação, correspondentes à dois domínios para H/C (um positivamente e um negativamente co-expressos), um para H (positivamente co-expresso) e um para C (positivamente co-expresso). Essa quantidade aumenta para seis em S2, com quantidade decrescente de SAS_IDs, sendo três para H/P (dois positivamente e um negativamente co-expresso), dois para C (positivamente co-expresso) e um para H (positivamente co-expresso).

Em S3 o número de SAS_IDs sobe para nove, sendo dois para H (dois positivamente e um negativamente co-expresso), três para H/C (um positivamente e dois negativamente co-expressos), seguido de ocorrências de um SAS_ID para H/P (um positivamente co-expresso), L (negativamente co-expresso) e P (negativamente co-expresso).

Já em S4, ocorre uma queda na quantidade de elementos de degradação, caindo para 3 SAS_IDs, sendo um para H/P (negativamente co-expresso), um para C (negativamente co-expresso) e por fim um para L (positivamente co-expresso).

Tabela 11: Resultado da análise de enriquecimento, incluindo integração de dados, para os dez transcritos selecionados pela análise de alvos efetores do processo de desenvolvimento do aerênquima. A expressão gênica está na escala TPM(log2), para a lista de transcritos mais expressos em S2, quando considerada a análise de expressão diferencial entre o primeiro e segundo segmento (S1xS2). As colunas posteriores indicam se o transcrito foi apontado nas listas de genes diferencialmente expressos entre segmentos 2 e 3 (S2xS3), destacando se essa ocorrência é de maior expressão em S2 (S2_S2xS3) ou em S3 (S3_S2xS3). Senão, para os SAS_IDs que não estão diferencialmente expressos, a coluna S3 mostra o valor da expressão gênica em S3. Aplicou-se a mesma análise ao arquivo de genes diferencialmente expressos entre S3 e S4 (S3_S3xS4 e S4_S3xS4), e na condição de não ocorrência nesta, verificou-se a expressão em S4. Segue-se uma coluna com a anotação funcional e por domínio de cada SAS_ID. Posteriormente, as colunas de integração de dados mostram a quantidade de layers que cada SAS_ID ocorreu por segmento (1#, 2# ou 3#), bem como a identidade da layers integradas (1=transcritoma, 2=proteoma, 3=enzimas) por segmento do primeiro ao quarto (S1 à S4).

Índice	SAS_ID	Valor de expressão gênica por condição analisada (TPM-log2)							Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4		S1	S2	S3	S4
1	SCJLRT1017C05.g	5.19	0	0	5.94	0	0	6.20	Alpha-L-arabinofuranosidase H/P_degradation	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
2	SCVPLB1020G05.g	4.93	0	0	5.70	0	0	5.72	Alpha-L-arabinofuranosidase H/P_degradation	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
3	SCCCCL4001H11.g	6.66	0	0	6.92	0	0	6.46	Beta-galactosidase H/P_degradation	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
4	SCCCCL4017A06.g	6.17	0	0	6.25	0	0	6.56	Glycosyl hydrolase family 1 C_degradation	2#1 3	2#1 3	3#1 2 3	2#1 3
5	SCAGLR1021F11.g	5.44	0	0	6.24	0	0	5.98	Glycosyl hydrolase family 1 C_degradation	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
6	SCEQHR1082B01.g	4.83	0	0	5.40	0	0	4.92	Glycosyl hydrolase family 1 C_degradation	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
7	SCVPRZ2042G03.g	4.01	0	5.53	0	0	0	5.65	Glycosyl hydrolase family 1 C_degradation	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
8	SCEQRT2025H02.g	3.39	0	0	3.40	0	0	3.84	Alpha-galactosidase H_degradation	2#1 3	2#1 3	3#1 2 3	2#1 3
9	SCUTLR1037C07.g	2.61	0	0	2.96	0	0	0	Alpha-galactosidase H_degradation	2#1 3	2#1 3	3#1 2 3	2#2 3
10	SCEPLR1051B04.g	5.08	0	0	5.30	5.30	0	0	Fasciclin-like arabinogalactan protein P_Pectin	2#1 2	2#1 2	2#1 2	2#1 2

Figura 23: Rede de co-expressão do segmento 1 (S1). Amplificação da rede de co-expressão com destaque para o transcrito correspondente à proteína arabinogalactana semelhante à fasciclina (SCEPLR1051B04.g), mostrando seus vizinhos positivamente (arestas em verde) e negativamente (arestas em vermelho) co-expressos. A cor dos nós indicam que a layer a qual os transcritos ocorrem, sendo amarelo para o transcriptoma, azul para o proteoma e verde para as enzimas.

Figura 24: Rede de co-expressão do segmento 2 (S2). Amplificação da rede de co-expressão com destaque para o transcrito correspondente à proteína arabinogalactana semelhante à fasciclina (SCEPLR1051B04.g), mostrando seus vizinhos positivamente (arestas em verde) e negativamente (arestas em vermelho) co-expressos. A cor dos nós indicam que a layer a qual os transcritos ocorrem, sendo amarelo para o transcriptoma, azul para o proteoma e verde para as enzimas.

Figura 25: Rede de co-expressão do segmento 3 (S3). Amplificação da rede de co-expressão com destaque para o transcrito correspondente à proteína arabinogalactana semelhante à fasciclina (SCEPLR1051B04.g), mostrando seus vizinhos positivamente (arestas em verde) e negativamente (arestas em vermelho) co-expressos. A cor dos nós indicam que a layer a qual os transcritos ocorrem, sendo amarelo para o transcriptoma, azul para o proteoma e verde para as enzimas.

Figura 26: Rede de co-expressão do segmento 4 (S4). Amplificação da rede de co-expressão com destaque para o transcrito correspondente à proteína arabinogalactana semelhante à fasciclina (SCEPLR1051B04.g), mostrando seus vizinhos positivamente (arestas em verde) e negativamente (arestas em vermelho) co-expressos. A cor dos nós indicam que a layer a qual os transcritos ocorrem, sendo amarelo para o transcriptoma, azul para o proteoma e verde para as enzimas.

5.4 Visualização microtomográfica de segmentos da raiz de cana-de-açúcar

O conjunto de imagens geradas na microtomografia permitiu o desenvolvimento de um vídeo (LEITE et al., 2017), o qual permite uma visualização precisa das transições entre os segmentos de raízes, as quais apontam diferenças morfológicas nos tecidos ao longo dos 5 segmentos de raiz seguindo o desenvolvimento da zona meristemática da raiz em direção à base e junto da formação do aerênquima.

O vídeo permitiu a observação de regiões das raízes onde o aerênquima não é formado, com interrupções dessa estrutura ao longo dos segmentos, mostrando que o aerênquima não é um canal único ao longo da raiz, mas que em certas regiões há a formação de raízes laterais que não permitem o aerênquima ser formado (LEITE et al., 2017).

O vídeo está disponível no site da publicação como material suplementar (<https://academic.oup.com/aob/article-abstract/doi/10.1093/aob/mcx050/3924498?redirectedFrom=fulltext>)

6 Discussão

6.1 Biologia de Sistemas

Um dos desafios da biologia de sistemas é compreender a forma pela qual um organismo integra os seus diversos níveis de regulação biológica e quais são os elementos envolvidos em cada etapa (LAPATAS et al., 2015; SAETZLER; SONNENSCHNEIN; SOTO, 2011). Muitos fenômenos biológicos são decorrentes de respostas sistêmicas, mediante a comunicação entre órgãos (MOLNAR; MELNYK; BAULCOMBE, 2011) ou da regulação de vários processos por um mesmo elemento (PHILP et al., 2005). Mas para todos estes casos, envolve-se a integração de dados biológicos.

Com destaque à área de biologia molecular, tecnologias de NGS (*Next Generation Sequencing*) têm contribuído para a geração de grandes conjuntos de dados de sequências biológicas, os quais podem ser gerados rapidamente tanto para novos organismos quanto para variadas condições experimentais. Neste sentido, estratégias computacionais de integração de dados podem auxiliar no processamento rápido e massivo de tais dados e condições, permitindo a transformação de dados em conhecimento biológico (TRIPATHI et al., 2016).

Em virtude desse crescente acesso aos dados e maior quantidade de informações já obtidas, cada vez mais organismos estão tendo seus genomas e outros níveis de regulação biológica, conhecidos. Mas ainda existem organismos cuja complexidade biológica impede que seja conhecida a completa sequência genômica, como é o caso da cana-de-açúcar, organismo de aplicação desse estudo. Embora sem genoma completo disponível, existem outras naturezas de dados

biológicos que também são gerados como o transcritoma, proteoma, metaboloma (CARDOSO-SILVA et al., 2014; GARCIA et al., 2013; NISHIYAMA et al., 2014; SOUZA et al., 2001; VICENTINI et al., 2015), além de estudos mais especializados como especificamente elementos da parede celular (DE SOUZA et al., 2013; GRANDIS, 2015; LEITE et al., 2017; TAVARES, 2015), os quais juntos, contribuem para a compreensão a cerca dos processos biológicos que ocorrem nesses organismos.

6.2 *Systems Integration Tool*, uma nova ferramenta de integração de dados

Motivados pela importância bioenergética da cana-de-açúcar, sob o aspecto das modificações que ocorrem na parede celular, criamos a ferramenta SIT (Figura 8), capaz de integrar diferentes naturezas de dados de uma só vez. Ferramentas de integração deste tipo apresentam amplas aplicações nos setores de bioenergia, indústrias farmacêuticas e biorremediação.

Há na literatura, ferramentas capazes de fazer a integração de dados entre dois níveis, como já é bastante conhecido na aplicação para *Homo sapiens* (CHO et al., 2008; GARCIA-ALCALDE et al., 2011; HEIJNE et al., 2005; HU et al., 2009; KUO; TIAN; TSENG, 2013; RITCHIE et al., 2015; SURAVAJHALA; KOGELMAN; KADARMIDEEN, 2016; TOKIMATSU et al., 2005), por exemplo, as quais possuem um grande arsenal de plataformas e ferramentas de dados.

Tais estudos geralmente envolvem o processamento de até duas *layers*, enquanto SIT é capaz de integrar até seis *layers* de dados. Com este enfoque na

integração de outros níveis de regulação biológica, além da transcritômica e proteômica, existem na literatura ferramentas como a Ondex (KOHLETER et al., 2006), MetNet (JAISWAL, 2011), Biozon (SHAFER et al., 2006), 3Omics (KUO; TIAN; TSENG, 2013) e VirtualPlant (KATARI et al., 2010), as quais são capazes de processar tais dados, incluindo a visualização gráfica dos mesmos.

Contudo, uma característica essencial de todas essas outras plataformas de integração de dados de várias *layers*, é a origem dos dados, que neste caso são obtidas de plataformas públicas. A vantagem atribuída neste caso, está relacionada ao fato de que é possível acessar em tempo real, as novas informações biológicas. A desvantagem é que isso limita a sua aplicação apenas para organismos cujos dados estão disponíveis em plataformas públicas.

Neste sentido, SIT foi desenvolvida especificamente para permitir essa condição de dados, o que foi desenvolvido através da plataforma gráfica de execução local (Figuras 9,10 e 11), que faz a importação dos dados, a integração e exportação na forma de redes complexas. A execução local é uma das principais características de SIT, que a diferencia das demais ferramentas disponíveis.

Sobre a forma de visualização de dados, SIT gera o resultado da sua análise na forma de redes complexas (Figura 12). Neste ponto, a ferramenta Ondex (KOHLETER et al., 2006) também possui a visualização na forma de redes complexas, mas de relacionamentos prontos e importados da literatura, o que diferencia esta da SIT. Isto porque com esta última se podem aplicar novas análises de redes, sob novos conjuntos de dados, podendo encontrar novos padrões de agrupamento e correlação que muitas vezes não foram vistos na literatura. SIT permite assim criar novas hipóteses de estudo.

Em SIT, o princípio do algoritmo para integração é o uso de expressões regulares e termos chave, o que torna-a mais flexível para o uso de termos idênticos mas também semelhantes, sem perder a consistência da integração de termos-

chave. Essa característica de SIT torna possível a integração de níveis que biologicamente não são sucessivos, ou diretamente conectados, como ocorre com a transcritômica e a fisiologia, por exemplo, que possuem diversos níveis de regulação intermediários entre essas duas variáveis medidas, que podem não estar sendo utilizados diretamente na integração de dados, pela complexidade das rotas de explicação dos dados entre si.

Diferentemente das ferramentas disponíveis atualmente na literatura, SIT é capaz de carregar e integrar ao mesmo tempo, até seis conjuntos de dados locais, colocados no que convencionou-se chamar de *layers*. SIT foi desenvolvido para realizar análises personalizadas, permitindo ao usuário escolher e construir o tipo de dado que deseja adicionar em cada *layer*.

O usuário de SIT pode definir como deseja aplicar a sua integração de dados. O tipo de dado que é carregado na ferramenta não é fixo. Especificamente no caso de genética molecular, SIT recebe as tabelas de transcritos em uma *layer* e proteínas em outra *layer*, não realizando a etapa de mapeamento e montagem de sequências de ômicas, que neste caso pode ser definido como processamento dos dados e uma etapa anterior ao uso da ferramenta. Contudo, no caso de dados de NGS, estamos falando de milhares de linhas por arquivo, o que através de uma abordagem automática como a de SIT, acelera tanto o processamento quanto a posterior interpretação dos dados.

Considerando o montante de dados de cada *layer* que deve ser integrado e analisado ao mesmo tempo para que seja possível recuperar as intersecções e exclusividades de cada camada, o processamento manual neste caso possibilitaria erros de processamento, além do maior tempo, inviabilizando determinadas análises de dados (NEWGARD et al., 2012)

Ainda sobre os módulos funcionais de SIT, a visualização integrada na forma de redes permite a identificação de propriedades emergentes (BHALLA AND RAVI

IYENGAR, 1999), que surgem mediante análises globais e não individuais dos elementos. Estas podem ser apontadas pelas centralidades (e.x: *degree* e *betweenness*). Além disso, a análise de propriedades das redes permite conhecer quais os elementos estão fortemente co-expressos com ele, informação esta obtida através de seus vizinhos mais fortemente conectados nas redes, os quais podem estar positivamente ou negativamente co-expressos. Geralmente, genes da mesma via metabólica são co-expressos (WEI et al., 2006). Dessa forma, a análise de redes nos traz a informação sobre possíveis vias de regulação do sistema em estudo.

Pensada para atingir também usuários sem expertise com linguagens de programação, SIT foi desenvolvida para ser executada inteiramente via interface gráfica (Figuras 9 e 12), mas gerando os arquivos intermediários em formato texto, para que também seja possível a sua utilização por profissionais que preferem o terminal de comando e assim utilizá-la em *pipelines* de análise.

6.3 A aplicação do SIT no processo de formação do aerênquima de raiz de cana-de-açúcar

Para explorar o potencial de SIT, utilizamos dados existentes no Laboratório de Fisiologia Ecológica de Plantas (LAFIECO). O modelo utilizado foi o desenvolvimento do aerênquima em raízes de cana-de-açúcar. Independente do tipo de formação, a aparência morfológica do aerênquima final é uma estrutura preenchida com lacunas de ar (EVANS, 2003). Entretanto, para o aerênquima do tipo lisígeno, como é observado em raízes de cana-de-açúcar, observam-se alguns remanescentes originados da morte celular, os quais podem ser identificados nas

lacunas de ar formadas (LEITE et al., 2017; SMIRNOFF; CRAWFORD, 1983).

O desenvolvimento do aerênquima em milho pode ser induzido por alagamento ou anoxia (SUBBAIAH; SACHS, 2003). Em comparação, a formação do aerênquima em arroz (REBOUILLAT et al., 2009), bem como em cana-de-açúcar (LEITE et al., 2017), é constitutivo e parecem ser parte do processo de desenvolvimento. Nestes dois últimos, o aerênquima é visível em todos os tipos de raízes, com exceção em locais onde formam as raízes laterais. No material suplementar de LEITE et al. (2017), há uma montagem de um filme baseado em microtomografia (parte desta tese) em que se pode acompanhar a estrutura do aerênquima em um segmento de 5cm de uma raiz de cana-de-açúcar.

Entre as vias metabólicas envolvidas na formação do aerênquima, comumente são reportados estudos sobre a sinalização de etileno e de espécies reativas de oxigênio (JACKSON; ARMSTRONG, 1999; NISHIUCHI et al., 2012; YAMAUCHI et al., 2014; YOO; CHOI; JUNG, 2015), envolvidos na ativação do processo e em resposta dele, respectivamente.

Adicionalmente, outros estudos têm focado também nas alterações estruturais que ocorrem no processo de desenvolvimento do aerênquima (CARDOSO; JIMÉNEZ; RAO, 2014; GAO et al., 2015; JUNG; LEE; CHOI, 2008; SHEVCHENKO; BRYKOV; IVANENKO, 2016) e entre essas alterações, a parede celular aparece intimamente relacionada à formação das lacunas de ar, de tal forma que a parede celular atua com papel chave na força mecânica do tecido e órgãos em questão (BUCKERIDGE et al., 2017b; FRY, 2011; YAMAUCHI et al., 2013).

Juntos, os polímeros da parede celular, incluindo a hemicelulose, celulose, pectinas, proteínas e alguns compostos fenólicos, são alterados espacialmente e temporalmente durante o crescimento celular. Isso mostra a estrutura dinâmica e multi-funcional da parede celular, capaz de afetar múltiplos processos de sinalização, ocasionando distintas características estruturais e funcionais

(BUCKERIDGE et al., 2017b; FLORES-BORGES; MASTROBERTI; BONA, 2016).

Neste sentido, este estudo trás a primeira análise de NGS para elementos da parede celular à ótica do processo de formação do aerênquima em raízes de cana-de-açúcar.

Sobre o processamento do RNA-Seq de parede celular para os quatro primeiros centímetros da raiz, representados pelos segmentos (S) de 1 centímetro, chamou a atenção para os três primeiros segmentos, a maior presença de transcritos anotados como β -1-3-glucanase família 17 (GH17) (Figura 18), que é uma enzima da família das glicosil hidrolases (EC 3.2.1). No caso do aerênquima, estas enzimas podem estar degradando o conteúdo de β -1-3-glucano (calose) mencionado por Leite et al. (2017), conforme o aerênquima se desenvolve, uma vez que esta é uma das atividades já reportadas para esta enzima (THOMAS et al., 2000).

Entre os polissacarídeos que possuem estruturas de β -1-3-glucano degradados pelas GH17 encontra-se a calose, encontrada na parede celular de uma ampla gama de plantas superiores (CHEN; KIM, 2009). Interessantemente, a calose é sintetizada por uma enzima de biossíntese chamada β -1-3-glucano sintase, que justamente é a segunda enzima mais representada em número de transcritos do RNA-Seq de parede celular na raiz da cana-de-açúcar, para os segmentos S2, S3 e S4 (Figura 18). Entre os diversos papéis aos quais a calose já foi reportada está a sua deposição no plasmodesmo, regulando o movimento de moléculas célula-célula pelo tamanho do limite de exclusão do plasmodesmo (BUCHER et al., 2001; IGLESIAS; MEINS, 2000).

Em aerênquima do tipo lisígeno, quando ocorre a morte celular programada de células do córtex para a sua formação (EVANS, 2003; GUNAWARDENA et al., 2001), também observado em raízes de cana-de-açúcar (GRANDIS et al., 2014; LEITE et al., 2017), foi reportado um aumento proeminente de plasmodesmos no

processo de formação do aerênquima (Leite et al., resultados não publicados), mas em um estágio anterior à visualização da degradação da parede celular propriamente dita, no que seria um estágio inicial de modificação da parede celular (GUNAWARDENA et al., 2001).

Esse resultado aponta para um equilíbrio entre biossíntese e degradação de β -1-3-glucano, que é um polissacarídeo do domínio das hemiceluloses, bem como um balanço de enzimas GH17. Na Tabela 12, estão reportados alguns transcritos diferencialmente expressos de β -1-3-glucanase família 17, incluindo resultados de integração com a proteômica e enzimas. Para melhor entendimento sobre quais destes transcritos estariam potencialmente envolvidos nesse processo de hidrólise de β -1-3-glucano, novos experimentos são necessários. Este é o foco da dissertação de um dos membros de nosso laboratório, Grayce Romin Silva.

Adicionalmente, buscou-se compreender o contexto da regulação genética que é responsável pela expressão do fenótipo observado, neste caso a presença ou ausência de aerênquima, sob a perspectiva de elementos da parede celular. Para isso, aplicou-se a ferramenta SIT desenvolvida em dados de transcrito, proteoma e enzimas, todos gerados para cada um dos quatro primeiros centímetros da raiz de cana-de-açúcar.

Por se tratar de uma análise em larga escala, um dos resultados obtidos é um conjunto de listas de transcritos que integraram também na lista de proteínas e enzimas (Tabela 5). Eles apontam vários potenciais alvos de degradação da parede celular. Tais enzimas, contudo, estão sob um complexo sistema regulador no interior das células (DE NADAL; AMMERER; POSAS, 2011; RIECHMANN, 2002; VOSS; HAGER, 2013; WRAY et al., 2003), atuando inclusive sob uma especificidade de sítios enzimáticos (HEDSTROM, 2010) e podendo portanto, estarem ativas em uma região da raiz mas não em outra. Um destes mecanismos foi identificado por Tavares (2015) em que uma pectinase (endo-poligalacturonase) tem sua regulação

feita pelo fator de transcrição RAV diretamente.

Assim, no contexto de produção de bioetanol 2G, conhecer a identidade de tais enzimas, além de direcionar estudos sobre as vias de modificação da parede celular, podem também auxiliar em estudos para o aumento da efetiva degradação da biomassa vegetal. Os níveis de expressão de tais enzimas podem indicar a quantidade requerida de enzimas exógenas em coquetéis enzimáticos (XIAO; POOVAIAH; COLEMAN, 2016).

Entre os resultados obtidos por SIT, foram identificados dois transcritos específicos por segmento (Tabela 6), no caso o SAS SCQSAM1030G04.g, anotado como β -xilosidase/ α -L-arabinofuranosidase, uma glicosil hidrolase da família 3, está integrado na proteômica e enzimas e somente em S1 da raiz. Este transcrito está associado à degradação dos domínios hemicelulose e pectina. Já o SAS SCQSAM1030G04.g, que é uma exoglucanase envolvida na degradação de celulose, foi integrado também na *layer* da proteômica e enzimas porém apenas no segmento 4.

Sobre as glicosil hidrolases, elas compreendem um grupo de enzimas que agem na degradação de polissacarídeos da biomassa vegetal e que são focos de estratégias para aumentar o poder de ação de coquetéis enzimáticos (MUNJAL et al., 2015; OLOFSSON et al., 2017; WANG et al., 2012).

A parede celular possui uma estrutura intrincada, composta de outros domínios como o de hemiceluloses, por exemplo. Especialmente em cana-de-açúcar, a proporção de hemiceluloses é superior à celulose (BUCKERIDGE et al., 2015, 2017b; DE SOUZA et al., 2014). Estes autores, assim como também Grandis (2014) e Tavares et al. (2015) vêm apontando a importância da existência de uma sequência de eventos e produção de enzimas na formação do aerênquima. Estes eventos têm sido correlacionados com a necessidades de produção de enzimas e sua ação sobre a biomassa por coquetéis para produção de bioetanol 2G.

Neste contexto, o presente estudo aponta pelo menos duas enzimas específicas para uma região sem aerênquima (S1) e uma região onde esta estrutura já está formada (S4).

Informações mais específicas sobre o papel de tais enzimas no processo de desenvolvimento do aerênquima requerem estudos experimentais. Isto já vem sendo feito em nosso laboratório. Através de estudos específicos sobre o desenvolvimento do aerênquima de cana-de-açúcar, além do trabalho de Leite et al. (2017), foram abordadas as atividades enzimáticas e as variações nas proteínas (proteômica) do processo (GRANDIS, 2015), morte celular programada (LEITE, 2013) e alguns dos mecanismos de regulação efetuados por hormônios vegetais e fatores de transcrição (TAVARES, 2015). Todos estes focos foram adotados na medida do desenvolvimento dos respectivos trabalhos. Para confirmar e estender estes focos, a ferramenta SIT pode ser utilizada.

Uma das formas de apontar transcritos alvos é através da metodologia de redes complexas, especificamente com análises de propriedades das redes, através das medidas de centralidades *degree* e *betweenness*. Dentro da teoria de redes, transcritos com altas centralidades (do inglês, hubs) são apontados como alvos candidatos e que tem sido associados com vias metabólicas chave dos processos (CARLSON et al., 2006; FULLER et al., 2007; IANCU et al., 2012; JUNKER; SCHREIBER, 2008).

Após a montagem do transcritoma dos quatro segmentos da raiz de cana-de-açúcar, os dados foram analisados de duas formas. A primeira foi uma análise de propriedades de redes de co-expressão e objetivou a identificação de *hubs* sob o cenário de transcritos de biossíntese e degradação de elementos da parede celular.

Na segunda forma, uma vez que o fenômeno de formação do aerênquima em cana-de-açúcar denota participação significativa de processos de degradação da parede celular (alterações morfológicas, expressão de genes e produção de

hidrolases), analisou-se a conectividade do ponto de vista apenas de elementos de degradação.

No primeiro caso, até o segmento 3 foram apontados como maiores centralidades predominantemente os transcritos de biossíntese, enquanto que em S4 as maiores centralidades foram todas de degradação de domínios da parede celular (Tabela 8). Não encontramos na literatura um estudo semelhante que pudesse relacionar esse perfil de centralidades ao desenvolvimento do aerênquima. Contudo, sugere-se que até S3 os genes da raiz, embora executando atividades de degradação, como é observado inclusive na formação do aerênquima em S2, as vias mais ativas seriam de biossíntese. Isto pode ser explicado pelo fato do ápice da raiz ser uma região meristemática (HAYASHI; HASEGAWA; MATSUNAGA, 2013) e portanto de intensa divisão celular. Apesar de interessante e significativo, o fato de haver dois processos paralelos impôs certa dificuldade na identificação de um potenciais alvos relacionados ao início do processo de formação do aerênquima.

Já no caso apenas de elementos de degradação (Tabela 10), buscou-se complementar a abordagem anterior. Assim, considerando que os espaços de ar do aerênquima começam a se formar em S2, recuperamos para este segmento as maiores centralidades relacionadas à degradação, que foram o SAS SCJLLR1102A08.g, uma glicosil-Hidrolase Família 28 (poligalacturonase), envolvida na degradação de pectina como maior *degree* e o SAS SCCCCL4001H11.g, uma β -Galactosidase como o maior *betweenness* também envolvida com a degradação de pectinas. Ou seja, transcrito de degradação da principal pectina da parede celular da cana-de-açúcar (poligalacturonano) é o *hub* mais conectado da rede de co-expressão de S2, portanto mais conectado da rede. Já o maior *betweenness* indica quem é o transcrito, considerado o caminho, mais conectado entre duas sub-redes. Este codifica para uma enzima que ataca galactano, polímero que já foi detectado como a principal ramificação neutra da pectina de cana-de-açúcar (DE SOUZA et al.,

2014) Esta última descoberta corrobora a observação de LEITE et al. (2017) de que há uma diminuição no nível de galactose nas raízes durante a formação do aerênquima. Estes resultados reafirmam que as pectinas são um foco importante e um alvo nos estudos do processo de degradação da parede celular (LATARULLO et al., 2016; TAVARES, 2015).

Adicionalmente, analisando exclusivamente os dados de transcritos diferencialmente expressos principalmente entre S1 e S2 (S1xS2), região que marca ausência e presença de aerênquima, respectivamente, a análise de integração apontou um outro transcrito interessante do ponto de vista funcional (Tabela 11). Isso porque, não é um transcrito anotado diretamente na degradação de domínios da parede celular, mas sim envolvido com a sua modificação.

Com base nisso, resolvemos aprofundar a relação dessa proteína, que é um arabinogalactano-proteína semelhante à fasciclina (SCEPLR1051B04.g), anotada no domínio das pectinas e que foi apontada nesta lista de transcritos diferencialmente expressos sobre o processo de formação do aerênquima.

Arabinogalactano-proteínas correspondem a uma das principais subfamílias de glicoproteínas ricas em hidroxiprolinas, ao lado das extensinas e proteínas ricas em prolinas (ESTÉVEZ et al., 2006). As duas últimas são proteínas altamente glicosiladas e abundantes na parede celular e membrana plasmática das plantas (NOTHNAGEL, 1997; SHOWALTER, 2001). Elas estão relacionadas à vários processos fisiológicos. Em animais são bastante associadas à adesão celular, contudo em plantas foram reportadas também em processos de expansão celular, incluindo em raízes de *Arabidopsis* (JOHNSON et al., 2011) e trigo (FAIK; ABOUZOUHAIR; SARHAN, 2006), além de sinalização célula-célula, alongamento celular, morte celular e resposta a estresse (FINCHER; STONE; CLARKE, 1983; KOTAKE et al., 2009; NOTHNAGEL, 1997; SHOWALTER, 2001).

O arabinogalactano-proteína apontado apresentou a expressão gênica

aumentada em S3 na comparação com S2. Em análise de domínios verificou-se que a sua sequência possui um domínio fasciclina (FAS1, PSSM: smart00554) conservado. Junto com outras proteínas como quinases receptoras, complexos de celulose sintase, quinases associadas à parede, também os arabinogalactanos-proteína semelhantes à fasciclinas atuam de forma dinâmica na interação da parede celular com a membrana plasmática (FAIK; ABOUZOUHAIR; SARHAN, 2006). Em plantas, acredita-se que as fasciclinas possam atuar na manutenção da integridade da membrana plasmática durante processos de expansão celular (JOHNSON et al., 2003).

Novos experimentos são necessários para confirmar a localização dessas proteínas em células do córtex da raiz e verificar se existe exclusividade do processo comparado à células do cilindro vascular.

Analizando as redes de co-expressão do arabinogalactano-proteína apontado, observou-se que entre seus vizinhos para S1 e S3, ocorreram pelos menos duas outras fasciclinas (SCQGRT1040D08.g e SCRUAD1133E06.g), ao lado de proteínas expansinas (SCEQLB1064E10.g (S1), SCCCST3144E01.g e SCRUAD1133E06.g (S3)) e que só foram vizinhos diretos nestes segmentos. Em todos estes casos a relação dos vizinhos foi positiva, indicando uma atividade sinérgica entre tais domínios da parede celular, uma vez que o arabinogalactano-proteína possui um papel também no processo de expansão celular (HOUSTON et al., 2016).

A descoberta de que a fasciclina pode ser um dos elementos centrais no processo de formação do aerênquima em cana-de-açúcar sugere que esta possa ter um papel específico no processo. Grandis e colaboradores (2014) reportaram uma sequência de módulos que ocorrem no processo de degradação da biomassa vegetal de cana-de-açúcar para a produção de bioetanol 2G. Esses módulos são: 1) percepção do balanço hormonal por células alvos; 2) separação celular; 3) expansão

celular; 4) morte celular programada. 5) hidrólise/afrouxamento de ligações entre hemiceluloses e celulose; 6) hidrólise de celulose. Com isso, os arabinogalactano-proteínas semelhante à fasciclinas estariam provavelmente agindo no módulo 3 da degradação, ou seja participando da expansão celular. É importante salientar que este foco ainda não havia sido abordado nos estudos em paralelo sobre o aerênquima de cana em nosso laboratório. Ele só foi notado após a utilização da ferramenta SIT.

É importante notar que os genes discutidos acima não foram apontados por SIT apenas através de dados de expressão gênica. Em todos os casos, seja na confirmação do foco nas transformações de enzimas relacionadas ao plasmodesmo, às pectinas e à celulose, seja no caso da descoberta do papel central da fasciclina no processo de expansão celular, todos os casos apontados aqui envolvem a integração entre expressão gênica, proteômica e atividade enzimática. Em outras palavras, a análise possibilitado por SIT vai além da simples atribuição de função somente através de um nível. Neste caso foram usados três níveis, o que diminui consideravelmente a incerteza de atribuição de função a certos genes.

6.4 Aplicabilidade de SIT

De forma geral, este estudo permitiu através da ferramenta SIT, a criação de uma estratégia para a integração de dados gerados por técnicas de NGS, bem como dados de diferentes naturezas (proteômica, atividade enzimática). A análise dos resultados gerados pela integração de dados sobre o desenvolvimento do aerênquima em raízes de cana-de-açúcar avançou principalmente a compreensão sobre os elementos de parede celular envolvidos em cada região da raiz

correspondente ao desenvolvimento do aerênquima.

Com os dados analisados de RNA-Seq já é possível gerar uma lista de hidrolases capazes de atacar a parede celular da cana-de-açúcar naturalmente. Estas serão informações cruciais para adquirir controle sobre o amolecimento do tecido e seu possível uso para a produção de etanol 2G.

A interação da transcritômica com a fisiologia, através das medidas enzimáticas, tem potencial de revelar caminhos importantes a seguir no sentido de compreender como a parede celular é construída e possivelmente dominar os processos de reengenharia da parede celular da cana-de-açúcar. Com a ferramenta SIT, já é possível integrar genes de outras categorias (fatores de transcrição e hormônios por exemplo) que podem ser candidatos a serem utilizados para modular a produção de parede celular na cana-de-açúcar.

Os resultados de integração gerados neste estudo poderão servir também para a criação de um modelo de formação do aerênquima, considerando os transcritos chave no processo. Os padrões de alterações de transcrição e tradução ainda precisam ser mais fortemente associados aos eventos fazem parte desse processo biológico.

Adicionalmente aos estudos com aerênquima, SIT foi aplicado com sucesso em colaborações com ao menos dois estudos. Na integração de dados de transcritos, proteínas e metabólitos obtidos em estudo de desenvolvimento embrionário de *Araucaria angustifolia*, realizado juntamente com a professora Dra. Eny Lochevet Segal Floh e doutorando Bruno Vianna, ambos do Instituto de Biociências da Universidade de São Paulo USP/SP. Uma segunda colaboração ocorreu para integração de dados das espécies *Senna multijuga* e *Hungerfordia aurea*, para variáveis metabólicas e ambientais sob os períodos de verão e inverno e para dois tipos diferentes de florestas (altitude e tabuleiro). Esta última ocorreu em colaboração com o professor Dr. Geraldo Rogério Faustini Cuzzuol e doutorando

Bernardo Pretti Becacici Macieira, ambos da Universidade Federal do Espírito Santo – UFES. Estas duas aplicações mostram que o SIT pode ser uma ótima ferramenta para aplicação com espécies nativas, que não possuem genoma sequenciado e para as quais há também relativamente poucos dados fisiológicos e bioquímicos. Assim, através de estudos intensivos, pesquisadores poderão produzir conjuntos de dados que, quando usados em SIT, lhes possibilitarão apontar focos de ação para novos experimentos, considerando uma perspectiva sistêmica de análise de dados.

7 Conclusões e perspectivas futuras

Em apenas cerca de duas décadas, os métodos de sequenciamento em larga escala, ao lado de técnicas de biologia de sistemas, têm permitido muitos novos achados biológicos, incluindo o conhecimento genômico de muitas novas espécies. Contudo, organismos não modelos acabam prejudicados quando as ferramentas estão adaptadas apenas para plataformas públicas. A cana-de-açúcar, por exemplo, diante da destacada importância biotecnológica, ainda não tem seu genoma completo. Motivados neste contexto, desenvolvemos a ferramenta SIT para integração de dados, que mostrou-se capaz de integrar entre outros, os dados de ômicas como a transcritômica, proteômica, além da sua capacidade em apontar potenciais alvos efetores do processo biológico em estudo.

SIT destaca-se entre as outras da literatura pela sua capacidade de processamento local, além da possibilidade de execução de novas análises de propriedades das redes, além das já disponíveis na literatura. A ferramenta possui a capacidade de processamento de dados de naturezas diversas e para até seis diferentes conjuntos de dados.

Entre as principais conclusões da análise de aplicação da ferramenta SIT podemos destacar os elementos relacionados ao domínio das pectinas, apontados na análise de transcritos específicos por segmento e posteriormente na análise de centralidades *degree* e *betweenness*. Além do arabinogalactano-proteína semelhante à fasciclina, onde sugerimos estar associada ao módulo de expansão da parede celular durante o desenvolvimento do aerênquima em raízes de cana-de-açúcar. Além destes, inúmeros potenciais alvos para novos estudos foram apresentados.

Até o momento, não havia uma ferramenta capaz de auxiliar neste sentido e

com tais características. SIT possibilitou os estudos de integração dos dados de cana-de-açúcar gerados previamente pelo grupo de pesquisa Lafieco, possibilitando uma visão integrada quanto a formação do aerênquima em diferentes segmentos da raiz. Agora, SIT está disponível para aplicação principalmente em organismos não modelo.

Referências

ALBERTS, B. et al. **Molecular Biology of the Cell**. 4th editio ed. [s.l.] New York: Garland Science, 2002.

ANDREWS, S. **FastQC: a quality control tool for high throughput sequence data**, 2010. Disponível em: <<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>>

AOKI, K.; OGATA, Y.; SHIBATA, D. Approaches for Extracting Practical Information from Gene Co-expression Networks in Plant Biology. **Plant and Cell Physiology**, v. 48, n. 3, p. 381–390, 19 jan. 2007.

ARBER, A. **Water Plants: a study of aquatic angiosperms**. [s.l.] Cambridge: Cambridge University Press, 1920.

ARIKADO, H. Different responses of soybean plants to an excess of water with special reference to anatomical observations. **Proceedings of the Crop Science Society of Japan**, v. 23, p. 28–36, 1954.

ARMSTRONG, W. Aeration in higher plants. **Advances in Botanical Research**, v. 7, p. 225–332, 1980.

AZUAJE, F. J. Selecting biologically informative genes in co-expression networks with a centrality score. v. 9, p. 1–23, 2014.

BALLOUZ, S.; VERLEYEN, W.; GILLIS, J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. **Bioinformatics**, v. 31, n. 13, p. 2123–2130, 1 jul. 2015.

BARABÁSI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. **Nature Reviews Genetics**, v. 5, n. 2, p. 101–113, fev. 2004.

BARRAT, A. et al. The architecture of complex weighted networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 101, n. 11, p. 3747–52, 16 mar. 2004.

BASSO, K. et al. Reverse engineering of regulatory networks in human B cells.

Nature Genetics, v. 37, n. 4, p. 382–390, 20 abr. 2005.

BAVELAS, A. **Communication patterns in task-oriented groups**, 1950.

BHALLA AND RAVI IYENGAR, U. S. Emergent Properties of Networks of Biological Signaling Pathways. **Science**, v. 283, n. 5400, p. 381–387, 15 jan. 1999.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. **Bioinformatics**, 2014.

BUCHER, G. L. et al. Local expression of enzymatically active class I β -1, 3-glucanase enhances symptoms of TMV infection in tobacco. **The Plant Journal**, v. 28, n. 3, p. 361–369, 23 dez. 2001.

BUCKERIDGE, M. .; DOS SANTOS, W. .; DE SOUZA, A. P. **Routes for cellulosic ethanol in Brazil**. [s.l.] In: Cortez LAB (ed) Sugarcane bio-ethanol: R&D for productivity and sustainability, 2010.

BUCKERIDGE, M. S. et al. The cell wall architecture of sugarcane and its implications to cell wall recalcitrance. In: LAM, E. et al. (Eds.). . [s.l.] Compendium of Bioenergy Plants: Sugarcane, 2015. p. 31–50.

BUCKERIDGE, M. S. et al. Advances of Basic Science for Second Generation Bioethanol from Sugarcane. p. 7–20, 2017a.

BUCKERIDGE, M. S. et al. Sugarcane Cell Wall Structure and Degradation: From Monosaccharide Analyses to the Glycomic Code. In: **Advances of Basic Science for Second Generation Bioethanol from Sugarcane**. Cham: Springer International Publishing, 2017b. p. 7–19.

BUCKERIDGE, M. S.; CAVALARI, A. A.; SILVA, G. B. Fisiologia Vegetal: Parede Celular. In: KERBAUY, G. B. (Ed.). . [s.l.] Guanabara Koogan, Rio de Janeiro, 2008. p. 165–181.

BUCKERIDGE, M. S.; DE SOUZA, A. P. Breaking the “Glycomic Code” of Cell Wall Polysaccharides May Improve Second-Generation Bioenergy Production from Biomass. **BioEnergy Research**, v. 7, n. 4, p. 1065–1073, 24 dez. 2014.

CARDOSO-SILVA, C. B. et al. De Novo Assembly and Transcriptome Analysis of Contrasting Sugarcane Varieties. **PLoS ONE**, v. 9, n. 2, p. e88462, 11 fev. 2014.

CARDOSO, J. A.; JIMÉNEZ, J. DE LA C.; RAO, I. M. Waterlogging-induced changes

in root architecture of germplasm accessions of the tropical forage grass *Brachiaria humidicola*. **AoB PLANTS**, v. 6, 8 abr. 2014.

CARLSON, M. et al. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. **BMC Genomics**, v. 7, n. 1, p. 40, 3 mar. 2006.

CARPITA, N. C. Fractionation of hemicelluloses from maize cell walls with increasing concentrations of alkali. **Phytochemistry**, v. 23, p. 1089–1093, 1984.

CARPITA, N. C. STRUCTURE AND BIOGENESIS OF THE CELL WALLS OF GRASSES. **Annual Review of Plant Physiology and Plant Molecular Biology**, v. 47, n. 1, p. 445–476, jun. 1996.

CARPITA, N. C.; GIBEAUT, D. M. Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. **The Plant journal: for cell and molecular biology**, v. 3, p. 1–30, 1993.

CHÁVEZ-BÁRCENAS, A. T. et al. Tissue-specific and developmental pattern of expression of the rice *sps1* gene. **Plant physiology**, v. 124, n. 2, p. 641–54, out. 2000.

CHEN, X.-Y.; KIM, J.-Y. Callose synthesis in higher plants. **Plant signaling & behavior**, v. 4, n. 6, p. 489–92, jun. 2009.

CHI, K. R. The dark side of the human genome. **Nature**, v. 538, n. 7624, p. 275–277, 17 ago. 2016.

CHO, K. et al. Integrated transcriptomics, proteomics, and metabolomics analyses to survey ozone responses in the leaves of rice seedling. **Journal of Proteome Research**, v. 7, p. 2980–2998, 2008.

CONESA, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, v. 21, n. 18, p. 3674–3676, 2005.

CONESA, A. et al. A survey of best practices for RNA-seq data analysis. **Genome Biology**, v. 17, n. 1, p. 13, 26 dez. 2016.

COX, J.; MANN, M. MaxQuant enables high peptide identification rates,

individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. **Nature Biotechnology**, v. 26, p. 1367–1372, 2008.

CRICK, F. Central dogma of molecular biology. **Nature**, v. 227, p. 561–563, 1970.

D'HONT, A. et al. Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. **Molecular and General Genetics**, v. 250, n. 4, p. 405–413, 1996.

DAUB, C. O. et al. Estimating mutual information using B-spline functions--an improved similarity measure for analysing gene expression data. **BMC bioinformatics**, v. 5, p. 118, 31 ago. 2004.

DE CARVALHO PIRES, T. (ED.). **Bioetanol combustível: uma oportunidade para o Brasil**. [s.l.] Centro de Gestão e Estudos Estratégicos, 2009.

DE NADAL, E.; AMMERER, G.; POSAS, F. Controlling gene expression in response to stress. **Nature Reviews Genetics**, v. 12, n. 12, p. nrg3055, 3 nov. 2011.

DE SOUZA, A. P. et al. Composition and structure of sugarcane cell wall polysaccharides: implications for second-generation bioethanol production. **Bioenergy Research**, v. 6, p. 564–579, 2013.

DE SOUZA, A. P. et al. Sugarcane as a Bioenergy Source: History, Performance, and Perspectives for Second-Generation Bioethanol. **Bioenergy Research**, v. 7, p. 24–35, 2014.

DE SOUZA, A. P. et al. CHANGES IN WHOLE-PLANT METABOLISM DURING GRAIN-FILLING STAGE IN SORGHUM BICOLOR L. (MOENCH) GROWN UNDER ELEVATED CO₂ AND DROUGHT. **Plant Physiology**, p. 1054, 2015.

DIAS, M. O. S. et al. Improving second generation ethanol production through optimization of first generation production process from sugarcane. **Energy**, p. 246–252, 2012.

DILLON, S. L. et al. Domestication to crop improvement: genetic resources for Sorghum and *Saccharum* (Andropogoneae). **Annals of botany**, v. 100, n. 5, p. 975–89, nov. 2007.

DRRKSRRRA, E. W. A Note on Two Problems in Connexion with Graphs. **Numerische Mathematik**, v. 1, p. 269–27, 1959.

EGERSTEDT, M. Complex networks: Degrees of control. **Nature**, v. 473, n. 7346, p. 158–159, 12 maio 2011.

EISEN MB, SPELLMAN PT, BROWN PO, B. D. Cluster analysis and display of genome-wide expression patterns. **Proc Natl Acad Sci U S A.**, v. 95, p. 14863–14868, 1998.

ESTÉVEZ, J. M. et al. Characterization of synthetic hydroxyproline-rich proteoglycans with arabinogalactan protein and extensin motifs in Arabidopsis. **Plant physiology**, v. 142, n. 2, p. 458–70, out. 2006.

EVANS, D. E. Aerenchyma formation. **New Phytologist**, v. 161, n. 1, p. 35–49, 24 out. 2003.

EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. **Genome research**, v. 8, n. 3, p. 186–94, 1 mar. 1998.

EZKURDIA, I. et al. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. **Human Molecular Genetics**, v. 23, n. 22, p. 5866–5878, 15 nov. 2014.

FAIK, A.; ABOUZOUHAIR, J.; SARHAN, F. Putative fasciclin-like arabinogalactan-proteins (FLA) in wheat (*Triticum aestivum*) and rice (*Oryza sativa*): identification and bioinformatic analyses. **Molecular Genetics and Genomics**, v. 276, n. 5, p. 478–494, nov. 2006.

FINCHER, G. B.; STONE, B. A.; CLARKE, A. E. Arabinogalactan-Proteins: Structure, Biosynthesis, and Function. **Annual Review of Plant Physiology**, v. 34, n. 1, p. 47–70, 28 jun. 1983.

FLORES-BORGES, D. N. A.; MASTROBERTI, A. A.; BONA, C. Ontogeny and distribution of cell-wall glycans during aerenchyma formation in roots of *Pistia stratiotes* (Araceae). **Botanical Journal of the Linnean Society**, v. 180, n. 3, p. 401–412, 1 mar. 2016.

FREEMAN, L. C. Centrality in Social Networks Conceptual Clarification. **Social Networks**, v. 179, p. 215–239, 1978.

FREEMAN, L. C.; BORGATTI, S. P.; WHITE, D. R. Centrality in valued graphs: A measure of betweenness based on network flow. **Social Networks**, v. 13, p. 141–

154, 1991.

FRY, S. C. Plant cell walls. From chemistry to biology. **Annals of Botany**, v. 108, n. 1, p. viii–ix, jul. 2011.

FULLER, T. F. et al. Weighted gene coexpression network analysis strategies applied to mouse weight. **Mammalian genome : official journal of the International Mammalian Genome Society**, v. 18, n. 6–7, p. 463–72, jul. 2007.

GAO, K. et al. A comprehensive analysis of root morphological changes and nitrogen allocation in maize in response to low nitrogen stress. **Plant, Cell & Environment**, v. 38, n. 4, p. 740–750, 1 abr. 2015.

GARCIA-ALCALDE, F. et al. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. **Bioinformatics**, v. 7, p. 137–139, 2011.

GARCIA, A. A. F. et al. SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. **Scientific Reports**, v. 3, n. 1, p. 3399, 2 dez. 2013.

GENOUD, T.; MÉTRAUX, J.-P. Crosstalk in plant cell signaling: structure and function of the genetic network. **Trends in plant Science**, v. 4, p. 503–507, 1999.

GILBERT, R. A. et al. Morphological Responses of Sugarcane to Long-Term Flooding. **Agronomy journal**, v. 99, p. 1622–1628, 2007.

GILLIS, J. et al. The Impact of Multifunctional Genes on “Guilt by Association” Analysis. **PLoS ONE**, v. 6, n. 2, p. e17258, 18 fev. 2011.

GILLIS, J.; PAVLIDIS, P. The role of indirect connections in gene networks in predicting function. **Bioinformatics**, v. 27, n. 13, p. 1860–1866, 1 jul. 2011.

GOMEZ, L. D. et al. Automated saccharification assay for determination of digestibility in plant materials. **Biotechnology for Biofuels**, v. 3, p. 23–35, 2010.

GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. **Nature Biotechnology**, v. 29, p. 644–652, 2011.

GRANBIO. **GranBio**. Disponível em: <<http://www.granbio.com.br/>>.

GRANDIS, A. et al. Using Natural Plant Cell Wall Degradation Mechanisms to Improve Second Generation Bioethanol. In: ET AL. (EDS.), M. C. M. (Ed.). . [s.l.]

Plants and BioEnergy, *Advances in Plant Biology* 4, 2014. p. 211–230.

GRANDIS, A. **Papel das enzimas de degradação da parede celular na formação do aerênquima em raízes de cana de açúcar.** [s.l.] Instituto de Biociências da Universidade de São Paulo. Departamento de Botânica., 2015.

GRIVET L.; ARRUDA, P. Sugarcane genomics: depicting the complex genome of an important tropical crop. **Current Opinion in Plant Biology**, v. 5, p. 122–127, 2002.

GUNAWARDENA, A. H. L. A. N. et al. Characterisation of programmed cell death during aerenchyma formation induced by ethylene or hypoxia in roots of maize(*Zea mays* L.). **Planta**, v. 212, n. 2, p. 205–214, 19 jan. 2001.

HAN, J. D. et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. **Nature**, v. 430, p. 88–93, 2004.

HASHIMOTO, R. F. et al. Growing genetic regulatory networks from seed genes. **Bioinformatics**, v. 20, p. 1241–1247, 2004.

HAYASHI, K.; HASEGAWA, J.; MATSUNAGA, S. The boundary of the meristematic and elongation zones in roots: endoreduplication precedes rapid cell expansion. **Scientific Reports**, v. 3, n. 1, p. 2723, 14 dez. 2013.

HE, C. J.; MORGAN, P. W.; DREW, M. C. Transduction of an ethylene signal is required for cell death and lysis in the root cortex of maize during aerenchyma formation induced by hypoxia. **Plant Physiology**, v. 112, p. 463–472, 1996.

HEDSTROM, L. Enzyme Specificity and Selectivity. In: **Encyclopedia of Life Sciences**. Chichester, UK: John Wiley & Sons, Ltd, 2010.

HEIJNE, W. H. et al. Systems toxicology: applications of toxicogenomics, transcriptomics, proteomics and metabolomics in toxicology. **Expert Review of Proteomics**, v. 7, p. 767–780, 2005.

HIMMEL M.E et al. Biomass recalcitrance: engineering plants and enzymes for biofuels production. **Science**, v. 315, p. 804–807, 2007.

HOUSTON, K. et al. The Plant Cell Wall: A Complex and Dynamic Structure As Revealed by the Responses of Genes under Stress Conditions. **Frontiers in plant science**, v. 7, p. 984, 2016.

HU, Z. et al. VisANT: data-integrating visual framework for biological networks and

modules. **Nucleic Acids Research**, v. 33, p. W352–W357, 2005.

HU, Z. et al. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. **Nucleic Acids Research**, v. 37, p. W115–121, 2009.

IANCU, O. D. et al. Utilizing RNA-Seq data for de novo coexpression network inference. **Bioinformatics (Oxford, England)**, v. 28, n. 12, p. 1592–7, 15 jun. 2012.

IGLESIAS, V. A.; MEINS, F. Movement of plant viruses is delayed in a beta-1,3-glucanase-deficient mutant showing a reduced plasmodesmatal size exclusion limit and enhanced callose deposition. **The Plant journal: for cell and molecular biology**, v. 21, n. 2, p. 157–66, jan. 2000.

JACKSON, M. B.; ARMSTRONG, W. Formation of aerenchyma and the processes of plant ventilation in relation to soil flooding and submergence. **Plant Biology**, v. 1, p. 274–287, 1999.

JAISWAL, P. Gramene Database: A Hub for Comparative Plant Genomics. In: **BMC Bioinformatics 2012 13:1**. [s.l.] BioMed Central, 2011. v. 678p. 247–275.

JANSEN, R.; GREENBAUM, D.; GERSTEIN, M. Relating whole-genome expression data with protein-protein interactions. **Genome Research**, v. 12, p. 37–46, 2002.

JI, L. et al. Global Electricity Trade Network: Structures and Implications. **PLOS ONE**, v. 11, n. 8, p. e0160869, 9 ago. 2016.

JIANG, J. et al. Construction and application of a co-expression network in *Mycobacterium tuberculosis*. **Scientific Reports**, v. 6, n. 1, p. 28422, 22 set. 2016.

JOHNSON, K. L. et al. The fasciclin-like arabinogalactan proteins of *Arabidopsis*. A multigene family of putative cell adhesion molecules. **Plant physiology**, v. 133, n. 4, p. 1911–25, dez. 2003.

JOHNSON, K. L. et al. A Fasciclin-Like Arabinogalactan-Protein (FLA) Mutant of *Arabidopsis thaliana*, *fla1*, Shows Defects in Shoot Regeneration. **PLoS ONE**, v. 6, n. 9, p. e25154, 22 set. 2011.

JUNG, J.; LEE, S. C.; CHOI, H.-K. Anatomical patterns of aerenchyma in aquatic and wetland plants. **Journal of Plant Biology**, v. 51, n. 6, p. 428–439, nov. 2008.

JUNKER, B. H.; KLUKAS, C.; SCHREIBER, F. VANTED: a system for advanced data analysis and visualization in the context of biological networks. **BMC Bioinformatics**,

v. 7, p. 109, 2006.

JUNKER, B. H.; SCHREIBER, F. **Analysis of biological networks**. [s.l: s.n.].

KARNOVSKY, A. et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. **Bioinformatics**, v. 7, p. 373–380, 2012.

KATARI, M. S. et al. VirtualPlant: a software platform to support systems biology research. **Plant physiology**, 2010.

KOHLER, J. et al. Graph-based analysis and visualization of experimental results with ONDEX. **Bioinformatics**, v. 22, n. 11, p. 1383–1390, 1 jun. 2006.

KOLBE, A. et al. Combined transcript and metabolite profiling of Arabidopsis leaves reveals fundamental effects of the thiol-disulfide status on plant metabolism. **Plant Physiology**, v. 7, p. 412–422, 2006.

KOTAKE, T. et al. Arabinogalactan-Proteins in The Evolution of Gravity Resistance in Land Plants. **Soc. Biol. Sci. Space; Article ID**, v. 23, p. 143–149, 2009.

KROUK, G. et al. Gene regulatory networks in plants: learning causality from time and perturbation. **Genome Biology**, v. 14, p. 123, 2013.

KUMARI, S. et al. Evaluation of Gene Association Methods for Coexpression Network Construction and Biological Knowledge Discovery. 2012.

KUO, T. C.; TIAN, T. F.; TSENG, Y. J. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. **BMC Systems Biology**, v. 7, p. 64, 2013.

LANGFELDER, P.; HORVATH, S. WGCNA: an R package for weighted correlation network analysis. **BMC bioinformatics**, v. 9, p. 559, 29 dez. 2008.

LAPATAS, V. et al. Data integration in biological research: an overview. **Journal of Biological Research-Thessaloniki**, v. 22, n. 1, p. 9, 2 dez. 2015.

LATARULLO, M. B. G. et al. Pectins, Endopolygalacturonases, and Bioenergy. **Frontiers in plant science**, v. 7, p. 1401, 2016.

LEITE, D. C. C. **Modificações da parede celular durante a formação de aerênquima em raízes de cana-de-açúcar**. [s.l.] University of São Paulo.

Bioscience Institute. Department of Botany. Brazil, 2013.

LEITE, D. C. C. et al. Cell wall changes during the formation of aerenchyma in sugarcane roots. **Annals of Botany**, v. 3, p. 818–833, 4 jul. 2017.

LEMMENS, K. et al. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in *Escherichia coli*. **Genome Biology**, v. 10, 2009.

LENG, N. et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. **Bioinformatics**, 2013.

LI, B.; DEWEY, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. **BMC Bioinformatics**, v. 12, p. 323, 2011.

LI, Y.; CHEN, L. Big Biological Data: Challenges and Opportunities. **Genomics, Proteomics & Bioinformatics**, v. 12, n. 5, p. 187–189, 2014.

LIANG, Y.-H. et al. Construction and validation of a gene co-expression network in grapevine (*Vitis vinifera* L.). **Horticulture Research**, v. 1, p. 14040, 13 ago. 2014.

LIMA, D. U. et al. Patterns of expression of cell wall related genes in sugarcane. **Genetics and Molecular Biology**, v. 24, p. 191–198, 2001.

LIU, Y.-Y.; SLOTINE, J.-J.; BARABÁSI, A.-L. Controllability of complex networks. **Nature**, v. 473, n. 7346, p. 167–173, 12 maio 2011.

LOBO, D.; LEVIN, M. Inferring Regulatory Networks from Experimental Morphological Phenotypes: A Computational Method Reverse-Engineers Planarian Regeneration. **Plos Computational biology**, v. 11, p. e1004295, 2015.

LOPES, F. . M.; DE OLIVEIRA, E. A.; CESAR-JR, R. M. **Analysis of the GRNs inference by using Tsallis entropy and a feature selection approach**. [s.l.] 14th Iberoamerican Congress on Pattern Recognition - CIARP, 2009.

LOPES, F.; DE OLIVEIRA, E. A.; CESAR, R. M. Inference of gene regulatory networks from time series by Tsallis entropy. **BMC Systems Biology**, v. 5, n. 1, p. 61, 2011.

LOPES, F. M. et al. A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks. **Information Sciences**, v. 272, p. 1–15, 2014a.

LOPES, F. M. et al. Entropic Biological Score: A cell cycle investigation for GRNs inference. **Gene**, 2014b.

MA'AYAN, A. et al. Lean Big Data integration in systems biology and systems pharmacology. **Trends in Pharmacological Sciences**, v. 35, n. 9, p. 450–460, set. 2014.

MAKAGON, M. M.; MCCOWAN, B.; MENCH, J. A. How can social network analysis contribute to social behavior research in applied ethology? **Applied animal behaviour science**, v. 138, n. 3–4, 1 maio 2012.

MARGOLIN, A. A. et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. **BMC Bioinformatics**, v. 7, n. Suppl 1, p. S7, 20 mar. 2006.

MARTINEZ, M. From plant genomes to protein families: computational tools. **Computational and structural biotechnology journal**, v. 8, p. e201307001, 2013.

MCCANN, M. C.; ROBERTS, K. The cytoskeletal basis of plant growth and form. In: LLOYD, C. W. (Ed.). . [s.l.] Academic Press: New York., 1991. p. 109–129.

MERELLI, I. et al. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. **BioMed research international**, v. 2014, p. 134023, 2014.

MEYER, V. R. **Practical High-Performance Liquid Chromatography**. 5th. ed. [s.l.] Wiley-VCH, 2004.

MICHAUT, M. et al. Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. **Scientific Reports**, v. 6, p. 18517, 5 jan. 2016.

MOLNAR, A.; MELNYK, C.; BAULCOMBE, D. C. Silencing signals in plants: a long journey for small RNAs. **Genome Biology**, v. 12, n. 1, p. 215, 11 jan. 2011.

MORTAZAVI, A. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nature Methods**, v. 5, n. 7, p. 621–628, 30 jul. 2008.

MOSIER, N. et al. Features of promising technologies for pretreatment of lignocellulosic biomass. **Bioresource Technology**, v. 96, p. 673–686, 2005.

MÜLLER, F.-J.; SCHUPPERT, A. Few inputs can reprogram biological networks.

Nature, v. 478, n. 7369, p. E4–E4, 19 out. 2011.

MUNJAL, N. et al. A Constitutive Expression System for Cellulase Secretion in *Escherichia coli* and Its Use in Bioethanol Production. **PLOS ONE**, v. 10, n. 3, p. e0119917, 13 mar. 2015.

NEKRUTENKO, A.; TAYLOR, J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. **Nature Reviews Genetics**, v. 13, n. 9, p. 667–672, 17 ago. 2012.

NEWGARD, C. D. et al. Electronic versus manual data processing: evaluating the use of electronic health records in out-of-hospital clinical research. **Academic emergency medicine : official journal of the Society for Academic Emergency Medicine**, v. 19, n. 2, p. 217–27, fev. 2012.

NEWMAN, M. E. J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. **PHYSICAL REVIEW**, v. 64, p. 16132, 2001.

NEWMAN, M. E. J. The Structure and Function of Complex Networks. **SIAM Review**, v. 45, n. 2, p. 167–256, jan. 2003.

NISHIUCHI, S. et al. Mechanisms for coping with submergence and waterlogging in rice. **Rice**, v. 5, n. 1, p. 2, 2012.

NISHIYAMA, M. Y. et al. Full-Length Enriched cDNA Libraries and ORFeome Analysis of Sugarcane Hybrid and Ancestor Genotypes. **PLoS ONE**, v. 9, n. 9, p. e107351, 15 set. 2014.

NOTHNAGEL, E. A. Proteoglycans and related components in plant cells. **International review of cytology**, v. 174, p. 195–291, 1997.

OECD/FOOD; OF THE UNITED NATIONS, A. O. **OECD-FAO Agricultural Outlook 2015**. [s.l.] OECD Publishing, Paris, 2015.

OLOFSSON, J. et al. Integrating enzyme fermentation in lignocellulosic ethanol production: life-cycle assessment and techno-economic analysis. **Biotechnology for Biofuels**, v. 10, n. 1, p. 51, 23 dez. 2017.

OPSAHL, T.; AGNEESSENS, F.; SKVORETZ, J. Node centrality in weighted networks: Generalizing degree and shortest paths. **Social Networks**, v. 32, n. 3, p. 245–251, jul. 2010.

- PEARSON, K. **The science of man, its needs and its prospects.** [s.l.] London :, 1920.
- PHILP, A. et al. Lactate--a signal coordinating cell and systemic function. **The Journal of experimental biology**, v. 208, n. Pt 24, p. 4561–75, 15 dez. 2005.
- PRINESS, I.; MAIMON, O.; BEN-GAL, I. Evaluation of gene-expression clustering via mutual information distance measure. **BMC bioinformatics**, v. 8, p. 111, 30 mar. 2007.
- PUNIYA, B. L. et al. Integration of Metabolic Modeling with Gene Co-expression Reveals Transcriptionally Programmed Reactions Explaining Robustness in Mycobacterium tuberculosis. **Scientific Reports**, v. 6, p. 23440, 22 mar. 2016.
- RAGAUSKAS, A. J. et al. The Path Forward for Biofuels and Biomaterials. **Science**, v. 311, p. 484–489, 2006.
- RAÍZEN. **Relatório de Sustentabilidade 2013/2014.** [s.l: s.n.]. Disponível em: <<http://www.raizen.com.br/>>.
- REBOUILLAT, J. et al. Molecular Genetics of Rice Root Development. **Rice**, v. 2, n. 1, p. 15–34, 23 mar. 2009.
- REDDYA, A. S. N. et al. Complexity of the Alternative Splicing Landscape in Plants. **The Plant Cell**, v. 25, p. 3657–3683, 2013.
- REID, K. E. et al. Identification of tissue-specific, abiotic stress-responsive gene expression patterns in wine grape (*Vitis vinifera* L.) based on curation and mining of large-scale EST data sets. **BMC Plant Biology**, v. 6, n. 1, p. 27, 2006.
- RIECHMANN, J. L. Transcriptional regulation: a genomic overview. **The Arabidopsis book**, v. 1, p. e0085, 2002.
- RITCHIE, M. D. et al. Methods of integrating data to uncover genotype–phenotype interactions. **Nature Reviews Genetics**, v. 16, n. 2, p. 85–97, 13 jan. 2015.
- ROY, S.; BHATTACHARYYA, D. K.; KALITA, J. K. Reconstruction of gene co-expression network from microarray data using local expression patterns. **BMC bioinformatics**, v. 15 Suppl 7, n. Suppl 7, p. S10, 2014.
- SACHS, J. A. **A text book of botany.** [s.l.] Oxford, UK: Oxford University Press., 1882.

SAETZLER, K.; SONNENSCHNEIN, C.; SOTO, A. M. Systems biology beyond networks: generating order from disorder through self-organization. **Seminars in cancer biology**, v. 21, n. 3, p. 165–74, jun. 2011.

SALEEM, M. et al. Big linked cancer data: Integrating linked TCGA and PubMed. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 27, p. 34–41, 2014.

SARASWATI, R.; MATOH, T.; SEKIYA, J. Nitrogen Fixation of Sesbania rostrata: Contribution of Stem Nodules to Nitrogen Acquisition. **Soil Science and Plant Nutrition**, v. 38, p. 775–780, 1992.

SCHNEIDER, M. V.; JIMENEZ, R. C. Teaching the Fundamentals of Biological Data Integration Using Classroom Games. **PLoS Computational Biology**, v. 8, n. 12, p. e1002789, 27 dez. 2012.

SCHULZE, W.; USADEL, B. Quantitation in Mass-Spectrometry-Based Proteomics. **Annual Review of Plant Biology**, v. 61, p. 491–516, 2010.

SHAFFER, P. et al. BIOZON: a system for unification, management and analysis of heterogeneous biological data. **BMC Bioinformatics**, v. 7, n. 1, p. 71, 15 fev. 2006.

SHANNON, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. **Genome Research**, v. 13, n. 11, p. 2498–2504, 1 nov. 2003.

SHEVCHENKO, G. V.; BRYKOV, V. A.; IVANENKO, G. F. Specific features of root aerenchyma formation in Sium latifolium L. (Apiaceae). **Cytology and Genetics**, v. 50, n. 5, p. 293–299, 11 set. 2016.

SHOWALTER, A. M. Arabinogalactan-proteins: structure, expression and function. **Cellular and Molecular Life Sciences**, v. 58, n. 10, p. 1399–1417, set. 2001.

SILVA, G. B. et al. Cell wall polysaccharides from fern leaves: Evidence for a mannan-rich Type III cell wall in Adiantum raddianum. **Phytochemistry**, v. 72, p. 2352–2360, 2011.

SMIRNOFF, N.; CRAWFORD, R. M. M. Variation in the Structure and Response to Flooding of Root Aerenchyma in some Wetland Variation in the Structure and Response to Flooding of Root Aerenchyma in some Wetland Plants. **Source: Annals**

of **Botany Ann. Bot**, v. 51, n. 51, p. 237–249, 1983.

SOCCOL, C. R. et al. Bioethanol from ligno-celluloses: status and perspectives in Brazil. **Bioresource Technology**, v. 101, p. 4820–4825, 2010.

SONG, L.; LANGFELDER, P.; HORVATH, S. Comparison of co-expression measures: mutual information, correlation, and model based indices. **BMC Bioinformatics**, v. 13, n. 1, p. 328, 2012.

SOUZA, G. M. et al. The sugarcane signal transduction (SUCAST) catalogue: prospecting signal transduction in sugarcane. **Genetics and Molecular Biology**, v. 24, n. 1–4, p. 25–34, dez. 2001.

SREENIVASAN, T. V; AHLOOWALIA, B. S.; HEINZ, D. J. **Cytogenetics**. [s.l.] Sugarcane Improvement Through Breeding, Elsevier Press, Amsterdam, 1987.

SRIVASTAVA, A.; KUMAR, S.; RAMASWAMY, R. Two-layer modular analysis of gene and protein networks in breast cancer. **BMC Systems Biology**, v. 8, p. 81, 2014.

STAIGERA, D.; BROWN, J. W. S. Alternative Splicing at the Intersection of Biological Timing, Development, and Stress Responses. **The Plant Cell**, v. 25, p. 3640–3656, 2013.

STEEN, K. V. Travelling the world of gene-gene interactions. **Briefings in Bioinformatics**, v. 13, p. 1–19, 2012.

STUART, J. M. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. **Science**, v. 302, n. 5643, p. 249–255, 10 out. 2003.

SUBBAIAH, C. C.; SACHS, M. M. Molecular and Cellular Adaptations of Maize to Flooding Stress. **Annals of Botany**, v. 91, n. 2, p. 119–127, 1 jan. 2003.

SURAVAJHALA, P.; KOGELMAN, L. J. A.; KADARMIDEEN, H. N. Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare. **Genetics, selection, evolution : GSE**, v. 48, n. 1, p. 38, 2016.

SZKLARCZYK, D. et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. **Nucleic Acids Research**, 2014.

TAVARES, E. Q. DE P. **REGULAÇÃO DA DEGRADAÇÃO DA PAREDE CELULAR DURANTE A FORMAÇÃO DO AERÊNQUIMA EM RAÍZES DE CANA-DE-**

AÇÚCAR. [s.l.] Universidade de São Paulo, 2015.

TAVARES, E. Q. P.; DE SOUZA, A. P.; BUCKERIDGE, M. S. How endogenous plant cell wall degradation mechanisms can help achieve higher efficiency in saccharification of biomass. **Journal of Experimental Botany**, v. 66, p. 4133–4143, 2015.

THE ARABIDOPSIS GENOME INITIATIVE. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. **Nature**, v. 408, n. 6814, p. 796–815, 14 dez. 2000.

THOMAS, B. R. et al. New perspectives on the endo-beta-glucanases of glycosyl hydrolase Family 17. **International Journal of Biological Macromolecules**, v. 27, n. 2, p. 139–144, abr. 2000.

TOKIMATSU, T. et al. KaPPA-view: a web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. **Plant Physiology**, v. 7, p. 1289–1300, 2005.

TRAPNELL, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. **Nature Biotechnology**, v. 28, p. 511–515, 2010.

TRAPNELL, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-Seq. **Nature Biotechnology**, v. 31, p. 26–53, 2012a.

TRAPNELL, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nature Protocols**, v. 7, p. 562–578, 2012b.

TRAPNELL, C.; PACHTER, L.; SALZBERG, S. L. TopHat: discovering splice junctions with RNA-Seq. **Bioinformatics**, v. 25, p. 1105–1111, 2009.

TREWAVAS, A. A brief history of systems biology. “Every object that biology studies is a system of systems.” Francois Jacob (1974). **The Plant cell**, v. 18, n. 10, p. 2420–30, out. 2006.

TRIPATHI, R. et al. Next-generation sequencing revolution through big data analytics. **Frontiers in Life Science**, v. 9, n. 2, p. 119–149, 2 abr. 2016.

TYANOVA, S.; TEMU, T.; COX, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. **Nature Protocols**, v. 11, n. 12, p. 2301–

2319, 27 out. 2016.

TZVELEV, N. N. The system of grasses (Poaceae) and their evolution. **Botanical Review**, v. 55, p. 141–203, 1989.

VAN DAM, S. et al. Gene co-expression analysis for functional classification and gene–disease predictions. **Briefings in Bioinformatics**, p. 1–18, 2017.

VETTORE, A. L. et al. The libraries that made SUCEST. **Genetics and Molecular Biology**, v. 24, p. 1–7, 2001.

VETTORE, A. L. et al. Analysis and Functional Annotation of an Expressed Sequence Tag Collection for Tropical Crop Sugarcane. **Genome Research**, v. 12, p. 2725–2735, 2003.

VICENTE, F. F. et al. Assessing the gain of biological data integration in gene networks inference. **BMC Genomics**, v. 13, n. Suppl 6, p. S7, 2012.

VICENTINI, R. et al. Large-Scale Transcriptome Analysis of Two Sugarcane Genotypes Contrasting for Lignin Content. **PLOS ONE**, v. 10, n. 8, p. e0134909, 4 ago. 2015.

VOSS, T. C.; HAGER, G. L. Dynamic regulation of transcriptional states by chromatin and transcription factors. **Nature Reviews Genetics**, v. 15, n. 2, p. 69–81, 17 dez. 2013.

WAGNER, G. P.; KIN, K.; LYNCH, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. **Theory in Biosciences**, v. 131, n. 4, p. 281–285, 8 dez. 2012.

WALTON, J.; BANERJEE, G.; CAR, S. GENPLAT: an Automated platform for biomass enzyme discovery and cocktail optimization. **Journal of Visualized Experiments**, v. 56, p. 3314, 2011.

WANDELT, S. et al. Data Management Challenges in Next Generation Sequencing. **Datenbank-Spektrum**, v. 12, n. 3, p. 161–171, 1 nov. 2012.

WANG, H.; LAN, P.; SHEN, R. F. Integration of transcriptomic and proteomic analysis towards understanding the systems biology of root hairs. **PROTEOMICS**, v. 16, n. 5, p. 877–893, mar. 2016.

WANG, M. et al. Cellulolytic Enzyme Production and Enzymatic Hydrolysis for

Second-Generation Bioethanol Production. In: **Advances in biochemical engineering/biotechnology**. [s.l: s.n.]. v. 128p. 1–24.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Review Genetics**, v. 10, p. 5763, 2009.

WARDE-FARLEY, D. et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. **Nucleic Acids Research**, v. 38, n. Web Server, p. W214–W220, 1 jul. 2010.

WEI, H. et al. Transcriptional coordination of the metabolic network in Arabidopsis. **Plant physiology**, v. 142, n. 2, p. 762–74, out. 2006.

WESTON, D. J. et al. Comparative physiology and transcriptional networks underlying the heat shock response in *Populus trichocarpa*, *Arabidopsis thaliana* and *Glycine max*. **Plant, Cell & Environment**, v. 34, n. 9, p. 1488–1506, set. 2011.

WRAY, G. A. et al. **The evolution of transcriptional regulation in eukaryotes** *Molecular Biology and Evolution*, 2003. Disponível em: <http://www.indiana.edu/~hahnlab/Publications/Wray_et al2003.pdf>. Acesso em: 18 out. 2017

XIAO, Y.; POOVAIAH, C.; COLEMAN, H. D. Expression of Glycosyl Hydrolases in Lignocellulosic Feedstock: An Alternative for Affordable Cellulosic Ethanol Production. **BioEnergy Research**, v. 9, n. 4, p. 1290–1304, 22 dez. 2016.

YAMAUCHI, T. et al. Aerenchyma formation in crop species: A review. **Field Crops Research**, v. 152, p. 8–16, 2013.

YAMAUCHI, T. et al. Ethylene and reactive oxygen species are involved in root aerenchyma formation and adaptation of wheat seedlings to oxygen-deficient conditions. **Journal of experimental botany**, v. 65, n. 1, p. 261–73, jan. 2014.

YOO, Y.-H.; CHOI, H.-K.; JUNG, K.-H. Genome-wide identification and analysis of genes associated with lysigenous aerenchyma formation in rice roots. **Journal of Plant Biology**, v. 58, n. 2, p. 117–127, 15 abr. 2015.

ZHANG, B.; HORVATH, S. A General Framework for Weighted Gene Co-Expression Network Analysis. **Statistical Applications in Genetics and Molecular Biology**, v. 4, n. 1, 12 jan. 2005.

Apêndice

Figura 28: Descrição das 20 primeiras subcategorias de parede celular identificadas no RNA-Seq do segmento 2 (2º cm) da raiz de cana-de-açúcar. Essa é uma região da raiz onde o aerênquima começa a se formar.

Figura 29: Descrição das 20 primeiras subcategorias de parede celular identificadas no RNA-Seq do segmento 3 (3º cm) da raiz de cana-de-açúcar. Segmento da raiz com alta taxa de formação do aerênquima.

Figura 30: Descrição das 20 primeiras subcategorias de parede celular identificadas no RNA-Seq do segmento 4 (4º cm) da raiz de cana-de-açúcar. Nessa região o aerênquima já está formado.

Figura 31: Descrição das 15 primeiras subcategorias de parede celular identificadas na proteômica para os segmentos de 1 à 4 de raiz de cana-de-açúcar. O aerênquima começa a ser formado na transição do segundo com o terceiro segmento da raiz.

Figura 32: Visualização dos agrupamentos das redes de co-expressão do segmento 1 da raiz de cana-de-açúcar. Visualização da ferramenta SIT com layout Cose.

Figura 34: Visualização do agrupamento da rede de co-expressão do segmento 2 da raiz de cana-de-açúcar. Visualização da ferramenta SIT com layout Cose.

Figura 33: Visualização do agrupamento da rede de co-expressão do segmento 3 da raiz de cana-de-açúcar. Visualização da ferramenta SIT com layout Cose.

Figura 35: Visualização dos agrupamentos das redes de co-expressão do segmento 4 da raiz de cana-de-açúcar. Visualização da ferramenta SIT com layout Cose.

Tabela 12: Potenciais elementos efetores do desenvolvimento do aerênquima em raízes de cana-de-açúca, apontados na integração de dados. Valor da expressão gênica está em TPM (Log2). O resultado de integração mostra quantidade de layes integradas (1#, 2# ou 3#) versus a identificação da mesma (1=transcritoma, 2=proteoma, 3 enzimas).

Índice	SAS_ID	Valor de expressão gênica por condição analisada							SUCEST_Anotação	Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4			S1	S2	S3	S4
1	SCAGLR1021 F11.g	5.44			6.24			5.98	glycosyl hydrolase family 1	Glycosyl hydrolase family 1 C_degradation	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
2	SCCCCL4001 H11.g	6.66			6.92			6.46	beta-galactosidase	Beta- galactosidase H/P_degradatio n	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
3	SCEQHR108 2B01.g	4.83			5.40			4.92	glycosyl hydrolase family 1	Glycosyl hydrolase family 1 C_degradation	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
4	SCJLRT1017 C05.g	5.19			5.94			6.20	alpha-l- arabinofuranosidase	Alpha-L- arabinofuranosi dase H/P_degradatio n	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
5	SCQSRT1034 D03.g	5.72			5.72			5.75	alpha-l- arabinofuranosidase	Alpha-L- arabinofuranosi dase H/P_degradatio n	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
6	SCVPLB1020 G05.g	4.93			5.70			5.72	alpha-l- arabinofuranosidase	Alpha-L- arabinofuranosi dase H/P_degradatio n	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
7	SCVPRZ2042 G03.g	4.01		5.53				5.65	glycosyl hydrolase family 1	Glycosyl hydrolase family 1 C_degradation	3#1 2 3	3#1 2 3	3#1 2 3	3#1 2 3
8	SCCCCL4017 A06.g	6.17			6.25			6.56	glycosyl hydrolase family 1	Glycosyl hydrolase family 1 C_degradation	2#1 3	2#1 3	3#1 2 3	2#1 3

Índice	SAS_ID	Valor de expressão gênica por condição analisada								SUCEST_Anotação	Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4	S1			S2	S3	S4	
9	SCEQRT2025 H02.g	3.39			3.40			3.84	alpha-galactosidase	Alpha-galactosidase H_degradation	2#1 3	2#1 3	3#1 2 3	2#1 3	
10	SCJFAD1010 G05.b	4.55			4.08			4.02	alpha-L-arabinofuranosidase	Alpha-L-arabinofuranosidase H/P_degradation	2#1 3	2#1 3	3#1 2 3	2#1 3	
11	SCQGLR104 1C07.g	3.38			2.92				alpha-L-arabinofuranosidase	Alpha-L-arabinofuranosidase H/P_degradation	2#1 3	2#1 3	3#1 2 3	1#3	
12	SCUTLR1037 C07.g	2.61			2.96				alpha-galactosidase	Alpha-galactosidase H_degradation	2#1 3	2#1 3	3#1 2 3	2#2 3	
13	SCBFLR1083 E01.g	6.63	6.63					6.51	alpha-mannosidase (gh family 38)	none	2#1 2	2#1 2	2#1 2	2#1 2	
14	SCBGLR1003 C11.g	3.45			4.06			4.15	alpha-mannosidase (gh family 38)	none	2#1 2	2#1 2	2#1 2	2#1 2	
15	SCCCLR1001 A05.g	9.61	9.61					9.22	sucrose synthase	none	2#1 2	2#1 2	2#1 2	2#1 2	
16	SCCCLR1048 D07.g	6.49			6.67			6.19	phenylalanine ammonia-lyase?(pal)	none	2#1 2	2#1 2	2#1 2	2#1 2	
17	SCCCRT1002 G03.g	6.33	6.33			6.09			glycosyl hydrolase family 3 protein (xylosidase)	Glycosyl hydrolase family 3 protein (Xylosidase) H_degradation	2#1 2	2#1 2	2#1 2	2#1 2	
18	SCCCRZ1002 G07.g	9.99			10.22			9.98	sucrose synthase	none	2#1 2	2#1 2	2#1 2	2#1 2	
19	SCCCRZ1C0 2F12.g	5.05			5.44			5.02	annexin	none	2#1 2	2#1 2	2#1 2	2#1 2	

Índice	SAS_ID	Valor de expressão gênica por condição analisada								SUCEST_Anotação	Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4	S1			S2	S3	S4	
20	SCEPCL6023 F02.g	2.32			2.06				sucrose synthase	none	2#1 2	2#1 2	2#1 2	1#2	
21	SCEPLR1051 B04.g	5.08			5.30	5.30			fasciclin-like arabinogalactan protein	Fasciclin-like arabinogalactan protein P_biosynthesis	2#1 2	2#1 2	2#1 2	2#1 2	
22	SCEPRT2043 B01.g	4.79			5.94			5.49	expansin	Expansin E_Expansin	2#1 2	2#1 2	2#1 2	2#1 2	
23	SCEQRT1025 C10.g	6.49			6.44	6.44			beta-1,3-glucanase glycosyl hydrolases family 17	beta-1-3- glucanase Glycosyl hydrolases family 17 H_degradation	2#1 2	2#1 2	2#1 2	2#1 2	
24	SCEZLB1007 A09.g	4.36			5.07			4.92	glycosyl hydrolase family 3 protein	glycosyl hydrolase family 3 protein H/C_degradatio n	2#1 2	2#1 2	2#1 2	2#1 2	
25	SCEZLR1052 C03.g	8.91	8.91					8.67	sucrose synthase	none	2#1 2	2#1 2	2#1 2	2#1 2	
26	SCEZRZ3099 D06.g	4.98			4.21			3.89	sucrose synthase	none	2#1 2	2#1 2	2#1 2	2#1 2	
27	SCJFLR1013 E04.g	7.22			7.08			7.00	udp-glucose dehydrogenase	none	2#1 2	2#1 2	2#1 2	2#1 2	
28	SCJFRT1008 G05.g	3.88			4.29			4.55	beta-1,3-glucanase glycosyl hydrolases family 17	beta-1-3- glucanase Glycosyl hydrolases family 17 H_degradation	2#1 2	2#1 2	2#1 2	2#1 2	

Índice	SAS_ID	Valor de expressão gênica por condição analisada							SUCEST_Anotação	Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4			S1	S2	S3	S4
29	SCJFRT1061 B07.g	4.84			5.01			4.43	caffeic acid 3-o- methyltransferase	caffeic acid 3- O- methyltransfera se L_biosynthesis	2#1 2	2#1 2	2#1 2	2#1 2
30	SCJFRZ1007 B01.g	3.20			4.81			3.14	sucrose synthase	none	2#1 2	2#1 2	2#1 2	2#1 2
31	SCJFRZ2029 F03.g	4.47			4.77			4.26	glycoside hydrolase family 31 (alpha- glucosidase or alpha- xylosidase)	Glycoside hydrolase family 31 (alpha- glucosidase or alpha- xylosidase) H_degradation	2#1 2	2#1 2	2#1 2	2#1 2
32	SCJLRT1023 B09.g	7.75			6.84			7.17	caffeic acid 3-o- methyltransferase	caffeic acid 3- O- methyltransfera se L_biosynthesis	2#1 2	2#1 2	2#1 2	2#1 2
33	SCQGLB103 8B02.g	7.35			7.66			7.23	sucrose synthase	none	2#1 2	2#1 2	2#1 2	2#1 2
34	SCRFLR1012 F12.g	8.09			7.95			7.29	caffeic acid 3-o- methyltransferase	caffeic acid 3- O- methyltransfera se L_biosynthesis	2#1 2	2#1 2	2#1 2	2#1 2
35	SCSBLB1033 B10.g	4.36			4.03			4.07	dirigent-like protein	none	2#1 2	2#1 2	2#1 2	2#1 2
36	SCUTLR2030 G02.g	7.47			7.52	7.52			endo-1,4-beta- glucanase glycosyl hydrolase family 9	endo-1-4-beta- glucanase Glycosyl hydrolase family 9 H/C_degradatio n	2#1 2	2#1 2	2#1 2	2#1 2

Índice	SAS_ID	Valor de expressão gênica por condição analisada								SUCEST_Anotação	Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4	S1			S2	S3	S4	
37	SCVPST1060 D08.g	2.13			2.21			2.07	laccase-like protein	laccase-like protein L_degradation	2#1 2	1#1	2#1 2	1#1	
38	SCACLR1129 A05.g	4.02			3.51			3.86	fasciclin-like arabinogalactan protein	Fasciclin-like arabinogalactan protein P_Pectin	1#1	1#1	2#1 2	1#1	
39	SCACLR2022 B05.g	5.45			5.81			4.87	glucosyltransferase like protein	glucosyltransferase like protein H/C_biosynthesis	1#1	1#1	2#1 2	1#1	
40	SCAGLR1064 F05.g	7.37			7.25			7.23	cellulose synthase (cesa)	Cellulose synthase (CesA) C_biosynthesis	1#1	1#1	2#1 2	1#1	
41	SCAGRT2038 D08.g	3.29			3.14			3.29	pectinacetylesterase family protein	Pectinacetylesterase family protein P_degradation	1#1	1#1	2#1 2	1#1	
42	SCAGRT2039 D04.g	3.87			3.28				laccase-like protein	laccase-like protein L_degradation	1#1	1#1	2#1 2	none	
43	SCBFRZ2017 F10.g	2.49			2.53			2.18	glycosyl transferase family 8 (cell wall metabolism, starch metabolism or lipopolysaccharide biosynthesis)	Glycosyl transferase family 8 (cell wall metabolism- starch metabolism or lipopolysaccharide biosynthesis) H_biosynthesis	1#1	1#1	2#1 2	1#1	
44	SCBFST3136 F06.g	4.33			4.55			3.41	pectinacetylesterase family protein	Pectinacetylesterase family protein P_degradation	1#1	1#1	2#1 2	1#1	
45	SCBGST310 6D08.g	4.75			5.04			5.42	isoflavone reductase/pinoresinol reductase (lignan biosynthesis)	none	1#1	1#1	2#1 2	2#1 2	
46	SCCCCL3120 E05.g	5.84		7.29				6.87	expansin	Expansin E_Expansin	1#1	1#1	2#1 2	1#1	

Índice	SAS_ID	Valor de expressão gênica por condição analisada								SUCEST_Anotação	Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4	S1			S2	S3	S4	
47	SCCCCL3140 B03.g	5.89			6.06	6.06				xylose isomerase	none	1#1	1#1	2#1 2	1#1
48	SCCCCL4007 E06.g	2.47			2.17			1.81		glucosyltransferase	glucosyltransferase H/C_biosynthesis	1#1	1#1	2#1 2	1#1
49	SCCCCL4007 H11.g	4.69			5.23			5.98		udp-d- xylose/glucose/galactose/arabinose epimerase	none	1#1	1#1	2#1 2	1#1
50	SCCCLB1026 G03.g	6.62			6.60			6.09		1,3-beta-glucan synthase (putative callose synthase)	1-3-beta-glucan synthase (putative callose synthase) H_biosynthesis	1#1	1#1	2#1 2	1#1
51	SCCCLR1065 F07.g	4.48			4.89			4.78		cellulose synthase (cesa)	Cellulose synthase (CesA) C_biosynthesis	1#1	1#1	2#1 2	1#1
52	SCCCLR1C0 6G07.g	5.67			5.86			5.21		sucrose-phosphate synthase	none	1#1	1#1	2#1 2	1#1
53	SCCCLR2C0 2H01.g	4.41			5.09	5.09				cellulose synthase (cesa)	Cellulose synthase (CesA) C_biosynthesis	1#1	1#1	2#1 2	1#1
54	SCCCRT2004 C04.g	3.27			3.21			1.34		beta-1,3-glucanase glycosyl hydrolases family 17	beta-1-3-glucanase Glycosyl hydrolases family 17 H_degradation	1#1	1#1	2#1 2	1#1

Índice	SAS_ID	Valor de expressão gênica por condição analisada							SUCEST_Anotação	Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4			S1	S2	S3	S4
55	SCCCRZ1001 B12.g	6.92			7.10			7.53	rhamnose synthase	none	1#1	1#1	2#1 2	1#1
56	SCCCRZ2002 G07.g	8.82	8.82					8.55	cellulose synthase (cesa)	Cellulose synthase (CesA) C_biosynthesis	1#1	1#1	2#1 2	1#1
57	SCCCRZ2C0 1A04.g	4.31		5.10				5.16	cinnamoyl-coa reductase (ccr)	Cinnamoyl CoA reductase (CCR) L_biosynthesis	1#1	1#1	2#1 2	1#1
58	SCCCRZ2C0 1F05.g	5.42			5.66			4.99	cinnamyl alcohol dehydrogenase (cad)	Cinnamyl alcohol dehydrogenase (CAD) L_biosynthesis	1#1	1#1	2#1 2	1#1
59	SCCCST3006 B11.g	3.11			3.30				glycosyl transferase family 8 (cell wall metabolism, starch metabolism or lipopolysaccharide biosynthesis)	Glycosyl transferase family 8 (cell wall metabolism- starch metabolism or lipopolysacchari de biosynthesis) H_biosynthesis	1#1	1#1	2#1 2	none
60	SCEPRZ1009 A11.g	3.17			3.15			2.74	1,3-beta-glucan synthase (putative callose synthase)	1-3-beta-glucan synthase (putative callose synthase) H_biosynthesis	1#1	1#1	2#1 2	1#1

Índice	SAS_ID	Valor de expressão gênica por condição analisada								SUCEST_Anotação	Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4	S1			S2	S3	S4	
61	SCEPRZ1011A02.g	3.58			4.21	4.21				cinnamyl alcohol dehydrogenase (cad)	Cinnamyl alcohol dehydrogenase (CAD) L_biosynthesis	1#1	1#1	2#1 2	1#1
62	SCEPSB1135A04.g	3.18			3.91			4.10		cinnamoyl-coa reductase (ccr)	Cinnamoyl CoA reductase (CCR) L_biosynthesis	1#1	1#1	2#1 2	1#1
63	SCEQLB1066C05.g	3.43			2.61					glucosyltransferase	glucosyltransferase H/C_biosynthesis	1#1	1#1	2#1 2	none
64	SCEQLR1029D03.g	3.51			4.15	4.15				cellulose synthase (cesa)	Cellulose synthase (CesA) C_biosynthesis	1#1	1#1	2#1 2	1#1
65	SCEZLR1031G01.g	6.25			6.48	6.48				udp-d-xylose/glucose/galactose/arabinose epimerase	none	1#1	1#1	2#1 2	1#1
66	SCEZRT2017G09.g	5.44		7.03		7.03				cellulose synthase like (csl)	Cellulose synthase like (CSL) H_biosynthesis	1#1	1#1	2#1 2	1#1
67	SCJFLR1017B09.g	7.38			7.41			7.42		cellulose synthase (cesa)	Cellulose synthase (CesA) C_biosynthesis	1#1	1#1	2#1 2	1#1
68	SCJFLR1073B10.g	6.29			6.48			6.09		1,3-beta-glucan synthase (putative callose synthase)	1-3-beta-glucan synthase (putative callose synthase) H_biosynthesis	1#1	1#1	2#1 2	1#1

Índice	SAS_ID	Valor de expressão gênica por condição analisada								SUCEST_Anotação	Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4	S1			S2	S3	S4	
69	SCJFRT2055 E11.b	5.16			6.16			6.27	expansin	Expansin E_Expansin	1#1	1#1	2#1 2	1#1	
70	SCJFRZ2005 H01.g	1.95			3.06				pectate lyase	Pectate lyase P_degradation	1#1	1#1	2#1 2	none	
71	SCJFRZ2014 H02.g	3.42			3.10			2.33	expansin	Expansin E_Expansin	1#1	1#1	2#1 2	1#1	
72	SCJFRZ3C01 G05.g	4.22			4.67			3.19	alpha-mannosidase (gh family 38)	none	1#1	1#1	2#1 2	1#1	
73	SCJFRZ3C05 E03.b	3.44			3.98			5.62	isoflavone reductase/pinoresinol reductase (lignan biosynthesis)	none	1#1	1#1	2#1 2	1#1	
74	SCJLLR1033 F06.g	3.73		4.54				4.10	1,3-beta-glucan synthase (putative callose synthase)	1-3-beta-glucan synthase (putative callose synthase) H_biosynthesis	1#1	1#1	2#1 2	1#1	
75	SCJLLR1106 H06.g	4.11			4.49	4.49			cellulose synthase (cesa)	Cellulose synthase (CesA) C_biosynthesis	1#1	1#1	2#1 2	1#1	
76	SCJLRT2050 C09.g	4.65			5.21			5.22	caffeoyl-coa 3-o-methyltransferase	caffeoyl-CoA 3-O-methyltransferase L_biosynthesis	1#1	1#1	2#1 2	1#1	
77	SCMCRT210 7D08.g	3.46			2.76				expansin	Expansin E_Expansin	1#1	1#1	2#1 2	none	
78	SCQGFL4080 C07.g	6.30			5.98			5.11	1,3-beta-glucan synthase (putative callose synthase)	1-3-beta-glucan synthase (putative callose synthase) H_biosynthesis	1#1	1#1	2#1 2	1#1	

Índice	SAS_ID	Valor de expressão gênica por condição analisada								SUCEST_Anotação	Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4	S1			S2	S3	S4	
79	SCQSRT1034 C09.g	6.96			6.94			7.33	caffeoyl-coa 3-o-methyltransferase	caffeoyl-CoA 3-O-methyltransferase L_biosynthesis	1#1	1#1	2#1 2	1#1	
80	SCQSRT1036 E09.g	3.02			2.70			2.94	putative p-coumarate 3-hydroxylase (c3h)	putative p-Coumarate 3-hydroxylase (C3H) L_biosynthesis	1#1	1#1	2#1 2	1#1	
81	SCQSST103 7B03.g	3.12			3.51	3.51			laccase-like protein	laccase-like protein L_degradation	1#1	1#1	2#1 2	1#1	
82	SCRFHR1008 D11.g	4.04			3.48				beta-1,3-glucanase glycosyl hydrolases family 17	beta-1-3-glucanase Glycosyl hydrolases family 17 H_degradation	1#1	1#1	2#1 2	none	
83	SCRFRZ3054 G02.g	4.87			4.25			4.34	laccase-like protein	laccase-like protein L_degradation	1#1	1#1	2#1 2	1#1	
84	SCRLST3163 A02.g	4.39			4.11			3.49	cellulose synthase (cesa)	Cellulose synthase (CesA) C_biosynthesis	1#1	1#1	2#1 2	1#1	
85	SCRUHR107 5E08.g	3.81			4.15			4.47	xyloglucan endotransglycosylase/hydrolase (xth)	Xyloglucan endotransglycosylase/hydrolase (XTH) H_degradation	1#1	1#1	2#1 2	1#1	
86	SCSBAD105 4A07.g	2.38			2.59			2.60	cinnamyl alcohol dehydrogenase (cad)	Cinnamyl alcohol dehydrogenase (CAD) L_biosynthesis	1#1	1#1	2#1 2	1#1	

Índice	SAS_ID	Valor de expressão gênica por condição analisada								SUCEST_Anotação	Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4	S1			S2	S3	S4	
87	SCSBSB1053C05.g	4.69			5.27	5.27				cellulose synthase (cesa)	Cellulose synthase (CesA) C_biosynthesis	1#1	1#1	2#1 2	1#1
88	SCSF4084A01.b	3.45			2.85					laccase-like protein	laccase-like protein L_degradation	1#1	1#1	2#1 2	none
89	SCSF4084B02.b	5.61			5.62			5.65		cinnamoyl-coa reductase (ccr)	Cinnamoyl CoA reductase (CCR) L_biosynthesis	1#1	1#1	2#1 2	1#1
90	SCUTRZ2008D10.g	5.23			5.39					glycosyltransferase family 47	glycosyltransferase family 47 H/P_biosynthesis	1#1	1#1	2#1 2	none
91	SCUTSD2027H09.g	6.08	6.08				5.85			cellulose synthase like (csl)	cellulose synthase like (CSL) H_biosynthesis	1#1	1#1	2#1 2	1#1
92	SCVPCL6064E05.g	4.33			3.95			4.31		cellulose synthase (cesa)	Cellulose synthase (CesA) C_biosynthesis	1#1	1#1	2#1 2	1#1
93	SCVPHR1092E06.g	3.85		5.04		5.04				cellulose synthase like (csl)	Cellulose synthase like (CSL) H_biosynthesis	1#1	1#1	2#1 2	1#1
94	SCVPRT2075E11.g	3.27			4.08			4.19		cellulose synthase (cesa)	Cellulose synthase (CesA) C_biosynthesis	1#1	1#1	2#1 2	1#1
95	SCVPST1061C06.g	2.63		4.77		4.77				sucrose synthase	none	1#1	1#1	2#1 2	1#1

Índice	SAS_ID	Valor de expressão gênica por condição analisada								SUCEST_Anotação	Domínio de parede celular	Integração no segmentos			
		S2_S1xS2	S2_S2xS3	S3_S2xS3	S3	S3_S3xS4	S4_S3xS4	S4	S1			S2	S3	S4	
96	SCACST3159 H02.g	3.35								fasciclin-like arabinogalactan protein	Fasciclin-like arabinogalactan protein P_biosynthesis	1#1	1#1	none	none
97	SCVPCL6044 H01.g	3.37								galactosyltransferase	Galactosyltrans ferase H/P_biosynthes is	1#1	1#1	none	none