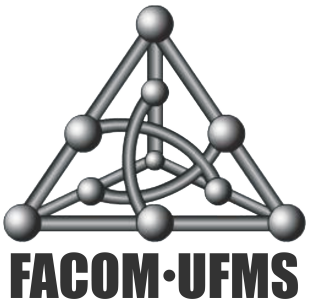
Engenharia reversa de redes de regulação gênica por meio de modelos gráficos probabilísticos

Mariana Caravanti de Souza

Dissertação apresentada
à
Faculdade de Computação
da
Universidade Federal de Mato Grosso do Sul
para
obtenção do título de Mestre
em
Ciência da Computação



Orientador: Prof. Dr. Carlos Henrique Aguena Higa

Área de Concentração: Ciência da Computação

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

Campo Grande, 4 de junho de 2018

Engenharia reversa de redes de regulação gênica por meio de modelos gráficos probabilísticos

Esta é a versão original da dissertação elaborada pela
Mariana C. de Souza, tal como
submetida à Comissão Julgadora.

Comissão Julgadora:

- Dr. Carlos Henrique Aguena Higa (presidente)
- Dr. Augusto Cesar de Aquino Ribas
- Dr. Marco Aurélio Stefanes

Agradecimentos

Agradeço primeiramente a Deus, pois sem Ele não chegaria até aqui. Meus mais profundos agradecimentos e admiração aos meus pais, Jorge e Marli, que sempre me deram todo o amor e suporte necessário, mesmo em meio às suas próprias lutas diárias. Agradeço aos meus padrinhos, Lourdes e Osano que, mesmo de longe, nunca deixaram de me amparar. Agradeço ao meu namorado, Thiago, por me fazer sorrir nos momentos de dificuldade. Agradeço ao meu orientador, Carlos, por todos os anos de trabalho juntos, por sua humanidade, e pelas vezes que conseguiu sanar minhas dúvidas disfarçadas de frases confusas. Agradeço também ao professor Francisco, que sempre esteve disponível para boas conversas e conselhos sobre o futuro. Agradeço à FACOM, por me fornecer toda a estrutura necessária e oportunidade de estudar em uma Universidade Federal. Obrigada a todos os professores que tive o privilégio de conhecer e aprender algo novo, e por sempre estarem dispostos a ajudar. Agradeço também ao Centro Tecnológico de Eletrônica e Informática (CTEI) da UFMS pela disponibilidade de clusters para a execução de meus algoritmos. Sem vocês, nada disso seria possível.

Resumo

DE SOUZA, M.C. **Engenharia reversa de redes de regulação gênica por meio de modelos gráficos probabilísticos**. 2018. 84 f. Dissertação (Mestrado) - Faculdade de Computação, Universidade Federal de Mato Grosso do Sul, Campo Grande, 2018.

As tecnologias de Sequenciamento de Nova Geração permitem que um grande volume de dados biológicos seja gerado. Sendo assim, metodologias matemáticas e computacionais tornaram-se essenciais para analisar e extrair informações relevantes desses dados, a fim de gerar conhecimento. Um problema importante estudado em Biologia Sistêmica e Bioinformática é o de engenharia reversa de redes de regulação gênica, conhecido também como inferência de redes, a partir de dados biológicos. Inferir redes nos leva a gerar hipóteses de regulação presentes em um determinado fenômeno biológico em questão. Neste trabalho, realizamos a inferência de redes utilizando os modelos de redes Bayesianas estático e dinâmico a partir de dados de expressão gênica temporais. Foram estudadas diferentes funções de pontuação responsáveis por selecionar a melhor estrutura de rede que representa o comportamento dos dados. Para a validação da metodologia, foram utilizados dados sintéticos, que já possuem a rede de regulação conhecida, e também utilizamos dados reais do ciclo celular da levedura. Além disso, propomos a incorporação de conhecimento biológico *a priori* ao algoritmo de inferência, com o intuito de aprimorar os resultados obtidos, e testamos a eficácia da metodologia conforme o número de amostras disponíveis aumenta. O trabalho foi desenvolvido estendendo as funcionalidades da biblioteca Pgmpy, uma biblioteca implementada na linguagem de programação Python, que lida com modelos gráficos probabilísticos.

Palavras-chave: Redes de Regulação Gênica, Redes Bayesianas, Redes Bayesianas Dinâmicas, Expressão Gênica, Engenharia Reversa.

Abstract

DE SOUZA, M. C. **Reverse engineering of gene regulatory networks through probabilistic graphical models**. 2018. 84 f. Dissertation (Mester) - Faculdade de Computação, Universidade Federal de Mato Grosso do Sul, Campo Grande, 2018.

Next-Generation Sequencing technologies allow a large volume of biological data to be generated. Thus, mathematical and computational methodologies have become essential to analyze and extract relevant information from this data in order to produce knowledge. An important problem that has been studied in Systems Biology and Bioinformatics is the reverse engineering of gene regulatory networks, also known as network inference, from biological data. Inferring networks can lead us to reason about regulation hypotheses present in a particular biological phenomenon in question. In this work, we performed the network inference using static and dynamic Bayesian networks as model for learning networks through time-series gene expression data. We studied different functions responsible for selecting the best network structure that represents the data. To validate the methodology, synthetic data were used, in which the true network is already known, and we also used real data of the yeast cell cycle. In addition, we incorporate prior biological knowledge in the inference algorithm to improve the results and tested the methodology effectiveness as the number of samples grows. This work was developed by extending the features of the Pgmpy library, developed in Python programming language, that deals with probabilistic graphical models.

Keywords: Gene Regulatory Network, Bayesian Network, Gene Expression, Reverse Engineering.

Sumário

Lista de Abreviaturas

BDe	*Bayesian Dirichlet likelihood-equivalence*
BIC	*Bayesian information criterion*
BN	Rede Bayesiana (*Bayesian Network*)
CPD	Distribuição de Probabilidade Condicional (*Conditional Probability Distribution*)
DAG	Grafo Acíclico Direcionado (*Directed Acyclic Graph*)
DBN	Redes Bayesianas Dinâmicas (*Dynamic Bayesian Network*)
DNA	Ácido Desoxirribonucleico (*Deoxyribonucleic Acid*)
FN	Falso Negativo (*False Negative*)
FP	Falso Positivo (*False Positive*)
GRN	Rede de Regulação Gênica (*Gene Regulatory Network*)
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
MLE	Estimativa de Máxima Verossimilhança (*Maximum Likelihood Estimation*)
mRNA	RNA Mensageiro (*Messenger RNA*)
NCBI	*National Center for Biotechnology Information*
PPV	Precisão (*Positive Predictive Value*)
RNA	Ácido Ribonucleico (*Ribonucleic acid*)
STRING	*Search Tool for the Retrieval of Interacting Genes/Proteins*
TN	Verdadeiro Negativo (*True Negative*)
TP	Verdadeiro Positivo (*True Positive*)

Lista de Símbolos

$\mathcal{G}$	Grafo acíclico direcionado
$\mathcal{X}$	Conjunto de todas as variáveis aleatórias no domínio
$\mathcal{B}$	Rede Bayesiana
X, Y, Z	Variáveis aleatórias
$\text{Val}(X)$	Conjunto de valores que uma variável X pode assumir
$P(\cdot)$	Distribuição de probabilidade
$P^*(\cdot)$	Distribuição de probabilidade desconhecida que gerou os dados
$(X \perp Y \mid Z)$	Independência condicional: X é independente de Y, dado Z
$\text{Pa}_{X_i}^{\mathcal{G}}$	Pais da variável X_i em $\mathcal{G}$
$\text{NonDescendants}_{X_i}$	Conjunto de variáveis não descendentes de X_i no grafo $\mathcal{G}$
$\mathcal{D}$	Conjunto de dados amostrais
$\mathcal{X}[i]$	i-ésima instância de dados do conjunto $\mathcal{D}$
M	Número total de instâncias do conjunto de dados amostrais $\mathcal{D}$
$\boldsymbol{\theta}$	Parâmetros de uma distribuição de probabilidade
$L(\boldsymbol{\theta} : \mathcal{D})$	Função de verossimilhança
$M[\boldsymbol{x}]$	Contagem do evento $\boldsymbol{x}$ nos dados
$\hat{\boldsymbol{\theta}}$	Parâmetros MLE
Θ	Espaço de parâmetros
$\text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$	Distribuição de Dirichlet
$\mathcal{G}^*$	Rede desconhecida que gerou os dados (*gold standard*)
$\text{score}_L(\mathcal{G} : \mathcal{D})$	Pontuação de verossimilhança
$\ell(\hat{\boldsymbol{\theta}}_{\mathcal{G}} : \mathcal{D})$	Logaritmo da função de verossimilhança
Δ	Grau máximo de entrada de um vértice no grafo (rede)
ω	Peso atribuído à equação de conhecimento biológico
Γ	Função Gamma

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Engenharia reversa ou inferência de redes de regulação gênica pode ser definido como o processo de identificar quais são as interações existentes entre genes a partir de dados experimentais, utilizando métodos computacionais (Hecker *et al.*, 2009). Neste caso, dados experimentais são dados de expressão gênica, em geral, provenientes de microarrays (Shalon *et al.*, 1996) e, mais recentemente, de experimentos de RNA-Seq (Wang *et al.*, 2009). Quando analisados, tais dados são capazes de revelar muito sobre a possível estrutura de rede que governa o comportamento dos genes, e este conhecimento impulsiona novas pesquisas. A engenharia reversa de redes tem auxiliado importantes pesquisas relacionadas a doenças como câncer e Alzheimer. Recentemente, Gendelman *et al.* (2017) descobriu um novo gene regulador, responsável pela progressão e sobrevivência do ciclo celular em células de câncer de mama humano. Além disso, estudos da doença de Alzheimer, capaz de manifestar uma série de alterações nas interações moleculares e em redes que definem o processo biológico, têm sido realizados, buscando-se por novas sub-redes e genes reguladores ligados à doença, onde tais redes são inferidas por meio de dados de expressão (Zhang *et al.*, 2016).

Alguns modelos matemáticos/computacionais foram propostos nos últimos anos para auxiliar no processo de inferência de GRNs (*gene regulatory networks*) (De Jong, 2002, Hecker *et al.*, 2009). Entre os mais utilizados, estão as redes Booleanas, redes Bayesianas e modelos baseados em equações diferenciais ordinárias. Redes Booleanas (Kauffman, 1969) tornam-se atraentes quando o objetivo é obter um modelo simplificado e robusto. Embora haja perda de informação ao utilizar esta formulação, pesquisas mostram que muitas questões biológicas podem ser respondidas, mesmo quando os dados são tratados de forma discreta. Por outro lado, redes Bayesianas se mostram mais promissoras em análises de padrões de expressão em genes (Friedman *et al.*, 2000). Enquanto redes Bayesianas são modelos acíclicos, de inferência causal, matematicamente definidos em termos de probabilidades e independência condicional, temos o modelo de redes Bayesianas dinâmicas, que além de apresentarem estas mesmas características, lidam muito bem com dados temporais e possuem uma representação capaz de incorporar relações cíclicas entre as variáveis, ao longo do tempo (Friedman *et al.*, 1998). Já equações diferenciais ordinárias (Goodwin *et al.*, 1963) permitem uma modelagem mais

detalhada, levando em conta o comportamento dinâmico e a organização temporal dos genes, necessitando de uma quantidade maior de dados a serem analisados.

Além de utilizar modelos que tentem captar as interações existentes entre os genes a partir dos dados de expressão gênica, pode-se incorporar conhecimento biológico *a priori*, proveniente de banco de dados públicos na metodologia de inferência (Hecker *et al.*, 2009). Exemplos são o *STRING* (Szklarczyk *et al.*, 2016), um banco de dados conhecido para predizer interações diretas (físicas) e indiretas (funcionais) entre proteínas; o *KEGG* (Kanehisa *et al.*, 2017), uma enciclopédia de genes e genomas que armazena funções moleculares e o *Jaspar* (Bryne *et al.*, 2008), que contém perfis de conjuntos de sequências nucleotídicas demonstrados experimentalmente.

Neste trabalho, realizamos a engenharia reversa de GRNs por meio de algoritmos de aprendizagem, utilizando os modelos de redes Bayesianas e redes Bayesianas dinâmicas e, adicionamos conhecimento biológico proveniente de bancos de dados públicos ao processo de inferência. Para isso, utilizamos uma biblioteca implementada em *Python* chamada Pgmpy, que já possui algumas funcionalidades necessárias, como a implementação do modelo de redes Bayesianas. Logo, esta biblioteca será estendida de forma a possuir o modelo de redes Bayesianas dinâmicas e permitir que estruturas de redes que possuam relações biológicas comprovadas experimentalmente sejam priorizadas. O conhecimento biológico será incorporado à funções responsáveis por atribuir pontuações a estruturas de redes candidatas, a fim de encontrar a melhor rede que represente o comportamento dos dados. As funções de pontuação estudadas neste trabalho são as funções BIC e BDe. Além disso, foram propostas variações em parâmetros que interferem no processo de inferência, como a dimensão máxima que a rede pode atingir e a atribuição de diferentes pesos à equação de conhecimento biológico definida. Também analisamos os efeitos das modificações propostas conforme o número de amostras disponíveis aumenta. Para a validação da metodologia, serão utilizados dados sintéticos e biológicos que possuam redes já conhecidas. Exemplos são os dados de *DREAM Challenge* (Greenfield *et al.*, 2010) e a rede do ciclo celular da levedura, proposta por Spellman *et al.* (1998).

Esta dissertação está organizada da seguinte forma: A fundamentação deste trabalho, incluindo conceitos de sequenciamento de nova geração, redes de regulação gênica, redes Bayesianas e estimativa de parâmetros, é definida no Capítulo 2. No Capítulo 3, realizamos a definição formal do problema de engenharia reversa de redes e uma breve revisão bibliográfica. O Capítulo 4 contém a metodologia deste trabalho e as modificações propostas para a biblioteca utilizada no processo de inferência de redes. Além disso, são estabelecidos os dados de expressão gênica utilizados, como adicionar conhecimento biológico ao algoritmo e como validar os resultados obtidos. As redes, análises e gráficos comparativos podem ser observados no Capítulo 5 e a conclusão do trabalho é fornecida no Capítulo 6.

Capítulo 2

Fundamentação Teórica

Nesta capítulo são definidos os conceito de expressão gênica e redes de regulação gênica. Além disso, apresentamos dois modelos gráficos probabilísticos: redes Bayesianas e redes Bayesianas dinâmicas, ambos utilizados no processo de engenharia reversa de redes. Também são definidas maneiras de se calcular o modelo de probabilidade local que representa o comportamento de uma rede, utilizando as estimativas de parâmetros de máxima verossimilhança e Bayesiana.

2.1 Expressão Gênica

Expressão gênica é o processo pelo qual a informação hereditária contida em um gene, como a sequência de DNA, é processada em um produto gênico final, na qual este produto gênico geralmente é uma proteína. Este processo é realizado por todas as formas de vida conhecidas para que os organismos se mantenham vivos e ocorre em duas etapas: a primeira consiste na produção de moléculas de RNA a partir do DNA (transcrição) e a segunda na produção de proteínas a partir deste RNA (tradução).

Uma molécula de **ácido desoxirribonucleico (DNA)** é composta por duas cadeias polipeptídicas, onde tais cadeias são conhecidas como *cadeias de DNA* ou *fitas de DNA*. As cadeias de DNA são mantidas unidas por meio de ligações de hidrogênio. Sua base é formada por adenina (A), citosina (C), guanina (G) e timina (T). Em uma célula, o DNA é replicado a partir de um molde preexistente formado por uma fita de DNA. As bases desta fita preexistente ligam-se com as bases da fita que está sendo sintetizada, onde As ligam-se com Ts, e Cs ligam-se com Gs.

Segundo Alberts *et al.* (2009), para cumprir a função de armazenamento de informação, o DNA deve ser capaz de fazer mais do que cópias de si mesmo. Ele também deve *expressar* sua informação, permitindo que esta guie a síntese de outras moléculas da célula. O processo começa com uma polimerização a partir de um molde chamada de transcrição, na qual segmentos da sequência de DNA são usados como moldes para guiar a síntese de moléculas de **ácido ribonucleico**, ou **RNA**. Assim, cada molécula de RNA produzida equivale a um

gene transcrito, ou seja, equivale à informação contida em uma sequência específica de RNA. Assim, quando um gene é transcrito, dizemos que este gene está *expresso*, caso contrário, ele está *inibido*.

O RNA possui estrutura distinta do DNA: ao invés de possuir desoxirribose em sua formação, este possui a ribose. Também há uma diferença na formação de suas bases, pois ele contém uracila (U) ao invés de timina (T). As três outras bases (A, C e G) são equivalentes ao DNA, e as bases se pareiam com suas contra-partes complementares no DNA: A, U, C e G do RNA com T, A, G e C do DNA, respectivamente.

Após o processo de transcrição do gene, há a síntese proteica. A síntese proteica é o principal objetivo da expressão gênica. Ela ocorre no citoplasma das células e é feita a partir das moléculas de RNA mensageiros (mRNA) (Figura 2.1). Segundo Claudia Moraes (2016), cada molécula de RNA contém a informação de um gene e assim cada proteína sintetizada a partir dele é produto deste gene. Ao processo de síntese de proteínas a partir de moléculas de mRNA, é dado o nome de tradução. Ao final dele, os mRNAs que já foram lidos algumas vezes são degradados. Caso seja necessária a síntese dessas proteínas novamente, novos mRNAs serão transcritos. Assim, a célula pode controlar a quantidade de proteínas que são sintetizadas.

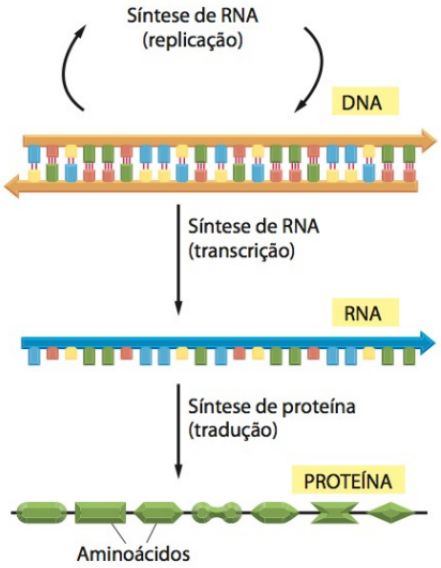


Figura 2.1: ***Do DNA à proteína.*** *A informação é lida e processada em duas etapas. Primeiro, na transcrição, os segmentos de uma sequência de DNA são usados para guiar a síntese de moléculas de RNA. Depois, na tradução, as moléculas de RNA são usadas para guiar a síntese de moléculas de proteínas. (Alberts et al., 2009)*

O mecanismo que controla a ativação de um grupo de genes em cada uma das diferentes células de um mesmo organismo é chamado controle da expressão gênica. Existem diversos mecanismos de controle de expressão gênica. O mais comum é o controle transcricional, ou seja, aquele que impede que o DNA seja transcrito em um mRNA. Este controle acontece pela ação de proteínas, que são chamadas proteínas reguladoras da expressão gênica.

As proteínas são capazes de se associar quimicamente a trechos específicos da molécula

de DNA. Normalmente, os trechos específicos envolvidos no controle de um determinado gene são próximos ao trecho do DNA que codifica aquele gene. Há proteínas reguladoras que são ativadoras ou repressoras da expressão gênica. As proteínas ativadoras, quando se ligam ao DNA, promovem a transcrição do gene. As repressoras inibem a transcrição do gene. É por meio desses movimentos de ligação e desligamento das proteínas reguladoras no DNA que elas controlam a expressão de um gene.

Os organismos são estruturados química e biologicamente para, em geral, otimizar os processos e uso de recursos (energia) para manutenção da vida. O controle da expressão gênica é um mecanismo fundamental para que essa otimização aconteça e, mais do que isso, para que o organismo funcione de maneira adequada. Os diferentes tipos celulares em um organismo multicelular diferem dramaticamente tanto em estrutura como em função (Figura 2.2). A diferenciação celular geralmente depende de mudanças de expressão gênica e não quaisquer alterações na sequência de nucleotídeos do genoma da célula (Claudia Moraes, 2016).

Uma forma de se obter os níveis de expressão de um determinado organismo é a utilização de mecanismos como microarrays (Shalon *et al.*, 1996) e, mais recentemente, por meio de Sequenciamento de Nova Geração. Mais detalhes a respeito desta técnica serão vistos na próxima seção.

2.1.1 Sequenciamento de Nova Geração

Para que níveis de expressão de genes sejam analisados, é necessário o estudo do transcriptoma, ou seja, da coleção de RNAs (transcritos) e proteínas que uma célula produz em um determinado período. Os tipos de RNAs podem ser RNAs mensageiros, RNAs não codificantes e microRNAs. Com o auxílio de ferramentas recentes, como microarrays e RNA-Seq, é possível realizar análises de perfis de transcriptomas de milhares de genes de forma rápida e eficaz. Um exemplo de uso dessas ferramentas, é conseguir uma comparação entre células saudáveis e células doentes, por meio de uma analogia dos níveis de expressão gênica presentes em ambas as amostras.

A Figura 2.3 ilustra os passos típicos de um processo de sequenciamento por RNA-Seq. Primeiro, o RNA (azul claro) é extraído (fase 1) de um organismo, e o DNA contaminante é removido (fase 2). O RNA restante é quebrado em pequenos fragmentos (fase 3). Os fragmentos são transcritos de maneira reversa em cDNA (amarelo, fase 4) e adaptadores (azul) são ligados (fase 5) às suas extremidades. É realizada a seleção dos fragmentos de acordo com seus tamanhos (fase 6). Finalmente, as duas extremidades são sequenciadas utilizando tecnologias de Sequenciamento de Nova Geração para produzir *paired-end reads* (Martin e Wang, 2011).

Após o sequenciamento, os *reads* são pré-processados e "montados" em transcritos. Os transcritos são processados e o nível de expressão de cada transcrito é estimado, contando o número de *reads* que foram alinhadas àquele transcrito. Assim, obtém-se os dados de

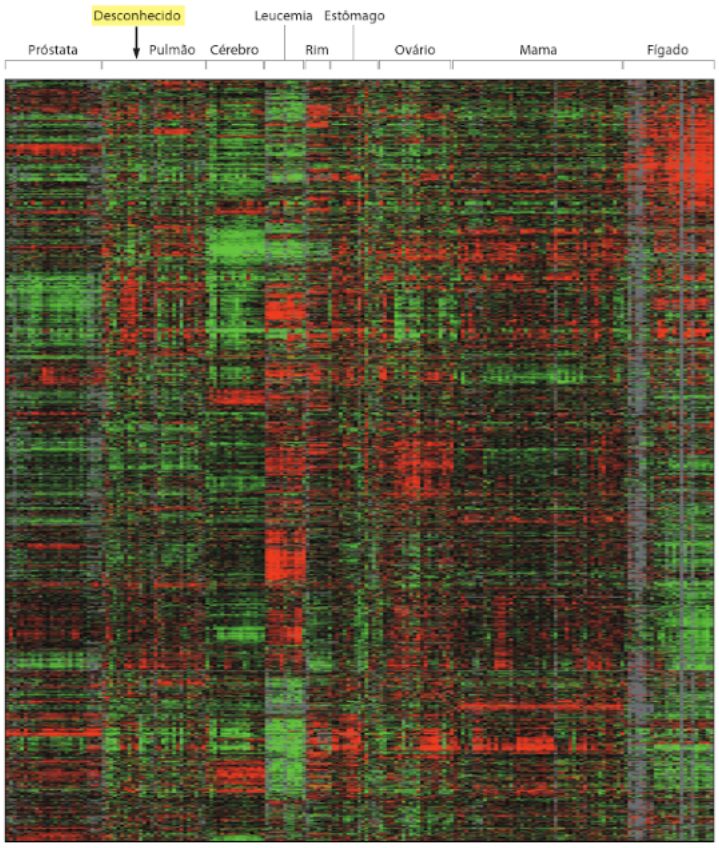


Figura 2.2: ***Diferenças no padrão de expressão de mRNAs entre diferentes tipos de células cancerosas humanas.*** *Esta figura resume o grande conjunto de medidas nas quais os níveis de mRNA de* 1800 *genes selecionados (arranjados de cima para baixo) foram determinados para 142 tumores humanos diferentes (arranjados da esquerda para a direita), cada um de um paciente diferente. Cada barra pequena vermelha indica que um determinado gene em um determinado tumor é transcrito em um nível significativamente maior do que a média entre todas as linhagens celulares. Cada barra verde pequena indica um nível de expressão menor do que a média, e cada barra negra indica um nível de expressão semelhante à média entre os diferentes tumores. O procedimento usado para gerar esses dados foi isolamento de mRNA seguido por hibridização de microarrays de DNA. A figura mostra que os níveis de expressão relativa de cada um dos* 1800 *genes analisados varia entre os diferentes tumores (visto seguindo-se um determinado gene da esquerda para a direita ao longo da figura.) Essa análise também mostra que cada tipo de tumor possui um padrão de expressão gênica característico. Essa informação pode ser usada para "tipar" células cancerosas de origem desconhecida pela comparação dos perfis de expressão gênica com os dos tumores conhecidos. Por exemplo, na figura uma amostra desconhecida foi identificada como um câncer de pulmão (Alberts et al., 2009).*

expressão gênica de um organismo.

Os dados de expressão gênica obtidos exigem recursos e técnicas computacionais que os manipulem, a fim de gerar conhecimento. A partir de dados de expressão, utilizando metodologias matemáticas e computacionais, é possível tirar conclusões a respeito de possíveis relações entre genes, já que muitas das atividades celulares são organizadas como uma rede de interações (Kauffman, 1969, Friedman *et al.*, 2000). Assim, muitos estudos buscam encontrar qual é a *rede de regulação gênica* que deu origem aos dados de expressão. Mais adiante, definiremos os conceitos de rede de regulação gênica e apresentaremos modelos matemáticos

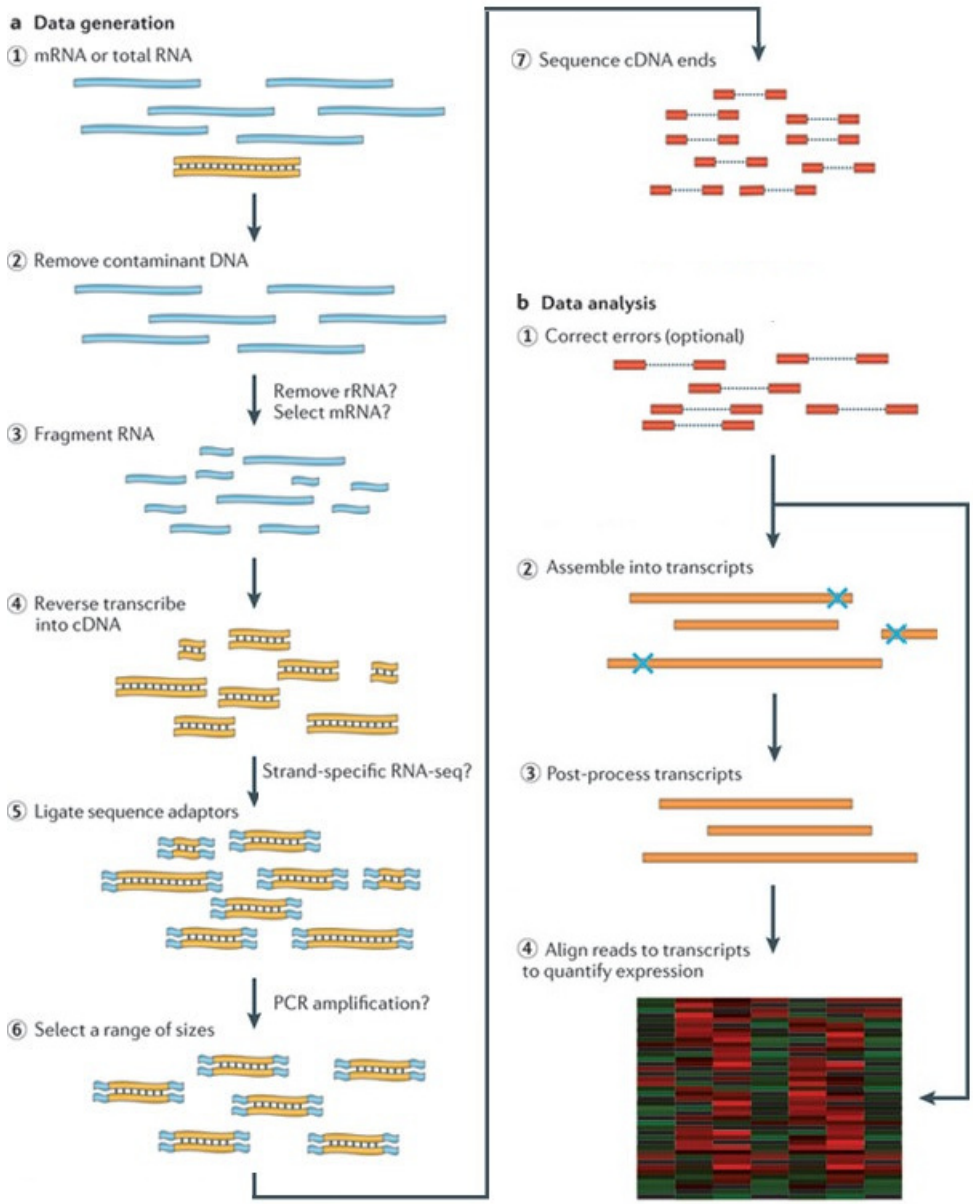


Figura 2.3: *Exemplo de um experimento utilizando sequenciamento por RNA-Seq (figura adaptada de Martin e Wang (2011)).*

que podem ser utilizados no processo de inferência de redes, por meio de dados de expressão.

2.2 Redes de Regulação Gênica

Os genes, as proteínas, incluindo os fatores de transcrição, e as interações entre estes elementos formam o que chamamos de Redes de Regulação Gênica, ou GRN (*gene regulatory networks*). Uma GRN é tipicamente representada por um diagrama onde nós representam os genes/proteínas/complexos proteicos e 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, por exemplo. Sendo

assim, as arestas representam as dependências entre os genes.

A estrutura de uma GRN é uma abstração da dinâmica química do sistema, ilustrada na Figura 2.4. Na parte superior da figura, temos três níveis que correspondem ao genoma (conjunto de genes do organismo), transcriptoma e proteoma (conjunto de proteínas encontrados em um organismo quando este está sujeito a algum estímulo). Neste exemplo, o gene A é transcrito e depois traduzido para a sua respectiva proteína, que serve como um fator de transcrição para o gene B, que por sua vez, interage com algum composto proteico e é necessário para a produção da proteína correspondente ao gene C. Estas interações são abstraídas no nível de GRN, na parte inferior da figura.

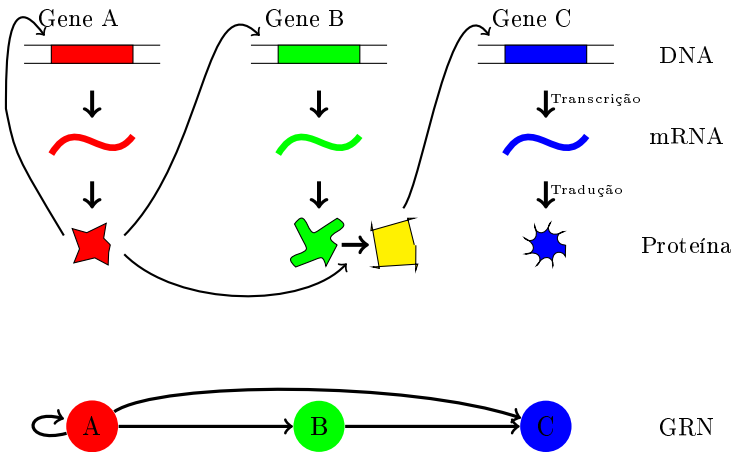


Figura 2.4: *Visão esquemática de uma rede de regulação gênica. Figura adaptada de Hecker et al. (2009).*

É possível fazer inferências sobre as interações existentes entre genes, por meio de medidas de expressão gênica. Neste trabalho, utilizamos um modelo gráfico probabilístico conhecido como redes Bayesianas. Este modelo é capaz de representar propriedades de independência condicional entre variáveis, além de fornecer uma metodologia clara para aprendizagem a partir de observações, mesmo quando os dados de expressão gênica utilizados no processo possuem ruídos (Friedman *et al.*, 2000).

2.3 Redes Bayesianas

Uma *rede Bayesiana*, ou BN (*Bayesian network*), é representada por um grafo acíclico direcionado $\mathcal{G}$, cujos nós correspondem a um conjunto de variáveis aleatórias $\mathcal{X} = \{X_1, \ldots, X_n\}$ e as arestas correspondem a influências diretas de um nó para outro. A Figura 2.5 representa uma rede Bayesiana, cujo cenário é composto por cinco variáveis aleatórias: A inteligência do estudante (I - *Intelligence*), a dificuldade do curso (D - *Difficulty*), a nota do estudante (G - *Grade*), a pontuação de um teste de habilidades acadêmicas do estudante (S - *SAT*) e a qualidade de uma carta de recomendação, realizada pelo professor do estudante (L - *Letter*).

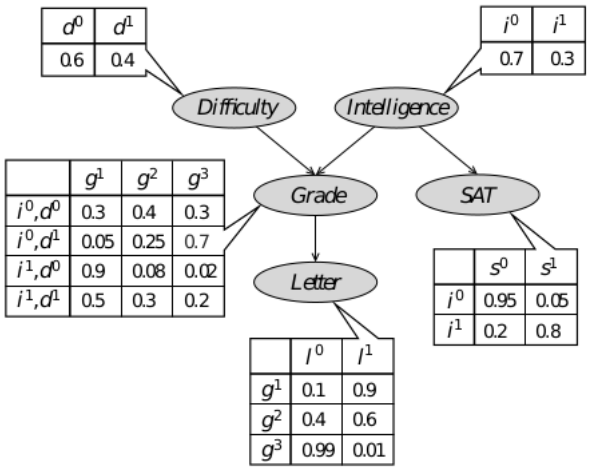


Figura 2.5: *Grafo de uma Rede Bayesiana para o exemplo do Estudante, com suas distribuições de probabilidade condicionais (CPDs) (Koller e Friedman, 2009).*

A inteligência do estudante pode assumir dois valores, onde $\text{Val}(I) = \{i^0, i^1\}$, que representam baixa inteligência (i^0) e alta inteligência (i^1). Similarmente, temos que $\text{Val}(D) = \{d^0, d^1\}$, que representam se o curso é fácil (d^0) ou difícil (d^1), $\text{Val}(S) = \{s^0, s^1\}$ representam pontuação baixa e alta e $\text{Val}(L) = \{l^0, l^1\}$ representam se a carta de recomendação é fraca (l^0) ou forte (l^1). A nota G do estudante pode assumir três valores distintos: $\text{Val}(G) = \{g^1, g^2, g^3\}$, que representam as notas A, B e C, respectivamente. Assim, a distribuição de probabilidade conjunta possui $2 \cdot 2 \cdot 2 \cdot 2 \cdot 3 = 48$ entradas.

Este grafo codifica um processo de amostragem generativo, onde o valor de cada variável é selecionado naturalmente, usando uma distribuição em que o valor que uma variável assume depende apenas de seus pais. Na Figura 2.5, as arestas codificam a intuição sobre o jeito que o universo funciona. Afirmações podem ser feitas, como "a carta de recomendação de um professor depende apenas da nota do estudante no curso". Esta afirmação se dá ao fato de não existirem arestas diretas ao nó L, exceto G, e pode ser representada formalmente como a declaração de que L é condicionalmente independente de todos os outros nós da rede, dado seu pai G:

$$(L \perp I, D, S \mid G) \ . \tag{2.1}$$

Em outras palavras, uma vez que se conhece a nota do estudante, a crença sobre a qualidade da carta de recomendação não é influenciada por nenhuma outra variável. Similarmente, a nota do teste de habilidades do estudante depende apenas de sua inteligência. Pode-se dizer que S é condicionalmente independente de todos os outros nós no grafo, dado seu pai I (Koller e Friedman, 2009).

$$(S \perp D, G, L \mid I) \ . \tag{2.2}$$

Podemos concluir que uma vez que se sabe o valor dos pais de uma variável, nenhuma informação relacionada direta ou indiretamente com seus pais ou outros descendentes pode influenciar nas crenças desta variável. Contudo, as informações sobre os descendentes desta

variável são capazes de mudar a crença sobre ela.

Definindo formalmente a semântica de uma rede Bayesiana: Uma estrutura de rede Bayesiana $\mathcal{G}$ é um grafo acíclico direcionado cujos nós representam variáveis aleatórias $\mathcal{X} = \{X_1, \ldots, X_n\}$. Seja $\text{Pa}_{X_i}^{\mathcal{G}}$ os pais de X_i em $\mathcal{G}$, e seja $\text{NonDescendants}_{X_i}$ as variáveis no grafo que não são descendentes de X_i. Logo, $\mathcal{G}$ representa o conjunto seguinte de atribuições de independência condicional, chamado de independências locais:

$$\text{Para cada variável } X_i : (X_i \perp \text{NonDescendants}_{X_i} \,|\, \text{Pa}_{X_i}^{\mathcal{G}}) \ . \tag{2.3}$$

Em outras palavras, as independências locais afirmam que cada nó X_i é condicionalmente independente de seus não descendentes dados seus pais (Koller e Friedman, 2009).

Com base nestas definições, na próxima seção apresentamos o segundo componente da representação de redes Bayesianas: o modelo de probabilidade local da rede, obtido por meio do processo de estimativa de parâmetros.

2.3.1 Estimativa de Parâmetros em Redes Bayesianas

O segundo componente da representação de redes Bayesianas é o conjunto de modelos de probabilidade local P que representa a natureza da dependência de cada variável com seus pais. No modelo do estudante, $P(I)$ representa a distribuição da população de estudantes inteligentes versus menos inteligentes. $P(D)$ representa a distribuição da dificuldade do curso, sendo fácil ou difícil. A distribuição sobre as notas dos estudantes é uma distribuição condicional $P(G \,|\, I, D)$. Esta atribuição especifica que a nota dos estudantes depende da inteligência do estudante e da dificuldade do curso. Assim, pode-se ter uma distribuição diferente para cada atribuição de valores i, d.

Em geral, cada variável X do modelo é associada a uma distribuição de probabilidade condicional, ou CPD (*conditional probability distribution*), que especifica a distribuição sobre os valores de X dado cada possível atribuição de valores para seus pais no modelo. Para um nó sem pais, a CPD está condicionada a um conjunto vazio de variáveis. Assim, a CPD transforma-se numa distribuição marginal, tal como $P(D)$ ou $P(I)$. Uma possível escolha de CPDs para este domínio é mostrada na Figura 2.5. A estrutura de rede juntamente com suas CPDs é uma rede Bayesiana $\mathcal{B}$; é usada a notação $\mathcal{B}^{\text{student}}$ para se referir a rede deste exemplo (Koller e Friedman, 2009). Assim, a especificação de distribuição de probabilidade conjunta deste exemplo é dada por:

$$P(I, D, G, S, L) = P(I)P(D)P(G \,|\, I, D)P(S \,|\, I)P(L \,|\, G) \ . \tag{2.4}$$

Para encontrar o modelo de probabilidade que governa o comportamento da rede, é necessário realizar a *estimativa de parâmetros*. Para realizar tal estimativa, é preciso assumir que há uma rede $\mathcal{G}$ existente, juntamente com um conjunto de dados amostrais $\mathcal{D}$. Este

conjunto de dados deve possuir instâncias totalmente observadas das variáveis que compõem a rede: $\mathcal{D} = \{\mathcal{X}[1], \ldots, \mathcal{X}[M]\}$. Uma das abordagens mais utilizadas para a estimativa de parâmetros é a abordagem Bayesiana. Porém, antes de compreendê-la, é importante entender primeiro alguns conceitos relacionados à máxima verossimilhança.

2.3.2 Estimativa de Máxima Verossimilhança

Considerando várias amostras independentes e identicamente distribuídas (IID) de um conjunto de variáveis $\mathcal{X} = \{X_1, \ldots, X_n\}$ a partir de uma distribuição desconhecida $P^*(\mathcal{X})$, tem-se que o conjunto de treinamento de dados é denotado como $\mathcal{D}$, em que $\mathcal{D}$ é formado por M instâncias das variáveis em $\mathcal{X}$: $\mathcal{X}[1], \ldots, \mathcal{X}[M]$.

Assume-se que há um modelo paramétrico para o qual deseja-se estimar parâmetros. Um modelo paramétrico é definido por uma função $P(X : \boldsymbol{\theta})$, especificada em termos de um conjunto de parâmetros. Dado um conjunto particular de valores de parâmetros $\boldsymbol{\theta}$ e uma instância $X \in \mathcal{X}$, o modelo atribui uma probabilidade a X. É exigido que para cada escolha dos parâmetros $\boldsymbol{\theta}$, $P(X : \boldsymbol{\theta})$ seja uma distribuição não negativa, de forma que a soma das probabilidades das observações dado os parâmetros seja igual a um:

$$\sum_X P(X : \boldsymbol{\theta}) = 1 \ . \tag{2.5}$$

A estimativa de máxima verossimilhança, ou MLE (*maximum-likelihood estimation*), tenta encontrar quais são os parâmetros ($\boldsymbol{\theta}$) que, dado um conjunto de observações ($\mathcal{D} = \{\mathcal{X}[1], \ldots, \mathcal{X}[M]\}$), melhor representam os dados, sendo possível aplicá-la em redes Bayesianas. A *função de verossimilhança* é dada por:

$$L(\boldsymbol{\theta} : \mathcal{D}) = \prod_m P(\mathcal{X}[m] : \boldsymbol{\theta}) \ , \tag{2.6}$$

onde $P(\mathcal{X}[m] : \boldsymbol{\theta})$ é a probabilidade atribuída pelo modelo a $\mathcal{X}[m]$ dado um conjunto particular de parâmetros $\boldsymbol{\theta}$.

Seja X uma variável que pode assumir valores $x^1, \ldots, x^K$. A mais simples representação de uma distribuição multinomial é um vetor $\boldsymbol{\theta} = \langle \theta_1, \ldots, \theta_K \rangle$ tal que $\sum_k \theta_k = 1$. A função de verossimilhança deste modelo é da forma:

$$L(\boldsymbol{\theta} : \mathcal{D}) = \prod_k \theta_k^{M[k]} \ , \tag{2.7}$$

e a máxima verossimilhança é alcançada quando:

$$\hat{\theta}_k = \frac{M[k]}{M} \ , \tag{2.8}$$

na qual $M[k]$ é o número de vezes que a variável X assumiu o valor x^k, e M é o número

total de amostras. Isto é, a probabilidade de cada valor de X corresponde a sua frequência nos dados de treinamento.

Para facilitar o cálculo da verossimilhança, é possível utilizar o *logaritmo da função de verossimilhança*. Assim, a função *log-verossimilhança* pode ser definida como:

$$\ell(\boldsymbol{\theta} : \mathcal{D}) = \sum_k M[k] \log \theta_k \ . \tag{2.9}$$

Maiores explicações podem ser encontradas em Koller e Friedman (2009).

Exemplo prático

Seja $x[1], \ldots, x[M]$ um conjunto de lançamentos de moedas. Cada lançamento é amostrado de forma independente a partir de uma mesma distribuição, em que $X[m]$ é igual a H (cara) ou T (coroa), cujas probabilidades são θ e $1 - \theta$. O objetivo é encontrar um bom valor para o parâmetro θ. Para isso, um espaço de hipóteses Θ e uma função objetivo são definidos. O espaço de hipóteses é um conjunto de valores a serem considerados para θ. A função objetivo define o quão bem diferentes hipóteses neste espaço correspondem ao conjunto de dados $\mathcal{D}$. Neste caso, o espaço de hipóteses Θ é o conjunto de todos os parâmetros $\theta \in [0, 1]$.

Uma maneira de pontuar diferentes parâmetros θ é avaliar o quão bem um parâmetro é capaz de predizer os dados. Se os dados são prováveis dado um parâmetro, o parâmetro é um bom preditor. Considere a sequência de lançamentos H, T, T, H, H. Caso o parâmetro θ seja conhecido, é possível atribuir uma probabilidade para esta sequência de observações. A probabilidade do primeiro lançamento é $P(X[1] = H) = \theta$. Como os lançamentos são independentes, a probabilidade do segundo lançamento é $P(X[2] = T) = 1 - \theta$. Logo, a função de verossmilhança da sequência é definida como:

$$L(\theta : \langle H, T, T, H, H \rangle) = P(\langle H, T, T, H, H \rangle : \theta) = \theta^3 (1 - \theta)^2 \ . \tag{2.10}$$

Os valores de parâmetros que possuem a mais alta verossimilhança são mais prováveis em gerar a sequência observada. Podemos utilizar a função de verossimilhança como medida de qualidade para diferentes valores de parâmetros e selecionar os valores que maximizam a função. A Figura 2.6 mostra que $\hat{\theta} = 0.6 = \frac{3}{5}$ maximiza a verossimilhança para a sequência H, T, T, H, H.

Seja $\mathcal{D}$ um conjunto de dados, que possui observações contendo $M[1]$ caras e $M[0]$ coroas. O objetivo é encontrar o valor $\hat{\theta}$ que maximiza a verossimilhança de θ correspondente a $\mathcal{D}$. A função de máxima verossimilhança neste caso é definida como:

$$L(\theta : \mathcal{D}) = \theta^{M[1]} (1 - \theta)^{M[0]}. \tag{2.11}$$

É mais simples maximizar o logaritmo da função de verossimilhança (*log-verossimilhança*).

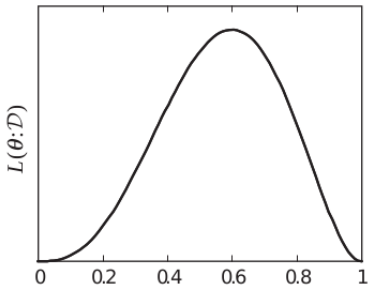


Figura 2.6: *A função de verossimilhança para a sequência de lançamentos* H, T, T, H, H.

Neste caso, a função pode ser definida como:

$$\ell(\theta : \mathcal{D}) = M[1] \log \theta + M[0] \log(1 - \theta) \ . \tag{2.12}$$

A função *log-verossimilhança* é relacionada com a verossimilhança. Maximizar uma destas funções é equivalente a maximizar a outra. Entretanto, é mais conveniente lidar com a função *log-verossimilhança*, já que os produtos são convertidos em somas (Koller e Friedman, 2009). Assim, o parâmetro que maximiza a função de máxima verossimilhança, denotado como $\hat{\theta}$ é:

$$\hat{\theta} = \frac{M[1]}{M[1] + M[0]} \ . \tag{2.13}$$

MLE e Redes Bayesianas

Podemos usar a máxima verossimilhança para estimar os parâmetros $\boldsymbol{\theta}$ de uma rede Bayesiana com estrutura $\mathcal{G}$. Seja um conjunto de dados $\mathcal{D} = \{\mathcal{X}[1], \ldots, \mathcal{X}[M]\}$. A função de verossimilhança é então definida como:

$$L(\boldsymbol{\theta} : \mathcal{D}) = \prod_i L_i(\boldsymbol{\theta}_{X_i \,|\, \mathrm{Pa}_{X_i}} : \mathcal{D}) \ , \tag{2.14}$$

na qual a *verossimilhança local* para X_i é

$$L_i(\boldsymbol{\theta}_{X_i \,|\, \mathrm{Pa}_{X_i}} : \mathcal{D}) = \prod_m P(x_i[m] \,|\, \mathrm{pa}_{X_i}[m] : \boldsymbol{\theta}_{X_i \,|\, \mathrm{Pa}_{X_i}}) \ . \tag{2.15}$$

Esta análise mostra que a verossimilhança é decomposta como um produto de termos independentes, um para cada CPD da rede. Esta propriedade é chamada de *decomposição global* da função de verossimilhança.

Proposição 1 *Seja* $\mathcal{D}$ *um conjunto de dados completos para* $X_1, \ldots, X_n$ *e* $\mathcal{G}$ *uma estrutura de rede sobre estas variáveis. Suponha que os parâmetros* $\boldsymbol{\theta}_{X_i \,|\, \mathrm{Pa}_{X_i}}$ *são disjuntos de* $\boldsymbol{\theta}_{X_j \,|\, \mathrm{Pa}_{X_j}}$, *para todo* $j \neq i$. *Seja* $\hat{\boldsymbol{\theta}}_{X_i \,|\, \mathrm{Pa}_{X_i}}$ *os parâmetros que maximizam* $L_i(\boldsymbol{\theta}_{X_i \,|\, \mathrm{Pa}_{X_i}} : \mathcal{D})$. *Logo,* $\hat{\boldsymbol{\theta}} = \langle \hat{\boldsymbol{\theta}}_{X_1 \,|\, \mathrm{Pa}_1}, \ldots, \hat{\boldsymbol{\theta}}_{X_n \,|\, \mathrm{Pa}_n} \rangle$.

Em outras palavras, é possível maximizar cada função local de verossimilhança de forma independente do restante da rede, e então combinar as soluções para se obter uma solução de MLE.

Seja X um conjunto de variáveis com pais $\mathbf{U}$. Se a CPD $P(X \mid \mathbf{U})$ for representada como uma tabela, então haverá um parâmetro $\theta_{x \mid \mathbf{u}}$ para cada combinação de $x \in \text{Val}(X)$ e $u \in \text{Val}(\mathbf{U})$. Logo, a função pode ser reescrita da seguinte forma:

$$\begin{aligned} L_X(\boldsymbol{\theta}_{X \mid \mathbf{U}} : \mathcal{D}) &= \prod_m \theta_{x[m] \mid \mathbf{u}[m]} \\ &= \prod_{\mathbf{u} \in \text{Val}(\mathbf{U})} \left[\prod_{x \in \text{Val}(X)} \theta_{x \mid \mathbf{u}}^{M[\mathbf{u},x]} \right] , \end{aligned} \tag{2.16}$$

na qual $M[\mathbf{u}, x]$ é referente ao número de vezes $\mathcal{X}[m] = x$ e $\mathbf{u}[m] = \mathbf{u}$ em $\mathcal{D}$. Isto é, o agrupamento de todas as ocorrências de $\theta_{x \mid \mathbf{u}}$ é realizado no produto de todas as instâncias, proporcionando uma decomposição local adicional na função de verossimilhança.

Assim, precisamos maximizar o termo $\theta_{x \mid \mathbf{u}}$ onde, para cada escolha de valores de pais $\mathbf{U}$, a probabilidade condicional seja válida, ou seja:

$$\sum \theta_{x \mid \mathbf{u}} = 1 \text{ para todo } \mathbf{u} \ . \tag{2.17}$$

Isso implica que a escolha do valor para $\theta_{x \mid \mathbf{u}}$ pode impactar a escolha de valores para $\theta_{x' \mid \mathbf{u}}$. Entretanto, a escolha de parâmetros dado diferentes valores $\mathbf{u}$ de $\mathbf{U}$ são independentes umas das outras. Portanto, é possível maximizar cada um dos termos presentes nos colchetes na Equação 2.16 de maneira independente.

Podemos, assim, decompor ainda mais a função de verossimilhança local para uma tabela de CPD em um produto de funções de verossimilhança simples. Cada uma dessas funções de verossimilhança é uma *verossimilhança multinomial*. As contagens nos dados para os diferentes resultados de x são simplesmente $\{M[\mathbf{u}, x] : x \in \text{Val}(X)\}$. Podemos então usar imediatamente a estimativa de máxima verossimilhança para a verossimilhança multinomial e observar que os parâmetros são:

$$\hat{\theta}_{x \mid \mathbf{u}} = \frac{M[\mathbf{u}, x]}{M[\mathbf{u}]} \ , \tag{2.18}$$

onde usamos o fato de que $M[\boldsymbol{u}] = \sum_x M[\mathbf{u}, x]$. Observe que o número de amostras usadas para estimar o parâmetro $\hat{\theta}_{x \mid \mathbf{u}}$ é $M[\mathbf{u}]$. As amostras que não estão de acordo com a atribuição de $\mathbf{u}$ não fazem parte do cálculo. À medida em que o número de pais $\mathbf{U}$ cresce, o número de configurações de pais diferentes cresce exponencialmente. Portanto, o número de amostras que se espera ter para uma única configuração diminui exponencialmente. Isso é chamado de *fragmentação de dados*.

Intuitivamente, quando temos um número muito pequeno de amostras a partir das quais estimamos um parâmetro, as estimativas que obtemos podem possuir muito ruído. Também

é mais provável obter um grande número de zeros na distribuição, o que pode levar a um desempenho muito ruim. A incapacidade de estimar parâmetros de forma confiável à medida que a dimensionalidade do conjunto pai cresce é um dos principais fatores limitantes no aprendizado de redes Bayesianas a partir de dados. Este problema é ainda mais grave quando as variáveis podem assumir um grande número de valores (Koller e Friedman, 2009).

Até agora, foram definidos os conceitos de redes Bayesianas e a forma de realizar a estimativa de parâmetros utilizando a estimativa MLE. Redes Bayesianas, apesar de serem boas representações de domínios que envolvem relações de incerteza entre variáveis aleatórias, possuem sérias limitações. A principal delas é ser um modelo que não permite relações cíclicas entre as variáveis. Uma alternativa para este problema pode ser encontrada no uso de redes Bayesianas dinâmicas. A seguir, mais detalhes serão definidos a respeito deste novo conceito e a forma de se realizar a estimativa MLE para o modelo.

2.4 Redes Bayesianas Dinâmicas

Redes Bayesianas dinâmicas, ou DBNs (*Dynamic Bayesian Networks*), permitem a construção de redes de regulação cíclicas, utilizando informação temporal, sendo inicialmente propostas no contexto genético por Friedman *et al.* (1998). Suas análises mostraram que DBNs apresentam bons resultados no processo de inferência de redes, principalmente por lidar bem com ruídos e dados incompletos. Além disso, DBNs lidam particularmente bem com dados temporais na construção de relações causais entre variáveis. A Figura 2.7 mostra como uma rede simplificada contendo ciclos pode ser representada por uma DBN, mesmo se tratando de um modelo acíclico.

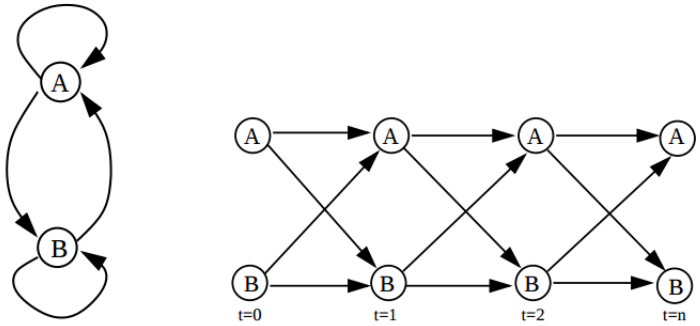


Figura 2.7: *Rede Bayesiana Dinâmica: A rede à esquerda não é uma rede Bayesiana, já que possui ciclos; Considerando os atrasos entre essas interações, é possível imaginar esta rede desdobrada no tempo, onde as interações dentro de qualquer instante de tempo t não são permitidas. O resultado é um DAG apropriado como representado pela rede à direita (figura retirada de Werhli et al. (2007)).*

Enquanto redes Bayesianas estáticas descrevem a distribuição de probabilidade sobre um conjunto fixo de variáveis, redes Bayesianas dinâmicas estendem estes conceitos a fim de modelar processos temporais.

Definindo formalmente a semântica de uma rede Bayesiana dinâmica: Uma estrutura de rede Bayesiana $\mathcal{G}$ é um grafo acíclico direcionado cujos nós representam variáveis aleatórias onde os processos sofrem mudanças. $X_i[t]$ é uma variável aleatória que representa o valor do atributo X_i no tempo t, e $\mathcal{X}[t]$ é o conjunto de variáveis aleatórias $X_i[t]$. Seja $\mathrm{Pa}^{\mathcal{G}}_{X_i[t]}$ os pais da variável $X_i[t]$ em $\mathcal{G}$, e seja $\mathrm{NonDescendents}_{X_i[t]}$ as variáveis no grafo que não são descendentes de $X_i[t]$. Logo, $\mathcal{G}$ representa o seguinte conjunto de atribuições de independência condicional:

$$\text{Para cada variável } X_i[t] : (X_i[t] \perp \mathrm{NonDescendants}_{X_i[t]} \,|\, \mathrm{Pa}^{\mathcal{G}}_{X_i[t]}) \ , \tag{2.19}$$

ou seja, as independências locais definem que cada nó $X_i[t]$ é condicionalmente independente de seus não descendentes dados seus pais.

Na próxima seção, apresentamos a forma de calcular o modelo de probabilidade local de redes Bayesianas dinâmicas, obtido por meio do processo de estimativa de parâmetros.

2.4.1 Estimativa de Parâmetros em Redes Bayesianas Dinâmicas

Para representar crenças sobre as trajetórias possíveis do processo, é necessário uma distribuição de probabilidade sobre variáveis aleatórias $\mathcal{X}[0] \cup \mathcal{X}[1] \cup \mathcal{X}[2] \cup \ldots$. O processo utilizado será Markoviano em $\mathcal{X}$, por exemplo, $P(\mathcal{X}[t+1] \,|\, \mathcal{X}[0], \ldots, \mathcal{X}[t]) = P(\mathcal{X}[t+1] \,|\, \mathcal{X}[t])$. A probabilidade de transição $P(\mathcal{X}[t+1] \,|\, \mathcal{X}[t])$ é independente de t. Logo, uma rede Bayesiana dinâmica que representa as distribuições conjuntas de todas as trajetórias possíveis do processo consiste em duas partes:

1. uma *rede primária* B_0 que especifica uma distribuição sobre os estados iniciais $\mathcal{X}[0]$; e
2. uma *rede de transição* $B_{\rightarrow}$ sobre as variáveis $\mathcal{X}[0] \cup \mathcal{X}[1]$ que especificam a probabilidade de transição $\mathrm{P}(\mathcal{X}[t+1] \,|\, \mathcal{X}[t])$ para todo t.

A Figura 2.8 ilustra um exemplo simples. Na rede de transição, Figura 2.8(b), as variáveis em $\mathcal{X}[0]$ não tem pais. A probabilidade da rede de transição é dada por:

$$P_{B_{\rightarrow}}(\boldsymbol{x}[1] \,|\, \boldsymbol{x}[0]) = \prod_{i=1}^{n} P_{B_{\rightarrow}}(x_i[1] \,|\, \mathbf{Pa}(X_i[1])) \ . \tag{2.20}$$

A DBN definida pelo par $(B_0, B_{\rightarrow})$ corresponde a rede semi-infinita sobre as variáveis $\mathcal{X}[0], \ldots, \mathcal{X}[\infty]$. Na prática, há um intervalo finito variando de $0, \ldots, T$. A estrutura da rede é "desenrolada" em uma rede Bayesiana sobre $\mathcal{X}[0], \ldots, \mathcal{X}[T]$. No tempo $t = 0$, os pais de $X_i[0]$ são aqueles especificados na rede B_0. No tempo $t+1$, os pais de $X_i[t+1]$ são aqueles nós no tempo t e $t+1$ correspondentes aos pais de $X_i[1]$ na rede $B_{\rightarrow}$ (Friedman *et al.*, 1998).

A Figura 2.8(c) mostra a rede desenrolada em três períodos de tempo de 2.8(a) e 2.8(b).

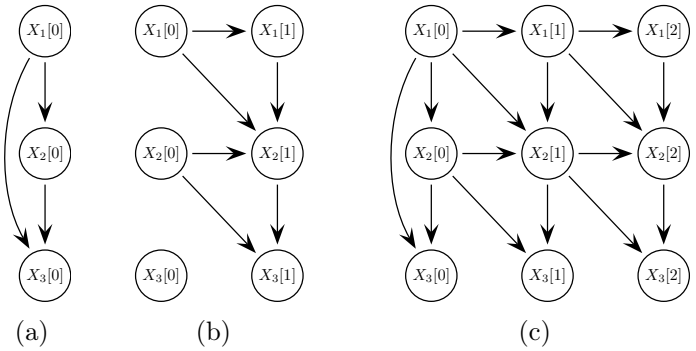


Figura 2.8: *Em (a) temos a rede primária B_0 e em (b) a rede de transição $B_{\rightarrow}$ para três variáveis. Em (c) temos a rede "desenrolada". Figura adaptada de Friedman et al. (2000).*

Dado um modelo DBN, a distribuição conjunta sobre $\mathcal{X}[0], \ldots, \mathcal{X}[T]$ é

$$P_B(\boldsymbol{x}[0], \ldots, \boldsymbol{x}[T]) = P_{B_0}(\boldsymbol{x}[0]) \prod_{t=0}^{T-1} P_{B_{\rightarrow}}(\boldsymbol{x}[t+1] \,|\, \boldsymbol{x}[t]) \ , \tag{2.21}$$

onde $P_{B_{\rightarrow}}(\boldsymbol{x}[t+1])$ é obtido a partir do modelo de transição.

Seja $\mathcal{D}$ um conjunto de dados de séries temporais, consistindo de M_{seq} sequências de observações completas, onde a l-ésima sequência possui M_l instâncias de $\mathcal{X} : \mathcal{X}^l[0], \ldots, \mathcal{X}^l[M_l]$. Como um conjunto de dados de expressão nos dá M_{seq} instâncias de períodos iniciais, nós podemos treinar a rede B_0 a partir delas, e $M = \sum_l M_l$ instâncias de transição que utilizaremos no treinamento da rede de transição $B_{\rightarrow}$.

A seguir, são apresentados os métodos para se obter o conjunto de modelos de probabilidade local, por meio da estimativa de parâmetros, em redes Bayesianas Dinâmicas.

2.4.2 Estimativa de Máxima Verossimilhança em Redes Bayesianas Dinâmicas

Assim como em redes Bayesianas, em redes Bayesianas dinâmicas também podemos utilizar a estimativa de máxima verossimilhança para estimar parâmetros $\boldsymbol{\theta}$ de uma rede com estrutura $\mathcal{G}$. Seja a verossimilhança definida como na Equação 2.6. Para definirmos a função correspondente em redes Bayesianas dinâmicas, vamos introduzir alguns conceitos. Seja:

$$\theta^{(0)}_{i,j_i',k_i'} = P(X_i[0] = k_i' \,|\, \mathbf{Pa}(X_i[0]) = j_i') \tag{2.22}$$

e similarmente:

$$\theta^{\rightarrow}_{i,j_i,k_i} = P(X_i[t] = k_i \,|\, \mathbf{Pa}(X_i[t]) = j_i) \ , \tag{2.23}$$

para $t = 1, \ldots, T$. Logo, precisaremos de uma notação que armazene a quantidade de vezes que uma variável X assume valor k_i sempre que seus pais assumem valores j_i, para cada

estrutura de rede. Seja então

$$M^{(0)}_{i,j_i',k_i'} = \sum_l I(X_i[0] = k_i', \mathbf{Pa}(X_i[0]) = j_i'; \mathcal{X}^l) \tag{2.24}$$

e

$$M^{\rightarrow}_{i,j_i,k_i} = \sum_l \sum_t I(X_i[t] = k_i, \mathbf{Pa}(X_i[t]) = j_i; \mathcal{X}^l) \ , \tag{2.25}$$

onde $I(\cdot\,; \mathcal{X}^l)$ é uma função indicadora que assume valor 1 se o evento $\cdot$ ocorre na sequência $\mathcal{X}^l$ e 0 caso contrário.

Analisando a Equação 2.6, temos que a função de verossimilhança se decompõe de acordo com a estrutura da DBN, logo temos:

$$L(\boldsymbol{\theta} : \mathcal{D}) = \prod_i \prod_{j_i'} \prod_{k_i'} (\theta^{(0)}_{i,j_i',k_i'})^{M^{(0)}_{i,j_i',k_i'}} \cdot \prod_i \prod_{j_i} \prod_{k_i} (\theta^{\rightarrow}_{i,j_i,k_i})^{M^{\rightarrow}_{i,j_i,k_i}} \ . \tag{2.26}$$

Para facilitar o cálculo da estimativa, usa-se a função logarítmica, *log-verossimilhança*, que pode ser definida como:

$$\ell(\boldsymbol{\theta} : \mathcal{D}) = \sum_i \sum_{j_i'} \sum_{k_i'} M^{(0)}_{i,j_i',k_i'} \log \theta^{(0)}_{i,j_i',k_i'} + \sum_i \sum_{j_i} \sum_{k_i} M^{\rightarrow}_{i,j_i,k_i} \log \theta^{\rightarrow}_{i,j_i,k_i} \ . \tag{2.27}$$

Assim, a função *log-verossimilhança* é expressa como uma soma de termos, onde cada termo depende apenas da probabilidade condicional da variável dado uma atribuição de valores particular de seus pais. Agora, é necessário encontrar quais são os parâmetros que maximizam esta função. Para isso, é possível maximizar cada termo de verossimilhança local de forma independente. Neste modelo, a rede B_0 é independente da rede $B_{\rightarrow}$. Logo, utilizando estimativa de máxima verossimilhança em distribuições multinomiais, tem-se a expressão seguinte para $\hat{\Theta}_{\mathcal{G}}$:

$$\hat{\theta}^{(0)}_{i,j_i',k_i'} = \frac{M^{(0)}_{i,j_i',k_i'}}{\sum_{k_i'} M^{(0)}_{i,j_i',k_i'}} \tag{2.28}$$

e similarmente para o caso da rede de transição.

A estimativa de parâmetros utilizando a abordagem de máxima verossimilhança, embora possua vantagens, possui muitas limitações. Por exemplo, levando-se em conta 10 lançamentos de moedas, se são observadas 3 caras, a estimativa de MLE será 0,3. A mesma estimativa será fornecida se forem observadas 300 caras de 1000 lançamentos. Claramente, os dois experimentos não são equivalentes. A intuição é que no segundo experimento a estimativa seja mais confiável, já que tem-se um maior *conhecimento prévio* a respeito de seu comportamento. Na próxima seção, definiremos uma abordagem que tenta minimizar este problema, por meio da utilização da estimativa Bayesiana de parâmetros, útil tanto para o modelo de BNs como para o modelo de DBNs.

2.5 Estimativa Bayesiana de Parâmetros

Na abordagem de estimativa Bayesiana de parâmetros, um conhecimento prévio (*prior knowledge*) sobre o parâmetro θ é representado por uma distribuição de probabilidade *a priori* $P(\theta)$. Uma vez que o conhecimento (ou a falta dele) a respeito de um parâmetro θ é quantificado, é possível criar uma distribuição conjunta sobre os parâmetros e os dados observados. Entre as distribuições utilizadas para representar $P(\theta)$ podemos citar a *distribuição Beta* e a *distribuição de Dirichlet*.

Considere um problema de inferência de redes, composto por um conjunto de treinamento $\mathcal{D}$, com M amostras IID de um conjunto de variáveis $\mathcal{X}$, a partir de uma distribuição desconhecida $P^*(\mathcal{X})$. Considere também um modelo paramétrico $P(\mathcal{X} \mid \boldsymbol{\theta})$, onde pode-se atribuir parâmetros a partir de um espaço paramétrico Θ.

Enquanto a abordagem MLE se preocupa em encontrar parâmetros $\hat{\boldsymbol{\theta}}$ em Θ que são "melhores" dado os dados, a abordagem Bayesiana tenta observar a crença sobre os valores de $\boldsymbol{\theta}$ e utiliza estas crenças para extrair conclusões. Na abordagem Bayesiana, $\boldsymbol{\theta}$ é tratado como uma variável aleatória, e requer o uso de probabilidade para descrever uma incerteza inicial sobre os parâmetros $\boldsymbol{\theta}$, utilizando o raciocínio probabilístico (teorema de Bayes) para levar em conta estas observações.

Uma vez especificada a função de verossimilhança e a distribuição *a priori*, podemos utilizar os dados para derivar a distribuição *a posteriori* sobre os parâmetros:

$$P(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathcal{D})} \ . \tag{2.29}$$

O primeiro termo, $P(\mathcal{D} \mid \boldsymbol{\theta})$, é a função de verossimilhança. O segundo termo, $P(\boldsymbol{\theta})$, é a distribuição *a priori* sobre os possíveis valores em Θ que captura a incerteza inicial sobre os parâmetros. O termo $P(\mathcal{D})$ é a verossimilhança marginal dos dados:

$$P(\mathcal{D}) = \int_{\Theta} P(\mathcal{D} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \ , \tag{2.30}$$

isto é, a integração da verossimilhança sobre todas as atribuições de parâmetros possíveis.

Seja X uma variável multinomial que pode assumir valores $x^1, \ldots, x^K$. É necessário descrever a incerteza sobre os parâmetros de uma distribuição multinomial. O espaço de parâmetros Θ é o espaço de vetores não negativos $\boldsymbol{\theta} = \langle \theta_1, \ldots, \theta_K \rangle$ tal que $\sum_k \theta_k = 1$. A função de verossimilhança deste modelo é da forma:

$$L(\boldsymbol{\theta} : \mathcal{D}) = \prod_k \theta_k^{M[k]} \ , \tag{2.31}$$

na qual $M[k]$ é o número de ocorrências de x^k. Uma vez que *a posteriori* é um produto entre *a priori* e a verossimilhança, é natural que *a priori* tenha uma forma similar a da verossimilhança.

Como mencionado, uma distribuição *a priori* utilizada para esta situação é a distribuição de Dirichlet. Esta distribuição é especificada por um conjunto de hiperparâmetros $\alpha_1, \ldots, \alpha_K$, que representam valores reais positivos, de modo que:

$$\boldsymbol{\theta} \sim \textit{Dirichlet}(\alpha_1, \ldots, \alpha_K) \text{ se } P(\boldsymbol{\theta}) \propto \prod_k \theta_k^{\alpha_k - 1} \ . \tag{2.32}$$

Intuitivamente, um hiperparâmetro α_k corresponde ao número de ocorrências imaginárias que foram "vistas" de x^k antes do experimento. Assim, quando utilizamos uma distribuição de Dirichlet como *a proiori*, *a posteriori* também é uma distribuição de Dirichlet:

Proposição 2 *Se $P(\boldsymbol{\theta})$ é Dirichlet$(\alpha_1, \ldots, \alpha_K)$ então $P(\boldsymbol{\theta} \mid \mathcal{D})$ é Dirichlet$(\alpha_1 + M[1], \ldots, \alpha_K + M[K])$, onde $M[k]$ é o número de ocorrências de x^k.*

Distribuições a *priori* como a distribuição de Dirichlet são úteis, desde que elas asseguram que a *a posteriori* possua uma distribuição compacta. Esta representação utiliza o mesmo formato da *priori*, facilitando o processo de computação.

2.6 Estimativa Bayesiana de Parâmetros em Redes

Suponha que queiramos estimar parâmetros para uma rede simples com duas variáveis X e Y, na qual X é pai de Y. O conjunto de treinamento consiste de observações $X[m], Y[m]$ para $m = 1, \ldots, M$. Além disso, há vetores de parâmetros desconhecidos $\boldsymbol{\theta}_X$ e $\boldsymbol{\theta}_{Y|X}$.

A estrutura de rede incorpora a suposição de que as distribuições *a priori* para os parâmetros são a princípio independentes, ou seja, conhecer o valor de um parâmetro não acrescenta nenhuma informação a respeito de outro parâmetro. Mais precisamente:

Proposição 3 *Seja $\mathcal{G}$ a estrutura de uma rede Bayesiana com parâmetros $\boldsymbol{\theta} = (\boldsymbol{\theta}_{X_1|\mathrm{Pa}_{X_1}}, \ldots, \boldsymbol{\theta}_{X_n|\mathrm{Pa}_{X_n}})$. Uma distribuição a priori $P(\boldsymbol{\theta})$ satisfaz a independência global de parâmetros se ela tem a forma:*

$$P(\boldsymbol{\theta}) = \prod_i P(\boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}}) \ . \tag{2.33}$$

Se duas variáveis de parâmetros são independentes *a priori*, elas serão independentes *a posteriori*. Utilizando a definição de independência condicional, conclui-se que:

$$P(\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y|X} \mid \mathcal{D}) = P(\boldsymbol{\theta}_X \mid \mathcal{D}) P(\boldsymbol{\theta}_{Y|X} \mid \mathcal{D}) \ , \tag{2.34}$$

ou seja, dado o conjunto de dados completo $\mathcal{D}$, pode-se determinar *a posteriori* sobre $\boldsymbol{\theta}_X$ independentemente da *posteriori* sobre $\boldsymbol{\theta}_{Y|X}$. Assim, uma vez que os problemas são resolvidos separadamente, combinam-se os resultados.

Caso geral

Suponha que seja dada uma rede com estrutura $\mathcal{G}$ e parâmetros $\boldsymbol{\theta}$. É preciso especificar uma distribuição *a priori* $P(\boldsymbol{\theta})$ sobre todas as possíveis parametrizações da rede. A distribuição *a posteriori* sobre os parâmetros dado o conjunto de amostras $\mathcal{D}$ é:

$$P(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathcal{D})} \ . \tag{2.35}$$

Como já foi discutido, podemos decompor a verossimilhança em verossimilhanças locais (Equação 2.14). Se assumirmos a independência global de parâmetros e combinando as Equações 2.14 e 2.33 temos:

$$P(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{1}{P(\mathcal{D})} \prod_i \left[L_i(\boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}} : \mathcal{D}) P(\boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}}) \right] \ . \tag{2.36}$$

Proposição 4 *Seja $\mathcal{D}$ um conjunto de dados completo para $\mathcal{X}$, e seja $\mathcal{G}$ uma estrutura de rede sobre estas variáveis. Se $P(\boldsymbol{\theta})$ satisfaz a independência global de parâmetros, então:*

$$P(\boldsymbol{\theta} \mid \mathcal{D}) = \prod_i P(\boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}} \mid \mathcal{D}) \ . \tag{2.37}$$

2.6.1 Decomposição Local

Dada a discussão anterior, devemos resolver problemas de estimativa Bayesiana locais para obter uma solução global.

Proposição 5 *Seja X uma variável com pais $\mathbf{U}$. Dizemos que a distribuição* a priori $P(\boldsymbol{\theta}_{X|\mathbf{U}})$ *satisfaz a* independência local de parâmetros *se:*

$$P(\boldsymbol{\theta}_{X|\mathbf{U}}) = \prod_{\mathbf{u}} P(\boldsymbol{\theta}_{X|\mathbf{u}}) \ . \tag{2.38}$$

Proposição 6 *Seja $\mathcal{D}$ um conjunto de dados completo para $\mathcal{X}$, seja $\mathcal{G}$ uma estrutura de rede sobre as variáveis com tabelas de CPDs. Se a distribuição* a priori $P(\boldsymbol{\theta})$ *satisfaz a independência global e local de parâmetros, então:*

$$P(\boldsymbol{\theta} \mid \mathcal{D}) = \prod_i \prod_{\mathrm{Pa}_{X_i}} P(\boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}} \mid \mathcal{D}) \ . \tag{2.39}$$

Além disso, se $P(\boldsymbol{\theta}_{X|\mathbf{u}})$ é uma *a priori Dirichlet* com hiperparâmetros $\alpha_{x^1|\mathbf{u}}, \ldots, \alpha_{x^K|\mathbf{u}}$, então *a posteriori* $P(\boldsymbol{\theta}_{X|\mathbf{u}} \mid \mathcal{D})$ é uma distribuição *Dirichlet* com hipermarâmetros $\alpha_{x^1|\mathbf{u}} + M[\mathbf{u}, x^1], \ldots, \alpha_{x^K|\mathbf{u}} + M[\mathbf{u}, x^K]$.

2.6.2 Distribuições *a priori* em Redes Bayesianas

Agora, abordaremos a questão de avaliar o conjunto de parâmetros *a priori* necessários para uma rede Bayesiana.

Proposição 7 *Seja $\mathcal{G}$ uma rede Bayesiana, onde cada nó X_i possui um conjunto de distribuições multinomiais $\boldsymbol{\theta}_{X_i|\mathrm{Pa}_{X_i}}$, uma para cada instanciação pa_{X_i} dos pais Pa_{X_i} de X_i. Cada um destes parâmetros terá um prior* Dirichlet*, governado pelos hiperparâmetros:*

$$\boldsymbol{\alpha}_{X_i|\mathrm{pa}_{X_i}} = (\alpha_{x_i^1|\mathrm{pa}_{X_i}}, \ldots, \alpha_{x_i^{K_i}|\mathrm{pa}_{X_i}}) \ , \tag{2.40}$$

onde K_i é o número de valores de X_i.

A atribuição dos valores $\boldsymbol{\alpha}$ pode ser feita com o auxílio de um especialista, baseado em seu conhecimento. Porém, esta tarefa é difícil de ser realizada. Uma alternativa é a utilização da distribuição de Dirichlet. Podem ser atribuídos hiperparâmetros α_{x^k} como contagens imaginárias em nossa experiência *a priori*.

Proposição 8 *Seja $\mathcal{D}'$ um conjunto de dados imaginários, podemos utilizar contagens a partir destes dados imaginários e atribuí-las como hiperparâmetros. Assim, temos que:*

$$\alpha_{x_i|\mathrm{pa}_{X_i}} = \alpha[x_i, \mathrm{pa}_{X_i}] \ , \tag{2.41}$$

onde $\alpha[x_i, \mathrm{pa}_{X_i}]$ é o número de vezes em que $X_i = x_i$ e $\mathrm{Pa}_{X_i} = \mathrm{pa}_{X_i}$ em $\mathcal{D}'$.

Um problema desta abordagem é que é preciso armazenar um grande conjunto de dados de pseudo-instâncias. Ao invés disso, podemos armazenar o tamanho do conjunto de dados α e uma representação $P'(X_1, \ldots, X_n)$ das frequências de eventos neste banco de dados *a priori*.

Proposição 9 *Seja $P'(X_1, \ldots, X_n)$ uma distribuição de eventos em $\mathcal{D}'$, então temos que:*

$$\alpha_{x_i|\mathrm{pa}_{X_i}} = \alpha \cdot P'(x_i, \mathrm{pa}_{X_i}) \ . \tag{2.42}$$

Podemos representar P' como uma rede Bayesiana, utilizando a inferência de redes Bayesianas para computar os valores de $P'(x_i, \mathrm{pa}_{X_i})$. Mais detalhes a respeito desta distribuição serão vistos no Capítulo 4, onde definiremos P' como um conjunto de independências marginais sobre as variáveis X_is e utilizaremos esta abordagem no cálculo da função de pontuação BDe.

A partir destas informações, somos capazes de obter o modelo de probabilidade global de uma rede Bayesiana, considerando uma rede $\mathcal{G}$ existente e dados de treinamento. No próximo capítulo definimos formalmente o problema de engenharia reversa de GRNs e fornecemos uma breve revisão bibliográfica a respeito do assunto.

Capítulo 3

Engenharia Reversa de Redes de Regulação Gênica

Neste capítulo definiremos um problema desafiador na Biologia Computacional: Realizar a engenharia reversa de redes de regulação gênica com base em dados de expressão. Para o processo de inferência, utilizaremos o modelo de Redes Bayesianas estático e dinâmico.

Como definido anteriormente, uma GRN é composta por um conjunto de genes X e suas interações, onde os genes são representados como nós, ou vértices, e as interações consistem de arestas direcionadas de um nó para outro em uma rede $\mathcal{G}$. Um gene pode assumir dois valores: 1 que representa o estado *ativo* e 0 que representa o estado *inativo*. Desta forma, o problema de engenharia reversa de GRNs pode ser definido como encontrar as relações que governam o comportamento da rede, a partir de dados de expressão gênica disponíveis. Trata-se de um problema inverso mal posto (*ill-posed inverse problem*), já que podem existir diversas redes igualmente capazes de representar os dados de entrada do problema. Logo, podemos definir o problema da inferência de redes da seguinte forma:

Definição do Problema da Engenharia Reversa de GRNs: Seja $\mathcal{X} = \{X_1, \ldots, X_n\}$ um conjunto de genes onde $X_i \in \{0, 1\}$, $i = 1, \ldots, n$, e $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_{M_{\text{seq}}}\}$ o conjunto de dados temporais de expressão gênica, onde $\mathcal{D}_j = \{\mathcal{X}[0], \ldots, \mathcal{X}[M_l]\}$, $j = 1, \ldots, M_{\text{seq}}$ e $\mathcal{X}[t] \in \{0, 1\}^n$, inferir as arestas $(X_i, X_j) \in E$ do grafo $\mathcal{G}(\mathcal{X}, E)$ de forma que $\mathcal{G}$ represente os dados $\mathcal{D}$ probabilisticamente.

Para inferir redes é necessário o estudo de abordagens que busquem pela melhor estrutura de rede, capaz de representar o comportamento dos dados de expressão. Para isso, utilizaremos funções de pontuação, cujo objetivo principal é atribuir uma pontuação a estruturas de redes candidatas, considerando um espaço de busca e estimativas de parâmetros. As funções de pontuação utilizadas neste trabalho serão as funções BIC e BDe, onde ambas possuem a estimativa Bayesiana de parâmetros como base. Além disso, propomos uma equação capaz de aumentar a pontuação de estruturas que possuam relações entre genes comprovadas experimentalmente. Esta informação será retirada de bancos de dados biológicos públicos. A

seguir, realizamos uma breve revisão bibliográfica referente ao tema de engenharia reversa de redes. No próximo capítulo, serão explicados os métodos utilizados para a inferência de redes e as funções de pontuação BIC e BDe. Além disso, veremos mais informações a respeito da biblioteca Pgmpy, utilizada para a implementação da metodologia, como adicionar conhecimento biológico ao processo de inferência de redes e formas de validar os resultados obtidos.

3.1 Revisão Bibliográfica

Algoritmos de inferência de redes são essenciais para auxiliar estudos que buscam identificar as relações causais existentes entre genes, a partir de dados biológicos. Nesta seção, citamos alguns trabalhos envolvendo o modelo de redes Bayesianas ou abordagens Bayesianas para a engenharia reversa de GRNs. Em 2000, Friedman *et al.* (2000) utilizou redes Bayesianas para descobrir interações entre genes baseando-se em várias amostras de expressão gênica, demonstrando o método em dados do ciclo celular da levedura (Spellman *et al.*, 1998). Cada medida de expressão foi tratada como uma amostra independente, não levando em conta o aspecto temporal das amostras. Também não foram considerados conhecimentos prévios sobre a rede, e mesmo assim, a abordagem se mostrou capaz de inferir uma rica estrutura, onde muitas conclusões biológicas plausíveis foram extraídas. Em 2007, Needham *et al.* (2007) publicou uma análise de métodos de inferência de redes e parâmetros Bayesianos, a fim de apresentar propriedades que auxiliem a produção de novos modelos.

Em 2016, Liu *et al.* (2016) propôs o uso de redes Bayesianas locais na reconstrução de redes de regulação gênica a partir de dados de expressão, utilizando informação mútua condicional para a construção de uma rede inicial, e decompondo-a em uma série de pequenas sub-redes, de acordo com a relação mais próxima entre os genes da rede. Para cada rede local, foi aplicado o modelo de redes Bayesianas, a fim de identificar as relações de regulação existentes nestas sub-redes. Por fim, as redes locais eram integradas em uma GRN candidata, onde a rede final apresentou uma quantidade menor de falsos positivos. Neste mesmo ano, pesquisas relacionadas à doença de Alzheimer foram auxiliadas por métodos de inferência de redes. Zhang *et al.* (2016) construiu uma GRN baseados em 1.647 tecidos cerebrais de pacientes com a doença de Alzheimer com início tardio e de pacientes não sujeitos a doença, demonstrando que o Alzheimer reconfigura partes específicas da estrutura de interação molecular. Foram utilizadas redes probabilísticas Bayesianas a fim de localizar potenciais genes reguladores da rede. No processo de inferência de rede, foi aplicado conhecimento *a priori*, onde as estruturas de redes obtidas foram ordenadas de acordo com suas relevâncias (*scores*) e conhecimentos prévios.

Recentemente, Gendelman *et al.* (2017) utilizou redes Bayesianas em redes moleculares de células cancerosas, onde tais redes são sistemas dinâmicos complexos que demonstram comportamentos não intuitivos. Foi utilizada uma nova estratégia computacional para inferir

redes probabilísticas de relacionamento causal baseada em expressão gênica, onde um modelo composto por um conjunto de redes usando dados multidimensionais foi capaz de identificar componentes já conhecidos do maquinário do ciclo celular, bem como revelar novos genes reguladores da doença.

Considerando a temporalidade dos dados biológicos, Friedman *et al.* (1998) apresentou um modelo de redes Bayesianas dinâmicas, mostrando como buscar por uma estrutura quando algumas das variáveis são desconhecidas. O modelo foi aplicado em pequenos exemplos artificiais conhecidos. A tecnologia se mostrou útil para predizer e classificar comportamentos dinâmicos entre as variáveis e aprender ordenações causais envolvidas na regulação gênica.

Em 2013, Chen Chen *et al.* (2013) aplicou redes Bayesianas dinâmicas para inferir GRNs a partir de dados temporais de expressão. Em seu algoritmo, foi adicionado conhecimento prévio, e utilizado o método de Monte Carlo via cadeias de Markov (Hastings, 1970). A validação foi realizada com dados sintéticos e redes já conhecidas como a rede da levedura.

Capítulo 4

Metodologia

Neste capítulo apresentamos a metodologia utilizada neste trabalho para realizar a engenharia reversa de GRNs, por meio dos modelos probabilísticos de redes Bayesianas e redes Bayesianas dinâmicas, e as funções de pontuação BIC e BDe. Além disso, definimos uma equação para adicionar conhecimento biológico ao processo de inferência, fornecemos uma visão geral dos dados de expressão gênica utilizados e a forma de validação dos resultados obtidos.

4.1 Engenharia Reversa de Redes

Uma das razões para se inferir redes é a descoberta do conhecimento: examinando as dependências da rede inferida, é possível aprender sobre a estrutura de dependência das variáveis. O ideal seria recuperar a rede desconhecida $\mathcal{G}^*$. Porém, podem existir mais de uma rede que seja da mesma classe de equivalência de $\mathcal{G}^*$, ou seja, redes que recebem a mesma pontuação, onde todas as redes encontradas sejam igualmente boas, não sendo possível distinguir apenas com base nos dados qual é a rede ótima. Logo, $\mathcal{G}^*$ não é identificável a partir dos dados. O melhor que se pode esperar, é que o algoritmo recupere uma rede de mesma classe de equivalência de $\mathcal{G}^*$.

Inferir a estrutura da rede é uma meta difícil de ser atingida. Os dados amostrais geralmente possuem ruídos e não é possível reconstruir perfeitamente a rede desejada. Geralmente, é necessário tomar decisões ao incluir arestas em uma estrutura cuja relação das variáveis causa incerteza. Por exemplo, se são incluídas arestas duvidosas, o modelo que será inferido conterá uma maior quantidade de arestas esporádicas. Se poucas destas arestas são incluídas, podem ser perdidas algumas dependências. Ambas as escolhas ladeiam para estruturas imprecisas que não representam a estrutura correta. Logo, decidir se é melhor ter correlações esporádicas ou independências falsas depende da aplicação (Koller e Friedman, 2009).

4.2 Inferência Baseada em Funções de Pontuação

Uma das abordagens utilizadas para inferência de redes Bayesianas é a abordagem baseada em funções de pontuação. Uma função de pontuação é definida, de modo que seja capaz de pontuar cada estrutura candidata com base nos dados de treinamento, e então é realizada uma busca pela estrutura que possui maior pontuação. O grande desafio é a escolha desta função.

Uma escolha para a função de pontuação é utilizar a função de máxima verossimilhança, também utilizada na estimativa de parâmetros. Esta função mede a probabilidade dos dados, dado um modelo. Assim, deseja-se encontrar uma rede que torne os dados os mais prováveis possíveis. A meta é encontrar a rede $\mathcal{G}$ e os parâmetros $\boldsymbol{\theta}_{\mathcal{G}}$ que maximizam a verossimilhança do modelo que, no caso, trata-se de um par $\langle \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}} \rangle$.

No Capítulo 2 foi explicado como se maximiza a verossimilhança para uma dada estrutura $\mathcal{G}$. Agora, os parâmetros de máxima verossimilhança $\hat{\boldsymbol{\theta}}_{\mathcal{G}}$ são utilizados para esta estrutura. Uma simples análise mostra que:

$$\begin{aligned} \max_{\mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}}} L(\langle \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}} \rangle : \mathcal{D}) &= \max_{\mathcal{G}} [\max_{\boldsymbol{\theta}_{\mathcal{G}}} L(\langle \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}} \rangle : \mathcal{D})] \\ &= \max_{\mathcal{G}} [L(\langle \mathcal{G}, \hat{\boldsymbol{\theta}}_{\mathcal{G}} \rangle : \mathcal{D})] \ . \end{aligned} \tag{4.1}$$

Em outras palavras, para se encontrar o par $\langle \mathcal{G}, \boldsymbol{\theta}_{\mathcal{G}} \rangle$ que maximiza a verossimilhança, deve-se encontrar a estrutura de rede $\mathcal{G}$ que atinge pontuação mais alta quando são utilizados parâmetros de MLE para a rede $\mathcal{G}$. Logo:

$$\text{score}_L(\mathcal{G} : \mathcal{D}) = \ell(\hat{\boldsymbol{\theta}}_{\mathcal{G}} : \mathcal{D}) \ , \tag{4.2}$$

onde $\ell(\hat{\boldsymbol{\theta}}_{\mathcal{G}} : \mathcal{D})$ é o logaritmo da função de verossimilhança e $\hat{\boldsymbol{\theta}}_{\mathcal{G}}$ são os parâmetros de máxima verossimilhança para $\mathcal{G}$.

A função de máxima verossimilhança, apesar de ser uma boa medida de ajuste de uma rede Bayesiana aos dados de treinamento, pode resultar em alguns problemas. Esta função de pontuação nunca preferirá uma rede mais simples sobre uma estrutura de rede mais complexa, a não ser em raras ocasiões.

Há situações em que é preferível uma estrutura de rede mais simples e a função de máxima verossimilhança nunca fará esta escolha. Logo, a rede inferida a partir desta função de pontuação tenderá a ser um rede densa, com muitas arestas. Em outras palavras, a função de máxima verossimilhança se ajusta demais (*overfitting*) aos dados de treinamento, inferindo um modelo que se encaixa especificamente aos dados empíricos contidos no conjunto de treinamento, falhando ao generalizar bem novos casos de dados: há amostras a partir da distribuição subjacente que não são idênticos aos da distribuição empírica, contidos no conjunto de treinamento. Uma boa alternativa então é a utilização da função de pontuação Bayesiana.

O princípio chave da abordagem Bayesiana é a incerteza. Neste caso, há incerteza em relação a estrutura e aos parâmetros da rede. Logo, definimos uma probabilidade *a priori* $P(\mathcal{G})$ que indica a probabilidade sobre diferentes estruturas de rede e uma probabilidade *a priori* $P(\boldsymbol{\theta} \mid \mathcal{G})$, que indica uma probabilidade em diferentes escolhas de parâmetros uma vez que a rede é fornecida. Pelo teorema de Bayes, tem-se:

$$P(\mathcal{G} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{G})P(\mathcal{G})}{P(\mathcal{D})} \ , \tag{4.3}$$

onde o denominador corresponde ao fator de normalização, que não auxilia a distinguir entre diferentes estruturas. Logo, a função de pontuação Bayesiana é definida como:

$$\text{score}_B(\mathcal{G} \ : \ \mathcal{D}) = \log P(\mathcal{D} \mid \mathcal{G}) + \log P(\mathcal{G}) \ . \tag{4.4}$$

A habilidade de atribuir conhecimento *a priori* sobre estruturas faz com que algumas redes sejam preferíveis sobre outras. Por exemplo, é possível penalizar estruturas densas mais do que estruturas esparsas. No entanto, o termo *a priori* da estrutura na função é quase irrelevante quando comparado ao primeiro termo. Este termo, $P(\mathcal{D} \mid \mathcal{G})$ leva em consideração nossa incerteza sobre os parâmetros:

$$P(\mathcal{D} \mid \mathcal{G}) = \int_{\Theta_\mathcal{G}} P(\mathcal{D} \mid \boldsymbol{\theta}_\mathcal{G}, \mathcal{G}) P(\boldsymbol{\theta}_\mathcal{G} \mid \mathcal{G}) d\boldsymbol{\theta}_\mathcal{G} \ . \tag{4.5}$$

onde $P(\mathcal{D} \mid \boldsymbol{\theta}_\mathcal{G}, \mathcal{G})$ é a verossimilhança dos dados dada a rede $\langle \mathcal{G}, \boldsymbol{\theta}_\mathcal{G} \rangle$ e $P(\boldsymbol{\theta}_\mathcal{G} \mid \mathcal{G})$ é a nossa distribuição *a priori* sobre diferentes valores de parâmetros para a rede $\mathcal{G}$. $P(\mathcal{D} \mid \mathcal{G})$ é chamada de *verossimilhança marginal* dos dados, dada a estrutura.

A verossimilhança marginal é um pouco diferente da função de máxima verossimilhança. Ambas examinam a verossimilhança dos dados, dada a estrutura. A função de máxima verossimilhança retorna o valor que maximiza a função. Em contraste, a verossimilhança marginal é o valor médio desta função, onde o valor médio é calculado baseado na fórmula $P(\boldsymbol{\theta}_\mathcal{G} \mid \mathcal{G})$.

A abordagem Bayesiana nos diz que, embora a escolha de parâmetros $\hat{\boldsymbol{\theta}}$ sejam os mais prováveis dado o conjunto de treinamento $\mathcal{D}$, eles não são a única alternativa. A *posteriori* sobre os parâmetros nos provê uma variedade de escolhas, juntamente com uma medida de quão provável é cada um deles. Integrando $P(\mathcal{D} \mid \boldsymbol{\theta}_\mathcal{G}, \mathcal{G})$ sobre diferentes escolhas de parâmetros $\boldsymbol{\theta}_\mathcal{G}$, estamos medindo a verossimilhança esperada, calculando a média sobre diferentes escolhas possíveis de $\boldsymbol{\theta}_\mathcal{G}$. Logo, estamos sendo mais conservadores em nossa estimativa do modelo ideal.

4.2.1 Score Bayesiano para Redes Bayesianas

Seja $\mathcal{G}$ uma estrutura de rede e $P(\boldsymbol{\theta}_{\mathcal{G}} \,|\, \mathcal{G})$ uma distribuição *a priori* que satisfaz a independência global de parâmetros. Então,

$$P(\mathcal{D} \,|\, \mathcal{G}) = \prod_i \int_{\Theta_{X_i|\mathrm{Pa}_{X_i}}} \prod_m P(x_i[m] \,|\, \mathrm{pa}_{X_i}[m], \boldsymbol{\theta}_{X_i \,|\, \mathrm{Pa}_{X_i}}, \mathcal{G}) P(\boldsymbol{\theta}_{X_i \,|\, \mathrm{Pa}_{X_i}} \,|\, \mathcal{G}) d\boldsymbol{\theta}_{X_i \,|\, \mathrm{Pa}_{X_i}} \ . \quad (4.6)$$

A seguir, definiremos os conceitos de duas funções de pontuação utilizadas neste trabalho para a inferência de redes, ambas baseadas na formulação Bayesiana: as funções de pontuação BIC e BDe.

4.2.2 Função de Pontuação BIC

Como visto anteriormente, a função de pontuação Bayesiana tende a escolher redes que possuam estruturas mais simples. Porém, conforme mais dados são analisados, torna-se necessária a inferência de estruturas que sejam mais complexas. Em outras palavras, deve haver um equilíbrio entre adequar os dados à estrutura do modelo ao mesmo tempo em que a extensão do *overfitting* é reduzida. Para entender este comportamento, segundo Koller e Friedman (2009), é necessário considerar uma aproximação da função de pontuação Bayesiana que melhor se aproxime das propriedades fundamentais da rede.

Utilizando parâmetros com a distribuição de Dirichlet para todos os parâmetros da rede, quando $M \to \infty$, temos que:

$$\log P(\mathcal{D} \,|\, \mathcal{G}) = \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{G}}} \,:\, \mathcal{D}) - \frac{\log M}{2} \mathrm{Dim}[\mathcal{G}] + O(1) \ , \quad (4.7)$$

onde $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{G}}}$ são os parâmetros em $\mathcal{G}$ que maximizam a verossimilhança dos dados, M é o número de instâncias de treinamento, $O(1)$ é um termo constante que é independente de M e $\mathcal{G}$ e, por fim, $\mathrm{Dim}[\mathcal{G}]$ é a dimensão do modelo (número de parâmetros), que pode ser definido como:

$$\mathrm{Dim}[\mathcal{G}] = \sum_i \sum_{p \in \mathbf{Pa}(X_i)} ||X_p|| \cdot (||X_i|| - 1) \ . \quad (4.8)$$

Assim, a função de pontuação Bayesiana tende a equilibrar a verossimilhança - *overfitting* - e a complexidade do modelo. Esta aproximação é chamada de BIC (*Bayesian information criterion*):

$$\mathrm{score}_{BIC}(\mathcal{G} \,:\, \mathcal{D}) = \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{G}}} \,:\, \mathcal{D}) - \frac{\log M}{2} \mathrm{Dim}[\mathcal{G}] \ . \quad (4.9)$$

4.2.3 Função de Pontuação BDe

A função de pontuação BDe (*Bayesian Dirichlet likelihood-equivalence*) também combina a verossimilhança dos dados de acordo com a rede, com alguma penalidade de acordo com a complexidade da rede. Ao aprender uma estrutura de rede, uma penalidade de acordo com

a complexidade é essencial já que a rede obtida pela estimativa de máxima verossimilhança tende a ter muitas arestas. Além disso, esta função também leva em conta a representação de parâmetros *a priori*, utilizando a distribuição de Dirichlet.

Considere uma rede que utiliza a distribuição de Dirichlet onde $P(\boldsymbol{\theta}_{X_i|\mathrm{pa}_{X_i}} \,|\, \mathcal{G})$ possui hiperparâmetros $\{\alpha^{\mathcal{G}}_{x_i^j|\boldsymbol{u}_i} : j = 1, \ldots, |X_i|\}$. Logo, é possível reescrever a Equação 4.6 utilizando a seguinte fórmula, considerando o espaço de variáveis e os hiperparâmetros:

$$P(\mathcal{D} \,|\, \mathcal{G}) = \prod_{\boldsymbol{u}_i \in Val(\mathrm{Pa}^{\mathcal{G}}_{X_i})} \frac{\Gamma(\alpha^{\mathcal{G}}_{X_i|\boldsymbol{u}_i})}{\Gamma(\alpha^{\mathcal{G}}_{X_i|\boldsymbol{u}_i} + M[\boldsymbol{u}_i])} \prod_{x_i^j \in Val(X_i)} \left[\frac{\Gamma(\alpha^{\mathcal{G}}_{x_i^j|\boldsymbol{u}_i} + M[x_i^j, \boldsymbol{u}_i])}{\Gamma(\alpha^{\mathcal{G}}_{x_i^j|\boldsymbol{u}_i})} \right] , \tag{4.10}$$

onde $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \, \mathrm{d}t$ é a função *Gamma* que satisfaz as propriedades $\Gamma(1) = 1$ e $\Gamma(x+1) = x\Gamma(x)$ e $\alpha^{\mathcal{G}}_{X_i|\boldsymbol{u}_i} = \sum_j \alpha^{\mathcal{G}}_{x_i^j|\boldsymbol{u}_i}$. Na prática, é utilizado o logaritmo desta equação a fim de simplificar o cálculo. Mais detalhes podem ser vistos em Koller e Friedman (2009).

A função de pontuação BDe assume que há uma distribuição P' sobre todo o espaço de probabilidade e uma quantidade de amostras equivalentes para o conjunto de amostras imaginárias. Logo, o conjunto de parâmetros é definido da seguinte forma:

$$\alpha^{\mathcal{G}}_{x_i^j|\boldsymbol{u}_i} = \alpha \cdot P'(x_i^j, \boldsymbol{u}_i) \ . \tag{4.11}$$

Como discutido na Seção 2.6.2, podemos representar P' como uma rede Bayesiana cuja estrutura pode representar nosso conhecimento *a priori* sobre a estrutura do domínio. Para este trabalho, atribuímos a P' uma rede Bayesiana inicialmente vazia com uma distribuição marginal uniforme para cada variável. É importante notar que a estrutura de rede é usada apenas para prover parâmetros *a priori*, não sendo utilizada para guiar a estrutura de busca.

4.2.4 Estrutura de Busca

As funções de pontuação de verossimilhança e Bayesiana são úteis na avaliação da qualidade de diferentes estruturas de redes candidatas. Agora, o problema discutido será de como encontrar uma estrutura de rede que possua a pontuação mais alta. A princípio, tem-se um problema de otimização bem definido, onde a entrada do problema é constituída por:

1. um conjunto de dados de treinamento $\mathcal{D}$;
2. uma função de pontuação (incluindo algum conhecimento a priori, se necessário);
3. um conjunto $\boldsymbol{\mathcal{G}}$ de possíveis estruturas de redes (incorporando algum conhecimento prévio).

A saída esperada é uma estrutura de rede (a partir de um conjunto de estruturas possíveis) que maximiza a função de pontuação.

Considere um espaço de busca, ou seja, um grafo composto por estruturas de redes candidatas, onde seja possível a movimentação do procedimento de busca entre diferentes

soluções.

Cada estado do grafo deve possuir poucos vizinhos, para garantir que o diâmetro do espaço de busca seja pequeno. Uma boa alternativa de vizinhança para os estados é representar sempre a mesma estrutura de rede, diferenciando uma rede de outra por meio de uma pequena modificação local. Logo, a conectividade do espaço de busca em termos de operadores é definida como:

1. adicionar uma aresta;
2. deletar uma aresta;
3. reverter uma aresta.

Os estados adjacentes a um estado $\mathcal{G}$ no grafo serão aqueles onde uma aresta estará alterada, adicionando-se, deletando-se, ou revertendo a sua orientação. As operações devem ser legais. Em Redes Bayesianas, por exemplo, o grafo deve permanecer acíclico. Esta propriedade garante que o diâmetro do espaço de busca seja no máximo n^2, onde n é o número de variáveis, havendo um caminho relativamente curto entre quaisquer duas redes escolhidas.

Como a função de pontuação de uma rede $\mathcal{G}$ é a soma de pontuações de redes locais, as operações consideradas resultam apenas em alterar um termo da função local (no caso de adicionar ou deletar uma aresta) ou dois termos (no caso de reverter uma aresta). Realizando-se uma mudança local na função de pontuação, a maioria dos componentes da função permanecem iguais. Isto implica que há algum senso de "continuidade" ao pontuar redes vizinhas.

Uma vez que é definido o espaço de busca, é necessário designar um procedimento para explorá-lo e pesquisar por estados de pontuação mais alta. No contexto de inferência, escolhe-se uma estrutura inicial de rede, onde tal estrutura pode ser uma rede vazia, uma rede aleatória, ou uma rede que contenha algum conhecimento prévio. É atribuída à rede uma pontuação. Então, considerando-se todos os vizinhos da rede no espaço - todas as redes legais obtidas aplicando-se um único operador a $\mathcal{G}$ - computamos a pontuação de cada uma delas. Este processo continua até que a pontuação obtida seja melhorada.

Quanto ao custo computacional, pode-se dizer que quando n é grande, considerar $O(n^2)$ vizinhos em cada iteração pode ser bem custoso. Entretanto, a maioria dos operadores tende a executar uma alteração ruim para a rede. Logo, para evitar o custo excessivo, é necessário utilizar procedimentos de busca que substituam a busca exaustiva como o *greedy hill climbing* (Buntine, 1991, Heckerman *et al.*, 1995).

Greedy hill climbing deve ser utilizado com cautela. Se uma rede é retornada pelo procedimento, é porque não há como melhorá-la aplicando-se um único operador (isto é, alterando-se uma das arestas). Isto pode implicar em duas situações. É possível que o algoritmo tenha atingido um *máximo local*, onde qualquer mudança na rede resulte em uma diminuição da pontuação. A outra opção é que o algoritmo tenha atingido um *plateau*: um grande conjunto de redes vizinhas que possuam mesma pontuação. Logo, o procedimento *greedy hill climbing* não é capaz de "navegar" pelo *plateau*, uma vez que se baseia na melhoria da pontuação para

orientá-lo para estruturas melhores. Pode ser que a partir de uma rede existente na região do *plateau* seja possível continuar o algoritmo, encontrando-se uma rede com maior pontuação, ou que esta classe de equivalência seja um máximo local. Porém, *greedy hill climbing* não é capaz de fazer esta diferenciação.

Uma estratégia é utilizar busca *tabu*. Este procedimento mantém uma lista de operadores que foram aplicados no algoritmo, e em cada passo, não são permitidos operadores que revertem o efeito de operações recentemente aplicadas. Uma vez que a busca decide adicionar uma aresta $X \rightarrow Y$, tal aresta não pode ser deletada nos próximos L passos. Tal estratégia permite que a busca continue sendo realizada, mesmo após atingir um máximo local, continuando a procura por uma estrutura que possua pontuação máxima (Koller e Friedman, 2009).

Para realizar o processo de engenharia reversa de redes, será utilizada a biblioteca Pgmpy, uma biblioteca implementada a partir da linguagem de programação Python, que já possui algumas funcionalidades como inferência de BNs utilizando a estimativa Bayesiana de parâmetros, a função de busca exaustiva *greedy hill climbing*, o recurso de busca *tabu* e as funções de pontuação BIC e BDe. Esta biblioteca será modificada, permitindo que o modelo de redes Bayesianas lide com dados de expressão gênica temporais. Além disso, será implementado o modelo de redes Bayesianas dinâmicas, permitindo a adição de conhecimento biológico à priori ao processo de inferência.

4.3 Biblioteca Pgmpy

Um modelo gráfico probabilístico (PGM) é uma técnica de representação de distribuição conjunta sobre variáveis aleatórias de uma maneira compacta, explorando as dependências entre as variáveis. PGMs utilizam uma estrutura de rede para codificar as relações entre as variáveis aleatórias e alguns parâmetros para representar a distribuição conjunta. Pgmpy é uma biblioteca implementada em Python, que trabalha com modelos gráficos probabilísticos (redes Bayesianas, Redes de Markov e variantes). Esta biblioteca, por ser bem documentada e de código aberto, possibilita ao usuário fácil extensibilidade para escrever seus próprios algoritmos ou editar o código existente. Pgmpy implementa o modelo de Redes Bayesianas estático proposto por Koller e Friedman (2009) e pode ser acessada por meio do endereço eletrônico https://github.com/pgmpy.

O objetivo deste trabalho é utilizar as funcionalidades oferecidas pela biblioteca no processo de engenharia reversa de GRNs, modificando o algoritmo de redes Bayesianas, permitindo que o modelo lide com dados temporais, bem como adicionar redes Bayesianas dinâmicas às suas funcionalidades. A partir disso, verificar o comportamento dos resultados diante de variações de parâmetros, alternando a função de pontuação utilizada, adicionando conhecimento biológico *a priori* ao algoritmo de pontuação das estruturas candidatas e diferenciando o grau máximo de entrada da rede, a fim de comparar as taxas de similaridade

das redes obtidas em relação às redes reais.

4.4 Modificações Propostas

A seguir, são descritas as modificações realizadas na biblioteca Pgmpy a fim de lidar com dados biológicos de expressão gênica temporais.

4.4.1 Considerando o fator tempo em redes Bayesianas

O algoritmo de redes Bayesianas estático contido na biblioteca Pgmpy assume que, dada uma amostra $\mathcal{X}[t]$, genes reguladores ativam o comportamento de um determinado gene alvo de forma instantânea. Considerando a Tabela 4.1 como exemplo, se analisamos a possibilidade do gene X_1 ser pai do gene X_2, assumimos que o gene alvo (X_2) muda seu comportamento com base no comportamento do gene regulador (X_1) sempre no mesmo instante de tempo t.

	X_1	X_2	X_3
t_1	0	1	1
t_2	1	0	0
t_3	1	0	1

Tabela 4.1: *Exemplo de dados de expressão contendo 3 genes (X_1, X_2 e X_3) e 3 amostras temporais.*

Quando lidamos com dados temporais, consideramos que genes reguladores podem ativar o comportamento de genes alvo em instantes de tempo distintos. Podemos observar na Tabela 4.2 um exemplo de dados de expressão gênica temporais, onde o gene X_1 no tempo t ativa o comportamento do gene alvo no tempo posterior $(t + 1)$.

	X_1	X_2	X_3
t_1	0	1	1
t_2	1	1	0
t_3	1	0	1

Tabela 4.2: *Exemplo do comportamento dos dados de expressão gênica temporais.*

Logo, para lidar com dados temporais, uma pequena modificação na biblioteca Pgmpy foi realizada. A fim de aumentar a acurácia das redes inferidas, um deslocamento é feito na coluna que contém o gene regulado no instante em que é calculada a função que atribui uma pontuação para uma estrutura candidata. Levando em conta o exemplo citado acima, os dados de expressão ficariam de acordo com a Tabela 4.3. Com o deslocamento, a última amostra do gene alvo é perdida, e são consideradas apenas as amostras completas durante

o cálculo de pontuação. Quando o cálculo é finalizado, os dados de expressão voltam a sua forma original.

	X_1	X_2	X_3
t_1	0	1	1
t_2	1	0	0
t_3	1	*null*	1

Tabela 4.3: *Exemplo de dados de expressão gênica temporais após deslocamento da coluna do gene alvo.*

4.4.2 Implementação de Redes Bayesianas Dinâmicas

A fim de tratar as limitações do algoritmo de redes Bayesianas estático, a biblioteca Pgmpy foi modificada de modo a permitir a execução do algoritmo de redes Bayesianas dinâmicas, conforme definido na Seção 2.4. O intuito é conseguir representar ciclos existentes em relações gênicas e além disso, permitir que um gene se auto-regule. Com esta modificação, buscamos lidar melhor com dados temporais e obter redes com maiores taxas de acerto e similaridade.

4.5 Dados Biológicos

Foram utilizados dados sintéticos e dados do ciclo celular da levedura no processo de engenharia reversa de GRNs. A seguir são descritos maiores detalhes a respeito desses dados.

4.5.1 Dados Sintéticos

Os dados de expressão gênica utilizados neste trabalho foram retirados de *DREAM4 Challenge - In Silico Network Challenge* (Marbach *et al.*, 2010). *DREAM Challenge* desafia usuários a inferir a estrutura de rede a partir de dados de expressão, disponibilizando a GRN que deu origem às amostras após o término da competição. Os arquivos utilizados foram os de séries temporais (*time series*), que mostram como uma rede responde a uma perturbação e como ela relaxa após a remoção da perturbação.

Os dados são divididos em duas categorias: arquivos de expressão gênica contendo 10 genes com 5 séries temporais diferentes e, arquivos que contém 100 genes e 10 séries temporais. Cada série temporal possui 21 amostras. A condição inicial sempre corresponde a um estado estacionário. No tempo $t = 0$, a perturbação é aplicada como descrita a seguir: a primeira metade das séries temporais mostra a resposta da rede à perturbação. A partir da segunda metade, a perturbação é removida, e então os níveis de expressão gênica voltam a partir da perturbação para o estado estacionário.

4.5.2 Dados do Ciclo Celular da Levedura

Foram utilizados dados de um catálogo de genes encontrados na levedura *Saccharomyces cereviseae*. Os dados contém cerca de 6.000 genes. Tais dados são divididos em experimentos, projetados para examinar os efeitos da indução de alguns componentes que interagem com os genes. Cada experimento pode ser denominado como uma série temporal, contendo uma determinada quantidade de amostras.

A primeira série temporal é composta por duas amostras com a adição do componente cln3. A segunda série temporal contém a indução do componente clb2, sendo composta por duas amostras. A terceira série temporal contém adição do fator alpha, sendo composta por 18 amostras. A quarta série temporal possui adição do componente cdc15, contando com 24 amostras. A quinta série temporal é caracterizada pela adição do componente cdc28, possuindo 17 amostras. Por último, temos a série temporal de elutriação, composta por 14 amostras. Mais informações a respeito dos experimentos podem ser vistas em Spellman *et al.* (1998).

Foram selecionados 11 genes presentes no ciclo celular da levedura, proposto por Li *et al.* (2004). Os genes são Cln3, SBF, Sic1, MBF, Cln1, Clb5, Cdh1, Clb1, Mcm1, Cdc20 e Swi5. A rede original possui arestas de ativação e inibição entre os genes. Neste trabalho, consideramos os casos de ativação e inibição como *relações* existentes entre os genes. Tais relações podem ser observadas na Figura 4.1.

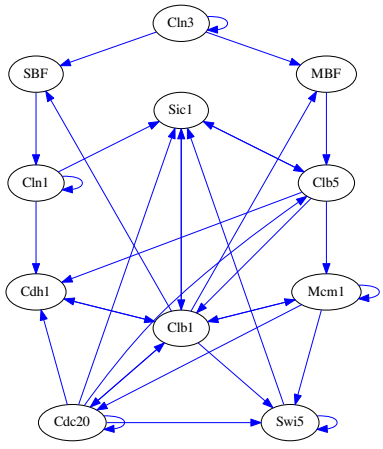


Figura 4.1: *Rede composta pelos 11 genes que governam o comportamento do ciclo celular da levedura, proposto por Li et al. (2004).*

No processo de inferência de redes utilizando os dados biológicos da levedura, foi proposta a adição de conhecimento biológico *a priori* ao algoritmo que pontua estruturas candidatas, tanto da função de pontuação BIC quanto da função BDe. A seguir, são citados alguns bancos de dados biológicos que podem nos auxiliar a aumentar a taxa de similaridade das redes obtidas e como esta informação é adicionada ao algoritmo.

4.6 Bancos de Dados Biológicos

Existem bancos de dados biológicos que armazenam informação sobre interações gênicas conhecidas. Tais bancos de dados podem ser de extrema importância para a inferência de GRNs, já que são capazes de enriquecer a estrutura. O objetivo é aumentar a pontuação de uma aresta que possua alta probabilidade de existir no modelo. Embora a adição de conhecimento prévio seja benéfica, são poucos os algoritmos de inferência que o aplicam.

Exemplos de bancos de dados são *NCBI Gene (National Center for Biotechnology Information)* (Brown *et al.*, 2015), um banco de dados cujos registros possuem nomenclaturas, localização genômica, produtos gênicos e seus atributos, expressão, interações e caminhos, contando com mais de 16 milhões de genes, incluindo vírus, procariontes e eucariontes; *KEGG* (*Kyoto Encyclopedia of Genes and Genomes*) (Kanehisa *et al.*, 2017), uma enciclopédia de genes e genomas, cujo objetivo é atribuir significados funcionais a genes e genomas, a nível molecular ou superior; *STRING* (*Search Tool for the Retrieval of Interacting Genes/Proteins*) que realiza a integração de interações funcionais entre proteínas, incluindo interações diretas (físicas), e indiretas (funcionais), desde que ambas sejam específicas e biologicamente significativas (Szklarczyk *et al.*, 2016); e *BioGRID* (*Biological General Repository for Interaction Datasets*), um repositório contendo interações gênicas e proteicas e associações químicas de organismos modelo (Chatr-aryamontri *et al.*, 2017).

Dos bancos de dados biológicos apresentados, utilizaremos neste trabalho o banco de dados *STRING* para extrair conhecimento a priori de relações entre genes, aplicando o conhecimento no algoritmo de inferência, de modo que redes que contenham as informações almejadas recebam uma pontuação mais alta. Dessa forma, o objetivo é fazer com que as redes obtidas sejam mais verossímeis.

4.6.1 Adicionando Conhecimento *a priori* ao Algoritmo

Foram pesquisados bancos de dados contendo interações conhecidas entre genes e proteínas que fornecessem uma pontuação que representasse o quanto dois genes (regulador e alvo) estavam relacionados. Na banco de dados *STRING* foram encontradas informações a respeito de interações conhecidas entre genes presentes no ciclo celular da levedura. Seja $s(X_i, X_j)$ o valor da evidência de associação entre estes dois genes. Primeiramente, as pontuações passaram por um processo de normalização. Dessa forma, $0 \leq s(X_i, X_j) \leq 1$. Assim, no instante em que a função de pontuação de uma determinada estrutura candidata é calculada, temos que o *score biológico* é definido como:

$$\beta(X_i, \mathrm{Pa}_{X_i}) = \frac{1}{|\mathrm{Pa}_{X_i}|} \sum_{X_j \in \mathrm{Pa}_{X_i}} s(X_i, X_j) \ . \tag{4.12}$$

A decomposição da Equação 2.9 (*log-verossimilhança*) facilita a computação das funções de pontuação BIC e BDe. Note que a verossimilhança é expressa como a soma de termos,

onde cada termo depende apenas da probabilidade condicional da variável X_i dado seus pais Pa_{X_i}. Logo, quando a computação das funções BIC ou BDe é realizada, o termo $\omega \cdot \beta(X_i, \mathrm{Pa}_{X_i})$ é adicionado, onde $\omega \in \mathbb{R}$ é o peso atribuído a equação de *score biológico*. Assim, $\omega = 0$ significa que não estamos levando em consideração a adição de conhecimento *a priori* ao cálculo da função que pontua estruturas candidatas, $\omega = 1$ significa que estamos atribuindo peso 1 à equação, e assim por diante.

4.7 Validação

A validação da metodologia é muito importante no processo de engenharia reversa. Na Figura 4.2 podemos observar um fluxograma com o processo de inferência a partir de dados sintéticos gerados por uma rede $\mathcal{G}^*$ (*gold standard*). Neste caso, como os dados de entrada são gerados por uma rede conhecida, é possível validar o algoritmo. O algoritmo é baseado em um modelo matemático - redes Bayesianas - e possui como saída uma rede $\mathcal{G}$. Para validar o algoritmo, podemos comparar a rede inferida $\mathcal{G}$ com a rede $\mathcal{G}^*$. Durante esse processo, o algoritmo pode sofrer alterações para que seja aperfeiçoado e validado novamente. O processo continua até que sejam obtidos resultados satisfatórios. A rede $\mathcal{G}^*$ usada como referência pode ser uma rede biológica onde as interações gênicas já foram comprovadas por meio de experimentos e encontram-se na literatura, ou uma rede sintética que pode ser construída incluindo algumas características desejadas (*e.g. scale free* (Khanin e Wit, 2006, Lopes *et al.*, 2011, Barabási, 2009)).

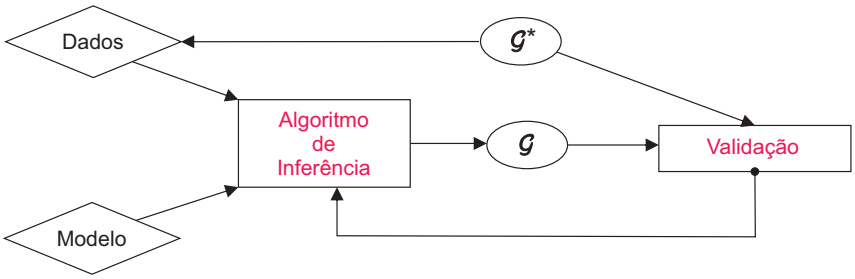


Figura 4.2: *Fluxograma da metodologia de inferência de GRNs.*

Uma das maneiras mais comuns para validar metodologias de inferência de redes é utilizar medidas como a taxa de verdadeiros positivos e falsos positivos, calculados a partir de uma matriz de confusão (Tabela 4.4).

Aresta	Inferida em $\mathcal{G}$	Não inferida em $\mathcal{G}$
Presente em $\mathcal{G}^*$	TP	FN
Ausente em $\mathcal{G}^*$	FP	TN

Tabela 4.4: *Matriz de confusão. TP = verdadeiro positivo, FN = falso negativo, FP = falso positivo e TN = verdadeiro negativo.*

A partir da matriz de confusão, a taxa de verdadeiro positivo (TPR) é definida por:

$$\mathrm{TPR}(k) = \frac{\mathrm{TP}(k)}{\mathrm{TP}(k) + \mathrm{FN}(k)} \tag{4.13}$$

e a taxa de falso positivo (FPR) é definida por:

$$\mathrm{FPR}(k) = \frac{\mathrm{FP}(k)}{\mathrm{FP}(k) + \mathrm{TN}(k)} \ , \tag{4.14}$$

onde $k > 1$ é um limiar indicando a quantidade de arestas a serem consideradas a partir de uma lista de arestas ordenadas por uma determinada pontuação. É possível calcular a similaridade entre a rede obtida $\mathcal{G}^*$ e a rede original $\mathcal{G}$ (Dougherty, 2007) da seguinte forma:

$$\mathrm{Similaridade}(k) = \sqrt{\mathrm{PPV}(k) \times \mathrm{Especificidade}(k)} \ , \tag{4.15}$$

onde

$$\mathrm{PPV(k)} = \frac{\mathrm{TP}(k)}{\mathrm{TP}(k) + \mathrm{FP}(k)} \tag{4.16}$$

e

$$\mathrm{Especificidade}(k) = \frac{\mathrm{TN}(k)}{\mathrm{TN}(k) + \mathrm{FP}(k)} \ . \tag{4.17}$$

Um algoritmo de inferência que apresenta bons resultados na validação pode ser utilizado para gerar novas hipóteses de regulação para serem testadas em laboratório (Figura 4.3). A análise de GRNs também pode auxiliar no desenvolvimento de novas drogas para o tratamento de doenças pois genes/proteínas alvo podem ser identificados no modelo e depois testados experimentalmente (Csermely *et al.*, 2013). Por exemplo, um estudo recente (Khurana *et al.*, 2013) sugere que em GRNs, genes que são essenciais tendem a ser altamente conectados (*hubs*).

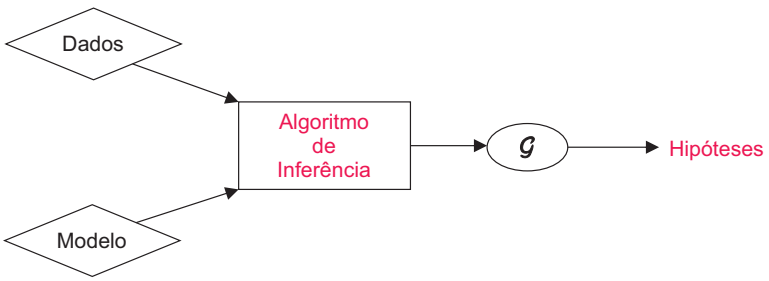


Figura 4.3: *Usando um algoritmo de inferência para gerar hipóteses.*

Capítulo 5

Resultados

A seguir, analisamos as redes obtidas a partir do processo de engenharia reversa por meio dos modelos de redes Bayesianas (BN), com deslocamento da coluna do gene alvo, e redes Bayesianas dinâmicas (DBN). A biblioteca Pgmpy recebe como entrada uma rede inicialmente vazia, o método de busca utilizado é o *greedy hill climbing*, com busca *tabu* para evitar que o algoritmo atinja regiões de *plateau*. As redes foram inferidas levando em conta dois tipos de funções de pontuação: BIC e BDe. Além disso, foram considerados diferentes valores do parâmetro que delimita a quantidade máxima de arestas que podem chegar a um nó, chamado de *grau máximo de entrada* da rede. Para representar este parâmetro, vamos utilizar a variável Δ, desta forma, consideramos $2 \leq \Delta \leq 4$. Os dados de expressão foram inicialmente discretizados, por meio do algoritmo *Bikmeans* proposto por Li *et al.* (2010).

Em relação às redes que foram inferidas por meio do modelo de redes Bayesianas dinâmicas, definido na Seção 2.4, foram consideradas somente as redes de transição ($B^{\rightarrow}$), sendo que as relações existentes entre genes que ocorrem do tempo t para o tempo $t+1$ foram unidas na rede. Assim, tais redes foram representadas contendo ciclos, permitindo por exemplo, uma relação da forma $X \rightarrow X$, onde um gene se auto regula - o que equivale a uma relação do gene X no tempo t com ele mesmo no tempo $t+1$ - e relações da forma $X \rightarrow Y$ e $Y \rightarrow X$, ou seja, uma relação em que um gene X seja capaz de regular um gene Y ao mesmo tempo que este mesmo gene Y seja regulador do gene X - o que equivale ao gene X no tempo t regular a variável Y no tempo $t+1$ ao mesmo tempo em que a variável Y no tempo t regula a variável X no tempo $t+1$.

As redes obtidas foram comparadas às redes *gold standard*. Foram calculadas as taxas de verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo, bem como a taxa de similaridade descrita na Seção 4.7.

5.1 Dados Sintéticos

Os dados de expressão gênica provenientes do *DREAM Challenge*, são compostos por 5 arquivos que contêm 10 genes e 5 séries temporais e, 5 arquivos que contêm 100 genes e

10 séries temporais, onde cada série temporal possui 21 amostras. A seguir, são apresentados os resultados, gráficos comparativos e discussões referentes aos experimentos realizados utilizando estes dados.

5.1.1 Função de Pontuação BIC

Os gráficos que comparam os valores de similaridade, considerando redes de 10 genes (redes $R1, \ldots, R5$) e os dois modelos propostos, BN e DBN, podem ser observados na Figura 5.1. Os três primeiros gráficos refletem os resultados obtidos utilizando-se a função de pontuação BIC no processo de inferência, com o parâmetro Δ variando de 2 a 4, que indica o grau máximo de entrada da rede (Figuras 5.1(a), 5.1(b) e 5.1(c), respectivamente). Conforme os resultados, temos que o modelo que se adequou melhor aos dados foi o modelo de redes Bayesianas dinâmicas, com taxa de similaridade média de 70% entre as redes, contra similaridade média de 56% observada pelo modelo de redes Bayesianas estático. Além disso, as redes inferidas com a função BIC se mostraram menos sensíveis a variação do parâmetro Δ, apresentando resultados equivalentes para os três valores testados.

Na Figura 5.2 podem ser vistos os valores de similaridade relacionados às redes que possuem 100 genes e utilizam a função de pontuação BIC (Figuras 5.2(a), 5.2(b) e 5.2(c)). Podemos observar que os resultados refletiram o comportamento da metodologia quando aplicada às redes pequenas. As redes permaneceram estáveis em relação a variação do parâmetro Δ, apresentando diferença mínima de resultados somente entre os gráficos das Figuras 5.2(a) e 5.2(b). As redes inferidas com o modelo de redes Bayesianas dinâmicas apresentaram melhores resultados, com taxa de similaridade média de 46%, contra similaridade média de 42% observada com a utilização do modelo de redes Bayesianas estático.

As redes obtidas pelo processo de engenharia reversa referentes à rede $R1$, que possuem 10 genes, podem ser observadas na Figura 5.3. A primeira rede, Figura 5.3(a), utiliza o modelo de redes Bayesianas e a segunda, Figura 5.3(b), o modelo de redes Bayesianas dinâmicas. Ambas as redes foram inferidas utilizando a função de pontuação BIC. As arestas azuis são as arestas verdadeiro positivas, ou seja, as arestas que foram corretamente inferidas. As arestas vermelhas são as falso positivas, ou seja, arestas que foram inferidas pelo algoritmo, porém não estão presentes na rede real. E por último, temos as arestas pontilhadas, arestas falso negativas que estão presentes na rede real, porém não foram inferidas pelo algoritmo. É possível perceber que as redes inferidas utilizando o modelo de redes Bayesianas dinâmicas apresentaram uma quantidade menor, tanto de arestas falso positivas como de arestas falso negativas, sendo mais similares à rede original.

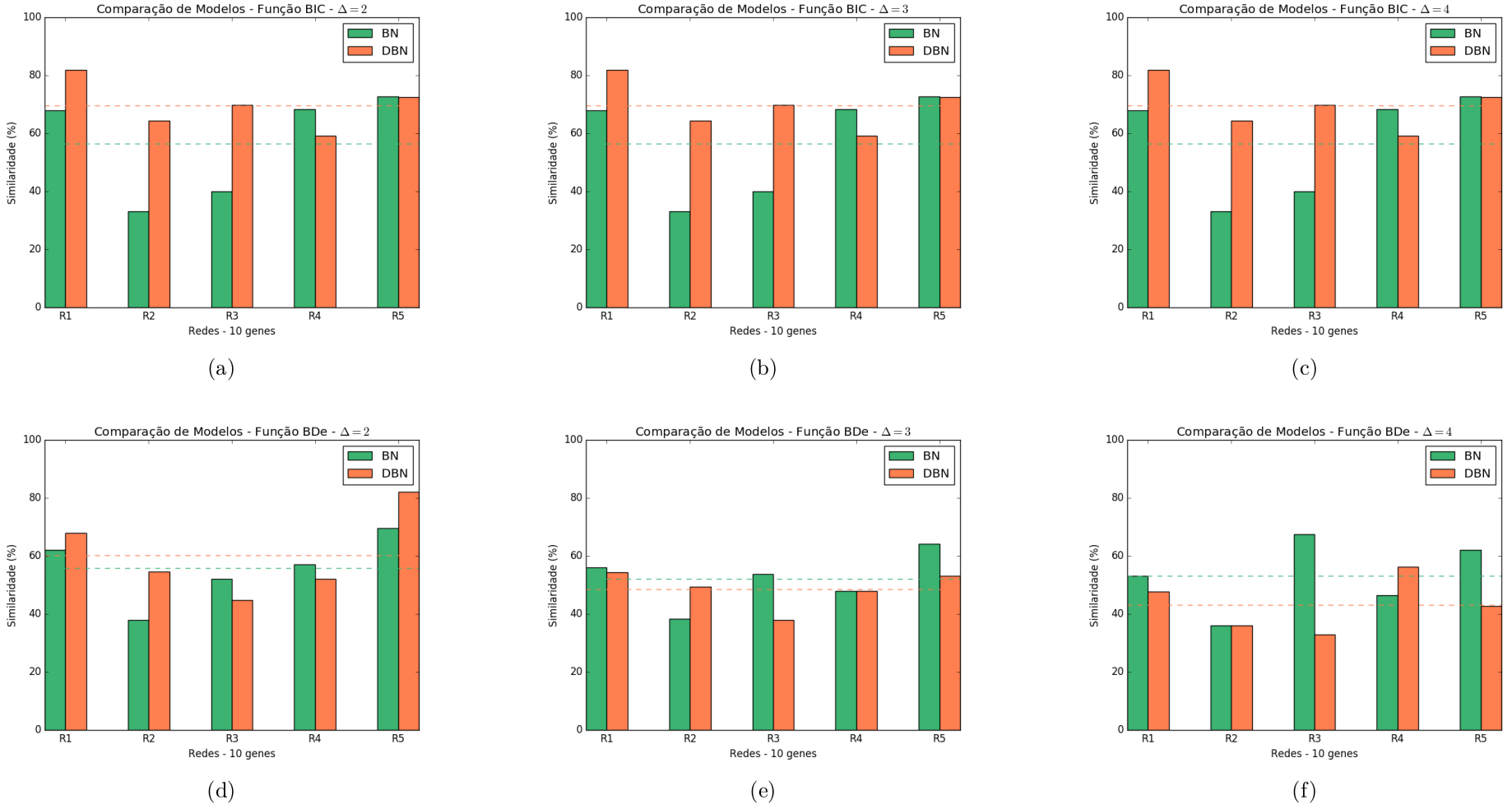


Figura 5.1: *Gráficos referentes aos resultados de similaridade obtidos no processo de inferência, utilizando os modelos de redes Bayesianas e redes Bayesianas dinâmicas para as redes de* 10 *genes (DREAM Challenge), considerando as funções de pontuação BIC e BDe e variando o parâmetro* Δ.

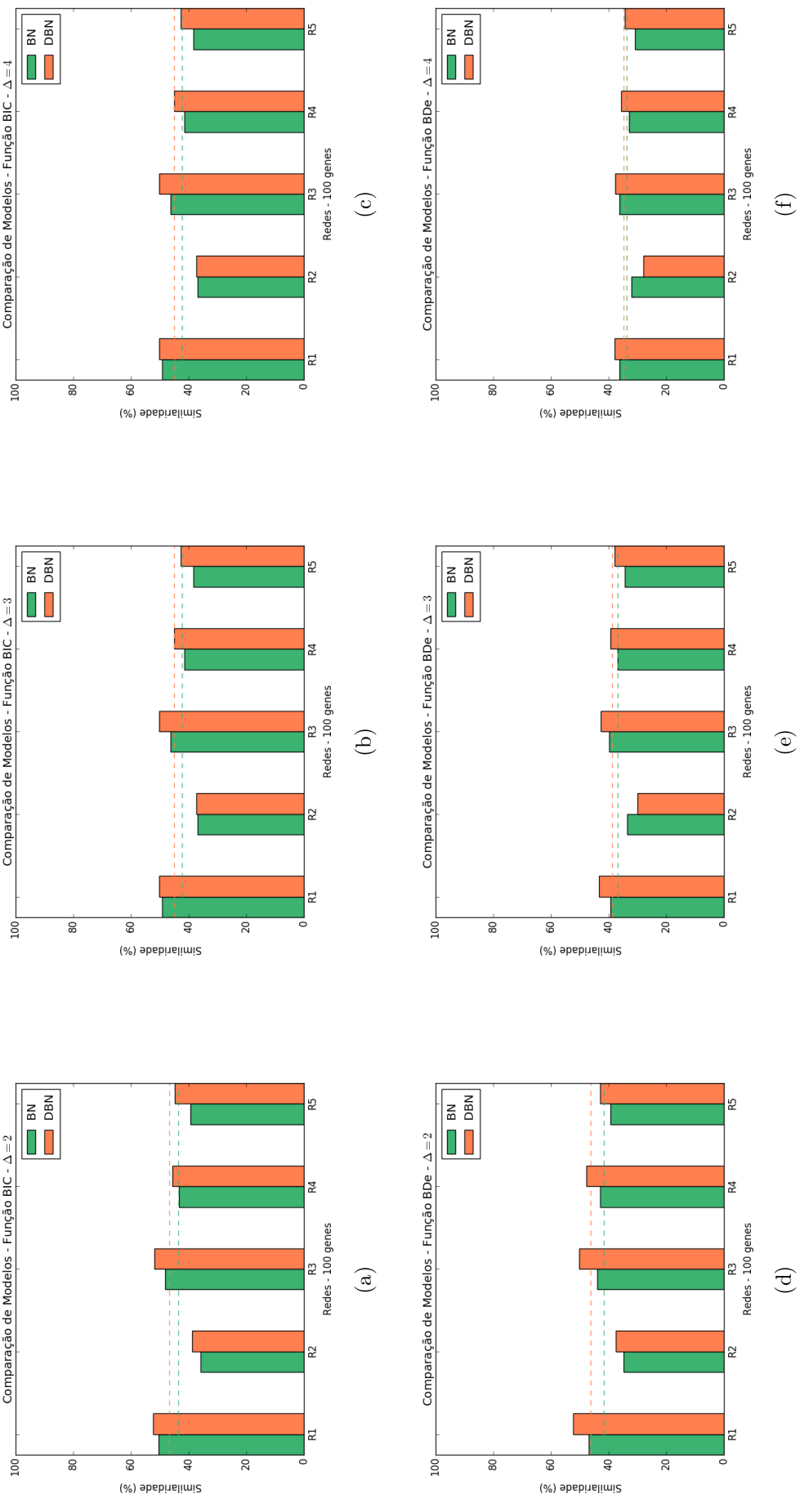
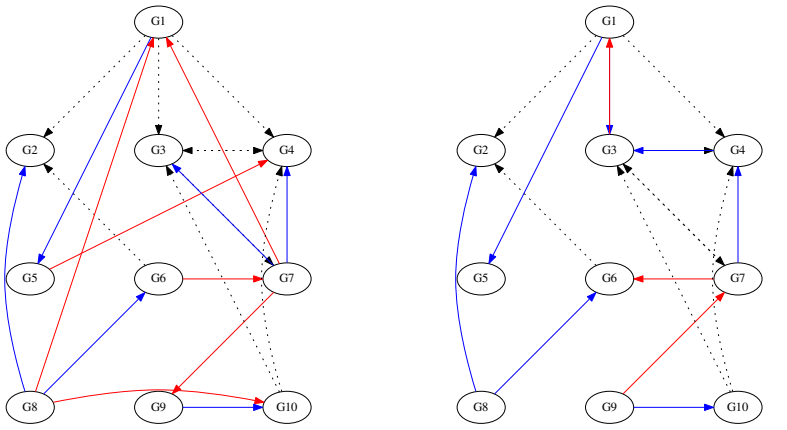


Figura 5.2: *Gráficos referentes aos resultados de similaridade obtidos no processo de inferência, utilizando os modelos de redes Bayesianas e redes Bayesianas dinâmicas para as redes de* 100 *genes (DREAM Challenge), considerando as funções de pontuação BIC e BDe e variando o parâmetro* Δ.



(a) Redes Bayesianas (b) Redes Bayesianas dinâmicas

Figura 5.3: *Grafos referentes à rede $R1$, de 10 genes, ambos considerando a função de pontuação BIC. Arestas azuis são verdadeiro positivas, vermelhas são falso positivas e pontilhadas são falso negativas.*

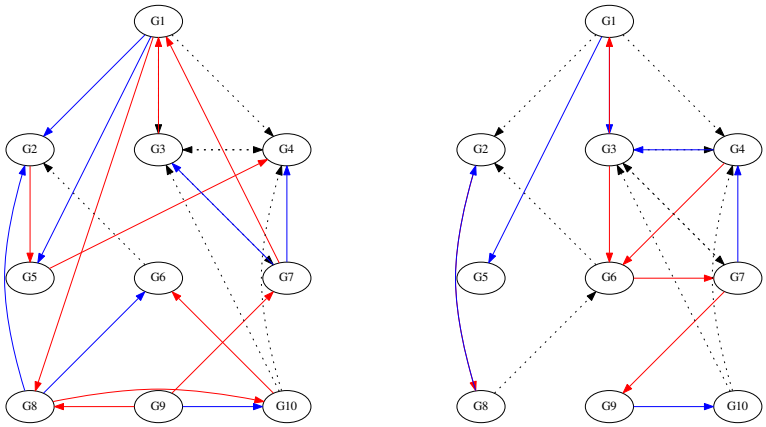
5.1.2 Função de Pontuação BDe

Os resultados de similaridade das redes obtidas no processo de engenharia reversa, utilizando a função de pontuação BDe e dados de expressão com 10 genes (redes $R1, \ldots, R5$), podem ser observados nas Figuras 5.2(d), 5.2(e) e 5.2(f). Diferente da função BIC, os resultados se mostraram mais sensíveis ao parâmetro Δ. Em redes pequenas, quando $\Delta = 2$, os resultados com maiores taxas de similaridade foram obtidos com o modelo de redes Bayesianas dinâmicas, com 60% de similaridade média, contra 55% de similaridade obtido pelo modelo de redes Bayesianas (Figura 5.1(d)). Já para os demais valores, $\Delta = 3$ e $\Delta = 4$, o modelo que apresentou melhor comportamento foi o de redes Bayesianas, com taxas de similaridade iguais a 52 e 53% respectivamente (Figuras 5.1(e) e 5.1(f)). O modelo estático pode ter apresentado melhor comportamento para estes casos devido à restrição de ser acíclico, limitando o número de arestas incidentes na rede mesmo quando o parâmetro Δ assume valores mais altos.

Em relação às redes com 100 genes (Figura 5.2), conforme o aumento do valor do parâmetro Δ, a similaridade das redes inferidas utilizando tanto o modelo de redes Bayesianas dinâmicas como o modelo de redes Bayesianas tende a decair (Figuras 5.2(d), 5.2(e) e 5.2(f)). Ainda levando em consideração a variação deste parâmetro, é possível observar que quanto maior o valor de Δ, mais próximas ficam as retas pontilhadas que informam os valores médios de similaridade, levando-nos a deduzir que conforme o grau da rede aumenta, os resultados se mostram mais favoráveis com o uso do modelo de redes Bayesianas estático.

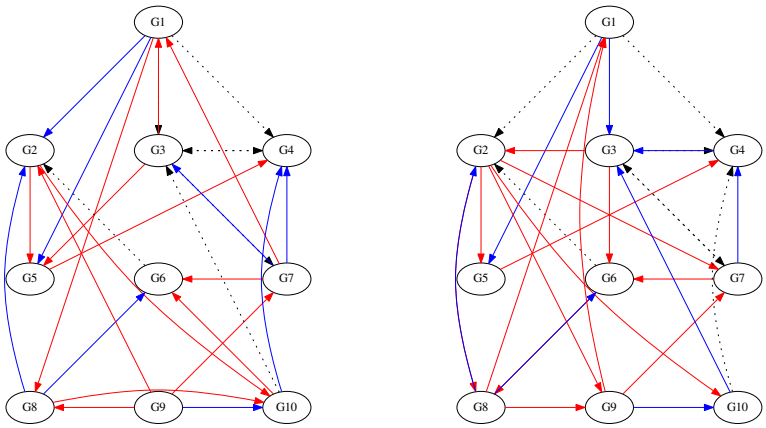
Nas Figuras 5.4, 5.5 e 5.6 podem ser vistas as redes inferidas com 10 genes, referentes à rede $R1$, utilizando a função de pontuação BDe. Conforme o valor do parâmetro Δ aumenta, o modelo de redes Bayesianas dinâmicas se mostra mais instável (Figuras 5.5 e 5.6). Ao

mesmo tempo em que o algoritmo acerta mais arestas, também há um número elevado de arestas falso positivas. Assim, para valores mais altos de Δ, as redes inferidas com o modelo de redes Bayesianas apresentaram taxas maiores de similaridade. Isso pode ocorrer devido à restrição do algoritmo de não permitir ciclos na rede, diminuindo assim o número de arestas que podem ser escolhidas no processo de engenharia reversa.



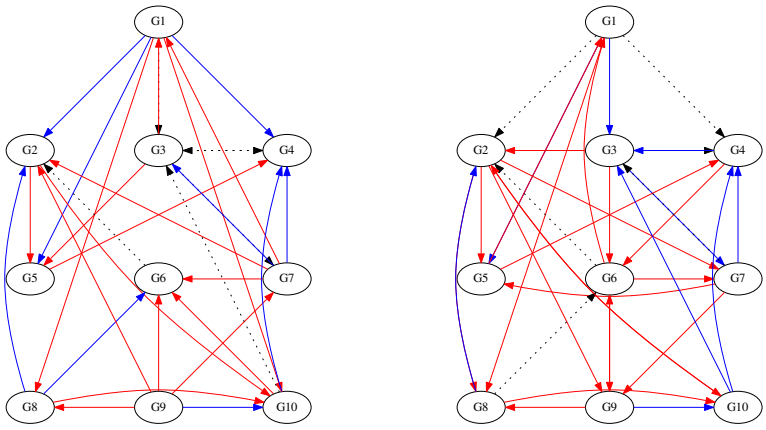
(a) Redes Bayesianas (b) Redes Bayesianas Dinâmicas

Figura 5.4: *Grafos referentes à rede R1, contendo 10 genes, ambos considerando a função de pontuação BDe e parâmetro $\Delta = 2$. Arestas azuis são verdadeiro positivas, vermelhas são falso positivas e pontilhadas são falso negativas.*



(a) Redes Bayesianas (b) Redes Bayesianas dinâmicas

Figura 5.5: *Grafos referentes à rede R1, contendo 10 genes, ambos considerando a função de pontuação BDe e parâmetro $\Delta = 3$. Arestas azuis são verdadeiro positivas, vermelhas são falso positivas e pontilhadas são falso negativas.*



(a) Redes Bayesianas (b) Redes Bayesianas dinâmicas

Figura 5.6: *Grafos referentes à rede $R1$, de 10 genes, ambos considerando a função de pontuação BDe e parâmetro $\Delta = 4$. Arestas azuis são verdadeiro positivas, vermelhas são falso positivas e pontilhadas são falso negativas.*

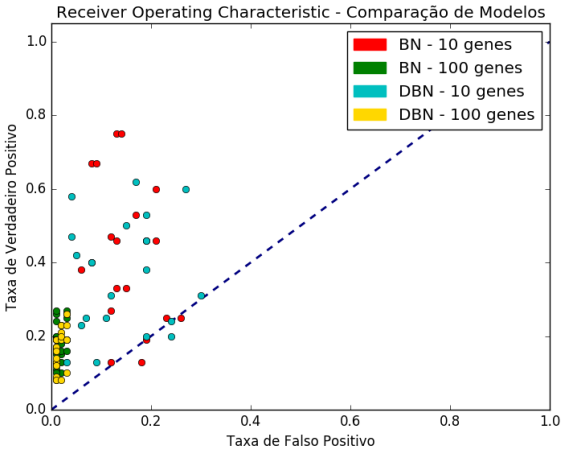
5.1.3 Discussão

A fim de comparar a qualidade das redes obtidas por meio dos modelos de redes Bayesianas e redes Bayesianas dinâmicas, levando em consideração a quantidade de arestas inferidas corretamente por ambos os modelos, foi utilizado o espaço ROC (*Receiver operating characteristic*). Trata-se de um espaço criado para representar as taxas verdadeiro positivas, no eixo y, em relação às taxas falso positivas, no eixo x. O ideal é que uma rede possua alta taxa de acertos ao mesmo tempo que apresenta baixo número de arestas inferidas de maneira incorreta. Os gráficos comparativos podem ser observados na Figura 5.7.

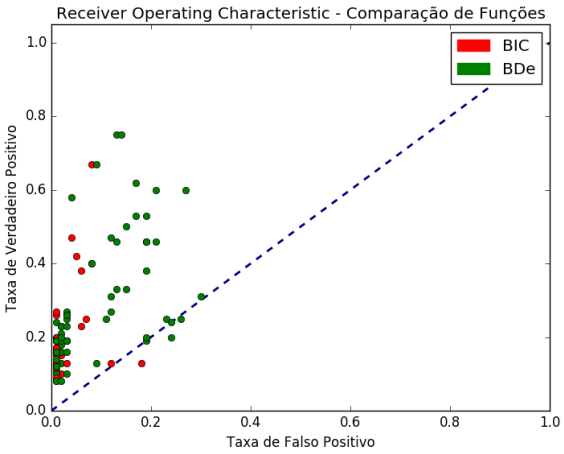
Conforme as Figuras 5.7(a) e 5.7(b), as redes de 10 genes apresentaram taxas de arestas verdadeiro positivas superiores às redes contendo 100 genes. O modelo que obteve redes com maiores taxas de arestas verdadeiro positivas foi o modelo de redes Bayesianas estático, utilizando a função de pontuação BDe. Por outro lado, redes inferidas com a função BIC apresentaram maior estabilidade por possuírem taxas menores de arestas falso positivas. Além disso, podemos observar na Figura 5.7(c) que redes inferidas utilizando o parâmetro $\Delta = 2$ obtiveram melhores resultados em relação à redes inferidas com os demais valores propostos, com baixas taxas de arestas falso positivas.

5.2 Dados do Ciclo Celular da Levedura

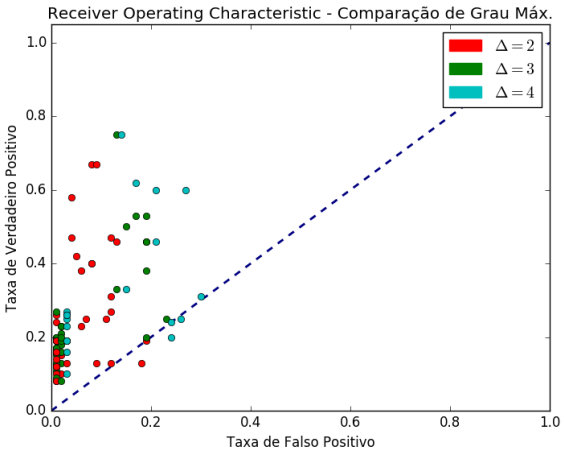
Considerando que os melhores resultados de similaridade de redes pequenas foram alcançados a partir do modelo de redes Bayesianas dinâmicas e o fato de que tais redes permitem ciclos, utilizamos este modelo no processo de engenharia reversa, contando agora com bancos de dados biológicos reais.



(a) Comparação de modelos.



(b) Comparação de funções de pontuação.



(c) Comparação de redes levando em conta o grau máximo.

Figura 5.7: *Espaço ROC, comparando as taxas de arestas verdadeiro positivas em relação as taxas de arestas falso positivas das redes inferidas utilizando os dados de expressão retirados de DREAM Challenge.*

Neste trabalho, foram utilizados dados de expressão da levedura *Saccharomyces cerevisiae*, como foi descrito na Seção 4.5.2, levando em consideração os 11 genes propostos por Li *et al.* (2004) que governam o comportamento do ciclo celular.

Foram analisadas cinco situações: Inferência de redes Bayesianas dinâmicas utilizando os dados de expressão gênica da série temporal alpha, utilizando os dados de expressão gênica da série temporal cdc15, utilizando a série temporal cdc28, a série temporal de elutriação e, por fim, utilizando o arquivo completo, que além de conter as amostras de todas as séries temporais citadas anteriormente, possui também as amostras pertencentes às séries temporais cln3 e clb2.

5.3 Adição de Conhecimento Biológico

Nesta etapa foi proposta a adição de conhecimento biológico ao algoritmo de inferência, como descrito na Seção 4.6.1. O intuito é obter melhorias nos resultados, elevando as taxas de similaridade obtidas a partir dos experimentos realizados. Assim, levando em consideração a Equação 4.12, foram atribuídos os seguintes valores para o parâmetro ω, que representa o peso atribuído ao conhecimento biológico: $0, 1, 2$ e 4, onde 0 indica a não adição de conhecimento biológico ao algoritmo.

Na Figura 5.8 é possível observar gráficos comparativos que mostram as porcentagens de similaridade das redes pertencentes a cada experimento (alpha, cdc15, cdc28, elutriação e completo). Ao utilizar a função de pontuação BIC, conforme o parâmetro ω é alterado, nota-se uma melhoria dos níveis de similaridade, principalmente considerando as redes obtidas a partir das séries temporais alpha e cdc28.

Ainda sobre as redes que utilizam a função de pontuação BIC, a melhoria geral dos resultados pode ser observada por meio das linhas de valores médios presentes nos gráficos (Figuras 5.8(a), 5.8(b) e 5.8(c)). Assim, quanto maior o parâmetro ω, maiores são as porcentagens médias de similaridade alcançadas. Além disso, os resultados são equivalentes para todos os valores de Δ propostos, onde o grau máximo de entrada da rede atingido é $\Delta = 2$.

Para as redes inferidas utilizando a função de pontuação BDe, Figuras 5.8(d), 5.8(e) e 5.8(f), também é possível notar a melhoria dos resultados conforme o aumento do parâmetro ω. Porém, quanto maior o parâmetro Δ, nota-se um declínio das taxas de similaridade. Assim, os melhores resultados são atingidos quando o grau máximo de entrada da rede é $\Delta = 2$.

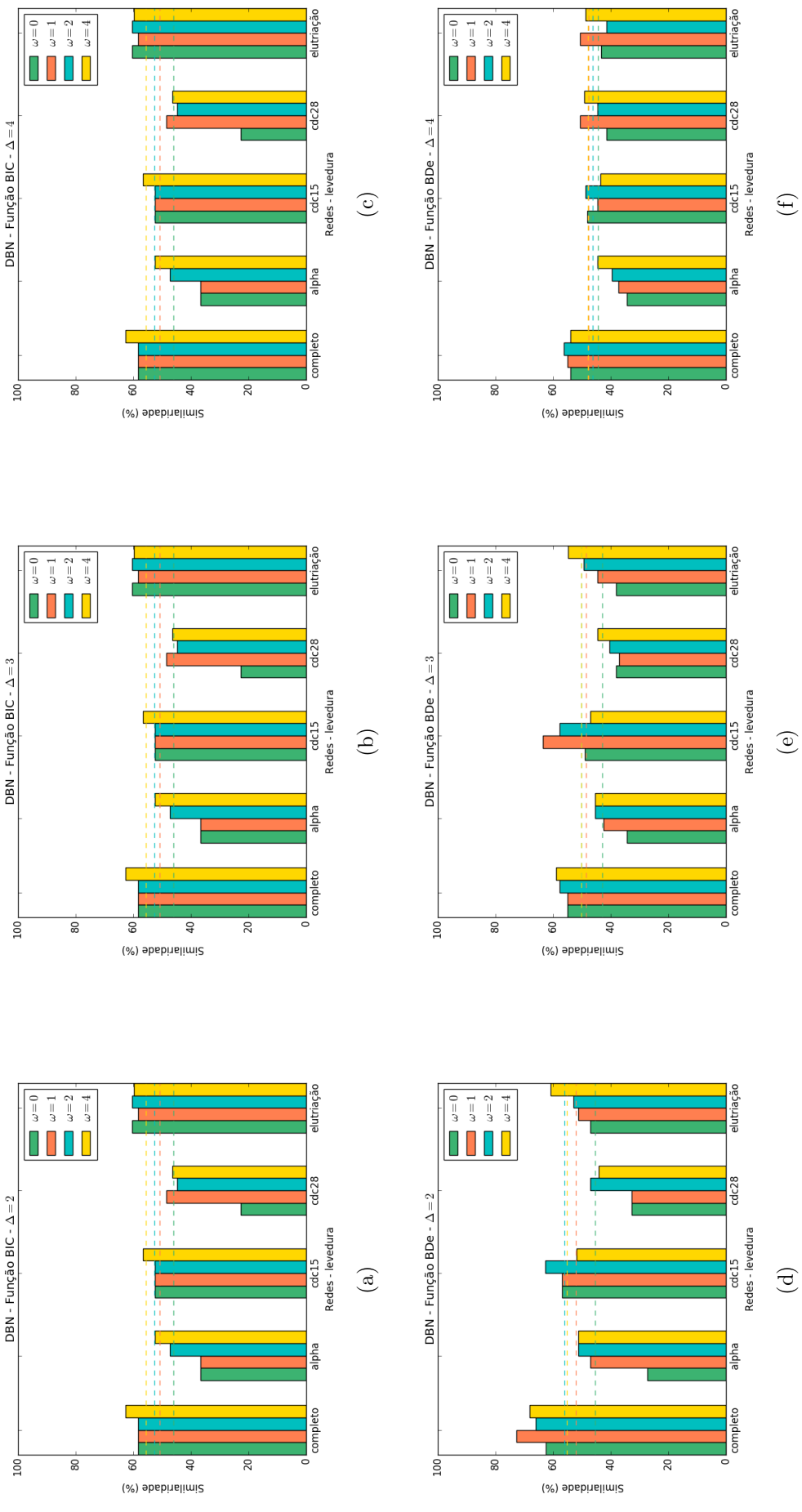
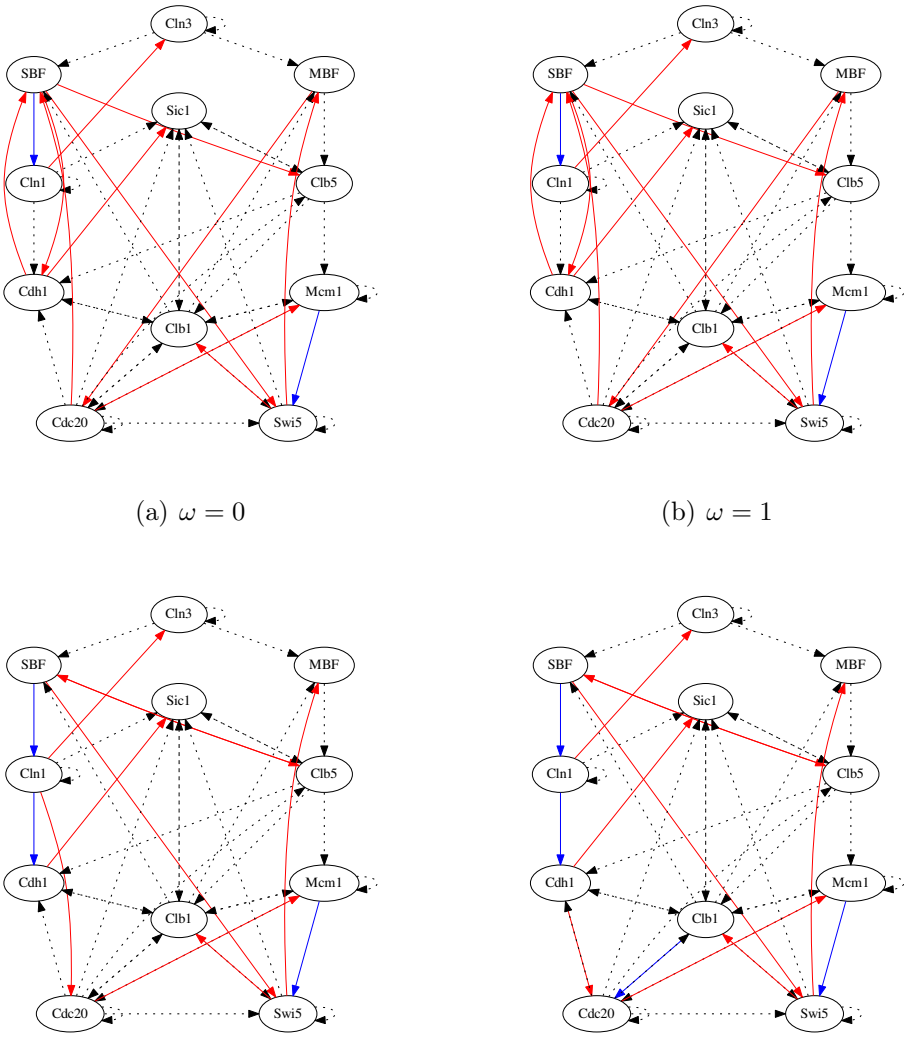


Figura 5.8: *Comparação dos níveis de similaridade das redes contendo os* 11 *genes responsáveis pelo ciclo celular da levedura, utilizando as funções de pontuação BIC e BDe para diferentes valores de* Δ*, levando em consideração a adição de conhecimento biológico ao algoritmo.*



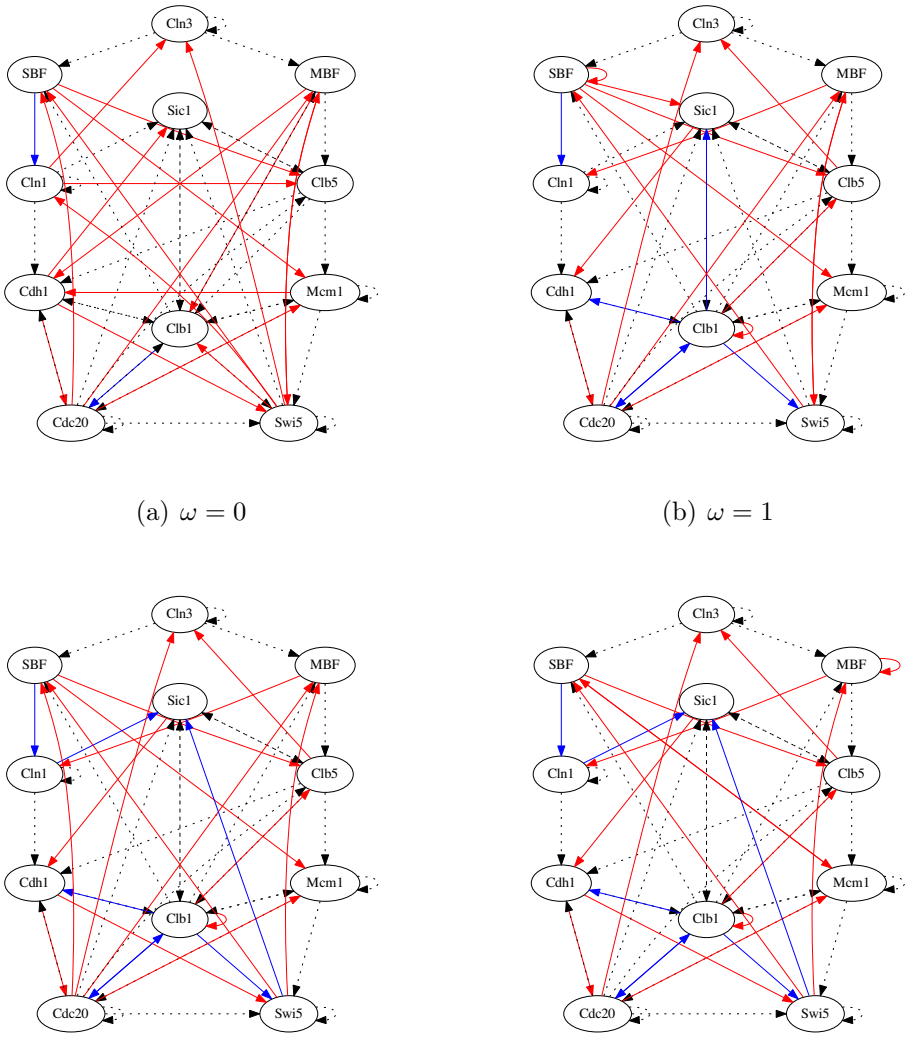
(a) $\omega = 0$ (b) $\omega = 1$

(c) $\omega = 2$ (d) $\omega = 4$

Figura 5.9: *Redes inferidas com base na série temporal* alpha*, utilizando o modelo de redes Bayesianas dinâmicas, com função de pontuação BIC,* $\Delta = 2$ *e adição de conhecimento biológico.*

Podemos observar nas Figuras 5.9 e 5.10 as redes inferidas utilizando os dados de expressão do experimento alpha, levando em conta o parâmetro $\Delta = 2$ e as variações do parâmetro ω propostas, onde a primeira figura utiliza a função de pontuação BIC e a segunda utiliza a função BDe.

Nas redes contidas na Figura 5.9, utilizando a função BIC, conforme a variação do parâmetro ω, é possível notar tanto a diminuição de arestas falso positivas (arestas vermelhas) como um aumento da taxa de arestas verdadeiro positivas (arestas azuis). Quanto às redes inferidas utilizando a função de pontuação BDe, é possível observar que há uma quantidade maior de acertos de arestas em relação a função BIC. Por outro lado, as redes apresentam número elevado de arestas falso positivas. Conforme valores mais altos são atribuídos ao parâmetro ω, o número de arestas inferidas incorretamente tende a diminuir (Figura 5.10).



(a) $\omega = 0$ (b) $\omega = 1$

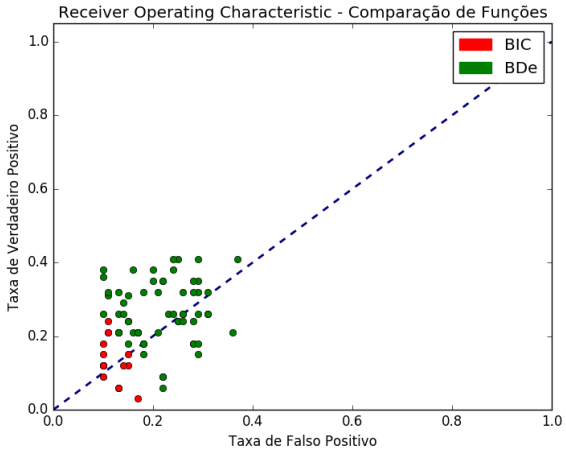
(c) $\omega = 2$ (d) $\omega = 4$

Figura 5.10: *Redes inferidas com base na série temporal* alpha*, utilizando o modelo de redes Bayesianas dinâmicas, com função de pontuação BDe e* $\Delta = 2$ *e adição de conhecimento biológico.*

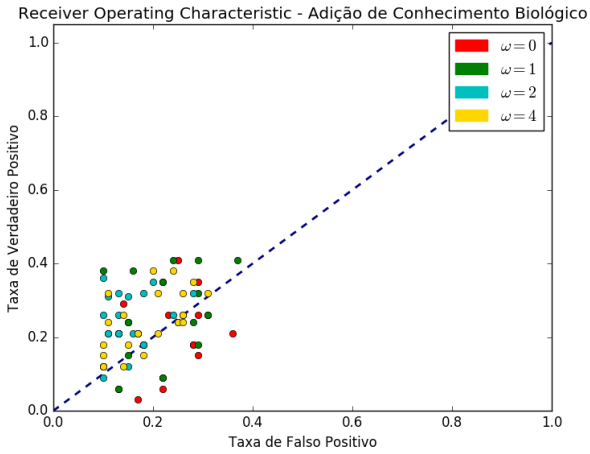
5.3.1 Discussão

A fim de comparar a qualidade das redes obtidas pelo processo de engenharia reversa, utilizando os 11 genes propostos por Li *et al.* (2004) que governam o comportamento do ciclo celular da levedura, foi utilizado o espaço ROC. Foram analisadas as taxas de arestas verdadeiro positivas em relação as taxas de arestas falso positivas obtidas de todas as redes inferidas, levando em conta a variação dos parâmetros Δ, ω e a função de pontuação utilizada. Os resultados destas análises podem ser verificados na Figura 5.11.

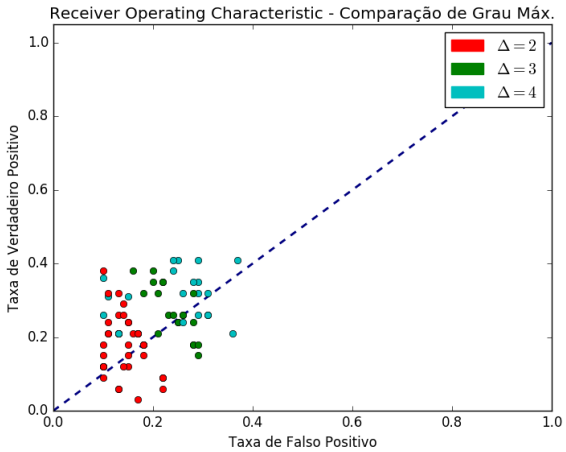
Conforme os resultados observados na Figura 5.11(a), podemos observar que embora redes inferidas utilizando a função de pontuação BDe tenham apresentado taxas maiores de arestas verdadeiro positivas, as redes inferidas utilizando a função BIC se mostraram mais estáveis por apresentarem menores taxas de arestas falso positivas.



(a) Comparação de funções de pontuação.



(b) Comparação de redes considerando o parâmetro ω.



(c) Comparação de redes considerando o parâmetro Δ.

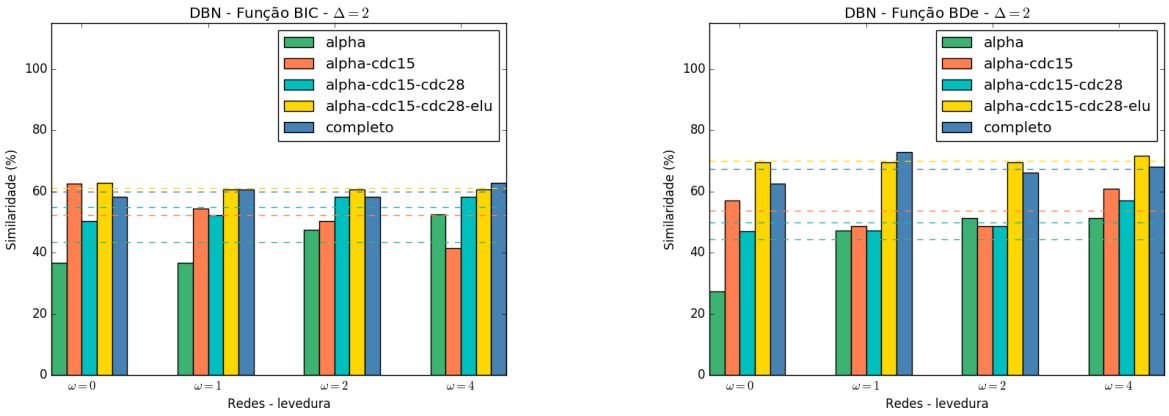
Figura 5.11: *Espaço ROC, comparando as taxas de verdadeiro positivo em relação as taxas de falso positivo das redes inferidas utilizando os dados de expressão gênica da levedura.*

Na Figura 5.11(b), podemos observar que quanto maior o parâmetro ω, que representa o valor que multiplica a equação de conhecimento biológico, melhores são os resultados. Ou

seja, as redes inferidas apresentam maiores taxas de arestas verdadeiro positivas e menores taxas de arestas falso positivas. Já na Figura 5.11(c), é possível observar que o grau máximo de entrada da rede que obteve melhor comportamento foi $\Delta = 2$.

5.4 Inferência de Redes Considerando a Quantidade de Amostras

Nesta etapa do trabalho, foi analisado como os resultados de similaridade das redes podem ser alterados em decorrência do número de amostras presentes nos dados de expressão gênica. Por meio dos dados da levedura *Saccharomyces cereviseae*, foram analisadas as seguintes situações: utilizar o arquivo com a série temporal alpha, adicionar a esse arquivo os dados da série temporal cdc15 e analisar os resultados. Após este procedimento, adicionar as amostras da série temporal cdc28. Posteriormente, adicionar as amostras de elutriação. E por último, analisar a rede inferida a partir do arquivo completo, adicionando as amostras de cln3 e clb2.



(a) (b)

Figura 5.12: *Comparação dos níveis de similaridade das redes utilizando dados de expressão gênica do ciclo celular da levedura, considerando a quantidade de amostras e a função de pontuação utilizada.*

No processo de inferência, foram utilizados o modelo de redes Bayesianas dinâmicas, considerando o grau máximo de entrada das redes $\Delta = 2$, e as funções de pontuação BIC e BDe. Os gráficos comparativos da Figura 5.12 nos mostram que conforme o número de amostras cresce, maiores taxas de similaridade são observadas nos resultados. Além disso, as redes que obtiveram maiores taxas de similaridade foram as que utilizaram parâmetro $\omega = 4$, alcançando taxa de similaridade média igual a 55,1% ao utilizar a função de pontuação BIC e 61,7% ao utilizar a função de pontuação BDe.

A rede inferida que apresentou melhores resultados foi a rede composta pelas séries temporais alpha, cdc15, cdc28 e elutriação, com similaridade média de 61,2% observada utilizando

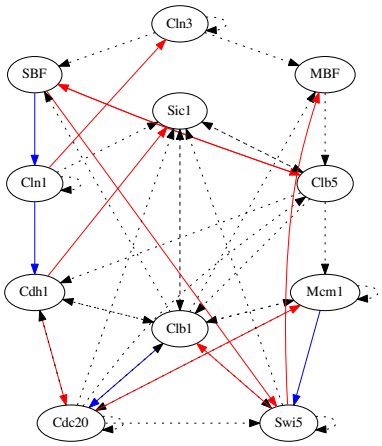
a função BIC e 70% utilizando a função de pontuação BDe.

As redes obtidas utilizando os parâmetros $\Delta = 2$, $\omega = 4$ e função de pontuação BIC podem ser vistas na Figura 5.13. A Rede 5.13(a) utiliza somente as amostras presentes na série temporal alpha no processo de inferência. A Rede 5.13(b) utiliza as amostras presentes na série temporal alpha em conjunto com a série cdc15. Ao adicionar a série cdc28 foi observada uma melhoria significativa nos resultados obtidos (Rede 5.13(c)). A Rede 5.13(d) adiciona a série temporal de elutriação e a Rede 5.13(e) adiciona as séries temporais cln3 e clb2. No caso geral, adicionar séries temporais no processo de inferência e utilizar conhecimento biológico faz com que as redes fiquem mais próximas do real.

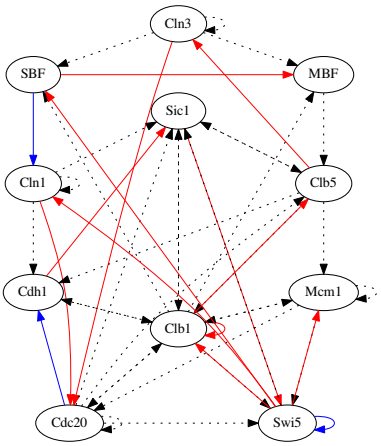
As redes contidas na Figura 5.14 foram inferidas por meio da função de pontuação BDe, também utilizando os parâmetros $\Delta = 2$ e $\omega = 4$. Conforme são adicionada séries temporais ao processo de inferência, maiores são as taxas de acerto obtidas. Além disso, também é observada uma queda no número de arestas falso positivas (arestas vermelhas) e verdadeiro negativas (arestas pontilhadas).

5.4.1 Discussão

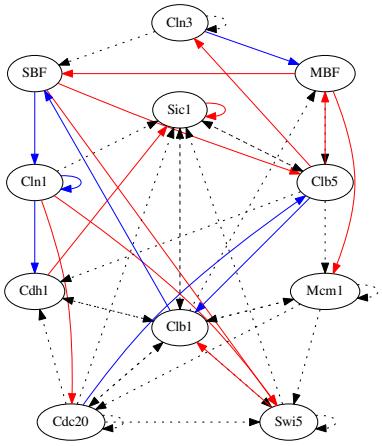
A fim de analisar a qualidade dos resultados obtidos e verificar qual função de pontuação obteve melhor comportamento, foi utilizado o espaço ROC. Nele são plotados as porcentagens de arestas verdadeiro positivas em relação as porcentagens de arestas falso positivas, obtidas no processo de engenharia reversa das redes da levedura, definida na Seção 5.4, considerando a quantidade de amostras. Conforme observado na Figura 5.15(a), a função de pontuação BDe apresentou redes com taxas mais altas de acertos em relação a função BIC. Além disso, analisando a Figura 5.15(b) é possível ver que conforme o parâmetro ω cresce, maiores taxas de acerto são alcançadas onde, os melhores resultados são obtidos quando $\omega = 4$.



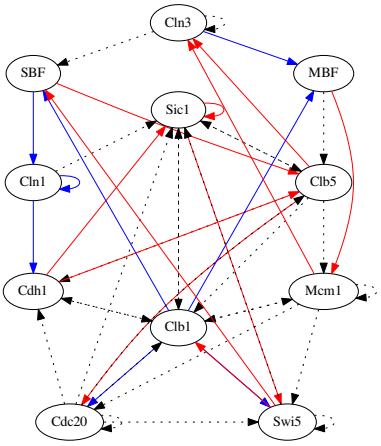
(a) Rede com série temporal alpha.



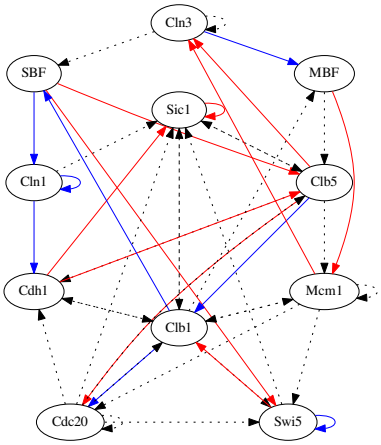
(b) Rede com séries temporais alpha e cdc15.



(c) Rede com séries temporais alpha, cdc15 e cdc28.

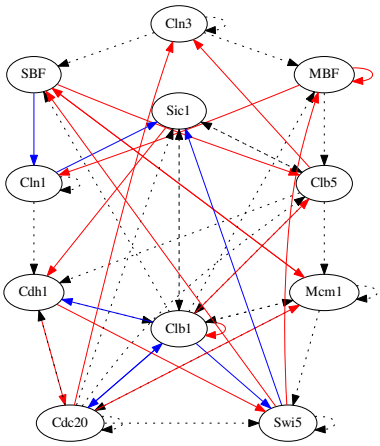


(d) Rede com séries temporais alpha, cdc15, cdc28 e elutriação.

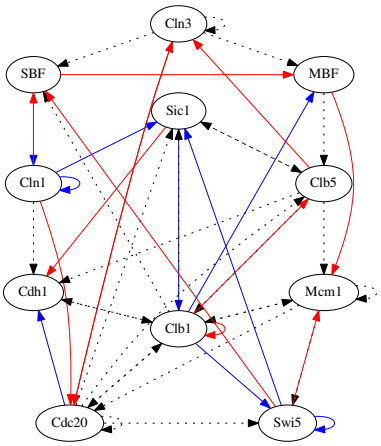


(e) Rede com todas as séries temporais.

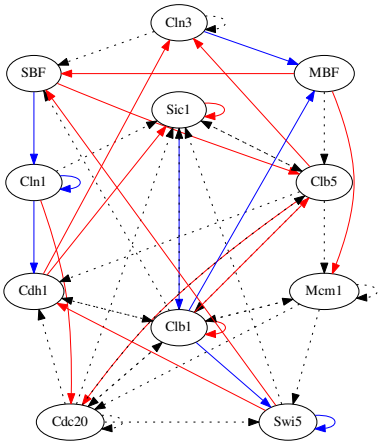
Figura 5.13: *Redes inferidas utilizando o modelo de redes Bayesianas dinâmicas, com função de pontuação BIC,* $\Delta = 2$ *e* $\omega = 4$.



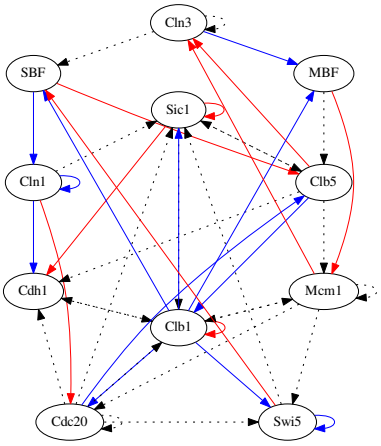
(a) Rede com série temporal alpha.



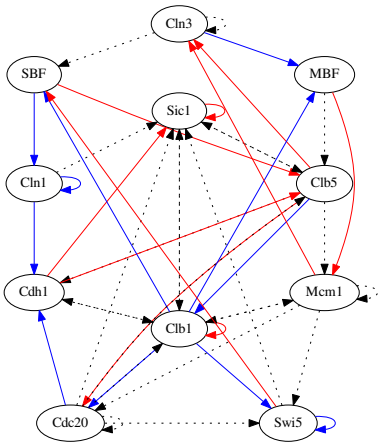
(b) Rede com séries temporais alpha e cdc15.



(c) Rede com séries temporais alpha, cdc15 e cdc28.

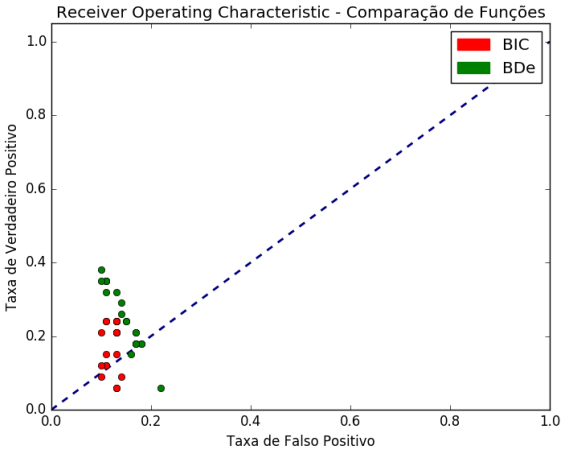


(d) Rede com séries temporais alpha, cdc15, cdc28 e elutriação.

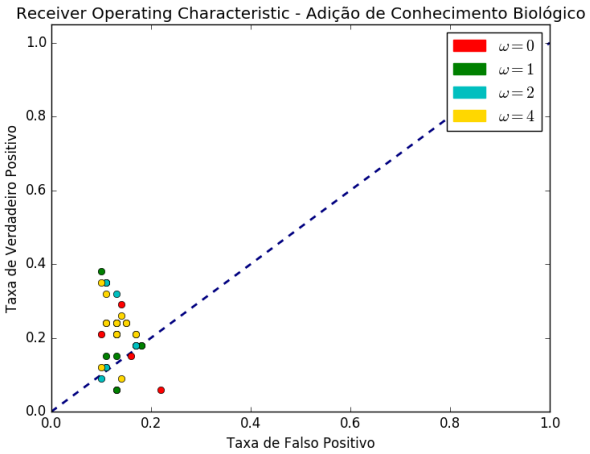


(e) Rede com todas as séries temporais.

Figura 5.14: *Redes inferidas utilizando o modelo de redes Bayesianas dinâmicas, com função de pontuação BDe,* $\Delta = 2$ *e* $\omega = 4$.



(a) Comparação de funções de pontuação.



(b) Comparação de redes considerando o parâmetro ω.

Figura 5.15: *Espaço ROC, comparando as taxas de verdadeiro positivo em relação as taxas de falso positivo de redes inferidas levando em consideração a quantidade de amostras.*

Capítulo 6

Conclusão

Neste trabalho, realizamos a engenharia reversa de GRNs por meio de dados de expressão gênica temporais. O objetivo foi analisar a similaridade e a taxa de acertos das redes obtidas, levando em consideração alguns aspectos como o modelo de inferência utilizado, a função responsável por pontuar estruturas de redes candidatas e a dimensão máxima permitida para a rede. Além disso, foi proposta a utilização de conhecimento biológico, de forma a colaborar no processo de escolha das arestas que compõe a rede. Adicionalmente, foram analisados os efeitos de se inferir redes com base na quantidade de amostras disponíveis.

Os modelos de inferência utilizados foram os modelos de redes Bayesianas e redes Bayesianas dinâmicas, sendo o primeiro modelo acíclico e o segundo cíclico. Foram analisados os comportamentos de redes contendo poucos genes e redes maiores, com 100 genes, quando inferidas utilizando as funções de pontuação BIC e BDe.

Quanto aos dados de expressão gênica sintéticos, retirados do *DREAM Challenge*, é possível concluir que as redes que apresentaram maiores taxas de similaridade foram as redes inferidas com o modelo de redes Bayesianas dinâmicas, utilizando o parâmetro $\Delta = 2$, que delimita o grau máximo de entrada da rede. Além disso, a função de pontuação BIC apresentou maiores taxas de similaridade e baixas taxas de arestas falso positivas em relação às redes inferidas com a função de pontuação BDe. As redes inferidas utilizando a função de pontuação BDe demonstraram maior instabilidade, pois ao mesmo tempo em que apresentaram altas taxas de acerto, também apresentaram elevadas taxas de arestas falso positivas.

Em relação à adição de conhecimento biológico ao algoritmo de redes Bayesianas dinâmicas, utilizando os dados de expressão gênica da levedura *Saccharomyces cereviseae*, podemos concluir que o procedimento beneficiou as taxas de similaridade obtidas onde, os melhores resultados foram observados quando o parâmetro ω atingiu valor $\omega = 4$. A função de pontuação que apresentou melhores resultados levando em conta este aspecto foi a função BDe, assim, quanto maior o valor de ω testado, maiores foram as taxas de acerto e de similaridade observadas.

Quanto a engenharia reversa de redes de regulação gênica, levando em conta a quantidade

de amostras e os dados biológicos da levedura, foi possível observar a melhoria dos resultados conforme séries temporais foram adicionadas ao processo e conforme o conhecimento biológico foi atribuído às funções que pontuam as estruturas candidatas. Novamente, a função que apresentou resultados mais satisfatórios considerando estes aspectos foi a função BDe.

Os algoritmos, resultados e redes obtidas neste trabalho podem ser acessados por meio do endereço eletrônico https://github.com/marianacaravanti/pgmpy-modified.git.

6.1 Trabalhos Futuros

Ao realizarmos o processo de engenharia reversa de redes utilizando o modelo de redes Bayesianas dinâmicas, adicionando conhecimento biológico ao algoritmo, houve um aumento considerável no tempo de processamento, não sendo possível testar a metodologia em redes de regulação gênica com grande quantidade de genes. Ao tentarmos processar arquivos contendo dados de expressão gênica de 125 genes e uma série temporal contendo cerca de 15 a 20 amostras, após dois meses de processamento, nenhum resultado foi alcançado. Um trabalho futuro a ser considerado seria realizar a paralelização do algoritmo de inferência proposto, considerando que o cálculo do modelo de probabilidade pode ser facilmente decomposto em termos independentes. Outra forma de minimizar o tempo de execução seria implementar a biblioteca Pgmpy com as alterações realizadas neste trabalho, utilizando uma linguagem de programação compilada, como C ou C++.

Quanto ao conhecimento biológico *a priori* adicionado ao algoritmo de inferência de GRNs, sugerimos que outras fontes de dados sejam utilizadas, que demonstrem o grau de evidência de associação entre genes de um organismo. Além disso, nossa metologia pode ser aplicada a dados de expressão gênica associados a problemas reais.

Referências Bibliográficas

Alberts *et al.***(2009)** Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts e Peter Walter. *Biologia molecular da célula*. Artmed Editora. Citado na pág. 3, 4, 6

Barabási(2009) Albert-László Barabási. Scale-free networks: a decade and beyond. *Science*, 325(5939):412–413. Citado na pág. 38

Brown *et al.***(2015)** Garth R Brown, Vichet Hem, Kenneth S Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D Pruitt, Donna R Maglott *et al.* Gene: a gene-centered information resource at NCBI. *Nucleic acids research*, 43 (D1):D36–D42. Citado na pág. 37

Bryne *et al.***(2008)** Jan Christian Bryne, Eivind Valen, Man-Hung Eric Tang, Troels Marstrand, Ole Winther, Isabelle da Piedade, Anders Krogh, Boris Lenhard e Albin Sandelin. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research*, 36(suppl 1):D102–D106. Citado na pág. 2

Buntine(1991) Wray Buntine. Theory refinement on Bayesian networks. Em *Uncertainty Proceedings 1991*, páginas 52–60. Elsevier. Citado na pág. 32

Chatr-aryamontri *et al.***(2017)** Andrew Chatr-aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam *et al.* The BioGRID interaction database: 2017 update. *Nucleic acids research*, 45(D1):D369–D379. Citado na pág. 37

Chen *et al.***(2013)** Haifen Chen, DAK Maduranga, Piyushkumar A Mundra e Jie Zheng. Integrating epigenetic prior in dynamic Bayesian network for gene regulatory network inference. Em *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2013 IEEE Symposium on*, páginas 76–82. IEEE. Citado na pág. 25

Claudia Moraes(2016) Ricardo Campos Claudia Moraes. *Ciências da natureza e suas tecnologias: Biologia*. CEJA (CECIERJ). Citado na pág. 4, 5

Csermely *et al.***(2013)** Peter Csermely, Tamás Korcsmáros, Huba JM Kiss, Gábor London e Ruth Nussinov. Structure and dynamics of molecular networks: a novel paradigm of

drug discovery: a comprehensive review. *Pharmacology & therapeutics*, 138(3):333–408. Citado na pág. 39

De Jong(2002) Hidde De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103. Citado na pág. 1

Dougherty(2007) Edward R Dougherty. Validation of inference procedures for gene regulatory networks. *Current genomics*, 8(6):351–359. Citado na pág. 39

Friedman *et al.***(1998)** Nir Friedman, Kevin Murphy e Stuart Russell. Learning the structure of dynamic probabilistic networks. Em *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, páginas 139–147. Morgan Kaufmann Publishers Inc. Citado na pág. 1, 15, 16, 25

Friedman *et al.***(2000)** Nir Friedman, Michal Linial, Iftach Nachman e Dana Pe'er. Using Bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4): 601–620. Citado na pág. 1, 6, 8, 17, 24

Gendelman *et al.***(2017)** Rina Gendelman, Heming Xing, Olga K Mirzoeva, Preeti Sarde, Christina Curtis, Heidi Feiler, Paul McDonagh, Joe W Gray, Iya Khalil e W Michael Korn. Bayesian network inference modeling identifies TRIB1 as a novel regulator of cell cycle progression and survival in cancer cells. *Cancer Research*, páginas canres–0512. Citado na pág. 1, 24

Goodwin *et al.***(1963)** Brian C Goodwin *et al.* Temporal organization in cells. A dynamic theory of cellular control processes. *Temporal organization in cells. A dynamic theory of cellular control processes*. Citado na pág. 1

Greenfield *et al.***(2010)** Alex Greenfield, Aviv Madar, Harry Ostrer e Richard Bonneau. DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PloS one*, 5(10):e13397. Citado na pág. 2

Hastings(1970) W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109. Citado na pág. 25

Hecker *et al.***(2009)** Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene Van Someren e Reinhard Guthke. Gene regulatory network inference: data integration in dynamic models - a review. *Biosystems*, 96(1):86–103. Citado na pág. 1, 2, 8

Heckerman *et al.***(1995)** David Heckerman, Dan Geiger e David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243. Citado na pág. 32

Kanehisa *et al.***(2017)** Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato e Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361. Citado na pág. 2, 37

Kauffman(1969) Stuart A Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, 22(3):437–467. Citado na pág. 1, 6

Khanin e Wit(2006) Raya Khanin e Ernst Wit. How scale-free are biological networks. *Journal of computational biology*, 13(3):810–818. Citado na pág. 38

Khurana *et al.***(2013)** Ekta Khurana, Yao Fu, Jieming Chen e Mark Gerstein. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol*, 9(3):e1002886. Citado na pág. 39

Koller e Friedman(2009) Daphne Koller e Nir Friedman. *Probabilistic Graphical Models: principles and techniques*. MIT press. Citado na pág. 9, 10, 12, 13, 15, 27, 30, 31, 33

Li *et al.***(2004)** F. Li, T. Long, Y. Lu, Q. Ouyang e C. Thang. The yeast cell-cycle network is robustly designed. *PNAS of the United States of America*, 101(14):4781–4786. Citado na pág. 36, 49, 52

Li *et al.***(2010)** Yong Li, Lili Liu, Xi Bai, Hua Cai, Wei Ji, Dianjing Guo e Yanming Zhu. Comparative study of discretization methods of microarray data for inferring transcriptional regulatory networks. *BMC bioinformatics*, 11(1):520. Citado na pág. 41

Liu *et al.***(2016)** Fei Liu, Shao-Wu Zhang, Wei-Feng Guo, Ze-Gang Wei e Luonan Chen. Inference of gene regulatory network based on local Bayesian networks. *PLoS Comput Biol*, 12(8):e1005024. Citado na pág. 24

Lopes *et al.***(2011)** Fabricio M Lopes, Roberto M Cesar Jr e Luciano Da F Costa. Gene expression complex networks: synthesis, identification, and analysis. *Journal of Computational Biology*, 18(10):1353–1367. Citado na pág. 38

Marbach *et al.***(2010)** Daniel Marbach, Robert J Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano e Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the national academy of sciences*, 107(14):6286–6291. Citado na pág. 35

Martin e Wang(2011) Jeffrey A Martin e Zhong Wang. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10):671–682. Citado na pág. 5, 7

Needham *et al.***(2007)** Chris J Needham, James R Bradford, Andrew J Bulpitt e David R Westhead. A primer on learning in Bayesian networks for computational biology. *PLoS Comput Biol*, 3(8):e129. Citado na pág. 24

Shalon *et al.***(1996)** Dari Shalon, Stephen J Smith e Patrick O Brown. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*, 6(7):639–645. Citado na pág. 1, 5

Spellman *et al.***(1998)** Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein e Bruce Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297. Citado na pág. 2, 24, 36

Szklarczyk *et al.***(2016)** Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, página gkw937. Citado na pág. 2, 37

Wang *et al.***(2009)** Zhong Wang, Mark Gerstein e Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63. Citado na pág. 1

Werhli *et al.***(2007)** Adriano V Werhli, Dirk Husmeier *et al.* Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat Appl Genet Mol Biol*, 6(1):15. Citado na pág. 15

Zhang *et al.***(2016)** Bin Zhang, Linh Tran, Valur Emilsson e Jun Zhu. Characterization of genetic networks associated with Alzheimer's disease. *Systems Biology of Alzheimer's Disease*, páginas 459–477. Citado na pág. 1, 24