

NATÁLIA FARAJ MURAD

**“REDES DE REGULAÇÃO GÊNICA DO
METABOLISMO DE SACAROSE EM CANA-DE-
AÇÚCAR UTILIZANDO REDES BAYESIANAS”**

CAMPINAS
2013

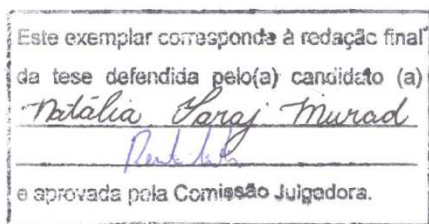


UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

NATÁLIA FARAJ MURAD

**“REDES DE REGULAÇÃO GÊNICA DO METABOLISMO DE
SACAROSE EM CANA-DE-AÇÚCAR UTILIZANDO REDES
BAYESIANAS”**



Tese apresentada ao Instituto de Biologia da UNICAMP para obtenção do Título de MESTRA em Genética e Biologia Molecular, na área de Genética Vegetal e Melhoramento.

Orientador: Prof. Dr. RENATO VICENTINI DOS SANTOS

CAMPINAS
2013

**FICHA CATALOGRÁFICA ELABORADA POR MARA JANAINA DE OLIVEIRA -
CRB8/6972
BIBLIOTECA DO INSTITUTO DE BIOLOGIA - UNICAMP**

Murad, Natália Faraj, 1989-

M931r Redes de regulação gênica do metabolismo de sacarose em cana-de-açúcar utilizando redes bayesianas / Natália Faraj Murad. – Campinas, SP: [s.n.], 2013

Orientador: Renato Vicentini dos Santos

Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Redes Reguladoras de Genes. 2. Cana-de-açúcar. 3. Redes bayesianas. I. Santos, Renato Vicentini. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em Inglês: Gene regulatory networks of the sucrose metabolism in sugarcane using bayesian networks

Palavras-chave em Inglês:

Gene Regulatory Networks

Sugar-cane

Bayesian networks

Área de concentração: Genética Vegetal e Melhoramento

Titulação: Mestra em Genética e Biologia Molecular

Banca examinadora:

Renato Vicentini dos Santos [Orientador]

Henrique Marques-Souza

Júlio Sílvio de Sousa Bueno Filho

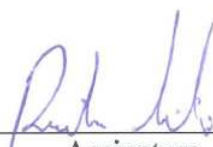
Data da defesa: 20-02-2013

Programa de Pós Graduação: Genética e Biologia Molecular

Campinas, 20 de fevereiro de 2013

BANCA EXAMINADORA

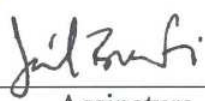
Prof. Dr. Renato Vicentini Dos Santos (orientador)


Assinatura

Prof. Dr. Henrique Marques Barbosa De Souza


Assinatura

Prof. Dr. Júlio Silvio De Sousa Bueno Filho


Assinatura

Dr. Marcelo Mendes Brandão

Assinatura

Dra. Paula Regina Kuser Falcão

Assinatura

*Aos meus pais Hiram (in memoriam)
e Cida, à minha irmã Isabela,
e ao Sergio pelo apoio, força, incentivo,
amor e companheirismo.*

AGRADECIMENTOS

À Unicamp e ao Instituto de Biologia pela oportunidade de realização do curso.

À CAPES e à FAPESP pela bolsa de estudo e suporte financeiro do projeto deste trabalho.

Ao Prof. Renato pela orientação, ensinamentos e paciência.

Aos professores que aceitaram participar da banca, em especial, ao Prof. Júlio por ter me apresentado à área de estatística e análise de expressão gênica durante a graduação.

Aos Profs. Henrique e Roberto pelas críticas e sugestões na pré-banca.

À Lourdes e à Sílvia pela disposição em ajudar com informações e trâmite de documentos.

A todas as pessoas que direta ou indiretamente colaboraram com a realização desse trabalho.

Redes de regulação gênica do metabolismo de sacarose em cana-de-açúcar utilizando redes bayesianas

RESUMO

Palavras-chave: Redes Reguladoras de Genes, Cana-de-açúcar, Redes bayesianas.

A cana-de-açúcar é uma das mais importantes plantas cultivadas no Brasil, que é o maior produtor e exportador mundial. Seu valor econômico é devido principalmente a sua capacidade de estocar sacarose nos colmos. Os padrões de expressão gênica podem regular processos de desenvolvimento da planta e influenciar no acúmulo de sacarose em tecidos de reserva. A regulação desses padrões ocorre através de complexos sistemas de interações entre muitos genes e seus produtos, resultando em uma complexa rede de regulação gênica. Modelos gráficos probabilísticos têm sido amplamente utilizados para inferência e representação dessas redes. Dentre eles, as redes bayesianas são o principal por serem consideradas o método mais flexível e também requererem um número reduzido de parâmetros para a descrição do modelo. Sendo assim, este estudo utilizou a metodologia de redes bayesianas para inferência de interações regulatórias entre genes de metabolismo e sinalização de sacarose a partir de dados de expressão gênica, obtidos através de *microarrays*, disponíveis no *Gene Expression Omnibus* (GEO). As redes foram obtidas através de *softwares* para inferência de redes e então analisadas quanto aos genes que as compõem e padrões de expressão. Os genes foram agrupados em clusters considerando-se seus padrões de co-expressão. Os genes mais representados no cluster da enzima sacarose fosfato sintase (SPS) em cana são genes relacionados à tradução, ligação ao DNA e genes de função desconhecida, enquanto os menos representados são de fotossíntese, resposta a hormônios, e outros eventos metabólicos. A rede do cluster da SPS apresentou sete genes principais (*hubs*) que aparentam ter um importante papel dentro do cluster. Foi obtida também uma rede considerando genes selecionados em estudos com experimentos de *microarrays* previamente publicados. Uma dessas redes possui 136 genes e apresentou 6 genes principais, sendo que a maioria deles são de fotossíntese. Na rede considerando genes diferencialmente expressos nesses experimentos (265 genes), genes que pertencem à mesma categoria funcional tenderam

a sofrer regulação por um único gene em comum, formando grupos de funções semelhantes em cada *hub*.

Gene regulatory networks of the sucrose metabolism in sugarcane using Bayesian networks

ABSTRACT

Keywords: Gene Regulatory Networks, Sugar-cane, Bayesian networks.

Sugarcane is one of the most important plants cultivated in Brazil which is the world's largest producer and exporter. Its economic yield is mainly due to its high sucrose content. The patterns of gene expression may regulate processes of plant development and influence the accumulation of sucrose by storage tissues. The regulation of these patterns occurs through complex systems of interactions between many genes and their products, resulting in a complex gene regulatory network. Probabilistic graphical models have been widely used for inference and representation of these networks. Among them, Bayesian networks are the main for being considered to be the most flexible method and also requiring a reduced number of parameters to the model description. Then, this work has used the Bayesian network methodology for inference of regulatory interactions between signaling and sucrose metabolism genes from gene expression data, obtained from microarrays, available on Gene Expression Omnibus (GEO). Networks were generated by networks inference softwares, and then analyzed observing their composing genes and expression patterns. The genes were grouped considering their coexpression patterns. The most represented genes in the saccharose phosphate syntase (SPS) cluster are related with translation, DNA binding and unknown function genes while the least represented are of photosynthesis, hormone response and other metabolic events. The SPS cluster network presented 7 main *hubs* that seem to play an important role in the cluster. It was also obtained a network considering genes selected from studies with microarray experiments previously published. One of these gene networks has 136 genes and it presented 6 main genes, being the most of them are from photosynthesis. In the network considering differential expressed in this experiments, genes that are from the same functional category tended to suffer regulation for one unique common gene, forming groups of genes with similar function on each *hub*.

SUMÁRIO

LISTA DE FIGURAS	xi
LISTA DE TABELAS.....	xiv
1. INTRODUÇÃO	1
2. OBJETIVOS.....	3
2.1. Objetivo geral.....	3
2.2. Objetivos específicos	3
3. REVISÃO BIBLIOGRÁFICA.....	4
3.1. Cana-de-açúcar	4
3.1.1. Importância econômica da cana-de-açúcar	5
3.1.2. Fisiologia da cana-de-açúcar	6
3.2. Sinalização por sacarose	7
3.3. Síntese de sacarose, translocação e acúmulo.	8
3.4. Relações fonte-dreno	12
3.5. Redes de regulação gênica	14
3.6. Teoria da Probabilidade	20
3.7. Modelos gráficos probabilísticos.....	22
3.8. Redes bayesianas	25
3.9. Softwares para redes bayesianas	30
3.9.1. <i>BaNJO: Bayesian Network Inference with Java Objects</i>	30
3.9.2. <i>BNFinder</i>	33
3.9.3. <i>PEBL: Python Environment for Bayesian Learning</i>	34
3.9.4. <i>R: Deal e BNArray</i>	35
4. MATERIAL E MÉTODOS	38
4.1. Conjunto de dados de <i>microarrays</i>	38
4.2. Experimento de metabolismo de parede celular	39
4.3. Experimento de manipulação da fotossíntese	40
4.4. Seleção dos genes de interesse pelo método <i>Heuristic Cluster Chiseling Algorithm</i> ...	41
4.5. Conjunto de dados	44
4.6. Normalização dos dados <i>Affymetrix</i>	45
4.7. Experimento de manipulação fonte e dreno.....	45
4.8. Montagem e análise inicial dos dados de RNA-seq	47
4.9. <i>Sugarcane Systems Biology Workplace</i>	49
4.10. Construção das redes bayesianas.....	51

4.11.	Visualização das redes	51
4.12.	Análise de ontologia e anotação dos genes	51
4.13.	Análise do padrão de expressão das redes	52
5.	RESULTADOS E DISCUSSÃO	53
5.1.	Redes de regulação gênica	53
5.2.	Análise do padrão de expressão das redes	79
5.3.	RNA-seq	81
6.	CONSIDERAÇÕES FINAIS	85
	REFERÊNCIAS BIBLIOGRÁFICAS	87

LISTA DE FIGURAS

- Figura 3-1 - Esquema da síntese de sacarose.** Dentro do cloroplasto ocorre o Ciclo de Calvin que produzirá trioses fosfato. Essas trioses poderão ser armazenadas sob forma de amido no cloroplasto ou serem transportadas para o citosol e seguir a via para síntese de sacarose. (Adaptado de Johansson, 2003). **11**
- Figura 3-2 - Esquema da planta de cana-de-açúcar e regulações fonte-dreno ao longo de seu desenvolvimento.** A) Planta jovem em desenvolvimento vegetativo, órgão em crescimento funcionando como dreno. B) 10-12 meses, planta adulta havendo produção de açúcares na fonte e acúmulo nos tecidos de reserva. Em A, com a planta em estado de desenvolvimento, folhas jovens funcionam como dreno enquanto em B, as folhas já desenvolvidas funcionam como fonte e os entrenós como órgãos de reserva, no caso drenos. (Adaptado de Ribeiro, 2012). **14**
- Figura 3-3 - Esquema da regulação da expressão gênica.** 1) Proteínas induzindo ou inibindo a transcrição. 2) Proteínas induzindo ou inibindo a tradução. 3) Proteínas induzindo vias metabólicas. 4) Vias metabólicas induzindo a expressão de proteínas. (Adaptado de Lima & Barrera, 2008). **15**
- Figura 3-4 - Rede booleana para morfogênese de flores em *A. thaliana*.** A) Fenótipos do tipo selvagem, mutantes *ap1* apenas com carpelos e pétalas, mutantes *ap3* ou *pi* com sépalas e carpelos, mutantes *ag* com sépalas e pétalas e estames e carpelos ausentes. B) Esquema da rede de regulação gênica para morfogênese da flor. Notação $\neg$ indica relações de repressão e $\rightarrow$ indica indução. (Adaptado de Mendonza & Alvarez-Buylla, 1998). **18**
- Figura 3-5 - Exemplo de uma rede bayesiana.** A figura apresenta uma rede bayesiana hipotética composta pelo conjunto de nós $V = \{A, B, C, D, E, F\}$ e arcos $E = \{(A, B), (A, C), (B, D), (C, D), (D, E), (D, F), (C, F)\}$. Aplicando as relações de independência representadas no grafo, a probabilidade conjunta pode ser escrita como $P(A, B, C, D, E, F) = P(A) P(B|A) P(C|A) P(D|B, C) P(E|D) P(F|D, C)$. (Werhli, 2007). **27**
- Figura 3-6 – Algoritmo de busca (*Searcher*) e seus componentes.** O *Proposer* sugere uma rede a partir dos dados iniciais, é verificado se essa rede contém ciclos pelo *Cycle Checker*, o *Evaluator* calcula o *score* da rede e o *Decider* avalia se o *score* é melhor que a rede atual ou não, e assim decide se continua com a rede atual ou a substitui pela nova. **32**
- Figura 3-7 - Representação do algoritmo.** Gera todas as combinações de trios possíveis a partir dos genes do conjunto de dados. Depois é selecionada a melhor rede bayesiana para cada conjunto de três genes, então essas redes são unidas em uma única rede incluindo todos os genes. (Adaptado de Watanabe *et al.*, 2012). **37**
- Figura 4-1 - Princípios do *Heuristic Cluster Chiseling Algorithm* (HCCA).** HCCA consiste de cinco passos; um: geração da NVN para cada nó na rede levando em conta os n-passos da vizinhança. 2: remoção dos nós mais conectados com nós fora da NVN resultante, 3: geração de SPCs, 4: seleção de clusters da população de SPCs, 5: remoção dos nós clusterizados da rede e retorno ao passo 1. (Adaptado de Mutwill, 2010). **42**

Figura 4-2 - Node Vicinity Networks. Exemplo da vizinhança de um gene (em cinza) considerando todos os genes a dois passos dele (http://aranet.mpimp-golm.mpg.de/aranet/FAQ).....	43
Figura 4-3 - Lista dos genes relacionados com <i>AtCesa1</i>. Genes ordenados e ranqueados de acordo com o valor do coeficiente de correlação de Pearson; em azul <i>ATCesa1</i> e em verde <i>Ctl1</i> (http://aranet.mpimp-golm.mpg.de/aranet/FAQ).....	44
Figura 4-4 - Lista dos genes relacionados com <i>Ctl1</i>. Genes ordenados e ranqueados de acordo com o valor do coeficiente de correlação de Pearson; em azul <i>ATCesa1</i> e em verde <i>Ctl1</i> (http://aranet.mpimp-golm.mpg.de/aranet/FAQ).....	44
Figura 4-5 - Desenho do experimento para manipulação da relação fonte-dreno. Em azul é indicado o período de incidência de luz, sendo que em cinza esta o período em que as plantas estão cobertas (com exceção da folha +3). Todos os tratamentos serão coletados ao final do experimento.	47
Figura 4-6 - Análise de qualidade dos <i>reads</i> sequenciados após trimagem inicial. Nota-se a elevada qualidade dos <i>reads</i> a serem utilizados na montagem do transcriptoma de cana-de-açúcar.	48
Figura 4-7 - Diagrama esquemático da montagem do transcriptoma. Representação da estratégia utilizada para a montagem de novo do transcriptoma de cana-de-açúcar.	49
Figura 4-8 - Interface principal para acesso ao sistema Sugarcane Systems Biology Workplace. (A), interface gráfica do site (B) rede da divisão de todos genes do <i>microarray</i> de cana-de-açúcar em clusters (C) rede bayesiana dos genes diferencialmente expressos no experimento de relação fonte-dreno.....	50
Figura 5-1 - Rede de correlação. Rede obtida pelo método de coexpressão para todas as sondas do <i>microarray</i> de cana da <i>Affymetrix</i>	53
Figura 5-2 - Clusters da rede de correlação. Rede representando os clusters obtidos pelo método de HCCA a partir da rede de correlação para todas as sondas do <i>microarray</i> de cana-de-açúcar da <i>Affymetrix</i>	54
Figura 5-3 - Rede bayesiana para o cluster da enzima SPS. Rede com 109 nós e 128 arcos. Cores dos nós representam sua função. Nós circulares representam os genes, variáveis contínuas e nós em forma de losango representam os tratamentos, variáveis discretas. Enzimas SPS indicadas em formato hexagonal.	55
Figura 5-4 - Genes da enzima SPS representados na rede. A) Gene BU925724 e os genes com os quais possui ligação direta. B) Gene CA232784 e os com os quais possui ligação direta.....	57
Figura 5-5 - Hubs presentes na rede do cluster de cana. Os sete <i>hubs</i> encontrados na rede do cluster da SPS em cana-de-açúcar.	59
Figura 5-6 - Rede dos genes selecionados dentro do cluster da SPS. Rede dos genes com HRR = 30 e NVN = 1 e HRR = 10 e NVN = 20, totalizando 32 nós sendo 7 discretos e 31 arcos.	60
Figura 5-7 - Rede bayesiana para os 136 genes definidos no trabalho de McCormick (2007). Rede de regulação 136 genes e sete tratamentos, 195 arcos. Cores indicam funções dos genes.....	64

Figura 5-8 - Hub CA068881 interagindo com 33 genes. CA068881 é um gene que apresentou um elevado número de interações constituindo um <i>hub</i>	67
Figura 5-9 - Hubs CA174395 e 50198865-54. Genes que apresentaram um elevado número de interações constituindo dois <i>hubs</i> conectados.....	68
Figura 5-10 - Hub CA119563. Genes que interagem com CA119563.	69
Figura 5-11 - Hub BU103687 e 50198865-28. Representação do gene BU103687 e suas interações e também do 50198865-28.	70
Figura 5-12 - Hubs CA248491 e CA221011. Genes CA248491 e CA221011 e os genes que interagem com eles.	71
Figura 5-13 - Quantificação dos carboidratos nas duas cultivares de cana-de-açúcar. Quantificação dos carboidratos galactose, glicose, sacarose e frutose nas variedades SP83-2847 e SP91-1049 às 7, 12, e 17 horas. Em SP83-2847, às 12 horas há produção de sacarase mas ao final do dia essa sacarose é consumida.	72
Figura 5-14 - Rede para os 265 genes diferencialmente expressos. Rede para os genes diferencialmente expressos tanto no trabalho de Casu <i>et al.</i> (2007) quanto McCormick <i>et al.</i> (2008) totalizando 120 nós e 118 arcos.....	73
Figura 5-15 - Hub CA271047. Gene CA271047 e os genes com os quais ele interage.	75
Figura 5-16 - Hub CA243069. Gene CA243069 e os genes com os quais ele interage.	76
Figura 5-17 - Hub CA151666. Gene CA151666 e os genes com os quais ele interage.	77
Figura 5-18 - Hub CA300310. Gene CA300310 e os genes com os quais ele interage.	78
Figura 5-19 - Hub CF572284. Gene CF572284 e os genes com os quais ele interage.....	79
Figura 5-20 - Análise do padrão de expressão para a rede de co-expressão de todos os genes do array. A) Experimento de manipulação da fotossíntese. B) Experimento de metabolismo de parede celular. Regiões em laranja representam alto nível de expressão e regiões em azul baixo nível de expressão.	80
Figura 5-21 - Análise do padrão de expressão para a rede de 143 genes. A) Experimento de manipulação de fotossíntese. B) Experimento de metabolismo de parede celular. Regiões em laranja representam alto nível de expressão e regiões em azul baixo nível de expressão...81	81
Figura 5-22 - Heat map para os genes diferencialmente expressos no experimento de manipulação fonte-dreno (n = 819). Esse gráfico considera os níveis de expressão para o genótipo 2847 durante todo o período de tratamento.	82
Figura 5-23 - Cluster hierárquico representando o sexto dia do experimento (n = 1644). Material coletado às 8, 12 e 17 horas.....	83

LISTA DE TABELAS

Tabela 4-1 - Lista de trabalhos que disponibilizaram experimentos com dados de cana-de-açúcar. Em negrito, são os trabalhos selecionados para serem utilizados nesse estudo.	39
Tabela 4-2 - Quantificação de sólidos solúveis totais (brix) nas plantas do experimento. Medidas realizadas em 26/05/2011 quando as plantas se apresentavam com 7 meses.	46
Tabela 4-3 - Resultado do sequenciamento do RNA-seq das 12 amostras de RNA de cana-de-açúcar no experimento de manipulação da relação fonte-dreno. A quantidade e a qualidade dos reads esta dentro do esperado pela tecnologia Illumina.	48
Tabela 5-1 - Classificação dos genes de acordo com sua função.	56
Tabela 5-2 - Classificação dos genes selecionados a partir do cluster conforme sua função.	61
Tabela 5-3 - Classificação dos 136 genes da rede de acordo com sua categoria funcional.	65
Tabela 5-4 - Classificação dos genes da rede de acordo com sua categoria funcional.	74

1. INTRODUÇÃO

O Brasil é um dos países mais apropriados para a produção da cultura da cana-de-açúcar e seus derivados, sendo que essa cultura assume um relevante papel no cenário nacional e mundial como principal matéria-prima para as agroindústrias do açúcar e álcool. A integração das diferentes condições a que a cultura ficou sujeita resulta na qualidade tecnológica e produtiva da cana, que são influenciadas por diversos fatores (Gilbert *et al.*, 2006). Assim, surge a necessidade da previsão das respostas da cultura a diferentes estímulos.

A maioria dos programas de melhoramento em cana deseja um aumento na capacidade de acúmulo de sacarose na planta, porém o rendimento atual é apenas 65% da capacidade total predita para o entrenó (Moore *et al.*, 1997; Jackson, 2005). Sendo assim, há um considerável incentivo à indústria do açúcar para desenvolver e utilizar as novas técnicas moleculares e atuar em parceria com programas de melhoramento visando o aumento do acúmulo de sacarose. Contudo, quando comparado a plantas C_3 há um baixo número de trabalhos objetivando o melhoramento da cultura (McCormick *et al.*, 2008).

Características únicas como armazenamento de altas concentrações de foto-assimilados na forma de sacarose como um soluto osmoticamente ativo e o fato do seu órgão de reserva ser o parênquima do colmo e não órgãos dreno fazem da cana um importante alvo de estudo das relações fonte e dreno (McCormick *et al.*, 2008).

Ainda há muitas questões importantes, mas sem respostas sobre o metabolismo e sinalização por açúcares, e com os avanços nas técnicas de obtenção e análise de dados essas questões podem ser solucionadas. O conhecimento sobre mecanismos endógenos usados pelas plantas para controlar vias metabólicas pode ser favorável para modificações na produção e utilização de carboidratos em plantas (Smeekens & Rook, 1997). No entanto, a maioria das aproximações para modificar esses processos através da manipulação de conversões metabólicas individuais tem encontrado sucesso limitado (Smeekens, 2000).

A elucidação do circuito completo de regulação que forma a base das respostas biológicas complexas para mudança dos níveis de açúcar tem encontrado dificuldades devido à existência de uma complexidade intrínseca de um organismo multicelular fotossintético com tecidos fonte e dreno e com atividades de produção e consumo de açúcar como é o caso da cana-de-açúcar. Além

disso, espécies diferentes também mostram diferentes relações fonte-dreno, transporte e produção de nutrientes e mecanismos de utilização. Diferentes plantas podem mostrar diferentes sensibilidades a sinais endógenos e exógenos em diferentes estágios de desenvolvimento (Rolland *et al.*, 2002). Esses fatores criam complicações em comparações diretas de resultados experimentais.

Visto a crescente identificação da expressão coordenada de genes codificadores de proteínas ou não, envolvidos no metabolismo de sacarose e na sinalização por açúcares, surge a necessidade de abordagens sistêmicas visando entender a complexidade da rede de regulação destes processos em espécies vegetais de interesse para a bioenergia. Novas abordagens de modelagem capazes de prever corretamente as associações entre fenótipo e genótipo podem ajudar a lidar com a complexidade da organização biológica e auxiliar o melhoramento destas espécies vegetais.

2. OBJETIVOS

2.1. Objetivo geral

Este trabalho visa inferir interações gênicas que regulam os processos de síntese e acúmulo de sacarose na cana-de-açúcar através das redes bayesianas, com o objetivo de desenvolver uma rede preditiva do comportamento do sistema em estudo a partir de perturbações ambientais.

2.2. Objetivos específicos

- 1) Definir um método de inferência de redes de regulação gênicas, com base em interações descritas em bases de dados de expressão gênica;
- 2) Estabelecer a rede de regulação gênica da síntese e sinalização de sacarose utilizando o modelo lógico de redes bayesianas;
- 3) Desenvolver uma interface de acesso ao modelo de modo que outros grupos de pesquisa possam utilizar a estratégia desenvolvida.

3. REVISÃO BIBLIOGRÁFICA

3.1. Cana-de-açúcar

A cana-de-açúcar é uma monocotiledônea pertencente à família Poaceae. Faz parte do gênero *Saccharum* L. que é composto por pelo menos seis espécies, sendo elas *S. officinarum*, *S. spontaneum*, *S. sinense*, *S. Barbieri*, *S. robustum* e *S. edule* (Palhares, 2009).

As cultivares modernas de cana-de-açúcar (*Saccharum* spp.) estão representadas por híbridos altamente poliplóides e aneuplóides derivados da hibridação interespecífica envolvendo *S. spontaneum* (2n=40 a 128), responsável pelo vigor vegetativo e resistência a estresses bióticos e abióticos, e *S. officinarum* (2n=80), a qual contribui com o alto teor de açúcar (Bortoleto, 2010; Ming *et al.*, 2001). Apesar de *S. officinarum* e *S. spontaneum* terem sido as espécies que mais contribuíram para a obtenção das atuais variedades comerciais, outras espécies, a exemplo de *S. sinense* (2n=111 a 120), *S. barberi* (2n=81 a 124) e *S. robustum* (2n=60 a 205), ainda que em menor proporção, também foram importantes para a composição genética das variedades modernas de cana (Marques, 2009).

Nesses híbridos multiespecíficos, *S. officinarum* contribuiu com 80%, *S. spontaneum* com 10% e as outras espécies do gênero com 10% dos cromossomos de seu genoma (D'Hont, 2005). Estes híbridos recebem a denominação *Saccharum* spp, possuem genoma altamente complexo e número de cromossomos variando de 100 a 120 (Branco, 2008).

A cana-de-açúcar é uma gramínea semi-perene e alógama e caracteriza-se por apresentar o sistema radicular fasciculado e caule tipo colmo que possui crescimento cilíndrico composto de nós e entrenós, folhas alternas, opostas, presas aos nós dos colmos, com lâminas de sílica em suas bordas, e bainha aberta (Jadoski *et al.*, 2010). A sacarose é predominantemente estocada nos colmos (Mantelatto, 2005). Suas flores são hermafroditas e dispõem-se numa inflorescência tipo panícula, também conhecida como flecha (Silva Júnior *et al.*, 2010). Quando cultivada comercialmente é propagada assexuadamente por propagação vegetativa, feita por intermédio de pedaços do colmo contendo uma ou mais gemas, porém, é uma planta de reprodução sexuada, e pode ser propagada através de sementes após a realização de cruzamentos principalmente em programas de melhoramento (Jadoski *et al.*, 2010; Casagrande, 2008).

Seu centro de origem provavelmente foi a região leste da Indonésia e Nova Guiné e então se difundiu para vários lugares do Oceano Pacífico, Golfo de Bengala, aparecendo como planta produtora de açúcar na Índia tropical (Ripoli *et al.*, 2007).

Seu cultivo pode ser em vários tipos de ambientes sendo que, atualmente, ocorre em regiões tropicais e subtropicais de mais de 90 países, é difundida em uma ampla faixa de latitude de 35°N a 30°S, e adapta-se a diversas condições de solos sujeitos a climas variados, exigindo precipitações pluviométricas entre 1500 a 2500 mm por ciclo vegetativo (Rodrigues, 1995). Isso resulta em uma grande variedade de ambientes para sua produção. Apesar disso, ela encontra-se nas suas melhores condições para fase de crescimento quando há um período quente e úmido e durante as fases de maturação num período seco, ensolarado e mais frio (Brunini, 2008).

As características morfológicas são utilizadas no melhoramento genético visando-se além das características agronômicas de produtividade, rusticidade, resistência a pragas e a doenças, algumas características industriais como alto teor de sacarose e teor médio de fibras (Stupiello, 1987).

3.1.1. Importância econômica da cana-de-açúcar

O Brasil lidera a produção mundial de cana com mais de 700 milhões de toneladas, seguido da Índia e China, sendo que em média, 45% da produção brasileira destina-se à produção de açúcar (FAO, 2012). A cana-de-açúcar é responsável por 60% da produção de açúcar no mundo (Rocha *et al.*, 2007). O açúcar e o álcool ocupam posições de destaque nas exportações do agronegócio brasileiro, pois pertencem a terceira mais importante cadeia produtiva, atrás apenas do complexo da soja (grão, farelo e óleo) e de carnes (Hamerski, 2009).

A importância econômica da cana-de-açúcar se dá devido sua habilidade única de estocar sacarose nos colmos. A sacarose é o componente de maior interesse no processamento da cana e, sua obtenção é desejada na forma cristalizada (Hamerski, 2009). Além disso, o processamento industrial da cana pode também ser dirigido visando-se a produção de álcool, que é utilizado como combustível e a partir daí, empregado em toda a indústria álcoolquímica (Santos, 2011).

A cultura da cana-de-açúcar é de grande versatilidade, sendo que todos seus componentes são utilizados. Desde o caldo do qual se obtêm o açúcar, a cachaça, o álcool, a rapadura e outros; do bagaço, o papel, a ração, o adubo ou o combustível; até as folhas que são convertidas em ração animal (Hamerski, 2009). Assim, a agroindústria da cana-de-açúcar, direciona-se a integrar os

sistemas de produção alimentar, não alimentar e energético, envolvendo atividades agrícolas e industriais, e ainda atua com vantagens comparativas em relação às outras matérias-primas, pelo fato de ser intensiva em mão-de-obra e o Brasil ter os menores custos de produção do mundo (Vasconcelos, 2002).

3.1.2. Fisiologia da cana-de-açúcar

A cana possui alta eficiência na conversão de energia solar em açúcar e fibras principalmente por ser uma planta com metabolismo C₄. Assim, quanto maior a captação solar, maiores serão as taxas fotossintéticas e consequentemente acúmulo de açúcares (Fortes, 2003; Brunini, 2008).

No decorrer do ciclo de crescimento e desenvolvimento, a cana-de-açúcar atravessa dois períodos distintos com relação ao teor de sacarose: o primeiro é assinalado por um intenso crescimento vegetativo acompanhado por uma gradual formação de sacarose, enquanto, no segundo, ocorre um predominante acúmulo de sacarose, motivado pela escassez dos principais fatores de desenvolvimento vegetativo, como temperatura e água disponível (Magalhães, 1987; Luchesi, 1995).

O estágio de maturação é verificado principalmente pelos teores de sacarose, de açúcares redutores e umidade que os colmos apresentam no decorrer do período da safra (Picoli, 2007). Segundo Silva (1989), a maturação da cana-de-açúcar pode ser considerada tanto do ponto de vista botânico quanto fisiológico. Botanicamente, a cana-de-açúcar está madura após a emissão de flores e formação de sementes que possam dar origem a novas plantas. Fisiologicamente, a maturação é alcançada quando os colmos atingem seu potencial de armazenamento de sacarose, ou seja, o ponto de máximo acúmulo possível. É a fase fisiológica que tem maior retorno econômico e é definida pelos fisiologistas como um estágio senescente, entre o crescimento rápido e a morte final da planta.

O processo de maturação da cana envolve um sistema metabólico complexo, que se inicia com a atividade fotossintética nos cloroplastos das células das folhas, culminando com o acúmulo de carboidratos fotossintetizados nos colmos (Rodrigues, 1995). Nesse estágio, a planta alcança o nível máximo de sacarose, que após esse período, começa a decair.

Outra etapa fisiológica importante no desenvolvimento da cana é o florescimento que é totalmente indesejado, pois são atribuídas ao florescimento, perdas substanciais em toneladas de

cana e teor de sacarose durante a colheita (Barbieri & Silva, 2008). Com o florescimento, os colmos da planta param de emitir novos entrenós e começam a se “isoporizar”, reduzindo consideravelmente a densidade e o peso da produção (Barbieri & Silva, 2008).

3.2. Sinalização por sacarose

Os açúcares desempenham um papel fundamental na vida das plantas, pois representam a principal fonte de carbono para a síntese de moléculas necessárias para a sobrevivência das células. Além disso, os açúcares podem ser importantes moléculas reguladoras de processos associados com crescimento, maturação e senescência (Koch, 1996; Jang & Sheen, 1997) e também das funções de controle do metabolismo, resistência a estresses, crescimento e desenvolvimento (Ramon *et al.*, 2008). São também importantes reguladores da expressão gênica (McCormick, 2007) e suas atividades regulatórias incluem ativação e repressão de diversos genes (Jang & Sheen, 1997). Assim como os hormônios, eles podem atuar como mensageiros primários e regularem sinais que controlam a expressão de vários genes envolvidos no metabolismo de açúcares (Gupta & Kaur, 2005).

A sinalização por açúcar pode ser entendida como a detecção da molécula de açúcar por uma proteína-sensor que dará origem a um sinal (Gupta & Kaur, 2005). A sensibilidade aos açúcares pode ser extracelular, nos níveis de transportadores e receptores de açúcares, ou intracelular através da ação de uma hexoquinase (HXK) ou uma quinase do complexo *sucrose-non-fermenting-1* (SNF1), que possui um papel crucial a ser desvendado, além de fosfatases e vários outros complexos proteicos tais como proteínas que se combinam ao DNA (Loreti *et al.*, 2001). Depois de gerado o sinal, iniciam-se cascatas de transdução que resultam em respostas celulares por meio da expressão alterada de genes e atividade enzimática (Gupta & Kaur, 2005).

A sacarose é o principal carboidrato transportado no floema das plantas (Magalhães, 1987) e sua função como molécula sinalizadora foi sugerida após ter sido observado que ela reprimiu níveis de mRNA e a atividade de transporte da *proton-sucrose symporter* (Barker *et al.*, 2000). O mecanismo de sinalização por sacarose é realizado por vias independentes da hexoquinase1 (HXK1) (Branco, 2008). Nessas vias a expressão gênica é regulada diretamente por dissacarídeos, sacarose ou trealose (Smeekens, 2000; Rolland *et al.*, 2002). Apesar dessas vias não estarem bem elucidadas, estima-se que alguns transportadores de sacarose poderiam atuar

como receptores (Choiu & Bush, 1998) e algumas SNFs quinases poderiam estar envolvidas (Purcell *et al.*, 1998; Tiessen *et al.*, 2003).

Foi mostrado que dissacarídeos de sacarose regulam a tradução do mRNA de um grupo de fatores de transcrição de *Arabidopsis* pertencentes a família do domínio bZIP (Wiese *et al.*, 2004) e o produto hidrolítico da sacarose, a glicose, interfere na atividade da hexoquinase1 que media a repressão de genes de expressão fotossintéticos e de controle de crescimento em *Arabidopsis* (Moore *et al.*, 2003).

Como revisado por Ribeiro (2012), mais recentemente, teve-se evidências genéticas de que a sacarose é um regulador global das respostas a falta de fosfato em *Arabidopsis* (Lei *et al.*, 2011a e 2011b) e também foi determinado que a sacarose induz a ativação rápida da CfSaPk, uma *mitogen activated kinase* em células de suspensão de *Cephalostachyum fuchiassun* Gamble (Li *et al.*, 2012).

3.3. Síntese de sacarose, translocação e acúmulo.

Na cana-de-açúcar, o CO₂ é fixado em um ácido orgânico composto por quatro carbonos, o malato, que é então transportado para as células da bainha e descarboxilado. Devido a essa compartimentalização das atividades, ela é uma planta com metabolismo do tipo C₄ (Hatch & Slack, 1966).

A síntese de sacarose em plantas C₄ é dependente do Ciclo de Calvin que é a etapa onde o CO₂ liberado será fixado em sacarose (Figura 3-1). Esse ciclo foi proposto por Calvin (1956) e é dividido em três fases: carboxilação, redução e regeneração. Na fase de carboxilação a ribulose-1,5-bifosfato, que é o acceptor de CO₂, é carboxilada formando duas moléculas de 3-fosfoglicerato, que é o primeiro intermediário estável do Ciclo de Calvin. Na próxima fase, o 3-fosfoglicerato é reduzido formando gliceraldeído-3-fosfato. Há gasto de um ATP fotoquimicamente derivado e dos equivalentes redutores na forma de NADPH. Em seguida, a ribulose-1,5-bifosfato (RuBisCO) é regenerada a partir do gliceraldeído-3-fosfato.

Conforme apresentado por Huber & Huber (1996), a síntese de sacarose ocorre no citosol a partir de trioses fosfato por uma rota via frutose-1,6-bifosfato e glicose-1-fosfato. Essas trioses fosfato são formadas na etapa de redução do Ciclo de Calvin e são exportadas para o cloroplasto, onde são convertidas em frutose-1,6-bifosfato pela enzima aldolase e depois hidrolisada para ser convertida em frutose-6-fosfato pela enzima frutose-1,6-bifosfatase. Para Magalhães (1987), este

último passo, irreversível, juntamente com o passo catalisado pela enzima sacarose fosfato sintase (SPS) são os pontos principais de controle na via de síntese de sacarose.

A frutose-6-fosfato é então convertida para glicose-6-fosfato e, logo após, glicose-1-fosfato que então é convertida em UDP-glicose por intermédio de uma UDP-glicose fosforilase. Esse é o ponto principal para que ocorra o início da síntese de sacarose, pois a glicose-1-fosfato é derivada de produtos fotossintéticos (Ribeiro, 2012). Primeiro, ocorre uma reação entre a UDP-glicose com a frutose-6-fosfato catalisada pela sacarose fosfato sintase (SPS), na qual, é sintetizada uma sacarose-6-fosfato e um UDP.

Em seguida, a sacarose-6-fosfato fosfatase, através de uma reação essencialmente irreversível, remove o fosfato da sacarose-6-fosfato produzindo a sacarose e Pi, que é reciclado e retorna ao cloroplasto via translocador de fosfato para o contínuo suprimento de triose-fosfato para o citosol (Lunn & ApRees, 1990). Como a atividade da sacarose fosfato fosfatase, que catalisa a etapa final, a produção da sacarose é elevada e mantém uma concentração muito baixa de sacarose fosfato, o produto da reação da SPS (Krause & Stitt, 1992).

A atividade da SPS é vista como contribuinte para o controle do fluxo da sacarose (Ferne *et al.*, 2002) sendo que em folha de cana está correlacionada com o seu teor de sacarose (Grof *et al.*, 1998). A expressão da SPS é regulada por sinais do desenvolvimento, ambientais e nutricionais e em alguns casos, em níveis transcricional, pós-transcricional, traducional e pós-traducional (Winter & Huber, 2000). Como revisado por Krause & Stitt (1992), vários trabalhos vêm evidenciando a sua importância. Worrell *et al.* (1991) superexpressaram a SPS em plantas de tomate transgênicas e foi observado aumento nas taxas de fotossíntese e síntese de sacarose. Galtier *et al.* (1993) reportaram uma correlação positiva entre atividade da SPS e a razão sacarose:amido em plantas transgênicas. Botha & Black (2000) também demonstraram uma correlação alta entre a atividade da SPS e concentração de sacarose no colmo, na variedade NCo376. Zhu *et al.* (1997) destacaram a importância de uma elevada atividade da SPS na determinação da concentração final de sacarose no colmo.

Recentemente, constatou-se através de estudos de análise de perfil de expressão gênica que transcritos de enzimas envolvidas na síntese e clivagem de sacarose (intermediários da glicólise), como a sacarose fosfato sintase (SPS) e a sacarose sintase (SS), foram significativamente reprimidos em tecidos de entrenó maduro, o mesmo ocorrendo em transcritos

para três enzimas regulatórias, sendo elas frutose-2,6-bifosfato, pirofosfato-frutose-6-fosfato (PFP) e fosfoglicomutase (Casu *et al.*, 2003). Apesar disso, o mecanismo por trás da regulação da SPS em cana-de-açúcar ainda é mal compreendido.

A sacarose produzida nas folhas é carregada para os vasos do floema por um *proton-sucrose-simporter* (Gahrtz *et al.*, 1994; Reismeier *et al.*, 1994) e transportada até os entrenós, onde é armazenada (Felix, 2006). No colmo, ela será passada para o espaço apoplástico e através da ação de invertases extracelulares será convertida em glicose e frutose que entrarão nas células do parênquima através de transportadores específicos (Krook *et al.*, 2000). O descarregamento para o parênquima do colmo é mediado por invertases da parede celular, com a clivagem de sacarose seguida pelo transporte de monossacarídeos através da membrana plasmática (Roitsch *et al.*, 1995). Frutose e glicose são transformadas em sacarose-fosfato por ação da sacarose fosfato sintase, e serão armazenadas nos vacúolos (Krook *et al.*, 2000).

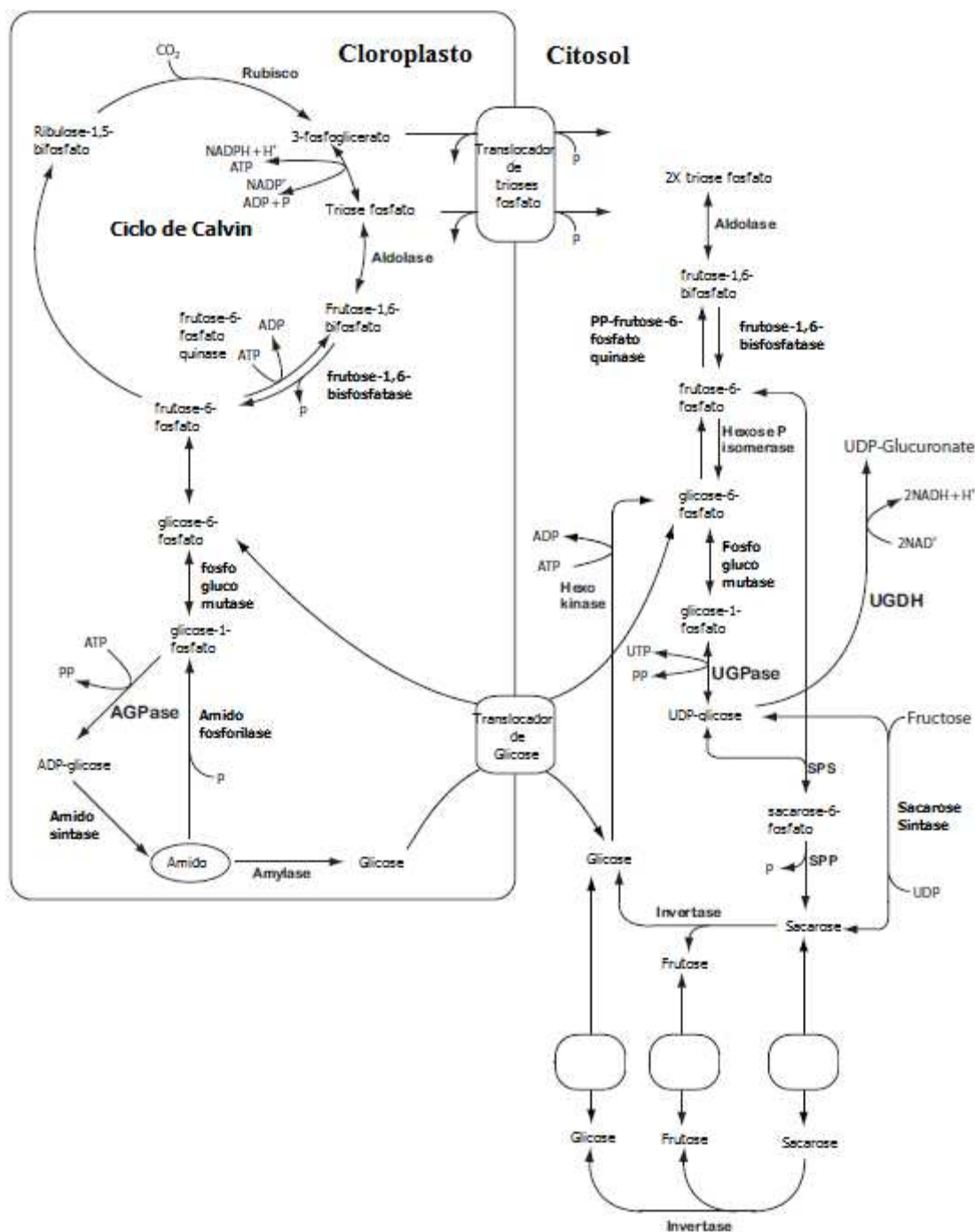


Figura 3-1 - Esquema da síntese de sacarose. Dentro do cloroplasto ocorre o Ciclo de Calvin que produzirá trioses fosfato. Essas trioses poderão ser armazenadas sob forma de amido no cloroplasto ou serem transportadas para o citosol e seguir a via para síntese de sacarose. (Adaptado de Johansson, 2003).

As enzimas invertases dirigem os carboidratos para o crescimento da planta ou para o acúmulo dos mesmos nos vacúolos, onde o aumento da sua concentração irá proporcionar o amadurecimento dos colmos, a qual ocorre quando a cultura apresentar a melhor produtividade

qualitativa e quantitativa de açúcares (Johansson, 2003). O acúmulo de sacarose ocorre contra um gradiente e a energia para tal é fornecida pela respiração (Johansson, 2003).

Cada entrenó acumula seu próprio açúcar, sendo os valores de sacarose mais elevados na direção do centro do colmo, declinando no sentido das pontas. Essas diferenças se acentuam mais nos entrenós mais jovens, refletindo provavelmente uma distribuição diferente de invertase, onde o meristema intercalar (anel de crescimento) contém muito mais invertase do que os tecidos centrais do entrenó (Rodrigues, 1995).

A atividade quase nula da invertase ácida vacuolar indica que a planta está obtendo um acúmulo efetivo de sacarose. À medida que as células se distanciam da região meristemática, estas se alongam com maior concentração de sacarose, atingindo o processo de maturação (Casagrande, 1991).

O mecanismo de acúmulo de sacarose em tecidos jovens e adultos é o mesmo, porém nos tecidos jovens ocorre o processo de hidrolização da sacarose pela invertase ácida vacuolar, movendo-se as hexoses resultantes rapidamente para o citoplasma, onde são utilizadas no crescimento e desenvolvimento celular (Toppa *et al.*, 2010).

3.4. Relações fonte-dreno

No floema, os açúcares são translocados das áreas de produção (fontes) para as áreas de metabolismo ou armazenamento (drenos). As fontes incluem órgãos exportadores que são capazes de produzir fotossintatos em excesso para suas necessidades, como folhas maduras ou órgãos de reserva (Turgeon, 1989).

Paul & Foyer (2001) comentam que relações fonte-dreno podem regular taxas fotossintéticas e consequentemente interferir no acúmulo de sacarose nos tecidos de reserva da planta. Essas atividades são causadas por uma desestabilização da homeostase dentro do cloroplasto, células do mesófilo ou mesmo na planta inteira, que pode ocorrer devido a mudanças significativas no ambiente e na fisiologia da planta.

Além do possível *feedback* através do acúmulo de produtos, há uma crescente evidência de que a atividade de enzimas relacionadas à fotossíntese e à expressão de genes transcritos associados são modificados pela demanda do dreno (Sheen, 1990, 1994; Black *et al.*, 1995; Koch, 1996; Pego *et al.*, 2000; Paul & Foyer, 2001; Rolland *et al.*, 2002).

Por serem organismos sésseis e autotróficos, as plantas possuem a necessidade de integrar o metabolismo e crescimento com sinais ambientais externos e assim as atividades fonte-dreno devem ser coordenadas rigorosamente (McCormick, 2007). Isso envolve tanto a expressão gênica quanto os mecanismos específicos de sinalização de açúcar (Rolland *et al.*, 2006).

O acúmulo de biomassa vegetal é um processo de *feedback* complexo, controlado tanto pelo aparato fotossintético (fonte) quanto pelos tecidos de armazenamento de carbono (drenos) (Ho, 1988). A cana evoluiu como um sistema de fonte-dreno único, em que o tecido de armazenamento é também o de crescimento, o colmo e, assim, o carbono acumulado deve ser dividido em respiração e crescimento do colmo, e a necessidade de armazenamento (McCormick, 2007).

Na cana-de-açúcar o principal órgão de fonte é composto por células do mesófilo foliar (Chandra, 2011). Os fotoassimilados produzidos na fotossíntese pelas folhas adultas são transportados para meristemas e órgãos em desenvolvimento, como exemplo, folhas jovens e o colmo que constituem os principais órgãos que dependem da importação de carboidratos mais importantes (Figura 3-2) (Chandra, 2011). Esse transporte de fonte para dreno acarreta em várias mudanças fisiológicas na folha, que podem ser responsáveis pela capacidade de acumular sacarose nos entrenós (Felix, 2006).

Em geral, as atividades fonte como fotossíntese, mobilização e exportação de nutrientes são induzidas em condições de baixa disponibilidade de açúcares, enquanto as atividades dreno como crescimento e armazenamento são induzidas quando as fontes de carbono estão disponíveis em abundância (Rolland, 2006). Embora a sacarose seja o principal produto fotossintético e açúcar de transporte nas plantas, muitos efeitos de sinalização de açúcar sobre o crescimento e metabolismo podem ser atribuídos aos produtos de sua hidrólise, glicose e frutose (Rolland, 2006).

Os mecanismos de regulação fonte-dreno não são caminhos lineares simples, mas um complexo sistema formado por diversos sinais atuando em momentos específicos e, em muitas vezes, através de pontos de controle de *feedbacks* formando, assim, redes de regulação gênica (Paul & Foyer, 2001; Fujita, 2007).

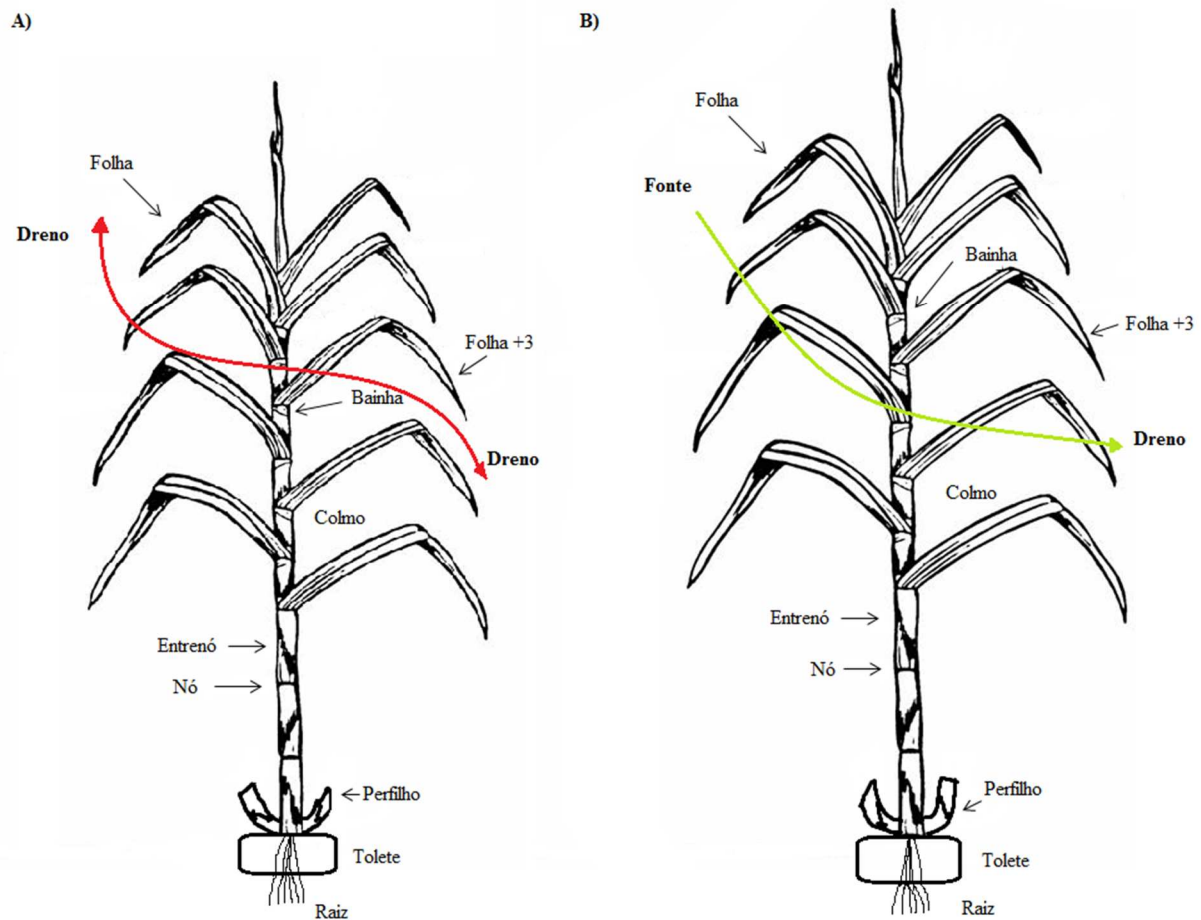


Figura 3-2 - Esquema da planta de cana-de-açúcar e regulações fonte-dreno ao longo de seu desenvolvimento. A) Planta jovem em desenvolvimento vegetativo, órgão em crescimento funcionando como dreno. B) 10-12 meses, planta adulta havendo produção de açúcares na fonte e acúmulo nos tecidos de reserva. Em A, com a planta em estado de desenvolvimento, folhas jovens funcionam como dreno enquanto em B, as folhas já desenvolvidas funcionam como fonte e os entrenós como órgãos de reserva, no caso drenos. (Adaptado de Ribeiro, 2012).

3.5. Redes de regulação gênica

O fluxo de informação durante a expressão gênica ocorre a partir do DNA, que serve de molde para a transcrição do RNA que é então traduzido em proteína (Figura 3-3). Esse processo resulta na produção de proteínas específicas para realização de funções e é vital para a manutenção das atividades celulares. Alterações nesse fluxo regulam a concentração de enzimas e isso garante que todos os intermediários das vias metabólicas estejam em concentrações adequadas e que processos metabólicos sejam desligados quando não se fazem necessários (Martin & Schnarrenberger, 1997).

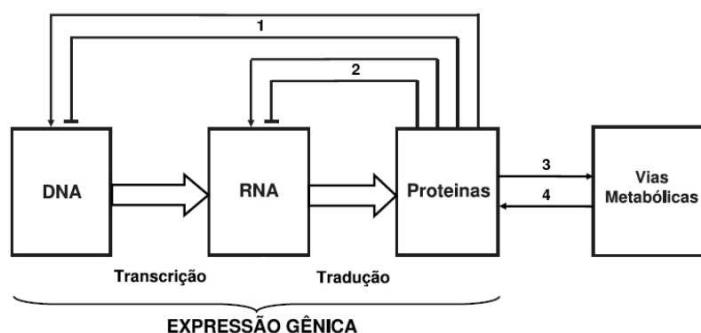


Figura 3-3 - Esquema da regulação da expressão gênica. 1) Proteínas induzindo ou inibindo a transcrição. 2) Proteínas induzindo ou inibindo a tradução. 3) Proteínas induzindo vias metabólicas. 4) Vias metabólicas induzindo a expressão de proteínas. (Adaptado de Lima & Barrera, 2008).

Segundo Pierce (2004), a expressão gênica pode ser regulada em vários pontos ao longo da via de fluxo de informações do genótipo para o fenótipo. Modificações no DNA ou em sua compactação podem influenciar quais sequências estão disponíveis para a transcrição assim como na taxa com que estas sequências são transcritas. A quantidade de proteínas produzidas pode depender da quantidade de mRNA sintetizado e da velocidade com a qual ele é degradado. Pode ocorrer regulação no nível pós-transcricional através do *splicing* alternativo, transporte de mRNAs para fora do núcleo assim como sua estabilização, além de vias de micro-RNAs. A disponibilidade de fatores necessários à tradução pode também regular a quantidade de mRNA que é traduzido, e proteínas e RNAs não-codificantes podem causar modificações pós-traducionais nos produtos da tradução afetando também a expressão gênica.

Contudo, o tipo de regulação mais estudado ocorre em nível transcricional (Cornelia, 2008). A transcrição é controlada por sequências de DNA (cis-reguladoras) que variam de simples interruptores a estruturas complexas que respondem a uma variedade de sinais criando circuitos combinatórios que ativam ou inativam a transcrição gênica. Os fatores de transcrição, um tipo específico de proteínas, reconhecem esses sítios e se ligam a eles, interagindo para mediar a transcrição dos RNAs. Algumas dessas proteínas são ativadores transcricionais enquanto outras são repressoras (Goutsias & Lee, 2007). Há milhares de fatores de transcrição conhecidos, cada um com sua estrutura única, contudo, a maioria deles contém um dos poucos motivos estruturais através dos quais se ligam no DNA (Cornelia, 2008). Os mais comuns são: *helix-turn-helix*, *homeodomain*, *leucineziper* os *zinc-fingers*.

De maneira geral, os sistemas de regulação gênica envolvem muitos genes conectados por interações de ativação ou repressão formando complexas redes. Uma rede de regulação gênica é um conjunto de espécies moleculares e suas interações, que juntas controlam genes e a quantidade de seus produtos e, dessa forma, afetam numerosos processos celulares (Karlach & Shamir, 2008). Elas representam o conjunto de interações entre genes, fatores de transcrição e outras biomoléculas.

Através de métodos estatísticos e computacionais, é possível identificar as principais relações de dependência nesse complexo sistema de interação de genes, que representam como os genes influenciam uns aos outros, de maneira ativadora ou inibidora. Redes gênicas naturais podem ser descritas como circuitos interconectados de módulos funcionais, cada um consistindo de interações especializadas entre proteínas, DNA, RNA e pequenas moléculas (Kaern *et al.*, 2003).

Todo gene possui um ou mais ativadores, que são sinais bioquímicos necessários para iniciar sua transcrição. Na ausência desses ativadores somente um baixo nível de expressão desse dado gene pode ocorrer. Da mesma forma, ele pode ter também inibidores que são os sinais bioquímicos que previnem ou diminuem sua expressão mesmo na presença de um ativador apropriado (Chen *et al.*, 1999).

Cada vez mais, as redes têm sido utilizadas como modelos para representar fenômenos no nível de expressão gênica e pesquisas sobre sua construção a partir de dados experimentais são abundantes (Brazhnik *et al.*, 2002). Apesar das redes descreverem somente a dinâmica da atividade gênica, interações nos níveis de proteoma e metaboloma estão implícitas uma vez que a dinâmica da atividade gênica depende deles (Brazhnik *et al.*, 2002).

Essas redes diferem-se das redes bioquímicas porque se concentram apenas em dados de transcrição e dessa forma, muitas conexões físicas entre macromoléculas podem ficar ocultas por atalhos, ou seja, muitas unidades intermediárias em cascatas regulatórias poderiam ser escondidas (Brazhnik *et al.*, 2002). A informação fornecida pela análise de expressão é o produto de todos os eventos regulatórios que a afetam (Wyrick & Young, 2002).

Dentre os objetivos desse tipo de estudo estão: trazer conhecimento sobre as interações moleculares e atividades metabólicas de um organismo; inferir quais genes determinam um tipo

específico de fenótipo; e modelar a atividade desses genes para a identificação de seu comportamento sob determinados estímulos ou situações.

Várias abordagens para a inferência de redes foram propostas como as Redes Bayesianas (Murphy & Mian, 1999; Friedman *et al.*, 2000; Hartemink *et al.*, 2001; Moler *et al.*, 2000; Imoto *et al.*, 2002; Tamada *et al.*, 2003; Dojer *et al.*, 2006), Redes Neurais (Weaver *et al.*, 1999), Modelos Lineares (van Someren *et al.*, 2000; D'Haeseleer *et al.*, 1999), Equações Estruturais (Xiong *et al.*, 2004), Redes Booleanas Probabilísticas (Akutsu *et al.*, 2000; Shmulevich *et al.*, 2002; Pal *et al.*, 2005), Modelos Gráficos Gaussianos (Schäfer & Strimmer, 2005), controles *Fuzzy* (Woolf & Wang, 2000) e Equações Diferenciais (Mestl *et al.*, 1995) e também modelos incluindo componentes estocásticos em nível molecular (Arkin *et al.*, 1998; Smolen *et al.*, 2000; Hasty *et al.*, 2001). Essas abordagens se diferem no grau de resolução que alcançam e nas suposições adjacentes (El Samad *et al.*, 2002).

Segundo Karlebach & Shamir (2008) todos esses modelos computacionais podem ser divididos em três classes. A primeira classe, os modelos discretos e lógicos, descreve as redes regulatórias qualitativamente. Esses modelos permitem que usuários obtenham uma compreensão básica das diferentes funcionalidades de uma determinada rede sob diferentes condições. A sua natureza qualitativa os torna flexíveis e fáceis de serem ajustados a fenômenos biológicos embora eles só possam responder a perguntas qualitativas. A segunda classe, modelos contínuos, foi desenvolvida para entender e manipular comportamentos que dependem do tempo e concentrações moleculares exatas, pois diferentemente dos modelos qualitativos, as variáveis não terão seus valores agrupados em classes. E a terceira classe de modelos foi introduzida depois da observação de que a funcionalidade das redes regulatórias é frequentemente afetada pelo ruído. Como a maioria dos modelos da terceira classe representam as interações entre as moléculas individuais, eles são referidos aqui como modelos em nível de *single-molecule* e explicam a relação entre estocasticidade e regulação dos genes.

Um exemplo bem conhecido de representação de redes de regulação gênica em plantas é a rede apresentada por Mendonza & Alvarez-Buylla (1998) para florescimento utilizando o modelo ABC, onde A representa a função que dá origem às pétalas, A e B originam as sépalas, B e C os estames e, finalmente, C é a função que origina os carpelos (Figura 3-4). Esses autores descrevem seis estados principais com base nessas funções.

No estado 1 somente a função A é exercida então tem-se que somente o gene AP1 (*apetala1*) está ativado e esse padrão de ativação é encontrado no primeiro verticilo floral. No estado 2 ocorre a ativação dos genes AP1, AP3 (*apetala3*) e PI (*pistillata*) que constituem as funções A e B que resultam na formação do segundo verticilo. O estado 3 é representado pela ativação dos genes AG (*agamous*), AP3 e PI e então tem-se a presença das funções B e C que formam o terceiro verticilo. O estado 4 se refere à ativação de AG, função C, formando o quarto verticilo. Há também um quinto estado que seria a expressão de dois genes que são inibidores florais EMF1 (*embryonic flower1*) e TFL1 (*terminal flower1*) característico de células que não fazem parte das flores. Finalmente, há um sexto estado que não é observado experimentalmente, onde os genes inibidores EMF1 e TFL1 são expressos juntamente com os genes indutores do florescimento.

Esse modelo foi obtido através da metodologia booleana onde as interações entre os genes são modeladas em termos de relações booleanas e circuitos lógicos combinatórios (Kauffman, 1969). Portanto, os estados são representados em dois níveis: 0 quando ele está reprimido e 1 quando o gene é expresso sendo que isso é definido dependendo se a concentração de mRNA está ou não acima de um dado *threshold*.

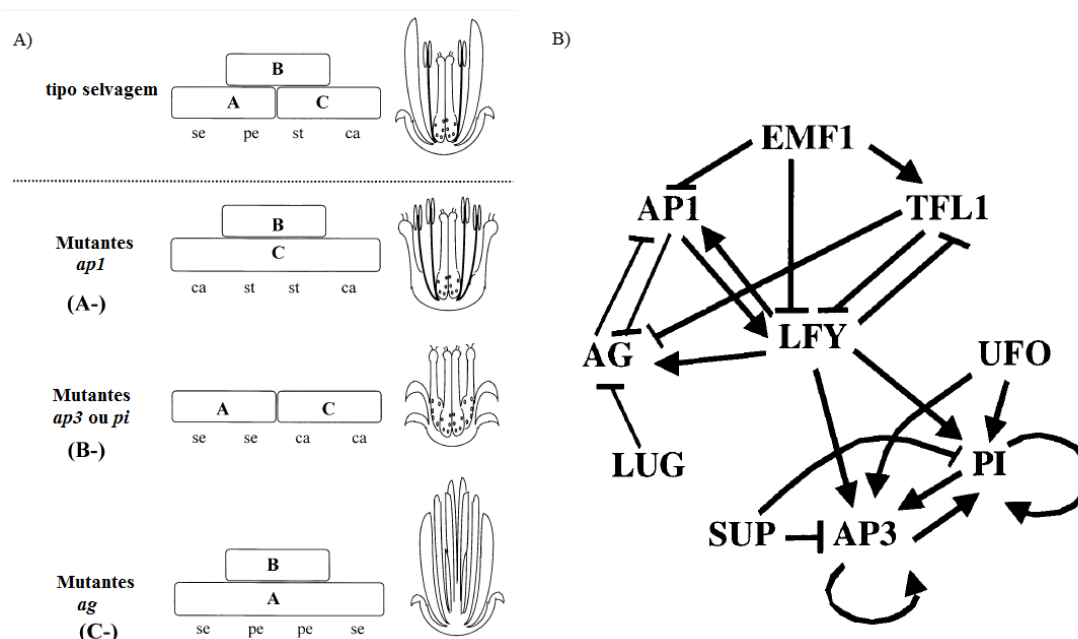


Figura 3-4 - Rede booleana para morfogênese de flores em *A. thaliana*. A) Fenótipos do tipo selvagem, mutantes *ap1* apenas com carpelos e pétalas, mutantes *ap3* ou *pi* com sépalas e carpelos, mutantes *ag* com sépalas e pétalas e estames e carpelos ausentes. B) Esquema da rede de regulação gênica

para morfogênese da flor. Notação $\text{---}|$ indica relações de repressão e $\text{---}\rightarrow$ indica indução. (Adaptado de Mendonza & Alvarez-Buylla, 1998).

Nesse trabalho, a abordagem utilizada para elaboração da rede é a bayesiana estática, portanto na rede não há representação de ciclos ou auto-regulações. Porém as relações de indução e repressão são também representadas de forma diferencial, sendo $\text{---}|$ a notação utilizada para repressão e $\text{---}\rightarrow$ para indução.

A principal vantagem do uso das redes regulatórias é que elas permitem uma visualização em larga-escala, representada através de uma maneira gráfica, do comportamento dos genes, o que talvez não fosse possível utilizando outras abordagens. Isso facilita a identificação de algumas propriedades da topologia da rede. Além disso, as redes gênicas não são apenas capazes de descrever um grande número das interações de uma forma concisa, mas também elas podem representar as propriedades dinâmicas subjacentes a essas interações ao nível de sistema (Brazhnik *et al.*, 2002). Como desvantagem, tem-se que experimentos de expressão gênica possuem muitos genes e poucas repetições e isso pode agregar ruído à rede dificultando a construção de redes de alta precisão.

Mutwill (2010) apresenta em seu trabalho a descrição de algumas propriedades gerais da topologia de redes biológicas com um foco maior em redes gênicas, mas que em alguns casos podem ser aplicadas a outros contextos. Uma delas é que a maioria dos nós possuem poucas ligações funcionais (uma ou duas) enquanto um pequeno número de nós, os *hubs*, possuem muitas ligações. Isso confere uma estrutura hierárquica em que os *hubs* desempenham um papel central em dirigir a resposta celular a um dado estímulo. O fato da maioria dos nós ser escassamente conectados confere a rede menor sensibilidade a perturbações aleatórias, porém isso torna a rede muito sensível à inativação dos *hubs*. Além disso, apenas poucos passos são necessários para juntar quaisquer dois nós, facilitando a propagação e integração de sinais.

Avanços nas técnicas modernas como *microarrays*, SAGE, e RNA-seq permitem que uma vasta quantidade de informações sobre essas redes sejam coletadas contemplando várias condições e diferentes períodos de tempo ou espécies, e isso, combinado com técnicas computacionais modernas, permite que redes de regulação sejam construídas a partir de dados de expressão gênica e modeladas *in silico* (Hickman & Hodgman, 2009).

Para a modelagem matemática e inferência dessas redes, assume-se que o padrão de expressão gênica contenha a informação sobre a quantidade de tradução e atividade das proteínas

e, além disso, que o nível de expressão de cada gene dependa tanto do seu próprio valor de expressão quanto dos valores de expressão de outros genes em instantes prévios de tempo (Fujita, 2007).

3.6. Teoria da Probabilidade

Esta seção é uma revisão breve de alguns conceitos que não são o foco do trabalho, mas serão úteis para o entendimento de outros tópicos e foi baseada em Spiegel *et al.* (2000).

Um experimento pode ser entendido como um ensaio científico para a verificação de um dado fenômeno. Quando o resultado desses fenômenos pode variar, mesmo em condições normais de experimentação, eles são considerados aleatórios e para a previsão de um resultado futuro sobre eles é necessária a adoção de modelos matemáticos probabilísticos. O conjunto de todos os resultados possíveis é chamado de espaço amostral (S). Se repetirmos o experimento indefinidamente sob as mesmas condições, seu resultado não será previsível, a priori.

Uma variável aleatória é uma função que aloca um ponto do espaço amostral para cada resultado de um experimento aleatório, ou seja, é uma função associada a um experimento, sendo que a realização do experimento leva essa variável a assumir um valor dependente do acaso, mas que faz parte do espaço amostral. Considerando-se que X seja um valor numérico cujo valor depende do resultado de um experimento, se X associa um resultado a um número, então ele é uma função cujo domínio é o conjunto de resultados e sua imagem é o conjunto dos números reais. Essas variáveis possuem probabilidades associadas. Essas variáveis podem ser discretas, quando seus valores podem ser representados por classes enumeráveis ou contínuas quando assumem valores numa escala contínua (na reta real), podendo ter valores fracionais.

Sendo X uma variável aleatória e D_X seu domínio, os valores da probabilidade de X equivalem a x , dado que X é um valor no espaço amostral denotado por $P(X = x)$, para todas as instâncias de valores x no espaço amostral em D_X , e constituem a distribuição de probabilidade de X . Agora, considerando-se duas variáveis aleatórias X e Y , a distribuição de probabilidade de X e Y constitui uma distribuição conjunta. A distribuição marginal é aquela dada por apenas uma das variáveis do conjunto. E por fim, quando a probabilidade de uma variável X está condicionada ao conhecimento de um valor para Y , ou seja, probabilidade de X equivale a x dado que Y equivale a y , denotado por $P(X = x|Y = y) = P(x|y)$, para todo x em D_X , tem-se a distribuição de probabilidade condicional de X dado que $Y = y$.

Duas variáveis X e Y são independentes se a probabilidade de uma delas dada a outra variável é igual à probabilidade dela sozinha, isto é, $P(x|y) = P(x)$ sempre que $P(y) > 0$, para todo $x \in D_x$ e $y \in D_y$. Isso implica que conhecer Y não altera a probabilidade de X . Essa independência pode ser expressa em termos da distribuição conjunta de X e Y , derivada da regra do produto, como $P(x, y) = P(x)P(y)$.

A distribuição de probabilidade conjunta para um conjunto de variáveis aleatórias $X_1, X_2, \dots, X_n$ é definida como $P(X_1, X_2, \dots, X_n)$ para os valores de $X_1, X_2, \dots, X_n$, isto é, o produto de todos os valores de probabilidades.

Para um dado experimento aleatório, a teoria da probabilidade é baseada em uma função P que a cada evento do espaço amostral associa um número real no intervalo de 0 a 1, representada pela probabilidade do evento e que satisfaz as seguintes propriedades (conhecidas como axiomas da probabilidade):

- 1 – $P(S) = 1$; isso determina que o resultado seja um ponto no espaço amostral.
- 2 – Se A e B são eventos disjuntos de S , então $P(A \cup B) = P(A) + P(B)$;
- 3 – Se $A_1, A_2, \dots, A_n$ são uma família de eventos de S , dois a dois, disjuntos, então $P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$.

Esses axiomas permitem manipulação de combinações de eventos.

Segundo Hajek (2007) existem três interpretações do conceito de probabilidade sendo eles o clássico, o frequentista e o subjetivo (bayesiano). Na definição clássica, é considerado um conjunto de eventos elementares que são considerados igualmente prováveis, então a probabilidade de um evento qualquer, formado pela união dos eventos elementares, é calculada como a razão entre o número de eventos elementares favoráveis aquele evento e o número total de eventos possíveis.

A definição frequentista leva em conta o número de eventos efetivamente obtidos em experimentos sob as mesmas condições. Se o número de eventos é grande o suficiente, a probabilidade é a razão entre o número de resultados favoráveis e o número total de experimentos. Quanto maior o número de experimentos, mais próxima do valor real está essa probabilidade.

Para o tratamento de sentenças matemáticas referentes a eventos que não se repetem, a noção de probabilidade é tratada, de maneira subjetiva, como grau de confiança sobre a veracidade de uma afirmação. Os parâmetros desconhecidos são vistos como variáveis aleatórias com uma densidade de probabilidade associada que é conhecida como distribuição a priori. Essa probabilidade a priori representa o conhecimento que se tem sobre a variável de interesse antes da realização do experimento. Depois que os dados são considerados, essa distribuição é recalculada e trata-se da distribuição a posteriori.

3.7. Modelos gráficos probabilísticos

Os modelos gráficos probabilísticos (MGP) têm sido amplamente utilizados na inferência de redes de regulação gênica a partir de dados experimentais, como por exemplo, dados de expressão gênica (Friedman *et al.*, 2000; Segal *et al.*, 2003; Friedman, 2004; Yeung *et al.*, 2002; Hoon *et al.*, 2003; Kishino & Waddell, 2000; Toh & Horimoto, 2002; Wu *et al.*, 2003; Dobra *et al.*, 2004). Isso se deve à necessidade de se descobrir e conhecer técnicas que permitam flexibilidade na inferência e representação de situações reais.

Modelos gráficos probabilísticos são grafos em que os nós representam as diferentes variáveis aleatórias envolvidas num problema e os arcos representam assunções condicionais de dependência, ou seja, relações probabilísticas entre elas, sendo que quando estes estão ausentes, é assumido que as variáveis são independentes. Isso proporciona uma representação compacta de distribuições de probabilidades conjuntas. Detalhes matemáticos como as equações ligando as variáveis e algoritmos para realizar e aproximar a inferência do modelo estão também associados aos grafos (Buntine, 1996).

De maneira resumida, eles são simplesmente modelos de distribuições conjuntas nos quais são assumidos paradigmas de amostragem aleatória (Dobra *et al.*, 2004). Eles apresentam uma arquitetura para estruturar, representar e decompor um problema usando a noção de independência condicional e possuem vários casos especiais e variações como as redes bayesianas, diagramas de influência, redes de Markov e redes probabilísticas causais (Buntine, 1996).

Através de um modelo probabilístico são descritas explicitamente as hipóteses que supomos sobre as relações entre dois tipos de dados e o processo de aprendizagem pode combinar evidências de vários conjuntos de dados e conclusões mais robustas pelo fato de tratarem

diferentes conjuntos de dados dentro um modelo (Friedman, 2004). Esses modelos descrevem a topologia (no sentido de dependências) dos componentes de um modelo de probabilidade complexo, esclarecem suposições sobre a representação e levam a algoritmos que fazem uso da topologia para atingir *speed-ups* exponenciais (Frey & Jojic, 2005).

Segundo Buntine (1996), eles são uma ferramenta de modelagem atrativa para descoberta do conhecimento por serem uma representação lúcida da variedade de problemas e isso permite que dependências chaves sobre um problema sejam expressas e irrelevâncias sejam ignoradas. Além disso, eles possuem técnicas bem entendidas para tarefas básicas na descoberta de processos. Os procedimentos para inferência em modelos gráficos são derivados dos princípios básicos da teoria da probabilidade e, além disso, a literatura em modelos gráficos fornece abordagens para aprendizagem de dados que são derivados de princípios básicos bem entendidos em estatística (Friedman, 2004).

Como descrito por Friedman (2004), para selecionar o modelo que melhor ajusta as observações atuais, é utilizado um procedimento de aprendizagem. Para conhecer os dados, fazer previsões e formular ideias e hipóteses, o modelo aprendido é usado. Esses modelos foram desenvolvidos nos campos de aprendizagem de máquina e estatística para modelagem de sistemas complexos com múltiplas entidades interagindo e apesar de serem relacionados com modelos de sequência probabilística, não são restritos a observações sequenciais. Através dos MGP as distribuições de probabilidade das observações são descritas por uma linguagem concisa.

Dentre os tipos de variáveis que podem ser representadas nos MGP, estão medidas reais ou discretas podendo ser variáveis cujos valores são desconhecidos, variáveis ocultas cuja existência é suposta ou parâmetros usados para especificar o modelo (Buntine, 1996). Essas variáveis podem ser representadas em diferentes ordens dependendo de que tipo de modelo o grafo está representando e ainda, eles podem ser manipulados para representar todas as diferentes visões de um modelo com uma base de conhecimento probabilística (Buntine, 1996).

A utilização dos grafos permite que o modelo de probabilidade seja visualizado do ponto de vista estrutural sem perda do detalhe matemático. Essas aproximações permitem o uso de observações para preencher muitos detalhes do modelo e também fornecem princípios para combinar modelos locais múltiplos em um modelo global conjunto. A flexibilidade ao construir modelos de inferência para novos conjuntos de dados ou desenhos experimentais existe porque

essas aproximações usam observações para a descrição de muitos detalhes do modelo e possuem princípios para combinação de modelos locais múltiplos em um modelo global conjunto (Friedman, 2004).

Dessa forma, ao usar modelos gráficos, é possível que sejam construídos primeiro submodelos simples, decompondo o problema e, depois estes subproblemas serão combinados para formular um modelo completo (Buntine, 1996). Por fim, os MGP possuem uma natureza declarativa, a qual fornece uma vantagem quando precisamos estender o modelo para que ele contemple aspectos adicionais dos sistemas ou novas observações (Dobra *et al.*, 2004).

Nos MGP, as distribuições de probabilidade conjuntas multivariadas são representadas por meio de um produto de termos, nos quais apenas algumas poucas variáveis estão envolvidas e um grafo que relata as variáveis que aparecem num termo comum representa a estrutura desse produto (Friedman, 2004). Esse grafo especifica a forma do produto da distribuição e também fornece ferramentas para inferência das propriedades vinculadas ao produto (Pearl, 1988).

Quando o modelo é um sistema biológico, o interesse são entidades que participam do sistema em estudo, no caso os genes, e seus diferentes atributos, nesse caso os níveis de expressão que são tratados como variáveis aleatórias (Friedman, 2004). Nas variáveis aleatórias, os níveis de expressão gênica de um experimento em particular são incluídos como atributos observados e alguns atributos ocultos são assumidos pelo modelo, como a atribuição do cluster de um gene particular e, nesse caso, a descrição da distribuição de probabilidade conjunta de todas as variáveis de interesse é incorporada pelo modelo (Friedman, 2004).

Através desse tipo de representação gráfica é possível explorar informações e gerar suposições sobre vias, pois ela oferece esse tipo de visão da expressão gênica (Dobra *et al.*, 2004). A montagem de distribuições multivariadas completas possibilita a percepção de informações além das simples interações de correlação, como abordagens de clusterização, onde é possível identificar grupos de genes relacionados com padrões de expressão (Dobra *et al.*, 2004).

Dobra *et al.* (2004) comenta que em um modelo gráfico, os vizinhos de primeira ordem de um gene específico G são os genes que juntos fornecem o conjunto de preditores de variações da expressão gênica em G e tornam G condicionalmente independente de todos os outros genes. Alguns genes podem ser pouco correlacionados com G , mas altamente correlacionados em

termos de correlação parcial considerando-se outros vizinhos e assim serão identificados como vizinhos de G.

Esses tipos de relacionamentos entre dois genes não poderiam ser captados por métodos baseados em correlação, e eles podem definir relações significantes numa via biológica. Assim, além da exploração de relações transitivas entre pares ou conjuntos de genes com base em caminhos curtos em grafos interessantes, um interesse chave aplicado nesses modelos reside na capacidade de expandir nossa capacidade de identificar genes candidatos que podem ter papéis importantes em um via específica sob investigação.

Os modelos probabilísticos usam estocasticidade para calcular medidas de ruído, variabilidade no sistema biológico, e aspectos do sistema que não são capturados no modelo (Friedman, 2004). Em um sistema estocástico, mais de um estado seguinte pode ser gerado por um dado nível de expressão e isso resulta que em diferentes células de uma mesma população um caminho de expressão gênica diferente possa ser seguido de um estado para outro (Lima, 2009). Essa propriedade é importante porque a resposta de indivíduos dentro de uma população de células geneticamente idênticas pode ser significativamente diferente da resposta da população média (Kaern *et al.*, 2003).

3.8. Redes bayesianas

Para métodos de inferência genética, há extensões de modelos de redes gênicas (Hickman & Hodgman, 2009) que podem ser usados para representá-las, sendo o método mais flexível as redes bayesianas, que podem ser descritas como um tipo de modelo probabilístico que representa relações causais entre variáveis (Pearl, 2000; Neapolitan, 2003).

Redes bayesianas consistem numa combinação entre teoria da probabilidade e teoria dos grafos. Uma rede bayesiana é especificada por uma estrutura gráfica, a qual define o relacionamento qualitativo causal entre os nós através das independências condicionais, uma família de distribuições de probabilidade condicional e seus parâmetros, que quantificam as relações probabilísticas causais entre os nós da estrutura (Husmeier, 2003).

Essas redes representam a estrutura de dependência entre entidades múltiplas que se interagem localmente, no caso, os níveis de expressão de diferentes genes e são úteis para descrever processos em que o valor de cada componente depende diretamente de valores de um número relativamente pequeno de componentes (Friedman *et al.*, 2000). Além disso, a natureza

probabilística desta abordagem é capaz de lidar com o ruído inerente tanto a processos biológicos quanto aos experimentos, tornando a inferência robusta e permitindo confiança nas estruturas de rede inferidas (Husmeier, 2003).

Sua estrutura gráfica é dada por um grafo acíclico direcionado, em que há somente arcos direcionados e não há ciclos direcionados. Essa estrutura acíclica garante que um nó não possa ser seu próprio pai. Os nós no grafo direcionado acíclico representam variáveis estocásticas e os arcos representam relações de dependências direcionadas entre variáveis que são quantificados por distribuições de probabilidade condicional (Sebastiani *et al.*, 2005). A família de distribuições de probabilidade condicional e seus parâmetros especificam a forma funcional das probabilidades condicionais associadas com os arcos, ou seja, eles indicam a natureza das interações entre os nós e a intensidade dessas interações (Werhli, 2007).

Lee *et al.* (2010) apresentam uma rede bayesiana para os genes de desenvolvimento da raiz lateral e regulador da sensibilidade à seca. Assim, eles mostram que a sensibilidade à seca é sinalizada pelo ácido abscísico, fatores de transcrição (DREB2A e ERD1) e pela enzima E3 SUMO ligase S1Z1 e que o desenvolvimento da raiz lateral é afetado pela proteína DDB1-Cu14 que degrada e regula o acúmulo de auxinas.

Bottcher & Dethlefsen (2003) apresentam as redes bayesianas considerando $G = (V, E)$ como um grafo acíclico direcionado que define a estrutura da rede bayesiana, onde V é um conjunto finito de nós e E é um conjunto finito de linhas direcionadas entre os nós. Para cada nó $v \in V$ no gráfico corresponde a uma variável aleatória X_v . O conjunto de variáveis associadas com o grafo G é então $X = (X_v)$ com $v \in V$. Para cada nó v , com pais $pa(v)$, uma distribuição de probabilidade local, $p(x_v | x_{pa(v)})$ é definida. Esse conjunto de distribuição de probabilidade local para todas as variáveis na rede é P . A rede bayesiana para um conjunto de variáveis aleatórias X é então o par (G, P) . As relações são representadas por arcos que se originam em um nó e atingem outro, sendo que o nó no qual a interação é originada é chamado nó pai e o nó atingido, nó filho ou descendente (Figura 3-5).

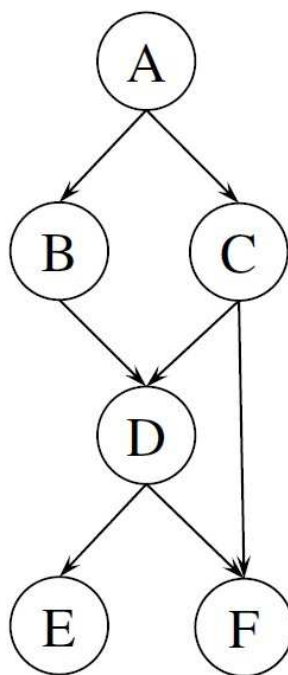


Figura 3-5 - Exemplo de uma rede bayesiana. A figura apresenta uma rede bayesiana hipotética composta pelo conjunto de nós $V = \{A, B, C, D, E, F\}$ e arcos $E = \{(A, B), (A, C), (B, D), (C, D), (D, E), (D, F), (C, F)\}$. Aplicando as relações de independência representadas no grafo, a probabilidade conjunta pode ser escrita como $P(A, B, C, D, E, F) = P(A) P(B|A) P(C|A) P(D|B, C) P(E|D) P(F|D, C)$. (Werhli, 2007).

A possível falta de arcos direcionados em G codifica independências condicionais entre as variáveis aleatórias (Sebastiani *et al.*, 2005). O conceito de independência condicional e marginal são conceitos substancialmente diferentes. Duas variáveis podem ser marginalmente independentes, mas podem ser dependentes, quando em uma terceira variável é adicionada. Por exemplo, duas variáveis pais são marginalmente independentes, mas tornam-se dependentes quando a terceira variável é adicionada e passam a ter um nó filho comum (Sebastiani *et al.*, 2005).

Uma rede bayesiana é caracterizada por uma simples e única regra para expandir a probabilidade conjunta em termos de probabilidades condicionais mais simples (Werhli, 2007). Segundo Werhli (2007), essa regra segue a propriedade local de Markov: “Um nó é condicionalmente independente de seus não descendentes dado seus pais”. Disso, surge a regra de fatoração ou regra da cadeia (3.1):

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i \mid \pi_M(X_i)) \quad (3.1)$$

Os mesmos símbolos são usados para representar os nós e as variáveis aleatórias que eles representam (X_i) e da mesma forma, o conjunto de nós pais e as variáveis aleatórias que eles representam também tem o mesmo símbolo ($\pi_M(X_i)$). Assim, na equação (3.1), $X_1, X_2, \dots, X_n$ são variáveis aleatórias representadas pelos nós $X_i \in 1, \dots, n$ e $\pi_M(X_i)$ é o conjunto de variáveis aleatórias representadas pelo conjunto de nós $\pi_M(X_i)$ que são os pais do nó X_i no modelo M .

A base de muitos algoritmos de busca para a aprendizagem de redes bayesianas a partir de dados se encontra nessa regra, pois, com esta decomposição, divide-se a distribuição total de probabilidade em módulos que podem interligar-se e a rede resumirá todas as dependências significativas sem que a informação se desintegre (Sebastiani *et al.*, 2005). Esta propriedade é usada para reduzir significativamente o número de parâmetros que são necessários para caracterizar a distribuição de probabilidade conjunta das variáveis (Ruggeri *et al.*, 2007).

O conjunto de nós pode possuir variáveis discretas e contínuas, então V é dado por $V = \Delta \cup \Gamma$, onde Δ e Γ são conjuntos de nós discretos e contínuos, respectivamente. O conjunto de variáveis X pode ser denotado como $X = (X_v)_{v \in V} = (I, Y) = ((I_\delta)_{\delta \in \Delta}, (Y_\gamma)_{\gamma \in \Gamma})$, onde I e Y são conjuntos de variáveis discretas e contínuas, respectivamente. Para uma variável discreta, δ , Γ_δ denota o conjunto de níveis.

Para garantir métodos de computação local exata, não é permitido que variáveis discretas tenham pais contínuos. A distribuição de probabilidade conjunta (3.2), então, pode ser fatorada numa parte discreta onde é considerado o produtório das probabilidades das variáveis discretas dados seus pais que só podem ser discretos e numa parte mista onde é considerado o produtório das probabilidades das variáveis contínuas dados seus pais que podem ser tanto discretos quanto contínuos, então:

$$p(x) = p(i, y) = \prod_{\delta \in \Delta} p(i_\delta \mid i_{pa(\delta)}) \prod_{\gamma \in \Gamma} p(y_\gamma \mid i_{pa(\gamma)}, y_{pa(\gamma)}) \quad (3.2)$$

As redes bayesianas estão aptas a capturar interações lineares e não lineares entre grupos de variáveis, ao mesmo tempo em que reduzem o número de parâmetros requeridos para descrever o modelo em comparação com outras modelagens (Woolf *et al.*, 2004).

Na aprendizagem de redes bayesianas, a partir de um conjunto de dados serão induzidas as distribuições condicionais, que quantificam a estrutura de dependência (estimativa de parâmetros), e as relações de interdependência entre as variáveis de um domínio de dados, que seria a criação gráfica de dependências condicionais (seleção de modelos) (Luna, 2004). O resultado será um grafo direcionado acíclico com um conjunto de probabilidades condicionais parametrizadas que melhor explica os dados (Werhli, 2007).

O processo de aprendizagem da rede consiste em dois estágios, nos quais, primeiro é selecionada uma estrutura de rede aleatória, com os arcos que ligam as entidades e, segundo, o conjunto de parâmetros associado a esses arcos é calculado e também se essas relações entre as entidades são ativadoras ou inibidoras tão bem como sua intensidade (Werhli, 2007). A identificação automática do modelo gráfico que melhor ajuste os dados é uma tarefa difícil que requer a identificação automática de dois componentes: a pontuação métrica (*score*) para escolher o melhor modelo e uma estratégia de pesquisa (algoritmo) para explorar o espaço de possíveis modelos alternativos (Sebastiani *et al.*, 2005).

Como descrito por Bansal *et al.* (2007), para fazer a inferência de um modelo de rede bayesiana para redes de regulação gênica, é necessário encontrar o grafo acíclico direcionado G (ou seja, os reguladores de cada transcrição) que melhor descreve o conjunto de dados de expressão d , onde d é assumido ser um conjunto de dados *steady-state*. Isto é realizado pela escolha de uma função que mede o *score* de cada grafo G (isto é, uma possível topologia da rede) em relação aos dados de expressão, e então busca pelo grafo que maximiza o *score*. A pontuação métrica (*score*) pode ser definida usando a regra de Bayes (3.3):

$$P(G | d) = \frac{P(d | G) * P(G)}{P(d)} \quad (3.3)$$

Onde $P(G)$ pode tanto conter algum conhecimento a priori sobre a estrutura da rede, se disponível, ou pode ser uma priori constante não informativa, e $P(d|G)$ é uma função, a ser escolhida pelo algoritmo que avalia a probabilidade de que os dados d tenham sido gerados pelo grafo G .

Os métodos mais populares para quantificação do *score* são *Minimal Description Length* (MDL) e a Equivalência bayesiana de *Dirichlet* (BDe). Ambos os *scores* incorporam uma penalidade para a complexidade a fim de evitar que os dados sejam super-ajustados. O BDe *score*

origina da estatística bayesiana e corresponde a probabilidade a posteriori de uma da rede para um conjunto de dados. Já *MDL score* origina da teoria da informação e corresponde ao tamanho dos dados comprimidos com o modelo de compressão derivado da estrutura da rede e também tem uma interpretação estatística como uma aproximação da probabilidade a posteriori (Wilczyński & Dojer, 2008).

Testar todas as possíveis combinações de interações entre os genes, isto é, todos os gráficos possíveis, e escolher o grafo com maior *score* bayesiano é um problema complicado. Para isso, métodos de busca heurísticos são utilizados, como a aproximação de *greedy-hill climbing*, método de Monte Carlo via cadeias de Markov ou *Simulated annealing*.

Um dos problemas geralmente encontrado na aprendizagem é que geralmente são encontradas várias redes com *score* alto ou indeterminado. Para contornar o problema, pode ser usada a média do modelo ou *bootstrap* para selecionar as interações regulatórias mais prováveis e obter estimativas de confiança para as interações. Por exemplo, se uma interação particular entre dois transcritos ocorre repetidamente nos modelos de maior *score*, provavelmente esse arco é uma dependência verdadeira. Alternativamente, pode-se aumentar um conjunto de dados incompletos com informação a priori para ajudar a selecionar o modelo de estrutura mais comum.

3.9. Softwares para redes bayesianas

Existem vários softwares para aprendizagem e inferência de redes bayesianas. Aqui serão revisados quatro deles, sendo eles *BaNJO* (Hartemink, 2006), *BNFinder* (Wilczyński & Dojer, 2008) e *PEBL* (Shah & Woolf, 2009) e o R, que possui dois pacotes específicos para isso: *Deal* (Bottcher & Dethlefsen, 2003) e *BNArray* (Chen *et al.*, 2006).

3.9.1. *BaNJO: Bayesian Network Inference with Java Objects*

BaNJO é um software desenvolvido (em Java) por um grupo da *Duke University*, sob direção de Hartemink, para inferência de redes de regulação gênica que utiliza a abordagem bayesiana. Eles propõem uma inferência eficiente da estrutura de redes ao analisar grandes conjuntos de dados. Vários trabalhos recentes têm utilizado esse software (Bose *et al.*, 2006; 2008; Cantone *et al.*, 2009; Eldawlatly *et al.*, 2010; Volinia *et al.*, 2010; Eckel *et al.*, 2009).

O *BaNJO* permite implementar tanto redes bayesianas estáticas, a partir de dados estacionários quanto dinâmicas a partir de dados de séries temporais, porém, em dados

dinâmicos, necessita-se de um número muito elevado de experiências (pontos de tempo) em comparação com o número de genes para que a estimativa seja mais precisa (Yu *et al.*, 2004).

Bansal *et al.* (2007) fizeram um estudo avaliando a performance de alguns programas para inferência de redes gênicas, dentre eles, o *BaNJO*. Eles testaram diferentes tamanhos de conjuntos de dados e diferentes estratégias para obtenção das redes. Eles encontraram limitações de memória ao executar o *BaNJO* com um conjunto de dados grande. De fato, mesmo trabalhando com poucos genes, se o número de experimentos for muito grande, o programa passa a exigir muita memória para ser executado. Isso pode ser uma desvantagem, pois algoritmos probabilísticos, como redes bayesianas em geral, fazem estimativas das distribuições de densidade de probabilidade, tarefa que exige muitas repetições no conjunto de dados.

Eles também observaram que o desempenho do *BaNJO* é ruim com poucos experimentos ($n = 10$), mas alcança boa precisão com um número maior de experimentos ($n = 100$), independente do número de genes, embora com uma sensibilidade muito baixa (apenas poucos arcos são achados). O *BaNJO* também mostrou baixo desempenho com conjuntos de dados onde havia ruído.

Este software dá ao usuário uma variedade de opções para seus parâmetros, por exemplo, o usuário pode determinar por quanto tempo o programa será executado e então mesmo que ele não tenha calculado o *score* de todas as redes possíveis, ele irá gerar um resultado com a rede de melhor *score* até o momento em que ele parou de ser executado. Isso pode ser uma vantagem, pois quando se trabalha com um número muito grande de genes, o número de redes possíveis tende a infinito, fazendo com que a busca pela rede de maior *score* dure por tempo indeterminado.

Outro recurso interessante é a opção “threads=N” onde são feitas N linhas de busca, uma em cada processador, utilizando as mesmas configurações e no final os resultados são combinados em um único *output*. Além de fornecer a rede com maior *score*, ele também pode fornecer a rede de consenso entre as melhores redes, sendo possível escolher o número de melhores redes para que ele monte a rede de consenso.

Este software aceita somente variáveis discretas e estas podem assumir até sete estados diferentes. Para dados contínuos, na parte de pré-processamento dos dados, ele fornece a opção de discretização das variáveis que pode ocorrer em no máximo 7 classes. Isso pode acarretar

perda de informação. O *BaNJO* permite um número maior de nós podendo fazer redes com cerca de até 1000 nós desde que o número de experimentos não seja grande.

Conforme a documentação do programa (Sladeczek *et al.*, 2005), ele consiste em um algoritmo de busca (*Searcher*), cuja principal tarefa é gerenciar todos os aspectos da busca num espaço de todas as redes possíveis para achar a rede com maior *score*. Essa busca é construída como um ciclo que é executado até que o programa chegue a alguma das condições de finalização definidas pelo usuário ou já tenha avaliado todas as estruturas de redes possíveis. Há duas opções de *Searchers*, o *Greedy* e o *Simulated Annealing*. Esse *Searcher* pode ser dividido em quatro componentes conforme a Figura 3-6.

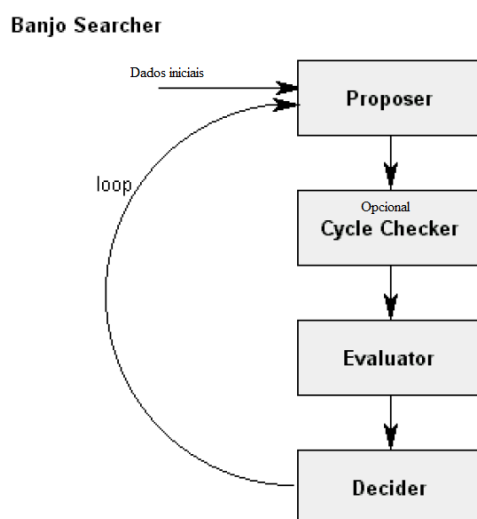


Figura 3-6 – Algoritmo de busca (*Searcher*) e seus componentes. O *Proposer* sugere uma rede a partir dos dados iniciais, é verificado se essa rede contém ciclos pelo *Cycle Checker*, o *Evaluator* calcula o *score* da rede e o *Decider* avalia se o *score* é melhor que a rede atual ou não, e assim decide se continua com a rede atual ou a substitui pela nova.

O primeiro deles é o *Proposer* que implementa a parte do algoritmo de busca que especifica quais mudanças acontecerão na rede em um único passo da iteração de busca. São oferecidas duas opções de *Proposers*, o *Random Local Moves* que simplesmente seleciona um movimento ao acaso na rede dentre todos os movimentos locais disponíveis e o *All Local Moves* que compõe uma lista de todos os movimentos locais disponíveis, dado o estado atual da rede e seleciona o movimento que produz a rede com maior pontuação. Em teste com alguns dados de levedura, foi percebido que o método *Random Moves* resulta em redes com maior *score* que o *All Local Moves* em combinação com quaisquer algoritmos de busca.

O próximo componente irá checar se a rede proposta contem ciclos. Se houver ciclos, essa rede é descartada e o *Searcher* volta a propor uma nova mudança na rede. O *Evaluator* vai computar o *score* de uma rede, baseado na medida de BDe. E depois o *Decider* vai determinar se a rede proposta na iteração atual será aceita como rede atual para a próxima iteração ou se será rejeitada, e nesse caso a busca continua com a rede atual.

3.9.2. *BNFinder*

O *BNFinder* é um software implementado em *Python* por Wilczyński & Dojer (2008) e vem sendo bastante utilizado em estudos recentes (Dabrowski *et al.*, 2010; Chai *et al.*, 2012; Bonn *et al.*, 2012; Wilczyński *et al.*, 2012). Ele possui uma versão online que suporta redes com até 20 nós e 500 observações experimentais (<http://bioputer.mimuw.edu.pl/software/bnf/>). Assim como o *BaNJO*, *BNFinder* pode também aprender tanto redes dinâmicas (a partir de séries temporais) quanto estáticas (a partir de dados experimentais independentes). Porém, no segundo caso é necessário especificar algumas limitações para a estrutura da rede, forçando sua aciclicidade.

Segundo Wilczyński & Dojer (2008) ele foi projetado para economizar velocidade e ter uma qualidade boa na aprendizagem em uma ampla classe de problemas que ocorrem em biologia. Ele trabalha sob o pressuposto de que não há necessidade de checar a aciclicidade do grafo, pois essa é satisfeita ao lidar com redes dinâmicas, nas quais as distribuições condicionais referem-se a variáveis aleatórias em pontos de tempo vizinhos e o grafo é sempre acíclico. E, no caso de redes estáticas, o conjunto de estruturas de redes possíveis deve ser restrito. *BNFinder* deixa o usuário dividir o conjunto de variáveis em um conjunto ordenado de subconjuntos disjuntos de variáveis, onde arcos podem ser colocados somente dos conjuntos *upstream* para *downstream*. Se tal ordenação não é conhecida de antemão, pode-se tentar executar o *BNFinder* com ordenações diferentes e escolher uma rede com a melhor pontuação geral.

BNFinder aprende redes ótimas com respeito a dois critérios de *score* geralmente usados: Equivalência bayesiana de *Dirichlet* (BDe) (que é o padrão) e *Minimal Description Length* (MDL). O algoritmo trabalha em tempo polinomial para ambos os *scores*, mas cálculos com o MDL são mais rápidos, especialmente para conjuntos grandes de dados. Essa é a principal vantagem do *BNFinder*, pois utilizando o MDL *score* é possível fazer redes maiores do que os outros softwares citados permitem. No entanto, o uso do BDe *score* deve ser preferido devido ao

seu melhor desempenho na interpretação estatística. Os *scores* MDL e BDe foram originalmente concebidos para variáveis discretas. As variáveis contínuas são tratadas com pontuações correspondentes, derivadas sob a assunção que distribuições condicionais pertencem a uma família de misturas Gaussianas.

Desde que perturbações mudam a estrutura das interações, procedimentos de aprendizagem tendem a usar dados seletivamente. *BNFinder* manipula perturbações, da seguinte maneira: para calcular o *score* de conjuntos de pais de uma variável v , ele em conta somente os experimentos onde v não é perturbado.

Uma distribuição a priori da estrutura da rede pode ser especificada através da atribuição de pesos para as interações potenciais entre as variáveis. Além disso, o tamanho do conjunto de reguladores de cada variável pode ser limitado a um dado número e o espaço de possíveis distribuições de probabilidade condicional de variáveis selecionadas pode ser restrito. É possível definir também o número máximo de pais para cada nó.

Este software não possui critérios para que a inferência seja finalizada, assim, a execução do programa só termina quando todo o espaço de redes possíveis, seguindo as restrições, foi acessado. Isso é vantajoso, pois ao contrário do *BaNJO*, o resultado será a rede de maior *score* dentre todas as redes possíveis.

3.9.3. PEBL: *Python Environment for Bayesian Learning*

O PEBL é uma biblioteca do Python para aprendizagem de redes bayesianas desenvolvida por Shah & Woolf (2009) e como requisito necessita da biblioteca *numpy* (Ascher *et al.*, 2001) para operações de matrizes de alto desempenho. O formato de entrada dos dados utilizados no programa devem ser arquivos de texto, delimitados por tabulações e ele é capaz de lidar com variáveis tanto discretas quanto contínuas. Sendo esta uma de suas vantagens, pois não ocorre perda de informação.

Para calcular o *score* das redes, o PEBL utiliza o *score* BDe e uma vantagem é que ele pode trabalhar com valores ausentes e variáveis ocultas usando marginalização exata e amostragem de *Gibbs*. O PEBL suporta o ajuste de probabilidades a priori e possui duas opções de algoritmos de busca, assim como o *BaNJO*, o *Greedy Hill-Climbing* e *Simulated Annealing*.

Assim como o *BaNJO*, o PEBL pode executar tarefas de aprendizagem em paralelo ao longo de múltiplos processadores ou núcleos de CPU. Com configurações adequadas e da utilização de execução paralela, PEBL pode ser usado para grandes tarefas de aprendizagem. Apesar de ser um programa semelhante ao *BaNJO*, ele possui menor sensibilidade, ou seja, as redes geradas por ele possuem poucos arcos. Essa característica pode ser tratada como vantajosa dependendo do objetivo do estudo.

3.9.4. R: *Deal* e *BNArray*

O *deal* é um pacote implementado por Bottcher & Dethlefsen (2003) em R (*Development Core Team*, 2003) para a aprendizagem de redes bayesianas, porém seu uso não é direcionado para dados de expressão gênica. Ele permite a aprendizagem somente de redes estáticas com variáveis contínuas e/ou discretas, o que é uma vantagem, pois não há necessidade de discretizar variáveis e não há perda de informação.

O *deal* inclui procedimentos para definir prioris, estimar parâmetros, calcular *score* das redes, executar a busca heurística, tão bem como, simular conjuntos de dados com uma dada estrutura de dependência. Para aprendizagem da rede, é necessário que o usuário forneça os dados de treino e representa qualquer conhecimento a priori disponível.

Esse pacote utiliza a rede bayesiana a priori para deduzir distribuições para todos os parâmetros no modelo. Então, isso é combinado com os dados de treino para calcular as distribuições de probabilidade a posteriori dos parâmetros. Um *score* é calculado para a rede e uma estratégia de busca é empregada para procurar a rede com *score* mais alto.

Chen *et al.* (2006) desenvolveram um pacote que serve de extensão para o *deal*, no contexto de dados de expressão gênica, o *BNArray*¹. Este pacote é uma ferramenta para implementar um processo sistematizado de reconstrução de sub-redes regulatórias a partir de dados de *microarrays*.

Para a construção das redes ele utiliza as funções disponíveis no *deal*, mas possui três ferramentas úteis quando se está trabalhando com dados de expressão gênica. Uma delas permite estimar valores ausentes nos dados com o algoritmo *Least Local Squares* (LLSimpute) para completar os dados de entrada para construção das redes. Outra é a re-amostragem do conjunto de dados para produzir dados mais confiáveis usando *bootstrap*. Essa função é implementada

¹ <http://www.cls.zju.edu.cn/binfo/BNArray/>

pelo procedimento *BootstrapBN*. E a terceira ferramenta, xCODENSE é uma ferramenta que foi estendida para grafos direcionados que permite a reconstrução de sub-redes coerentes e significantes. Essa ferramenta utiliza uma função interna para mineração dos subgrafos altamente conectados, assim, são derivados subgrafos frequentes de arcos altamente confidentes previamente selecionados, que oferecem as redes potenciais para relações regulatórias reais entre os genes.

A principal desvantagem do *deal* e *BNArray*, é que eles exigem muito esforço computacional, o que torna a inferência da rede muito lenta a medida que o número de genes aumenta. Isso acontece porque o espaço de busca, isto é, número de redes possíveis, aumenta exponencialmente com o aumento do número de nós. O número de nós que o *BNArray* permite na rede para que seja viável usá-lo na inferência de uma rede não ultrapassa 20. Isso porque esses pacotes não apresentam parâmetros como critérios de parada, assim o programa obrigatoriamente acessa todas as estruturas de redes possíveis para obter a de maior *score*.

Watanabe *et al.* (2012) propõem um tipo de paralelização para contornar esse problema, pois uma paralelização eficaz permite a construção de redes em grande escala. O método apresentado por eles pode ser dividido em três passos (Figura 3-7):

Passo 1: Divide-se o conjunto de todos os genes V em subconjuntos com todas combinações de três genes.

Passo 2: O *score* de todas as redes possíveis para cada trio é calculado de maneira independente para o obter a melhor rede para cada trio.

Passo 3: A melhor rede para cada trio é unida em uma única rede.

Mas essa abordagem não aumenta muito o número de genes que pode ser trabalhado no *BNArray*, pois a proposta principal desse trabalho foi apenas utilizar um método que estime as redes mais precisamente gastando o mesmo tempo que o *BNArray* gastaria sem a paralelização. Então, a execução do programa levaria o mesmo tempo paralelizado ou não, porém com a paralelização o resultado seria mais preciso, ou seja, haveria uma maior chance de encontrar a rede que realmente corresponde à realidade.

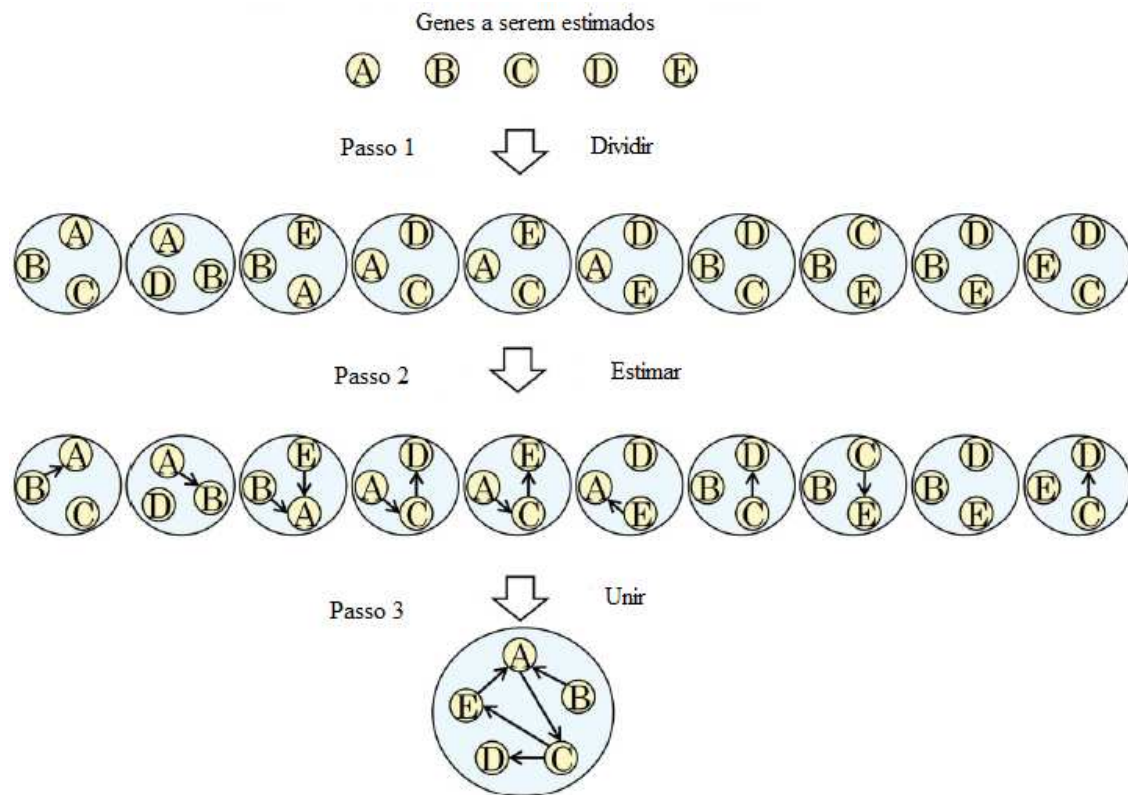


Figura 3-7 - Representação do algoritmo. Gera todas as combinações de trios possíveis a partir dos genes do conjunto de dados. Depois é selecionada a melhor rede bayesiana para cada conjunto de três genes, então essas redes são unidas em uma única rede incluindo todos os genes. (Adaptado de Watanabe *et al.*, 2012).

4. MATERIAL E MÉTODOS

4.1. Conjunto de dados de *microarrays*

Diversos projetos de sequenciamento de *expressed sequence tags* (EST) de cana-de-açúcar forneceram uma base gênica para estudos de perfil de expressão possibilitando a identificação de genes que possuem expressão relacionada com diversas situações. Porém, esses estudos foram baseados em lâminas de *microarray* de nylon ou vidro personalizados utilizando subconjuntos selecionados de ESTs e fragmentos de DNA correspondentes à necessidade de cada trabalho (Tabela 4-1). O material utilizado e a composição de genes de cada experimento são diferentes, e isso impossibilita a comparação dos perfis de transcrição obtidos por diferentes laboratórios (Casu *et al.*, 2007).

Apesar da vantagem de se poder combinar informações de diferentes fontes na elaboração das redes, não é possível integrar as informações de experimentos que possuem alguns genes ou não possuem outros, pois o conjunto de dados ficará com valores faltantes. Sendo assim, nesse trabalho optamos por utilizar os trabalhos de Casu *et al.* (2007) e McCormick *et al.* (2008), pois os dois utilizaram a plataforma da *Affymetrix* denominada Gene Chip *Sugarcane Genome Array*² para obtenção de seus dados de expressão gênica e portanto são os únicos trabalhos que apresentam conjunto de dados comparáveis.

² *Affymetrix*, Santa Clara, CA, EUA; <http://www.affymetrix.com>

Tabela 4-1 - Lista de trabalhos que disponibilizaram experimentos com dados de cana-de-açúcar.
Em negrito, são os trabalhos selecionados para serem utilizados nesse estudo.

Órgão	Tratamento ou Caractere estudado	Referência
Folha e Colmo	Sacarose e especificidade de tecidos	Carson <i>et al.</i> , 2002
Folha	Resposta a frio e seca	Nogueira <i>et al.</i> , 2003
Folha	Inducibilidade por etanol para descoberta de promotores	Camargo <i>et al.</i> , 2007
Colmo	Especificidade no colmo para descoberta de promotores	Damaj <i>et al.</i> , 2010
Colmo	Acúmulo de sacarose e especificidade genotípica	Casu <i>et al.</i> , 2004
Colmo	Acúmulo de sacarose e especificidade genotípica	Casu <i>et al.</i> , 2005
Raiz	Metil jasmonato e indução a defesa	Bower <i>et al.</i> , 2005
Múltiplos	Genes expressos constitutivamente e em tecidos específicos	Papini-Terzi <i>et al.</i> , 2005
Múltiplos	Metil jasmonato, ABA, seca, deficiência de fosfato	Rocha <i>et al.</i> , 2007
Folha	Resposta a CO ₂	de Souza <i>et al.</i> , 2008
Colmo	Acúmulo de sacarose e especificidade genotípica	Papini-Terzi <i>et al.</i> , 2009
Folha	Fonte de sacarose e especificidade genotípica	Felix <i>et al.</i> , 2009
Colmo	Síntese de parede celular	Casu <i>et al.</i>, 2007
Folha	Relações fonte-dreno	McCormick <i>et al.</i>, 2008

4.2. Experimento de metabolismo de parede celular

Esse trabalho foi conduzido por Casu *et al.* (2007), sendo a primeira avaliação da *Gene Chip Sugarcane Genome Array* produzido pela *Affymetrix* visando identificar transcritos expressos diferencialmente em relação ao desenvolvimento e acúmulo de sacarose no colmo. Esse experimento leva em conta a relação entre crescimento da planta e acúmulo de lignina e o acúmulo de sacarose. Isso porque quando as células do colmo param de crescer, começa a ocorrer aumento da espessura da parede celular, lignificação e partes basais do colmo, que já terão completado seu crescimento, começam a acumular sacarose. Porém, os entrenós superiores ainda estão jovens e em crescimento. Como o alongamento do colmo é sensível a temperatura, nos meses frios há aumento do acúmulo de sacarose nos colmos jovens, pois, fotoassimilados são desviados do crescimento para acúmulo, e isto coincide com a fase de maturação.

Esse experimento consiste de amostra de três tecidos diferentes, portanto três tratamentos, colhidos de plantas da variedade Q117 crescidas no campo, com quatro repetições. Essas plantas foram cultivadas como uma cultura comercial na Austrália e colhidas após 11 meses de plantio.

Foram recolhidas amostras do colmo dos tecidos:

Jovem – meristema e entrenós 1-3 (M, I1-3);

Colmo em maturação – entrenó 8 (I8);

Maduro – entrenó 20 (I20).

O *array* de cana possui 8.236 sondas que correspondem a expressão de cerca de 6,024 genes. Os dados foram obtidos do *Gene Expression Omnibus*³ onde se encontravam disponíveis sob número de acesso GSE5021. Com esse trabalho, Casu *et al.* (2007) identificaram 124 sondas diferencialmente expressas nos três tratamentos.

4.3. Experimento de manipulação da fotossíntese

Os dados desse experimento foram publicados por McCormick *et al.* (2008). Nesse trabalho, eles utilizaram o frio em folhas crescidas no campo para aumentar artificialmente concentrações de sacarose foliares e determinar se existia evidência de alterações na expressão gênica associada à regulação da fotossíntese. O objetivo deles era testar a hipótese de que o acúmulo de sacarose em plantas C₄ suprime fotossíntese na folha através da expressão diferencial de genes relacionados à sinalização de açúcar como ocorre em plantas C₃.

Para isso, plantas de *Saccharum* spp. (L.) híbridas da variedade N19 com 12 meses de idade cultivadas no Monte Edge Combe, KwaZulu-Natal (SASRI) foram usadas no estudo conduzido em 2006. Duas semanas antes do tratamento de frio, as plantas foram irrigadas duas a três vezes por semana para assegurar hidratação adequada. Em todos os tratamentos, foram utilizadas as segundas e terceiras folhas.

O experimento foi projetado para que as plantas fossem expostas ao frio toda manhã, às 8:00 horas por um período de 80 horas, e todas as folhas foram coletadas às 16:00 horas, no final do dia. Havia dois tratamentos: plantas expostas ao frio e controle, com quatro repetições cada um.

³ GEO <http://www.ncbi.nlm.nih.gov/geo/>

A abundância de transcritos em cada amostra de RNA da folha foi analisada usando também o *Affymetrix Gene Chip Sugarcane Genome Array* (Affymetrix, Inc., Santa Clara, CA, USA) da mesma maneira que foi descrito para o experimento de metabolismo de parede celular. Os dados contendo os arquivos CEL e PHP foram obtidos do *Gene Expression Omnibus* onde estavam disponíveis sob número de acesso GSE11934.

Das 8236 sondas do chip, McCormick *et al.* (2008) identificaram 673 como significativamente diferencialmente expressas nos dois tratamentos. Para focar o resultado em eventos relacionados com carboidrato na folha e fotossíntese eles selecionaram um conjunto de 143 sondas que foram identificadas nas categorias: 1) metabolismo de carboidrato; 2) fotossíntese; 3) sinalização de açúcar; 4) metabolismo mitocondrial; 5) metabolismo de parede; e 6) resposta a estresse. Esse grupo representa 74 genes.

4.4. Seleção dos genes de interesse pelo método *Heuristic Cluster Chiseling Algorithm*

Como a metodologia de redes bayesianas não permite que redes com um grande número de genes sejam inferidas devido ao grande esforço computacional que exigem. Assim, foi necessário selecionar alguns conjuntos menores de genes. Para isso, utilizou-se o método de *Heuristic Cluster Chiseling Algorithm* (HCCA) que separa uma rede grande em grupos menores de genes mais correlacionados (Figura 4-1).

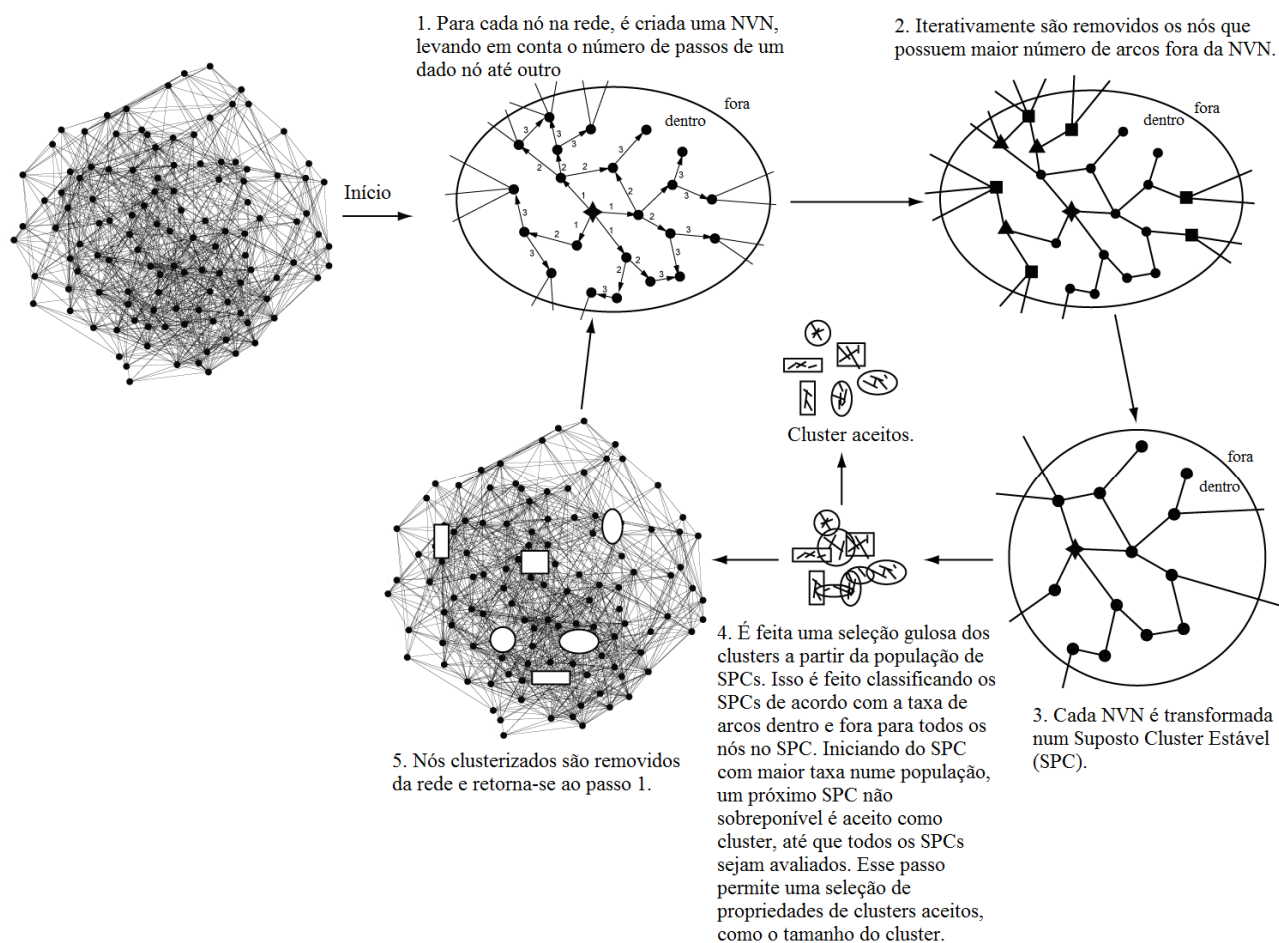


Figura 4-1 - Princípios do *Heuristic Cluster Chiseling Algorithm* (HCCA). HCCA consiste de cinco passos; um: geração da NVN para cada nó na rede levando em conta os n-passos da vizinhança. 2: remoção dos nós mais conectados com nós fora da NVN resultante, 3: geração de SPCs, 4: seleção de clusters da população de SPCs, 5: remoção dos nós clusterizados da rede e retorno ao passo 1. (Adaptado de Mutwill, 2010).

Essa metodologia para divisão da rede em pequenas unidades gerenciáveis ou cluster de genes mais relacionados foi proposta por Mutwill *et al.* (2010) com o objetivo de produzir efeitos visuais de redes que podem ser facilmente interpretados. Essa metodologia, apesar da divisão, visa manter o maior número possível de relações biológicas relevantes.

Para que o método seja empregado, é necessário ter a rede dos genes do conjunto de dados, que pode ser obtida de forma fácil e rápida através do cálculo da co-expressão de todos os genes contra todos utilizando a correlação de Pearson. Sendo que o valor mínimo de coeficiente de regressão linear utilizado foi 0,8. A correlação de Pearson captura apenas relações lineares entre quaisquer dois componentes de dados e seu coeficiente define a relação mútua de co-expressão entre dois genes.

Após ter gerado a rede, é medida a *Node Vicinity Networks* (NVN, também chamada de vizinhança) que é uma parte da rede gerada levando em conta todos os nós a um número de passos de um dado gene (Figura 4-2).

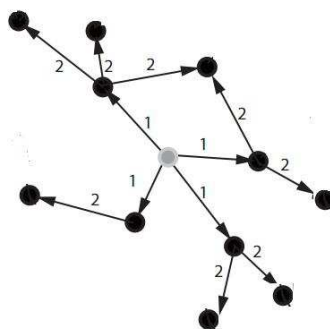


Figura 4-2 - Node Vicinity Networks. Exemplo da vizinhança de um gene (em cinza) considerando todos os genes a dois passos dele (<http://aranet.mpimp-golm.mpg.de/aranet/FAQ>).

O número de passos da NVN vai determinar o tamanho do cluster. Para cada nó da rede, o algoritmo gera NVNs considerando todos os nós a n passos de distância do dado nó. Nós com maior conectividade fora da NVN são iterativamente removidos. Os clusters resultantes são então classificados pela taxa de conectividade de fora-para-dentro e filtrados de acordo com o tamanho de cluster definido. Clusters que não se sobrepõem são mantidos pelo algoritmo, e os nós desses grupos são removidos da rede. Nós associados com cluster rejeitados são devolvidos à rede e reavaliados. O HCCA recursivamente cria clusters que não se sobrepõem até que nenhum dos nós sejam deixados na rede ou não seja mais possível obter clusters estáveis (Figura 4-1). Neste último caso, os nós restantes são associados com os clusters que se mostrarem maior conectividade.

Outro parâmetro proposto por Mutwill *et al.* (2010) é o cálculo do *Highest Reciprocal Rank* (HRR), que mede a similaridade da expressão e, é assumido ter como valor máximo 30, ou seja, o número máximo de arcos que um nó pode ter é limitado a 30. Esse índice é calculado para cada par de nós e como exemplo, vamos considerar os genes *AtCesal* e *Ctl1* de *Arabidopsis*. Primeiro, vamos considerar o *AtCesal*. Todos os genes relacionados com ele são ordenados do maior para o menor de acordo com o valor do coeficiente de regressão linear (R), e é atribuído um valor que corresponde à posição desse gene (*rank*), conforme a Figura 4-3.

valor de R	sonda	gene	descrição	rank
1	253428_at	at4g32410	AtCesa1	0
0.88646	247552_at	at5g60920	Cobra	1
0.8779	250827_at	at5g05170	AtCesa3	2
0.81694	247251_at	at5g64740	AtCesa6	3
0.77653	261279_at	at1g05850	Ctl1	4
0.76322	248573_at	at5g49720	Korrigan	5

Figura 4-3 - Lista dos genes relacionados com *AtCesa1*. Genes ordenados e ranqueados de acordo com o valor do coeficiente de correlação de Pearson; em azul *ATCesa1* e em verde *Ctl1* (<http://aranet.mpimp-golm.mpg.de/aranet/FAQ>).

O mesmo é feito para o outro gene do par, no caso *Ctl1* (Figura 4-4).

valor de R	sonda	gene	descrição	rank
1	261279_at	at1g05850	Ctl1	0
0.8396	250827_at	at5g05170	AtCesa3	1
0.83488	250645_at	at5g06700	Expressed protein	2
0.82953	247552_at	at5g60920	Cobra	3
0.80783	254239_at	at4g23400	Major intrinsic family protein	4
0.78417	248393_at	at5g52060	BAG-domain containing protein	5
0.77625	253428_at	at4g32410	ATCesa1	6
0.77118	251962_at	at3g53420	plasma membrane intrinsic proteir	7
0.76532	250317_at	at5g12250	tubulin beta-6 chain	8
0.76417	248573_at	at5g49720	Korrigan	9

Figura 4-4 - Lista dos genes relacionados com *Ctl1*. Genes ordenados e ranqueados de acordo com o valor do coeficiente de correlação de Pearson; em azul *ATCesa1* e em verde *Ctl1* (<http://aranet.mpimp-golm.mpg.de/aranet/FAQ>).

Então o cálculo do HRR entre um par de genes A e B é dado pela posição do gene do par que tiver maior valor absoluto na lista de *ranks*, seguindo a fórmula:

$$HRR(A,B) = \max(r(A,B), r(B, A))$$

No caso do exemplo, o HRR seria igual a 6. O HRR é dividido em três classes: 10 (para arcos com HRR menor ou igual a 10), 20 (para arcos com HRR maior que 10 e menor ou igual a 20) e 30 (para genes com HRR maior que 20 e menor ou a 30).

4.5. Conjunto de dados

Considerando as limitações em se trabalhar com um grande número de genes, foram selecionados conjuntos de dados menores. Assim, para a construção das redes de cana foram utilizados os seguintes conjuntos:

- Genes que foram diferencialmente expressos tanto no trabalho de Casu *et al.* (2007) quanto no trabalho de McCormick *et al.* (2008) totalizando 265 genes.
- Conjunto de genes que foram diferencialmente expressos no trabalho de McCormick *et al.* (2008) e ele definiu como fazendo parte do grupo de genes relacionados com sinalização e acúmulo de açúcar além de poderem ter importante papel na regulação de relações fonte e dreno, totalizando 143 sondas. Dessas algumas correspondiam a um só gene e então foi tirada a média dos valores de expressão, totalizando 136 genes.
- Genes que faziam parte do cluster no qual se encontrava a enzima Sacarose fosfato sintase, definido pelo método de HCCA sendo composto por 107 genes.
- Genes dentro do cluster da Sacarose fosfato sintase com NVN = 1 e HRR = 30 e NVN = 2 e HRR = 10, totalizando 25 genes.

4.6. Normalização dos dados *Affymetrix*

A normalização dos dados de cana foi feita utilizando o R 2.9.2 com a função “*rma*” do pacote *affy* do Bioconductor. A correção do *background* foi feita pelo método *Robust Multi-array Average* (RMA). Essa correção é baseada na assunção de que o sinal de *probe match* observado consiste de um *background* normalmente distribuído ou componente de ruído e um sinal exponencialmente distribuído. As intensidades observadas são ajustadas pela sua substituição pelo sinal verdadeiro esperado.

4.7. Experimento de manipulação fonte e dreno

Neste experimento, variedades contrastantes para acúmulo de sacarose (SP83-2847 e SP91-1049) foram avaliadas quanto à plasticidade que apresentam em suprir a um aumento na necessidade de sacarose Tabela 4-2. Esta perturbação foi realizada aumentando-se a quantidade de tecidos drenos e reduzindo a de tecidos fontes na planta de cana-de-açúcar. O plantio deste material em campo contou com uma parcela para cada genótipo, na configuração de 6 sulcos x 10 metros por parcela.

Tabela 4-2. Quantificação de sólidos solúveis totais (brix) nas plantas do experimento. Medidas realizadas em 26/05/2011 quando as plantas se apresentavam com 7 meses.

Amostra / Genótipo	Quinto internó		Terceiro internó	
	SP83-2847	SP91-1049	SP83-2847	SP91-1049
Amostra 1	14,6	21	16,8	20
Amostra 2	14,8	21	17	21
Amostra 3	15,4	20,4	17,4	20,8
Amostra 4	16,4	20,8	17,8	20,6
Amostra 5	17,4	21	16,8	20,8
Amostra 6	17,6	20	16	20,8
Amostra 7	16,8	21	15,4	19,2
Amostra 8	17,2	21,6	16	19
Amostra 9	17,4	22	17,2	21,8
Amostra 10	16,2	22	17,8	21,6
Média	16,4	21,0	16,7	20,4
Desvio padrão	1,1	0,6	0,8	0,9
Coeficiente de variação (%)	6,4	2,9	4,6	4,3

Uma vez que as plantas acima atingiram a maturidade no campo foi iniciado o experimento de manipulação fonte-dreno, utilizando-se de mecanismos de sombreamento de folhas. Este experimento foi conduzido em junho de 2011, e os tratamentos (14, 6, 3, 1 e 0 dias) foram conduzidos como indicado na Figura 4-5. Note que no tratamento com 6 dias de sombra foram coletados materiais em três períodos ao longo do dia (7h, 12h e 17h). Após o término do experimento, a folha +3 de cada planta foi coletada, congelada e teve seu mRNA extraído para análise do perfil de expressão gênica utilizando-se o sequenciamento de RNA em larga escala.

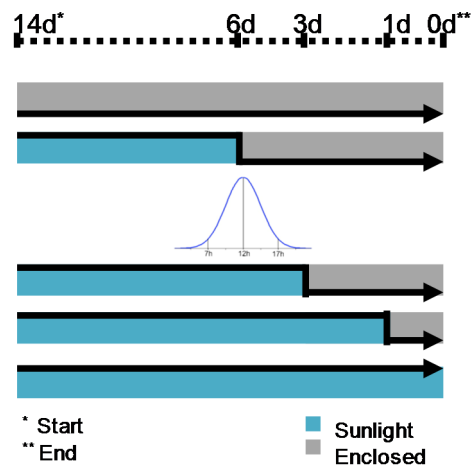


Figura 4-5 - Desenho do experimento para manipulação da relação fonte-dreno. Em azul é indicado o período de incidência de luz, sendo que em cinza esta o período em que as plantas estão cobertas (com exceção da folha +3). Todos os tratamentos serão coletados ao final do experimento.

4.8. Montagem e análise inicial dos dados de RNA-seq

A partir das amostras da folha +3 do experimento descrito acima, foram extraídos os RNAs que então foram sequenciados utilizando-se a plataforma Illumina HiSeq. As bibliotecas produziram altíssima profundidade de *reads* (Tabela 4-3) e com ótima qualidade (Figura 4-6). A contaminação por sequências bacterianas foi de cerca de 0.1% e a de rRNA foi de aproximadamente 5%.

Tabela 4-3. Resultado do sequenciamento do RNA-seq das 12 amostras de RNA de cana-de-açúcar no experimento de manipulação da relação fonte-dreno. A quantidade e a qualidade dos *reads* está dentro do esperado pela tecnologia Illumina.

Biblioteca	Número de reads antes da remoção do adaptador	Número de reads após a remoção do adaptador	% de reads sem adaptador
GXQ-1	74.252.988	71.100.312	95.75
GXQ-2	52.500.726	50.452.725	96.10
GXQ-3	41.670.747	38.500.137	92.39
GXQ-4	55.264.264	52.285.357	94.61
GXQ-5	61.038.447	58.935.357	96.55
GXQ-6	46.122.987	43.658.690	94.66
GXQ-7	101.168.376	96.952.938	95.83
GXQ-8	40.576.375	38.585.366	95.09
GXQ-9	44.173.762	42.087.869	95.28
GXQ-10	48.655.973	46.600.795	95.78
GXQ-11	63.869.521	60.780.973	95.16
GXQ-12	62.206.724	60.384.906	97.07



Figura 4-6 - Análise de qualidade dos *reads* sequenciados após trimagem inicial. Nota-se a elevada qualidade dos *reads* a serem utilizados na montagem do transcriptoma de cana-de-açúcar.

Os *reads* foram montados utilizando os softwares Velvet e Oases, com diferentes tamanhos de *hash* (Figura 4-7). Desta forma, esperava-se obter um balanço entre a profundidade e tamanho dos *contigs*, assim como a manutenção da variabilidade (SNPs, por exemplo). Ao final do processo os *contigs* foram remontados utilizando também as sequências Sanger existentes nos

bancos de dados. Desta forma espera-se que este transcriptoma sirva como referência não apenas para a análise de expressão dos genes do experimento de manipulação da relação fonte-dreno, mas também como referência para outros estudos com cana-de-açúcar.

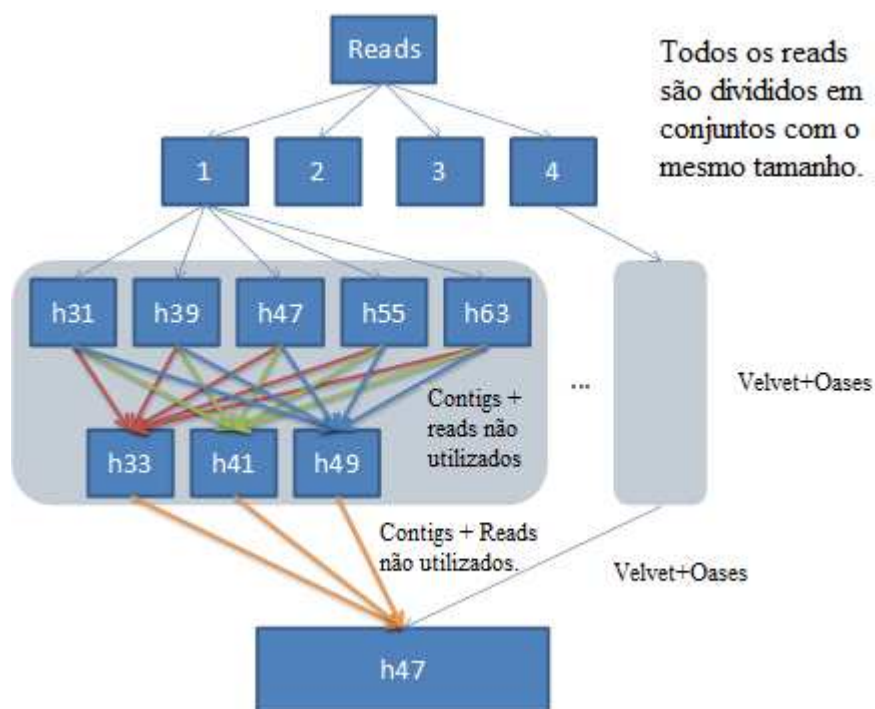


Figura 4-7 - Diagrama esquemático da montagem do transcriptoma. Representação da estratégia utilizada para a montagem de novo do transcriptoma de cana-de-açúcar.

4.9. *Sugarcane Systems Biology Workplace*

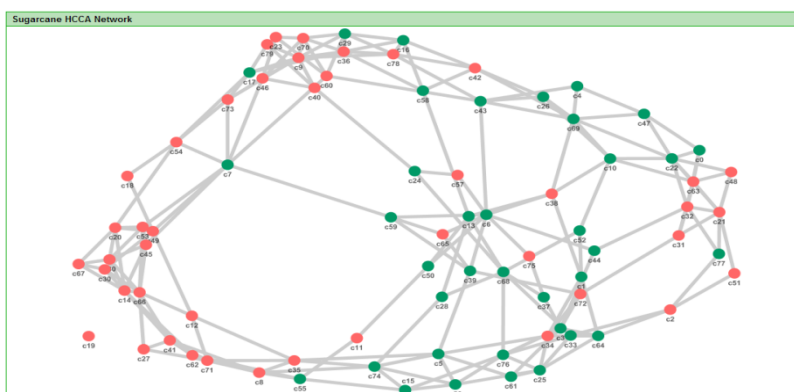
Há uma busca crescente por novos modelos e a formulação de novas hipóteses que sejam capazes de explicar os fenômenos observados. Desta forma, está sendo criada uma base de dados que atualmente armazena dados de anotação, organização genômica, ortologia e expressão gênica, todos eles integrados com diferentes redes de regulação gênica e redes metabólicas.

A interface de acesso ao banco de dados já foi desenvolvida por Vicentini (dados não publicados), e permite a navegação entre diferentes categorias funcionais, além do acesso a rede de co-expressão em larga escala que inclui dados de experimentos de relação fonte-dreno e possui mais de 8.000 genes de cana-de-açúcar (Figura 4-8).

A)



B)



C)

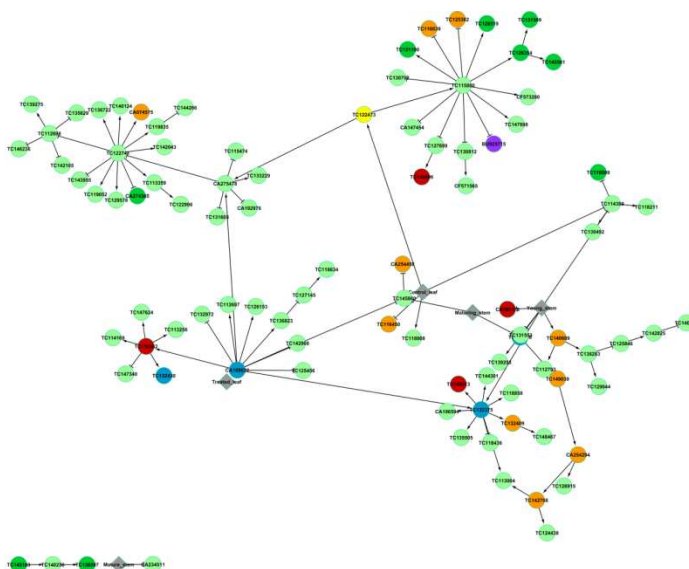


Figura 4-8 - Interface principal para acesso ao sistema Sugarcane Systems Biology Workplace. (A), interface gráfica do site (B) rede da divisão de todos genes do *microarray* de cana-de-açúcar em clusters (C) rede bayesiana dos genes diferencialmente expressos no experimento de relação fonte-dreno.

Esse banco de dados incluirá as redes bayesianas e análises desenvolvidas durante esse trabalho de mestrado. Desta forma, espera-se a conclusão e liberação do banco de dados para acesso por toda a comunidade científica de cana-de-açúcar.

4.10. Construção das redes bayesianas

Para construção das redes, inicialmente, tentou-se utilizar os pacotes do R: *deal* e *BNArray*. O *BNArray* funciona até a versão 2.9.2 do R em Windows. Porém, exigiu muito esforço computacional e tempo sendo viável sua utilização apenas para um número restrito de nós (até aproximadamente 10 nós) e mesmo com a paralelização pelo método dos tripletos, esse número de genes ainda permanece muito baixo em relação ao número que gostaríamos que estivessem nas redes. Houve tentativas de se utilizar o *BaNJO* e o *PEBL* também, porém como o número de experimentos em cana eram somente dois e com poucas repetições esses programas não mostraram boa precisão e sensibilidade na construção das redes.

Dessa forma, o programa que teve melhor desempenho e se adequou melhor as necessidades desse trabalho, considerando o esforço computacional exigido, tempo de execução, topologias de redes obtidas no que se refere à questão de sensibilidade e precisão, foi o *BNFinder*. As redes foram calculadas utilizando-se o *BDe score*, com todos os genes podendo ser reguladores de todos os outros e o parâmetro $-l$ que define o número máximo de conjuntos de pais com melhor *score* que será retornado para cada nó, foi definido como 5. A rede dos 265 genes diferencialmente expressos foi obtida com o *MDL score* devido ao número de genes.

4.11. Visualização das redes

A comparação das redes foi feita utilizando-se o software Cytoscape 2.8.2 que é o ambiente padrão para análise e visualização de dados em Biologia de Sistemas, e tem a capacidade de integrar redes gênicas a partir de dados de expressão (Shannon *et al.*, 2003). Ele permite integrar interações moleculares e medidas de diferentes estados em uma estrutura comum e assim conectar os dados com um amplo conjunto de parâmetros e outros atributos biológicos (Shannon *et al.*, 2003).

4.12. Análise de ontologia e anotação dos genes

O *Gene Ontology* (GO) Consortium (Ashburner *et al.*, 2000) é um projeto que visa a necessidade de descrições consistentes de produtos gênicos em diferentes bases de dados para

isso ele classifica os genes de acordo com categorias aplicáveis a qualquer organismo. Ele apresenta três categorias básicas: processo biológico que se refere ao objetivo para qual o gene ou produto gênico contribui; função molecular que se refere à atividade bioquímica de um produto gênico; e componente celular que indica o local na célula onde o produto gênico é ativo.

A anotação de genes em categorias facilita a interpretação das informações funcionais e pode ser feita automaticamente através de vários softwares, sendo o mais utilizado o Blast2Go (Conesa *et al.*, 2005). Para usá-lo é necessário um arquivo contendo as sequências no formato fasta e então é executado um BLAST para encontrar genes homólogos a essas sequências. Depois de feito isso, é executado um mapeamento para recuperar os termos GO. As sequências utilizadas neste trabalho foram retiradas de *The Gene Index Project* mantido pelo *Computational Biology and Functional Genomics Laboratory* de Harvard⁴ e a partir disso foi obtida a identificação das funções e categorias de GO para os genes.

4.13. Análise do padrão de expressão das redes

O *ViaComplex* é um software desenvolvido por Castro *et al.* (2009) e é uma ferramenta generalizada para avaliar a expressão de genes. Ele possui interface gráfica e permite o cálculo tanto a partir de dados de expressão como para um modelo de interesse de redes de regulação.

Sua principal vantagem é que ele disponibiliza a visualização dos dados como paisagens funcionais projetadas em mapas de redes gênicas através da distribuição dos dados nas redes, assim ele sobrepõe a informação funcional com a informação de interação dos genes. Além disso, ele possui recursos que permitem a comparação estatística entre conjuntos de dados diferentes.

Essa análise foi obtida para a rede do conjunto de 136 genes diferencialmente expressos e para todos os genes presentes no *microarray* de cana-de-açúcar.

⁴ http://compbio.dfci.harvard.edu/cgi-bin/tgi/est_ann.pl?gudb=s_officinarum

5. RESULTADOS E DISCUSSÃO

5.1. Redes de regulação gênica

A Figura 5-1, mostra a rede obtida pelo método de correlação para todos os genes presentes no *microarray* de cana-de-açúcar da *Affymetrix*. Um dos objetivos do trabalho era obter a rede utilizando-se a metodologia de redes bayesianas para todos esses genes, porém devido ao grande esforço computacional exigido por essa estratégia só é possível trabalhar com conjuntos menores de genes e assim, a rede com todos os genes do *microarray* só pôde ser feita pela estratégia de correlação.

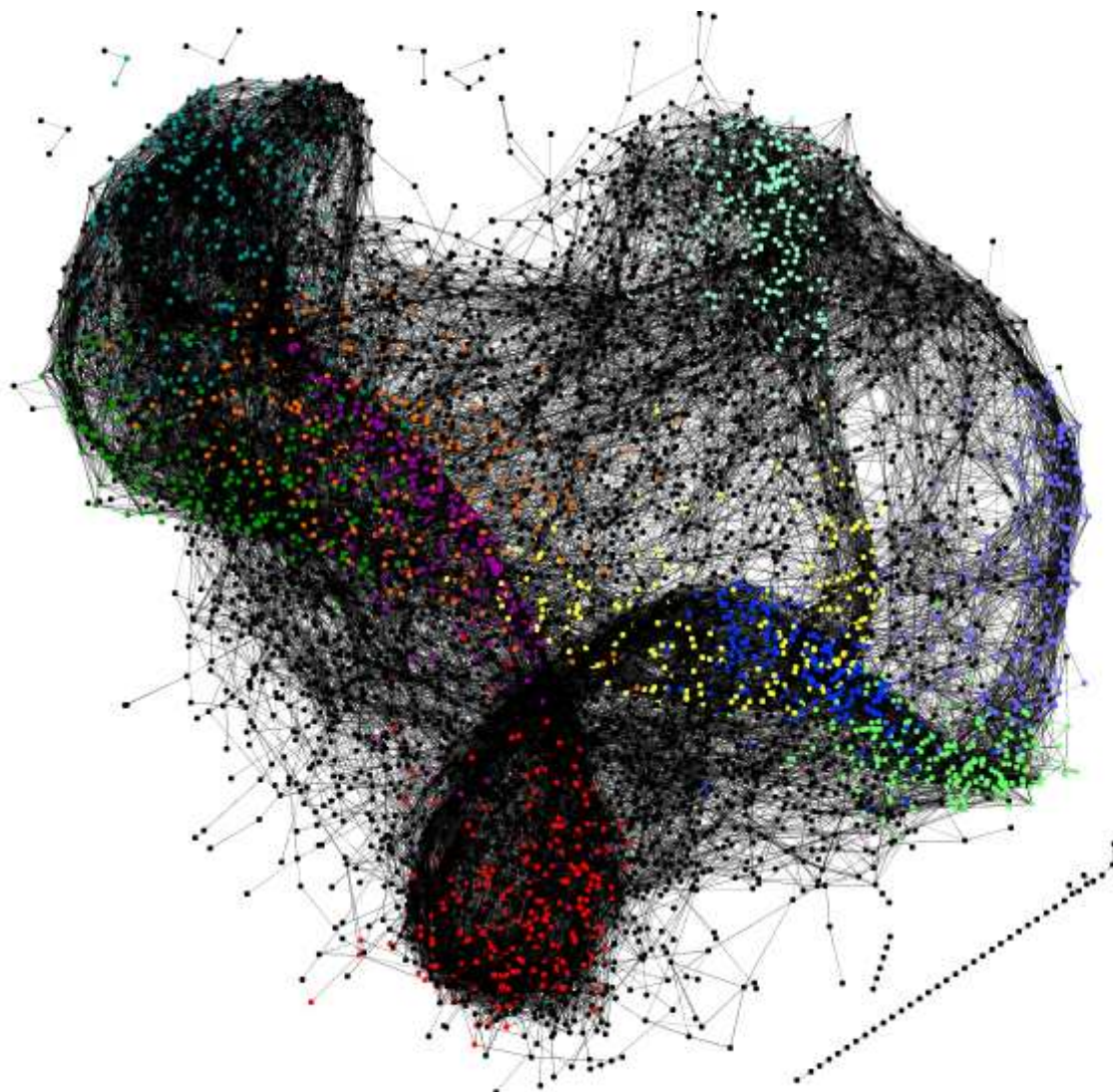


Figura 5-1 - Rede de correlação. Rede obtida pelo método de coexpressão para todas as sondas do *microarray* de cana da *Affymetrix*.

A partir dessa rede, foi possível fazer a divisão dos clusters através do método de HCCA. O HCCA agrupa os genes com níveis de expressão mais correlacionados. Para as 8236 sondas, foram encontrados 85 clusters interligados por 381 arcos (Figura 5-2). Vale a pena frisar que nem sempre o fato de dois genes terem níveis de expressão correlacionados implica que os níveis de expressão sejam realmente dependentes um do outro. Por exemplo, à medida que a planta avança em seu estágio de crescimento, os genes relacionados à produção de lignina aumentam seus níveis de expressão para a formação do colmo, e como a planta está passando para o estágio de maturação seus níveis de produção de sacarose também aumentam devido à expressão de genes relacionados à produção de sacarose, porém a regulação desses dois grupos de genes não necessariamente está relacionada.

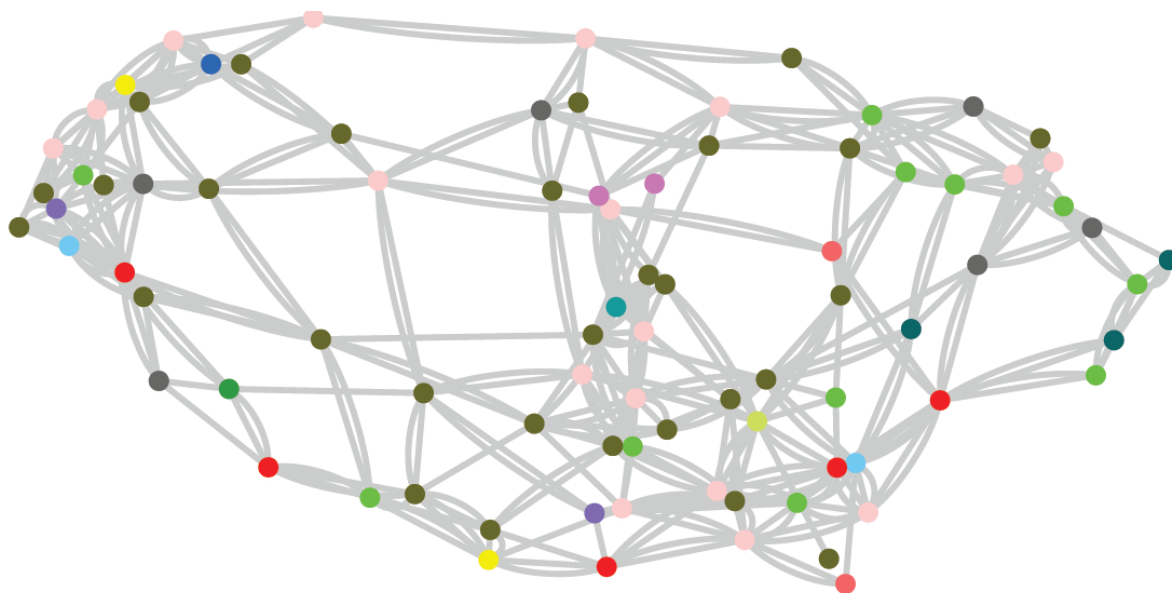


Figura 5-2 - Clusters da rede de correlação. Rede representando os clusters obtidos pelo método de HCCA a partir da rede de correlação para todas as sondas do *microarray* de cana-de-açúcar da *Affymetrix*.

A rede bayesiana obtida para os genes do cluster ao qual pertence à enzima SPS em cana está representada na Figura 5-3.

Essa rede é composta por 109 nós, sendo que dois deles são discretos, correspondentes aos tratamentos do experimento do McCormick *et al.* (2008), sendo eles folhas controle (*Control Leaf*) e folhas expostas ao frio (*Treated Leaf*). Dos 107 nós contínuos, que representam os genes, 27 são transportadores, 27 estão relacionados a DNA, tradução de RNA e ligação, 11 ao metabolismo de aminoácido e proteínas, 2 a metabolismo de lipídios, 3 a metabolismo de parede celular, 2 ao metabolismo de citoesqueleto, 3 a resposta a estresses, 4 a metabolismo de carboidratos, 1 a resposta a hormônios, 2 de metabolismo mitocondrial, apenas 1 de fotossíntese, e 1 é da classe de outros eventos metabólicos (Tabela 5-1). Além disso, 22 dos genes presentes na rede possuem função não identificada.

Tabela 5-1. Classificação dos genes de acordo com sua função.

Categoria	Número de Genes	Porcentagem
Metabolismo de Aminoácidos e Proteínas	11	10.28%
Tradução, ligação ao DNA	27	25.23%
Metabolismo de Parede Celular	3	2.80%
Metabolismo de Lipídios	2	1.87%
Resposta a Estresses	3	2.80%
Metabolismo de Carboidratos	5	4.67%
Metabolismo de Citoesqueleto	2	1.87%
Resposta a Hormônios	1	0.93%
Metabolismo Mitocondrial	2	1.87%
Fotossíntese	1	0.93%
Transportadores	27	25.23%
Outros Eventos Metabólicos	1	0.93%
Função Não Identificada	22	20.56%
Total	107	100%

O nó BU925724 é uma enzima que apresenta atividade de sacarose fosfato sintase. Na rede, este nó apresenta ligação direta com sete outros genes formando um *hub* (Figura 5-4-a). A expressão desse gene está sendo induzida por um único gene, o CF573514, que codifica uma proteína mitocondrial de transporte transmembrana. Ele também se relaciona com outros dois transportadores, o CA154088 que é localizado na membrana e serve para ligação de nucleotídeos e o CA093554 que possui atividade catalítica e se localiza nas membranas vacuolar e plasmática.

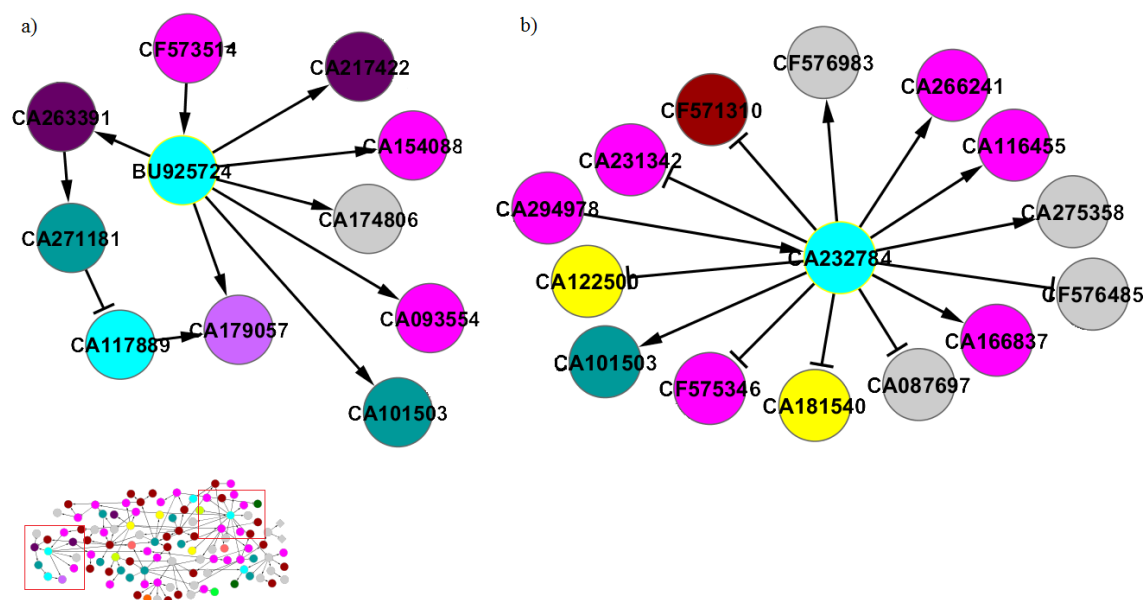


Figura 5-4 - Genes da enzima SPS representados na rede. A) Gene BU925724 e os genes com os quais possui ligação direta. B) Gene CA232784 e os com os quais possui ligação direta.

Dois genes relacionados a respostas a estresses são induzidos pelo gene BU925724, sendo eles o CA217422 que é uma protease associada à isoforma1 de uma proteína com domínio da família *zinc finger* e a CA263391 que possui atividade de *glutathione transferase* e regula resposta à luz vermelha localizada tanto no núcleo quanto membrana e citoplasma. Outro gene induzido pelo BU925724 é o CA10153, que está localizado na membrana, citoplasma, mitocôndria e vacúolo e tem função de processar proteínas, proteólise além de regular sinais de transdução e ter também atividade peptidase. CA179057, também induzido pelo BU925724, se encontra no cloroplasto e responde ao ácido abscísico e possui também função na transdução de sinais e morfogênese de tricomas além do crescimento celular. O gene CA174806 não possui função definida ainda.

O gene CA232784 (Figura 5-4-b) codifica a enzima SPS apresenta relação direta com 14 genes, sendo que seis dessas relações são repressivas. O gene CA294978 localizado em vesículas ligadas a membrana citoplasmática e possui papel de ligação de lipídios e cálcio é o único que induz diretamente a expressão da SPS. Dois dos genes que possuem expressão induzida pela SPS não possuem função identificada. O CA232784 está inibindo a expressão de dois genes relacionados ao metabolismo de parede, o CA181540 que possui função de molécula estrutural e o CA122500 que codifica uma fosfolipase d que possui atividade catalítica de hidrolase, participa do catabolismo de lipídios e ligação de íons

cálcio e está localizada na membrana. Ele também inibe a expressão do CF571310 que codifica um constituinte estrutural do ribossomo e participa da ligação de íons metal.

O gene CA232784 também atua induzindo a expressão de alguns genes, como o CA101503 que integra membranas vacuolares e plasmáticas e também está presente no retículo endoplasmático e mitocôndria participando do processamento de proteínas e regulação da transdução de sinais. Também induz transportadores como CA266241 localizado na membrana plasmática, o CA116455 que participa da divisão celular e processos catabólicos de GTP, tendo, portanto atividade de GTPase, e o CA166837 que possui atividade de *glutathione transferase* no citoplasma.

Nessa rede é possível notar que um aumento na expressão da SPS, causa repressão de CF575643 que é responsável pela indução de CF571746 e CA120893 que são expressos na folha controle, mas não na tratada. Assim, pode-se concluir que um aumento na produção de sacarose ocorre na folha tratada, assim como observado pela análise dos dados de expressão por McCormick *et al.* (2008).

Essa rede do cluster apresenta oito genes que possuem relação direta com um número grande de outros genes (sete ou mais) (Figura 5-5). Dentre eles, o CA186500 que participa do processo de biossíntese de celulose; o CA114531 que é uma proteína de transporte intracelular; o CA248525 que participa da resposta ao estímulo por auxina, ubiquitinação de proteínas, morfogênese da inflorescência, desenvolvimento de raiz e ligação de íons zinco; CA230194 que é uma proteína de transporte intracelular; CA067632 que é um repressor da atividade de tradução e transcrição; CF576485 que não possui função definida e CA167795 que possui atividade de proteína quinase e ligação de ATP.

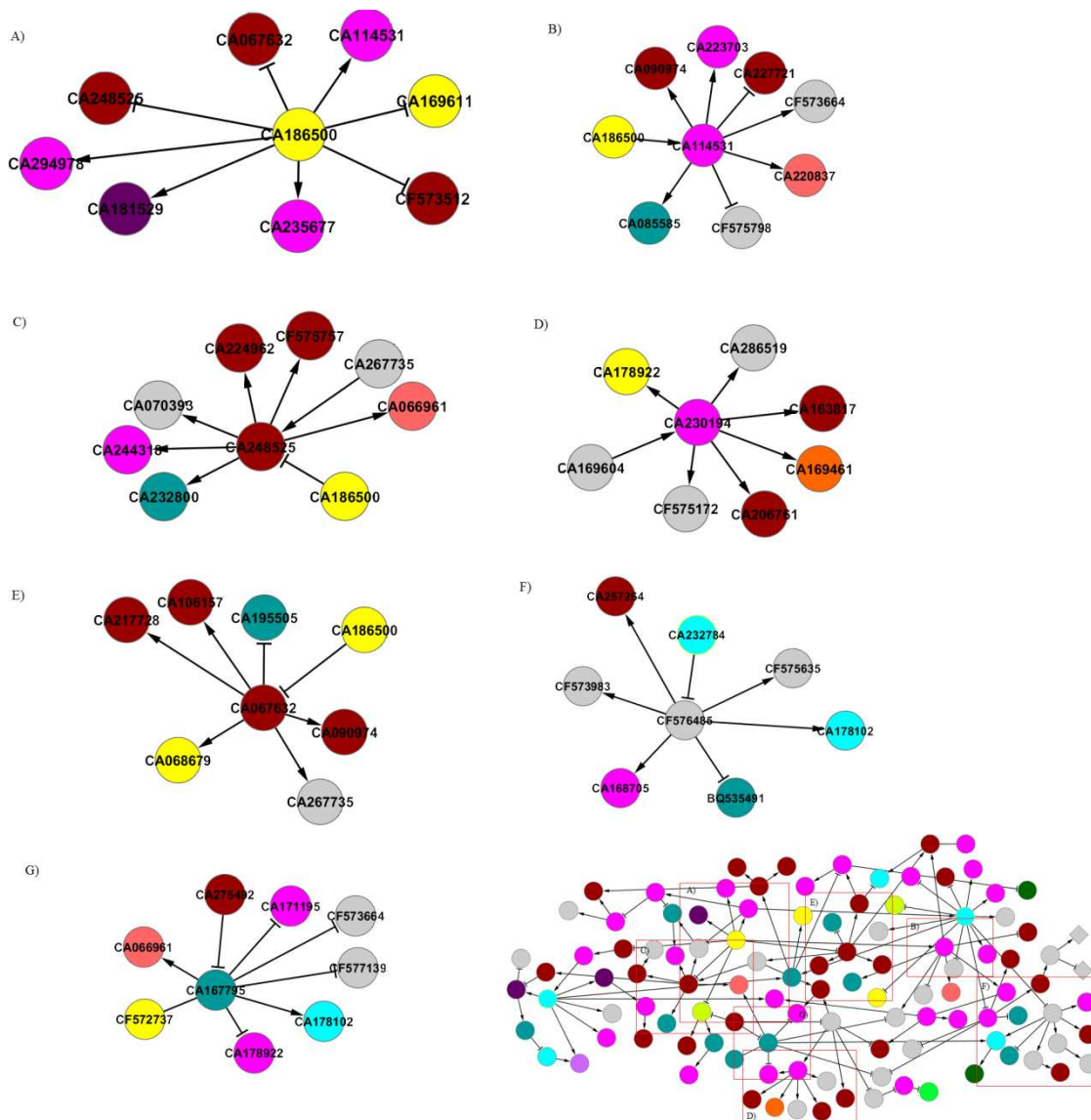


Figura 5-5 - Hubs presentes na rede do cluster de cana. Os sete hubs encontrados na rede do cluster da SPS em cana-de-açúcar.

A rede para os genes selecionados dentro do cluster da SPS (Figura 5-6) ficou com 32 nós, sendo sete deles discretos relativos aos tipos de tratamento dos trabalhos de Casu *et al.* (2007) e McCormick *et al.* (2008) e 25 nós contínuos referentes aos genes. O número de arcos interligando esses nós foi 31.

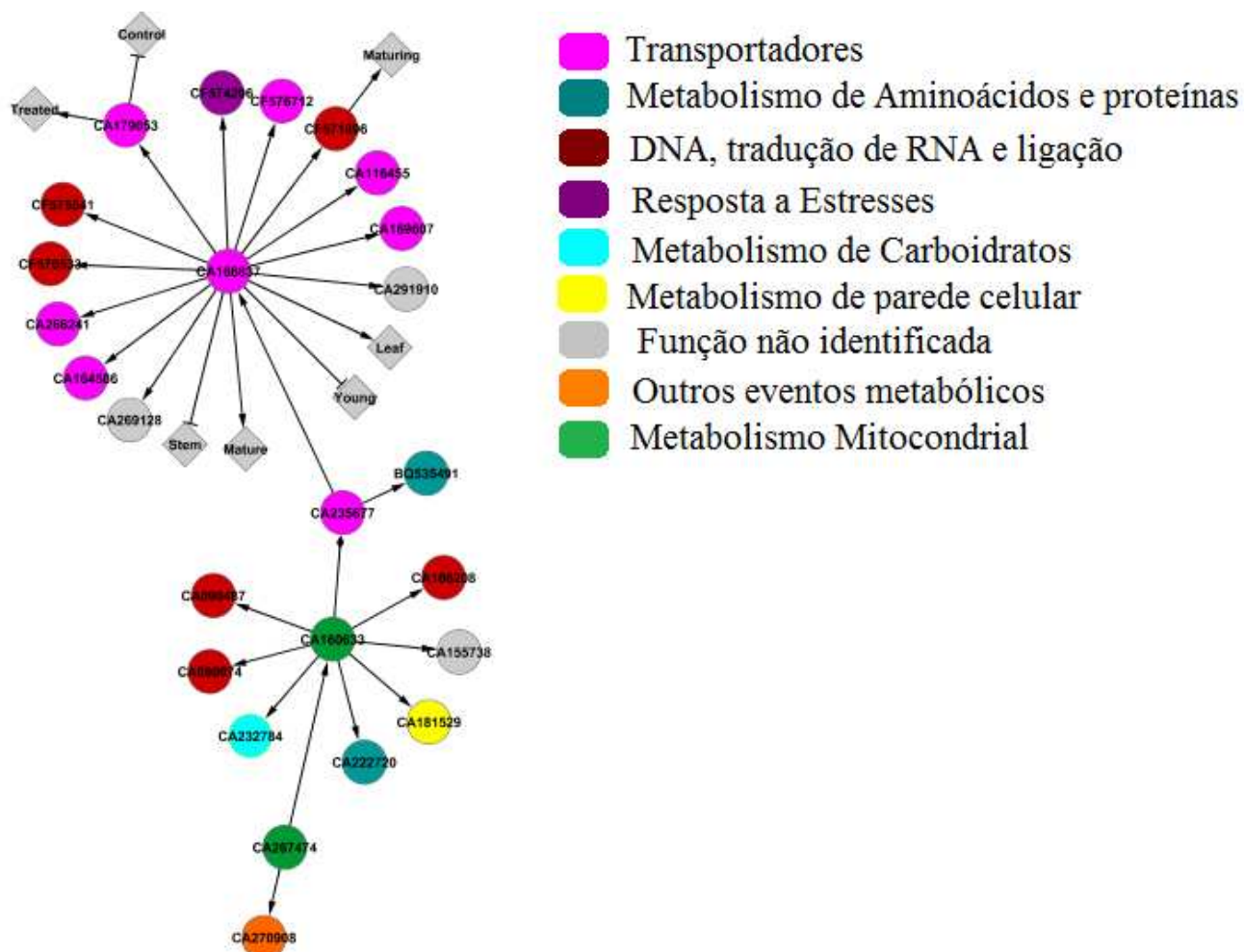


Figura 5-6 - Rede dos genes selecionados dentro do cluster da SPS. Rede dos genes com HRR = 30 e NVN = 1 e HRR = 10 e NVN = 20, totalizando 32 nós sendo 7 discretos e 31 arcos.

Conforme a Tabela 5-2, a maior parte dos genes representados nessa rede são transportadores (32%), seguido dos genes relacionados a Tradução, ligação ao DNA (24%). Outras categorias apresentaram representação similar (8%) como Resposta a estresse, Metabolismo mitocondrial e Metabolismo de aminoácidos, enquanto Metabolismo de carboidratos só representa 4% dos genes presentes na rede.

Tabela 5-2. Classificação dos genes selecionados a partir do cluster conforme sua função.

Categoria	Número de Genes	Porcentagem
Metabolismo de Aminoácidos e Proteínas	2	8.00%
Tradução, ligação ao DNA	6	24.00%
Resposta a Estresses	2	8.00%
Metabolismo de Carboidratos	1	4.00%
Metabolismo Mitocondrial	2	8.00%
Transportadores	8	32.00%
Outros Eventos Metabólicos	1	4.00%
Função Não Identificada	3	12.00%
Total	25	100%

Essa rede apresenta dois *hubs* principais. Um tem como nó central o gene CA160833 que é similar a uma proteína de função desconhecida em *Arabidopsis* (AT5G06970), mas faz parte do metabolismo mitocondrial. O outro tem como nó central um transportador, o CA166837 que possui atividade de *glutathione transferase*.

As *glutathione transferases* ainda possuem papel mal compreendido em plantas. Elas possuem função de desintoxicação de metabólitos de estresse oxidativo, biossíntese de leucotrienos e parecem estar envolvidas na ligação intracelular e estabilização de flavonoides, além disso, parecem proteger as plantas de danos oxidativos funcionando como peroxidases e podem também funcionar na tolerância a estresses (Dixon *et al.*, 2002).

Dos genes ligados ao primeiro *hub* (CA160633), três estão relacionados a Tradução, ligação ao DNA. O CA090974 que regula *splicing* de RNA, tradução e tem papel na ligação de RNAs sendo localizado tanto no núcleo quanto no corpo processador de mRNA citoplasmático; o CA090487 que possui atividade de tRNA-aminoacilase; e o CA106208 que está localizado no citoplasma e possui função na terminação traducional. Fazendo parte do *hub* encontra-se também um gene de resposta a estresse, o CA181529 que codifica uma proteína de reparo de DNA rad23, que também responde ao frio localizada no núcleo. Por último, há o CA222720 que possui localização intracelular e tem atividade de aminoácido ligase. CA155738 não possui função definida.

No segundo *hub* também se ligaram alguns genes relacionados a Tradução, ligação ao DNA. O CF575541 que possui atividade de nucleosídeo-trifosfatase; o CF570533 que codifica uma proteína constituinte estrutural do ribossomo; e o CF571896 que tem papel na

ligação de DNA. Porém a maioria dos genes desse *hub* estão relacionados a transportadores, sendo 6. CA116455 que possui atividade de GTPase; CA164586 que é uma proteína de fusão vacuolar; CA169607 que é um transportador transmembrana de de carboidratos e açúcares; CA179053 que possui atividade ceramidase; CA266241 localizado na membrana plasmática; CF576712 que media o transporte de vesículas; e por fim, o CA235677 que possui função no transporte intracelular de proteínas. Além desses, há um gene de resposta a estresse, o CF574296 localizado na membrana e na mitocôndria e participa da ligação de íons zinco.

O gene da SPS, CA232784, está sendo induzido diretamente pelo nó de um dos *hubs*, CA160833, que também induz um gene que faz a ligação dos dois *hubs* (CA235677). O nó central do outro *hub*, não é expresso nem no colmo e nem na planta no estado jovem. Sua expressão está bastante relacionada com a expressão do gene da SPS, pois, eles dependem do mesmo gene para ter sua expressão induzida. E é possível perceber que a expressão desse gene tem efeito positivo nos estados Maduro e Maturando, indicando que a expressão da SPS também aumenta nesses estados, isso pode ser observado no desenvolvimento da planta de cana, no qual, a planta em estágio de maturação começa a acumular açúcar até alcançar a maturação máxima.

Um ponto interessante nessa rede, é que um aumento de CA160633 resulta num aumento da expressão da SPS e CA235677 que por sua vez induz CA166837, que aumenta a expressão de CA179053 que é expresso na folha tratada, mas não na folha controle, ou seja, o aumento da expressão dos genes que induzem a SPS e logo aumentam a produção de açúcar induz também a expressão de gene um que não é expresso na folha de tratamento controle. E ainda, CA166837 é expresso na folha, mas não no colmo. No trabalho de McCormick *et al.* (2008) é observado um aumento no acúmulo de sacarose de folhas expostas ao frios (tratadas) assim como representado na rede.

E como a expressão do mesmo gene que induz a SPS, induz também o CA166837, mesmo que indiretamente, pode-se observar que o CA166837 não é expresso na planta jovem, pois a planta jovem dispõe a energia produzida para seu crescimento e não para acúmulo de sacarose.

O nó que faz a ligação entre os dois *hubs*, é o CA235677 que codifica uma proteína de transporte intracelular localizada na membrana e cloroplasto. Essa proteína provavelmente é responsável por levar o sinal de um *hub* até o outro, permitindo que eles se comuniquem.

A Figura 5-7 representa a rede bayesiana obtida para o grupo de genes definidos por McCormick *et al.* (2008).

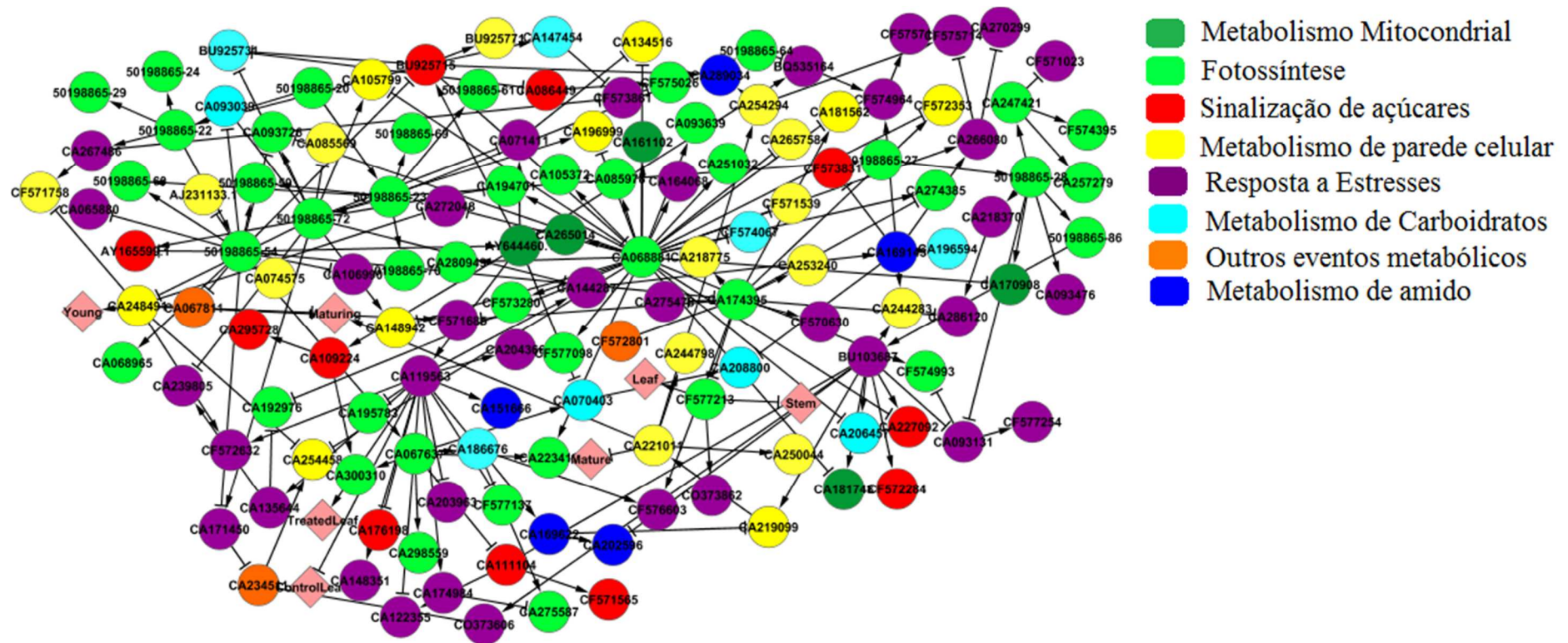


Figura 5-7 - Rede bayesiana para os 136 genes definidos no trabalho de McCormick (2007). Rede de regulação 136 genes e sete tratamentos, 195 arcos. Cores indicam funções dos genes.

Essa rede possui 143 nós sendo sete deles discretos referentes aos tratamentos, (Folhas expostas ao frio, folhas controle, colmo jovem, colmo maturando e colmo maduro) 136 nós contínuos referentes aos genes e um total de 195 arcos.

A categoria funcional com maior representação dentre os genes da rede foi Fotossíntese, a qual pertence 31.61% dos genes, em seguida vem Resposta a estresses com 27.2% e Metabolismo de Parede Celular com 16.91% (Tabela 5-3). As outras categorias representadas apresentaram porcentagem bem menor como Metabolismo de carboidratos englobando 6.61% dos genes, 8.82% dos genes referentes à Sinalização de açúcar, e 3.67% referentes a metabolismo de amido e a mesma porcentagem para Metabolismo Mitocondrial. A categoria menos representada foi Outros eventos metabólicos com apenas 2.21% dos genes totais da rede.

Tabela 5-3. Classificação dos 136 genes da rede de acordo com sua categoria funcional.

Categoria	Número de Genes	Porcentagem
Metabolismo de Parede Celular	23	16.91%
Resposta a Estresses	37	27.20%
Metabolismo de Carboidratos	9	6.61%
Sinalização de açúcares	12	8.82%
Metabolismo Mitocondrial	5	3.67%
Fotossíntese	43	31.61%
Metabolismo de Amido	5	3.67%
Outros Eventos Metabólicos	3	2.21%
Total	136	100%

A grande presença de genes de resposta a estresses e fotossíntese é devida ao tipo de experimento que foi utilizado para selecionar esse grupo de genes. A composição gênica dessa rede é bem diferente quando comparada as redes anteriores. As redes com genes selecionados pelo método de HCCA eram ricas em transportadores e genes relacionados a Tradução, ligação ao DNA enquanto apenas um número restrito de genes tinha função nos diversos tipos de metabolismo. Já nesse grupo de genes diferencialmente expressos, há um número elevado de genes de fotossíntese, que o grupo de HCCA não possui além de um número maior de genes de metabolismo de parede celular e carboidratos, e resposta a estresses. E nesse grupo, é possível observar genes de sinalização de açúcares, que não haviam no cluster. O método de HCCA selecionou genes relacionados a funções mais basais da célula, o que é interessante,

pois um gene pode estar diretamente relacionado com um gene de interesse, mas não necessariamente ser diferencialmente expresso.

Essa rede possui um *hub* principal, que aparenta ter um papel muito importante para seu funcionamento (Figura 5-8). Esse *hub* está interagindo com 33 genes. Ele é o gene CA068881 que codifica uma proteína da superfamília do citocromo p450 e sendo um citocromo faz parte de processos de oxirredução e possui atividade de carregador de elétrons. Sua expressão está relacionada com a expressão de vários outros genes de fotossíntese e metabolismo mitocondrial. Dos genes de fotossíntese, ele regula negativamente o CA192976 que é uma enzima málica dependente de NADP, que possui função de malato desidrogenase e se liga a NADP e metais. Outro gene de fotossíntese regulado negativamente é o CA194701 que é uma fosfoglicerato quinase com função no ciclo redutivo da pentose-fosfato, glicólise e fosforilação e também responde a íons cádmio e liga ao ATP, localizada nos tilacóides, estroma do cloroplasto e mitocôndria.

Todos os outros genes de fotossíntese são regulados positivamente. Dentre eles: CA194701 e CA085978 que também são fosfoglicerato quinases com as mesmas função da anterior, porém além daquelas funções, responde também ao frio e está localizada em mais lugares como núcleo, apoplasto, parede celular, envelope do cloroplasto; CA105372 e CA223413 são geranilredutases que fazem parte do processo de oxi-redução sendo acceptor de NAD ou NADP e doador de grupos CH, processa tRNA, participa da biossíntese de clorofila, vitamina E, e phytyl difosfato e se localizam nos plastídeos; CA093039 é uma gliceraldeído-3-fosfato desidrogenase que faz parte do metabolismo de glicose. CA274385 e CF574993 codificam proteínas do fotossistema I; CA192976 é uma enzima málica dependente assim como já foi descrito acima; CA280949 e CA251032 localizadas na membrana do tilacóide e faz parte da estabilização do fotossistema II; CF577098 que participa do processo de biossíntese de clorofila e possui atividade de nucleosídeo fosfatase.

CA068881 inibe todos os genes relacionados a metabolismo de parede celular e carboidratos, inclusive uma frutose bisfosfatasealdolase (CA206457 e CF574067) e CA272048 que é uma proteína quinase com função de transdução de sinal. Ele é inibido por um sinalizador de açúcares, BU925731 que é uma glicosil hidrolase que faz parte do metabolismo de carboidratos e atua como uma sacarose alfa-glucosidase.

CA068881 induz dois genes de resposta a estresse CA164068 que é uma proteína quinase e CA266080 que responde a estresse oxidativo, possui atividade peroxidase e faz parte do processo de oxi-redução.

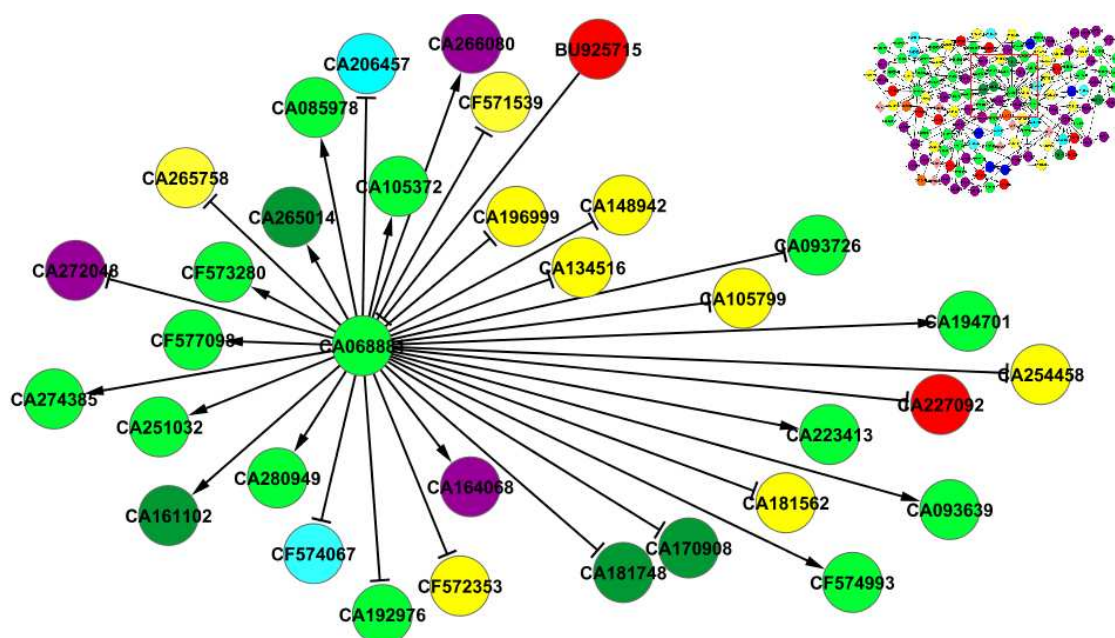


Figura 5-8 - Hub CA068881 interagindo com 33 genes. CA068881 é um gene que apresentou um elevado número de interações constituindo um *hub*.

O gene 50198865-54 também faz parte de um *hub* que é conectado diretamente com outro *hub* pelo gene CA174395 (Figura 5-9). CA174395 é um citocromo p450. 50198865-59 reprime três genes de resposta a estresse: CA171450 que possui atividade de glutamato descarboxilase; CA106990 que é uma proteína que responde a estresse; e CA065880 que é uma álcool desidrogenase. Além desses, ele também reprime genes de sinalização de açúcares, o BU925715 que é um transportador de açúcar e AY166699 que participa do transporte transmembrana de carboidratos sendo que todos os dois integram a membrana. Ele também inibe um gene de metabolismo de carboidratos, o CA093039 localizado no plastídio e que participa do metabolismo de glicose. Os genes de metabolismo de parede celular também são inibidos AJ231133 e CA074575 que atua como xiloglucano:xiloglucosil transferase.

Esse *hub* atua inibindo o outro, que tem como gene central o CA174395. O CA174395, ao contrário do gene anterior, induz a expressão da maioria dos genes de metabolismo de parede presente no *hub*: CA253240 que é um centro de reação do fotossistema I; CA181562

que possui atividade chitinase e atua no processo catabólico de macromoléculas de parede celular; CF572353 que participa da biossíntese de lignina e metilação; CA218775 que tem atividade de hidrolase; e inibe apenas um deles, o CA274385. Dos genes de resposta a estresse, apenas CF570630 é induzido. Os outros, CA286120 que faz parte do processo catabólico de clorofila e ligação de metais, CF571685 que responde a lesões e possui atividade peptidase; CA275478 são todos reprimidos. Há um gene de fotossíntese sendo reprimido também, o CF577213.

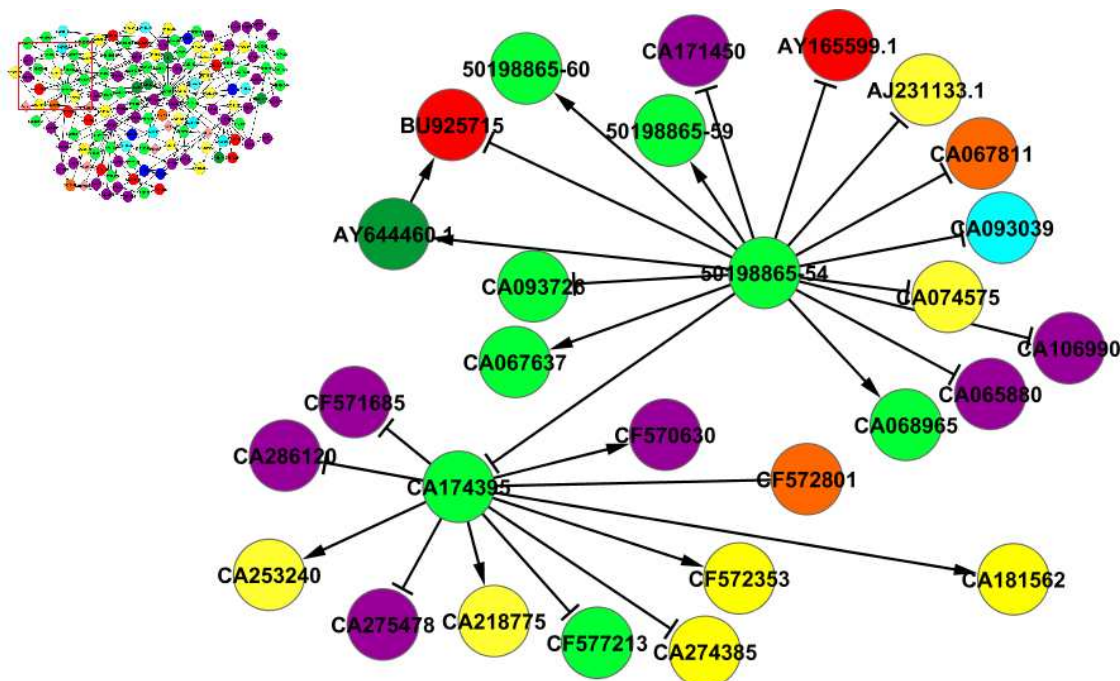


Figura 5-9 - Hubs CA174395 e 50198865-54. Genes que apresentaram um elevado número de interações constituindo dois *hubs* conectados.

Outro *hub* é o do gene CA119563 (Figura 5-10) que interage com 16 variáveis, sendo duas delas referentes a tratamentos e as outras 14 aos genes. CA119563 é um gene de resposta a estresse. Ele regula positivamente sete outros genes de resposta a estresse: CA204366 localizada no cloroplasto e na mitocôndria, responde ao estresse salino e possui atividades de aldeído desidrogenase; CA203963; CA148351 também possui atividade de aldeído desidrogenase; CA174984; CA144287 localizada no citoplasma e se liga a íons metal; CF572632 que é uma proteína induzida por baixas temperaturas; CA135644 que é uma proteína induzida por jasmonato e se liga a manose. Os genes de resposta a estresse que sofreram repressão são CA203963 e CA122355.

Quando o gene de resposta a estresse é expresso, os genes de fotossíntese do *hub* são reprimidos CF577137 com atividade de ribulose bisfosfato carboxilase e função na fotorrespiração e CA195783. Porém os genes de amido são induzidos CA151666 que faz parte do processo de biossíntese de glicogênio e amido e CA169622 que tem atividade fosforilase, ambos localizados nos plastídios. Um gene de sinalização de açúcar, o CA176198 também localizado nos plastídios é reprimido. Ele possui função de biossíntese de trealose. Os genes de estresse são expressos na folha tratada com baixas temperaturas mas não na folha controle, conforme o esperado.

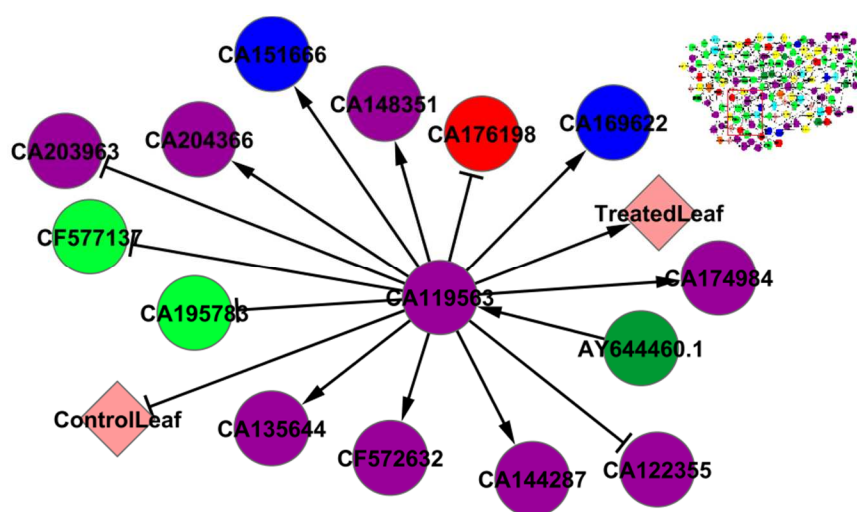


Figura 5-10 - Hub CA119563. Genes que interagem com CA119563.

Outra vez ocorre de dois *hubs* estarem conectados. A Figura 3-1 mostra o *hub* composto pelo gene BU103687 que é um gene de resposta a estresse, e atua na biossíntese de flavonoides e o outro é um *hub* menor com um nó central relacionado à fotossíntese, o 50198865-28.

BU103687 induz três genes de sinalização de açúcares: CA227092 que codifica uma proteína translocadora de fosfato; CA111104; e CF572284, que também possui atividade de transportador transmembrana. Além desses, ele induz também dois genes de metabolismo de parede celular: CA244283 e CA219099. Também são induzidos por ele um gene de metabolismo de amido, o CA206457 (frutose bisfosfatase aldolase) e um de metabolismo mitocondrial, o CA181748 que tem atividade chitinase e atua no processo catabólico de macromoléculas de parede celular.

BU103687 é induzido por um gene de fotossíntese e sua expressão reprime apenas dois genes, o CA202596 que participa do metabolismo de carboidratos e CA093131 o qual responde a níveis de nitrogênio. Para que BU103687 seja expresso, é necessária a expressão do nó central do outro *hub* ligado a ele, o 50198865-28. Esse nó reprime o mesmo gene de resposta a estresse que BU103687, porém induz outros dois: CA218370 (participa do processo catabólico de clorofila e se liga a íons metal como cobre e ferro) e CA093476 (responde a falta de água, sal, calor).

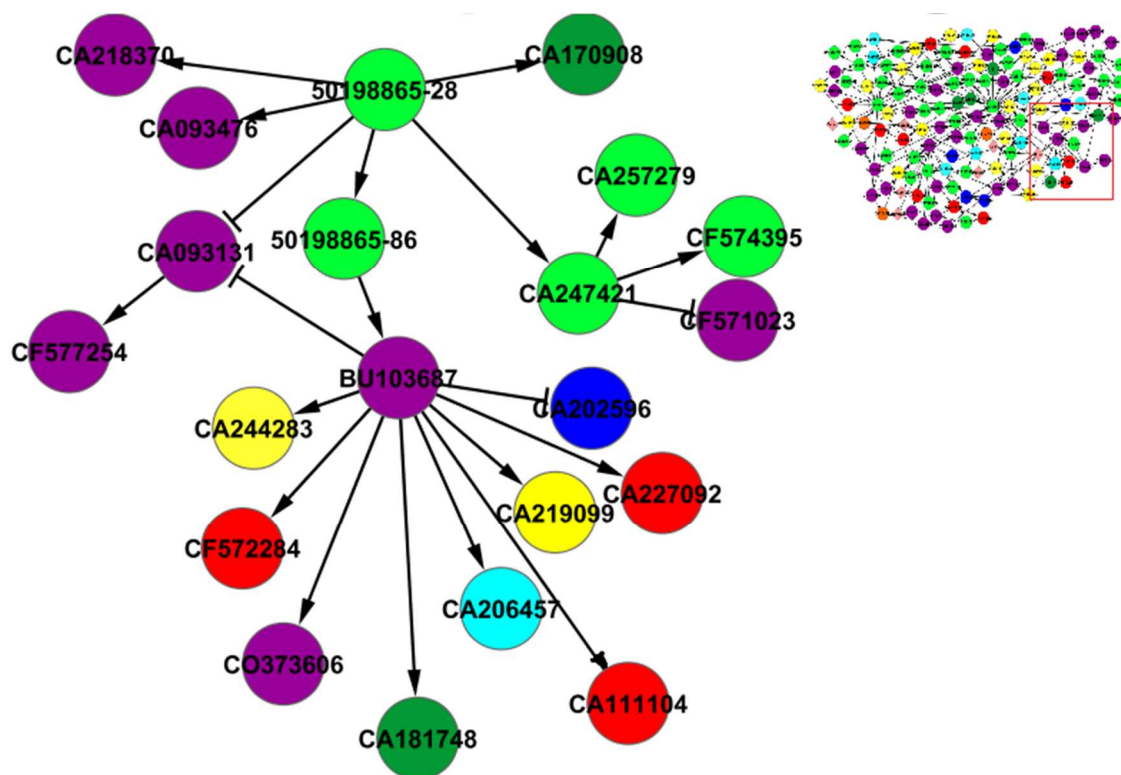


Figura 5-11 - Hub BU103687 e 50198865-28. Representação do gene BU103687 e suas interações e também do 50198865-28.

Outros genes que apresentaram um número considerável de interações foram CA248491 e CA221011 (Figura 5-12). Esses são genes relacionados ao metabolismo de parede celular. CA221011 é uma proteína da família celulose sintase e possui atividade transferase, transferindo grupos glicosil. Ela é induzida pelo gene CA219099 que possui atividade de celulase. Sua expressão acarreta a expressão de outros três genes também de metabolismo de parede celular. CA254294 que também possui atividade de celulose sintase, assim como, CA250044 e CA244798. É interessante observar que esses genes estão sendo

expressos na planta no estado Maturando, que é quando a planta se encontra em fase de crescimento e consequentemente deve haver produção de celulose para o seu desenvolvimento. Já na planta madura, não ocorre expressão desses genes, pois a planta não tem necessidade de gastar energia para produzir celulose, já que seus órgãos já estão desenvolvidos.

CA248491 está induzindo dois genes de resposta a estresse, CF572632 (resposta a baixas temperaturas) e CF571685 (resposta a lesões) e inibindo CA272048 (transdução de sinais). Ele também inibe a expressão de dois genes de metabolismo de parede, CF571758 (atividade catalítica) e CA254458 (processo metabólico de d-xilose, atividade de UDP-glucoronato descarboxilase). Esses genes são inibidos provavelmente por apresentarem função catalítica, pois CA248491 é expresso na planta jovem, e a planta jovem, por estar em fase de crescimento, necessita da expressão de proteínas para produção de parede. Esse gene não é expresso na planta no estágio maturando, talvez porque nesse estágio inicia-se a expressão de outros genes para biossíntese de parede celular, não sendo mais necessária sua expressão.

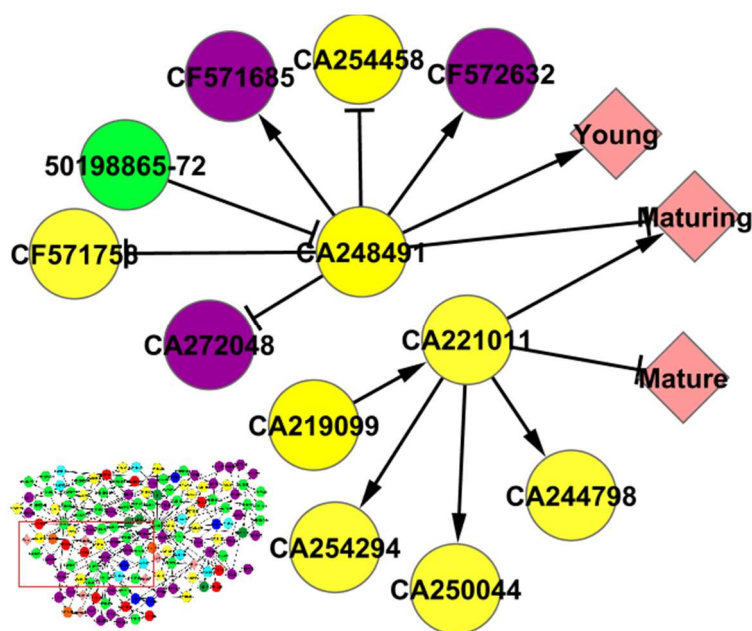


Figura 5-12 - Hubs CA248491 e CA221011. Genes CA248491 e CA221011 e os genes que interagem com eles.

Um padrão observado nessa rede é que os genes de fotossíntese inibem os genes de metabolismo de parede. Em outro trabalho do grupo, foram quantificados os carboidratos nas cultivares de cana-de-açúcar utilizadas para o experimento de fonte-dreno na folha +1 que

representa um tecido dreno em três horários diferentes ao longo do dia (Figura 5-13). Observou-se que às 12 horas, as duas cultivares possuem sacarose acumulada, porém ao final do dia somente a cultivar SP91-1049 mantém essa quantidade acumulada. A cultivar SP83-2847 apresenta maior crescimento no campo que a cultivar SP91-1049. Isso sugere que na SP83-2847 esses genes de fotossíntese provavelmente não estão reprimindo os genes de metabolismo de parede, e a planta utiliza essa energia que acumula durante o dia para a construção de parede celular e consequentemente seu crescimento.

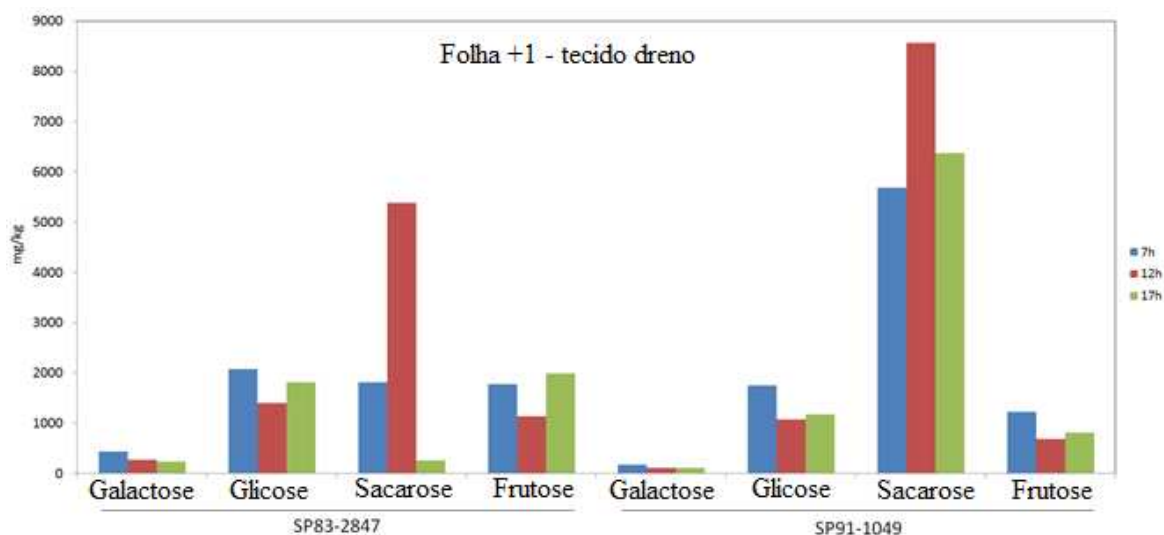


Figura 5-13 - Quantificação dos carboidratos nas duas cultivares de cana-de-açúcar. Quantificação dos carboidratos galactose, glicose, sacarose e frutose nas variedades SP83-2847 e SP91-1049 às 7, 12, e 17 horas. Em SP83-2847, às 12 horas há produção de sacarase mas ao final do dia essa sacarose é consumida.

A Figura 5-14 representa a rede de regulação para os 265 genes que foram diferencialmente expressos tanto no trabalho de Casu *et al.* (2007) quanto McCormick *et al.* (2008). Essa rede ficou com apenas 120 nós dos 265, ligados por 118 arcos. Essa redução do número de nós ocorreu devido ao método utilizado para o cálculo do *score* da rede no BNFinder que acabou considerando grande parte dos genes do grupo condicionalmente independentes dos outros genes do grupo. Como 265 genes é um número grande de genes, optou-se por usar o *score* MDL que exige menos esforço computacional e torna possível que a rede seja obtida.

Conforme a Tabela 5-4 a maior parte dos genes representados nessa rede é de Fotossíntese (33.33%), há também 11.67% de genes de Resposta a Estresses, 10% de genes de Metabolismo de Parede Celular, 9.16% de genes de Sinalização de Açúcar. As outras categorias aparecem em menor quantidade, sendo que Metabolismo de Aminoácidos e Proteínas e Metabolismo de Lipídios representam apenas 0.83% dos genes da rede. Nessa rede é possível observar a presença de genes Transportadores e de Metabolismo de Carboidratos, com 6.67%. 5.83% dos genes possuem função não identificada e 2.5% estão classificados em Outros Eventos Metabólicos. Metabolismo Mitocondrial representa 3.33% dos genes na rede.

Tabela 5-4. Classificação dos genes da rede de acordo com sua categoria funcional.

Categoria	Número de Genes	Porcentagem
Metabolismo de Aminoácidos e Proteínas	1	0.83%
Tradução, ligação ao DNA	10	8.33%
Metabolismo de Parede Celular	12	10.00%
Metabolismo de Lipídios	1	0.83%
Resposta a Estresses	14	11.67%
Metabolismo de Carboidratos	8	6.67%
Metabolismo Mitocondrial	4	3.33%
Fotossíntese	40	33.33%
Transportadores	8	6.67%
Outros Eventos Metabólicos	3	2.50%
Sinalização de Açúcares	11	9.16%
Função Não Identificada	7	5.83%
Total	120	100%

A estrutura dessa rede é composta por 5 *hubs* principais. Considerando-se uma organização hierárquica dessa rede, que seria no sentido da direita para esquerda na Figura 5-14, genes de funções essenciais para a atividade celular e atividades basais como Tradução, ligação ao DNA, Transportadores, Metabolismo de Aminoácidos e Proteínas ocupam a posição de início da rede, enquanto genes com funções não tão essenciais como Metabolismo de Parede Celular e Resposta a estresses estão representados nos níveis finais da rede. Genes de uma mesma categoria funcional apresentam a regulação por um mesmo gene e, portanto tenderam a constituir um mesmo *hub*.

O primeiro *hub* (Figura 5-15) é controlado pelo gene CA271047 que tem função de hidrólise de ATP acoplado com transporte de prótons. Esse gene está interagindo com 13 genes sendo que 10 das interações são de repressão. Ele está reprimindo genes de Sinalização

de Açúcares (CF571106 – manose *binding*); Tradução, ligação ao DNA (CO373196 – constituinte estrutural da cromatina, CA244277 – iniciação da tradução, CA126364 – histona h2a, CA077455 – histona h2a); Fotossíntese (CA089534 - fosfatase ácida, 50198865-94, 50198865-53), Transportadores (CA168705). Ele induz apenas três genes um de fotossíntese (CA151256 – malato desidrogenase), um sem função definida (CA267665) e um transportador (CF577408).

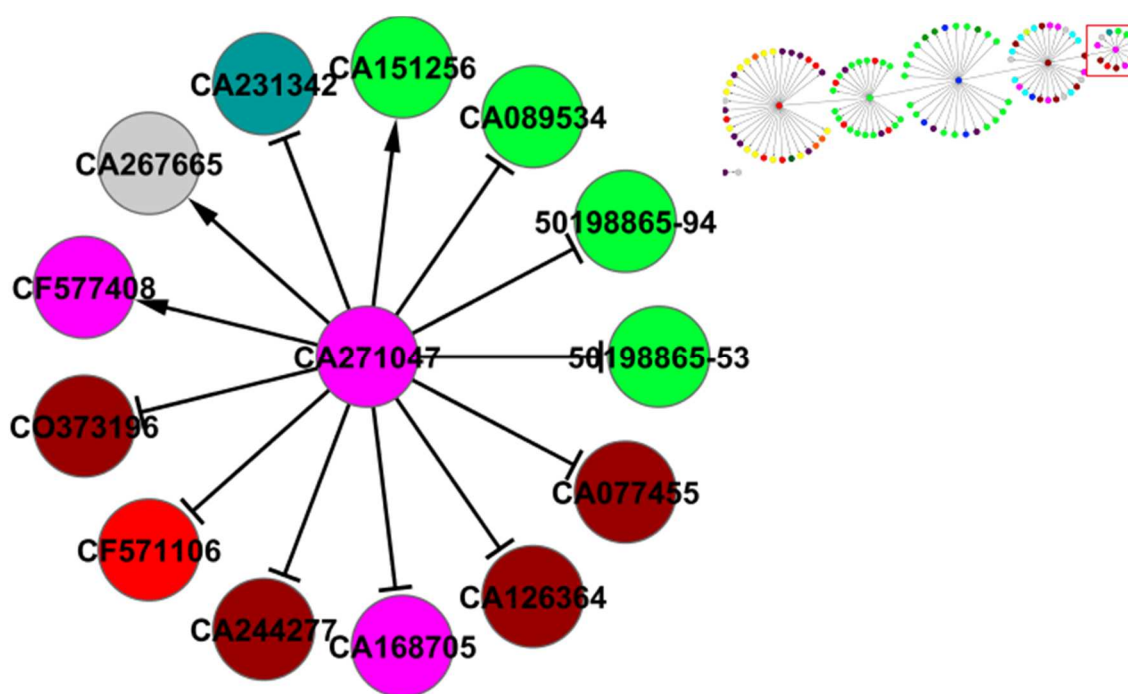


Figura 5-15 - Hub CA271047. Gene CA271047 e os genes com os quais ele interage.

O segundo agrupamento de genes (Figura 5-16) ainda apresenta uma composição que se assemelha a do primeiro, no que se refere às categorias funcionais. A expressão do gene central é induzida pelo gene CA077455 que codifica uma histona h2a. O nó central do grupo é o CA243079 que também codifica uma histona h2a. Ele induz a expressão dos genes CA128590 que é também uma histona, CA093743 que é uma catalase, CA085955, CA070403 que possui atividade de fosfofrutoquinase na glicólise, BU925731 que é uma glicosil hidrolase com atividade de sacarose alfa-glicosidase, CA208800, CA206457 que é uma frutose-bisfosfato aldolase também com atividade na glicólise, CA171058 que é uma peptidase que responde a íons cádmio, CA133751 que codifica histona h3 e CA087711 que tem atividade ligase.

Por outro lado, ele reprime a expressão de CF574067 que tem atividade de frutose bisfosfatase aldolase, CA221011 que é uma proteína da família das celulosas sintases, CA202596 que possui atividade de hidrolase e participa do processo metabólico de glucanos, CA148883 e CA065094 que codificam vesículas ligadas à membrana citoplasmática, CA112866, CA093039 e CA196594 que são gliceraldeído-3-fosfato desidrogenases, CA081293, CA260047, CA205440 participa da regulação da transcrição, CA160989 realiza transporte de lipídios, e CA087714.

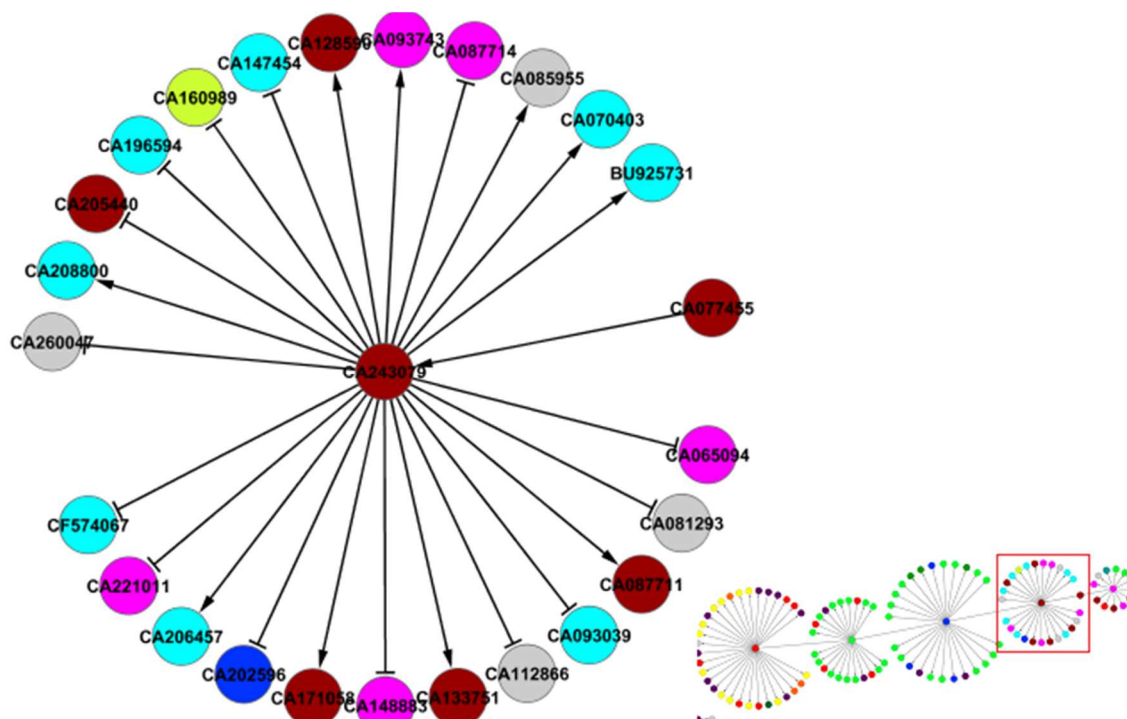


Figura 5-16 - Hub CA243069. Gene CA243069 e os genes com os quais ele interage.

No terceiro agrupamento (Figura 5-17), começam a aparecer genes relacionados com fotossíntese, sendo que eles representam a maior parte dos genes presentes nesse *hub*. O gene central é o CA151666 que codifica a subunidade grande da ADP-glicose fosforilase e participa da biossíntese de glicogênio e amido. Ele é responsável pela repressão de todos os genes de Fotossíntese presentes nesse agrupamento, exceto um deles, o CA186676 que é um precursor da apoproteína cp26 de ligação à clorofila.

Além disso, ele causa a indução de genes Metabolismo de Amido como o CA289034 que possui atividade de isoamilase, CA169622 que possui atividade fosforilase e CA169143 que é uma amilase e se liga a íons cálcio. Ele também causa a expressão de genes de

Metabolismo Mitocondrial: CA181748, uma NAD malato desidrogenase, CA170908 proteína de transporte mitocondrial, AY644460 e reprime dois genes de resposta a estresses, CA265014 e CA161102 que são proteínas que causam o final do desenvolvimento embrionário em sementes dormentes.

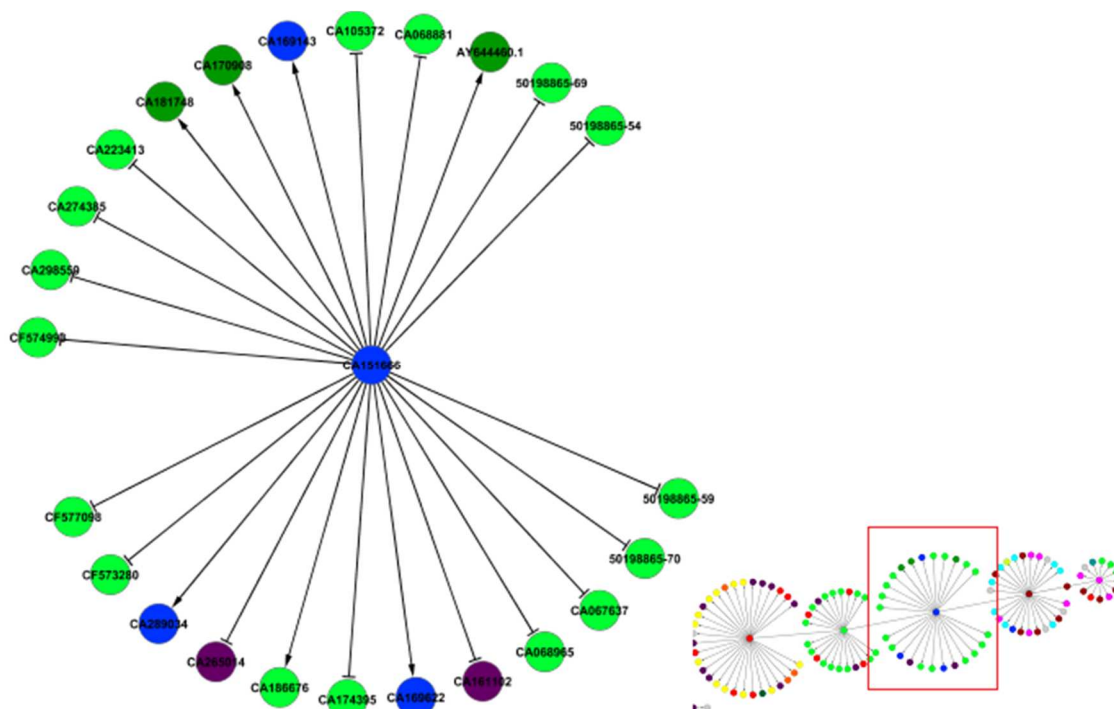


Figura 5-17 - Hub CA151666. Gene CA151666 e os genes com os quais ele interage.

A Figura 5-18 representa o terceiro grupo de genes. Nesse grupo, dos 21 genes interagindo com o nó central, somente cinco não são de Fotossíntese. O nó central, CA300310 é uma proteína de membrana do fotossistema I, localizada no cloroplasto. Ele está reprimindo os quatro genes de sinalização de açúcares aos quais está ligado. São estes: CA111104, CF571565, CA295728 e CA109224, sendo esses dois últimos translocadores de glicose-6-fosfato. Além disso, ele induz um gene de resposta a estresse (CA164068 – proteína quinase, fosforilação) e reprime outro (CA272048 – proteína quinase dependente de calmodulina, transdução de sinal).

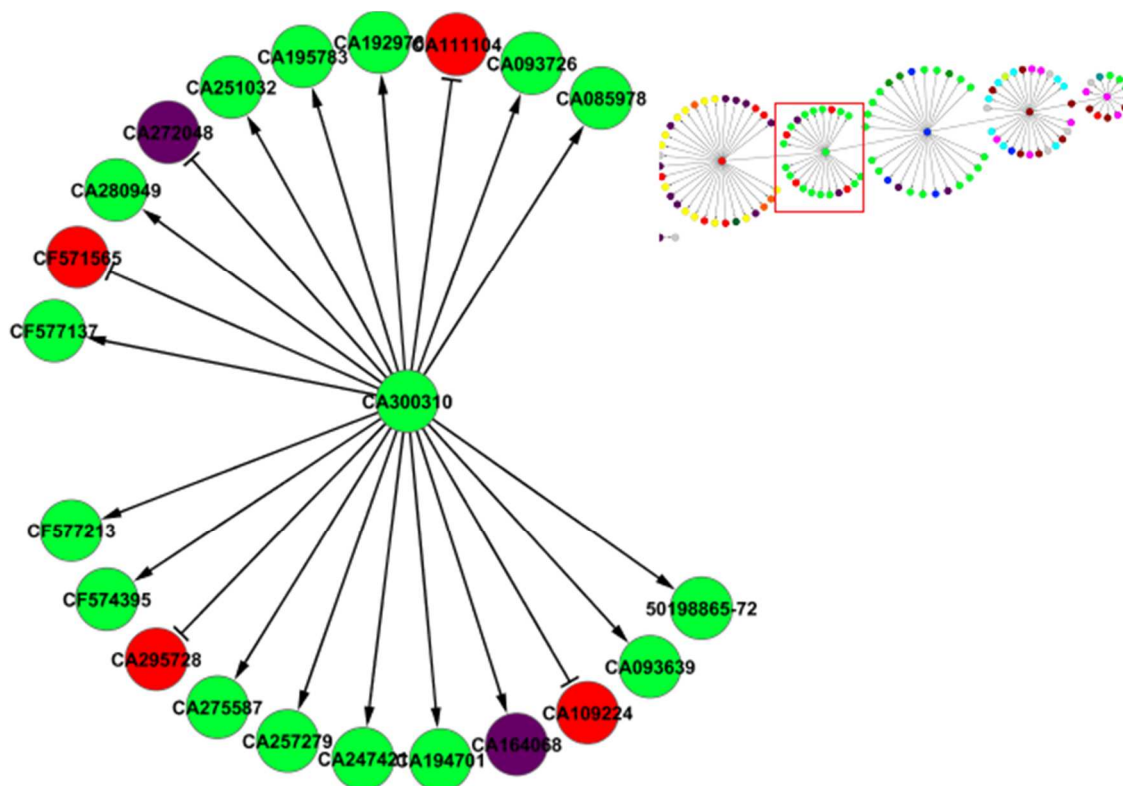


Figura 5-18 - Hub CA300310. Gene CA300310 e os genes com os quais ele interage.

No último *hub* (Figura 5-19), aparecem genes de Metabolismo de Parede Celular. O nó central desse agrupamento é o CF572284, de Sinalização de açúcares, que é um translocador de fosfato nos plastídios. Esse gene reprime a atividade apenas de outros três genes, o CF573831 e CA176198 que são trealose-6-fosfato sintases 2 com atividade catalítica, e o CA093476 que responde ao estresse por falta de água, sal e calor. Três genes de sinalização de açúcar são induzidos, CA227092 que codifica um translocador de fosfato, CA086449 sendo uma trealose fosfato sintase 2, e BU925715 um transportador transmembrana de açúcares.

Os genes de Metabolismo de Parede Celular induzidos são: CA218775 com atividade de hidrolase; CA181562 com atividade de chitinase; CA254294 celulose sintase; CA265758 com atividade celulase; CF571539; CF572353; CF571758; CA244283 vesícula ligada à membrana citoplasmática; CA196999 proteína da família glicosil transferase; CA105799 atividade hidrolase e AJ231133. E os genes de resposta a estresse induzidos são: CA093131 proteína quinase induzida por AMP em resposta a níveis de nitrogênio; CA267486 proteína de resposta à água e estresse; CF571023 proteína da família álcool desidrogenase; CF577254

também possui atividade de proteína quinase ativada por AMP; CA171450 e CA270299 glutamato descarboxilase; BQ535164 proteína induzida em baixas temperaturas; CA122355.

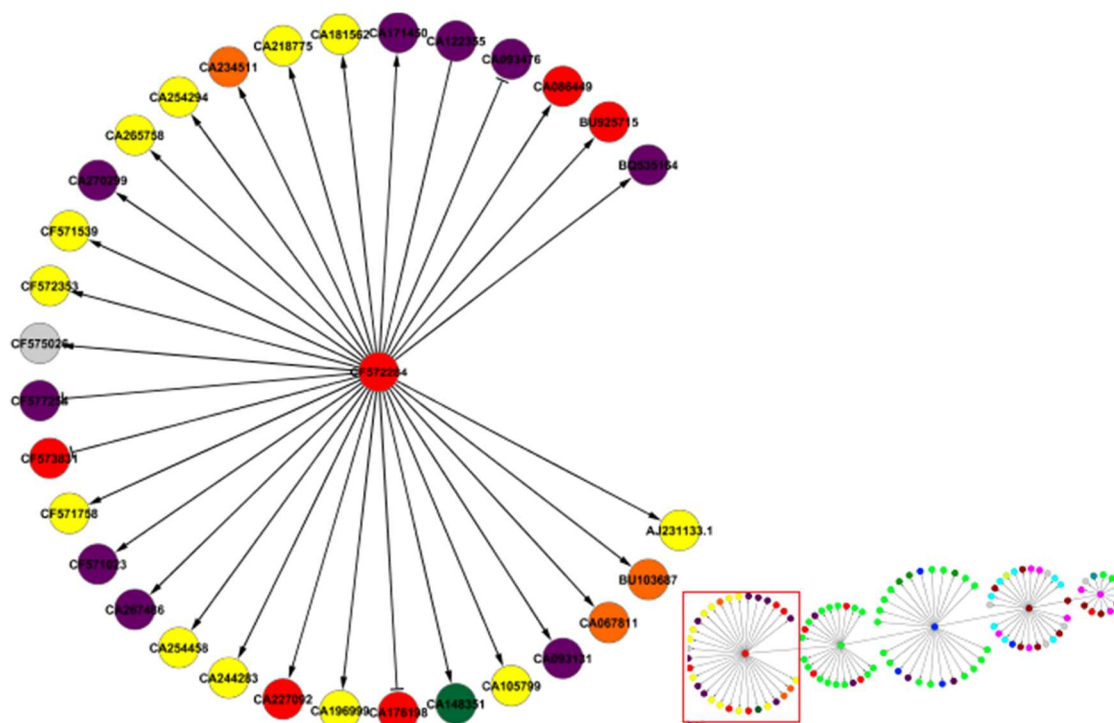


Figura 5-19 - Hub CF572284. Gene CF572284 e os genes com os quais ele interage.

É importante ressaltar que alguns dos genes representados em todas as redes, marcados em cor cinza são genes que não possuem anotação. Esses genes podem ser ncRNAs, miRNAs ou genes novos e talvez possuam funções regulatórias importantes.

5.2. Análise do padrão de expressão das redes

A Figura 5-20 mostra o resultado da análise obtida pelo ViaComplex para a rede de co-expressão de todos os genes do *microarray*. A Figura 5-20-a, mostra o padrão de expressão dos genes de resposta às manipulações da fotossíntese, do trabalho do McCormick *et al.* (2008). Já a Figura 5-20-b mostra o padrão de expressão dos genes de resposta ao experimento de metabolismo de parede celular (Casu *et al.*, 2007). As regiões em azul mostram genes com expressão reprimida e as regiões em laranja mostram genes que aumentaram sua expressão. Assim pode ser percebido que os genes que respondem a modificações nas condições de fotossíntese estão localizados na parte superior da rede enquanto genes que respondem a estágio de desenvolvimento da planta estão agrupados na parte esquerda da rede.

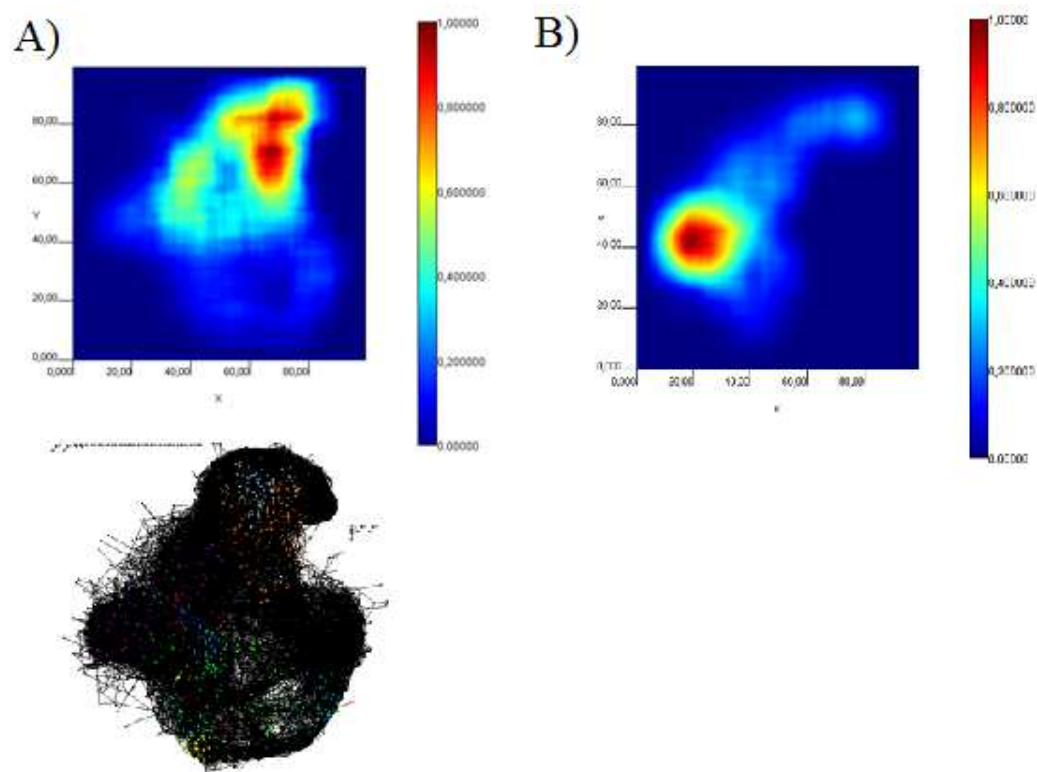


Figura 5-20 - Análise do padrão de expressão para a rede de co-expressão de todos os genes do *array*. A) Experimento de manipulação da fotossíntese. B) Experimento de metabolismo de parede celular. Regiões em laranja representam alto nível de expressão e regiões em azul baixo nível de expressão.

A Figura 5-21 mostra o padrão de expressão gênica para a rede do grupo dos 143 genes diferencialmente expressos. A Figura 5-21-a representa o padrão de expressão para o experimento de manipulação da fotossíntese. Quando comparado à rede, pode-se perceber que as regiões que demonstram expressão aumentada coincidem com os nós que possuem função relacionada com fotossíntese e resposta a estresse. A Figura 5-21-b mostra o padrão de expressão gênica para o experimento de metabolismo de parede, e nessa representação as regiões que possuem expressão gênica aumentada coincidem com locais onde se encontram genes relacionados a metabolismo de parede celular e de carboidratos.

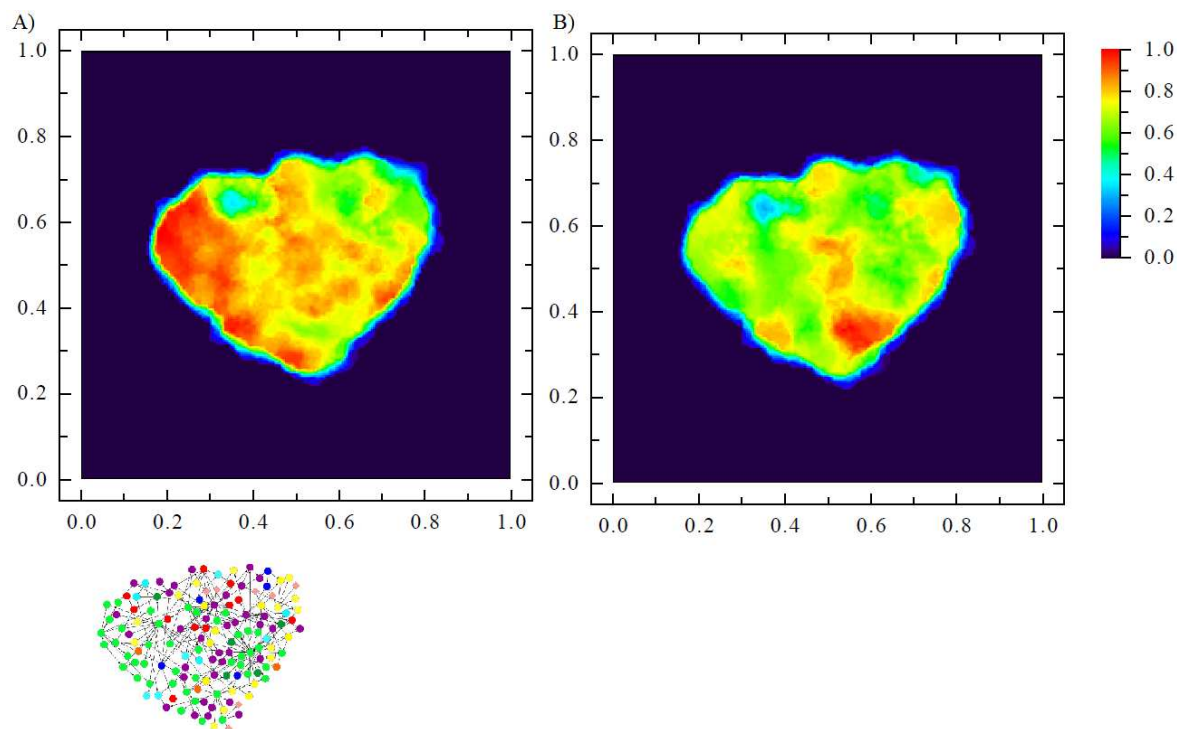


Figura 5-21 - Análise do padrão de expressão para a rede de 143 genes. A) Experimento de manipulação de fotossíntese. B) Experimento de metabolismo de parede celular. Regiões em laranja representam alto nível de expressão e regiões em azul baixo nível de expressão.

5.3. RNA-seq

Foram obtidos dois clusters hierárquicos a partir do experimento de RNA-seq. Eles foram elaborados com base nos genes diferencialmente expressos para o experimento do genótipo 2847. A Figura 5-22 mostra o gráfico representando as diferenças dos níveis de expressão entre a planta que ficou sob sombreamento durante todo o experimento, a planta que ficou sob sombreamento durante 6 dias, planta que ficou sob sombreamento durante 3 dias e a planta controle. Na figura, é possível observar que a planta que ficou sob sombreamento durante todo o tempo de experimento e a planta que não esteve sob sombreamento apresenta padrões de expressão contrários.

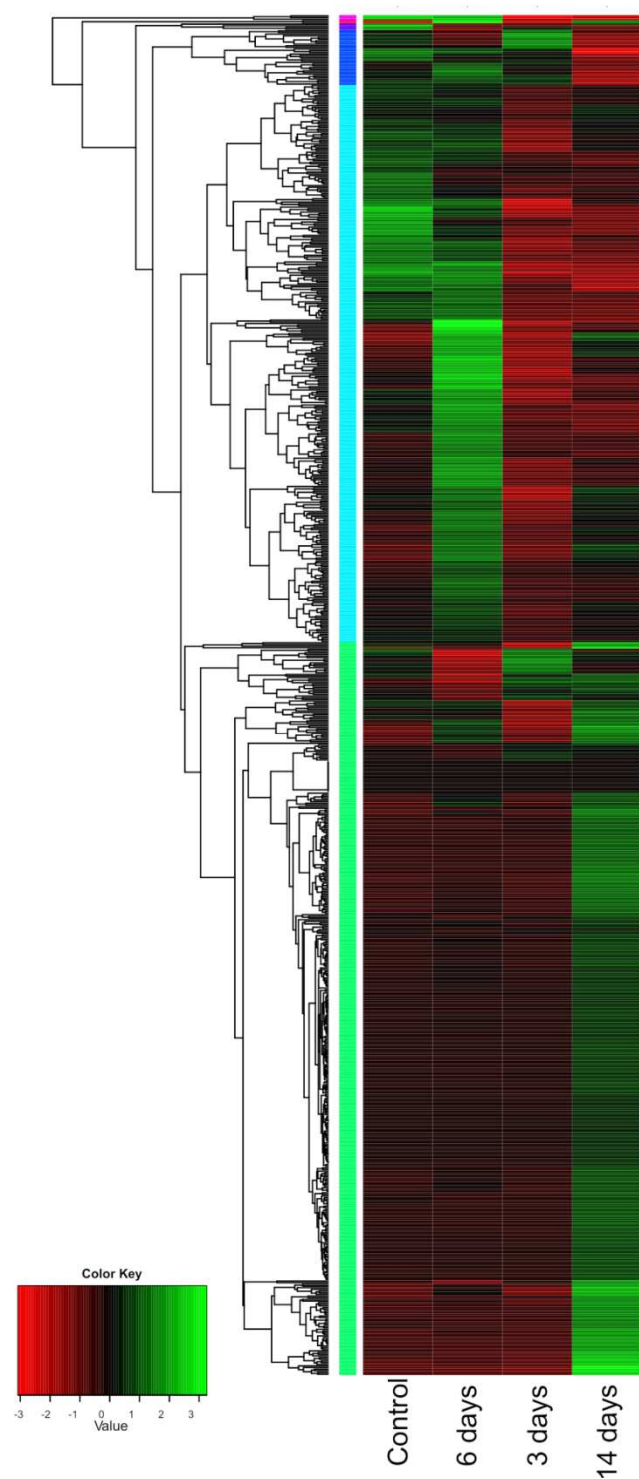


Figura 5-22 - *Heat map* para os genes diferencialmente expressos no experimento de manipulação fonte-dreno (n = 819). Esse gráfico considera os níveis de expressão para o genótipo 2847 durante todo o período de tratamento.

Na Figura 5-23 é mostrado o cluster hierárquico representando o sexto dia do experimento ($n = 1644$). O material coletado às 8, 12 e 17 horas e portanto, a variação da expressão é mostrada nesses horários ao longo do sexto dia somente. Cores tendendo a verde indicam genes expressos e vermelho reprimidos.

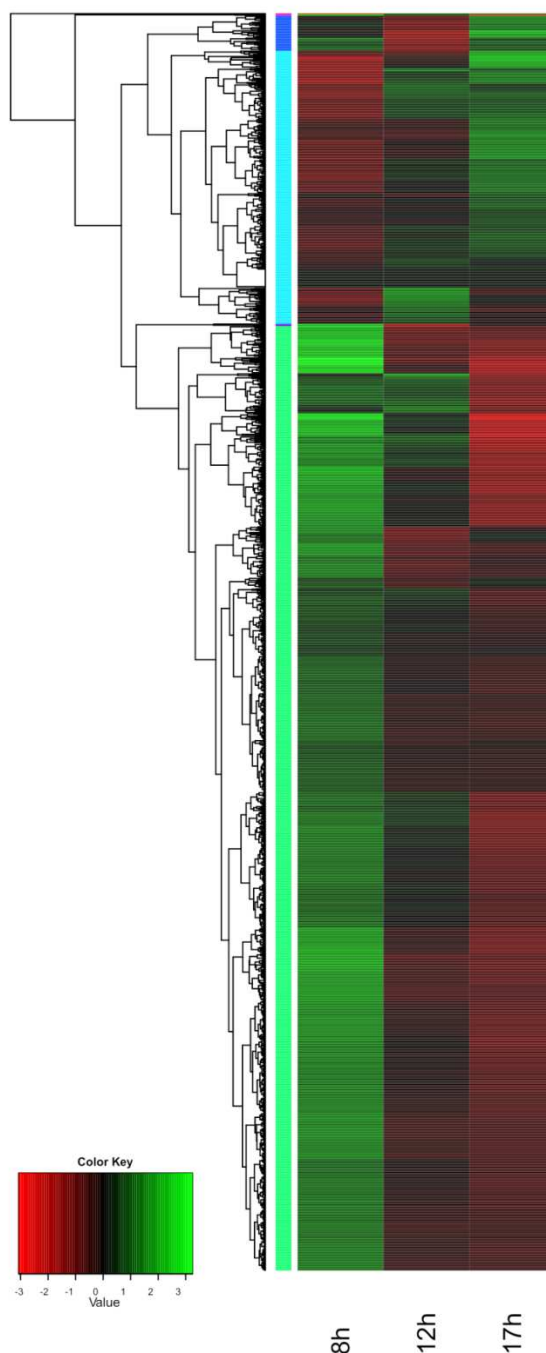


Figura 5-23 - Cluster hierárquico representando o sexto dia do experimento ($n = 1644$). Material coletado às 8, 12 e 17 horas.

Pretende-se usar os dados desse experimento de RNA-seq para elaboração de redes, pois a cana-de-açúcar possui aproximadamente 32000 genes e no chip de *microarray* estão representados aproximadamente 8200 apenas. Dessa forma, muitos dos genes que podem ter papel na sinalização e metabolismo de açúcar não se encontram representados no chip e consequentemente não são englobados nas redes.

6. CONSIDERAÇÕES FINAIS

As redes bayesianas ainda são uma estratégia que exige muito esforço computacional, impossibilitando a elaboração de redes com muitos genes e fazendo necessária a seleção de genes de interesse. Isso torna necessária a implementação de novos algoritmos para sua inferência assim como a utilização de outras estratégias para obtenção de redes. A metodologia de co-expressão é rápida, porém menos precisa e não permite a visualização da direcionalidade da interação.

É necessária também a busca por estratégias que permitam a seleção de genes mais relacionados a um dado gene de interesse. Por exemplo, nesse estudo tentou-se, porém sem sucesso o uso de técnicas de *Incremental Growth* onde a partir de um conjunto inicial de genes é calculado qual o melhor gene para integrar o grupo, utilizando-se algum tipo de *score* e isso é repetido até que o número de genes de interesse seja alcançado. O problema dessa abordagem é que não havia como determinar se o gene escolhido para ser adicionado ao grupo era realmente o melhor gene.

Utilizar a seleção dos genes pelo método de HCCA permitiu selecionar genes relacionados à SPS com funções primárias ou desconhecidas, que seriam genes que não despertam interesse num primeiro momento, mas possuem relações próximas com a SPS e consequentemente são importantes nas vias de produção de açúcares. A vantagem disso é que podem ser representados na rede fatores de transcrição, RNAs não codantes, miRNAs e até mesmo genes novos.

Através dessa abordagem, constatou-se que SPS interage diretamente com genes de transportadores, tradução, ligação ao DNA e genes de função não identificada. Os genes mais representados no cluster da SPS em cana são genes de tradução e ligação ao DNA e genes de função desconhecida enquanto os menos representados são de fotossíntese, resposta a hormônios, e outros eventos metabólicos. Desses genes de função não identificada 11 não apresentaram alinhamento com nenhuma sequência já identificada, o que é uma evidência de que esses genes possam ser genes novos.

A rede do cluster da sacarose apresentou 7 genes principais (*hubs*) que aparentam ter importante função dentro do cluster, sendo essenciais para o funcionamento correto da rede. A

rede dos 136 genes apresentou 6 genes principais, sendo que a maioria deles eram de fotossíntese. A rede dos genes diferencialmente expressos (265 genes) apresentou 5 *hubs* principais, sendo que genes com funções semelhantes tenderam a sofrer regulação por um mesmo gene, formando grupos de genes com funções semelhantes em cada *hub*.

O padrão de expressão gênica para os genes do experimento de fotossíntese mostrou uma maior expressão de genes de fotossíntese e resposta a estresse enquanto o padrão de expressão para o experimento de metabolismo de parede celular foi detectada maior expressão dos genes que pertenciam ao grupo de metabolismo de parede celular condizendo com o tipo de experimento dos quais os dados foram retirados.

Devido à limitação da quantidade de genes representados no chip de *microarray*, pretende-se utilizar os dados do experimento de RNA-seq para a obtenção de redes, e assim espera-se que novos genes que antes não haviam sido amostrados sejam também representados nas redes.

REFERÊNCIAS BIBLIOGRÁFICAS

- Akutsu, T.; Miyano, S.; Kuhara, S. **Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function.** Journal of Computational Biology, 7(3-4), 331-343. 2000.
- Arkin, A.; Ross, J.; McAdams, H.H. **Stochastic kinetic analysis of developmental pathway bifurcation in phage-infected Escherichia coli cells.** Genetics, 149, 1633-1648. 1998.
- Ascher, D.; Dubois, P.F.; Hinsén, K.; Hugunin, J.; Oliphant, T. **An open source project: Numerical python.** Technical Report UCRL-MA-128569, Lawrence Livermore National Laboratory. 2001.
- Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T. **Gene ontology: tool for the unification of biology.** Nat. Genet., 25, 25–29. 2000.
- Bansal, M.; Belcastro, V.; Ambesi-Impiombato, A.; Bernardo, D. **How to infer gene networks from expression profiles.** Molecular System Biology, 3:78, 2007.
- Barbieri, V.; Silva, F.C. da. **Influence of sugarcane burning on the temperatures of the cane stalks and the soil surface.** Holos Environment, v. 8. 2008.
- Barker, L.; Kühn, C.; Weise, A.; Schulz, A.; Gebhardt, C.; Hirner, B.; Hellmann, H.; Schulze, W.; Ward, J.M.; Frommer, W.B. **SUT2, a putative sucrose sensor in sieve elements.** Plant Cell 12, 1153–1164. 2000.
- Black, C.C.; Loboda, T.; Chen, J.; Sung, S. **Can sucrose cleavage enzymes serve as markers for sink strength and is sucrose a signal molecule during plant sink development?** American Society of Plant Physiologists 1995: 49–64. 1995.
- Bonn, S.; Zinzen, R.P.; Girardot, C.; Gustafson, E.H.; Perez-Gonzalez, A.; Delhomme, N.; Furlong, E.E. **Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development.** Nature genetics, 44(2), 148-156. 2012.
- Bortoleto, J. F. **A biotecnologia no melhoramento da cana-de-açúcar.** Seminário em Genética e Melhoramento de Plantas. ESALQ/USP. 2010.
- Bose, R.; Molina, H.; Patterson, A.S.; Bitok, J.K.; Periaswamy, B.; Bader, J.S.; Cole, P.A. **Phosphoproteomic analysis of Her2/neu signaling and inhibition.** Proceedings of the National Academy of Sciences, 103(26), 9773-9778. 2006.

- Botha, F.C.; Black, K.G. **Sucrose phosphate synthase and sucrose synthase activity during maturation of internodal tissue in sugarcane**. Australian Journal of Plant Physiology, 27, 81–85. 2000.
- Bottcher, S. G. **Learning Bayesian Networks with Mixed Variables**. Tese de Phd. Departamento de Ciências Matemáticas, Aalborg University, 2004.
- Bottcher, S. G.; Dethlefsen, C. **Learning bayesian networks with R**. Proceedings of the 3rd International Workshop on Distributed Statistical Computing. Vienna, Austria, Kurt Hornik, Friedrich Leisch&AchimZeileis, 2003.
- Bower, N.I.; Casu, R.E.; Maclean, D.J.; Reverter, A.; Chapman, S.C.; Manners, J.M. **Transcriptional response of sugarcane roots to methyl jasmonate**. Plant Sci, 8:761–772. 2005.
- Branco, D.S. **Sinalização por carboidratos em cana-de-açúcar e divergência evolutiva**. 203 p. Dissertação (Mestrado em Genética e Biologia Molecular) - Instituto de Biologia, Universidade Estadual de Campinas, Campinas, 2008.
- Brazhnik, P.; de la Fuente, A.; Mendes, P. **Gene networks: how to put the function in genomics**. TRENDS in Biotechnology, 20(11), 467-472. 2002.
- Brunini, O. **Ambientes climáticos e exploração agrícola da cana-de-açúcar**. In: DINARDO-MIRANDA, L.L.; VASCONCELOS, A.C.M; LANDELL, M.G.A. (Ed.). Cana-de-açúcar. Campinas: Instituto Agrônômico, 882p, 2008.
- Buntine, W. **Graphical Models for Discovering Knowledge**. Research Institute for Advanced Computing Sciences, Computational Sciences Division, NASA Ames Research Center, p. 65-86, 1996.
- Calvin, M. **The photosynthetic carbon cycle**. J. Chem. Soc. 1956:1895–1915. 1956.
- Camargo, S.R.; Cancado, G.M.A.; Ulian, E.C.; Menossi, M. **Identification of genes responsive to the application of ethanol on sugarcane leaves**. Plant Cell Rep, v. 26, p. 2119–2128. 2007.
- Cantone, I.; Marucci, L.; Iorio, F.; Ricci, M.A.; Belcastro, V.; Bansal, M.; Cosma, M.P. **A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches**. Cell, 137(1), 172. 2009.
- Carson, D.; Huckett, B.; Botha, F. **Differential gene expression in sugarcane leaf and intermodal tissues of varying maturity**. S Afr J Bot 68:434–442. 2002.

- Casagrande, A. A. **Tópicos de morfologia e fisiologia da cana-de-açúcar**. Jaboticabal, FUNEP, 157p. 1991.
- Casagrande, M. **Evaluation précoce des performances du blé biologique (rendement et teneur en protéines): une approche combinée de diagnostic agronomique, de modélisation à l'aide d'indicateurs de nuisibilité et d'études des pratiques dans les exploitations agricoles**. PhD Thesis - l'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech). 2008.
- Castro, M.A.; Rybarczyk Filho, J.L.; Dalmolin, R.J.; Sinigaglia, M.; Moreira, J.C.; Mombach, J.C.; de Almeida, R. M. **ViaComplex: software for landscape analysis of gene expression networks in genomic context**. *Bioinformatics*, 25(11), 1468-1469. 2009.
- Casu, R.E.; Dimmock, C.M.; Chapman, S.C.; Grof, C.P.L.; McIntyre, C.L.; Bonnett, G.D.; Manners, J.M. **Identification of differentially expressed transcripts from maturing stem of sugarcane by in silico analysis of stem expressed sequence tags and gene expression profiling**. *Plant MolBiol* 54:503–517. 2004.
- Casu, R.E.; Grof, C.P.L.; Rae, A.L.; McIntyre, C.L.; Dimmock, C.M.; Manners, J.M. **Identification of a novel sugar transporter homologue strongly expressed in maturing stem vascular tissues of sugarcane by expressed sequence tag and microarray analysis**. *Plant Mol. Biol.*, 52, pp. 371–386. 2003.
- Casu, R.E.; Jarmey, J.M.; Bonnett, G.D.; Manners, J.M. **Identification of transcripts associated with cell wall metabolism and development in the stem of sugarcane by AffymetrixGeneChip Sugarcane Genome Array expression profiling**. *FunctIntegr Genomics* 7:153–167. 2007.
- Casu, R.E.; Manners, J.M.; Bonnett, G.D.; Jackson, P.A.; McIntyre, C.L.; Dunne, R.; Chapman, S.C.; Rae, A.L.; Grof, C.P.L. **Genomics approaches for the identification of genes determining important traits in sugarcane**. *Field Crops Res* 92:137–147. 2005.
- Chai, L.; Mohamad, M.; Deris, S.; Chong, C.; Choon, Y.; Ibrahim, Z.; Omatu, S. **Inferring Gene Regulatory Networks from Gene Expression Data by a Dynamic Bayesian Network-Based Model**. *Distributed Computing and Artificial Intelligence*, 379-386. 2012.
- Chandra, A.; Jain, R.; Rai, R.K.; Solomon, S. **Revisiting the source-sink paradigm in sugarcane**. *Curr Sci*, 100(7), 978-980. 2011.

- Chen, T.; He, H. L.; Church, G. M. **Modeling gene expression with differential equations.** Pacific Symposium Biocomputing, p. 29-40, 1999.
- Chen, X.; Anantha, G.; Wang, X. **An effective structure learning method for constructing gene networks.** Bioinformatics, v. 22, n. 11, p. 1367-1374, doi:10.1093/bioinformatics/btl090, 2006.
- Chen, X.; Chen, M.; Ning, K. **BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network.** Bioinformatics, v. 22, n. 23, p. 2952-2954, 2006.
- Chiou T.-J.; Bush D. R. **Sucrose is a signal molecule in assimilate partitioning.** Proc. Natl. Acad. Sci. USA 95, 4784-4788. 1998.
- Conesa, A.; Gotz, S.; Garcia-Gomez, J.M.; Terol, J.; Talon, M.; Robles, M. **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** Bioinformatics. 21:3674---3676. 2005.
- Cornelia, A. **Modeling gene regulatory networks.** Internship report, 50 p. Institute of Advanced Biosciences, Keio University, Japão, 2008.
- D'Haeseleer, P.; Wen, X.; Fuhrman, S.; Somogyi, R. **Linear modeling of mRNA expression levels during CNS development and injury.** Pac. Symp. Biocomput., 3, 41–52. 1999.
- D'Hont, A. **Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana.** Cytogenet.Genome Res. 109, 27–33, 2005.
- Dabrowski, M.; Dojer, N.; Zawadzka, M.; Mieczkowski, J.; Kaminska, B. **Comparative analysis of cis-regulation following stroke and seizures in subspaces of conserved eigensystems.** BMC systems biology, 4(1), 86. 2010.
- Damaj, M.B.; Beremand, P.D.; Buenrostro-Nava, M.T.; Ivy, J.; Kumpatla, S.P.; Jifon, J.; Beyene, G.; Yu, Q.; Thomas, T.L.; Mirkov, T.E. **Isolating promoters of multigene family members from the polyploidy sugarcane genome by PCR-based walking in BAC DNA.** Genome 53:840–847. 2010b.
- De Hoon, M.J.L.; Imoto, S.; Kobayashi, K.; Ogasawara, N.; Miyano, S. **Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations.** Pac. Symp. Biocomput., 8, 17–28. 2003.
- de Souza, A.P.; Gaspar, M.; da Silva, E.A.; Ulian, E.C.; Waclawovsky, A.J.; Nishiyama, M.Y. Jr; Dos Santos, R.V.; Teixeira, M.M.; Souza, G.M.; Buckeridge, M.S. **Elevated CO₂**

- increases photosynthesis, biomass and productivity, and modifies gene expression in sugarcane.** *Plant Cell Environ* 31:1116–1127. 2008.
- Dixon, R.A.; Achnine, L.; Kota, P.; Liu, C.J.; Reddy, M.S.S.; Wang, L.J. **The phenylpropanoid pathway and plant defence - a genomics perspective.** *Molecular Plant Pathology* 3, 371–390. 2002.
- Dobra, A.; Hans, C.; Jones, B.; Nevins, J. R.; Yao, G.; West, M. **Sparse graphical models for exploring gene expression data.** *Journal of Multivariate Analysis*, v. 90, p. 196-212, 2004.
- Dojer, N.; Gambin, A.; Mizera, A.; Wilczyński, B.; Tiuryn, J. **Applying dynamic Bayesian networks to perturbed gene expression data.** *BMC bioinformatics*, 7(1), 249. 2006.
- Eckel, C.C.; El-Gamal, M.A.; Wilson, R.K. **Risk loving after the storm: A Bayesian-Network study of Hurricane Katrina evacuees.** *Journal of Economic Behavior & Organization*, 69(2), 110-124. 2009.
- El Samad, H.; Khammash, M.; Petzold, L.; Gillespie, D. **Stochastic Modeling of Gene Regulatory Networks.** *International Journal of Robust and Nonlinear Control*, 00:1-6, 2002.
- Eldawlatly, S.; Zhou, Y.; Jin, R.; Oweiss, K.G. **On the use of dynamic Bayesian networks in reconstructing functional neuronal networks from spike train ensembles.** *Neural computation*, 22(1), 158-189. 2010.
- FAO. Food and Agriculture Organization. Statistical Database. Disponível em: <<http://faostat.fao.org/>>. Acesso em: 15 set 2012.
- Felix, J.M. **Análise da expressão gênica envolvida no metabolismo de sacarose em cana-de-açúcar (Saccharum spp).** PhD thesis defended March 10th, 2006 Universidade Estadual de Campinas, Campinas, São Paulo, Brasil <http://libdigi.unicamp.br/document/?allowbreakview=vtls000380284>. 2006.
- Felix, J.M.; Papini-Terzi, F.S.; Rocha, F.R.; Vencio, R.Z.N.; Vicentini, R.; Nishiyama, M.Y. Jr; Ulian, E.C.; Souza, G.M.; Menossi, M. **Expression profile of signal transduction components in a sugarcane population segregating for sugar content.** *Trop Plant Biol* 2:98–109. 2009.
- Fernie, A.R.; Willmitzer, L.; Trethewey, R.N. **Sucrose to starch: a transition in molecular plant physiology.** *Trends Plant Sci.*, 7, 35–42. 2002.

- Fortes, C. **Discriminação varietal e estimativa de produtividade agroindustrial de cana-de-açúcar pelo sensor orbital ETM+/LANDSAT 7.** 131 p. Dissertação (Mestrado em Agronomia) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2003.
- Frey, B. J.; Jojic, N. **A comparison of Algorithms for inference and learning in Probabilistic Graphical Models.** IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 27, n. 9, p. 1392-1416, 2005.
- Friedman, N. **Inferring Cellular Networks Using Probabilistic Graphical Models.** Science, v. 303, p. 799-805, 2004.
- Friedman, N.; Linial, M.; Nachman, I.; Pe’er, D. **Using Bayesian Networks to Analyse Expression Data.** Journal of Computational Biology, v. 7, n. 3/4, p. 601-620, 2000.
- Fujita, A. **Análise de dados de expressão gênica: normalização de microarrays e modelagem de redes regulatórias.** Tese de doutorado em Ciências. USP, São Paulo. 2007.
- Gahrtz, M.; Stolz, J.; Sauer, N. **A phloem specific sucrose-H⁺ symporter from Plantago major L. supports the model of apoplastic phloem loading.** Plant J. 6, 697-706. 1994.
- Galtier, N.; Foyer, C.H.; Huber, J.; Voelker, T.A.; Huber, S.C. **Effects of elevated sucrose-phosphate-synthase activity of photosynthesis, assimilate partitioning, and growth in tomato (Lycopersicon esculentum var UC82B).** Plant Physiology 101, 535-43. 1993.
- Gilbert, R.A.; Shine Júnior, J.M.; Miller, J.D.; Rainbolt, C.R. **The effect genotype, environmental and time of harvest on sugarcane yields in Florida, USA.** Field Crops Research, v.95, p.156-170, 2006.
- Goutsias, J.; Lee, N. H. **Computational and Experimental Approches for Modeling Gene Regulatory Networks.** Current Pharmaceutical Design, v. 13, p. 1416-1436, 2007.
- Grof, C.P.L.; Knight, D.P.; McNeil, S.D.; Lunn, J.E.; Campbell, J.A. **A modified assay method shows leaf sucrose-phosphate synthase activity is correlated with leaf sucrose content across a range of sugarcane varieties.** Australian Journal of Plant Physiology, Collingwood, v. 25, n. 4, p. 499-502, 1998.
- Gupta, A.K.; Kaur, N. **Sugar signaling and gene expression in relation to carbohydrate metabolism under abiotic stresses in plants.** Journal of Biosciences 30: 761-776. 2005.

- Hajek, A. **Interpretations of Probability**. 2007. <http://plato.stanford.edu/entries/probability-interpret>. Acesso em: 03/2013.
- Hamerski, F. **Estudo de variáveis no processo de carbonatação do caldo de cana-de-açúcar**. Curitiba, 149p. Dissertação (Mestrado em Tecnologia de Alimentos). Setor de Tecnologia, Universidade Federal do Paraná (UFP). 2009.
- Hartemink, A. BANJO: Bayesian Network Inference with Java Objects. Available at: <https://www.cs.duke.edu/~amink/software/banjo/>. Acesso: 16 set 2012. 2006.
- Hartemink, A.; Gifford, D. K.; Jaakkola, T. S.; Young, R. A. **Using Graphical Models and Genomic Expression Data to Statistically Validate Models of Genetic Regulatory Networks**. Pacific Symposium on Biocomputing, v. 6, p.422-433, 2001.
- Hasty, J.; McMillen, D.; Isaacs, F.; Collins, J. J. **Computational studies of gene regulatory networks: in numero molecular biology**. Nature Reviews, Genetics, v. 2, p. 268-279, 2001.
- Hatch M.D.; Slack C.R. **Photosynthesis by sugar-cane leaves. A new carboxylation reaction and the pathway of sugar formation**. Biochem J. Oct;101(1):103–111. 1966.
- Hickman, G. J.; Hodgman, T. C. **Inference of gene regulatory networks using Boolean-network inference methods**. Journal of Bioinformatics and Computacional Biology, 2009. v. 7, n. 6, p. 1013-1029.
- Ho, L.C. **Metabolism and compartmentation of imported sugars in sink organs in relation to sink strength**. Annual Review of Plant Physiology and Plant Molecular Biology.39, 355-78. 1988.
- Huber, S.C.; Huber, J.L. **Role and regulation of sucrose-phosphate synthase in higher plants**. Annu. Rev. Plant Physiol. Plant Mol. Biol. 47:431–44. 1996.
- Husmeier, D. **Reverse engineering of genetic networks with Bayesian networks**. Biochemical Society Transactions, v. 31, part 6, p. 1516-1518, 2003.
- Imoto, S.; Tomoyuki, H.; Goto, T.; Tashiro, K.; Kuhara, S.; Miyano, S. **Combining microarrays and biological knowledge for estimating gene networks via bayesian networks**. Journal of Bioinformatics and Computational Biology, v. 11:35, p. 77-98, 2004.
- Jackson P.A. **Breeding for improved sugar content in sugarcane**. Field Crops Res 92:277–290, 2005.

- Jadoski, S.O.; Maggi, M.F.; Lima, A. dos Santos; Vieira, D.J.; Wazne, R. **Avaliação da fórmula NPK 8-30-20 com adição de gesso agrícola em comparação à adubação convencional para produção de batata (*Solanum tuberosum* L.)**. Pesquisa Aplicada & Agrotecnologia, v3, n.1, p.111-123, 2010.
- Jang, J.C.; Sheen, J. **Sugar sensing in higherplants**. Trends in Plant Science, v. 2, 1997.
- Johansson, H. **Gene regulation of UDP-glucose synthesis and metabolism in plants**. PhD Thesis. Umeå University, 2003.
- Kaern, M.; Blake, W. J.; Collins, J. J. **The engineering of gene regulatory networks**. Annual review of biomedical engineering, 5(1), 179-206. 2003.
- Karlebach, G.; Shamir, R. **Modeling and analysis of gene regulatory networks**. Nature Reviews, v. 9, p. 770-780, 2008.
- Kauffman, S.A. **Metabolic stability and epigenesis in randomly constructed genetic nets**. J. Theor. Biol., 22, 437-467. 1969.
- Kishino, H.; Waddell, P.J. **Correspondence analysis of genes and tissue types and finding genetic links from microarray data**. Genome Informatics, 11, 83-95. 2000.
- Koch, K.E. **Carbohydrate-modulated gene expression in plants**. Annual Review of Plant Physiology and Plant Molecular Biology 47, (in press). 1996.
- Krause K.P.; Stitt M. **Sucrose-6-phosphate levels in spinach leaves and their effects on sucrose-phosphate synthase**. Phytochemistry 31: 1143-1146. 1992.
- Krook, J.; Vreugdenhil, D.; van der Plas Linus, H.W. **Uptake and phosphorylation of glucose and fructose in *Daucus carota* cell suspensions are differently regulated**. Plant PhysiolBiochem 38: 603-612. 2000.
- Lee, I.; Ambaru, B.; Thakkar, P.; Marcotte, E.M.; Rhee, S.Y. **Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana***. Nature Biotechnology, v. 28, n. 2, p. 149-158, 2012.
- Lee, I.; Ambaru, B.; Thakkar, P.; Marcotte, E.M.; Rhee, S.Y. **Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana***. Nature Biotechnology, vol. 28 no. 2. 2010.
- Lei M.; Liu D. **Sucrose regulates plant responses to deficiencies in multiple nutrients, (aust)**, Plant Signaling & Behavior, Oxford, v.10 p.1247-1249, 2011a.

- Lei M.; Liu Y.; Zhang B.; Zhao Y.; Wang X.; Zhou Y.; Raghothama K. G. **Genetic and genomic evidence that sucrose is a global regulator of plant responses to phosphate starvation in Arabidopsis**. Plant Physiology, Washington, v.156,n.3, p.1116-1130, 2011b.
- Li L.; Li Y.; Zhang L.; Xu C.; Su T.; Ren D.; Yang H. **Sucrose induces rapid activation of CfSAPK, a mitogen-activated protein kinase, in Cephalostachyum fuchsianum Gamble cells**. Plant, Cell & Environment, Oxford In Press, 2012.
- Lima, A. D. **Otimização do aproveitamento do palhico da cana-de-açúcar**. Tese (Doutorado) - Faculdade de Ciências Agrônômicas, Universidade Estadual Paulista "Júlio De Mesquita Filho", Botucatu-SP. 2009.
- Lima, L. A. **Um algoritmo eficiente para o crescimento de redes sobre o grafo aleatório completo de um sistema de regulação gênica**. 78 p. Dissertação (Mestrado em Ciências), Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2009.
- Lima, L.A.; Barrera, J. **Técnicas de identificação de redes de regulação gênica**. IV Simpósio de Iniciação Científica e Pós-Graduação. Instituto de Matemática e Estatística - IME. Universidade de São Paulo - USP. 2008.
- Loreti, E.; De Bellis, L.; Alpi, A.; Perata, P. **Why and how plant cells sense sugars?** Ann Bot 88:803–812. 2001.
- Lucchesi, A.A. **Processos fisiológicos da cultura da cana-de-açúcar (Saccharum spp)**. Boletim Técnico ESALQ/CENA, n.7, p.1-50, 1995.
- Luna, J.E.O. **Algoritmos EM para aprendizagem de redes bayesianas a partir de dados incompletos**. Dissertação de Mestrado (Ciência da Computação), Departamento de Computação e Estatística, Centro de Ciências Exatas e Tecnologia, Universidade Federal do Mato Grosso do Sul, 2004.
- Lunn J.E.; ApRees T. **Apparent equilibrium constant and mass-action ratio for sucrose-phosphate synthase in seeds of Pisum sativum**. Biochemical Journal 267, 739–743. 1990.
- Magalhães, A.C.N. **Ecofisiologia da cana-de-açúcar: aspectos do metabolismo do carbono da planta**. In: Castro, P. R. C.; Ferreira, S. O.; Yamada, T. (Coord.) Ecofisiologia da Produção. Piracicaba: Associação Brasileira para pesquisa do Potássio e do Fósforo. p. 113-118. 1987.

- Mantelatto, P.E. **Estudo do processo de cristalização de soluções impuras de sacarose de cana-de-açúcar por resfriamento.** Dissertação (Mestrado) – Universidade Federal de São Carlos, São Carlos, SP, 2005.
- Marques, D. **Guia da cana.** Conselho de Informações sobre Biotecnologia – CIB. São Paulo, SP. 2009.
- Martin, W.; Schnarrenberger, C. **The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis.** Curr.Genet. 32, 1–8. 1997.
- McCormick, A.J. **Sink regulation of photosynthesis in sugarcane.** PhD Thesis. School of Biological and Conservation Sciences, University Of KwaZulu-Natal, Durban, South Africa, July, 2007.
- McCormick, A.J.; Cramer, M.D.; Watt, D.A. **Differential expression of genes in the leaves of sugarcane in response to sugar accumulation.** Tropical Plant Biology, v.1, p.142-158, 2008.
- Mendoza, L.; Alvarez-Buylla, E.R. **Dynamics of the genetic regulatory network for Arabidopsis thaliana flower morphogenesis.** Journal of Theoretical Biology, 193, 307-319. 1998.
- Mestl, T.; Plahte, E.; Omholt, S.W. **A mathematical framework for describing and analysing gene regulatory networks.** J. Theor. Biol., 176, 291–300. 1995.
- Ming, R.; Liu, S.C.; Moore, P.H.; Irvine, J.E.; Paterson, A.H. **QTL Analysis in a complex autopolyploid: genetic control of sugar content in sugarcane.** Genome Research, v. 11, p. 2075-2084, 2001.
- Moler, E.J.; Radisky, D.C.; Mian, I.S. **Integrating naive Bayes models and external knowledge to examine copper and iron homeostasis in S. cerevisiae.** Physiol. Genom., 4, 127-135. 2000.
- Moore, B.; Zhou, L.; Rolland, F.; Hall, Q.; Cheng, W.H.; Liu, Y.X.; Hwang, I.; Jones, T.; Sheen, J. **Role of the Arabidopsis Glucose Sensor HXK1 in Nutrient, Light, and Hormonal Signaling.** Science 300:332-336. 2003.
- Moore, P.H.; Botha, F.C.; Furbank, R.T.; Grof, C.P.L. **Potential for overcoming physio-biochemical limits to sucrose accumulation.** In: Keating BA, Wilson JR, editors. Intensive sugarcane production: meeting the challenges beyond 2000. Wallingford: CAB International. p. 141-155. 1997.

- Murphy, K.; Mian, S. **Modelling Gene Expression Data using Dynamic Bayesian Networks**. Computer Science Division, University of California, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, 12 p., 1999.
- Mutwil, M.; Usadel, B.; Schütte, M.; Loraine, A.; Ebenhöf, O.; Persson, S. **Assembly of an Interactive Correlation Network for the Arabidopsis Genome Using a Novel Heuristic Clustering Algorithm**. *Plant Physiology*, v. 152, p. 29-43, 2010.
- Neapolitan, R. E. **Learning bayesian networks**. Prentice Hall, Harlow, 2003.
- Nogueira, F.T.S.; De Rosa, V.E. Jr; Menossi, M.; Ulian, E.C.; Arruda, P. **RNA expression profiles and data mining of sugarcane response to low temperature**. *Plant Physiol* 132:1811–1824. 2003.
- Pal, R.; Datta, A.; Bittner, M.L.; Dougherty, E.R. **Intervention in context-sensitive probabilistic Boolean networks**. *Bioinformatics*, 21(7), 1211-1218. 2005.
- Palhares, A.C. **Estudos genéticos em cana-de-açúcar**. LGN 5799 - Seminários de "Genética e melhoramento de plantas". Departamento de Genética, Esalq/USP. Piracicaba/SP, 2009.
- Papini-Terzi, F.S.; Rocha, F.R.; Vencio, R.Z.N.; Felix, J.M.; Branco, D.S.; Waclawovsky, A.J.; Del Bem, L.E.V.; Lembke, C.G.; Costa, M.D.L.; Nishiyama, M.Y. Jr; Vicentini, R.; Vicentz, M.G.A.; Ulian, E.C.; Menossi, M.; Souza, G.M. **Sugarcane genes associated with sucrose content**. *BMC Genomics* 10:120. 2009.
- Papini-Terzi, F.S.; Rocha, F.R.; Vencio,Z.; Oliveira, K.C.; Felix, J.M.; Vicentini, R.; Rocha, C.S.; Simoes, A.C.; Ulian, E.C.; Zingaretti Mauro, S.M.; Da Silva, A.M.; de Braganca Pereira, C.A.; Menossi, M.; Mendes Souza, G. **Transcription profiling of signal transduction-related genes in sugarcane tissues**. *DNA Res* 12:27–38. 2005.
- Paul, M. J.; Foyer, C. **Sink regulation of photosynthesis**. *Journal Experimental of Botany*, London, v.52, n. 360,p.1.383-1.400, 2001.
- Pearl, J. **Causality: Models, reasoning, and inference**. Cambridge University Press, New York, Cambridge, UK. 2000.
- Pearl, J. **Probabilistic reasoning in intelligent systems: networks of plausible inference**. Morgan Kaufmann. 1988.
- Pego, J.V.; Kortsee, A.J.; Huijser, C.; Smeekens, S.C.M. **Photosynthesis, sugars and the regulation of gene expression**. *Journal of Experimental Botany* 51: 407–416. 2000.

- Picoli, M.C.A. **Estimativa da produtividade agrícola da cana-de-açúcar utilizando agregados de redes neurais artificiais: estudo de caso Usina Catanduva**. Dissertação (Mestrado em Sensoriamento Remoto) – Instituto de Pesquisas Espaciais, São José dos Campos, 2007.
- Pierce, B. A. **Genética: um enfoque conceitual**. Rio de Janeiro: Guanabara Koogan, p. 423, 546-548, 339. 2004.
- Purcell, P.C.; Smith, A.M.; Halford, N.G. **Antisense expression of a sucrose non-fermenting-1-related protein kinase sequence in potato results in decreased expression of sucrose synthase in tubers and loss of sucrose-inducibility of sucrose synthase transcripts in leaves**. Plant J. 14, 195–202. 1998.
- R Development Core Team. **R: A language and environment for statistical computing**. Disponível em: <http://www.r-project.org>. 2003.
- Ramon, M.; Rolland, F.; Sheen, J. **Sugar sensing and signaling**. The Arabidopsis Book. The American Society of Plant Biologists. pp. 1–22. 2008.
- Ribeiro, C. **Análise de expressão gênica e metabolismo primário em cana-de-açúcar: sinalização de açúcares**. Dissertação (Mestrado em Ciências) - Escola Superior de Agricultura "Luiz de Queiroz" - ESALQ/USP. 2012.
- Riesmeier, J.W.; Willmitzer, L.; Frommer, W.B. **Evidence for an essential role of the sucrose transporter in phloem loading and assimilate partitioning**. EMBO J. 13: 1-7. 1994.
- Ripoli, T. C. C.; Ripoli, M. L. C.; Casagrandi, D. V.; Ide, B. Y. **Plantio de cana-de-açúcar: Estado da arte**. 2.ed. Piracicaba: Ed. dos Autores, 198p. 2007.
- Rocha F.R.; Papini-Terzi F.S.; Nishiyama M.Y. Jr; Vencio R.Z.N.; Vicentini R.; Duarte R.D.C.; De Rosa V.E. Jr; Vinagre F.; Barsalobres C; Medeiros A.H.; Rodrigues F.A.; Ulian E.C.; Zingaretti S.M.; Galbiatti J.A.; Almeida R.S.; Figueira A.V.O.; Hemerly A.S.; Silva-Fihlo M.C.; Menossi M.; Souza G.M. **Signal transduction-related responses to phytohormones and environmental challenges in sugarcane**. BMC Genomics 8:71, 2007.
- Rodrigues, J. D. **Fisiologia da cana-de-açúcar**. Botucatu: Instituto de Biociências, 101 p. Apostila. 1995.

- Roitsch, T.; Bittner, M.; Godt, D.E. **Induction of apoplastic invertase of *Chenopodium rubrum* by d-glucose and a d-glucose analog and tissue-specific expression suggest a role in sink source regulation.** *Plant Physiol* 108: 285–294. 1995.
- Rolland F.; Moore B.; Sheen J. **Sugar sensing and signaling in plants.** *The Plant Cell* 14, S185-S205, 2002.
- Rolland, F.; Baena-Gonzalez, E.; Sheen, J. **Sugar sensing and signaling in plants: Conserved and novel mechanisms.** *Annu. Rev. Plant Biol.*, 57, pp. 675–709. 2006.
- Ruggeri, F.; Kenett, R.S.; Faltin, F. **Encyclopedia of Statistics in Quality and Reliability.** John Wiley and Sons, Chichester, UK. 2007.
- Santos, N. B. **Identificação dos fatores críticos da colheita mecanizada de cana-de-açúcar (*Saccharum spp.*).** 86p. Dissertação (Mestrado em Máquinas Agrícolas) - Escola Superior de Agricultura “Luiz de Queiroz”, 2011.
- Schäfer, J.; Strimmer, K. **An empirical Bayes approach to inferring large-scale gene associations networks.** *Bioinformatics*, v. 21, n. 6, p. 754-764, 2005.
- Sebastiani, P.; Abad, M.M.; Ramoni, M.F. **Bayesian networks for genomic analysis.** *Genomic Signal Processing and Statistics*, 281-320. 2004.
- Segal, E.; Shapira, M.; Regev, A.; Pe’er, D.; Botstein, D.; Koller, D.; Friedman, N. **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nature Genetics*, v. 34, n. 2, p.166-176, 2003.
- Shah, A.; Woof, P. **Python Environment for bayesian learning: Inferring the structure of bayesian networks from knowledge and data.** *Journal of Machine Learning Research*, v. 10, p. 159-162, 2009.
- Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 13:2498-2504. 2003.
- Sheen, J. **Feedback control of gene expression.** *Photosynthesis Research* 39: 427–438. 1994.
- Sheen, J. **Metabolic repression of transcription in higher plants.** *The Plant Cell* 2: 1027–1038. 1990.
- Shmulevich, I.; Dougherty, E. R.; Kim, S.; Zhang, W. **Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks.** *Bioinformatics*, v. 18, n. 2, p. 261-274, 2002.

- Silva Júnior, J. R.; Bastos, G. Q.; Souza, A. E. R. **Desempenho agroindustrial das novas variedades de cana-de-açúcar (*Saccharum*spp), em âmbito regional**. IX Jornada de ensino, pesquisa e extensão. Recife, 2010.
- Silva, M.S.L. **Efeito do cultivo da cana-de-açúcar em propriedades do solo de tabuleiro do estado de Alagoas**. Recife, Universidade Federal de Pernambuco, 106p. (Tese de Mestrado). 1989.
- Sladeczek, J.; Hartemink, A.J.; Robinson, J. **Bayesian Network Inference with Java Objects, Version 2.2 - User Guide**. Disponível em: <https://www.cs.duke.edu/~amink/software/banjo/documentation/banjo.user.pdf>. 2005.
- Smeekens S. **Sugar-induced signal transduction in plants**. Annu Rev Plant Physiol Plant Mol Biol 51:49–81, 2000.
- Smeekens S.; Rook F. **Sugar sensing and sugar-mediated signal transduction in plants**. Plant Physiol. 115:7–13, 1997.
- Smolen, P.; Baxter, D.; Byrne, J. **Mathematical modeling of gene networks**. Neuron, 26, 567-580. 2000.
- Spiegel, M.R.; Schiller, J.; Srinivasan, R.A. **Probability and Statistics, Schaum's outline series**, McGraw-Hill, New York. 2000.
- Srinivasasainagendra V.; Page G.P.; Mehta T.; Coulibaly I.; Loraine A.E. **CressExpress: a tool for large-scale mining of expression data from Arabidopsis**. Plant Physiol. 147 1004–1016. 2008.
- Stupiello, J. P. **A cana-de-açúcar como matéria-prima**. In: PARANHOS, S. B. Cana-de-açúcar: cultivo e utilização. v. 2. Campinas: Fundação Cargill, p. 30-51, 1987.
- Tamada, Y.; Kim, S.; Bannai, H.; Imoto, S.; Tashiro, K.; Kuhara, S.; Miyano, S. **Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection**. Bioinformatics, 19(suppl 2), ii227-ii236. 2003.
- Tiessen, A.; Prescha, K.; Branscheid, A.; Palacios, N.; McKibbin, R.; Halford, N.G.; Geigenberger, P. **Evidence that SNF1-related kinase and hexokinase are involved in separate sugar-signalling pathways modulating post-translational redox activation of ADP-glucose pyrophosphorylase in potato tubers**. Plant J. 34, 490–500. 2003.
- Toh, H.; Horimoto, K. **Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling**. Bioinformatics, 18, 287–297. 2002.

- Toppa, E.V.B.; Jadoski, C.J.; Julianetti, A.; Hulshof, T.; Ono, E.O. **Aspectos da fisiologia de produção da cana-de-açúcar (*Saccharum Officinarum* L.)**. Revista Brasileira de Tecnologia Aplicada nas Ciências Agrárias, v. 3, n. 3, 2011.
- Turgeon, R. **The sink-source transition in leaves**. Annu. Rev. Plant Physiol. Plant Mol. Biol. 40, 119-138.1989.
- van Someren, E.; Wessels, L.; Backer, E.; Reinders, M. **Genetic network modeling**. Pharmacogenomics, 3, 507–525. 2002.
- Vasconcelos, A.C.M. **Desenvolvimento do sistema radicular e da parte aérea de socas de cana-de-açúcar sob dois sistemas de colheita: crua mecanizada e queimada manual**. 140f. Tese (Doutorado em Produção Vegetal) – Faculdade de Ciências Agrárias e Veterinárias - Universidade Estadual Paulista, Jaboticabal, 2002.
- Volinia, S.; Galasso, M.; Costinean, S.; Tagliavini, L.; Gamberoni, G.; Drusco, A.; Rosenberg, A. **Reprogramming of miRNA networks in cancer and leukemia**. Genome research, 20(5), 589-599. 2010.
- Watanabe, Y.; Seno, S.; Takenaka, Y.; Hideo Matsuda. **An estimation method for inference of gene regulatory net-work using Bayesian network with uniting partial problems**. BMC Genomics, v. 13 <http://www.biomedcentral.com/1471-2164/13/S1/S12>, 2012.
- Weaver, D.C.; Workman, C.T.; Stormo, G.D. **Modeling regulatory networks with weight matrices**. Pac. Symp. Biocomput., 4, 112–123. 1999.
- Werhli, A. V. **Reconstruction of gene regulatory networks from post genomic data**. PhD Thesis. Institute for Adaptive and Neural Computation, School of Informatics. University of Edinburgh. 2007.
- Wiese, A.; Elzinga, N.; Wobbes, B.; Smeekens, S. **A Conserved Upstream Open Reading Frame Mediates Sucrose-Induced Repression of Translation**. Plant Cell 16:1717-1729. 2004.
- Wilczyński, B.; Dojer, N. **BNFinder: exact and efficient method for learning Bayesian networks**. Bioinformatics, 25(2), 286-287. 2009.