Dinâmica da fermentação alcoólica: aplicação de redes booleanas na dinâmica da expressão gênica em linhagens de *Saccharomyces cerevisiae* durante o processo fermentativo

Melline Fontes Noronha

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Programa: Mestrado em Bioinformática

Orientador: Prof. Dr. Ronaldo Fumio Hashimoto

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da
CNPq

São Paulo, novembro de 2012

Dinâmica da fermentação alcoólica: aplicação de redes booleanas na dinâmica da expressão gênica em linhagens de *Saccharomyces cerevisiae* durante o processo fermentativo

Este trabalho trata-se da versão revisada da aluna (Melline Fontes Noronha).

“O sucesso é ir de fracasso
em fracasso sem perder
entusiasmo.” *Winston
Churchill*

Agradecimento

Primeiramente, agradeço à minha mãe e à minha avó Lais por todo apoio durante realização deste curso, inclusive pela compreensão por todo tempo que estive fora. Agradeço também ao meu pai e Kátia por todo carinho e apoio. Agradeço também à minha querida irmazinha Bia que, apesar da pouca idade, ser bastante compreensiva com sua irmã sempre ausente.

Ao meu namorado, Osvaldo, por tem me ajudando bastante nesse trabalho. Sempre muito compreensivo e companheiro durante esse tempo. Não tenho palavras para agradecer.

Ao meu orientador, prof. Ronaldo Hashimoto, por todos os ensinamentos e paciência em me passar todo o seu conhecimento. Obrigada pelos tempos que reservou pra mim todos esses anos.

Aos meus colegas do Laboratório de Gênomica e Expressão, em especial aos colegas da Bioinformática: Marcelo, Ramon, Herai, Leandro e Luciana, e também ao Osmar, por me passar vários conhecimentos biológicos e disponibilizar seus dados na realização desse trabalho. Gostaria de agradecer também ao prof. Gonçalo Pereira por me dar a oportunidade de aprender muito sobre a área durante todo esse tempo em seu laboratório. O conhecimento que adquiri lá não pode ser substituído por nenhuma bibliografia.

Gostaria de agradecer também à secretária do IME, a Patrícia, por todo o seu apoio durante todo o curso e também pelos papos descontraídos na USP. Com certeza, ganhei uma amiga e tenho muito de agradecer a ela.

Gostaria de agradecer a alguns amigos e amigas que me apoiaram durante todo esse tempo, em especial à minha grande amiga Claryanne. Aos colegas de curso do IME. Aos meus eternos professores, Methanias e Sara Cuadros. Enfim, a todos o meu muito obrigado.

Obrigada também ao Cnpq por financiar esse trabalho.

Resumo

Na busca por soluções que maximizem a produção de etanol, o melhoramento genético de diferentes linhagens de levedura tornou-se foco de investigação em diversos centros de pesquisa. Com o recente sequenciamento de uma linhagem selvagem utilizada nas usinas sucroalcooleiras brasileiras, a linhagem PE-2 da espécie *Saccharomyces cerevisiae*, surgiu o interesse em estudar sua dinâmica durante o processo de fermentação a fim de encontrar aspectos que possam explicar como estas se tornaram mais adaptadas às dornas de fermentação mantendo a alta produtividade de bioetanol. A partir da análise transcricional da linhagem PE-2, Buscamos por métodos de inferência de redes que possam representar a dinâmica dessa levedura. Propomos nesse trabalho a modelagem de dados experimentais temporais das linhagens PE-2 e S288c (utilizada como referência) baseado em um modelo de Redes Booleanas. Trata-se de um modelo onde convertemos dados contínuos em dados discretos (0 or 1) no qual, de acordo com restrições ditadas pelo modelo, são inferidas redes que representem interações gênicas ao longo do tempo baseados nas amostras temporais. Conseguimos modelar, com sucesso, algumas redes utilizando conjuntos com 11 e 12 genes relacionados a genes pertencentes à via da glicólise e fermentação da levedura.

Palavras-chave: bioetanol, levedura, linhagem PE-2, redes de regulação gênica, modelo booleano.

Abstract

Ethanol production improvements give rise to the breeding of yeast strains, that became the investigation focus in several research centers. Recently, a wild strain used in Brazilian sugarcane industry was sequenced, the PE-2 strain of *Saccharomyces cerevisiae*, and this event brought an interest in studying the dynamics of the fermentation of this strain in order to understand which aspects this strain become more adapted to the fermentation conditions, maintaining a high capacity to produce bioethanol. From the analysis of transcriptional strain PE-2, we seek for inference networks methods that can represent the dynamics of this yeast. In this work, we model an experimental temporal data of strain PE-2 and strain S288c (used as a reference) based on Boolean networks model. In this model, the data are converted from continuous into discrete data (0 or 1) and, based on constraints rules of Boolean Network model, networks are inferred to represent gene interactions over time based on temporal data. We successfully model networks using a set with 11 and 12 genes related to yeast glycolysis and fermentation pathways.

Key-words: bioethanol, yeast, PE-2 strain, genes regulatory networks Boolean Network.

Lista de Abreviaturas

AA	Ausente em Ambas as matrizes
BIOGRID	<i>Biological General Repository for Interaction Datasets</i>
BN	Redes booleanas (<i>Boolean Networks</i>)
BNp	Redes booleanas com perturbação (<i>Boolean Networks with perturbation</i>)
cDNA	DNA complementar (<i>complementary DNA</i>)
DNA	Ácido desoxirribonucleico (<i>Deoxyribonucleic acid</i>)
GEO	<i>Gene Expression Omnibus</i>
GRN	Redes de regulação gênica (<i>Gene regulatory networks</i>)
KEGG	<i>Kyoto Encyclopedia of Genes and Genomes</i>
mRNA	Ácido ribonucleico mensageiro (<i>Messenger ribonucleic acid</i>)
NCBI	<i>National Center for Biotechnology Information</i>
OD	Densidade óptica (<i>Optical density</i>)
PA	Presente em Ambas as matrizes
PP	Presente na Primeira Matriz
PPV	<i>Positive Predictive Value</i>
OS	Presente na Segunda matriz
RNA	Ácido ribonucleico (<i>Ribonucleic acid</i>)
RNA-seq	<i>RNA-sequencing</i>
SAGE	<i>Serial analysis of gene expression</i>
SGD	<i>Saccharomyces Genome Database</i>
SNP	Polimorfismo de único nucleotídeo (<i>single-nucleotide polymorphism</i>)

Lista de Símbolos

A	Matriz de regulação
a_{ij}	Aresta de regulação do gene x_j para x_i
D	Matriz de dados de entrada
F	Funções booleanas
G	Rede booleana
H	Entropia
μ	Distância
k_i	Conectividade de entrada para o gene x_i
S	Vetor de estados
t	Tempo
X	Conjunto de genes
$x_i(t)$	Estado de uma rede no tempo t

Lista de Figuras

Figura 1. Matriz de produção mundial de etanol (Jensen, 2012)	2
Figura 2. Experimento fermentativo e estresse com SC288c e PE-2	4
Figura 3. Ilustração simplificada da Transcrição e Tradução.....	8
Figura 4. Ilustração de interações gênicas e sua representação em uma rede de regulação gênica	10
Figura 5. Diagrama de transição de estados de uma rede Booleana (Shmulevich, et al., 2002).	12
Figura 6. Rede de regulação da amostra da Equação 3	13
Figura 7. Diagrama de estados para a rede da Figura 6	14
Figura 8. Estados estacionários encontrados na via do ciclo celular. Fonte: (Li, et al., 2004)	16
Figura 9. O diagrama de estados da maior bacia de atração encontrada no ciclo celular. Fonte: (Li, et al., 2004)	17
Figura 10. Ilustração dos passos da técnica de microarray double-channel.....	20
Figura 11. Tabela das concentrações de glicose do experimento de (Derisi, et al., 1997)	22
Figura 12. Gráfico representando as variações nas concentrações de etanol e açúcares redutores presentes na fermentação em escala de laboratório com dados obtidos da Tabela 3. (Netto, 2010).....	23
Figura 13. Concentração de Etanol para as amostras coletas na Usina Nova América (Netto, 2010).....	24
Figura 14. Exemplos de limiar para binarização dos dados	26
Figura 15. Esquema de janelas para o experimento de S288c	28
Figura 16. Genes iniciais na via da glicólise e da fermentação (Conant, et al., 2007) .	29
Figura 17. Sequência de processos metabólico dos 19 genes.....	31
Figura 18. Esquema de seleção das bibliotecas de dados.....	33
Figura 19. Esquema de seleção das bibliotecas de dados até gerar as redes aleatórias	34
Figura 20. Gráficos de distribuição de entropia para cada biblioteca	39
Figura 21. Diagrama de estados da rede de RNA-seq com 11 genes.....	40
Figura 22. Dinâmica dos genes do conjunto de 11 genes nas amostras de RNA-seq em PE-2.....	41
Figura 23. Diagrama de estados da rede de microarray em PE-2 com 11 genes.....	42

Figura 24. Diagrama de estados da rede da Janela 1 em S288c com 11 genes.....	43
Figura 25. Diagrama de estados da rede da Janela 2 em S288c com 11 genes.....	44
Figura 26. Diagrama de estados da rede da Janela3 em S288c com 11 genes.....	45
Figura 27. Diagrama de estados da rede da Janela4 em S288c com 11 genes.....	46
Figura 28. Comparativo dos estados de RNA-seq PE-2 e dos atratores das janelas de S288c.	47
Figura 29. Comparativo dos atratores de RNA-seq PE-2 e da janela 4 em S288c.....	48
Figura 30. Gráfico comparativo usando medida de distância baseada em regras entre as bibliotecas do conjunto de 11 genes.	49
Figura 31. Sobreposição de estados da Janela 2 na rede de RNA-seq PE-2.	50
Figura 32. Rede de regulação de RNA-seq PE-2 do conjunto de 11 genes.	51
Figura 33. Rede de regulação da Janela 2 de S288c do conjunto de 11 genes.....	52
Figura 34. Maior bacia da rede do conjunto 12 para dados de RNA-seq PE-2.....	53
Figura 35. Maior bacia da rede do conjunto 12 para dados de microarray em PE-2. ..	53
Figura 36. Maior bacia da rede do conjunto 12 para dados da Janela 1 em S288c. ...	54
Figura 37. Maior bacia da rede do conjunto 12 para dados da Janela 2 em S288c. ...	54
Figura 38. Maior bacia da rede do conjunto 12 para dados da Janela 3 em S288c. ...	55
Figura 39. Maior bacia da rede do conjunto 12 para dados da Janela 4 em S288c. ...	55
Figura 40. Comparação dos estados das redes de RNA-seq PE-2 e os atratores das redes de S288c no conjunto de 11 genes e no conjunto 12 (conjunto com 12 genes).	56
Figura 41. Gráfico comparativo usando medida de distância baseada em regras entre as bibliotecas do conjunto 12 com 12 genes.....	57
Figura 42. Sobreposição de estados correspondentes aos dados de entrada das Janelas 1 e 3 na rede de RNA-seq PE-2.	58

Lista de Tabelas

Tabela 1. Conjunto de soluções dado o conjunto de amostras da matriz <i>A</i> (Equação 4)	15
Tabela 2. Exemplo de matrizes soluções para conjunto da Tabela 1	15
Tabela 3. Concentrações de etanol (°Gl) e açúcares residuais no mosto do experimento de fermentação em escala de laboratório. (Netto, 2010)	23
Tabela 4. Tabela correspondência genes/alias para os 11 genes iniciais	30
Tabela 5. Tabela com nomes e alias dos 19 genes selecionados.....	30
Tabela 6. Matriz confusão para medir a distância topológica das redes identificadas, (PA = presente em ambas, PP, presente na primeira rede apenas, OS = presente na segunda rede apenas e AA = ausente em ambas)	37
Tabela 7. Tabela comparativa usando medida de similaridade entre as bibliotecas do conjunto de 11 genes.....	49
Tabela 8. Tabela comparativa usando medida de similaridade entre as bibliotecas do conjunto 12 com 12 genes.....	57

Sumário

LISTA DE ABREVIATURAS	VI
LISTA DE SÍMBOLOS	VII
LISTA DE FIGURAS	VIII
LISTA DE TABELAS.....	X
INTRODUÇÃO	1
1.1. O BIOETANOL.....	1
1.1.1. Leveduras.....	2
1.2. MOTIVAÇÃO	3
1.3. OBJETIVOS	6
1.4. ORGANIZAÇÃO	6
FUNDAMENTAÇÃO TEÓRICA.....	8
2.1. EXPRESSÃO GÊNICA	8
2.2. INFERÊNCIA DE REDES GÊNICAS	9
2.3. O MODELO DE REDES BOOLEANAS.....	11
2.4. REDES BOOLEANAS RESTRITAS	12
2.5. O ALGORITMO <i>TANGINFERENCE</i>	14
2.6. MODELO DE INFERÊNCIA POR REDES BOOLEANAS.....	16
MATERIAIS E MÉTODOS.....	18
3.1 O PROBLEMA BIOLÓGICO.....	18
3.2 EXPRESSÃO GÊNICA	19
3.2.1 <i>As técnicas de análise de expressão gênica utilizadas</i>	19
3.3 OS CONJUNTOS DE DADOS	21
3.3.1 <i>Dados da linhagem S288c em microarray</i>	21
3.3.2 <i>Dados de PE-2 em microarray</i>	22
3.3.3 <i>Dados de PE-2 em RNA-seq</i>	24
3.4 TRATAMENTO DOS DADOS BIOLÓGICOS	25
3.4.1 <i>Normalização dos dados</i>	25
3.4.2 <i>Binarização dos dados</i>	26
3.4.3 <i>Janelas</i>	27
3.5 SELEÇÃO DOS CONJUNTOS DE GENES.....	28
3.5.1 <i>Conjunto inicial de genes</i>	28
3.5.2 <i>Conjuntos de genes utilizados</i>	29
3.5.3 <i>Breve revisão sobre os genes utilizados</i>	31

3.5.4.	<i>Esquema dos conjuntos de genes</i>	32
3.5.5.	<i>Seleção das bibliotecas de dados</i>	33
3.6.	GERANDO REDES CONSISTENTES	33
3.7.	ANÁLISE COMPARATIVA DAS REDES	34
3.7.1.	<i>Entropia</i>	34
3.7.2.	<i>Medida de distância lógica</i>	35
3.7.3.	<i>Medida de distância topológica</i>	36
RESULTADOS E DISCUSSÕES		38
4.1.	REDES COM APENAS 11 GENES	38
4.2.	RESULTADO DAS REDES UTILIZANDO 12 GENES	52
4.3.	OUTROS CONJUNTOS DE GENES	58
4.4.	DISCUSSÕES	59
CONCLUSÃO		60
APÊNDICE A		62
APÊNDICE B		64
REFERÊNCIAS		66

Capítulo 1

Introdução

1.1. O Bioetanol

Por muitos anos, as fontes de energia não-renováveis como petróleo e o carvão vem sendo utilizadas como principais fontes de energia no mundo. Além de não-renováveis, essas fontes de energia liberam gases poluentes na atmosfera contribuindo fortemente com o efeito estufa. Recentemente, iniciou-se uma campanha mundial para substituir essas fontes de energia por fontes renováveis e limpas, principalmente as derivadas de biomassa. O Etanol é uma dessas fontes de energia alternativas com grande potencial principalmente no Brasil.

O Etanol (álcool etílico) é uma substância obtida através de fermentação de açúcar, amido ou celulose, sendo a forma de melhor aproveitamento o da cana-de-açúcar. No Brasil, a fermentação da cana-de-açúcar é a mais economicamente viável devido aos baixos custos de obtenção da matéria-prima e à fácil adaptação ao clima e ao solo fértil. Atualmente, o Brasil é o mais eficiente produtor de etanol no mundo e compete diretamente com os Estados Unidos (Figura 1). Essa eficiência brasileira se justifica pela sua enorme capacidade produtiva, e também graças ao Proálcool (Programa Nacional do Álcool) iniciado na década de 70, que desencadeou o desenvolvimento de toda uma tecnologia de produção, distribuição e utilização de etanol (Netto, 2010). Além disso, o governo brasileiro vem investindo em pesquisas para melhoramento da produção de biocombustíveis apoiando iniciativas como o BIOEN (programa da Fapesp em bioenergia), CTBE (Laboratório Nacional de Ciência e Tecnologia do Bioetanol), entre outras (Pagliari, 2010).

Figura 1. Matriz de produção mundial de etanol (Jensen, 2012)

Atualmente, os esforços de pesquisa no setor estão focados em aumentar o rendimento e a produtividade para que o etanol se torne ainda mais competitivo internacionalmente. Algumas das principais iniciativas na área visam à busca ou produção de melhores variedades de cana-de-açúcar, incluindo plantas transgênicas, a expansão da área agrícola, inovações na linha de produção das usinas e seleção de linhagens de leveduras mais eficientes.

1.1.1. Leveduras

As leveduras são microrganismos responsáveis pela transformação da glicose (proveniente da cana-de-açúcar) em etanol. Ao longo dos anos, o melhoramento de linhagens de leveduras para uso industrial foi foco de diversas pesquisas, uma vez que estas precisam apresentar características que permitam a sua sobrevivência em ambiente altamente estressante como o encontrado nas dornas de fermentação. Dentre estas características destacam-se: estabilidade genética (capacidade de conservar informação genética ao longo das gerações), alta tolerância ao etanol, capacidade de produzir elevadas concentrações de etanol, osmotolerância (habilidade de fermentar soluções concentradas de carboidratos), viabilidade celular, tolerância a repetidos ciclos fermentativos e tolerância à temperatura.

A variabilidade genética em organismos é um processo natural ao longo da escala evolutiva e, nas leveduras, essa variabilidade gerou aproximadamente 350 espécies distintas. Estas espécies se diferenciam entre

si geneticamente por características como inserção de íntrons, mudança de *códon usage*, diploidia, duplicação de genoma, entre outras diferenciações. Essas variações fizeram com que cada espécie tenha sua particularidade variando desde sua capacidade de utilização como agentes de fermentação alcoólica, utilização na produção de substâncias de interesse industrial, e até mesmo prejudicando a saúde de animais e seres humanos e comprometendo a conservação de alimentos.

A espécie utilizada nas usinas sucroalcooleiras para produção do bioetanol é a *Saccharomyces cerevisiae*. Dentre as suas linhagens também é possível observar particularidades que as fazem ter diferentes potenciais fermentativos. Diversos trabalhos foram realizados analisando comparativamente linhagens de *Saccharomyces cerevisiae* quanto à produção de etanol, resistência a altas temperaturas, entre outros aspectos (Brooks, 2008; Population genomics of domestic and wild yeasts, 2009). No Brasil, a linhagem PE-2 utilizada na produção de etanol foi a que se mostrou melhor adaptada e com alto desempenho na grande maioria das indústrias do estado de São Paulo. Nas últimas safras, juntamente com a linhagem CAT-1, essa levedura foi utilizada por cerca de 150 usinas, que correspondem por 60% do álcool combustível produzido no Brasil (Netto, 2010). Em testes preliminares de fermentação, observou-se que esta linhagem produz cerca de 50% mais etanol do que a linhagem S288c (usada como referência). Além disso, esta linhagem mostra-se mais tolerante a altas temperaturas e ao estresse (Argueso, et al., 2009)(Figura 1).

1.2. Motivação

Em um projeto em andamento do Laboratório de genômica e expressão do Instituto de biologia da UNICAMP, no qual a aluna é bioinformata colaboradora, em parceria com a Duke University Medical Center – EUA, foi realizado o sequenciamento de cepas de leveduras da linhagem PE-2 (Argueso, et al., 2009). Esses dados já estão disponíveis publicamente no banco de dados do NCBI (<http://www.ncbi.nlm.nih.gov/>). Além do sequenciamento, montagem e disponibilização desses dados, foram realizadas análises de predição de genes e análise de *SNPs* (polimorfismos). Este trabalho teve grande impacto científico uma vez que foi a primeira vez que uma

Figura 2. Experimento fermentativo e estresse com S288c e PE-2. (A) Experimento fermentativo: produção de etanol entre linhagem referência S288c e PE-2. (B) Estresse (temperatura e oxidativo): entre linhagem S288c e a PE-2 (Argueso, et al., 2009).

linhagem brasileira utilizada em larga escala na produção de etanol teve sua estrutura gênica descrita.

Com o genoma desta linhagem sequenciado, foi possível observar algumas mudanças genotípicas em relação à linhagem S288c, incluindo a inserção e deleção de alguns genes. Avançando nesse projeto, surgiu o interesse em conhecer o perfil de expressão gênica (transcriptoma) dessa linhagem durante a fermentação e conhecer as possíveis mudanças metabólicas que as tornaram mais eficientes quanto à produção do bioetanol.

Medir o nível de expressão gênica de um gene é quantificar os transcritos gerados por esse gene, dada uma condição/instante de tempo. A análise de expressão gênica em larga escala (vários genes simultaneamente) pode ser realizada utilizando tecnologias como *microarray* (Schena, et al., 1995), SAGE (Velculescu, et al., 1995) e *RNA-sequencing* (Wang, et al., 2009). Devido à quantidade de dados gerados por essas tecnologias, normalmente são utilizadas ferramentas matemáticas e computacionais a fim de facilitar a análise desses dados. Projetos envolvendo transcriptômica, em geral, visam não só conhecer o nível de expressão gênica de um conjunto de genes como também entender as relações entre os genes do sistema biológico em uma ou mais condições estudadas. Nesse sentido, alguns modelos matemáticos foram propostos para modelar sistemas biológicos a partir de dados de expressão

gênica. Esse problema é conhecido como inferência de redes de regulação gênica (GRN - *gene regulatory networks*).

A inferência de redes de regulação gênica é um problema bastante abordado em Biologia de Sistemas. Há diversos modelos matemáticos propostos para auxiliar o processo de inferência das redes, sendo um dos mais populares o modelo de redes Booleanas ou BNs (*Boolean networks*) proposto por Kauffman (Kauffman, 1969). Posteriormente este modelo foi aprimorado por (Shmulevich, et al., 2002), dando origem ao modelo de Redes Booleanas Probabilísticas (PBN - do inglês *Probabilistic Boolean Networks*). Existem ainda outros modelos como as redes Bayesianas (Friedman, et al., 2000), inclusive há revisões sobre alguns dos modelos de redes de regulação gênica (Karlebach, et al., 2008; Hecker, et al., 2009).

Nosso foco é no modelo discreto conhecido como redes Booleanas (*Boolean network*). Nesse modelo, os dados de expressão contínuos são transformados em dados discretos com apenas dois valores: 0 ou 1. Por tratar os dados de forma “grosseira” (binária), alguns consideram essa discretização como uma desvantagem para o modelo em relação aos modelos contínuos. Entretanto, essa discretização possibilita uma visão mais simplificada dos dados. Além disso, iremos trabalhar com amostras extraídas utilizando técnicas diferentes, e essa discretização também poderá se tornar uma vantagem por minimizar as possíveis divergências entre as técnicas. Esse modelo de inferência de redes já foi utilizado com sucesso para representar diversos sistemas biológicos como o desenvolvimento da planta *Arabidopsis thaliana* (Espinosa-Soto, et al., 2004), o ciclo celular da levedura (Li, et al., 2004), entre outros.

Baseado no modelo do dogma central da Biologia celular, o nível de expressão gênica medido num instante de tempo é reflexo das interações entre os genes/proteínas naquele instante. Os métodos de inferência de redes tentam propor soluções matemáticas baseados nos valores de expressão encontrados. Porém, inferir redes a partir de expressões gênicas é um problema mal posto (*ill-posed*), pois podem existir várias redes capazes de explicar os dados observados (também pode ocorrer de não existir uma rede). Outro problema do método de inferência de redes são os dados disponíveis. As técnicas de análises de expressão gênica são bastante custosas e, por

consequência, a maioria dos dados possuem poucas amostras temporais disponíveis.

1.3. Objetivos

O objetivo principal desse trabalho é inferir redes gênicas a partir de conjuntos de dados transcrpcionais temporais de levedura da linhagem PE-2 e de dados públicos de levedura da linhagem S288c (usada como referencial). Além disso, é nosso objetivo utilizar metodologias já descritas na literatura para analisar as redes inferidas. O desenvolvimento desse trabalho seguiu por duas abordagens: a biológica e a abordagem matemática e computacional. A abordagem biológica desse trabalho aborda o estudo do processo de fermentação em leveduras, seleção dos conjuntos de genes a serem utilizados e tratamento dos dados experimentais. A abordagem matemática e computacional desse trabalho consiste inicialmente na inferência e no estudo do modelo determinístico de redes Booleanas. Utilizamos e desenvolvemos também ferramentas computacionais baseadas nesse modelo e aplicamos metodologias já descritas em literatura para analisar as redes geradas, a fim de extrair informações biológicas relevantes para o problema da produção do bioetanol.

1.4. Organização

Este trabalho está organizado da seguinte forma:

No **Capítulo 2** apresentamos uma revisão teórica sobre alguns conceitos biológicos e descrevemos alguns conceitos e métodos matemáticos utilizados nesse trabalho. Na seção 2.1 iremos abordar alguns conceitos de expressão gênica e na seção 2.2 iremos introduzir a inferência de redes gênicas. Nas seções 2.3 e 2.4 iremos descrever o modelo de Redes Booleanas e de Redes Booleanas Restritas e também o Problema de Satisfação de Restrições na seção 2.5. Na seção 2.6 iremos descrever um pouco sobre as particularidades do algoritmo utilizado nesse trabalho: *Tanginference* e na seção 2.7 iremos descrever brevemente um trabalho utilizando esse modelo de redes em leveduras.

No **Capítulo 3** iremos abordar os métodos e materiais utilizados tanto biologicamente como matematicamente. Na seção 3.1 e 3.2, descrevemos um pouco sobre o problema biológico e uma breve descrição das técnicas de análise de expressão gênica utilizadas. Nas seções 3.3 e 3.4, descrevemos como foram extraídos, normalizados e tratados os dados biológicos utilizados. Em seguida, na seção 3.5, descrevemos como foram selecionados os genes utilizados nesse trabalho. Já na parte matemática, descrevemos o esquema utilizado para gerar as redes gênicas nesse trabalho (seção 3.6) e como foi realizada a análise comparativa dessas redes (seção 3.7).

No **Capítulo 4** iremos apresentar os resultados e discussões deste trabalho. Inicialmente iremos apresentar os resultados encontrados da análise das redes modeladas utilizando os genes principais da via da fermentação (seção 4.1). Em seguida vamos apresentar mais detalhadamente algumas redes obtidas para os 19 conjuntos de genes propostos (seção 3.5) e apresentar os resultados encontrados nessas redes para cada biblioteca de dados (seção 4.2). E, na seção 4.3 e 4.4, iremos descrever resumidamente a análise geral sobre as redes inferidas.

No **Capítulo 5** iremos apresentar nossas conclusões sobre este trabalho.

Capítulo 2

Fundamentação teórica

2.1. Expressão gênica

O DNA é um composto presente em todos os seres vivos e em alguns vírus. Ele é formado por um conjunto de sequências de nucleotídeos (Adenina, Citosina, Timina, Guanina) onde a informação genética está contida. Algumas regiões no DNA, conhecidas como genes, são codificadas gerando um produto gênico que geralmente é uma proteína (em alguns casos, o produto pode ser um RNA funcional). Quando um gene é codificado, ele é denominado transcrito (*mRNA*) e posteriormente é traduzido em uma proteína. Uma ilustração simplificada pode ser observada na Figura 3.

Figura 3. Ilustração simplificada da Transcrição e Tradução

Apesar de estarem contidos no DNA, os genes não codificam todos os genes a todo o momento. Há mecanismos que podem “ligar” ou “desligar” a expressão de genes de acordo com a sua necessidade (e em casos de organismos multicelulares os genes ainda podem ser específicos ao tipo de célula). Há mecanismos que podem “ligar” ou “desligar” a expressão de genes de acordo com diversos estímulos, por exemplo, em resposta a nutrientes, à estresse, à uma condição ambiental, entre outros. A análise de expressão gênica consiste em analisar o comportamento de um organismo dada uma

condição, ou seja, será feita a quantificação da expressão dos genes numa determinada condição.

Os dados de expressão gênica podem ser de dois tipos: estacionários ou temporais. Nos dados estacionários, uma amostra não possui uma relação de tempo com as outras. Por exemplo, um experimento comparando tecido com câncer e normal. Por outro lado, os dados temporais representam uma análise das expressões ao longo do tempo. Por exemplo, efeito da aplicação de um fármaco ao longo do tempo.

Para realizar uma análise de expressão gênica é necessário utilizar técnicas que possam mensurar o nível de expressão gênica de um organismo. Essas técnicas são capazes de medir a quantidade de *mRNA* presentes em uma amostra. As técnicas de análise de expressão gênicas em larga escala (vários genes simultaneamente) mais conhecidas são *microarray* (Schena, et al., 1995), *SAGE – Serial Analysis of Genes Expression* (Velculescu, et al., 1995) e *RNA-sequencing (Whole Transcriptome Shotgun Sequencing)* (Wang, et al., 2009). Sendo a *RNA-sequencing* (ou RNA-seq) a mais recente.

2.2. Inferência de Redes Gênicas

Os sistemas biológicos são bastante complexos, possuindo diversos mecanismos para controle de funções orgânicas como diferenciação celular, controle de componentes químicos, reações ao ambiente externo, etc. Quando há necessidade de produzir algum composto ou realizar alguma atividade orgânica, um conjunto de genes específicos é ativado. Os genes se relacionam entre si através de suas proteínas ou compostos resultantes. Essas reações podem ser realizadas em cadeia onde o substrato da reação anterior é dependente da seguinte e com ou sem necessidade de outros cofatores. Essa reação em cadeia e seus genes são conhecidos como via metabólica.

Os modelos de Inferência de rede gênicas ajudam a modelar matematicamente fenômenos biológicos. Assim, podemos representar a interação de genes/proteínas/compostos por um diagrama representado por nós; onde arestas entre pares de genes indicam que existe uma interação entre estes genes, ou seja, o produto de um gene afeta o produto do outro. Estas arestas podem ser direcionadas, indicando ativação ou inibição. Dessa forma,

os modelos matemáticos e computacionais de redes de regulação gênica podem auxiliar o estudo das interações entre genes/proteínas de sistemas biológicos, permitindo uma análise macro dos dados biológicos (Higa, 2012). Uma ilustração dessas interações e a sua representação em uma rede de regulação gênica pode ser observada na Figura 4.

Figura 4. Ilustração de interações gênicas e sua representação em uma rede de regulação gênica. Figura adaptada de (HECKER, ET AL., 2009)

Neste trabalho utilizamos um modelo discreto conhecido como redes Booleanas ou simplesmente BN (*Boolean Networks*). Apesar dos dados biológicos serem contínuos, é possível descrever simplificadaamente esses dados de maneira binária como “ligado” e “desligado” ou “ativo” e “inativo”. Uma vantagem em utilizar modelos discretos é a possibilidade de inferir redes a partir de um conjunto pequeno de dados e, em muitos casos, os dados de expressão gênica disponíveis possuem uma pequena quantidade de amostras disponíveis e uma grande quantidade de genes a serem analisados. Além disso, as amostras biológicas utilizadas nesse trabalho foram extraídas utilizando técnicas de análises de expressão gênicas distintas e a discretização desses dados pode se tornar uma vantagem nesse tipo de rede, já que a isso minimizaria as possíveis divergências entre os resultados. Acreditamos que o modelo de redes Booleanas pode nos auxiliar no processo de inferência dessas redes de regulação.

2.3. O modelo de redes Booleanas

Uma rede booleana $G(X, F)$ é definida como um conjunto $X = \{x_1, x_2, \dots, x_n\}$ de n variáveis booleanas e um conjunto $F = \{f_1, f_2, \dots, f_n\}$ de n funções booleanas. Sendo $x_i = 0$ quando inativo e, caso contrário, $x_i = 1$. O valor do gene x_i no instante de tempo $t + 1$ é determinado por uma função booleana $f_i: \{0,1\}^{k_i} \rightarrow \{0,1\}$ associada aos valores das variáveis $x_{j_1(i)}, x_{j_2(i)}, \dots, x_{j_{k(i)}(i)}$ no instante t (Li, et al., 2004; Andrade, 2010). Assim,

$$x_i(t + 1) = f_i(x_{j_1(i)}(t), x_{j_2(i)}(t), \dots, x_{j_{k(i)}(i)}(t)) \quad \text{Equação 1}$$

Neste modelo consideramos que todos os genes são atualizados de forma síncrona através da função F . Um estado de uma rede booleana (BN) no tempo t é um vetor binário $S(t) = (x_1(t), x_2(t), \dots, x_n(t))$. Assim, temos 2^n estados possíveis, $x_0, x_1, \dots, x_{2^n-1}$. Aplicando essas funções booleanas, um estado no tempo t leva a rede para o próximo estado $t + 1$ de forma determinística. A transição sequencial entre os estados representa a dinâmica da rede e pode ser representada graficamente através de um diagrama de transição de estados.

Pelo fato de ser um modelo determinístico e possuir um número finito de estados, alguns estados podem ser visitados de forma cíclica. Estes estados são chamados de atratores e representam os estados estacionários de um sistema dinâmico. Atratores com um único estado são chamados de *singletons*. Os estados que não estão nos atratores são chamados de *estados transientes* e são visitados somente uma única vez. O conjunto de todos os estados que conduzem a este atrator é chamado de *bacia de atração*. Por exemplo, na Figura 5 temos cinco bacias de atração com seis atratores, sendo quatro *singletons*: os atratores *singletons* (00000) e (00100) com 8 estados e os atratores (10011) e (11111) com 2 estados cada; e o atrator de tamanho 2 (11010-11110) com 12 estados.

Figura 5. Diagrama de transição de estados de uma rede Booleana (Shmulevich, et al., 2002).

2.4. Redes Booleanas Restritas

Neste trabalho iremos utilizar um algoritmo baseado no modelo de Redes Booleanas descrito por (Li, et al., 2004) conhecido como Redes Booleanas Restritas. Neste modelo, a função booleana f_i é definida de acordo com uma matriz de regulação $A_{n \times n}$ com $a_{ij} = \{-1, 0, 1\}$, $i, j = 0, \dots, n$ e o valor dos genes x_{ji} no instante t , para $j = 1, 2, \dots, n$, de acordo com a Equação 2.

$$x_i(t+1) = \begin{cases} 1, & \text{se } \sum_{j=1}^n a_{ij}x_j(t) > 0 \\ x_i(t), & \text{se } \sum_{j=1}^n a_{ij}x_j(t) = 0 \\ 0, & \text{se } \sum_{j=1}^n a_{ij}x_j(t) < 0 \end{cases} \quad \text{Equação 2}$$

Chamamos o somatório $\sum_{j=1}^n (a_{ij}x_j(t))$ de entrada (*input*) de x_i no instante t . Além da regulação definida pela matriz A , esse modelo também permite que os genes possuam uma característica de auto-degradação, onde um gene x_i que possua tal característica recebe $a_{ii} = -1$. Podemos observar que nem toda função Booleana pode ser representada pela Equação 2, e por

isso a rede Booleana é chamada de “Restrita”. Além disso, podemos notar também que a entrada de cada gene x_i depende exclusivamente da i -ésima linha da matriz A (Andrade, 2010).

Portanto, a partir de um conjunto de amostras (Equação 3) deduzimos a rede regulação que explique as interações desse conjunto (Figura 6). A partir da rede de regulação (interações dos genes) podemos montar a matriz de regulação correspondente (Equação 4). Nessa matriz (genes por genes) representamos a influência entre os genes utilizando valores: -1 (inibição), 1 (ativação) ou 0 (sem ativação/inibição).

$$M = \begin{matrix} & \begin{matrix} T_1 & T_2 & T_3 \end{matrix} \\ \begin{matrix} G_1 \\ G_2 \\ G_3 \end{matrix} & \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \end{matrix} \quad \text{Equação 3}$$

Figura 6. Rede de regulação da amostra da Equação 3

$$A = \begin{matrix} & \begin{matrix} G_1 & G_2 & G_3 \end{matrix} \\ \begin{matrix} G_1 \\ G_2 \\ G_3 \end{matrix} & \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \end{pmatrix} \end{matrix} \quad \text{Equação 4}$$

Aplicando então a função definida na Equação 2 construímos o diagrama de estados (Figura 7). Neste caso, o diagrama de estados é encontrado multiplicando todos os estados possíveis (000 - 111) pela matriz de regulação definida pela Equação 4, aplicando a regra descrita pela Equação 2 para cada gene. Por exemplo, multiplicando a matriz de regulação da Equação 4 pelo estado “001”, teremos os seguintes somatórios (Equação 5), aplicando as regras definidas pela Equação 2, obtemos resultados da Equação 6. Portanto, a partir do estado “001” o próximo estado será o estado “001”.

$$\begin{matrix} G_1 & G_2 & G_3 \\ G_1 & \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \end{pmatrix} & * \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} \end{matrix} \quad \text{Equação 5}$$

$$\begin{matrix} G_1 & G_2 & G_3 \\ G_1 & \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \end{pmatrix} & * \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \end{matrix} \quad \text{Equação 6}$$

Figura 7. Diagrama de estados para a rede da Figura 6

2.5. O algoritmo *Tanginference*

Como mostrado na seção 2.4, a partir de um dado conjunto de amostras (dados de entrada), buscamos uma matriz de regulação que explique os dados amostrais observados. O objetivo desse programa é encontrar a(s) matrizes de regulação que expliquem os meus dados amostrais.

O modelo Booleano restrito, apresentado pela Equação 2, as linhas da matriz de regulação são independentes entre si. Essa propriedade (genes independentes) permite encontrar a partir de um conjunto de dados de entrada (amostras) todas as possíveis soluções (linhas) para cada gene, ao invés de buscar por todas as soluções possíveis para a matriz (todos os genes). Isto é, esse algoritmo irá encontrar soluções que expliquem a regulação de cada gene separadamente. Com isso, o tempo de complexidade do algoritmo é reduzido pela diminuição do número de combinações possíveis de 3^{n^2} (número de matrizes possíveis) para $n \cdot 3^n$ (número total de linhas possíveis), sendo n o número de genes. (Higa, et al., 2010).

Vamos supor o seguinte conjunto de amostras da Equação 7. Aplicando essa abordagem de genes independentes, temos todas as soluções consistentes para cada um dos três genes apresentados na Tabela 1. Com essas soluções podemos combinar os resultados obtidos para cada gene e formar matrizes soluções para o problema. Na Tabela 2 é possível observar algumas matrizes soluções utilizando os resultados independentes obtidos na Tabela 1.

$$D = \begin{matrix} & t_1 & t_2 & t_3 \\ \begin{matrix} G_1 \\ G_2 \\ G_3 \end{matrix} & \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \end{matrix} \quad \text{Equação 7}$$

Soluções para Gene1	Soluções para Gene2	Soluções para Gene3
-1	-1	-1
0	-1	0
1	0	-1
	0	-1
	0	-1
	0	0
	0	-1
	1	-1
	1	0
	0	-1

Tabela 1. Conjunto de soluções dado o conjunto de amostras da matriz A (Equação 7)

Matriz solução 1	Matriz solução 2	Matriz solução 3
0	0	1
1	0	-1
-1	1	-1

Tabela 2. Exemplo de matrizes soluções para conjunto da Tabela 1

Utilizando essa abordagem, foi proposto um algoritmo, que chamaremos nesse trabalho de *Tanginference*, baseado no algoritmo desenvolvido por (Higa, et al., 2010). Ele utiliza os modelos de Redes Booleanas Restritas descritas e faz uma busca exaustiva por soluções dado um conjunto de dados temporal. O objetivo desse algoritmo é encontrar todas as linhas consistentes para cada gene.

Um problema desse algoritmo é tamanho do conjunto de genes. Por se tratar de um problema mal-posto (*ill-posed*), um conjunto de dados com poucas amostras temporais e o número grande de genes pode gerar um grande número de soluções. Uma solução para esse problema é reduzir o número de genes do conjunto empregando técnicas de agrupamento, algoritmos de seleção de recursos, construção de sub-redes pequenas usando o paradigma de genes de sementes de cultura e/ou selecionando um grupo pequeno de genes envolvidos em um processo biológico como genes envolvidos no ciclo celular, via da respiração, etc (Higa, et al., 2010). Neste trabalho focamos na modelagem utilizando apenas genes envolvidos com o processo de fermentação (via da glicólises/fermentação da levedura).

2.6. Modelo de inferência por Redes Booleanas

O trabalho realizado por (Li, et al., 2004) realizou com sucesso a modelagem de dados temporais de levedura utilizando o modelo de redes booleanas restritas. Eles selecionaram 11 principais genes envolvidos no processo do ciclo celular de levedura (*Saccharomyces cerevisiae*) baseada em evidências na literatura (são 800 genes envolvidos).

Aplicando o modelo de redes booleanas restritas ao conjunto de dados do ciclo celular, (Li, et al., 2004) encontraram como solução um rede formada por sete bacias com atratores *singletons* (estados estacionários) descritos na Figura 8. O atrator da maior bacia é composto por 86% dos estados da rede. Estudos apontavam que este estado atrator corresponde ao estado estacionário biológico do ciclo celular, conhecido como G_1 (estado inicial/final do ciclo celular).

Basin size	Cln3	MBF	SBF	Cln1,2	Cdh1	Swi5	Cdc20	Clb5,6	Sic1	Clb1,2	Mcm1
1,764	0	0	0	0	1	0	0	0	1	0	0
151	0	0	1	1	0	0	0	0	0	0	0
109	0	1	0	0	1	0	0	0	1	0	0
9	0	0	0	0	0	0	0	0	1	0	0
7	0	1	0	0	0	0	0	0	1	0	0
7	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	0	0	0	0	0	0

Figura 8. Estados estacionários encontrados na via do ciclo celular. Fonte: (LI, ET AL., 2004)

O diagrama de estados da maior bacia de atração encontrada pode ser observada na Figura 9, onde os estados estão representados nas cores verdes, as arestas na cor vermelha e os principais estados que convergem ao atrator estão representados na cor azul. Esses estados correspondem ao sentido da via biológica do ciclo celular.

Figura 9. O diagrama de estados da maior bacia de atração encontrada no ciclo celular. Fonte: (Li, et al., 2004)

Capítulo 3

Materiais e Métodos

Neste capítulo apresentaremos os materiais e métodos utilizados ao longo deste trabalho. Por se tratar de um trabalho multidisciplinar, abordaremos tanto os métodos biológicos como os computacionais. Apresentaremos uma breve introdução sobre o conteúdo desse capítulo na seção 3.1 e uma breve descrição sobre análise de expressão gênica e as técnicas existentes (seção 3.2). Em seguida, serão apresentadas informações sobre a origem dos dados biológicos (seção 3.3), apresentando as ferramentas e métodos utilizados na análise e tratamento dos dados (seção 3.4) e os métodos utilizados para seleção do conjunto de genes utilizados (seção 3.5). Na parte computacional, abordaremos as ferramentas e métodos utilizados para gerar as redes booleanas, objetivo do nosso trabalho (seção 3.6) e as métricas utilizadas para realizar a análise comparativa (3.7).

3.1 O problema biológico

O processo de fermentação modifica o metabolismo das leveduras devido às condições estressantes a que são submetidas. Entretanto, algumas linhagens de leveduras fermentadoras conseguem se adaptar melhor a esse tipo de ambiente. Entender os mecanismos de adaptação dessas leveduras é de bastante importância para melhoramento das linhagens de levedura industrial. Combinar informações genéticas que maximizem a produção e minimizem os problemas de adaptação das leveduras é o foco principal dos projetos de pesquisa em bioetanol.

Utilizamos então conjuntos de dados distintos para modelar os dados biológicos utilizando o modelo de redes booleanas e analisar a possibilidade de encontrar diferenças metabólicas que expliquem a melhor adaptação de uma linhagem em relação à outra utilizando esse modelo.

Os dados biológicos utilizados neste trabalho são de duas linhagens distintas de levedura da espécie *Saccharomyces cerevisiae*: a linhagem S288c (domesticada) e a linhagem PE-2 (selvagem) com amostras extraídas em

condições ambientais distintas: mescladas entre escala laboratorial e industrial e entre as condições de fermentação e respiração. Esses dados foram extraídos de trabalhos experimentais já realizados. Os dados da linhagem S288c do trabalho de (Derisi, et al., 1997) e os dados da linhagem PE-2 são de trabalhos realizados pelo aluno Osmar Carvalho Netto do Laboratório de Genômica e Expressão – UNICAMP, ainda não disponíveis publicamente. São três conjuntos de bibliotecas (amostras) e cada um desses conjuntos será detalhado nas seções seguintes.

3.2 Expressão gênica

3.2.1 As técnicas de análise de expressão gênica utilizadas

Nesta seção iremos descrever as técnicas de análise de expressão gênica de *microarray* e de *RNA-seq*, que foram utilizadas nos conjuntos de dados biológicos utilizados neste trabalho.

Microarray

Microarray ou microarranjos de DNA é uma técnica de análise de expressão gênica de transcritos em larga escala (Schena, et al., 1995), na qual se utiliza uma lâmina com oligonucleotídeos de DNA fita simples ou cDNAs. Cada oligonucleotídeo possui uma sequência pré-definida que representa um determinado gene. Esses oligonucleotídeos estão aderidos a pontos conhecidos na lâmina chamados de *spots*. Uma placa de *microarray* pode conter dezenas ou até centenas de milhares de *spots*, cobrindo assim todos os genes conhecidos de um determinado genoma (Mayer, 2009).

Para a construção das placas de *microarray* é necessário conhecer previamente o genoma do organismo de interesse (sequenciar seu genoma ou obter lâminas pré-montadas com genomas já conhecidos – há empresas que já fornecem) já que é necessário um oligonucleotídeo de DNA fita simples de cada gene deste organismo. As moléculas de *mRNA* extraídas do experimento biológico são marcadas com um corante fluorescente. Quando esses *mRNAs* hibridizam (colam) com seus oligonucleotídeos correspondentes fixados na lâmina, liberam uma fluorescência de forma a ser detectado por um *scanner*

(leitor). A intensidade da fluorescência é proporcional à concentração das moléculas de *mRNAs* hibridizadas. Desta forma, é possível detectar e quantificar o nível de expressão proveniente das amostras biológicas.

É possível analisar duas amostras na mesma placa utilizando corantes distintos (*double-channel microarray*). Como, por exemplo, quando estudamos duas condições distintas para o mesmo organismo. Neste caso, o *cDNA* de uma amostra (condição) é marcado com corante vermelho (Cy5) e a outra amostra (controle) com corante verde (Cy3). No caso do *double-channel microarray*, podemos entender que spots na cor amarela são comuns à duas amostras e os spots verde (controle) e vermelho (condição) são exclusivos para cada uma das condições. Uma ilustração desse processo pode ser observado na Figura 10.

Figura 10. Ilustração dos passos da técnica de microarray double-channel.

A análise da intensidade dos spots envolve uma fase de processamento da imagem do *microarray* para extrair o sinal (expressão) de cada spot, analisar sua qualidade e quantificar. O objetivo desta fase é extrair o sinal de cada spot e gerar uma matriz numérica contendo as intensidades dos sinais.

RNA-seq

O RNA-seq também é uma técnica de análise de expressão gênica em larga escala (RNA-Seq: a revolutionary tool for transcriptomics, 2009). Esta técnica é, na verdade, uma nova abordagem para as novas tecnologias de sequenciamento de DNA em larga escala (*high-throughput*), como: *Illumina*, *SOLiD*, *454* e *Helicos tSMS*. Porém, ao invés de extrair dados de DNA, são extraídos os RNAs e estes são convertidos em bibliotecas de fragmentos de *cDNA*, conforme o protocolo de preparação de amostras de cada tecnologia, e posteriormente é sequenciado. Com os dados provenientes do sequenciamento desses *cDNAs*, os fragmentos gerados são alinhados ao genoma ou o transcriptoma (conjunto de transcritos) do organismo de interesse ou até mesmo podem ser montadas *de novo* (sem genoma de referência), de maneira a produzir um mapa transcricional em escala genômica. (Ribeiro, 2010).

As principais vantagens dessa técnica são: i) não haver necessidade de conhecimento anterior do genoma do organismo, ii) não necessitar de clonagens ou amplificações, iii) uma maior acurácia (por ser um método quantitativo) e iv) a possibilidade de detectar variações na sequência genômica nas regiões transcritas (como, por exemplo, SNPs – *Single-nucleotide Polymorphisms*) (Ribeiro, 2010).

3.3 Os conjuntos de dados

Neste trabalho utilizamos três conjuntos de dados temporais de levedura da espécie *Saccharomyces cerevisiae*, sendo dois da linhagem PE-2 e um da linhagem S288c. Iremos detalhar cada um deles nessa seção.

3.3.1 Dados da linhagem S288c em microarray

Os dados da linhagem de S288c foram extraídos de um experimento realizado e publicado por (Derisi, et al., 1997) disponível publicamente pelo NCBI (National Center for Biotechnology Information) e SGD (*Saccharomyces Genome Database*). Este experimento consiste em uma análise de expressão gênica temporal dessa linhagem de levedura em um intervalo de tempo de 21 horas com temperatura constante de 30°C. O experimento possui tempo inicial

durante a fermentação, passando pelo *diauxic shift* (transição da fermentação para respiração) até chegar ao ponto final durante respiração. Foram extraídas amostras em sete tempos distintos, com concentração de glicose inicial de 20 g/l e final de 0 g/l, isto é, toda a glicose do meio foi consumida. A técnica utilizada para análise de expressão gênica foi a de *microarray*. A Figura 11 mostra a curva das concentrações de glicose e de crescimento (OD) durante todo o experimento.

Figura 11. Tabela das concentrações de glicose do experimento de (DERISI, ET AL., 1997)

3.3.2 Dados de PE-2 em microarray

Os dados da linhagem de PE-2 em *microarray* foram extraídos de experimentos realizados pela equipe do Laboratório de Genômica e Expressão da UNICAMP em 2009. Este experimento visa analisar o perfil de expressão gênica dessa linhagem durante a fermentação. Entretanto, essa análise foi realizada em escala laboratorial, isto é, fermentação assistida em laboratório. Como controle na análise diferencial nos três pontos de fermentação, utilizamos uma cultura submetida à condição de respiração (referência). As amostras de fermentações foram realizadas sob condições controladas de

temperatura (30°C), utilizando-se xarope de cana-de-açúcar . Foram coletadas três amostras de células em cada coleta (réplicas), sendo a primeira coleta após três horas de fermentação, a segunda em 9 horas e a última após 15 horas (Netto, 2010).

As amostras foram hibridadas em lâminas de *microarray* contendo as regiões codantes do genoma de *Saccharomyces cerevisiae*. Estas lâminas foram sintetizadas pelo laboratório do Prof. Juan Lucas Argueso, na Universidade de Duke (EUA). Após a hibridização, as lâminas foram lavadas e a intensidade das fluorescências foi captada pelo scanner *GenePix 4000B* e as imagens analisadas pelo programa *GenePix Pro 6.0*. Os valores das concentrações de etanol e de açúcares residuais do experimento realizado podem ser visualizados na Tabela 3 e na Figura 12.

Horas/Réplica	°GL	% açúcar residual
3/1	3,06	11,27
3/2	3,01	11,35
3/3	2,99	11,38
9/1	7,25	4,80
9/2	7,37	4,62
9/3	7,37	4,62
15/1	8,78	2,44
15/2	8,81	2,39
15/3	8,85	2,33

Tabela 3. Concentrações de etanol (°GL) e açúcares residuais no mosto do experimento de fermentação em escala de laboratório. (Netto, 2010)

Figura 12. Gráfico representando as variações nas concentrações de etanol e açúcares redutores presentes na fermentação em escala de laboratório com dados obtidos da Tabela 3. (Netto, 2010)

3.3.3 Dados de PE-2 em RNA-seq

Os dados da linhagem de PE-2 em *RNA-seq* foram extraídos de experimentos realizados também pela equipe do Laboratório de Genômica e Expressão da UNICAMP em parceria com a Usina Nova América, com o Dr. Piotr Mieczkowski na Universidade da Carolina do Norte (EUA) e do Prof. Juan Lucas Argueso na Universidade do Colorado (EUA) em 2010. Esse trabalho pertence a uma pesquisa realizada pelo aluno Osmar V. de Carvalho Netto onde foram coletadas amostras biológicas durante o processo de fermentação em escala industrial, isto é, nas dornas de fermentação de usinas sucroalcooleiras. O experimento foi realizado na Usina Nova América, em Maracaí – SP em condições típicas de usina, com temperatura da dorna em torno de 31°C. As amostras foram coletadas a cada 3 horas (totalizando seis amostras), desde o início da fermentação até o completo esgotamento da glicose (15 horas) (Netto, 2010). A Figura 13 ilustra as concentrações de etanol nas seis amostras.

O preparo da amostra foi realizado na Universidade do Colorado e as amostras foram enviadas para a *High Throughput Sequencing Facility* do Dr. Mieczkowski para o sequenciamento. Para o sequenciamento, foram utilizados os protocolos e equipamentos *Illumina Analyzer II* (Netto, 2010).

Figura 13. Concentração de Etanol para as amostras coletas na Usina Nova América (Netto, 2010).

3.4 Tratamento dos dados biológicos

Nesta seção serão descritos os processos de normalização dos dados, binarização (discretização) e o esquema de janelas que utilizamos nos dados de *microarray* em S288c.

3.4.1. Normalização dos dados

A normalização dos dados foi realizada de acordo com a técnica de análise de expressão gênica utilizada. Nessa seção iremos detalhar como foi realizada cada uma delas.

Dados de S288c em *microarray*

Como citado anteriormente, os dados de S288c em *microarray* foram extraídos diretamente do banco de dados público NCBI e estes se encontram normalizados e em escala $\log_2$ ratios. A normalização foi realizada utilizando a amostra do tempo 1 (fermentação) como referência. Esses dados são públicos e estão acessíveis pelo site do NCBI (GEO: GSE28).

Dados de PE-2 em *microarray*

A normalização dos dados de *microarray* em PE-2 foram realizados utilizando pacotes *R* incluídos no *Bioconductor* (repositório) baseados no tutorial descrito (Filho, et al., 2009). Utilizamos o pacote *limma* e ajustes para scanners *GenePix* (no caso dos nossos dados) para leitura das placas, correção de background, normalização e a extração dos dados. Os gráficos dos dados antes e após a normalização encontram-se no Apêndice deste trabalho.

Dados de PE-2 em *RNA-seq*

Como citado anteriormente, o RNA-seq é uma técnica *high-throughput* cujo resultado são as sequências de nucleotídeos dos fragmentos de cDNA de entrada (*reads*). Estes podem ser alinhados no genoma ou transcriptoma para quantificar o nível de expressão gênica para cada gene. No nosso caso, alinhamos as sequências de saída ao genoma da levedura PE-2.

Para analisar esses dados utilizamos um pacote do *Bioconductor* chamado *DEGseq* (Wang, et al., 2010). O *DEGseq* é um pacote *R* utilizado principalmente para identificar genes diferencialmente expressos em

bibliotecas de *RNA-seq* comparando bibliotecas de dados de cDNAs. Para comparar bibliotecas, ele necessita do mapeamento dos *reads* provenientes dos dados de *RNA-seq* com a anotação de genes correspondentes em um genoma ou com valores de expressão de genes fornecidos por outros programas. No nosso caso, como alinharmos nossos dados ao genoma da levedura, utilizamos um programa chamado *bowtie* (Langmead, et al., 2009) para alinhamento das sequências com o genoma da PE-2.

Utilizamos os dados do *tempo 1* (inicial) como referência para as amostras nos tempos seguintes (4, 7, 9, 10, 12 e 15 horas). Utilizamos os dados em escala $\log_2$ ratios, já que foi a mesma escala utilizada nas outras bibliotecas.

3.4.2 Binarização dos dados

As redes booleanas são redes formadas por conjuntos de dados binários, portanto, é um pré-requisito para aplicação desse modelo que os dados contínuos sejam transformados em dados binários. Para isso, é necessário um algoritmo que, observando os valores de expressão gênica ao longo do tempo (amostras), encontre um valor limiar que separe expressões ativas e inativas (0,1). Exemplos de como seria essa binarização pode ser observado na Figura 14, onde os pontos em azuis representam os valores de expressão ao longo do tempo (amostras) e a linha em vermelho representa o limiar encontrado.

Figura 14. Exemplos de limiar para binarização dos dados

Portanto, para binarizar dos dados, desenvolvemos um programa em *perl* utilizando o algoritmo proposto por (Shmulevich, et al., 2002). O algoritmo utilizado pode ser observado no Algoritmo 1.

Binarize: INPUT G_i , OUTPUT B_i

```

 $S_i \leftarrow \text{sort}(G_{i,1}, G_{i,2}, \dots, G_{i,k})$ 
for  $j = 1$  to  $k - 1$  do
     $D_{i,j} \leftarrow (S_{i,j+1} - S_{i,j})$ 
end for
 $t \leftarrow (S_{i,k} - S_{i,1}) / (k - 1)$ 
 $m = \min\{j : D_{i,j} > t\}$ 
for  $j = 1$  to  $k$  do
    if  $G_{i,j} \geq S_{i,m+1}$  then
         $B_{i,j} \leftarrow 1$ 
    else
         $B_{i,j} \leftarrow 0$ 
    end if
end for

```

Algoritmo 1. Binarização de dados temporais (Shmulevich, et al., 2002)

Este algoritmo tem como entrada o vetor dos dados temporais (G_i) e como saída um vetor com os dados binarizados (B_i). Para binarizar os dados, esse algoritmo calcula um limiar m onde se $G_i \geq m$ então $B_i = "1"$ senão $B_i = "0"$. Para encontrar esse limiar é criado um vetor (D) com a subtração sequencial do tempo i pelo $i - 1$ do vetor (S), sendo (S) o vetor ordenado de (G). O valor m é o menor valor encontrado que atenda a condição: $m > (S_{i,k} - S_{i,1}) / (k - 1)$, sendo k o número de amostras temporais. Este algoritmo foi utilizado para binarizar dados dos três conjuntos de dados (bibliotecas).

3.4.3 Janelas

Os dados da linhagem S288c são provenientes de um experimento que inicia em fermentação e termina em respiração (*diauxic shift*). Desenhamos então um esquema para modular cada etapa (fermentação/respiração) com mais precisão, dividindo os dados em um esquema que denominamos *janelas*. Esse esquema pode ser observado na Figura 15.

Chamamos de janela o conjunto de 4 tempos sequenciais do experimento, sendo o tempo x de cada janela o tempo seguinte ao tempo x da janela anterior, isto é, iremos formar 4 janelas: a primeira do tempo 1 ao 4, a

segunda do tempo 2 ao 5, a terceira do tempo 3 ao 6 e a quarta do tempo 4 ao 7.

Figura 15. Esquema de janelas para o experimento de S288c

3.5 Seleção dos conjuntos de genes

Nesta seção serão descritos os processos que utilizamos para selecionar o conjunto de genes utilizados.

3.5.1 Conjunto inicial de genes

O foco deste trabalho é analisar genes envolvidos no durante o processo de fermentação. Portanto, iniciamos a seleção de genes pesquisando os genes pertencentes às vias da fermentação e da glicólises e selecionamos 11 genes desse conjunto baseados em dados na literatura. Os genes selecionados foram HXK1, PGI1, PFK1, FBA1, TDH1, PGK1, ENO1, ENO2, CDC19, PDC1 e ADH1. Na Figura 16 podemos observar a disposição desses genes nas vias de fermentação e glicólises e a Tabela 4 mostra correspondências do *alias* com o nome oficial de cada gene.

Figura 16. Genes iniciais na via da glicólise e da fermentação (Conant, et al., 2007)

3.5.2 Conjuntos de genes utilizados

Para formar o nosso conjunto final de genes, utilizamos uma ferramenta chamada BIOGRID (*Biological General Repository of Interaction Datasets*). O BIOGRID é um repositório de dados público online onde são depositados dados de interações gênicas e proteicas em alguns organismos modelos, sendo *Saccharomyces cerevisiae* um deles. A partir dos nossos 11 genes iniciais, selecionamos através do BIOGRID todos os genes que se relacionam (gene-proteína) com nossos genes, independente da via ou processo a que pertençam. De todos os genes encontrados, selecionamos apenas os 19 genes mais frequentes, isto é, os genes que se relacionaram com o maior número de

genes do nosso conjunto inicial. Os 19 genes selecionados podem ser observados na Tabela 5.

GENE	ALIAS
YOL086C	ADH1
YLR044C	PDC1
YGR254W	ENO1
YHR174W	ENO2
YCR012W	PGK1
YAL038W	CDC19
YJL052W	TDH1
YGR240C	PFK1
YBR196C	PGI1
YFR053C	HXK1
YKL060C	FBA1

Tabela 4. Tabela correspondência genes/alias para os 11 genes iniciais

Para maiores informações sobre cada um desses genes é possível acessar através do site do (KEGG).

GENE	ALIAS
YJR009C	TDH2
YGR192C	TDH3
YLR134W	PDC5
YDR510W	SMT3
YER151C	UBP3
YGR087C	PDC6
YHR200W	RPN10
YHR027C	RPN1
YFR004W	RPN11
YCL040W	GLK1
YOR127W	ARO1
YNR051C	BRE5
YLL039C	UBI4
YBR126C	TPS1
YGL122C	NAB2
YBR194W	AIM4
YGL253W	HXK2
YMR205C	PFK2
YKL152C	GPM1

Tabela 5. Tabela com nomes e alias dos 19 genes selecionados

3.5.3. Breve revisão sobre os genes utilizados

Os 11 genes iniciais utilizados foram extraídos da via da glicólise de levedura, portanto todos os genes estão ligados à fermentação. Além desses, mais outros sete genes selecionados através do BIOGRID pertencem às vias da glicólise e fermentação: TDH2, TDH3, PDC5, PDC6, HXK2, PFK2 e GPM1. Esses 19 genes (11 genes iniciais mais os 7 genes do conjunto do BIOGRID pertencentes às vias glicólise/fermentação) estão relacionados entre si através de seus substratos, onde o substrato do gene anterior é necessário para a produção do substrato do gene seguinte. Na Figura 17 podemos observar a sequência de processos metabólicos dos 19 genes.

Figura 17. Sequência de processos metabólico dos 19 genes

Podemos observar que algumas reações possuem dois ou mais genes relacionados. Numa reação geralmente há um gene preferencial, onde os outros genes servem como substitutos caso haja uma deleção ou “não-ativação” do gene principal, como no caso do PDC1 e PDC5. Em outros casos, como do ENO1 e ENO2, essa escolha é feita por fatores externos: durante a fermentação o gene ENO2 é o preferencial. Outro exemplo também é com os genes TDH1, TDH2, e TDH3. Onde o gene TDH1 é ativado normalmente quando o organismo se encontra em estado estacionário e o TDH2 e TDH3 durante crescimento (KEGG).

Os outros 12 genes selecionados através do BIOGRID vêm de processos distintos. Muitos desses genes estão relacionados à replicação, transporte, degradação de compostos, fosforilação e/ou interações proteicas, como SMT3, NAB2 e a família dos RPN (RPN1, RPN10 e RPN11). Os genes AIM4, UBI4, UBP3 e BRE5 estão relacionados à resposta ao estresse. O gene TPS1 também está relacionado ao estresse, porém esse gene está relacionado também à produção de trealose. A trealose é um composto que serve como uma reserva de açúcar e como proteção para a própria levedura durante processos de estresse. Já o gene ARO1 está relacionado à resistência ao etanol que, apesar de ser um composto produzido pela própria levedura, um nível elevado de etanol no meio pode ser tóxico para a levedura.

3.5.4. Esquema dos conjuntos de genes

Com os 11 genes iniciais mais os 19 genes selecionados através do BIOGRID, temos um total de 30 genes. Como algoritmo utilizado neste trabalho é limitado quanto ao número de genes, um conjunto com 30 genes seria impraticável. Decidimos então formar grupos de 19 genes, onde os 11 genes iniciais (genes importantes para processo de fermentação e respiração da levedura) estão presentes em todos os conjuntos e para cada conjunto 1 gene foi selecionado a partir do conjunto de 19 genes restantes.

3.5.5. Seleção das bibliotecas de dados

Para cada conjunto de genes, selecionamos os dados de expressão gênica em cada biblioteca. Chamaremos de biblioteca os conjuntos de dados de expressão gênica dos experimentos de microarray em PE-2 (micrope2), RNA-seq em PE-2 (rnaseqpe2) e microarray em S288c (janela 1- 4). Ainda para facilitar o entendimento, chamaremos de conjunto de dados os conjuntos de genes em cada biblioteca (20 conjuntos para cada). Resumidamente, temos seis bibliotecas de dados (quatro delas de S288c e duas de PE-2) divididas em 20 conjuntos de genes (1 conjunto com 11 genes iniciais e 19 conjuntos com 12 genes), totalizando 120 conjuntos de dados a serem analisados. Podemos observar na Figura 18 o resumo desse *pipeline*.

Figura 18. Esquema de seleção das bibliotecas de dados.

3.6. Gerando redes consistentes

Como detalhado no capítulo 2 (seção 2.5), utilizamos o algoritmo baseado no algoritmo desenvolvido por (Higa, et al., 2010) para gerar redes consistentes para os nossos dados amostrais. Portanto, para cada conjunto de dados foram geradas todas as linhas consistentes para cada gene. Entretanto, o fato das bibliotecas possuírem poucas amostras temporais fez com que o número de linhas consistentes para cada gene ainda fosse extenso e seria impossível trabalhar com todas as linhas consistentes encontradas.

Modelamos então nosso trabalho da seguinte forma. Para cada conjunto de dados, geramos apenas 1.000 redes aleatórias para representar esse conjunto de dados. Essas redes foram geradas utilizando programas desenvolvidos em *perl*, sendo que cada linha de cada gene foi escolhida

randomicamente, tendo esta que seguir apenas uma restrição: a linha selecionada deve ter grau de entrada: $Input < 8$. Essa restrição visa evitar gerar matrizes onde um ou mais genes sejam regulados por quase todos os outros genes. Gerando essas 1.000 redes para cada um dos conjuntos de dados, totalizamos 120.000 redes para as seis bibliotecas. Na Figura 19 podemos visualizar o esquema desde a seleção dos dados até o esquema das redes aleatórias.

Figura 19. Esquema de seleção das bibliotecas de dados até gerar as redes aleatórias.

3.7. Análise comparativa das redes

Para analisar esse grande número de redes, utilizamos alguns métodos já descritos em literatura para analisar e comparar as redes. Utilizamos os seguintes métodos: o cálculo de entropia (Shannon, 1948), medidas de distância lógica (*rule-based distance*) (Dougherty, 2007) e uma medida de distância topológica baseadas nos trabalhos de (Dougherty, 2007; Lopes, 2011). Nesta seção iremos descrever cada um desses métodos.

3.7.1. Entropia

Entropia é um conceito utilizado na termodinâmica desde 1865 por Rudolf Clausius e posteriormente foi aplicada à teoria da informação por Claude Shannon (Shannon, 1948). Essa medida é frequentemente utilizada para medir a quantidade de informação contida em um determinado conjunto de dados. A entropia também pode ser utilizada para medir a desordem (incerteza) de um

conjunto de dados (Lopes, 2011). A entropia proposta por Shannon (Shannon, et al., 1963) é definida pela Equação 8.

$$H(X) = - \sum_{i=1}^x p(i) \log p(i) \quad \text{Equação 8}$$

Ou seja, dada a variável aleatória $p(i)$, $H(X)$ é o valor de incerteza associado a esta variável. Quanto maior a entropia, maior incerteza em prever o valor dessa variável. Portanto, considerando x o número de bacias de uma rede e $p(i)$ a probabilidade de um estado na bacia x , o valor de $H(X)$ reflete o grau de incerteza de um estado pertencer à bacia x . Dessa forma, podemos associar uma maior entropia às redes mais estáveis (com maiores bacias de atração). Utilizando o valor de entropia da informação da Equação 8, calculamos então o valor da entropia normalizada dos dados (Equação 9).

$$\bar{H} = \frac{H}{H_{max}} \quad \text{Equação 9}$$

Sendo H_{max} o valor de entropia máxima definida por um valor de distribuição de probabilidade que maximize a entropia da informação, isto é, supondo que todas as bacias possuem o mesmo tamanho. A entropia máxima então pode ser definida pela Equação 10.

$$H_{max} = - \sum_{i=1}^x p_i \log p_i = - \sum_{i=1}^x \frac{2^n}{x} \log \frac{2^n}{x} = -x \left(\frac{1}{x} \log \frac{1}{x} \right) = - \log \frac{1}{x} \quad \text{Equação 10}$$

onde x o número de bacias da rede e n o número de genes.

3.7.2. Medida de distância lógica

Podemos ainda medir a distância lógica das redes utilizando o método comparação de distância baseado em regras (do inglês, *Rule-based distance*) proposto por (Dougherty, 2007). Esse método calcula a distância entre duas redes através da proporção de linhas erradas dadas as funções definidas pela tabela verdade. Dadas funções dos estados para as redes H e M por $f = (f_1, f_2, \dots, f_n)$ e $g = (g_1, g_2, \dots, g_n)$ respectivamente

e supondo que n sejam tabelas verdades consistindo 2^n linhas cada, a distância é dada pela Equação 11.

$$\mu_{fun}(M, H) = \frac{1}{n2^n} \sum_{i=1}^n \sum_{k=1}^N I[g_i(x_k) \neq f_i(x_k)] \quad \text{Equação 11}$$

Sendo N o total de colunas e n o total de linhas das matrizes. Portanto, $I[A]$ pode assumir valor 1 para uma sentença verdadeira ou 0 caso a sentença seja falsa. Por exemplo, vamos supor duas matrizes T e B (Equação 12).

$$T = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{Equação 12}$$

Observando as matrizes podemos concluir que $\sum_{i=1}^3 \sum_{k=1}^3 I[A] = 2$. Portanto, distância entre as redes T e B é definida por $\mu_{fun}(T, B) = 0,0833$.

3.7.3. Medida de distância topológica

O método de medida de distância topológica proposto por (Dougherty, 2007) propõe medir a distância entre duas redes (supostamente, uma modelo e outra identificada), utilizando como base a matriz de confusão (Kohavi, et al., 1998; Webb, 2002). Uma matriz de confusão funciona como um classificador onde cada elemento da matriz é uma medida entre as arestas das redes de entrada. No nosso trabalho, iremos utilizar esse método para comparar duas redes identificadas, portanto, iremos utilizar um modelo adaptado.

No método proposto por (Dougherty, 2007) são utilizadas as matrizes de adjacências das redes de entrada para comparar as redes, entretanto, como estamos interessados em manter o sentido das arestas, iremos utilizar as matrizes direcionais. Para computar os elementos da matriz confusão aplicamos as seguintes regras: PA (presente em ambas as redes) para arestas de mesmo sinal na mesma direção, AA (ausente em ambas as redes) para arestas ausentes, PP (presente apenas na primeira) para arestas presente

apenas na primeira rede de entrada (independente do sinal), PS (presente apenas na segunda) para arestas presente apenas na segunda rede de entrada e, para o caso de arestas presentes em ambas as redes com sinal invertido, foram computadas em PP e PS. Um resumo da matriz de confusão utilizada encontra-se na Tabela 6. Este método é uma adaptação do método utilizado em (Lopes, 2011).

	Presente na rede 2	Ausente na rede 2
Presente na rede 1	PA	PP
Ausente na rede 1	OS	AA

Tabela 6. Matriz confusão para medir a distância topológica das redes identificadas, (PA = presente em ambas, PP, presente na primeira rede apenas, OS = presente na segunda rede apenas e AA = ausente em ambas)

Para avaliar os resultados obtidos na matriz confusão, utilizamos medidas de que visam quantificar as arestas inferidas. Elas podem ser definidas por PPV (do inglês *Positive Predictive Value*) definida pela Equação 13 e a medida de Sensibilidade pela Equação 14. Essas medidas não são independentes e, portanto, é adequado utilizar sua média geométrica quando utilizadas para medir similaridade entre duas redes. Essa métrica, chamada medida de similaridade, é definida pela Equação 15. É importante ressaltar que a similaridade máxima será obtida por valores próximos a 1 (Lopes, 2011).

$$PPV = \frac{PA}{PA+PS} \quad \text{Equação 13}$$

$$\text{Sensibilidade} = \frac{PA}{PA+PP} \quad \text{Equação 14}$$

$$\text{Similaridade}(A, B) = \sqrt{PPV * \text{Sensibilidade}} \quad \text{Equação 15}$$

Capítulo 4

Resultados e discussões

Neste capítulo apresentaremos alguns resultados obtidos utilizando as bibliotecas de S288c e PE-2 utilizando os conjuntos de genes descritos. Os métodos utilizados nas redes apresentadas foram descritos seção 3.6 e os métodos de comparação foram descritos na seção 3.7. Na seção 4.1 descrevemos uma análise inicial utilizando apenas 11 genes principais da via da fermentação. Na seção 4.2 iremos descrever alguns resultados dos 19 conjuntos de genes propostos. Nas seções 4.3 e 4.4 faremos uma breve discussão sobre resultados encontrados.

4.1. Redes com apenas 11 genes

Nessa análise geramos redes utilizando apenas os 11 genes iniciais nas seis bibliotecas. Foram geradas 1.000 redes aleatórias para cada biblioteca e os valores de entropia de cada rede foram calculados. O que descrevemos aqui como valor de entropia é o valor de entropia relativa descrito na seção 3.7.3. Para todas as redes (1.000 redes de cada biblioteca), agrupamos os resultados na faixa de valores (0,1 à 0,9) e analisamos a distribuição de entropias. Esses gráficos podem ser observados na Figura 20. Cada gráfico representa o número de redes por faixa de valor de entropia.

Figura 20. Gráficos de distribuição de entropia para cada biblioteca

Como primeiro critério de seleção, buscamos por redes com maior estabilidade (menor número de bacia e uma grande com muitos estados). Essa característica é ideal para que possamos chegar o mais próximo possível de redes biológicas reais. Portanto, nosso objetivo é selecionar redes que apresentam as menores entropias do conjunto. Observando as distribuições de entropias das 1.000 redes para cada biblioteca (Figura 20) concluímos que a maioria das redes nas seis bibliotecas encontra-se na faixa de valor de entropia entre 0,5 e 0,7. E que redes com entropia menores que 0,1 ($H < 0,1$) são menos que 6% das redes encontradas. Portanto, buscamos como melhores redes as que apresentam entropia menor ou igual a 0,1.

Como segundo critério para seleção das redes é o número de estados da maior bacia de atração. Seguindo esse critério, selecionamos as seguintes redes representadas nas: Figura 21, Figura 23, Figura 24, Figura 25, Figura 26 e Figura 27.

Figura 21. Diagrama de estados da rede de RNA-seq com 11 genes. Os pontos em verde representam os estados da rede, os pontos em vermelho os estados do dado de entrada, sendo o de cor amarela o estado atrator da rede. O estado final dos dados de entrada e o estado atrator da rede são o mesmo ponto.

A rede baseada nos dados de RNA-seq em PE-2 (Figura 21) possui duas bacias de atração: uma com 2047 estados e uma com 1 estado apenas “00000000000” (ou 0). O atrator da rede de RNA-seq em PE-2 com 2047 estados coincide com o estado final dos dados de entrada “01111101111” (ou 1007 em decimal). Para analisar a dinâmica dos genes ao longo do tempo, utilizamos a via metabólica da fermentação para plotar os genes e sua ativação/inativação ao longo do tempo (Figura 22). Plotamos os estados correspondente aos valores de entrada dos dados amostrais. Os 11 genes utilizados nesse conjunto estão marcados em amarelo e os genes ativos em cada um dos estados estão em cor vermelha. Podemos observar que da amostra 1 (T1) para a amostra 2 (T2) apenas 3 genes encontram-se ativos (HXK1, ENO1, PGK1) e é quando a sacarose do meio é transformada em

glicose/frutose (b). Do tempo T2 para o tempo T3 já não há mais sacarose no meio, mas ainda há um aumento na quantidade de frutose no meio e podemos observar que todos os genes estão ativos, exceto o gene TDH1 (conhecido na literatura por estar ativo apenas em estados estacionários). Do tempo T3 para T4, o gene TDH1 se torna ativo e então há o consumo da glicose e da frutose do meio. Do T4 para o T5, os genes TDH1 e HXK1 encontram-se desligados. É esperado que o gene HXK1 fique inativo já que esse gene é responsável pela fosforilação do açúcar e já não há mais açúcar no meio. No tempo T5, apenas os genes HXK1 e ENO1 estão desligados.

Figura 22. Dinâmica dos genes do conjunto de 11 genes nas amostras de RNA-seq em PE-2

Figura 23. Diagrama de estados da rede de microarray em PE-2 com 11 genes. Os pontos em verde representam os estados da rede, os pontos em vermelho representam os dados de entrada e o estado em amarelo é o atrator.

A rede baseada nos dados de microarray em PE-2 (Figura 23) também possui 2 atratores: um com 2046 estados com atrator “11000100011”(1571) e um atrator “00000000000” (ou 0) com 2 estados. O atrator dessa rede não coincide com o último estado dos dados de entrada e é o estado “11000100011”. Nota-se que o maior atrator das duas redes não são iguais mesmo estando nas mesmas condições (fermentação), porém há outros fatores que podem influenciar nesse resultado: a) os dados de RNA-seq são dados industriais enquanto os de *microarray* são laboratoriais, b) O meio da fermentação industrial é composto por sacarose (glicose e frutose) e o meio laboratorial contém apenas glicose, c) a última amostra dos dados de microarray ainda contém açúcar no meio (Figura 12).

Em seguida, observamos as redes para dos dados de S288c.

Figura 24. Diagrama de estados da rede da Janela 1 em S288c com 11 genes. Os pontos em verde representam os estados da rede. Os pontos em vermelho representam os estados dos dados de entrada e os estados em são os atratores (sendo um deles o estado final dos dados de entrada).

Figura 25. Diagrama de estados da rede da Janela 2 em S288c com 11 genes. Os pontos em verde representam os estados da rede. Os pontos em vermelho representam os estados dos dados de entrada e os estado em amarelo é o atrator (também é o estado final dos dados de entrada).

Figura 26. Diagrama de estados da rede da Janela3 em S288c com 11 genes. Os pontos em verde representam os estados da rede. Os pontos em vermelho representam os estados dos dados de entrada e o estado em amarelo é o atrator (e também estado final dos dados de entrada).

Figura 27. Diagrama de estados da rede da Janela4 em S288c com 11 genes. Os pontos em verde representam os estados da rede. Os pontos em vermelho representam a trajetória que liga os estados dos dados de entrada, sendo o ponto amarelo o atrator da bacia.

As redes das janelas 1 (Figura 24), 2 (Figura 25) e 3 (Figura 26) compartilham o mesmo atrator “11111111111”(2047) que corresponde ao último estado dos dados de entrada. A rede da Janela 1 possui quatro bacias de atração, sendo a maior com atrator de tamanho 2 (os estados “11111111111”(2047) e “11111110000”(2032)) com 2027 estados. As redes da Janela 2 e 3 possuem dois atratores, sendo o maior de tamanho 2047. A rede da janela 4 tem 3 atratores, sendo o maior com atrator no estado “01101010000” (ou 848).

O que podemos observar nessas redes é que as redes das janelas 1 a 3 de S288c parecem pertencer à mesma condição (fermentação) onde todos os genes aparentemente estão ligados a todo tempo. Quando a condição ambiental muda (respiração), esta via tende a um estado onde apenas os genes PGI1, PFK1, TDH1 e ENO1 permanecem ativos.

Analisando os estados dos dados de RNA-seq PE-2 e os atratores das janelas de S288c (Figura 28), vimos que os atratores dos tempos finais 15, 17 e 19 horas de S288c correspondem ao estado das 10 horas em PE-2 e o atrator das 21 horas de S288c parece corresponder aos estados das 12 e 15 horas de PE-2, mostrando que a linhagem S288c parece ser mais lenta quanto à produção do etanol.

Figura 28. Comparativo dos estados de RNA-seq PE-2 e dos atratores das janelas de S288c.

Entretanto, quando comparamos o atrator dos dados de PE-2 (RNA-seq) e da janela 4 (S288c) vimos que o conjunto de genes ativos/inativos são divergentes apesar das amostras finais de ambos não conterem açúcar no meio. Podemos observar na Figura 29 os genes ativos em vermelho e os inativos em amarelo, sendo os genes com expressão divergente destacados em linhas pontilhadas. Observamos que os primeiros genes (HXK1, PGK1 E PFK1) possuem a mesma expressão e são genes mais relacionados à absorção do açúcar do meio, porém os outros genes possuem expressão completamente inversa. Na PE-2, apesar de não haver mais açúcar no meio, ela continua a fermentar o açúcar por muito tempo. Baseado nessa observação, sugerimos então que a capacidade de absorver mais rapidamente o açúcar do meio parece não acompanhar o mesmo ritmo da transformação desse açúcar em etanol na PE-2.

Figura 29. Comparativo dos atratores de RNA-seq PE-2 e da janela 4 em S288c.

Analizamos comparativamente as redes encontradas utilizando a medida de similaridade descrita na seção 3.7.3 entre todas as bibliotecas (Tabela 7) e também utilizando a medida de distância baseada em regras (*rule-based distance*) descrita na seção 3.7.2 (Figura 30). Observamos que os dados de microarray em PE-2 são topologicamente mais distantes das outras redes, inclusive da rede de RNA-seq. Já na análise lógica, observamos que as redes de microarray e RNA-seq estão mais próximas, porém ainda bem distante do esperado já que se tratam de dados da mesma linhagem de levedura nas mesmas condições ambientais. Acreditamos que esse efeito possa ser causado pela quantidade de amostras desse experimento ou pela quantidade de açúcar no meio ao final da fermentação nos dados de microarray em PE-2.

Com o intuito de encontrar redes da PE-2 e da S288c com maior similaridade entre si, realizamos as análises lógicas e topológicas descritas nas seções 3.7.2 e 3.7.3. A rede da PE-2 (RNA-seq) é mais similar à rede da janela 2 (11 à 17 horas de fermentação) tanto lógica como topologicamente. Analisando apenas as distâncias lógicas poderíamos concluir que as redes da Janela 3 e 4 são mais similares (Tabela 7, Tabela 8).

	RNA-seq (PE2)	Microarray (PE2)	Janela1 (S288c)	Janela2 (S288c)	Janela3 (S288c)	Janela4 (S288c)
RNA-seq(PE2)	1	-	-	-	-	-
Microarray(PE2)	0,61	1	-	-	-	-
Janela 1 (S288c)	0,82	0,70	1	-	-	-
Janela 2 (S288c)	0,83	0,69	0,88	1	-	-
Janela 3 (S288c)	0,81	0,61	0,81	0,81	1	-
Janela 4 (S288c)	0,80	0,69	0,82	0,86	0,77	1

Tabela 7. Tabela comparativa usando medida de similaridade entre as bibliotecas do conjunto de 11 genes.

Figura 30. Gráfico comparativo usando medida de distância baseada em regras entre as bibliotecas do conjunto de 11 genes.

Para comparar a dinâmica das redes de RNA-seq PE-2 e da Janela 2 (diagrama de estados), utilizamos o diagrama de estados da rede de PE-2 gerada e plotamos os estados de entrada da rede da Janela 2 na figura a fim de analisar se há sobreposição de estados e como estes estão dispostos na rede da PE-2 (Figura 31).

Figura 31. Sobreposição de estados da Janela 2 na rede de RNA-seq PE-2. Os pontos em vermelho correspondem aos dados de entrada apenas de PE-2, os em azul claro correspondem aos dados de entrada apenas de S288c e os estados em roxo correspondem aos dados de entrada presentes nas duas redes.

Podemos observar nessa figura que três estados correspondentes aos dados de entrada da janela 2 (2 estados idênticos) encontram-se sobrepostos em 2 estados de RNA-seq, sendo estes estados possuem um grande número de estados direcionados a ele. Apesar de não possuírem o mesmo atrator, os atratores de ambos são visitados durante o ciclo.

Os estados com maior número de estados visualizados nessa figura são “1111111111”, “0111111111” e “0111101111”, onde o primeiro e o terceiro são os atratores da janela 2 e do PE-2 respectivamente. Nessa sequência, vemos todos os genes ligados, em seguida o gene HXK1 é desativado e em seguida o gene ENO1.

Observando os dados amostrais de S288c vimos que o estado “1111111111” permanece durante três amostras temporais (durante a fase entre 13-17 horas de fermentação). Já em PE-2 esse estado está presente apenas no tempo de 9 horas, isto é, o estado atrator de S288c é apenas um estado transiente em PE-2. Portanto, baseado nessas análises, podemos

concluir que a sequência de estados entre as duas linhagens parece ser similar, porém, através de algum mecanismo biológico, a PE-2 passa por esse estado mais rapidamente que a linhagem S288c.

As redes de regulação de RNA-seq em PE-2 (diagrama de estados da Figura 21) e da Janela 2 de S288c (diagrama de estados da Figura 25) podem ser observadas nas Figura 32 e Figura 33.

Figura 32. Rede de regulação de RNA-seq PE-2 do conjunto de 11 genes. As arestas em azul representam ativação e as arestas em vermelho representam repressão.

Figura 33. Rede de regulação da Janela 2 de S288c do conjunto de 11 genes. As arestas em azul representam ativação e as arestas em vermelho representam repressão.

4.2. Resultado das redes utilizando 12 genes

Para essa nova análise com conjuntos contendo 12 genes cada, geramos então 1.000 redes para os 19 conjuntos nas seis bibliotecas. Neste resultado (conjuntos com 12 genes) encontramos dificuldade em conseguir redes cuja maioria dos estados, incluindo os dados de entrada (estados), pertencesse à mesma bacia de atração. Outro efeito causado pela inserção de mais um gene foi o número muito pequeno de redes com entropia menor que 0,1 ($H < 0,1$).

Selecionamos então o conjunto de genes 12. Os diagramas de estados desse conjunto em cada biblioteca podem ser observados nas Figura 34 (RNA-seq PE-2), Figura 35 (microarray PE-2), Figura 36 (Janela 1 –S288c), Figura 37 (Janela 2- S288c), Figura 38 (Janela 3- S288c), Figura 39 (Janela 4- S288c).

Figura 34. Maior bacia da rede do conjunto 12 para dados de RNA-seq PE-2. Os pontos em verde representam os estados, os pontos em vermelho são os estados dos dados de entrada e o ponto em amarelo é o atrator. As arestas em vermelho mostram a trajetória do primeiro estado dos dados de entrada até o atrator da bacia.

Essa bacia possui 3.710 estados (dos 4096 possíveis) e seu atrator é o estado “010110001110” (ou 1422).

Figura 35. Maior bacia da rede do conjunto 12 para dados de microarray em PE-2. Os pontos em verde representam os estados, os pontos em vermelho são os estados dos dados de entrada e o ponto em amarelo é o atrator. As arestas em vermelho mostram a trajetória do primeiro estado até o atrator.

Essa bacia possui 2546 estados e seu atrator final é o estado “100001000111” (2119).

Figura 36. Maior bacia da rede do conjunto 12 para dados da Janela 1 em S288c. Os pontos em verde representam os estados, os pontos em vermelho são os estados dos dados de entrada e o ponto em amarelo é o atrator. As arestas em vermelho mostram a trajetória do primeiro estado até o atrator.

Figura 37. Maior bacia da rede do conjunto 12 para dados da Janela 2 em S288c. Os pontos em verde representam os estados, os pontos em vermelho são os estados dos dados de entrada e o ponto em amarelo é o atrator. As arestas em vermelho mostram a trajetória do primeiro estado até o atrator.

Figura 38. Maior bacia da rede do conjunto 12 para dados da Janela 3 em S288c. Os pontos em verde representam os estados, os pontos em vermelho são os estados dos dados de entrada e o ponto em amarelo é o atrator. As arestas em vermelho mostram a trajetória do primeiro estado até o atrator.

Figura 39. Maior bacia da rede do conjunto 12 para dados da Janela 4 em S288c. Os pontos em verde representam os estados, os pontos em vermelho são os estados dos dados de entrada e o ponto em amarelo é o atrator. As arestas em vermelho mostram a trajetória do primeiro estado até o atrator.

As redes das janelas 1 a 2 apresentaram o diagrama de estados das redes onde a maior bacia possui um número menor de atratores (1200 a 1700), as maiores bacias da janela 3 e 4 possuem 3567 e 3386 estados respectivamente. Os atratores são respectivamente; estado “111111100011” (4067), “111111111111” (4095), “111101111111” (3967), “111010100001” (3745).

Observando os atratores encontrados nos conjuntos de 11 genes e no conjunto de 12 genes (Figura 40), vimos que a inserção de apenas 1 gene causou pequenas mudanças nos estados dos genes nos atratores encontrados. Essa mudança é pequena (poucos estados), porém o último atrator das redes de S288c possuem vários estados divergentes. Isso pode ter sido causado pelo fato da última rede de S288c conter dados de amostras no processo conhecido como *diauxic shift* (mudança de fermentação para respiração) e, por não se tratar de um estado estacionário, a expressão de um gene aliado à um número pequeno de dados gera soluções distintas. Isso mostra que esse modelo de inferência é bastante sensível ao conjunto de dados de entrada.

Conjunto com 11 genes

RNA-seq:

4 hrs: 10000110000
7 hrs: 11110111111
10 hrs: 11111111111
12 hrs: 01110111111
15 hrs: 01111101111

Atratores (S288c):

15 hrs: 111111100011
17 hrs: 111111111111
19 hrs: 111101111111
21 hrs: 111010100001

X

Conjunto com 12 genes

RNA-seq:

4 hrs: 100001100000
7 hrs: 111101111111
10 hrs: 111111111111
12 hrs: 011111111111
15 hrs: 010110001110

Atratores (S288c):

15 hrs: 11111111111 – 11111110000
17 hrs: 11111111111
19 hrs: 11111111111
21 hrs: 01101010000

Figura 40. Comparação dos estados das redes de RNA-seq PE-2 e os atratores das redes de S288c no conjunto de 11 genes e no conjunto 12 (conjunto com 12 genes).

Analisando comparativamente as redes pelos métodos de similaridade e distância baseada em regras (Tabela 8) (Figura 41), podemos observar que a rede de RNA-seq PE-2 parece ser mais similar aos dados da Janela 1 e 3 (medida de distância). Quanto à topologia, as redes de PE-2 e a da Janela 2 são mais similares entre si.

	RNA-seq (PE2)	Microarray (PE2)	Janela1 (S288c)	Janela2 (S288c)	Janela3 (S288c)	Janela4 (S288c)
RNA-seq(PE2)	1	-	-	-	-	-
Microarray(PE2)	0,85	1	-	-	-	-
Janela 1 (S288c)	0,74	0,77	1	-	-	-
Janela 2 (S288c)	0,84	0,87	0,74	1	-	-
Janela 3 (S288c)	0,80	0,80	0,76	0,76	1	-
Janela 4 (S288c)	0,82	0,83	0,76	0,78	0,80	1

Tabela 8. Tabela comparativa usando medida de similaridade entre as bibliotecas do conjunto 12 com 12 genes.

Figura 41. Gráfico comparativo usando medida de distância baseada em regras entre as bibliotecas do conjunto 12 com 12 genes.

Como no conjunto anterior, sobrepomos os pontos das janelas 1 e 3 no diagrama de estados RNA-seq PE-2 e encontramos o mesmo efeito da sobreposição do conjunto anterior. Os dados de S288c parecem encaixar em uma pequena parte da sequência de dados de PE-2, evidenciando o metabolismo mais lento essa linhagem.

Além disso, outra observação foi encontrada quando analisamos os dados de entrada (estados) de S288c. Enquanto os dados de PE-2 parecem seguir um fluxo contínuo, os estados de S288c parecem estar “presos” há um pequeno ciclo de estados (a interseção dos estados de PE-2 e S288c) e portanto, esses estados estão mais presentes durante as janelas em S288c.

Figura 42. Sobreposição de estados correspondentes aos dados de entrada das Janelas 1 e 3 na rede de RNA-seq PE-2. Os pontos em vermelho são estados apenas de PE-2, os em rosa são estados apenas de S288c e os estados em amarelo são estados presentes nas duas redes.

4.3. Outros conjuntos de genes

Analisamos os dados gerados a partir dos outros conjuntos de genes e concluímos que o modelo mostrou-se bastante sensível ao conjunto de dados e por isso, a modelagem utilizando os diferentes conjuntos de genes produziram diferentes resultados. Um aspecto que identificamos é que com a adição desse

novo gene, poucas redes atingiram o valor de entropia abaixo de 0,1 (como o definido anteriormente).

4.4. Discussões

Conseguimos modelar com sucesso redes utilizando o conjunto com 11 genes principais pertencentes ao processo de fermentação, gerando soluções consistentes com os dados de entrada e com o processo biológico da fermentação. Quando adicionamos um gene em cada conjunto pertencente à outras vias ou processos, o modelo não foi capaz de soluções satisfatórias, encontrando muitas redes com alta entropia

Acreditamos que o modelo aplicado é bastante sensível aos dados e a inserção de genes de outras vias a um conjunto conservado de genes desestabilizou redes em algumas bibliotecas. A maior parte desses genes adicionados à rede inicial estão envolvidos em processos secundários ao processo de fermentação e, portanto, acreditamos que para uma modelagem satisfatória seria necessário acrescentar à rede genes que estejam juntamente envolvido no processo ao qual o gene inicial pertence ou que todos os genes envolvidos à essa nova via (famílias gênicas) sejam utilizados.

Capítulo 5

Conclusão

Neste trabalho analisamos dados de duas linhagens de levedura da espécie *Saccharomyces cerevisiae*: PE-2 e a S288c. Utilizamos dados experimentais realizados nas condições de fermentação e respiração (Derisi, et al., 1997; Netto, 2010) para analisar a dinâmica da fermentação dessas duas linhagens utilizamos o modelo de Redes Booleanas proposto por (Li, et al., 2004). No trabalho de (Li, et al., 2004), foi possível modelar com sucesso a rede do ciclo celular de levedura. Portanto, baseado neste trabalho aplicamos o modelo de Redes Booleanas para gerar redes a partir dos dados amostrais dessas linhagens.

Entretanto, a inferência de redes a partir de dados biológicos é um problema mal-posto. Dado um conjunto de dados experimentais é possível encontrar mais de uma solução ou até nenhuma solução que valide os dados amostrais. Esse problema torna-se mais complexo quando tentamos inferir redes com um número de amostras pequeno para um número maior de genes. Devido ao elevado custo em realizar experimentos de expressão gênica, é comum encontrarmos experimentos com poucas amostras temporais, como no caso dos nossos dados. O baixo número de amostras dos nossos dados resultou num grande número de soluções (linhas possíveis). Sendo assim, adotamos alguns critérios para seleção de genes como estabilidade da rede (entropia) e número de estados da maior bacia.

A via da fermentação é bastante conservada e, portanto, reduzimos o número de genes para nossa análise ao máximo (11 genes). Com esse número de genes modelamos com sucesso redes para as seis bibliotecas de dados. Entretanto, quando tentamos adicionar genes envolvidos mais indiretamente na via da fermentação, as redes perderam a estabilidade e poucas soluções satisfaziam as restrições impostas (entropia e número de estados da maior bacia).

Quanto aos dados biológicos, observamos que as redes de PE-2 (RNA-seq) possuem maior estabilidade, provavelmente pelo número maior de

amostras. Vimos também que as redes geradas da linhagem S288c parecem pertencer às redes encontradas da linhagem PE-2, porém estas parecem ciclar em estados que são apenas transitórios na linhagem PE-2 sugerindo então que a linhagem PE-2 fermenta mais rapidamente que a linhagem S288c. Observando os valores dos genes encontrados nos atratores do conjunto de 11 genes sugerimos que a PE-2 consegue captar açúcar do meio mais rapidamente que na linhagem S288c e, apesar de também fermentar mais rapidamente que a S288c, há um “gargalo” onde a velocidade de captação do açúcar do meio parece ser bem maior que a velocidade de fermentação desse açúcar.

Dada a complexidade do algoritmo (exponencial a o número de genes), não foi possível gerar redes com um número maior de genes e buscar pelas diversas combinações de genes necessárias que possam explicar melhor as interações dos genes dessa via devido ao tempo e número de combinações que seriam necessárias. Apenas na etapa para gerar as matrizes de regulação, um conjunto com 12 genes consome em média de 40 minutos por biblioteca de dados para as 1.000 redes (esse valor varia de acordo com o número de amostras – tempo aproximado utilizando um servidor com dois processadores *quadricore*). Esse tempo elevado para gerar as redes limitou o nosso conjunto de testes.

Como sugestão para futuros trabalhos então sugiro a utilização de um número maior de genes mesmo com o tempo elevado de processamento e também diferentes combinações de genes envolvidos direta e indiretamente à via da fermentação. Quanto a sugestões biológicas, sugiro um foco maior no mecanismo de transporte e fosforilação do açúcar (dadas à expressão divergentes entre as linhagens ao longo do tempo) e os genes iniciais do processo de fermentação (“gargalo”) na linhagem PE-2.

Apêndice A

Dados normalização de placas de *microarray* em PE-2

Neste apêndice, apresentamos os gráficos dos dados das nove placas (três tempos com três réplicas cada) antes e após a normalização. (seção 3.3.1)

Figura A.1 Gráfico da normalização em escala para as nove placas. (a) Dados das nove placas antes da normalização. (b) Dados das nove placas após a normalização.

Figura A.2 Gráfico de blocos antes da normalização para as nove placas. (a) Tempo 3: réplicas 1, 2 e 3.(b) Tempo 6: réplicas 1,2 e 3.(c) Tempo 9: réplicas 1,2, e 3.

Figura A.3 Gráfico de blocos após a normalização para as nove placas. (a) Tempo 3: réplicas 1, 2 e 3.(b) Tempo 6: réplicas 1,2 e 3.(c) Tempo 9: réplicas 1,2, e 3.

Apêndice B

Lista dos genes para cada conjunto

Conjunto de genes 1

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YJL009C.

Conjunto de genes 2

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YGR192C.

Conjunto de genes 3

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YLR134W.

Conjunto de genes 4

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YDR510W.

Conjunto de genes 5

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YER151C.

Conjunto de genes 6

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YGR087C.

Conjunto de genes 7

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YHR200W.

Conjunto de genes 8

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YHR027C.

Conjunto de genes 9

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YFR004W.

Conjunto de genes 10

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YCL040W.

Conjunto de genes 11

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YDR127W.

Conjunto de genes 12

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YNR051C.

Conjunto de genes 13

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YLL039C.

Conjunto de genes 14

YFR053C, YBR196C, YGR240C, YKL060C, YJL052W, YCR012W, YGR254W, YHR174W, YAL038W, YLR044C, YOL086C, YBR126C.

Conjunto de genes 15

YFR053C,YBR196C,YGR240C,YKL060C,YJL052W,YCR012W,YGR254W,YH
R174W,YAL038W,YLR044C,YOL086C,YGL122C.

Conjunto de genes 16

YFR053C,YBR196C,YGR240C,YKL060C,YJL052W,YCR012W,YGR254W,YH
R174W,YAL038W,YLR044C,YOL086C,YBR194W.

Conjunto de genes 17

YFR053C,YBR196C,YGR240C,YKL060C,YJL052W,YCR012W,YGR254W,YH
R174W,YAL038W,YLR044C,YOL086C,YGL253W.

Conjunto de genes 18

YFR053C,YBR196C,YGR240C,YKL060C,YJL052W,YCR012W,YGR254W,YH
R174W,YAL038W,YLR044C,YOL086C, YMR205C.

Conjunto de genes 19

YFR053C,YBR196C,YGR240C,YKL060C,YJL052W,YCR012W,YGR254W,YH
R174W,YAL038W,YLR044C,YOL086C, YKL152C.

Referências

- Andrade, Tales Pinheiro.** 2010. *Interações gênicas usando Redes Booleanas Restritas no contexto de Problema de Satisfação de Restrições*. Universidade São Paulo. São Paulo : s.n., 2010.
- Argueso, Juan Lucas, Carazzolle, Marcelo and Mieczkowski, Piotr.** 2009. Genome structure of a *Saccharomyces Cerevisiae* strain widely used in bioethanol production. *Genome Research*. 2009.
- Bioconductor. [Online] <http://www.bioconductor.org/>.
- BIOGRID. *Biological General Repository of Interaction DataSets*. [Online] <http://thebiogrid.org/>.
- Brooks, Akan A.** 2008. Ethanol production potential of local yeast strains isolated from ripe banana peels. *African Journal of Biotechnology*. 2008, Vol. 7, 20, pp. 3749–3752.
- Conant, G and Wolfe, K.** 2007. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Molecular Systems Biology*. 2007, Vol. 129.
- Derisi, Joseph L, Iyer, Vishwanath R and Brown, Patrick O.** 1997. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. 1997, Vol. 5338, pp. 680–686.
- Dougherty, Edward R.** 2007. Validation of Inference Procedures for Gene Regulatory Networks. *Current Genomics*. 2007, pp. 351–359.
- Empresa de pesquisa energética. [Online] [Cited: janeiro 17, 2011.] <http://www.epe.gov.br>.
- Espinosa-Soto, Carlos, Padilla-Longoria , Pablo and Alvarez-Buylla, Elena.** 2004. A gene regulatory network model for Cell-Fate determination during arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. *Plant Cell*. 2004, Vol. 16, pp. 2923–2939.
- Derisi, Joseph L, Iyer, Vishwanath R and Brown, Patrick O.** 1997. *Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale..* 1997, Science, Vol. 5338, pp. 680–686.
- Filho, Diógenes Ferreira and Leandro, Roseli Aparecida.** 2009. *Tutorial: Análise de Microarray usando o R e o Bioconductor*. São Carlos : s.n., 2009.

- Friedman, N, et al. 2000.** Using Bayesian Network to Analyze Expression Data. *Journal of Computational Biology*. 2000, Vol. 7, pp. 601–620.
- Romero, Thiago. 2007.** *Genoma da levedura de cana de açúcar é sequenciado*. outubro 2007, FAPESP.
- Grupo de processos biotecnológicos. 2004.** Microorganismos: Levedura. [Online] 2004. [Cited: janeiro 11, 2011.] http://www.enq.ufsc.br/labs/probio/disc_eng_bioq/trabalhos_grad2004/microorganismos/leveduras.htm.
- Hecker, Michael, et al. 2009.** Gene regulatory network inference: Data integration in dynamic models – a review. *BioSystems*. 1, 2009, Vol. 96, pp. 86–103.
- Higa, Carlos Henrique. 2012.** Inferência de redes de regulação gênica utilizando o paradigma de crescimento de sementes. São Paulo : Universidade São Paulo, 2012.
- Higa, Carlos, et al. 2010.** Constraint-based analysis of gene interactions using re-stricted Boolean networks and time-series data. 2010.
- Illumina. 2011.** RNA-seq data comparison with Gene Expression Microarrays. *White paper: sequencing*. 2011.
- Jensen, Niels. 2012.** The Case For Agriculture As The Next Big Asset Class. *Business Insider*. [Online] março 2012. <http://www.businessinsider.com/the-case-for-agriculture-as-the-next-big-asset-class-2012-3>.
- Jong, H D. 2002.** Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *Journal of Computational Biology*. 2002, Vol. 9, pp. 67–103.
- Karlebach, Guy and Shamir, Ron. 2008.** Modelling and analysis of gene regulatory network. *Nature*. 2008, Vol. 9, pp. 770–780.
- Kauffman. 1969.** Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets. *Journal of Theoretical Biology*. 1969, Vol. 22, pp. 437–467.
- Kauffman, S A. 1969.** Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets. *Journal of Theoretical Biology*. 1969, Vol. 22, pp. 437–467.
- KEGG.** Kyoto Encyclopedia of Genes and Genomes. [Online] <http://www.genome.jp/kegg/>.
- Kohavi, R and Provost, F. 1998.** Glossary of terms. *Machine Learning*. 30, 1998, pp. 271–274.

Langmead, Ben, et al. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. 2009, Vol. 10, p. 212.

Lau, Kai-Yeung, Ganguli, Surya and Tang, Chao. 2007. Function constrains network architecture and dynamics: A case study on the yeast cell cycle Boolean network. *Physical review*. 2007, Vol. 75.

Li, F, et al. 2004. The Yeast Cell-Cycle Network is Robustly Designed. *Proceedings of the National Academy of Sciences of the United States of America*. 2004, Vol. 101, pp. 4781–4786.

Lopes, Fabrício Martins. 2011. Redes complexas de expressão gênica: síntese, identificação, análise e aplicações. Universidade São Paulo : s.n., 2011.

Mayer, Lauri. 2009. Analisando a expressão de milhares de genes simultaneamente. *Webartigos*. [Online] 2009. [Cited: junho 2011, 29.] <http://www.webartigos.com/articles/17028/1/Tecnologia-microarray-biochips-de-DNA/pagina1.html>.

Mestl, T, Plahte, E and Omholt, S. 1995. A Mathematical Framework for describing and Analyzing Gene Regulatory Networks. *Journal of Theoretical Biology*. 1995, Vol. 176, pp. 291–300.

NCBI. *National Center for Biotechnology Information*. [Online] <http://www.ncbi.nlm.nih.gov/>.

Netto, Osmar V C. 2010. *Fisiologia da Fermentação Alcoólica: Análise da expressão gênica de uma linhagem industrial de Saccharomyces cerevisiae durante o processo fermentativo*. FAPESP. Campinas : s.n., 2010.

Pagliari, E. 2010. Brasil busca liderança na produção do biocombustível. *Portal brasileira de energias renováveis*. [Online] 2010. [Cited: janeiro 13, 2011.] <http://www.energiarenovavel.org/>.

Liti, G, Carter, D and Moses, A. 2009. *Population genomics of domestic and wild yeasts*. 2009, Nature.

Ribeiro, Antonio Cláudio Bello. 2010. Artigo “RNA-Seq: a revolutionary tool for transcriptomics.”. *System Biology and Computacional Biology*. [Online] 2010. [Cited: julho 29, 2011.] <http://systemsbiology.biowebdb.org/?p=634>.

Wang, Zhong, Gerstein, Mark and Snyder, Michael. 2009. *RNA-Seq: a revolutionary tool for transcriptomics*. 2009, Nature Reviews Genetics 10 , pp. 57–63.

Russel, S J and Norvig, P. 2002. *Artificial Intelligence: A modern approach*. 2. s.l. : Prentice Hall, 2002. pp. 1-24.

Saccharomyces Genome Database. [Online] <http://www.yeastgenome.org/>.

Schena, Mark, et al. 1995. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*. 1995, Vol. 270, pp. 467-470.

Velculescu, Victor E, et al. 1995. *Serial Analysis of Gene Expression*. outubro 20, 1995, Science, Vol. 270, pp. 484-487.

Shannon, C and Weaver, W. 1963. The mathematical theory of communication. *University of Illinois Press*. 1963, 23.

Shannon, Claude. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*. 1948, Vol. 27.

Shmulevich, I, et al. 2002. Probabilistic Boolean Networks: A rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 2002, Vol. 18, pp. 261-274.

Shmulevich, Ilya and Zhang, Wei. 2002. Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics*. 2002, Vol. 18, pp. 555-565.

Smyth, Gordon. Limma: Linear Models for Microarray Data.

Thomas, R. 1991. Regulatory Networks seen as Asynchronous Automata: A Logical Description. *Journal of Theoretical Biology*. 1991, Vol. 153, pp. 01-23.

Tsang, Edward. 1993. *Foundations of Constraint Satisfaction*. 1993. pp. 1-30.

Velculescu, Victor E, et al. 1995. Serial Analysis of Gene Expression. *Science*. outubro 20, 1995, Vol. 270, pp. 484-487.

Wang, Likun and Wang, Xi. 2010. How to use the DEGseq Package. *Bioconductor*. [Online] 2010. <http://bioconductor.org/packages/2.6/bioc/vignettes/DEGseq/inst/doc/DEGseq.pdf>.

Wang, Likun, et al. 2010. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*. 2010, Vol. 26, pp. 136-138.

Wang, Zhong, Gerstein, Mark and Snyder, Michael. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009, pp. 57-63.

Webb, A R. 2002. Statistical Pattern Recognition. 2nd edition, 2002.