Universidade Federal do ABC
Centro de Matemática, Computação e Cognição (CMCC)
Pós-Graduação em Ciência da Computação

Carlos Fernando Montoya Cubas

AGRUPAMENTO DE INSTÂNCIAS EM CLASSES DE EQUIVALÊNCIA PARA LIDAR COM O PROBLEMA DA DIMENSIONALIDADE EM INFERÊNCIA DE REDES GÊNICAS

Tese de Doutorado

Santo André - SP

Dezembro de 2020

Carlos Fernando Montoya Cubas

AGRUPAMENTO DE INSTÂNCIAS EM CLASSES DE EQUIVALÊNCIA PARA LIDAR COM O PROBLEMA DA DIMENSIONALIDADE EM INFERÊNCIA DE REDES GÊNICAS

Tese

Tese apresentada ao Curso de Pós-Graduação da Universidade Federal do ABC como requisito parcial para obtenção do grau de Doutorado em Ciência da Computação

David Correa Martins Junior

Santo André - SP

Dezembro de 2020

Sistema de Bibliotecas da Universidade Federal do ABC
Elaborada pelo Sistema de Geração de Ficha Catalográfica da UFABC
com os dados fornecidos pelo(a) autor(a).

Montoya Cubas, Carlos Fernando
Agrupamento de instâncias em classes de equivalência para lidar com o problema da dimensionalidade em inferência de redes gênicas / Carlos Fernando Montoya Cubas. — 2020.

129 fls. : il.

Orientador: David Correa Martins Junior

Tese (Doutorado) — Universidade Federal do ABC, Programa de Pós Graduação em Ciência da Computação, Santo André, 2020.

1. redes de regulação gênica. 2. redes booleanas. 3. inferência de redes. 4. problema da dimensionalidade. 5. reticulados booleanos. I. Martins Junior, David Correa. II. Programa de Pós-Graduação em Ciência da Computação, 2020. III. Título.

Este exemplar foi revisado e alterado em relação à versão original, de acordo com as observações levantadas pela banca examinadora no dia da defesa, sob responsabilidade única do(a) autor(a) e com a anuência do(a) (co)orientador(a).

Santo André , 14 de dezembro de 2020 .

Carlos Fernando Montoya Cobas

Nome completo e Assinatura do(a) autor(a)

Daniel Correa Martins Jr.

Nome completo e Assinatura do(a) (co)orientador(a)

FOLHA DE ASSINATURAS

Assinaturas dos membros da Banca Examinadora que avaliou e aprovou a Defesa de Tese de Doutorado do candidato CARLOS FERNANDO MONTOYA CUBAS, realizada em 7 de Dezembro de 2020:

Dr. DAVID CORREA MARTINS JUNIOR, UFABC

Presidente - Interno ao Programa

Dr. JUNIOR BARRERA, USP

Membro Titular - Examinador(a) Externo à Instituição

Dr. ULISSES DE MENDONÇA BRAGA NETO, Texas A&M

Membro Titular - Examinador(a) Externo à Instituição

Dr. RONALDO FUMIO HASHIMOTO, USP

Membro Titular - Examinador(a) Externo à Instituição

Dr. CARLOS DA SILVA DOS SANTOS, UFABC

Membro Titular - Examinador(a) Externo ao Programa

Dr. LUIZ CARLOS DA SILVA ROZANTE, UFABC

Membro Suplente - Examinador(a) Interno ao Programa

”O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001”

Aos meus queridos pais.

Agradecimentos

Eu nunca teria chegado até aqui se não fosse por meu pai, embora eu tenha sido um filho problemático, ele sempre soube guiar o meu caminho na base da sua confiança e o seu carinho. Obrigado pai, este trabalho é por você e para você! Gostaria também de agradecer a minha mãe, meus irmãos e meus gatos por serem um suporte para mim e por aguardarem todo ano meu retorno à casa, obrigado por tudo família, este trabalho também é seu trabalho.

Gostaria também de agradecer ao professor David Correa Martins Jr, pela sua amizade, dedicação e seu suporte para realizar este trabalho, muito obrigado professor, você sempre será um grande amigo e um exemplo a seguir para mim. Do mesmo modo quero agradecer também à galera do laboratório, ao Leonardo, ao Rodrigo e ao Carlos, por suas dicas e por me ajudar nas ideias e discussões sobre este trabalho.

Também agradeço à Universidade Federal do ABC e a todos os seus professores, que me ajudaram a obter todos os conhecimentos necessários para a realização do presente trabalho. Agradeço também ao governo do Brasil que, por meio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, financiaram este trabalho.

Gostaria também de agradecer a todos os meus amigos de Cusco que me acompanharam nesta aventura longe de casa, à Yanina, ao Christian, ao Ray, ao Juanito Gil, e ao Carlos, obrigado pessoal! Também agradeço aos meus novos amigos que fiz no Brasil, aos meninos da república Rep Zeppelin, aos meus amigos brasileiros, à galera peruana, e aos meus outros amigos de outras nacionalidades que conheci no Brasil, e com quem compartilhei muitos bons momentos, obrigado caras! Eu não posso nomear a todos aqui porque são muitos mesmo, vocês trouxeram um pedaço do meu lar ao Brasil, mesmo estando longe. Sempre levarei comigo muitas boas lembranças de vocês.

Por último, e não menos importante, gostaria de dedicar este trabalho também à galera dos *cuchiullas* e os *cumas* de Cusco, meus amigos que a cada ano esperaram pelo meu retorno, e sempre torceram por mim para que tudo dê certo. Obrigado *cuchiullas* e *cumas*, eu sempre levei e levo vocês no meu coração!

"Educação não transforma o mundo.
Educação muda as pessoas.
Pessoas transformam o mundo."
(Paulo Freire)

Resumo

A inferência de redes de interação gênica a partir de perfis de expressão é um dos problemas importantes pesquisados em biologia sistêmica, sendo considerado um problema em aberto. Diversas técnicas matemáticas, estatísticas e computacionais têm sido desenvolvidas para modelar, inferir e simular mecanismos de regulação gênica, sendo o problema de inferência o foco desta proposta. Tal proposta tem por objetivo continuar as pesquisas realizadas no mestrado, as quais envolveram o estudo de métodos de inferência de redes gênicas baseados em seleção de características (seleção do melhor conjunto de genes preditores do comportamento de um dado gene alvo em termos de suas expressões temporais de mRNA), propondo alternativas para aumentar o poder de estimação estatística em situações típicas nas quais o conjunto de amostras com perfis de expressão gênica é bem limitado e possuem elevada dimensionalidade (número de genes). Mais concretamente, no mestrado foram propostos métodos para aliviar o problema da dimensionalidade na inferência de redes Booleanas, através de partições no reticulado Booleano induzidas por combinações lineares dos valores dos genes preditores (instâncias dos preditores). Cada valor de combinação linear determina uma classe de equivalência entre as instâncias dos genes preditores. Neste trabalho de doutorado, o problema de agrupamento de instâncias foi reformulado como um problema de busca no reticulado de partições, além de formular estratégias de busca nesse reticulado com base em informações a priori (por exemplo: que uma rede gênica tende a ser composta majoritariamente por funções lineares e de canalização) para examinar um subespaço de partições potencialmente relevantes sem abrir mão da eficiência computacional. Adicionalmente desenvolvemos um método de transferência de aprendizado supervisionado obtido da inferência de redes geradas aleatoriamente (sintéticas) que busca estimar as dimensões corretas (graus) dos conjuntos de genes preditores para os respectivos genes alvos. Resultados experimentais através de dados simulados e dados reais de *microarray* do *Plasmodium falciparum*, um agente causador da malária, indicam que os métodos desenvolvidos, especialmente o método que busca por funções de canalização, obtêm redes competitivas tanto do ponto de vista topológico, como do ponto de vista da dinâmica da expressão gênica gerada pelas redes inferidas. A principal vantagem desses métodos de agrupamento é a superior capacidade de generalização para gerar o próximo estado do sistema com base em estados iniciais sorteados e que não estejam no conjunto de amostras de treinamento. Além disso, a adoção da estratégia de transferência de aprendizado dos graus se mostrou efetiva, conferindo uma vantagem a todos os métodos de inferência de redes gênicas considerados, incluindo o método original sem agrupamento de instâncias.

Palavras-chave: redes de regulação gênica, redes booleanas, inferência de redes, problema da dimensionalidade, reticulados booleanos, seleção de características.

Abstract

The inference of gene interaction networks from expression profiles is one of the relevant problems in systems biology, being considered an open problem. Several mathematical, statistical and computational techniques have been developed to model, infer and simulate gene regulation mechanisms, whereas the inference problem is the focus of this work. Our proposal is a continuation of the research conducted during the masters, which involved the study of gene networks inference based on feature selection (selection of the best subset of genes for predicting the behavior of a given target in terms of their temporal mRNA expressions), proposing alternatives to increase the statistical estimation power in typical situations where the set of samples with gene expression profiles is very limited and presents high dimensionality (number of genes). More concretely, during the masters we proposed methods to alleviate the curse of dimensionality in Boolean Networks inference, through Boolean lattice partitions induced by a linear combination of the predictor genes values (predictor instances). Each linear combination value determines an equivalence class between the predictor instances. In this work, the problem of instances grouping was reformulated as a partition lattice search problem, besides idealizing search strategies in this lattice based on prior information (eg. gene networks tend to be mostly composed by linear and canalizing functions) to examine a partition subspace potentially relevant without forgetting computational efficiency. In addition, we developed a method which transfers the supervised learning achieved from randomly generated (synthetic) networks inference aiming to estimate the correct dimension (degree) of the predictor gene sets for the corresponding target genes. Experimental results through simulated data and real *microarray* data from *Plasmodium falciparum*, a malaria agent, indicate that the developed methods, especially the method which searches for canalizing functions, achieves competitive networks considering both topology and gene expression dynamics generated by the inferred networks. The main advantage of these methods is the superior capacity of generalization to predict the next system state based on randomly chosen initial states which are not in the training set. Besides, the adoption of the strategy for transfer learning of the degrees sounds effective, benefitting all gene network inference methods considered, even the original method which does not group instances.

Keywords: gene regulatory networks, Boolean networks, network inference, dimensionality problem, boolean lattice, feature selection.

Sumário

Lista de Abreviaturas e Siglas

BA	Redes livres de escala propostas por Barabási e Albert
BEI	Busca Exaustiva Incremental.
BEIF	Busca Exaustiva Incremental com Grau Fixo.
BN	Rede Booleana (*Boolean Network*).
CG	Agrupamento por canalização (*canalizing grouping*).
DNA	Ácido desoxirribonucleico (*Deoxyribonucleic acid*).
EQM	Erro Quadrático Médio.
ER	Redes aleatórias de Erdös-Rényi.
EX-SFS	Busca Hibrida Exaustiva Sequencial para Frente.
FN	Falso Negativo (*False Negative*).
FP	Falso Positivo (*False Positive*).
GLSFS	Agrupamento por busca SFS no reticulado de partições.
GM	Grau Médio de preditores.
GRN	Rede de Regulação Gênica (*Gene Regulatory Network*).
IDC	Ciclo de desenvolvimento intraeritrocítico do Plasmodium falciparum.
IM	Informação Mútua.
LG	Agrupamento linear (*linear grouping*).
PBN	Rede Booleana Probabilística (*Probabilistic Boolean Network*).
PGN	Rede Gênica Probabilística (*Probabilistic Gene Network*)
RNA	Ácido Ribonucleico (*Ribonucleic Acid*).
RNA-Seq	Sequenciamento de RNA (*RNA sequencing*).
RNBio	Rede Nacional de Bioinformática.
SA	Sem agrupamento.
SAGE	Análise Serial de Expressão Gênica (*Serial Analysis of Gene Expression*).
SBC	Sociedade Brasileira de Computação.
SFS	Busca Sequencial para Frente (*Sequential Forward Search*).
TN	Verdadeiro Negativo (*True Negative*).
TP	Verdadeiro Positivo (*True Positive*).
TPC	Tabela de probabilidades condicionais.
k-nn	k vizinhos mais próximos.
mRNA	RNA mensageiro (*messenger RNA*).

Lista de Símbolos

V	Conjunto de nós (genes) de uma rede Booleana
n	Número de nós (genes) de uma rede
g_i	i-ésimo gene
i, j	Índices
Φ	Conjunto de funções Booleanas
ϕ_i	Função Booleana preditora do gene i
ψ	Classificador
c	Número de classes (rótulos)
v_i	i-ésimo nó (gene) de uma rede booleana
t	instante de tempo
k	Número de genes preditores (grau) de um determinado gene alvo
k_i	Número de genes preditores (grau) de i-ésimo gene
k^-	Número de genes preditores inibidores de um determinado gene alvo
k^+	Número de genes preditores ativadores de um determinado gene alvo
$\langle k \rangle$	Número médio de preditores por gene (grau médio) em uma rede
v_{ki}	k-ésimo gene preditor que possui aresta incidente ao gene v_i (alvo)
$\vec{s}$	Um estado de uma rede Booleana
$\mathbf{X}$	Vetor de características
a, b	Estado de um gene numa rede Booleana
$\|C\|$	Numero de funções canalizadoras existentes para n variáveis booleanas
w_i	i-ésimo coeficiente de una variavel booleana X_i numa função linearmente separavel
Y	Variável de classes (rótulos)
$H(\cdot)$	Entropia
$H(Y\|\mathbf{x})$	Entropia condicional de Y dado $\mathbf{X} = \mathbf{x}$
$H(Y\|\mathbf{X})$	Entropia condicional média de Y dado $\mathbf{X}$
$IM(\cdot)$	Informação mútua
$IM_d(\cdot)$	Informação mútua normalizada pela dimensão
$F(\mathbf{X}\|Y_j)$	Função de densidade dos preditores $\mathbf{X}$ para cada classe Y_j
P	Probabilidade
γ	Constante de decaimento de uma lei de potência
$\mathbf{Z}$	Subconjunto de características de $\mathbf{X}$
m	Número de amostras temporais
B_n	Número total de partições de um conjunto de n elementos (número de Bell)
$\mathbf{A}$	Vetor de coeficientes de uma combinação linear
$\mathcal{C}$	Conjunto dos possíveis valores dos coeficientes $\mathbf{A}$

a_i	Um dos componentes do vetor de coeficientes de uma combinação linear **A**
$\mathbf{A}^*$	Vetor de coeficientes da combinação linear que otimize uma determinada função critério
$\mathbf{z}$	Instância de $\mathbf{Z}$
L	Número inteiro de mapeamento linear de agrupamento de configurações (instâncias) dos preditores
$f(Y = y)$	Frequência de observações do valor $Y = y$
ECM_{min}	Entropia condicional média mínima
ECM_{AL}	Entropia condicional média por agrupamento linear
$\mathcal{F}$	Função critério

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Contextualização

A célula é a unidade fundamental da vida, constituindo todos os seres vivos. Existem diversos tipos de células com diferentes formas e funções, embora nos organismos multicelulares todas as células possuam essencialmente o mesmo DNA no seu núcleo. As células se diferenciam por sua função e pelas proteínas que elas geram. A produção de uma determinada proteína depende do gene associado a proteína em questão se encontrar ativo ou inativo em um determinado momento do processo celular. Tal processo confere a funcionalidade especifica de cada célula. O entendimento dos processos celulares é atualmente um dos principais focos de pesquisa em biologia sistêmica. A busca por esse entendimento envolve descobrir quais são as causas pelas quais alguns genes se encontram ativos ou inativos em um determinado contexto e momento. Ou seja, procura-se entender os mecanismos que regulam a expressão desses genes, por meio de envio e recepção de sinais [Snoep and Westerhoff, 2005].

Uma maneira de entender melhor estes mecanismos de controle regulatório é considerar a evolução temporal dos níveis de expressão gênica, ou seja, sua dinâmica. Em particular, o desenvolvimento de técnicas massivas de extração de informação molecular, como os DNA Microarrays [Shalon et al., 1996], SAGE (do inglês *Serial Analysis of Gene Expression*) [Velculescu et al., 1995], RNA-Seq [Wang et al., 2009], e mais recentemente o RNA-Seq de única célula (*single cell* RNA-Seq) [Eberwine et al., 2014], têm possibilitado estimar o nível de expressão de milhares de genes simultaneamente e em múltiplos instantes de tempo. As pesquisas nesse campo vêm recebendo forte atenção de pesquisadores do mundo todo na esperança de impactar positivamente no desenvolvimento de novos tratamentos e medicamentos contra doenças, no entendimento da biologia do câncer, na produção de bioenergia, dentre outras aplicações.

Os genes são elementos importantes dos sistemas de controle de organismos, consti-

tuindo uma forma de rede de comunicação que processa informação biológica e regula as vias metabólicas das células. Os genes apresentam a propriedade de se expressarem, produzindo cópias de segmentos de DNA na forma de RNA mensageiro (mRNA). Os mRNAs passam dos orifícios do núcleo celular para o citoplasma, onde são traduzidos em proteínas. Algumas dessas proteínas atuam como enzimas que catalisam reações metabólicas para manutenção das atividades da célula. Outras voltam para o núcleo atuando como fatores de transcrição de um gene e regulando sua síntese de mRNAs [Crick, 1970, D'haeseleer et al., 1999].

Nesse cenário podemos entender o sistema de regulação gênica como um grafo dirigido no qual os vértices são os genes e as arestas representam a dependência funcional entre os genes [Hecker et al., 2009]. Essa dependência funcional é indireta através da produção de proteínas que atuam como fatores de transcrição. Tipicamente essas dependências possuem uma influência ativadora ou inibidora dos genes preditores em relação ao gene alvo, mas tal influência também pode ser resultante de uma combinação não linear dos genes preditores na determinação da expressão de um gene alvo [Anastassiou, 2007, Martins-Jr et al., 2008, Marbach et al., 2010]. Por exemplo, um gene pode ser ativado apenas quando ambos os seus genes preditores estejam ativos produzindo proteínas que juntas formam um complexo proteico que atua como fator de transcrição do gene alvo, como exemplificado na Figura 1.1 [Hecker et al., 2009]. O conjunto de genes e suas respectivas dependências é denominado uma rede de regulação gênica (do inglês: *Gene Regulatory Network* - GRN).

Muitos modelos matemáticos e computacionais vêm sendo desenvolvidos para explicar interações gênicas [Hecker et al., 2009, Ristevski, 2013, Kotiang and Eslami, 2020] e existe um número considerável de tentativas para modelar redes de expressão gênica incluindo *grafos dirigidos* [Jong, 2002], modelos gráficos probabilísticos [Kotiang and Eslami, 2020] incluindo especialmente as *redes Bayesianas* [Friedman et al., 2000, Friedman, 2004], *redes lógicas generalizadas* [Thomas, 1991, Song et al., 2009], *equações diferenciais ordinárias* [Mestl et al., 1995, Ma et al., 2020], *redes Booleanas* (BN) [Kauffman, 1969, Tovar et al., 2019, Montagna et al., 2020, Shi et al., 2020], *redes Booleanas probabilísticas* (PBN) [Shmulevich et al., 2002] e *redes gênicas probabilísticas* (PGN) [Barrera et al., 2007]. Uma revisão sobre modelos de redes de regulação gênica pode ser vista em [Karlebach and Shamir, 2008, Hecker et al., 2009, Marbach et al., 2010, Ristevski, 2013, Martins-Jr et al., 2016, Banf and Rhee, 2017, Huynh-Thu and Sanguinetti, 2019, Delgado and Gómez-Vela, 2019].

No contexto dos modelos discretos, as redes Booleanas (do inglês: *Boolean Networks* - BNs) representam um modelo adequado para generalizar e capturar o comportamento dos sistemas biológicos em nível global (qualitativo), em face ao número limitado de experimentos (amostras), da alta dimensionalidade de variáveis (genes) e da natureza ruidosa

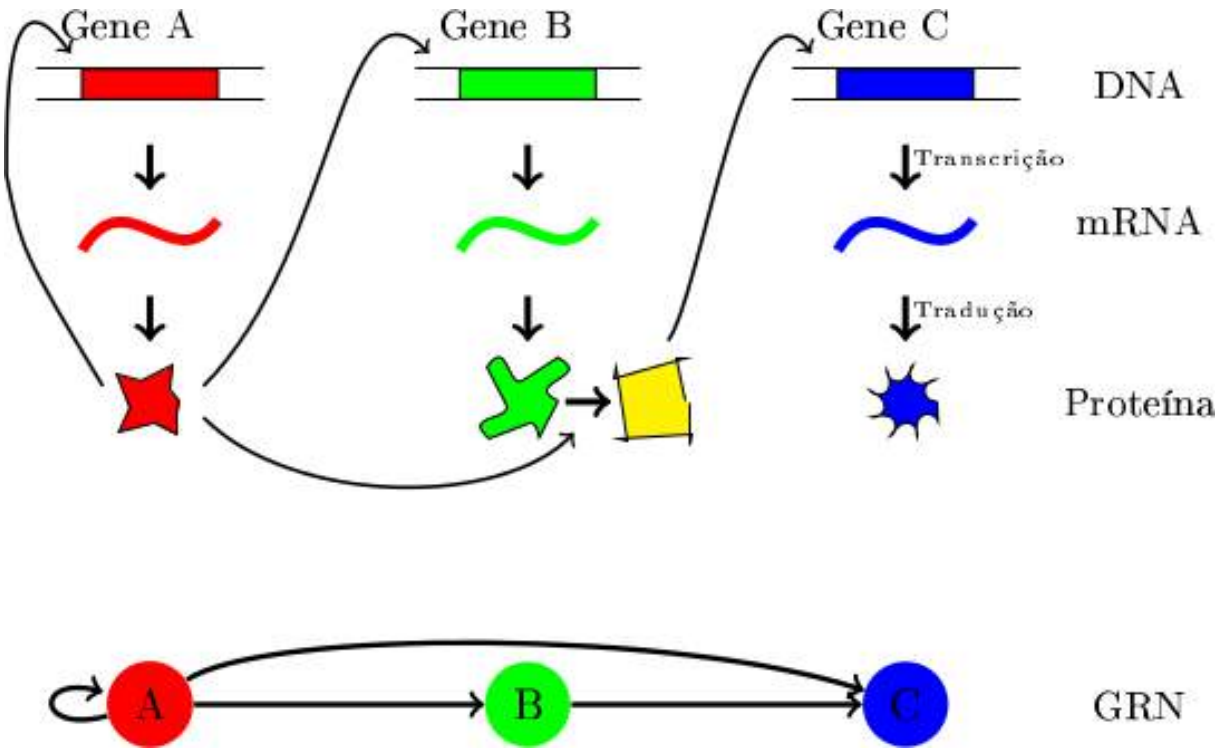


Figura 1.1: Exemplo de uma rede de regulação gênica (GRN) com 3 genes representada como um grafo. A proteína produzida pelo gene A (vermelha) atua como fator de transcrição do próprio gene A e do gene B. Essa mesma proteína vermelha produzida pelo gene A forma um complexo com a proteína verde produzida pelo gene B. Tal complexo atua como fator de transcrição do gene C. Sendo assim, o gene A é preditor de si mesmo, do gene B e do gene C, enquanto o gene B é preditor do gene C. Adaptado de [Hecker et al., 2009].

das medidas de expressão [Kauffman, 1969]. Embora as BNs sejam úteis em diversos casos, uma limitação importante é o seu determinismo inerente, que faz a suposição de um ambiente sem incerteza. Além disso, é importante considerar uma célula como um sistema aberto, o qual pode receber estímulos externos. Dependendo das condições externas em um dado instante de tempo, a célula pode alterar sua dinâmica [Shmulevich and Dougherty, 2014]. Para lidar com esse problema é proposto um tipo especial de BNs, as redes Booleanas probabilísticas (PBNs), as quais além de considerar genes com valores binários, associa a cada um deles um conjunto de funções Booleanas preditoras, atribuindo uma probabilidade a cada função específica [Shmulevich et al., 2002]. Embora esta abordagem tenha também desvantagens importantes, a desvantagem mais apontada é a perda de informação decorrente da discretização dos dados. Mas isso faz os modelos Booleanos mais simples e mais fáceis de serem tratados computacionalmente [Styczynski and Stephanopoulos, 2005, Ivanov and Dougherty, 2006]. Nossa proposta pretende se concentrar no modelo de redes Booleanas e redes Booleanas probabilísticas.

Embora haja muitos métodos de inferência de GRNs na literatura [Barrera et al., 2007,

Markowetz and Spang, 2007, Hecker et al., 2009, De-Smet and Marchal, 2010, Marbach et al., 2012, Ristevski, 2013, Lopes et al., 2014, Martins-Jr et al., 2016, Banf and Rhee, 2017, Carastan-Santos et al., 2017, Jacomini et al., 2017, Huynh-Thu and Sanguinetti, 2019, Delgado and Gómez-Vela, 2019, Marco et al., 2019, Kotiang and Eslami, 2020, Shi et al., 2020], a inferência de GRNs é considerada um problema mal posto (*ill-posed*), uma vez que, para um determinado conjunto de dados de perfis de expressão gênica, existem muitas redes (senão infinitas) capazes de explicar este mesmo conjunto de dados. Este problema é ainda mais dificultado devido a um número tipicamente limitado de amostras, uma enorme dimensionalidade (número de variáveis, ex., genes), além da presença de ruído, conforme já mencionado anteriormente [Barrera et al., 2007, Hecker et al., 2009, Shmulevich and Dougherty, 2014, Lopes et al., 2014, Martins-Jr et al., 2016, Jacomini et al., 2017, Banf and Rhee, 2017, Huynh-Thu and Sanguinetti, 2019, Delgado and Gómez-Vela, 2019, Shi et al., 2020].

Para auxiliar o processo de inferência, diversas metodologias têm sido utilizadas, baseadas em diversas áreas de estudo, tais como reconhecimento de padrões, inteligência artificial, teoria da informação, teoria de controle, redes complexas, inferência estatística, sistemas dinâmicos, entre outras. Uma técnica de reconhecimento de padrões comumente usada para inferir GRNs é a seleção de características [Liang et al., 1998, Barrera et al., 2007, Lopes et al., 2008, Lopes et al., 2014, Montoya-Cubas et al., 2015, Martins-Jr et al., 2016, Jacomini et al., 2017]. Técnicas baseadas nessa abordagem selecionam um subconjunto de genes que sejam bons preditores do padrão de expressão do gene alvo.

Em inferência de redes de regulação gênica a seleção de caraterísticas consiste em associar uma determinada variável (um gene alvo) a um conjunto de características (genes preditores), de forma que seja possível classificar o valor do gene alvo com base nos valores (instâncias) dos genes preditores. Para construir e avaliar esse classificador, estima-se uma tabela de probabilidades condicionais a partir dos dados observados, na qual cada possível instância resultará em uma distribuição de probabilidades (histograma) dos possíveis valores do gene alvo. A partir daí, uma função critério adotada é aplicada sobre essa tabela para avaliar a qualidade de predição (classificação) dessa tabela. O principal problema a ser tratado nesta tese é que o número de instâncias do conjunto dos candidatos a preditores cresce exponencialmente com o tamanho desse conjunto (número de candidatos a preditores), o que faz com que a estimação das probabilidades condicionais seja muito pobre para situações nas quais o número de candidatos a preditores é moderadamente grande e o número de amostras disponíveis é limitado. Esse problema é conhecido como maldição da dimensionalidade [Jain et al., 2000]. Mesmo considerando o modelo discreto mais simples no qual os genes possuem valores binários, o numero de instâncias da tabela de probabilidades condicionais é igual a 2^n, sendo n o numero de candidatos a preditores

Visto que o número de instâncias do conjunto de preditores candidatos cresce ex-

ponencialmente com o número de preditores, é preciso desenvolver técnicas de seleção de características para amenizar o problema de estimação estatística existente ao inferir redes gênicas a partir de um pequeno número de amostras. Em linhas gerais, uma proposta seria desenvolver uma estratégia que agrupe as instâncias em classes de equivalência de forma a obter um bom balanço entre a função critério, o poder de estimação das probabilidades condicionais, e a perda de informação inerente ao processo de agrupamento.

Durante o mestrado, foi desenvolvida uma primeira abordagem de agrupamento de instâncias dos candidatos a preditores em suas respectivas combinações lineares, fazendo com que o número de instâncias cresça linearmente com a dimensionalidade (número de preditores) ao invés de crescer exponencialmente quando as instâncias originais são consideradas. Pelo fato do valor de cada gene ser uma combinação linear de outros genes, cada gene é considerado um perceptron, mas com restrição nos pesos para que tenham valores -1 (inibição), 0 (ausência de dependência) e +1 (ativação). Esta abordagem mostrou-se promissora para inferir redes de regulação a partir de dados simulados [Montoya-Cubas, 2014, Montoya-Cubas et al., 2014, Montoya-Cubas et al., 2015], e a partir de dados reais de *Plasmodium falciparum*, um agente causador da malária[Montoya-Cubas et al., 2015]. Foram desenvolvidas variantes com base nessa abordagem para agrupar instâncias somente quando o número de observações das instâncias originais são consideradas insuficientes.

O problema do agrupamento de instâncias pode ser visto como um problema de análise multiresolução, cujo objetivo é encontrar a melhor resolução para resolver uma tarefa específica [Dougherty et al., 2001]. O interesse aqui é na tarefa de prever a expressão de um gene alvo com base na expressão de seus genes preditores selecionados. O objetivo então é encontrar a melhor configuração do espaço de todas as possíveis partições das instâncias de um conjunto de genes candidatos a preditores de um gene alvo, e avaliar esse conjunto de candidatos com base nessa configuração. Tal espaço também é conhecido como reticulado de partições (do inglês: *partition lattice*). No mestrado, essa ideia foi explorada aplicando-se agrupamentos lineares, implicando em uma forte restrição no espaço do reticulado de partições.

1.2 Objetivos

A proposta deste trabalho consistiu em continuar as pesquisas realizadas durante o mestrado, as quais envolveram o estudo e o desenvolvimento de técnicas de seleção de características para inferir redes gênicas, cujo princípio se baseia em reduzir o número de parâmetros de estimação (instâncias) dos valores dos preditores através de agrupamentos em classes de equivalência. Como dito anteriormente, no trabalho de mestrado foi abordado o agrupamento linear, para o qual as classes de equivalência são definidas por uma

combinação linear dos valores das instâncias. Entretanto, o agrupamento linear examina apenas um subconjunto bem restrito de possíveis agrupamentos (partições).

Já nesta proposta de doutorado, propomos investigar mais a fundo o problema do agrupamento de instâncias, visto como um problema de encontrar a melhor configuração do espaço de todas as possíveis partições dessas instâncias. Tal espaço também é conhecido como reticulado de partições (do inglês: *partition lattice*). Portanto, um dos objetivos principais foi o de formular o problema de encontrar um agrupamento ótimo de instâncias como um problema de busca no reticulado de partições, propondo assim funções objetivo e algoritmos de busca nesse reticulado que possam implicar em uma melhora na inferência de redes gênicas, tanto do ponto de vista topológico como do ponto de vista da dinâmica gerada pelas redes inferidas, quando comparados ao método original sem agrupamento e às variantes de agrupamento linear desenvolvidas durante o mestrado e em parte do doutorado. Além disso, outro objetivo principal dentro desse contexto foi o desenvolvimento de métodos para restringir o espaço de busca por meio de inclusão de informações a priori sobre características intrínsecas típicas das funções de predição encontradas em redes gênicas reais, tais como a própria linearidade, como também o fenômeno da canalização.

Os objetivos específicos são:

1. Reformulação do problema de agrupamento como um problema de busca no reticulado de partições;

2. Desenvolvimento de funções objetivo que orientem a busca no reticulado de partições de modo a encontrar agrupamentos que resultem em boas tabelas de probabilidades condicionais entre o alvo e os candidatos a preditores tanto do ponto de vista da qualidade de estimação, quanto do ponto de vista da predição dos valores do gene alvo, compensando o custo da perda da informação inerente ao processo de agrupamento;

3. Desenvolvimento de algoritmos que percorrem o espaço de partições com base nas funções objetivo desenvolvidas;

4. Desenvolvimento de métodos que restringem a busca no espaço de partições através de conhecimento a priori sobre as características típicas das funções de predição encontradas em redes gênicas, tais como linearidade e canalização;

5. Desenvolvimento um método de transferência de aprendizado supervisionado sobre as dimensões corretas dos conjuntos de preditores com base nos rótulos fornecidos por redes sintéticas ideais (*groundtruths*) e nos perfis de evolução dos valores da função critério de seleção de características com o aumento da dimensão;

6. Avaliação dos métodos desenvolvidos em dados simulados e em dados reais de *Plasmodium falciparum*, um agente causador da malária;

1.3 Justificativa

Este projeto está relacionado a um dos Grandes Desafios da Pesquisa em Computação no Brasil 2006-2016 [Carvalho, 2006] conforme proposto pela Sociedade Brasileira de Computação (SBC):

- *Desafio 3 - Modelagem Computacional de Sistemas Complexos Artificiais, Naturais e Sócio-culturais e da Interação Homem-natureza:* as técnicas desenvolvidas neste projeto serão úteis na identificação e modelagem de interação gênica, por meio de métodos de inferência e análise de seu comportamento dinâmico.

Um dos objetivos mais importantes das pesquisas em biologia sistêmica é o de entender o comportamento dinâmico dos genes por atuarem como agentes de controle, regulando grande parte dos fenômenos que ocorrem nos organismos. Há ainda muita pesquisa a ser realizada nesse sentido devido à complexidade dos processos celulares, incluindo as redes de regulação gênica.

A biologia sistêmica é considerada um dos últimos estágios das pesquisas em bioinformática, já que isso pressupõe um entendimento global de como os organismos funcionam. Trata-se de um campo interdisciplinar que envolve diversas áreas do conhecimento. Vale ressaltar que no Brasil há um crescente interesse de pesquisa em biologia sistêmica, sendo que no contexto de pesquisas em genômica e proteômica, existe o interesse do agronegócio que contribui com uma parte significativa da economia nacional, além de pesquisas envolvendo biocombustíveis tendo a cana-de-açúcar como um dos principais produtos investigados. O combate a doenças tropicais como a malária ou a dengue também é considerado estratégico, sendo um importante alvo de pesquisas nacionais. A criação da Rede Nacional de Bioinformática (RNBio)[1] é apenas mais um exemplo dos esforços recentes da ciência nacional na produção de conhecimento sobre sistemas biológicos. Finalmente, existe um grande interesse mundial voltado para aplicações médicas, tais como o entendimento da biologia do câncer e o tratamento e cura de doenças como a AIDS e doenças neurodegenerativas e do neurodesenvolvimento como a esquizofrenia, o transtorno do espectro autista (TEA), do déficit de atenção com hiperatividade (TDAH), o Alzheimer, dentre outros. Especialmente no contexto atual de pandemia, há um grande esforço mundial envolvendo o novo coronavirus SARS-COV-2 responsável pela COVID-19, caracterizada

[1] http://bioinfo.lncc.br/

como uma doença complexa que pode apresentar um variado grau de severidade e diversos desfechos clínicos dependendo de uma combinação de fatores ainda não muito bem compreendida [Wicik et al., 2020].

Embora a principal motivação deste trabalho seja o estudo do problema da inferência de redes gênicas, é importante notar que as contribuições deste trabalho podem ser úteis para problemas de outros domínios, já que o tema central abordado é o de análise multiresolução, o qual tem sido aplicado com sucesso, por exemplo, para projetar filtros digitais de imagens [Dougherty et al., 2001].

1.4 Síntese das contribuições

- Proposta de uma função critério de seleção de características baseada na informação mútua normalizada pela dimensão do subconjunto de características, com o objetivo de estimar a dimensão ideal do subconjunto de características resultante de um processo de seleção de características (Seção 2.3.6);
- Reformulação do problema de agrupamento de instâncias como um problema de busca no espaço de partições estruturado sobre um reticulado com ordem parcial entre as partições (Seção 3.1);
- Proposta de um método de agrupamento linear de instâncias baseado na hipótese de linearidade nas funções de predição que compõem as redes gênicas, com o objetivo de reduzir o numero de instâncias a serem estimadas. O desenvolvimento dessa proposta foi realizado durante o mestrado e aperfeiçoado durante o doutorado (Seção 3.2);
- Proposta de um método de agrupamento de instâncias baseado em busca sequencial para frente no reticulado de partições, procurando priorizar o agrupamento de instâncias com menos amostras, de modo a obter um numero mínimo de amostras por grupo e ao mesmo tempo reduzir o número de parâmetros de estimação (Seção 3.3);
- Proposta de um método de agrupamento de instâncias baseado na hipótese de presença de canalização nas funções de predição (Seção 3.4);
- Proposta de um método de transferência do aprendizado das dimensões (graus) dos genes sobre redes geradas artificialmente, para um cenário de dados reais, como os dados de *microarray* de *Plasmodium falciparum* (Seção 3.5);
- Publicação do artigo [Montoya-Cubas et al., 2015] sobre o método de agrupamento linear e variantes propostas, bem como uma avaliação desse método em dados simulados e dados reais do *Plasmodium falciparum*;

- Finalização de um artigo a ser submetido no periódico *Bioinformatics* sobre os métodos de agrupamento desenvolvidos, bem como a estratégia de transferência de aprendizado supervisionado dos graus com base em redes simuladas geradas artificialmente (previsão de submissão: entre dezembro/2020 e janeiro/2021).

1.5 Organização do texto

Esta tese está organizada do seguinte modo.

O Capítulo 2 apresenta uma revisão sobre os conceitos necessários para o entendimento do restante da tese:

- Sinais de expressão gênica e a tecnologia de *microarray* de medição desses sinais (Seção 2.1);
- Revisão sobre modelos de redes de regulação gênica, com foco em redes Booleanas (BN), redes Booleanas probabilísticas (PBN) e redes gênicas probabilísticas (PGN) (Seção 2.2);
- Na Seção 2.3 são apresentados tópicos de reconhecimento de padrões com foco em análise de expressão gênica, incluindo os conceitos de classificador Bayesiano, reticulados Booleanos, o problema da dimensionalidade, algoritmos de busca e funções critério de seleção de características, incluindo a informação mútua normalizada por dimensão proposta nesta tese. Também serão discutidos algoritmos de busca e funções critério de seleção de características adotados nesta tese. Finalmente, será discutido o método de classificação por k vizinhos mais próximos, adotado para o aprendizado dos graus dos genes (dimensões dos conjuntos de preditores dos genes).
- A Seção 2.4 apresenta modelos de topologias de redes complexas, incluindo o modelo de geração de redes aleatórias de Erdös-Rényi como também o modelo de geração de redes livres de escala de Barabási-Albert.
- A Seção 2.5 apresenta a metodologia adotada de avaliação das redes inferidas, incluindo critérios adotados para a avaliação topológica e da dinâmica gerada pela rede inferida.

Os métodos de agrupamento são introduzidos no Capítulo 3:

- A Seção 3.1 apresenta uma reformulação do problema de agrupamento de instâncias como um problema de busca no espaço de partições estruturado sobre um reticulado com imposição de ordem parcial entre as partições;

- A Seção 3.2 apresenta o método de agrupamento linear;
- A Seção 3.3 apresenta o método de agrupamento por busca sequencial para frente no reticulado de partições;
- A Seção 3.4 apresenta o método de agrupamento por canalização;
- Finalmente, a Seção 3.5 apresenta uma estratégia para a transferência de aprendizado supervisionado dos graus dos genes baseado em redes gabaritos geradas artificialmente.

O Capítulo 4 apresenta as análises e os resultados experimentais dos métodos de agrupamento propostos e do método original sem agrupamento em dados simulados.

O Capítulo 5 apresenta os resultados experimentais da aplicação de todos os métodos com a adoção da estratégia de transferência de aprendizado dos graus a partir de redes gabaritos geradas aleatoriamente. Os resultados desse capítulo envolvem uma comparação com os resultados do capítulo anterior (Capítulo 4).

O Capítulo 6 apresenta os resultados da aplicação dos métodos com e sem a adoção da estratégia de transferência de aprendizado para dados de *microarray* do *Plasmodium falciparum*.

Finalmente, o Capítulo 7 encerra a tese com conclusões e perspectivas futuras.

Capítulo 2

Revisão e conceitos básicos

2.1 Sinais de expressão gênica

Dado que uma funcionalidade específica de uma célula é fortemente determinada pelos genes que ela expressa, e sendo a transcrição o primeiro passo no processo de converter a informação armazenada no DNA do organismo em proteínas, é de se esperar que esse processo seja regulado pela rede de controle que coordena a atividade celular [Shmulevich and Dougherty, 2014]. Um meio primário de regulação da atividade celular é o controle de produção de proteína através da quantidade de RNAs mensageiros (mRNA) expressos pelos genes. Os mRNAs passam pelos orifícios do núcleo celular chegando no ribossomo, no citoplasma, onde eles são traduzidos em sequências de proteínas, construindo enzimas que catalisam reações metabólicas ou retornam ao núcleo para interagir com o DNA na regulação da síntese de RNA. A Figura 2.1 ilustra esse processo dentro da célula. Portanto, conjuntos de genes constituem redes de comunicação bastante complexas que controlam vias metabólicas.

Com o advento das tecnologias *microarray* [Shalon et al., 1996], SAGE (*Serial Analysis of Gene Expression*) [Velculescu et al., 1995], RNA-Seq [Wang et al., 2009], e mais recentemente o RNA-Seq de única célula (*single cell* RNA-Seq) [Eberwine et al., 2014], a quantidade de dados públicos disponíveis de expressão gênica de várias espécies e em diversas condições vem aumentando consideravelmente a cada ano. Qualquer que seja a tecnologia utilizada, no final os dados de expressão são constituídos essencialmente de uma matriz com os genes dispostos nas linhas e os experimentos (condições ou instantes de tempo distintos) dispostos nas colunas.

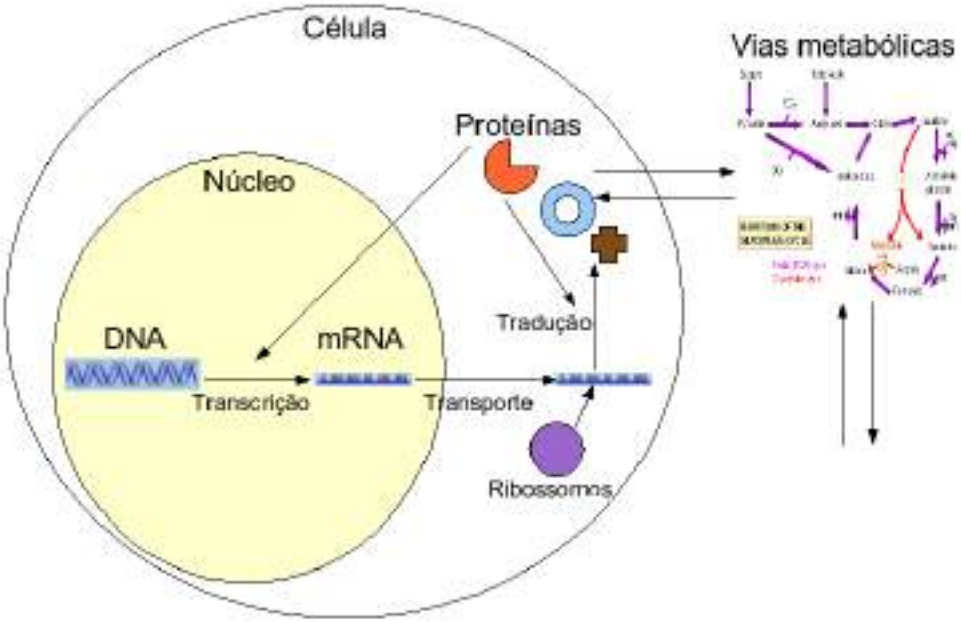


Figura 2.1: Esquema simplificado da dinâmica celular (Fonte: [Martins-Jr., 2008]).

2.2 Modelagem de redes gênicas

2.2.1 Visão geral

Existem duas abordagens principais para modelar matematicamente as redes complexas de interações gênicas [Shmulevich and Dougherty, 2014]. Uma das abordagens é a que considera variáveis no domínio contínuo, empregando equações diferenciais e suas variações para construir modelos quantitativos detalhados de redes bioquímicas com funções celulares de interesse [Mestl et al., 1995, Jong, 2002, Ma et al., 2020]. A segunda abordagem baseia-se na construção de modelos qualitativos discretos de interação gênica, incluindo os modelos baseados em grafos tais como as redes Bayesianas [Friedman et al., 2000, Friedman, 2004], redes Booleanas [Kauffman, 1969], e redes Booleanas probabilísticas [Shmulevich et al., 2002] que incluem as redes gênicas probabilísticas [Barrera et al., 2007]. Embora a abordagem contínua possibilite um entendimento detalhado do sistema em questão, em geral ela necessita de um conjunto considerável de amostras, além de informações sobre determinadas características das reações [Karlebach and Shamir, 2008], o que a torna apropriada em um número reduzido de situações. Em contrapartida, as abordagens discretas podem ser facilmente modeladas computacionalmente, tendo sido empregadas com sucesso na modelagem e simulação de algumas redes e processos biológicos, tais como Drosophila melanogaster [Sánchez and Thieffry, 2001, Albert and Othmer, 2003], ciclo celular da levedura [Li et al., 2004, Zhang et al., 2006, Davidich and Bornholdt, 2008], Arabdopsis thaliana [Espinosa-Soto et al., 2004, Li et al., 2006], Saccharomyces cerevisiae [Li and Lu, 2005], ciclo celular de mamíferos [Faure et al., 2006], Plasmodium

falciparum [Barrera et al., 2007, Lopes et al., 2008, Montoya-Cubas et al., 2015, Jacomini et al., 2017], dentre outros.

Em particular, as redes Bayesianas constituem um modelo probabilístico capaz de representar uma rede gênica causal. Nesse tipo de modelagem, as entidades biológicas (genes, proteínas e outras moléculas) são representadas como nós de um grafo que são conectados por arestas que indicam um determinado relacionamento entre eles. As redes Bayesianas são amplamente utilizadas para representar redes gênicas [Friedman et al., 2000, Kelemen et al., 2008]. Tal modelo utiliza distribuições de probabilidades, teoria dos grafos e propriedade local de Markov (cada variável é condicionalmente independente de seus não-ancestrais) para representar relações entre variáveis e estados com o objetivo de realizar inferências. A inferência de redes Bayesianas com base em um número pequeno de amostras, como é o caso dos dados de expressão gênica, é um importante desafio.

O modelo de redes Booleanas (do inglês: *Boolean Networks* - BNs) constitui um tipo de rede Bayesiana dinâmica discreta que representa um modelo adequado para generalizar e capturar o comportamento dos sistemas biológicos em nível global (qualitativo), face ao número limitado de experimentos (amostras), da alta dimensionalidade de variáveis (genes) e da natureza ruidosa das medidas de expressão [Kauffman, 1969, Lähdesmäki et al., 2006, Montagna et al., 2020]. Embora as BNs sejam úteis em diversos casos, uma limitação importante é o seu determinismo, que faz a suposição de um ambiente sem incerteza. Além disso, é importante considerar uma célula como um sistema aberto, o qual pode receber estímulos externos. Dependendo das condições externas em um dado instante de tempo, a célula pode alterar sua dinâmica [Shmulevich and Dougherty, 2014]. Para lidar com esse problema é proposto um tipo especial de BNs, as redes Booleanas probabilísticas (PBNs), nas quais além de considerar genes com valores binários, associa a cada um deles um conjunto de funções Booleanas preditoras, atribuindo uma probabilidade a cada função específica [Shmulevich et al., 2002]. Embora esta abordagem tenha também desvantagens importantes, a desvantagem mais apontada é a perda de informação decorrente da discretização dos dados. Mas isso faz os modelos Booleanos mais simples e mais fáceis de serem tratados computacionalmente [Styczynski and Stephanopoulos, 2005]. Uma discussão a respeito disso pode ser vista em [Ivanov and Dougherty, 2006]. Este trabalho adota os modelos de redes Booleanas e redes Booleanas probabilísticas para modelagem de redes gênicas. A próxima seção discute particularidades importantes desses modelos.

2.2.2 Redes Booleanas

As redes Booleanas (BNs) foram introduzidas por Kauffman [Kauffman, 1969] para a modelagem da dinâmica de sistemas complexos e em particular de GRNs. As BNs são definidas por um conjunto de vértices $V = \{v_1, v_2, ..., v_n\}$ e um conjunto de funções Bo-

oleanas $\Phi = \{\phi_1, \phi_2, ..., \phi_n\}$, uma para cada gene, também conhecidas como funções de transição Booleanas [D'haeseleer et al., 1999].

Cada gene $v_i \in \{0, 1\}, i = 1, 2, ..., n$ representa uma variável binária, e seu valor no instante de tempo $t + 1$ é completamente determinado pelos valores dos seus k genes preditores no instante de tempo t. Mais formalmente, podemos representar esta dinâmica como $v_i(t + 1) = \phi_i(v_{1i}(t), v_{2i}(t), ..., v_{ki}(t))$, na qual $v_{1i}, v_{2i}, ..., v_{ki}$ representam os k genes preditores ou regulatórios que possuem arestas incidentes ao gene v_i (alvo).

Desta forma, as funções Booleanas Φ são usadas para atualizar os genes, considerando iterações discretas no tempo, sendo todos os genes atualizados de forma sincronizada de acordo com a função atribuída a ele. Este processo síncrono simplifica a computação e preserva características gerais da dinâmica da rede [Kauffman, 1969].

Neste tipo de rede a dinâmica é determinística, ou seja, a escolha dos k preditores e respectivas funções lógicas para cada gene permanecem inalteradas durante todos os instantes de tempo. Quando as funções Booleanas ϕ_i são escolhidas de forma aleatória para cada um dos genes, a BN recebe o nome de rede Booleana aleatória (do inglês: *Random Boolean Network*) [Shmulevich and Dougherty, 2014].

O estado de um gene v_i em uma BN é definido pelo valor assumido por ele, $v_i = 1$ representa que o gene está ativo e $v_i = 0$ inativo. Um estado $\vec{s}$ de uma BN é definido pelos valores de todos os genes em um dado instante de tempo, $\vec{s}(t) = (v_1, v_2, ..., v_n), v_i \in \{0, 1\} \forall i = 1, 2, ..., n$.

Para cada BN temos 2^n possíveis estados definidos por $S = \{\vec{s}_1, \vec{s}_2, ..., \vec{s}_z\}, z = 2^n$, sendo n o número de genes da rede (BN). Mas nem sempre todos os 2^n estados são observados (estão presentes) numa rede, enquanto certos estados serão frequentemente observados, dependendo do estado inicial escolhido. Esses estados que são observados periodicamente compõem os chamados atratores (ciclos), e os estados que levam até os atratores são chamados estados transientes, os quais compõem a bacia de atração representada pelo atrator correspondente.

Os atratores são estados estacionários de um sistema dinâmico que capturam o comportamento deste sistema a longo prazo [Shmulevich and Dougherty, 2014, Montagna et al., 2020]. Os atratores são sempre cíclicos e podem ser formados por um ou mais estados. O número de estados que um sistema pode visitar antes de retornar a um estado já visitado é denominado tamanho do ciclo. Kauffman interpreta que os atratores de uma BN podem ser vistos como diferentes tipos celulares, argumentando que diferentes células são caracterizadas por seu padrão recorrente de expressão gênica, de certa forma correspondente aos atratores de uma BN [Kauffman, 1969].

GRNs reais são altamente estáveis na presença de perturbações ocasionadas por fatores externos, seja de algum gene isoladamente ou de vários genes. Considerando o formalismo

das BNs, isto significa que, quando um número mínimo de genes são perturbados, estes genes mudam de valores (estados), mas os estados da rede continuam na mesma bacia de atração e acabam por chegar no mesmo atrator [Shmulevich and Dougherty, 2014]. Neste sentido, atratores com grandes bacias de atração conferem uma alta estabilidade para o sistema [Li et al., 2004, Zhang et al., 2006]. Esta estabilidade das redes regulatórias em organismos vivos permite que as células mantenham seu estado funcional no organismo mesmo quando submetidas a perturbações externas. Entretanto, dependendo do grau de perturbação, algumas células podem acabar passando por situações anormais, como por exemplo as células cancerosas capturadas em ciclos atratores erráticos denominados "*cancer attractors*" [Huang et al., 2009].

2.2.3 Redes Booleanas Probabilísticas

As Redes Booleanas Probabilísticas, as quais incluem as Redes Gênicas Probabilísticas [Barrera et al., 2007] (ver Seção 2.2.5), são tipos específicos de Redes Bayesianas dinâmicas, nas quais cada gene em um determinado instante de tempo tem o seu valor de expressão binária determinado por um conjunto de funções Booleanas de outros genes no instante de tempo anterior, onde cada função tem uma probabilidade de ser aplicada [Shmulevich et al., 2002]. Normalmente, modela-se o quase-determinismo inerente em GRNs simplesmente impondo que uma dessas funções tenha uma probabilidade bastante próxima de 1, enquanto as funções de menor probabilidade fazem o papel de perturbações ou mudanças de contexto biológico [Brun et al., 2005, Dougherty et al., 2007]. Como consequência, as Redes Booleanas representam um tipo particular de redes Booleanas Probabilísticas em que cada gene possui uma única função preditora.

2.2.4 Funções booleanas em redes de regulação gênica

Dentre as funções booleanas, existem dois conjuntos de funções biologicamente relevantes: as funções canalizadoras e as funções linearmente separáveis.

Funções canalizadoras

As funções de canalização são consideradas biologicamente relevantes [Waddington, 1942, Kauffman et al., 2004, Layne et al., 2012, Li et al., 2013]. Uma função booleana ϕ é canalizadora se houver pelo menos uma variável $X_i \in \{0,1\}$ em $\mathbf{X}$ de modo que $\phi(\mathbf{X}) = b \in \{0,1\}$ sempre que $X_i = a \in \{0,1\}$. Em outras palavras, uma função ϕ é canalizadora caso existam $i \in \{1,...,n\}$, $\{a,b\} \in \{0,1\}^2$ tal que $X_i = a => \phi(\mathbf{X}) = b$.

Por exemplo, $\phi(\mathbf{X}) = X_1 \wedge (X_2 \oplus X_3)$ é uma função de canalização porque $X_1 = 0$

induz $\phi(\mathbf{X}) = 0$. Já a função $\phi(\mathbf{X}) = X_1 \oplus X_2$ (XOR) não é canalizadora, tendo em vista que essa função é um teste de diferença entre X_1 e X_2, ou seja, resulta em 0 se as variáveis forem iguais, ou em 1 se forem diferentes. Assim, o resultado dessa função sempre depende de conhecer os valores de ambas as variáveis, fazendo com que nenhuma das variáveis seja canalizadora.

Vale notar que o numero de funções canalizadoras existentes para n variáveis booleanas é dado por:

$$|C| = 2((-1)^n - n) + \sum_{i=1}^{n} (-1)^{i+1} S_i.$$

em que $S_i = \binom{n}{i} 2^{i+1} 2^{2^{n-i}}$ [Just et al., 2004]

Funções linearmente separáveis

Funções linearmente separáveis (em inglês: *threshold function*) são frequentemente encontrados em processos biológicos tais como os ciclos celulares de *S. cerevisiae* e *S. pombe* [Li et al., 2004, Davidich and Bornholdt, 2008]. Um estudo detalhado sobre a dinâmica das redes biológicas com funções linearmente separáveis pode ser encontrado em [Zanudo et al., 2011].

Sejam $w_1, ..., w_n$ e θ números reais. Uma função booleana ϕ é linearmente separável se ϕ é representada como:

$$\phi(x) = \begin{cases} 1, & \text{se } \sum w_i X_i \geq \theta, \\ 0, & \text{se } \sum w_i X_i < \theta. \end{cases}$$

Muitas funções booleanas são linearmente separáveis. Por exemplo, $x \wedge y$ pode ser representado como $x + y \geq 2$, $x \wedge \bar{y}$ pode ser representado como $x - y \geq 1$, $x \vee y$ pode ser representado como $x + y \geq 1$ e $x \vee \bar{y}$ pode ser representado como $x - y \geq 0$. Por outro lado, existem também funções booleanas que não são linearmente separáveis. Por exemplo, $x \oplus y$ (XOR) não é uma função linearmente separável. Funções linearmente separáveis têm sido amplamente utilizadas em modelos discretos de redes neurais [Rani et al., 2018].

2.2.5 Redes gênicas probabilísticas

Os perfis de expressão de genes preditores em uma rede de regulação oferecem um conteúdo informativo relevante (individualmente ou em conjunto com outros preditores) sobre o perfil de expressão de um dado gene alvo. Métodos de seleção de características podem

ser empregados para encontrar o subconjunto de genes (preditores) apresentando o maior conteúdo informativo sobre os valores do gene alvo.

Em particular, a abordagem de redes gênicas probabilísticas (do inglês: *Probabilistic Gene Networks* - PGN) [Barrera et al., 2007, Lopes et al., 2008] segue o princípio de seleção de características: para cada alvo, é realizada uma busca pelo subconjunto de preditores que melhor descreve o comportamento do alvo de acordo com seus sinais de expressão. Barrera *et al* discute essa abordagem no contexto da análise de sinais dinâmicos de expressão do Plasmodium *falciparum* (um dos agentes da malária), provendo resultados biológicos interessantes [Barrera et al., 2007]. Essa abordagem assume que as amostras temporais seguem uma cadeia de Markov de primeira ordem em que cada valor do gene alvo em um dado instante de tempo depende apenas dos valores dos seus preditores no instante de tempo anterior. A função de transição é homogênea (é sempre a mesma para todos os instantes de tempo), quase determinística (de qualquer estado, existe um estado preferencial para o sistema ir no próximo instante de tempo) e condicionalmente independente (ou seja, o valor de um determinado gene é dependente apenas dos valores de seus preditores). Outra propriedade simplificadora e biologicamente plausível do modelo PGN considera que as funções preditoras são caracterizadas por combinações lineares [Barrera et al., 2007]. Em um dos métodos de inferência de redes gênicas descrito na Seção 3.2, proposto durante o mestrado, a abordagem PGN é aplicada com um agrupamento linear das configurações dos preditores de cada alvo de modo a amenizar o problema da dimensionalidade. Tal método foi o primeiro envolvendo o modelo PGN que considerou essa propriedade simplificadora de modo direto.

Alguns conceitos fundamentais de reconhecimento de padrões e seleção de características para compreender o restante do texto são apresentados a seguir (Seção 2.3).

2.3 Reconhecimento de padrões e seleção de características

Em análise de sinais genômicos, uma grande variedade de problemas envolve reconhecimento de padrões. Por exemplo, *microarrays* e RNA-Seq contêm medidas de expressão de milhares de transcritos e um dos principais objetivos é a classificação de padrões a partir desses perfis de expressão. Exige-se, portanto, o projeto de um classificador ψ que receba como entrada um vetor de níveis de expressão gênica $\mathbf{X} = (x_1, x_2, ..., x_n)$ e devolva um rótulo que prediz a classe $Y = \{0, 1, ..., c-1\}$ à qual o vetor considerado pertence. Um problema típico em análise de expressão gênica é a classificação de diferentes tipos de câncer ou diferentes estágios de desenvolvimento de um tumor [Porter et al., 2001]. Classificadores são projetados com base em um conjunto de amostras (vetores de expressão)

que podem ser provenientes de tecidos diferentes ou de um mesmo tecido, que em geral é submetido a diversas condições ou em diferentes estágios de ciclo celular.

O problema de seleção de características consiste em selecionar um subconjunto de características que represente adequadamente os objetos em estudo. Uma técnica de seleção de características é dividida em duas partes principais: um algoritmo de busca e uma função critério [Theodoridis and Koutroumbas, 1999]. Em análise de expressão gênica, as características são os genes, cujos valores são dados pela expressão gênica. Os conjuntos de dados de expressão gênica usualmente apresentam milhares de características. Alguns métodos de inferência de GRNs que aplicam técnicas de seleção de características foram propostos na literatura [Liang et al., 1998, Butte and Kohane, 2000, Hashimoto et al., 2004, Peng et al., 2005, Margolin et al., 2006, Faith et al., 2007, Barrera et al., 2007, Zhao et al., 2008, Dougherty et al., 2008, Lopes et al., 2008, Borelli et al., 2013, Lopes et al., 2014, Montoya-Cubas et al., 2015, Martins-Jr et al., 2016, Jacomini et al., 2017].

2.3.1 Problema da dimensionalidade

Um importante passo para o desenho de um sistema de classificação é a avaliação do desempenho de um classificador, no qual a probabilidade de erro de classificação é estimada. Além da complexidade computacional, outra motivação para a seleção de características é a existência do problema da dimensionalidade, no qual o erro do classificador em função do número de características que descrevem os padrões (dimensionalidade) forma uma "curva em U" (ver Figura 2.2) [Jain et al., 2000, Bishop, 2006]. Observando essa figura, constata-se que para as dimensões menores que d_1, a adição de características implica em uma melhora no desempenho esperado do classificador. Entre as dimensões d_1 e d_2, a inclusão de características passa a não causar qualquer impacto significativo em seu desempenho. O problema da dimensionalidade começa a ocorrer após o ponto d_2 em que novas características passam a afetar negativamente o desempenho esperado do classificador.

O número de amostras necessárias para que um classificador tenha um desempenho satisfatório é exponencial com relação à dimensão do vetor de características [Jain et al., 2000]. Devido a isso, é muito comum que um classificador se torne excessivamente ajustado aos dados de treinamento (*overfitting*) caso o número de características selecionadas para o projeto do classificador seja muito maior face ao tamanho do conjunto de amostras de treinamento.

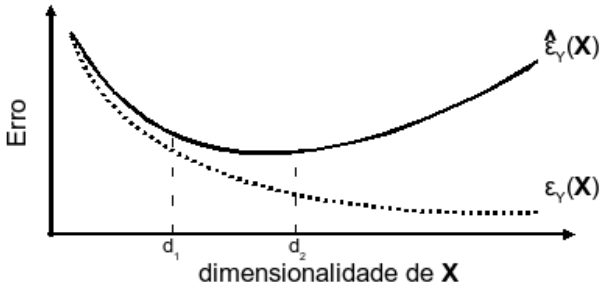


Figura 2.2: Gráfico das taxas de erro em função da dimensionalidade com número fixo de amostras ilustrando o problema da dimensionalidade. A curva do erro Bayesiano (erro do classificador ótimo) é dada por $\varepsilon_Y(\mathbf{X})$, enquanto a curva do erro esperado ao aplicar um classificador projetado a partir de um número finito de amostras é dada por $\hat{\varepsilon}_Y(\mathbf{X})$.

2.3.2 Reticulados Booleanos

Na grande maioria dos casos, e particularmente no contexto de seleção de características, os reticulados Booleanos são estruturas algébricas que normalmente representam o conjunto potência (*power-set*) de um conjunto de elementos, ou seja, todos os conjuntos propriamente contidos nesse conjunto, incluindo o conjunto vazio e o conjunto total. Em seleção de características, normalmente os elementos são as características dos objetos em estudo. Desse modo, o reticulado Booleano representa o espaço de busca de todos os subconjuntos de características possíveis. A Figura 2.3(a) ilustra um reticulado Booleano representando todos os possíveis subconjuntos do conjunto $\mathbf{X} = \{X_1, X_2, X_3\}$, sendo que a ilustração da Figura 2.3(b) mostra as cadeias binárias correspondentes dos subconjuntos, nos quais cada bit representa ausência (0) ou presença (1) de um dos elementos em um dado subconjunto. Por exemplo, o subconjunto $\{X_1, X_3\}$ é representado pela cadeia 101, já que os elementos X_1 e X_3 estão presentes, enquanto o elemento X_2 está ausente do subconjunto de características em questão. Vale notar ainda que as arestas do reticulado Booleano representam a vizinhança entre dois subconjuntos, de tal forma que dois subconjuntos são vizinhos se a diferença entre eles é de apenas um elemento. Em outras palavras, uma aresta representa um mapeamento de um subconjunto a outro pela adição ou remoção de um elemento específico.

Os algoritmos de busca para seleção de características determinam um passeio ao longo do reticulado Booleano de modo a procurar pelo subconjunto que otimize uma determinada função critério. Tal função critério mapeia cada subconjunto em um determinado valor que quantifica a sua qualidade em representar os objetos em estudo.

Neste trabalho, como o foco é em relação a função critério, os reticulados Booleanos apresentados na Seção 3.1 para explicar os métodos desenvolvidos representam um conceito diferente do apresentado até agora. Tais reticulados representam todos os possíveis valores (instâncias ou configurações) de um conjunto de elementos, em que cada

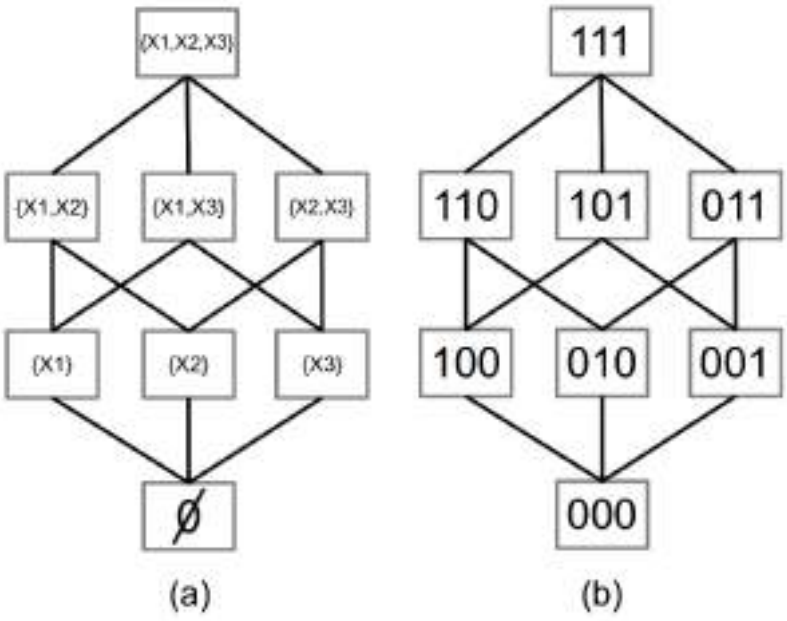


Figura 2.3: Reticulado Booleano de grau 3 representando todos os possíveis subconjuntos de 3 elementos: (a) descrição dos subconjuntos; (b) cadeias binárias correspondentes.

elemento possui dois valores possíveis no modelo de redes Booleanas para representação de redes gênicas. E as arestas desses reticulados representam uma relação de vizinhança entre as instâncias de acordo com a distância de Hamming[1]. Ou seja, existe uma aresta entre duas configurações se, e somente se duas configurações possuem distância de Hamming igual a 1. Por exemplo, na Figura 2.3(b), o vértice 101 corresponde à instância $\{X_1, X_2, X_3\} = \{1, 0, 1\}$. Isto é, são instâncias que diferem apenas pelo estado de um gene.

2.3.3 Algoritmos de busca para seleção de características

Como discutido anteriormente, os algoritmos de seleção de características percorrem parte do conjunto potência do conjunto total de características em busca de um subconjunto que otimize uma determinada função custo. Até o momento, não se conhece um algoritmo polinomial para resolver o problema da seleção de características [Pudil et al., 1994, Somol et al., 1999, Nakariyakul and Casasent, 2009]. Consequentemente, a busca exaustiva, a qual percorre todo o espaço de busca, é o único algoritmo capaz de obter a solução ótima em geral, embora existam algoritmos do tipo *branch-and-bound* que garantem otimalidade para funções critério com estruturas específicas [Jain et al., 2000, Somol and Pudil, 2004, Ris et al., 2010, Reis et al., 2019]. A Figura 2.4 apresenta a taxonomia dos principais métodos utilizados em reconhecimento de padrões. Tais algoritmos são categorizados de

[1] A distância de Hamming entre duas cadeias é o número de posições nas quais elas diferem entre si

acordo com dualidades tais como ótimo (devolve a melhor solução) *versus* sub-ótimo, determinístico (devolve sempre a mesma solução) *versus* estocástico (pode devolver soluções diferentes em execuções distintas), única solução *versus* várias soluções.

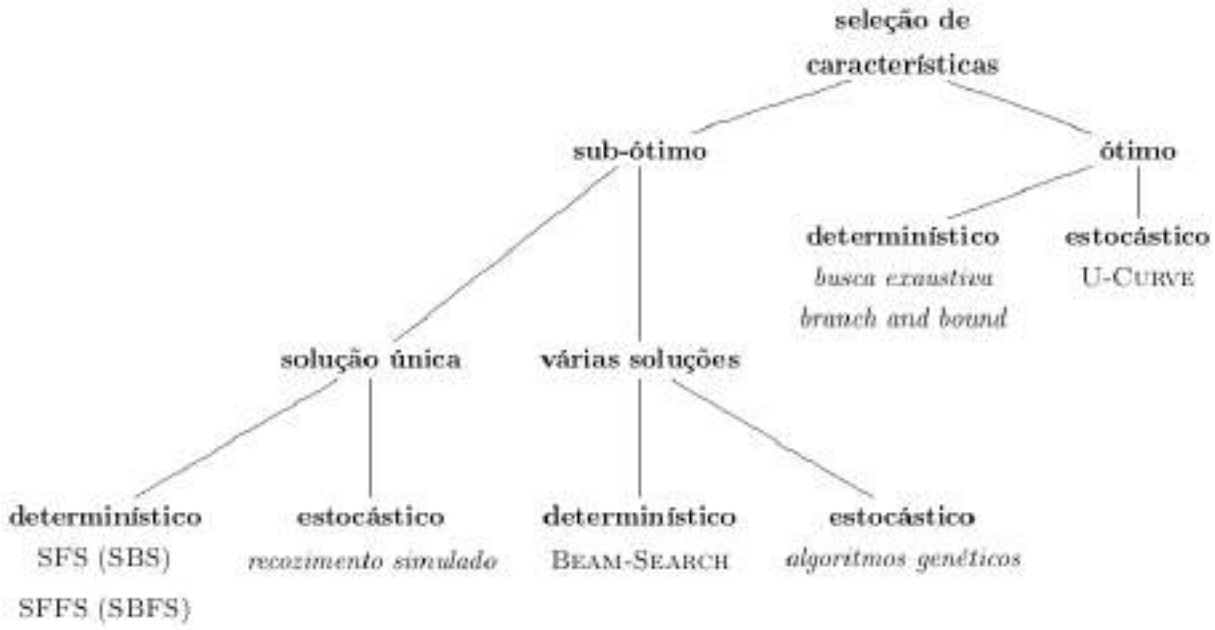


Figura 2.4: Categorização dos algoritmos de seleção de características comumente empregados em reconhecimento de padrões (Fonte: [Reis, 2012]).

Nesta seção serão apresentados apenas o algoritmo SFS e o algoritmo EX-SFS, tendo em vista que estes dois foram utilizado nos experimentos dos Capítulos 4, 5 e 6.

Busca Sequencial para Frente (SFS)

A busca sequencial para frente (*Sequential Forward Search* - SFS) é um algoritmo de busca guloso que começa com um subconjunto resultado vazio e adiciona a melhor característica encontrada a esse subconjunto. Em seguida, adiciona uma segunda característica que, em conjunto com a primeira, forma o melhor par de características, e assim sucessivamente. [Pudil et al., 1994]

Busca hibrida Exaustiva - Sequencial para Frente (EX-SFS)

A busca hibrida Exaustiva - Sequencial para Frente (EX-SFS) realiza uma busca exaustiva até uma determinada dimensão, a qual pode ser fixa segundo a complexidade dada pelo numero de variáveis. Uma vez que a busca exaustiva atinge essa dimensão, a busca continua através do algoritmo SFS até atingir o critério de parada.

2.3.4 Funções critério

O problema central em reconhecimento de padrões é projetar classificadores a partir de um conjunto de treinamento, neste caso os dados são fornecidos por uma distribuição conjunta de probabilidades das configurações de um conjunto de características e de seus respectivos rótulos. Geralmente, tal distribuição é estimada a partir de um número limitado de amostras, o que justamente é um caso recorrente em inferência de redes gênicas. O erro de estimação depende da seleção de um conjunto de características que tente predizer os rótulos a partir dos padrões observados. Uma função critério é uma medida de qualidade dessa distribuição conjunta estimada. A escolha da função critério é fundamental, pois seu papel é orientar os algoritmos de busca por um subconjunto de características que melhor predizem os rótulos a partir das amostras com o objetivo de projetar classificadores que cometam poucos erros de rotulação (classificação). Existem diversas funções critério propostas na literatura, dentre as quais pode-se destacar:

- Distância de Mahalanobis [Theodoridis and Koutroumbas, 1999]
- Distância de divergência [Duda et al., 2000]
- Distância Kullback-Leibler [Theodoridis and Koutroumbas, 1999]
- Distância de Bhattacharyya [Theodoridis and Koutroumbas, 1999]
- Coeficiente de determinação (CoD) [Dougherty et al., 2000]
- Entropia condicional média [Lin, 1991, Martins-Jr et al., 2006, Montoya-Cubas et al., 2015]

Essas funções critério estimam a distribuição conjunta de um subconjunto de características, fazendo com que sejam suscetíveis ao fenômeno da curva em U, conforme discutido na Seção 2.3.1: para um número fixo de amostras, o aumento no número de características pode induzir a um aumento no erro de estimação. De fato, na prática o número de amostras disponíveis geralmente não é suficiente para realizar boas estimações, requerendo fatores de penalização ou maneiras de reduzir o número de parâmetros estatísticos a serem estimados. Por outro lado, o ruido nas amostras pode comprometer a estimação da dimensão (grau) ideal do subconjunto de características, induzindo a uma superestimação da dimensão como será discutido durante a apresentação dos resultados experimentais dos métodos de inferência considerados aplicados a dados simulados (Capítulo 4).

2.3.5 Entropia condicional média

A entropia condicional média tem sido aplicada com sucesso como função critério para seleção de características no contexto da inferência de redes gênicas [Barrera et al., 2007,

Lopes et al., 2008, Lopes et al., 2011, Lopes et al., 2014, Montoya-Cubas et al., 2015, Martins-Jr et al., 2016, Jacomini et al., 2017] A entropia mede o grau de desordem de uma variável, ou seja, quanto maior a entropia de uma variável, mais difícil prever o seu comportamento. A entropia de uma variável Y é dada por:

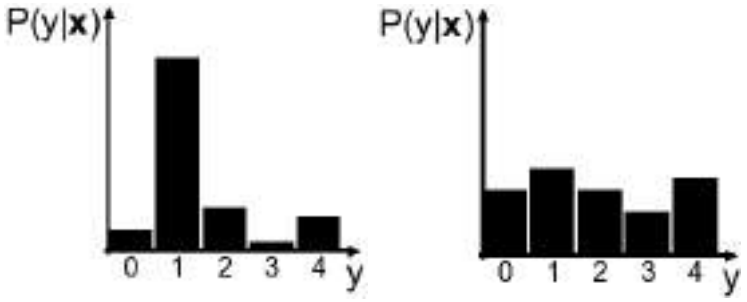
$$H(Y) = -\sum_{y \in Y} P(Y = y) log P(Y = y) \tag{2.1}$$

em que $P(Y = y)$ é a probabilidade de $Y = y$. Já a entropia condicional de uma variável Y dada uma instância $\mathbf{x} \in \mathbf{X}$ é definida por:

$$H(Y|\mathbf{X} = \mathbf{x}) = -\sum_{y \in Y} P(Y = y|\mathbf{X} = \mathbf{x}) log P(Y = y|\mathbf{X} = \mathbf{x}) \tag{2.2}$$

em que $P(Y = y|\mathbf{X} = \mathbf{x})$ é a probabilidade condicional de $Y = y$ dado que $\mathbf{X} = \mathbf{x}$.

A entropia condicional diz o quanto uma instância $\mathbf{x} \in \mathbf{X}$ consegue predizer o comportamento da variável Y. Quanto menor a entropia condicional de Y dado $\mathbf{X} = \mathbf{x}$, melhor será a predição de Y através de $\mathbf{X} = \mathbf{x}$. A Figura 2.5 ilustra dois histogramas, um para baixa entropia condicional e outro para alta entropia condicional de Y dado $\mathbf{X} = \mathbf{x}$. O histograma da esquerda (baixa entropia condicional) configura o caso em que $\mathbf{X} = \mathbf{x}$ realiza uma boa predição dos valores de Y, já que a distribuição de probabilidades condicionais apresenta massa altamente concentrada sobre $Y = 1$. Já o histograma da esquerda (alta entropia condicional) representa o caso em que $\mathbf{X} = \mathbf{x}$ não é adequado para predizer o comportamento de Y, pois a distribuição de probabilidades condicionais é próxima da uniforme.



$H(Y|\mathbf{X} = \mathbf{x})$ baixa $H(Y|\mathbf{X} = \mathbf{x})$ alta

Figura 2.5: O histograma da esquerda configura uma situação em que Y é bem predito por $\mathbf{X} = \mathbf{x}$ porque a massa de probabilidades condicionais está bem concentrada em $Y = 1$ (entropia condicional baixa). Já para o histograma da direita, a massa de probabilidades está melhor distribuída ao longo das classes, o que faz com que o padrão $\mathbf{X} = \mathbf{x}$ não seja um bom preditor de Y (entropia condicional alta). (Fonte: [Martins-Jr., 2008]).

A entropia condicional média é definida como a média ponderada das entropias con-

dicionais de todas as possíveis instâncias $\mathbf{x} \in \mathbf{X}$. Sua equação é dada por:

$$H(Y|\mathbf{X}) = \sum_{\mathbf{x}\in\mathbf{X}} P(\mathbf{X} = \mathbf{x})H(Y|\mathbf{X} = \mathbf{x}) \tag{2.3}$$

em que $H(Y|\mathbf{X} = \mathbf{x})$ é a entropia condicional dada pela Equação 2.2. Quanto menor o valor de $H(Y|\mathbf{X})$, maior será o ganho de informação sobre Y através do conjunto de características $\mathbf{X}$. Em base a este resultado podemos calcular a informação mútua média como a diferença da entropia *a priori* da entropia condicional média, ou seja:

$$IM(\mathbf{X}, Y) = H(Y) - H(Y|\mathbf{X}) \tag{2.4}$$

2.3.6 Informação mútua normalizada por dimensão

A informação mútua definida na Equação 2.4 quantifica a informação que um vetor de variáveis aleatórias $\mathbf{X}$ contém sobre a variável aleatória Y. Embora esta medida seja adequada para quantificar o grau de relação entre duas variáveis, um problema é que ela não leva em conta a dimensão de $\mathbf{X}$. Normalizar a informação mútua pela dimensão do subconjunto de características pode ajudar na estimação da dimensão ideal do subconjunto de características resultante de um processo de seleção de características, sendo esta então uma função critério proposta nesta tese.

Com o objetivo de quantificar o ganho de informação que se obtém ao passar do melhor subconjunto de dimensão $d-1$ ($\mathbf{X_{d-1}}$) obtido por um dado algoritmo de seleção de características para o melhor subconjunto de dimensão d ($\mathbf{X_d}$) pelo mesmo algoritmo, propomos a informação mútua normalizada pela dimensão (IM_d) definida pela Equação 2.5:

$$IM_d(\mathbf{X_d}, Y) = \frac{H(Y|\mathbf{X_{d-1}}) - H(Y|\mathbf{X_d})}{H(Y|\mathbf{X_{d-1}})}, \forall d \geq 1 \tag{2.5}$$

em que $H(Y|\mathbf{X_d})$ é a entropia condicional media do melhor subconjunto de características de dimensão d resultante de um dado algoritmo de busca considerado. Vale notar que $0 \leq IM_d \leq 1$, tendo em vista que $H(Y|\mathbf{X_{d-1}}) \geq H(Y|\mathbf{X_d})$.

Um caso especial a ser considerado é quando $d = 0$. Para esse caso, $IM_d(\mathbf{X_0}, \mathbf{Y}) = H(Y|\mathbf{X_0}) = H(Y)$, pois a entropia condicional de Y dado um conjunto vazio de características é a própria entropia *a priori* de Y ($H(Y)$).

2.3.7 Classificação por k-vizinhos mais próximos (k-nn)

Para melhorar a estimação da dimensão ideal de um subconjunto de gene preditores que melhor classificam um determinado gene alvo, adotamos o método k-nn para classificar perfis de evolução da função critério ao longo das dimensões geradas por um dado algoritmo de seleção de características. Tais perfis são obtidos através da aplicação de um dado algoritmo de seleção de características que toma como entrada amostras de expressão gênica temporais geradas por redes simuladas, cujos rótulos são justamente as dimensões dos subconjuntos de preditores dadas por essas redes (ver Seção 5.1). Sendo assim, esta seção apresenta uma breve descrição do método k-nn.

O método k - vizinhos mais próximos (do ingles *k-nearest neighbors*, abreviado *k*-nn) [Cover and Hart, 1967] é um método de classificação supervisionado que serve para estimar a função de densidade $F(\mathbf{X}|Y_j)$ dos preditores $\mathbf{X}$ para cada classe Y_j.

Este é um método de classificação não paramétrico, que estima o valor da função densidade de probabilidade ou diretamente a probabilidade *a posteriori* de que uma amostra $\mathbf{X}$ pertença à classe Y_j a partir das informações fornecidas pelo conjunto de amostras. Nenhuma suposição é feita no processo de aprendizado sobre a distribuição das variáveis preditivas.

As amostras de treinamento são vetores de características $\mathbf{X} = (x_1, x_2, ..., x_n)$, descritos em termos de n atributos considerando c classes para sua classificação. Os valores da i-ésima amostra $(1 \leq i \leq m)$ são representados pelo vetor n-dimensional:

$$\mathbf{X_i} = (x_{1i}, x_{2i}, ..., x_{ni})$$

O espaço é particionado em regiões por locais e rótulos das amostras de treinamento. Um ponto no espaço é atribuído à classe Y_i se esta for a classe mais frequente entre as k amostras de treinamento mais próximas. Geralmente a distância euclidiana, dada pela equação a seguir, é adotada.

$$d(\mathbf{X_i}, \mathbf{X_j}) = \sqrt{\sum_{r=1}^{n} (x_{ri} - x_{rj})^2}$$

A fase de treinamento do algoritmo consiste em armazenar os vetores de características e os rótulos das classes das amostras de treinamento. Na fase de classificação, uma nova amostra (cuja classe é desconhecida) a ser classificada é representada por um vetor de características. A distância entre os vetores armazenados e o novo vetor é calculada e os k vetores mais próximos são selecionados. O novo vetor é classificado com a classe mais frequente dos vetores selecionados.

O método k-nn pode ser sumarizado em dois algoritmos:

- **Algoritmo de treinamento:** Para cada amostra $\langle \mathbf{X}, f(\mathbf{X}) \rangle$, adicione o amostra à estrutura que representa as amostras de treinamento.
- **Algoritmo de classificação**: Dada uma amostra $\mathbf{X_i}$ que deve ser classificada, sejam $\mathbf{X_1}, ..., \mathbf{X_k}$ os k vizinhos mais próximos de $\mathbf{X_i}$ nas amostras de treinamento. Devolva:

$$\hat{f}(\mathbf{X_i}) \leftarrow \arg\max_{y \in Y} \sum_{i=1}^{k} [y = f(\mathbf{X_i})]$$

em que $[y = f(\mathbf{X_i})] = 1$ e $[y \neq f(\mathbf{X_i})] = 0$.

O valor $\hat{f}(\mathbf{X_i})$ devolvido pelo algoritmo como um estimador de $\hat{f}(\mathbf{X_i})$ é apenas o valor mais comum de f entre os k vizinhos mais próximos de $\mathbf{X_i}$.

2.4 Modelos de topologias de redes complexas

A teoria de redes complexas estende o formalismo da teoria dos grafos por acrescentar medidas e métodos fundamentados em propriedades reais de um sistema [Costa et al., 2008]. Os modelos de redes complexas apresentam topologias distintas e propriedades bem definidas, as quais podem ser usadas para representar redes gênicas, bem como caracterizá-las em termos de medidas de redes complexas. Desta forma, a teoria de redes complexas permite a caracterização, análise e representação dos mais variados sistemas complexos, como por exemplo sistemas biológicos [Kauffman, 1993, Jeong et al., 2000, Guelzim et al., 2002, Farkas et al., 2003, Przulj et al., 2004, Albert, 2005, Costa et al., 2008, Narasinham et al., 2009, Barabasi, 2009].

O primeiro modelo de redes complexas foi o de redes aleatórias, proposto por Paul Erdös e Alfréd Rényi em 1959 [Erdös and Rényi, 1959]. Desde então, outros modelos de redes complexas foram propostos para a representação de sistemas reais, tais como: livre de escala (*scale-free*) [Barabási and Albert, 1999], geométrico [Przulj et al., 2004], mundo pequeno (*small-world*) [Watts and Strogatz, 1998] e geográfico [Gastner and Newman, 2006].

A seguir serão apresentados os modelos aleatório e livres de escala, os quais foram empregados na geração das redes gabarito simuladas nos experimentos do Capítulo 4.

2.4.1 Redes aleatórias

No modelo de redes aleatórias de Erdös-Rényi (ER) [Erdös and Rényi, 1959] cada par de vértices possui uma probabilidade p de possuir uma aresta entre eles. Assim, o grau médio de cada nó é dado por $\langle k \rangle = p(n-1)$, sendo n o número de vértices do grafo.

A distribuição do número de conexões por vértices $P(k)$ em um grafo gerado pelo modelo aleatório aproxima-se por uma distribuição de Poisson [Costa et al., 2008], dada por:

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \tag{2.6}$$

O modelo ER apresenta um padrão de conexões aleatórias contendo um número de conexões k similar entre seus vértices, ou seja, a maioria dos vértices terão um grau k próximo da média $\langle k \rangle$. Parte dos experimentos do Capítulo 4 foram gerados a partir de redes ER com grau médio $\langle k \rangle = 3$.

2.4.2 Redes livres de escala

Barabási e Albert [Barabási and Albert, 1999], procurando entender a dinâmica e a estabilidade topológica de grandes redes reais, perceberam que em muitos sistemas, a probabilidade $P(k)$ de um vértice da rede interagir com k outros vértices decai como uma lei de potência, na forma:

$$P(k) \sim k^{-\gamma} \tag{2.7}$$

em que o parâmetro γ é uma constante de decaimento.

Esse modelo tem sido utilizado para simular e descrever o comportamento das redes de interação gênica [Jeong et al., 2000, Guelzim et al., 2002, Farkas et al., 2003, Albert, 2005, Costa et al., 2008, Barabasi, 2009]. Com relação à constante γ, diversos trabalhos verificaram que redes biológicas seguem uma lei de potência com $2 < \gamma < 3$ [Jeong et al., 2000, Albert, 2005, Lopes et al., 2014].

As redes livres de escala (*scale-free*) não apresentam uma distribuição homogênea de conexões entre seus vértices, apresentando poucos vértices altamente conectados a outros vértices da rede, e um grande número de vértices com poucas conexões [Costa et al., 2008]. Esses vértices altamente conectados são chamados de *hubs*.

O modelo de construção de redes livres de escala proposto por Barabási e Albert (BA) [Barabási and Albert, 1999] é baseado em duas regras: crescimento e preferência linear de

ligação. A geração de redes BA é iniciada com a inclusão de $n_0 < n$ vértices conectados aleatoriamente, em geral usando o modelo ER apresentado na seção anterior.

Na etapa de crescimento da rede, em cada iteração $t = 1, 2, ..., n - n_0$, um novo vértice v_i contendo $\langle k \rangle \leq n_0$ arestas é adicionado na rede, seguindo uma preferência linear de ligação. Ou seja, a probabilidade de um vértice v_j já existente na rede ser conectado ao novo vértice v_i é linearmente proporcional ao grau k_j do vértice v_j tal que:

$$P(v_i \leftrightarrow v_j) = \frac{k_j}{\sum_u k_u}, \forall v_u \in V \tag{2.8}$$

em que V é o conjunto de vértices do grafo. Essa preferência de ligação pelos vértices mais conectados resulta no fenômeno "rico fica mais rico".

2.5 Validação da inferência de redes gênicas

Uma vez obtida a rede inferida, há duas formas para avaliar a qualidade da rede: avaliação topológica e avaliação da dinâmica [Dougherty, 2011]. No caso da avaliação topológica, compara-se a rede inferida com a rede modelo (ou gabarito), contando quantas arestas da rede gabarito foram recuperadas (verdadeiros positivos, *true positive* – TN), quantas delas não foram recuperadas (falsos negativos – FN), quantas arestas foram obtidas e que não estão presentes na rede gabarito (falsos positivos – FP), e finalmente quantas arestas não foram obtidas e que não estão presentes na rede gabarito (verdadeiros negativos, *true negative* – TN).

Redes gênicas são caracterizadas por serem esparsas, ou seja, a matriz de adjacências do grafo de uma rede gênica possui um número muito maior de zeros do que de uns (há muito mais "não-arestas" do que arestas). Nessa situação, a proporção de negativos acaba respondendo por quase 100% do número de pares de nós do grafo. Sendo assim, não faz muito sentido levar em conta a taxa TN na avaliação. Uma boa métrica nessa situação é o F-SCORE, pois ela avalia o balanço entre as taxas TP, FP e FN, sem levar em conta a taxa TN, sendo definida pela média harmônica ponderada entre a precisão e a sensibilidade, conforme a Equação 2.9:

$$F - SCORE = \frac{2TP}{2TP + FP + FN} \tag{2.9}$$

Mesmo que uma rede não tenha bom escore topológico, ainda assim é possível que ela gere uma dinâmica de expressão similar aquela gerada pela rede gabarito, pois é possível que múltiplas redes possam explicar o mesmo conjunto de dados. Então também é necessário examinar o poder de explicação dos dados de uma rede inferida. Como esta

proposta se concentra no modelo de redes Booleanas, uma maneira natural de avaliar a dinâmica da rede inferida é por meio da distância de Hamming normalizada, que é definida pelo número de bits de diferença entre o sinal ideal e o sinal inferido dividido pelo tamanho do sinal. Assim, quanto menor a distância de Hamming normalizada, melhor será o sinal inferido.

Capítulo 3

Agrupamento em classes de equivalência

3.1 Agrupamento como um problema de busca no espaço de partições

Conforme discutido na Seção 2.3.1, a chamada "maldição da dimensionalidade" [Bishop, 2006] é o fenômeno no qual o número de amostras de treinamento requeridas para uma classificação ou predição satisfatória de uma variável alvo é dada por uma função exponencial da dimensão do espaço de características. Esse problema consiste na divisão do espaço de características, observando as amostras disponíveis no conjunto de treinamento. Se, por exemplo, cada característica X_i for dividida em M divisões (classes), cada uma delas associada a uma determinada classe Y_j, então o numero total de divisões é M^d a qual passa a crescer exponencialmente com a dimensionalidade do espaço de características. O aumento no número de divisões no espaço de características pode aumentar a precisão com o qual cada objeto é especificado, no entanto, cada divisão deve conter pelo menos uma amostra, assim a quantidade de amostras necessárias para o treinamento cresce exponencialmente ao mesmo tempo que cresse o numero de divisões.

Estimar uma função booleana de k variáveis a partir de dados experimentais requer estimar 2^k parâmetros: para cada uma das 2^k possíveis instâncias das variáveis de entrada, é necessário dizer se a saída deve ser 0 ou 1. Mesmo para valores moderados de k, o número de parâmetros tende a ser muito maior do que o número de amostras. Nesta situação, a maior parte das instâncias da entrada nunca são observadas em dados experimentais e, mesmo para uma instância observada, o número de ocorrências pode não ser suficiente para uma estimação confiável do valor de saída. Nesta proposta serão avaliadas abordagens para reduzir o número de instâncias realizando agrupamentos de instâncias. Cada modo

de agrupamento corresponde a um jeito de particionar o conjunto de instâncias, Uma partição de um conjunto é um agrupamento dos elementos do conjunto em subconjuntos não vazios do conjunto original, de tal forma que cada elemento pertença a um único subconjunto.

O número total de partições de um conjunto de n elementos é dado pelo número de Bell B_n, o qual satisfaz a seguinte recursão [Wilf, 1994]:

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k \tag{3.1}$$

Por exemplo, os primeiros 6 números de Bell são: $B_0 = 1$, $B_1 = 1$, $B_2 = 2$, $B_3 = 5$, $B_4 = 15$, $B_5 = 52$ e $B_6 = 203$. Os números de Bell crescem como uma função exponencial de n, o que indica que o número de partições possíveis de um conjunto de n elementos cresce exponencialmente. Adicionalmente, é importante notar que o número n de interesse aqui é relacionado com o número de instâncias de um conjunto de preditores, o qual cresce exponencialmente com a dimensão d. Ou seja, considerando uma rede Booleana, um conjunto de d preditores possui 2^d possíveis instâncias. Então como o número de partições n cresce exponencialmente em função do número de instâncias, o número de possíveis partições de um conjunto de instâncias cresce super-exponencialmente em função do número de preditores considerados. A Tabela 3.1 ilustra a velocidade de crescimento do número de partições em função da dimensão do conjunto de preditores. Note que para apenas 4 preditores, o número de partições já é da ordem de 10^{10}, enquanto que para 5 preditores, esse numero explode para aproximadamente 10^{26}.

Tabela 3.1: Ilustração do crescimento do número de partições em função do número de preditores considerando preditores binários.

# preditores	# instâncias	# partições
2	$2^2 = 4$	15
3	$2^3 = 8$	4.140
4	$2^4 = 16$	10.480.142.147
5	$2^5 = 32$	128.064.670.052.407.582.646.900.609

Portanto, uma busca exaustiva no espaço de partições se torna impraticável para conjuntos de preditores de dimensão 4 em diante. Assim, é importante desenvolver estratégias de busca no espaço de partições que encontrem boas partições mesmo percorrendo uma ínfima parte do espaço.

Nesse sentido, o espaço de partições pode ser estruturado em um reticulado de partições (*partition lattice*), impondo uma ordem parcial entre as partições que possa auxiliar no desenvolvimento de estratégias de busca. Essa ordem parcial é obtida a partir de duas operações: união de dois subconjuntos e desmembramento de um subconjunto em dois.

Dessa forma, uma dada partição é imediatamente anterior a outras partições se estas forem resultantes da união de dois subconjuntos da partição original. De modo análogo, uma dada partição é imediatamente posterior a outras partições se estas forem resultantes do desmembramento de um subconjunto da partição original em dois. A Figura 3.1 ilustra um reticulado de partições de um conjunto de 4 elementos, formando então 15 possíveis partições conforme o número de Bell para $n = 4$.

Fazendo uma correspondência do reticulado de partições da Figura 3.1 com as 4 instâncias de 2 preditores binários $(0,0),(0,1),(1,0),(1,1)$, a partição inferior corresponde às instâncias originais $\{(0,0)\},\{(0,1)\},\{(1,0)\},\{(1,1)\}$. A partição superior corresponde a um único conjunto com todas as instâncias agrupadas $\{(0,0),(0,1),(1,0),(1,1)\}$. A partição em formato de um "X" corresponde a uma das partições com 2 grupos de 2 instâncias cada $\{(0,0),(1,1)\},\{(0,1),(1,0)\}$.

A abordagem de agrupamento linear desenvolvida durante o mestrado percorre uma parte bastante restrita do espaço de partições, conforme será discutido logo a seguir, na Seção 3.2

3.2 Agrupamento linear

Com base na hipótese que grande parte das funções nas redes de regulação genica apresentam um comportamento linearmente separável *(threshold functions)* como foi estudado em [Tran et al., 2013], este método de agrupamento visa priorizar este tipo de funções construindo grupos de tal modo que a maioria das funções geradas sejam linearmente separáveis, ao mesmo tempo reduzindo o número de parâmetros a serem estimados.

Para a estimação das distribuições de probabilidades condicionais $P(Y|\mathbf{Z})$, onde Y é uma variável binária e $\mathbf{Z} \subseteq \mathbf{X}$ é um vetor binário em $\{0,1\}^k$, o método de agrupamento linear consiste no mapeamento linear das instâncias de entrada $\mathbf{Z} \in \{0,1\}^k$ em um número inteiro L, conforme a Equação 3.2:

$$L = a_1 Z_1 + a_2 Z_2 + \ldots + a_k Z_k \tag{3.2}$$

em que $a_i \in \mathcal{C}$, e $\mathcal{C} \subset \mathbb{Z}$ que representa os possíveis pesos que o gene i pode assumir, para $i \in \{1,2,\ldots,k\}$. Podemos definir um vetor de coeficientes $\mathbf{A} = \{a_1, a_2, ..., a_k\} \in \mathcal{C}^k$ reescrevendo a Equação 3.2 em formato vetorial: $L = \mathbf{A}^T\mathbf{Z}$, sendo l o número de valores que L pode assumir, o que é equivalente ao numero de classes de equivalência. Nessa modelagem, supõe-se que um preditor possa ser um ativador (caso seu coeficiente seja maior que 0) ou um inibidor (caso seu coeficiente seja menor que 0) do gene alvo. Como cada coeficiente a_i pode assumir um dos valores de $\mathcal{C}$, sendo c o numero de elementos

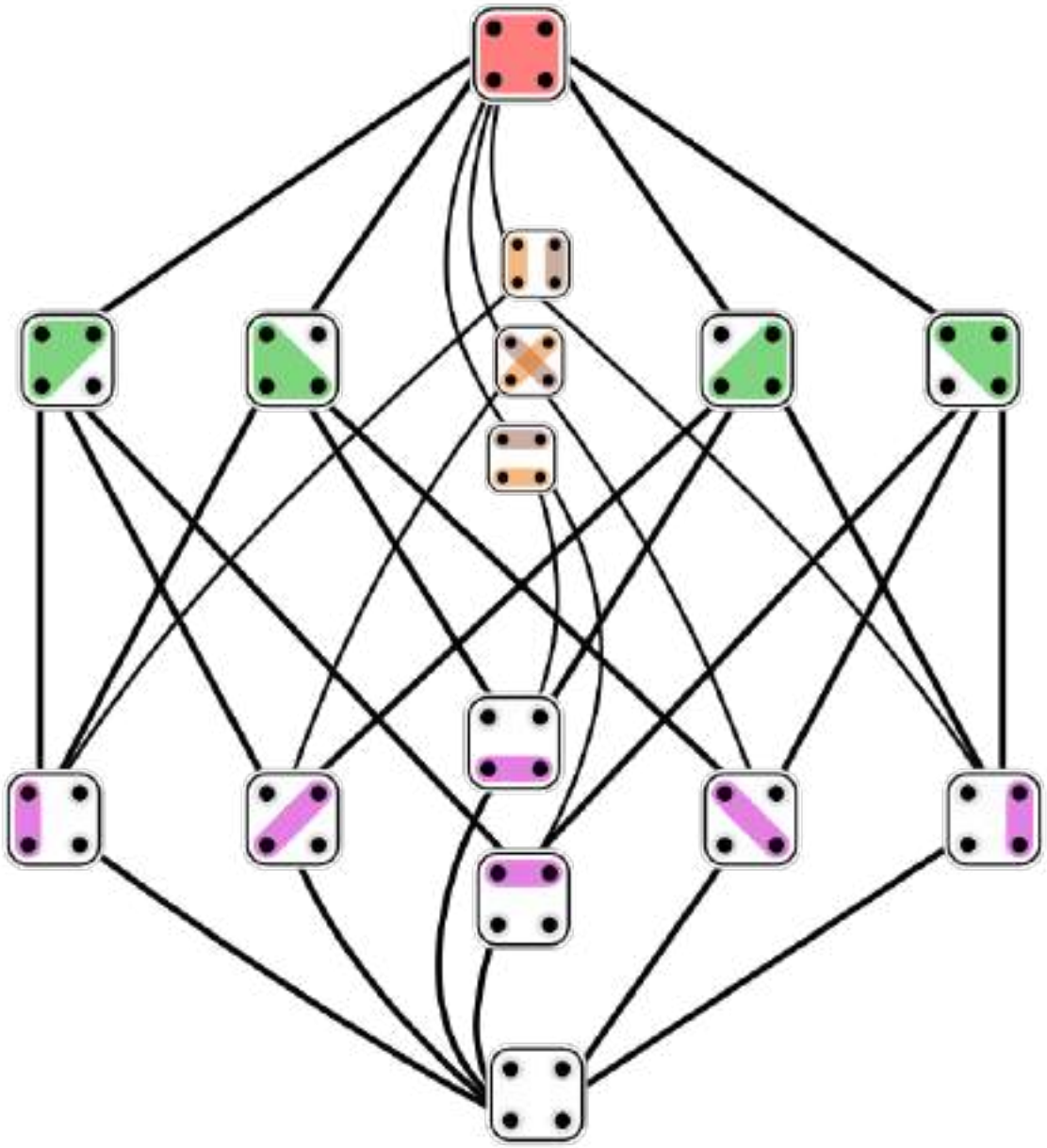


Figura 3.1: Reticulado de partições para um conjunto de 4 elementos. Duas partições são ligadas por uma aresta se uma delas é a união de dois subconjuntos da outra partição. A partição inferior corresponde a todos os conjuntos unitários, enquanto a partição superior corresponde a um único conjunto com todos os elementos. Fonte: [Commons, 2014]

de $\mathcal{C}$, existem c^k combinações lineares possíveis a serem avaliadas para cada conjunto de k preditores. Entretanto uma mesma classe de equivalência pode ser resultante de mais de uma combinação linear. Para um determinado conjunto de preditores, adota-se a configuração de coeficientes $\mathbf{A}^*$ que resulta no melhor valor de função critério. A

comparação entre diferentes conjuntos preditores é baseada no valor da função critério para a melhor combinação linear de cada conjunto.

O resultado acima implica que as 2^k instâncias do vetor original de preditores são mapeadas para l classes de equivalência, de acordo com os valores resultantes de uma dada combinação linear dos valores dos preditores. No método proposto, a estimação direta de $P(Y|\mathbf{Z})$ é substituída pela estimação de $P(Y|L)$. Desse modo o número de parâmetros a serem estimados a partir do conjunto de amostras torna-se bastante reduzido. Em seguida, pode-se aplicar qualquer função critério para a avaliação dos preditores, por exemplo, a entropia condicional média (Equação 2.3), utilizando-se $P(Y|L)$ em lugar de $P(Y|\mathbf{Z})$.

Para ilustrar o método, consideremos a tarefa de estimar o valor de $Y \in \{0, 1\}$ a partir das variáveis $Z_1, Z_2, Z_3 \in \{0, 1\}^3$. Isto leva a $2^3 = 8$ instâncias distintas, o que implica na necessidade de estimar $P(y|z_1, z_2, z_3)$ para 8 instâncias de z_1, z_2, z_3 e dois valores de y.

Tomando como exemplo $\mathcal{C} = \{-1, 1\}$, existem $2^3 = 8$ combinações lineares possíveis, mas apenas 4 diferentes combinações lineares precisam ser investigadas: $(-1, -1, -1)$, $(-1, -1, 1)$, $(-1, 1, -1)$, $(-1, 1, 1)$. As outras 4 combinações restantes geram exatamente as mesmas partições como já discutido anteriormente. A Figura 3.2 ilustra graficamente esses agrupamentos no reticulado Booleano, cujos nós representam as possíveis instâncias das 3 variáveis binárias, conforme explicado previamente na Seção 2.3.2.

O Algoritmo 1 apresenta o processo de inferência de redes gênicas pelo agrupamento linear.

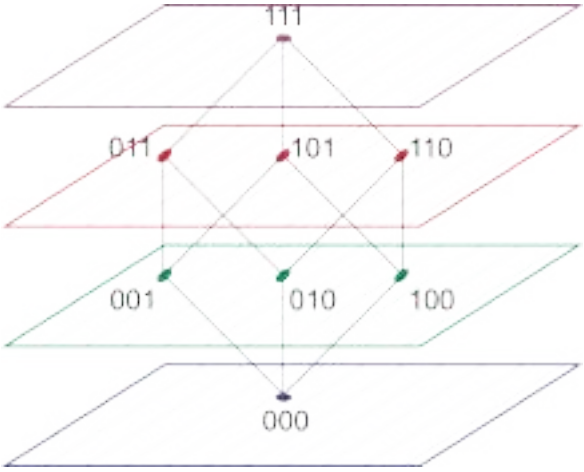
Algorithm 1 Agrupamento Linear

Require: Conjunto de preditores candidatos $\mathbf{Z} = \{z_1, \ldots, z_k\}$, gene alvo Y, matriz de expressão gênica com M amostras temporais e N genes, e os 2^{k-1} possíveis agrupamentos.

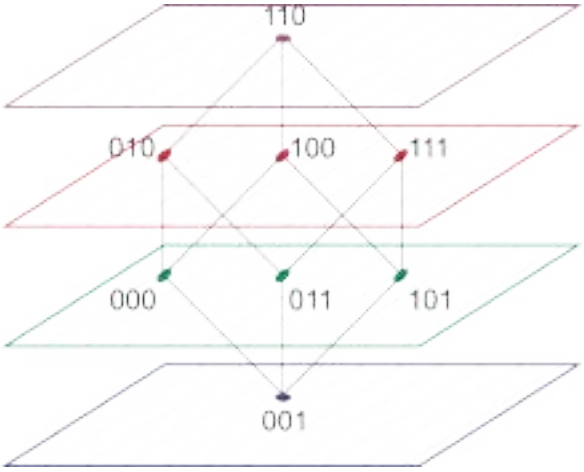
Ensure: Entropia condicional média de Y dado $\mathbf{Z} = \{z_1, \ldots, z_k\}$.

1: $ECM_{min} \leftarrow \infty$
2: **for** cada possível agrupamento definido por uma instância do vetor de coeficientes lineares **A do**
3: Preencher a tabela TPC de probabilidades condicionais de acordo com a matriz de expressão gênica, os valores de $Y[t+1]$, e os valores resultantes das combinações lineares dos valores de $\mathbf{Z}[t]$ para todo $1 \leq t \leq M-1$, e dos coeficientes lineares.
4: Calcular a entropia condicional média ECM com base na tabela TPC
5: **if** $ECM < ECM_{min}$ **then**
6: $ECM_{min} \leftarrow ECM$
7: **end if**
8: **end for**
9: **return** ECM_{min}

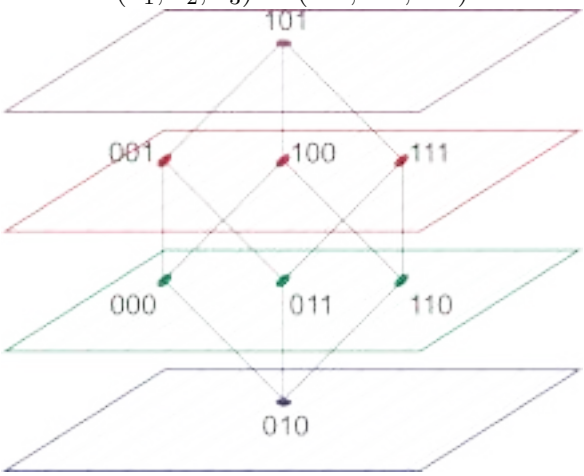
Ao assumir uma relação linear entre os genes e $\mathcal{C} = \{-1, 1\}$, essa estratégia considera 2^{k-1} partições possíveis do total do espaço de partições. Como o espaço de partições



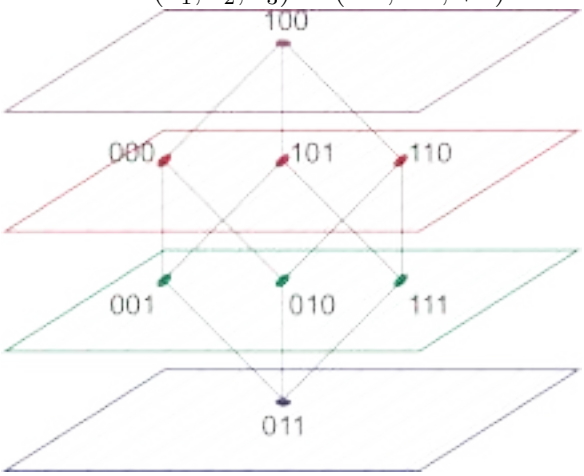
$(a_1, a_2, a_3) = (-1, -1, -1)$



$(a_1, a_2, a_3) = (-1, -1, +1)$



$(a_1, a_2, a_3) = (-1, +1, -1)$



$(a_1, a_2, a_3) = (-1, +1, +1)$

Figura 3.2: Particionamentos no reticulado Booleano para os coeficientes lineares $(a_1, a_2, a_3) = \{(-1, -1, -1); (-1, -1, +1); (-1, +1, -1); (-1, +1, +1)\}$.

possíveis cresce de forma super-exponencial em função de k (número de preditores), normalmente essa forma de agrupamento considera um conjunto de partições bastante restrito do espaço total de partições. Por exemplo, para $k = 3$, o agrupamento linear considera $2^{k-1} = 4$ partições possíveis, conforme ilustrado na Figura 3.2. Entretanto, o número de possíveis partições para $k = 3$ é dado pelo 8º número de Bell ($2^3 = 8$ instâncias possíveis): 4.140 (Tabela 3.1).

Um problema do agrupamento linear é que suas partições consideram apenas grupos desbalanceados de instâncias para $k > 1$. Por exemplo, observe que para $k = 3$, há quatro grupos, sendo dois deles com uma instância cada, e outros dois com três instâncias cada, conforme ilustrado na Figura 3.2.

O método de agrupamento linear foi projetado com a intenção de priorizar as funções linearmente separáveis, porém o conjunto destas funções que têm prioridade depende diretamente dos possíveis valores de $\mathcal{C}$. Dessa forma, apenas um conjunto reduzido de funções linearmente separáveis são examinadas. O tamanho do espaço de busca da função

preditora cresce seguindo a relação 2^{2^d}, em que d é a dimensão do conjunto de preditores, contudo apenas um subconjunto destas funções é linearmente separável. Assim, este método aproveita esse fato para priorizar este subconjunto de funções.

3.3 Agrupamento por busca sequencial para frente no reticulado de partições

Esta é uma primeira estrategia proposta para a busca no reticulado de partições que não assume qualquer hipótese sobre a regulação entre os genes, diferentemente do que ocorre com as estratégias de agrupamento linear apresentadas anteriormente, as quais assumem a hipótese de que os genes não regulados por uma combinação linear entre eles. Basicamente, a abordagem aqui proposta é gulosa do tipo Busca Sequencial para Frente (*Sequential Forward Search* – SFS) [Pudil et al., 1994] sobre o reticulado de partições. Inicialmente, examina-se a partição original das instâncias, sem agrupamento. A partir daí, une-se as duas instâncias com o menor número de amostras observadas (em caso de empates, une-se o par de instâncias que resulte no melhor valor da função critério adotada), gerando uma nova partição com $n-1$ grupos, sendo 1 grupo com duas instâncias e as instâncias restantes formando grupos unitários. Em seguida, repete-se o mesmo raciocínio, unindo-se os dois grupos com menor número de amostras em um único grupo. A busca termina quando todos os grupos possuírem um número mínimo pré-determinado de amostras. Esse procedimento está mais precisamente descrito no Algoritmo 2. A Figura 3.3 ilustra um exemplo de busca sequencial para frente (SFS) no reticulado de partições com 4 instâncias.

Algorithm 2 Agrupamento por Busca Sequencial para Frente no Reticulado de Partições

Require: Conjunto de preditores candidatos $\mathbf{Z} = \{z_1, \ldots, z_k\}$, gene alvo Y, matriz de expressão gênica com M amostras temporais e N genes, e um limiar θ de número máximo de observações que uma instância ou um grupo de instâncias deve possuir para que ele seja considerado pouco observado.

1: Preencher a tabela de contagem de observações das instâncias dos preditores de acordo com a matriz de expressão gênica, os valores de $Y[t+1]$, e os valores de $\mathbf{Z}[t]$ para todo $1 \leq t \leq M-1$.
2: **while** Pelo menos um grupo não tiver mais do que θ observações **do**
3: Unir o grupo com menos observações com aquele grupo que for o segundo menos observado. Em caso de empates, escolha aquele que minimiza a entropia condicional média (ECM);
4: **end while**
5: **return** ECM do agrupamento resultante

É importante observar que a quantidade de partições examinadas por essa busca é ao mesmo tempo significativamente maior do que o número de partições examinadas pelo agrupamento linear, mas significativamente menor do que o espaço total de partições.

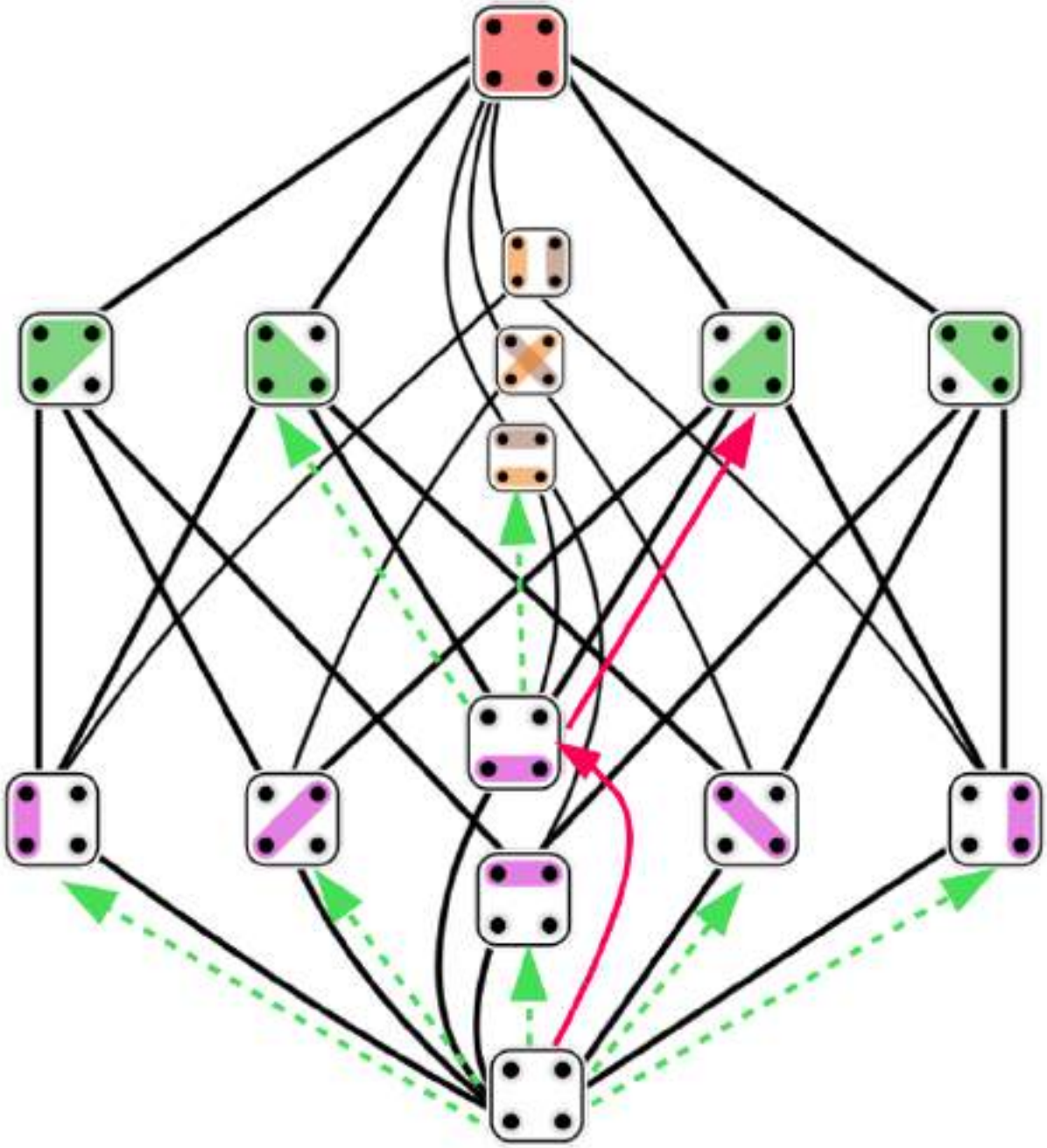


Figura 3.3: Exemplo de busca sequencial para frente (SFS) no reticulado de partições para um conjunto de 4 instâncias. As setas verdes tracejadas indicam as partições examinadas pela busca, enquanto as setas vermelhas indicam as partições selecionadas em cada passo.

Considerando um conjunto de n instâncias, as partições examinadas são restritas a, no máximo:

- Partição original (n grupos unitários);
- $\binom{n}{2}$ partições com $n-2$ grupos unitários e 1 grupo com 2 instâncias;
- $\binom{n-1}{2}$ partições com $n-2$ grupos;

- $\binom{n-2}{2}$ partições com $n-3$ grupos;
- e assim por diante, até formar um único grupo contendo todas as instâncias.

Ou seja, o número mínimo de partições examinadas equivale a $\binom{n}{2}+1$ partições, caso nenhuma partição com um único grupo de 2 elementos seja melhor do que a partição original de acordo com a função critério adotada (neste caso, a partição original é devolvida). Já o número máximo de partições examinadas equivale a:

$$\sum_{i=0}^{n-2}\binom{n-i}{2}+1 \leq n \times \binom{n}{2} \leq n^3$$

o que significa que o número de partições examinadas é algo entre ordem de n^2 e ordem de n^3. Já no agrupamento linear, o número de partições examinadas é da ordem de n. Por exemplo, para $n=8$ ($k=3$), o número de partições examinadas pelo agrupamento SFS é de 84, enquanto o número de partições examinadas pelo agrupamento linear é de 4. Lembrando que o número total de partições para $n=8$ é de 4.140 conforme previamente discutido. Portanto, o agrupamento por SFS examina um número polinomial de partições significativamente maior que o agrupamento linear, mas significativamente menor do que o número de partições possíveis. Ou seja, o agrupamento por SFS apresenta um bom balanço entre espaço de partições percorrido e eficiência computacional.

3.4 Agrupamento por canalização

No agrupamento por canalização os grupos são formados de tal modo que representem funções canalizadoras (Seção 2.2.4). Ou seja, para cada preditor X_i em um agrupamento, ficarão em um grupo todas as instâncias com $X_i=0$ e ficarão separadas aquelas instâncias com $X_i=1$. Em outro agrupamento, ficarão juntas as instâncias com $X_i=1$ e separadas as instâncias com $X_i=0$. Neste método examina-se a possibilidade de cada gene preditor ser um possível canalizador. Assim em um grupo de k preditores serão examinadas duas formas de agrupamento para cada um dos k (para $X_i=0$ e $X_i=1$), gerando no total $2n$ formas diferentes de agrupar as instâncias. O agrupamento com o melhor valor de função critério. O Algoritmo 3 apresenta esse procedimento.

Para ilustrar este procedimento a Tabela 3.2 apresenta um exemplo de canalização em que todas as instâncias com $x_1=0$ ficam em um mesmo grupo, enquanto as instâncias com $x_1=1$ ficam insoladas.

Algorithm 3 Agrupamento por Canalização

Require: Conjunto de preditores candidatos $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$, gene alvo Y, matriz de expressão gênica com M amostras temporais e N genes.

Ensure: Entropia condicional média de Y dado $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$ após o agrupamento por canalização das instâncias de $\mathbf{Z}$

1: $ECM_{min} \leftarrow \infty$;
2: **for** cada gene preditor $Z_i \in \mathbf{Z}$ **do**
3: Preencher a tabela TF de frequências (contagens) dos valores de Y dadas todas as possíveis instâncias de $\mathbf{Z}$ considerando $\mathbf{Z}[t]$ e $Y[t+1]$ para todo $1 \leq t \leq M-1$;
4: $TF_0 \leftarrow$ Agrupar em apenas uma instância as instâncias com $Z_i = 0$ de TF;
5: $TF_1 \leftarrow$ Agrupar em apenas uma instância as instâncias com $Z_i = 0$ de TF;
6: Calcular as entropias condicionais médias ECM_0 e ECM_1 com base na tabelas TPC_0 e TPC_1, respectivamente (TPC é a tabela de probabilidades condicionais obtida com base na tabela de frequências TF);
7: $ECM \leftarrow \min(ECM_0, ECM_1)$
8: **if** $ECM < ECM_{min}$ **then**
9: $ECM_{min} \leftarrow ECM$
10: **end if**
11: **end for**
12: **return** ECM_{min}

3.5 Transferência de aprendizado supervisionado dos graus sobre redes gênicas artificiais

Um outro problema com que se tem que lidar na inferência de redes de regulação genica é a estimação do grau (dimensão) dos preditores para um dado gene alvo. Algoritmos de busca como a busca sequencial para frente ou a busca exaustiva incremental precisam de um critério de parada para deixar de prosseguir a busca pelo melhor subconjunto em dimensões superiores. Trabalhos como [Montoya-Cubas et al., 2015] utilizam como critério de parada o não incremento da função critério, isto é, parar de procurar pelo melhor subconjunto de preditores em dimensões superiores quando a função critério deixar de melhorar. Entretanto, em situações com poucas amostras, como é típico em dados de expressão gênica, dificilmente todos os preditores de um gene alvo que possui um grau relativamente alto serão recuperados com essa estratégia. Sendo assim, o ideal seria estimar corretamente a dimensão dos preditores, mesmo em situações com poucas amostras.

Neste contexto idealizamos uma estratégia baseada no aprendizado supervisionado das dimensões dos subconjuntos de preditores com base nos perfis de evolução da função critério ao longo das dimensões. Um perfil de evolução da função critério é resultante da aplicação de um dado algoritmo de seleção de características para dimensões $d = 1, 2, ..., k_{max}$, sendo que o valor da função critério em cada dimensão d é o melhor obtido pelo algoritmo para d. Para isso, naturalmente são necessários conjuntos de amostras temporais, de tal modo que a aplicação de uma algoritmo de seleção de características

Tabela 3.2: Esquerda: Tabela de frequências antes de aplicar o agrupamento por canalização; Direita: Tabela de frequências resultante da aplicação do agrupamento de canalização para $X_1 = 0$

X_1	X_2	X_3	$f(Y=0)$	$f(Y=1)$
0	**0**	**0**	3	0
0	**0**	**1**	0	0
0	**1**	**0**	1	0
0	**1**	**1**	0	0
1	0	0	2	1
1	0	1	0	4
1	1	0	0	1
1	1	1	1	5
Total			7	11

Grupos	$f(Y=0)$	$f(Y=1)$
000,001,010,011	4	0
1 0 0	2	1
1 0 1	0	4
1 1 0	0	1
1 1 1	1	5
Total	7	11

para um dado conjunto de amostras resulte em um perfil de evolução dos valores da função critério. Portanto, quanto mais conjuntos de amostras estiverem disponíveis, mais perfis de evolução serão gerados, consequentemente implicando em um melhor aprendizado das dimensões ideais.

Tendo em vista que o foco desta tese é no modelo de Redes Booleanas Probabilísticas (PBN), a partir de uma PBN é possivel gerar uma amostra temporal simplesmente através do sorteio de um estado qualquer e da geração do próximo estado aplicando as funções lógicas definidas para todos os genes em relação aos seus genes preditores. Portanto, para gerar um conjunto de M amostras temporais, basta aplicar esse procedimento M vezes. Vale notar que esse procedimento permite gerar diversos conjuntos de amostras temporais a partir de uma PBN.

Para sumarizar, a estratégia consiste então em:

1. **Geração de PBNs:** Gerar aleatoriamente diversas PBNs através de algum modelo topológico, como por exemplo o aleatório de Erdos-Renyi ou o livre de escala de Barabasi-Albert (Seção 2.4);

2. **Geração de conjuntos de amostras:** Para cada PBN gerada na etapa anterior, gerar um certo número de conjuntos de amostras de tamanho M, sendo que cada amostra consiste de um par $(S,\ \Phi(S))$ no qual S é um estado qualquer da rede sorteado, e $\Phi(S)$ é o próximo estado resultante da aplicação das funções Booleanas definidas aleatoriamente para cada gene da rede;

3. **Aplicação de um algoritmo de seleção de características:** Para cada conjunto de amostras gerado na etapa anterior, aplicar um algoritmo de seleção de características para cada gene Y, de forma a obter os valores dos melhores subconjuntos de preditores para cada dimensão $d = 1, 2, ..., d_{max}$, sendo um valor por dimensão. Tais valores comporão um perfil de evolução da função critério ao longo das dimensões.

Na etapa 3, um algoritmo de seleção de características tipicamente avalia diversos conjuntos de características com base em uma dada função critério, como por exemplo a Equação 2.5 que é baseada na entropia condicional, tomando como entrada uma tabela de probabilidades condicionais (TPC) de Y dadas todas as possíveis instâncias de $\mathbf{Z}$. Essa TPC é derivada de uma tabela de frequências (contagens), cuja construção é descrita a seguir.

Geração da tabela de frequências de Y dado Z: Dado um conjunto de amostras temporais (pares $(S, \Phi(S))$), para cada par $(S_i, \Phi(S_i))$, $i = 1, 2, ..., M$, sendo $\mathbf{Z_j} = \mathbf{z_j}$ em S_i e $Y_j = y_j$ em $\Phi(S_i)$, incrementa-se o contador de vezes com que um dado valor $Y_j = y_j$ foi observado em $\Phi(S)$ dada a observação da instância $\mathbf{Z_j} = \mathbf{z_j}$ em S.

Cada perfil de evolução da função critério tem um rótulo associado a ele correspondente à dimensão verdadeira do subconjunto de características, conforme a PBN da qual esse perfil de evolução foi derivado. Sendo assim, cada PBN gerada aleatoriamente na primeira etapa serve como um gabarito que fornece a dimensão ideal do subconjunto de preditores de cada gene.

Com os perfis de evolução da função critério e seus rótulos (graus ou dimensões dos preditores), pode-se treinar algoritmos de classificação supervisionada para associar conjuntos de perfis aos seus respectivos rótulos. Tal aprendizado pode servir de base para a classificação (estimação do grau correto) em outros processos de inferência com características similares, incluindo inferência de redes a partir de dados temporais reais de expressão. Esse processo é denominado *transferência de aprendizado* (do inglês: *Transfer Learning*) [Konidaris and Barto, 2007, Taylor et al., 2007]. De fato, aplicamos esse processo de transferência de aprendizado para inferência de redes gênicas do ciclo intraeritrocítico do *Plasmodium falciparum*, um agente causador da malária, através de dados de *microarrays* (Capítulo 6).

Capítulo 4

Resultados experimentais para redes simuladas

4.1 Protocolo experimental

Para avaliar a eficácia dos métodos de agrupamento propostos para inferir redes gênicas a partir de dados de expressão, foram realizados alguns experimentos com dados simulados. Redes Booleanas artificiais foram geradas aleatoriamente e as dinâmicas produzidas pelas redes foram simuladas ao longo do tempo para criar o conjunto de dados de entrada. A Seção 4.1.1 descreve o procedimento de geração das redes artificiais e dos respectivos dados de expressão gênica simulados. As métricas de avaliação de similaridade topológica adotadas estão descritas na Seção 4.1.2. Os métodos de busca, bem como a função critério adotada, são apresentados nas Seções 4.1.3 e 4.1.4, respectivamente. Os valores dos parâmetros adotados para a execução dos experimentos podem ser encontrados na Seção 4.1.5. Finalmente, os resultados experimentais envolvendo a comparação dos métodos sem agrupamento (SA), agrupamento linear (LG), agrupamento por busca SFS no reticulado de partições (GLSFS) e agrupamento por canalização (CG) são expostos e discutidos na Seção 4.2.

4.1.1 Geração de redes booleanas e dos dados simulados

Para gerar uma rede Booleana, foram fixados: o número total de genes (N), o número médio de preditores por gene ($\langle k \rangle$), e a topologia da rede (Erdös-Rényi). Primeiramente é gerada a topologia da rede empregando o modelo Erdös-Rényi descrito na Seção 2.4.1. Após definida a topologia da rede, para cada gene g_i, um conjunto de funções Booleanas ϕ_i é selecionado aleatoriamente do conjunto de $2^{2^{k_i}}$ possíveis funções de k_i preditores, em que k_i é o número de preditores do gene g_i. O método de Quine-McCluskey [McCluskey, 1956]

foi empregado para verificar se uma determinada função Booleana sorteada é mínima (ou seja, se depende ou não de todas as variáveis estipuladas na rede gabarito). Caso a função não dependa de todas as variáveis, novas funções vão sendo sorteadas até encontrar uma que de fato dependa de todas as variáveis. No final, cada gene tem um número fixo de funções preditoras, cada qual com uma probabilidade de ser aplicada. No caso de redes Booleanas, cada gene tem uma única função preditora (comportamento determinístico). Já no caso de redes Booleanas probabilísticas (PBNs), foram fixadas duas funções preditoras por gene, associando a uma delas probabilidade próxima de 1, simulando assim um comportamento quase-determinístico inerente a sistemas biológicos reais.

Para gerar os dados simulados, fixa-se um número M de instantes de tempo (número de amostras) e para cada PBN, um estado inicial $\vec{s}_0$ é escolhido aleatoriamente do conjunto de todos os 2^N estados possíveis. Então, a evolução dos estados da rede é simulada a partir de $\vec{s}_0$ até $\vec{s}_{M-1}$ através de repetidas aplicações do conjunto de funções Φ. Caso algum estado $\vec{s}_i$ seja repetido no processo de simulação (no caso da simulação ter percorrido totalmente um ciclo atrator antes de gerar M estados), os dados são descartados e a simulação é reiniciada a partir de um outro estado $\vec{s}_0$ escolhido aleatoriamente. Dessa forma, garante-se que cada simulação passe por M estados distintos. Para garantir a diversidade dos dados simulados é fixado um limiar l, de proporção mínima de diversidade dos dados, ou seja, no mínimo uma proporção l dos dados são diferentes entre os possíveis valores.

4.1.2 Métricas de avaliação adotadas

As redes inferidas foram comparadas com as redes gabaritos através de uma métrica de similaridade topológicas baseada no número de verdadeiros positivos, falsos positivos e falsos negativos. O valor dos verdadeiros negativos não é levado em conta pelo fato das redes serem esparsas, fazendo com que o numero de verdadeiros negativos respondam por quase 100% do total de pares de genes. Nesse sentido, usamos a métrica F-SCORE conforme definido anteriormente na Equação 2.9.

Também adotamos a taxa de acerto como uma forma de avaliar a dinâmica do sinal gerado pela rede inferida. Trata-se do número de bits de diferença entre o sinal gerado pela rede inferida e o sinal gerado pela rede gabarito, normalizado pelo tamanho do sinal. Essa medida de distância foi adotada para comparar o próximo estado inferido a partir de um estado inicial sorteado, com o próximo estado gerado pela rede gabarito a partir do mesmo estado inicial.

Para avaliar a capacidade de generalização dos métodos em relação a geração da dinâmica, diversos estados iniciais são sorteados. Em seguida, esses estados são fornecidos para os métodos gerarem o próximo estado a partir de cada estado inicial sorteado com base na topologia e regras lógicas inferidas. Esses estados então são comparados com

os estados gerados pela topologia e pelas regras lógicas da rede gabarito. Quanto mais próximos os estados inferidos estiverem dos estados gerados pelo gabarito em termos de bits de diferença, melhor.

4.1.3 Algoritmo de seleção de características adotado

Adotamos a busca exaustiva incremental (BEI) como algoritmo de seleção de características, a qual consiste em aplicar o algoritmo de busca exaustiva procurando pelo melhor subconjunto de características de um grau fixo k, de acordo com uma função critério adotada. Inicialmente, esse método obtém a melhor característica individual ($k = 1$) e verifica se a melhor característica obteve uma melhora em relação a um conjunto vazio de atributos (no caso $k = 1$, utiliza-se para comparação o valor da entropia *a priori* do gene alvo considerado). Caso haja melhora, procura pelo melhor par de características ($k = 2$), comparando esse subconjunto com a melhor característica individual obtida ($k = 1$), e assim por diante. O procedimento termina quando o melhor subconjunto de grau k' não melhora a função critério em relação ao melhor subconjunto de grau $k' - 1$, devolvendo assim o melhor subconjunto de grau $k' - 1$. Porém, dessa forma o custo computacional da busca exaustiva a partir de um certo grau k'' começa a ficar inviável. Por isso, a fim de continuar o procedimento a partir desse grau, troca-se o algoritmo BEI para a busca sequencial para frente (SFS) conforme descrito na Seção 2.3.3. Foi adotado $k'' = 4$ em todos os experimentos. O algoritmo 4 apresenta uma descrição mais precisa desse procedimento.

Algorithm 4 Algoritmo de busca exaustiva incremental (BEI)

Require: Conjunto de todos os genes $\mathbf{X}$, gene alvo Y, conjunto de dados de expressão gênica, função critério $\mathcal{F}$ que deve ser minimizada, e grau k'' a partir do qual o algoritmo passa a executar a busca sequencial para frente (SFS)
Ensure: Conjunto de preditores $\mathbf{Z} = \{z_1, \ldots, z_k\}$ para Y

$\mathbf{Z} \leftarrow \emptyset$
$k \leftarrow 1$
$\mathbf{Z}' \leftarrow$ melhor subconjunto de tamanho 1 de $\mathbf{X}$ como preditor de Y de acordo com $\mathcal{F}$
while $\mathcal{F}(\mathbf{Z}', Y) < \mathcal{F}(\mathbf{Z}, Y)$ **do**
  $\mathbf{Z} \leftarrow \mathbf{Z}'$
  if $k \leq k''$ **then**
    $\mathbf{Z}' \leftarrow$ melhor subconjunto de tamanho k de $\mathbf{X}$ como preditor de Y
    $k \leftarrow k + 1$
  else
    $\mathbf{Z} \leftarrow$ SFS($\mathbf{X}$,Y,$\mathbf{Z}$,$\mathcal{F}$)
    return $\mathbf{Z}$
  end if
end while
return $\mathbf{Z}$

Como uma alternativa ao critério de parada do algoritmo de busca, foi aplicada uma

variante do algoritmo 4, aplicando-se o processo de transferência de aprendizado idealizado neste trabalho, apresentado na Seção 2.3.7, para tentar estimar o grau correto $\hat{k}$ de cada gene. Assim, caso $\hat{k} \leq k''$, a busca exaustiva é realizada apenas para grau $\hat{k}$. Já caso $\hat{k} > k''$, realiza-se uma busca exaustiva pelo melhor conjunto de grau k'' e o algoritmo continua através do SFS até atingir o grau $\hat{k}$ fornecido a priori. Com isso, não é levada em consideração a evolução da função critério, pois o grau $\hat{k}$ desejado é fixado externamente. O algoritmo 5 apresenta a descrição deste procedimento.

Algorithm 5 Algoritmo de Busca Exaustiva Incremental com Grau Fixo (BEIF)

Require: Conjunto de todos os genes $\mathbf{X}$, gene alvo Y, conjunto de dados de expressão gênica, função critério $\mathcal{F}$ que deve ser minimizada, grau $\hat{k}$ fixo do gene alvo Y, e grau k'' a partir do qual o algoritmo passa a executar a busca sequencial para frente (SFS) até atingir o grau $\hat{k}$

Ensure: Conjunto de preditores $\mathbf{Z} = \{Z_1, \ldots, Z_{\hat{k}}\}$ para Y

1: **if** $\hat{k} \leq k''$ **then**
2: $\quad \mathbf{Z} \leftarrow$ busca exaustiva pelo melhor subconjunto de tamanho $\hat{k}$ de $\mathbf{X}$ como preditor de Y de acordo com $\mathcal{F}$
3: **else**
4: $\quad \mathbf{Z} \leftarrow$ busca exaustiva pelo melhor subconjunto de tamanho k'' de $\mathbf{X}$ como preditor de Y de acordo com $\mathcal{F}$
5: $\quad \mathbf{Z} \leftarrow SFS(\mathbf{X}, Y, \mathbf{Z}, \mathcal{F}, \hat{k})$
6: **end if**
7: **return** $\mathbf{Z}$

4.1.4 Funções critério adotadas

Para avaliar a qualidade da predição de um conjunto de genes preditores em relação ao comportamento de um determinado gene alvo foi adotada a entropia condicional média conforme a Equação 2.3 como função critério. Já para estimar o grau correto dos preditores através do procedimento proposto descrito na Seção 3.5, foi considerada a evolução da informação mútua normalizada (Equação 2.5). E finalmente, para o aprendizado supervisionado dos graus dos perfis de evolução da informação mútua normalizada ao longo das dimensões foi considerado o algoritmo K vizinhos mais próximos K-nn (Seção 2.3.7).

4.1.5 Parâmetros adotados

A Tabela 4.1 apresenta os parâmetros usados para gerar os experimentos. Os desempenhos dos métodos de inferência por agrupamento foram comparados entre si e com a inferência sem agrupamento, todos seguindo o modelo de redes gênicas probabilísticas (PGN) [Barrera et al., 2007] considerando genes Booleanos.

Tabela 4.1: Parâmetros utilizados nos experimentos.

Parâmetro	Valores
Tamanho da rede gabarito (número de genes N)	50
Grau médio da rede gabarito $\langle k_{gab} \rangle$	3
Número de amostras (tamanho do sinal M)	{30,50}
Probabilidades das funções Booleanas das PBNs	$(0,98; 0,02)$
Modelo topológico da rede gabarito	Erdös-Rényi

4.2 Resultados

Cada experimento apresentado a seguir corresponde a 250 redes gabarito obtidas aleatoriamente, cada qual tendo gerado 4 conjuntos de amostras a partir de estados iniciais distintos sorteados, resultando então em 1000 resultados de inferência. Sendo assim, cada gráfico "violino" (*Violin plot*) corresponde à distribuição de 1000 resultados de inferência de rede por um dos métodos seguintes: sem agrupamento (SA), agrupamento linear (LG), agrupamento por busca sequencial para frente no reticulado de partições (GLSFS), e agrupamento por canalização (CG). Para observar o impacto dos métodos de agrupamento com conhecimento a priori das funções das redes biológicas reais, os métodos foram testados para redes com funções geradas aleatoriamente (sem nenhum conhecimento a priori), redes com funções canalizadoras, e redes com funções linearmente separáveis (Seção 2.2.4).

4.2.1 Redes com funções aleatórias

Cada experimento apresentado a seguir corresponde a 250 redes gabarito obtidas aleatoriamente, cada uma tendo gerado 4 conjuntos de amostras a partir de estados iniciais distintos sorteados, resultando então em 1000 resultados de inferência. Lembrando que as funções preditoras são minimais, ou seja, dependem de todos os preditores para o gene alvo. Essa é a única restrição imposta na geração das funções.

Avaliação topológica das redes inferidas

A Figura 4.1 apresenta os *Violin plots* para o F-Score obtido para 1000 redes inferidas pelos métodos comparados, para 30 e 50 amostras.

Primeiramente, observa-se que o método mais prejudicado é o método de canalização (CG) tanto para $M = 30$ e $M = 50$ amostras, como esperado, já que este método considera apenas um conjunto reduzido de funções como possíveis funções preditoras em comparação aos outros métodos. Por outro lado, o método de agrupamento linear (LG), mesmo considerando também um conjunto reduzido de funções preditoras, apresenta resultados competitivos em relação aos outros dois métodos que não possuem essa restrição, sendo

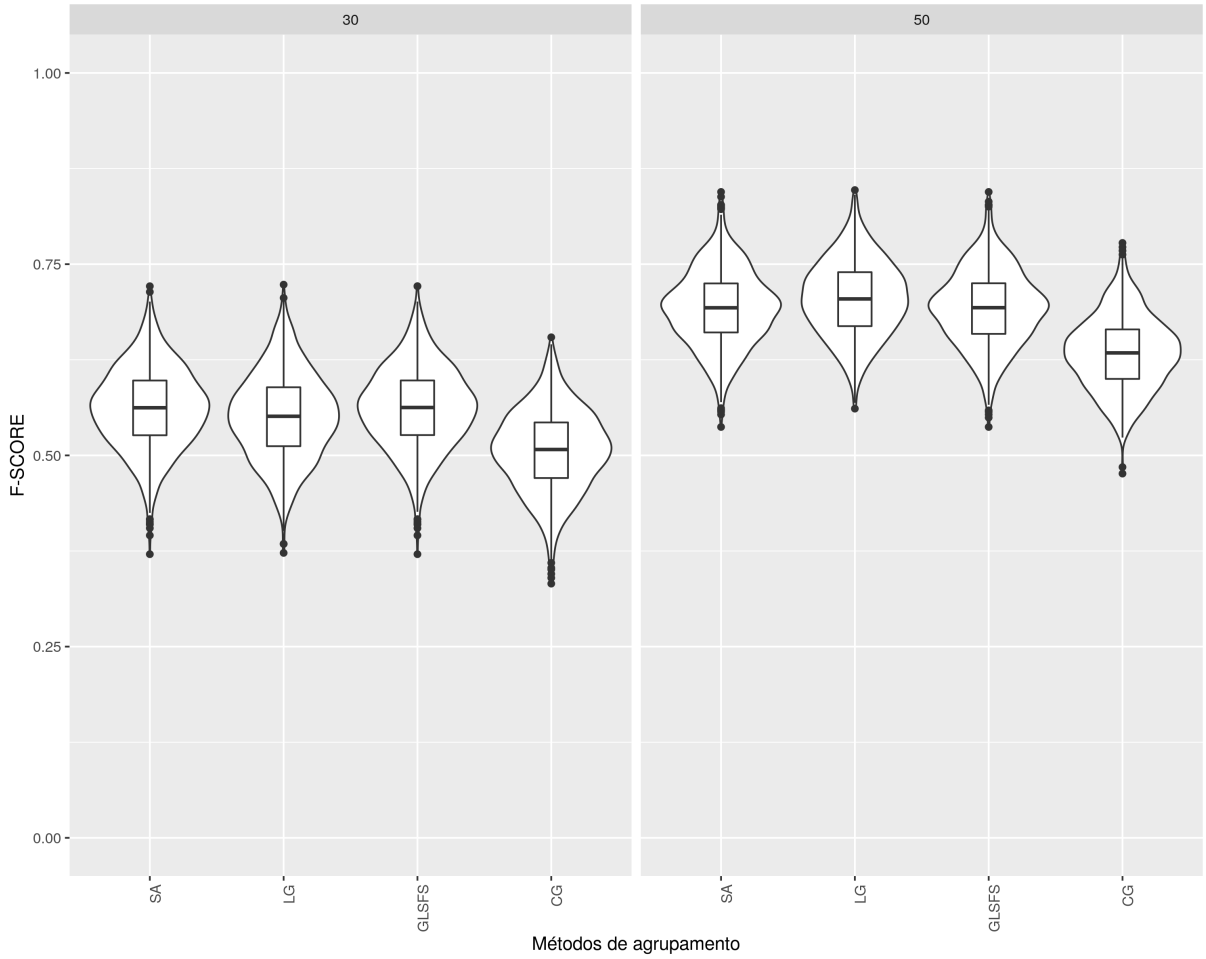


Figura 4.1: *Violin plots* dos valores de F-Score para as topologias das redes inferidas, para 30 amostras (à esquerda) e 50 amostras (à direita). Cada gráfico contém 4 *Violin plots*, um para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, e CG: agrupamento por canalização). Cada *Violin plot* corresponde a uma distribuição de valores de F-Score de 1000 redes inferidas.

ainda ligeiramente superior no caso de $M = 50$ amostras. Uma possível explicação é o poder que este método tem para lidar com dados ruidosos, que compensa o fato de não lidar com todas as possíveis funções. Tal comportamento foi observado em [Montoya-Cubas et al., 2015]. No caso de 30 amostras, o método que apresentou melhor desempenho foi o GLSFS, com uma média de 0.5631. Já considerando 50 amostras, o método LG obteve a melhor média: 0.7052. A Tabela 4.2 apresenta um sumário dos valores apresentados na Figura 4.1.

Una caraterística importante a ser estudada em relação a inferência de redes gênicas, que estão também ligadas a topologia da rede, é descobrir a dimensão ideal do conjunto de preditores. No caso de um número pequeno de amostras, é difícil se obter um conjunto de preditores de dimensão grande, já que isso implica em um volume grande de instâncias não observadas, ocasionando falsos negativos na inferência. Por outro lado, se a dimensão

Tabela 4.2: Sumário dos valores de F-Score para redes inferidas, para 30 amostras e 50 amostras, incluindo média, desvio padrão, o valor mínimo e o valor máximo dos valores de F-Score de 1000 redes inferidas, para cada método de inferência.

Método	30 amostras				50 amostras			
	Média	DP	Min	Max	Média	DP	Min	Max
SA	0.5617	0.0535	0.3708	0.7212	0.6925	0.0495	0.5371	0.8444
LG	0.5512	0.0567	0.3725	0.7232	0.7052	0.0493	0.5611	0.8469
GLSFS	0.5621	0.0535	0.3708	0.7212	0.6925	0.0495	0.5371	0.8444
CG	0.5060	0.0532	0.3323	0.6544	0.6338	0.0480	0.4762	0.7778

ideal for muito pequena (graus 0 e 1), é comum que os métodos obtenham uma dimensão maior do que a ideal (superpopulação), implicando em falsos positivos para esses casos. Visando analisar esse problema com mais detalhes, a Figura 4.2 apresenta as comparações dos histogramas de graus das redes gabaritos com os histogramas de graus das redes inferidas pelos 4 métodos considerados, variando-se o número de amostras $(30, 50)$. As barras correspondentes ao grau 5 representam o acumulado das frequências dos graus 5 e superior.

No caso de 30 amostras, pode-se observar que os métodos SA, LG e GLSFS tendem a superestimar o grau a partir do grau 3, apresentando picos no grau 4. Além disso, há uma queda no grau 5 (e superior), o que pode ser explicado pela troca do algoritmo de busca para SFS, uma vez que o algoritmo CG consegue ter um melhor perfil até o grau 3, mas ele também superestima o grau a partir do grau 4. No caso de 50 amostras, todos os métodos têm perfis semelhantes ao gabarito até o grau 3, sendo que os métodos SA, GLSFS e CG apresentam ligeira superestimação nos graus 4 e 5. Já o método LG apresenta um comportamento mais conservador concentrando a maior parte dos genes de grau alto no grupo de grau 4, sendo que poucos genes superam este grau. A Tabela 4.3 apresenta o grau médio e o erro quadrático médio de cada método tomando as redes gabarito como referência (padrão ouro).

Um resultado que à primeira vista pode parecer contraditório é o fato de que o erro quadrático é maior para o caso de 50 amostras em comparação a 30 amostras, isto é, quanto mais amostras, maior o erro ao tentar encontrar o grau correto dos preditores. Pode-se observar também que todos os métodos apresentam maior grau médio no caso de 50 amostras. Logo, o aumento do número de amostras tende a implicar em aumento no número de preditores (grau) por genes alvo. Justamente esse aumento do grau em alguns casos é devido a superestimação, implicando em um maior erro quadrático médio. Por outro lado, é possível observar na análise do F-SCORE que o método melhora com a inclusão de mais amostras, o que significa que apesar da superestimação do grau fazer com que haja mais falsos positivos, a taxa de verdadeiros positivos também aumenta a ponto de compensar ao menos em parte essa superestimação.

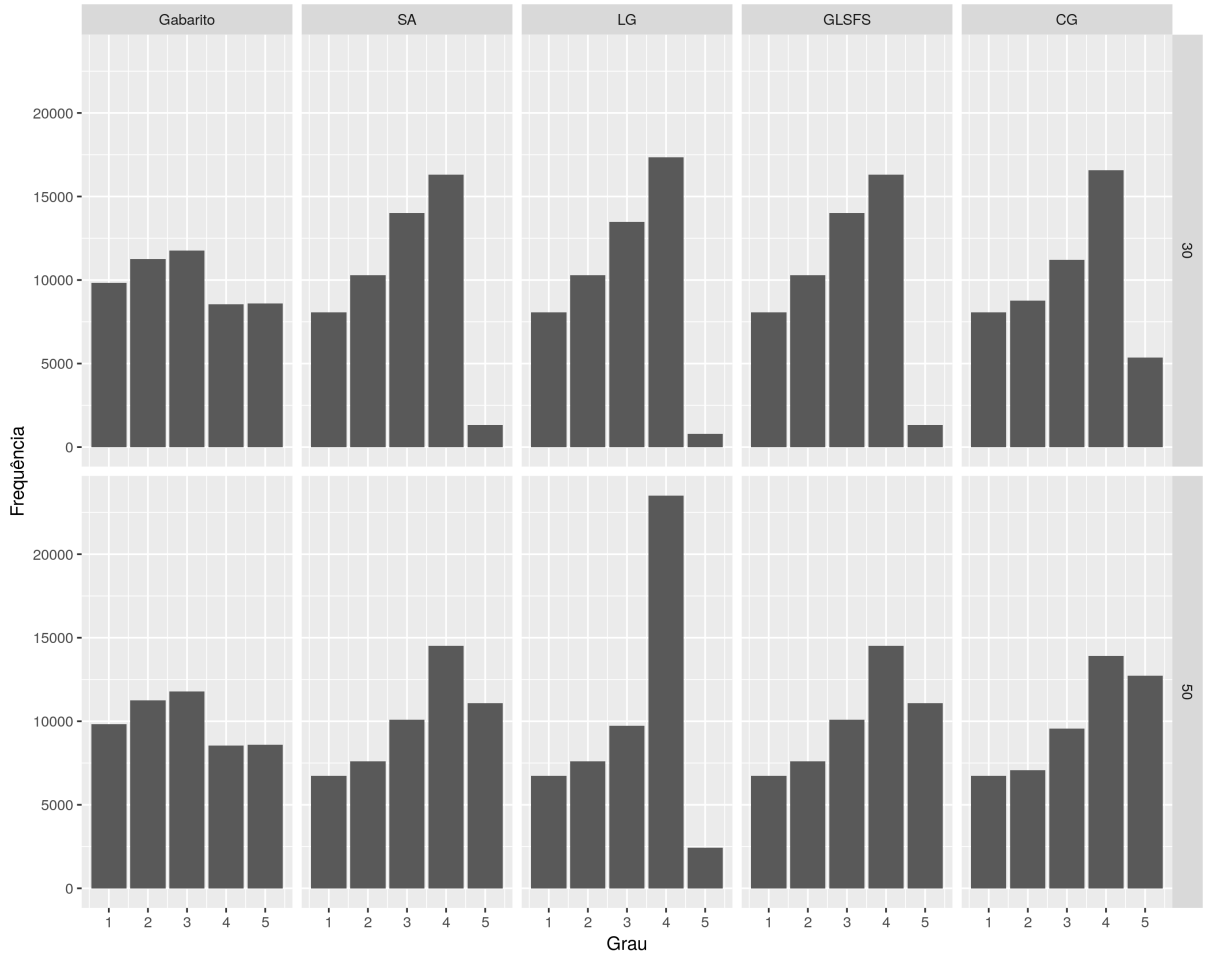


Figura 4.2: Histogramas de graus das redes gabaritos (à esquerda) e das redes inferidas pelos 4 métodos considerados com base em conjuntos de 30 amostras (topo) e 50 amostras (embaixo). SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização. As barras correspondentes ao grau 5 representam o acumulado das frequências dos graus 5 e superiores.

Tabela 4.3: Grau médio (GM) e erro quadrático médio (EQM) considerando os graus dos genes das redes gabaritos e das redes inferidas pelos 4 métodos considerados com base em conjuntos de 30 amostras e 50 amostras. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições; CG: agrupamento por canalização.

Método	GM.30	EQM.30	GM.50	EQM.50
Gabarito	2.897	—	2.897	—
SA	2.850	0.676	3.313	0.995
LG	2.850	0.695	3.147	1.026
GLSFS	2.850	0.676	3.313	0.995
CG	3.048	0.805	3.378	1.149

Ainda na linha de verificar se os métodos tendem a obter as dimensões ideais dos conjuntos de preditores, aprofundamos as análises de modo a obter uma estatística de quantas vezes um método devolveu conjuntos de dimensão i para genes alvos dos gabaritos

com conjuntos de dimensão j, para todo $0 \leq i, j \leq k_{max}$, sendo k_{max} a maior dimensão dentre os conjuntos de preditores das redes gabaritos. Isso gera uma matriz de confusão (matriz de contagens) na qual as colunas representam as dimensões dos conjuntos de preditores dos gabaritos e as linhas representam as dimensões dos conjuntos de preditores das redes inferidas. Essa matriz de contagens normalizada pelo número de vezes com que cada dimensão ocorreu no gabarito pode ser visualizada por um mapa de calor (*heatmap*). Assim, de forma análoga ao que foi feito em relação aos histogramas de grau, aqui também para uma melhor visualização dos dados, graus maiores ou iguais a 5 foram agrupados (a coluna do grau 5 corresponde ao acúmulo de frequências de graus 5 ou superior). A Figura 4.3 apresenta os heatmaps dos 4 métodos comparados aos heatmaps ideais dos gabaritos (com 100% das ocorrências nas células da diagonal principal), variando o número de amostras em $(30, 50)$. Tons mais escuros representam proporções mais altas, enquanto tons mais claros representam proporções mais baixas.

Células escuras embaixo da diagonal representam superestimação do grau (graus inferidos superiores aos graus das redes gabarito), enquanto células escuras acima da diagonal representam subestimação do grau (graus inferidos inferiores aos graus das redes gabarito). Em todos os cenários, os métodos apresentam superestimação para graus inferiores a 4, sendo que esse problema é naturalmente acentuado para o caso de 50 amostras, o que está de acordo com o incremento do erro quadrático médio apresentado na Tabela 4.3. Isso pode ser explicado por uma dificuldade dos métodos de lidarem com dados ruidosos (a quantidade de ruído tende aumentar com o número de amostras). Para genes alvos de dimensão superior a 3, a proporção de graus subestimados pelos métodos de inferência aumenta, especialmente para 30 amostras. De fato, raramente os métodos obtêm conjuntos de preditores de dimensão 5 no cenário de poucas amostras, exceto pelo método de agrupamento por canalização, que consegue obter uma proporção substancialmente maior de preditores com dimensão superior a 4 do que os demais métodos. No caso de 50 amostras, todos os métodos conseguem estimar melhor o grau, exceto o método de agrupamento linear que dificilmente atinge graus maiores que 5.

Isso sugere que o principal problema das funções critério é a superestimação do grau devido ao ruído. Por isso, uma estrategia que propomos para tentar lidar com esse problema será apresentada no Capítulo 5.

Em termos gerais, os métodos de agrupamento linear (LG) e agrupamento por canalização (CG) são os que apresentam os piores resultados de estimação das dimensões dos conjuntos de preditores, sendo o método LG o mais conservador, priorizando dimensões menores. Apesar disso, o LG apresenta os melhores resultados topológicos para 50 amostras, apresentando maior precisão em obter os preditores corretos, compensando a menor precisão na estimação dos graus corretos.

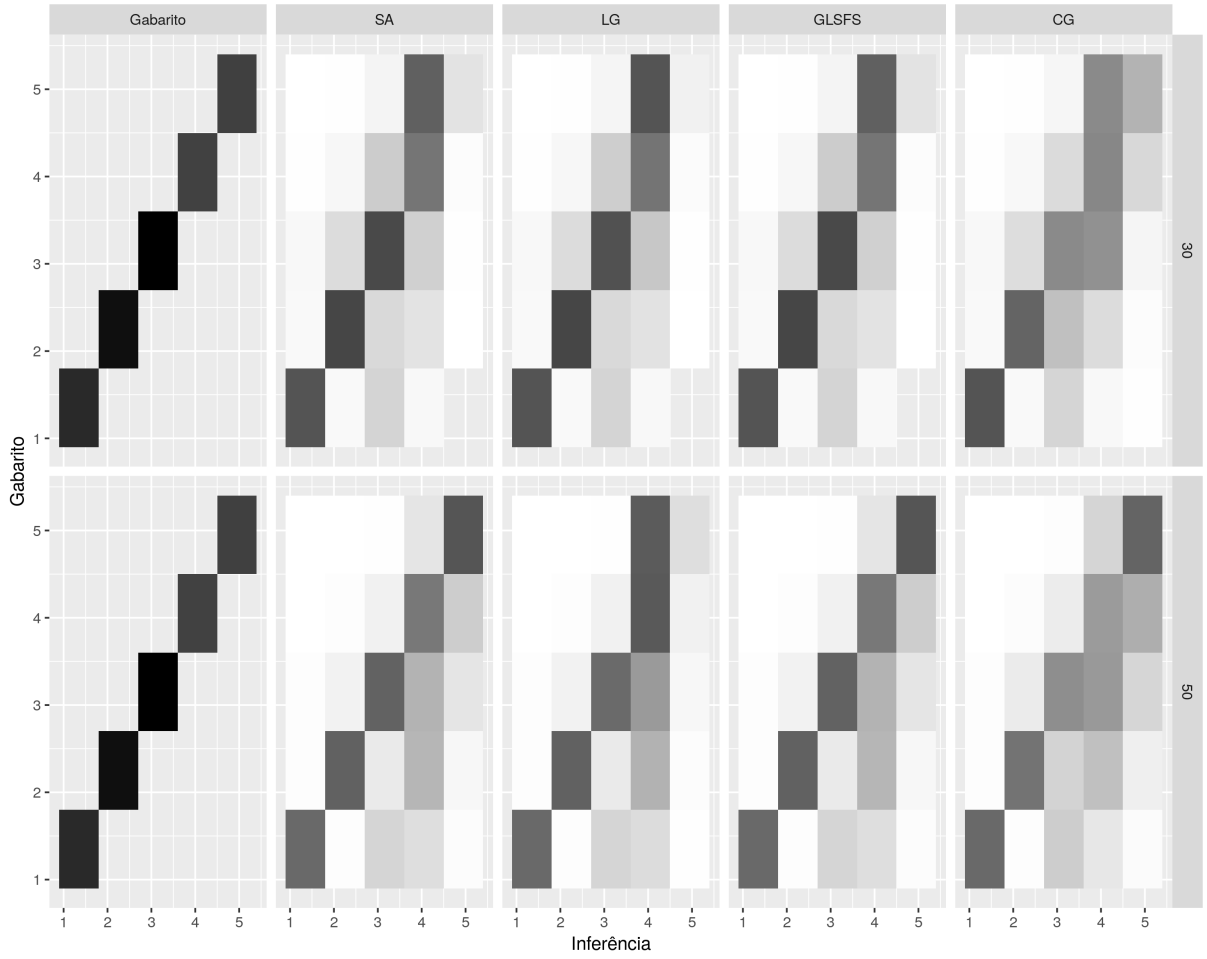


Figura 4.3: Heatmaps onde cada célula (i, j) representa a quantidade de vezes que foram inferidos conjuntos de preditores de dimensão i (linhas) para genes alvos que possuem conjuntos de preditores de dimensão j (colunas) nas redes gabaritos. Os heatmaps indicados por "Gabarito" representam os heatmaps ideais. Quanto mais escuro o tom de preto, maior é a proporção. Número de amostras $= \{30, 50\}$. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

Avaliação da dinâmica gerada pelas redes inferidas

A Figura 4.4 mostra os *Violin plot* correspondentes às taxas de acerto obtidas para 1000 redes inferidas pelos métodos comparados. Cada uma das redes inferidas gerou o próximo estado a partir de 1000 estados iniciais sorteados. A Tabela 4.4 apresenta um resumo dos resultados apresentados na Figura 4.4.

Assim como ocorreu com a avaliação topológica, o método de agrupamento por canalização (CG) apresentou o pior desempenho tanto para 30 como para 50 amostras. Por sua vez, os outros três métodos apresentaram resultados equivalentes, sendo o agrupamento por busca SFS no reticulado de partições (GLSFS) o que apresentou um desempenho ligeiramente superior em todas as situações.

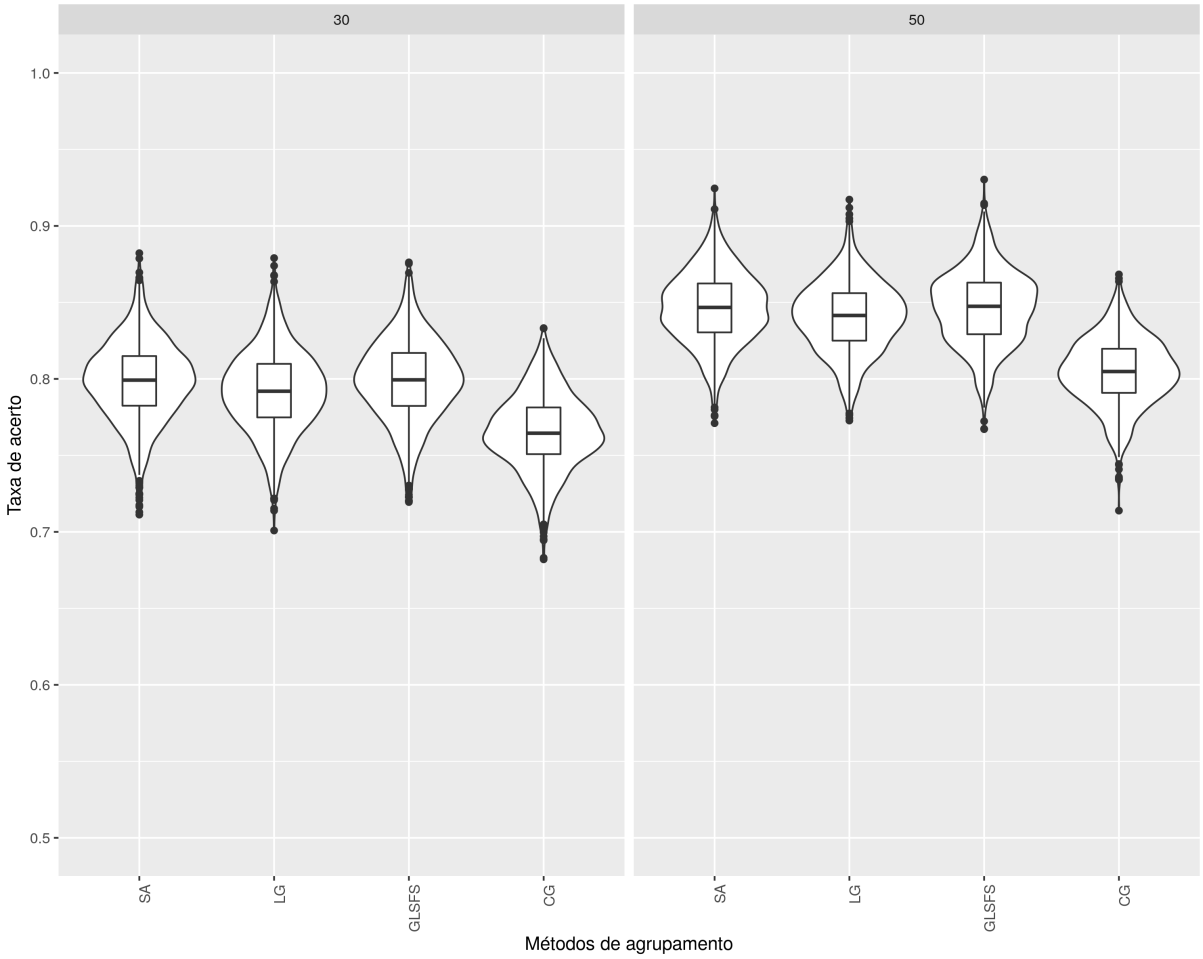


Figura 4.4: *Violin plot* dos valores de taxa de acerto sobre as dinâmicas geradas pelas redes inferidas para 30 amostras (à esquerda) e 50 amostras (à direita). Cada gráfico contém 4 *Violin plots*, um para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de valores de taxa de acerto para 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado considerando 1000 estados iniciais sorteados.

Tabela 4.4: Sumário dos valores de taxa de acerto sobre as dinâmicas geradas pelas redes inferidas, para 30 amostras e 50 amostras. Cada dado corresponde a média o desvio padrão, o valor mínimo e o valor máximo dos valores de taxa de acerto de 1000 redes inferidas,sendo que cada rede inferida gerou o próximo estado de 1000 estados iniciais sorteados.

Método	30 amostras				50 amostras			
	Media	DP	Mín	Máx	Média	DP	Mín	Máx
SA	0.7984	0.0263	0.7113	0.8822	0.8464	0.0235	0.7710	0.9245
LG	0.7918	0.0266	0.7009	0.8790	0.8405	0.0232	0.7728	0.9172
GLSFS	0.7991	0.0267	0.7195	0.8762	0.8463	0.0243	0.7671	0.9303
CG	0.7653	0.0247	0.6820	0.8332	0.8044	0.0224	0.7139	0.8683

Como nesse experimento o conjunto de amostras de teste consistiu de 1000 estados ini-

ciais sorteados, muito provavelmente nenhum desses estados foram observados no conjunto de treinamento, tendo em vista que os conjuntos de treinamento possuem no máximo 50 estados diferentes de um total de $2^{50} \approx 10^{15}$ estados possíveis. Ou seja, esse experimento testa efetivamente a capacidade de generalização dos métodos. Pode-se medir a capacidade de generalização de um método realizando a estatística de quantas vezes instâncias não observadas foram exigidas na predição dos valores dos genes alvos, normalizada pelo total de valores preditos (o que equivale ao número de genes da rede, $N = 50$, multiplicado por 1000 estados iniciais sorteados, ou seja, 50.000). Quanto maior a proporção de instâncias não observadas na predição dos genes alvos, pior é a capacidade de generalização do método.

A Figura 4.5, apresenta os *Violin plots* referentes às distribuições das proporções de instâncias não observadas para 1000 conjuntos de amostras (250 gabaritos $\times$ 4 conjuntos de amostras), enquanto a Tabela 4.5 apresenta um sumário dessas proporções. Observa-se que, de modo geral, o método sem agrupamento (SA) é o que apresenta as maiores proporções, como esperado, já que a ausência de agrupamento implica em um crescimento exponencial do número de instâncias em função da dimensão do conjunto de preditores obtido. O método de agrupamento por canalização apresenta um numero maior em comparação aos outros dois, porém esta proporção é bem menor do que a obtida pelo método SA. O método de agrupamento linear (LG) apresenta proporções mais baixas, as quais não aumentam com o número de amostras, o que implica em uma maior capacidade de generalização independentemente do numero de amostras. É importante notar também que o método de busca no reticulado de partições (GLSFS) apresenta proporções sempre nulas, já que sua função critério foi especificamente projetada para que não haja grupos sem observação.

Tabela 4.5: Sumário das proporções de instâncias não observadas exigidas na predição dos valores dos genes alvos, para 30 amostras e 50 amostras. Cada dado corresponde a média, o desvio padrão, o valor mínimo e o valor máximo das proporções de 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado de 1000 estados iniciais sorteados.

Método	30 amostras				50 amostras			
	Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	0.0848	0.0178	0.0419	0.1505	0.1482	0.0267	0.0582	0.2317
LG	0.0043	0.0029	0.0000	0.0181	0.0060	0.0032	0.0000	0.0177
GLSFS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CG	0.0288	0.0091	0.0035	0.0637	0.0496	0.0117	0.0199	0.0920

Um fato que à primeira vista pode parecer contraintuitivo é que os métodos sem agrupamento (SA) e de canalização (CG) possuem uma proporção de instâncias não observadas exigidas maior justamente para um maior número de amostras (50). Embora o aumento do número de amostras favoreça a diminuição do número de instâncias não observadas exi-

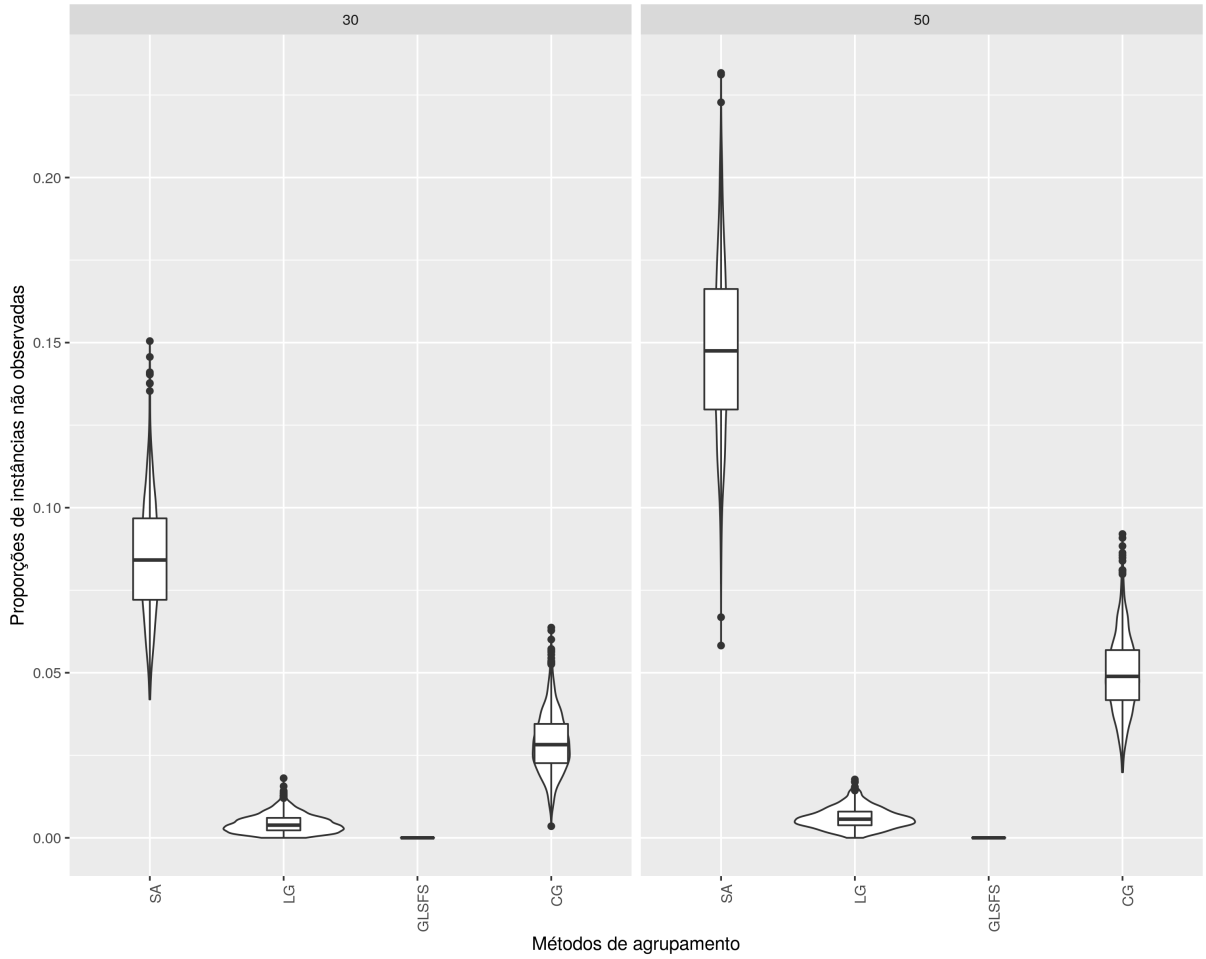


Figura 4.5: *Violin plots* das proporções de instâncias não observadas exigidas na predição dos valores dos genes alvos para redes inferidas para 30 amostras (à esquerda) e 50 amostras (à direita). Cada gráfico contém 4 *Violin plots*, um para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado de 1000 estados iniciais sorteados.

gidas na predição, por outro lado um maior número de amostras faz com que os métodos tendam a obter conjuntos de preditores de dimensões maiores (superestimação do grau), como observado anteriormente na Tabela 4.3. Dessa forma a superestimação do grau faz com que o número de instâncias a serem estimadas acabe aumentando, implicando então em um aumento no número de instâncias não observadas exigidas na predição. Mesmo assim, vale destacar que o método sem agrupamento (SA) foi o mais prejudicado com o aumento de amostras, como esperado, tendo em vista que um maior número de amostras induz dimensões maiores e, com isso, o número de instâncias dos preditores cresce exponencialmente, implicando em uma tabela de probabilidades condicionais com um número muito maior de instâncias não observadas.

4.2.2 Resultados para redes com funções canalizadoras

Com o propósito de testar o desempenho dos métodos de inferência em redes com características biológicas realistas, nesta seção a proposta foi impor uma restrição às funções de predição, admitindo apenas funções canalizadoras (vide Seção 2.2.4).

Seguindo o mesmo protocolo experimental apresentado na Seção 4.1, cada experimento apresentado corresponde a 250 redes gabarito obtidas aleatoriamente, com 4 conjuntos de amostras para cada rede, obtendo assim 1000 resultados de inferência. Para manter consistência com a topologia da rede, apenas funções canalizadoras mínimas (que dependem de todos os preditores para o gene alvo) são admissíveis.

Avaliação das topologias das redes inferidas

A Figura 4.6 apresenta os *Violin plots* para o F-Score obtido para 1000 redes inferidas pelos métodos comparados, para 30 e 50 amostras. Os resultados correspondentes são sumarizados na Tabela 4.6. Primeiramente, observa-se que o método de canalização (CG) tanto para $M = 30$ e $M = 50$ amostras, apresenta resultados similares aos outros métodos, representando uma melhora significativa em relação aos resultados apresentados para redes com funções aleatórias conforme Seção 4.2.1, sendo o método com melhores resultados para $M = 30$. Este resultado é esperado já que o método de canalização é projetado justamente para reconhecer redes gabarito geradas com funções canalizadoras. Ressalta-se então a importância de embutir *informação a priori* sobre a topologia e funções lógicas típicas em redes gênicas para o desenvolvimento de métodos mais eficazes para a inferência das redes [Lopes et al., 2014, Montoya-Cubas et al., 2015, Martins-Jr et al., 2016].

Mesmo assim, ao comparar o método de canalização com os outros métodos do ponto de vista topológico, esse método não se mostrou significativamente superior em termos de F-SCORE, sendo inclusive ligeiramente superado pelo método de agrupamento linear (LG) para $M = 50$ amostras. Tal resultado pode ser explicado pelo fato de que o método LG também é capaz de identificar um conjunto considerável de funções canalizadoras especialmente para graus pequenos[1], além de apresentar um bom poder de generalização considerando que ele também apresentou resultados competitivos no cenário de redes aleatórias (ver Seção 4.2.1).

Do mesmo modo que na análise de redes compostas por funções aleatórias, analisamos a eficácia dos métodos em identificar a dimensão ideal do conjunto de preditores. A Figura 4.7 apresenta as comparações dos histogramas de graus das redes gabaritos com os histogramas de graus das redes inferidas pelos 4 métodos considerados, variando-se o

[1]Por exemplo, para grau 2, das 10 funções Booleanas minimais, 8 são canalizadoras e linearmente separáveis, sendo que as outras duas (XOR e NXOR) não são canalizadoras e nem linearmente separáveis

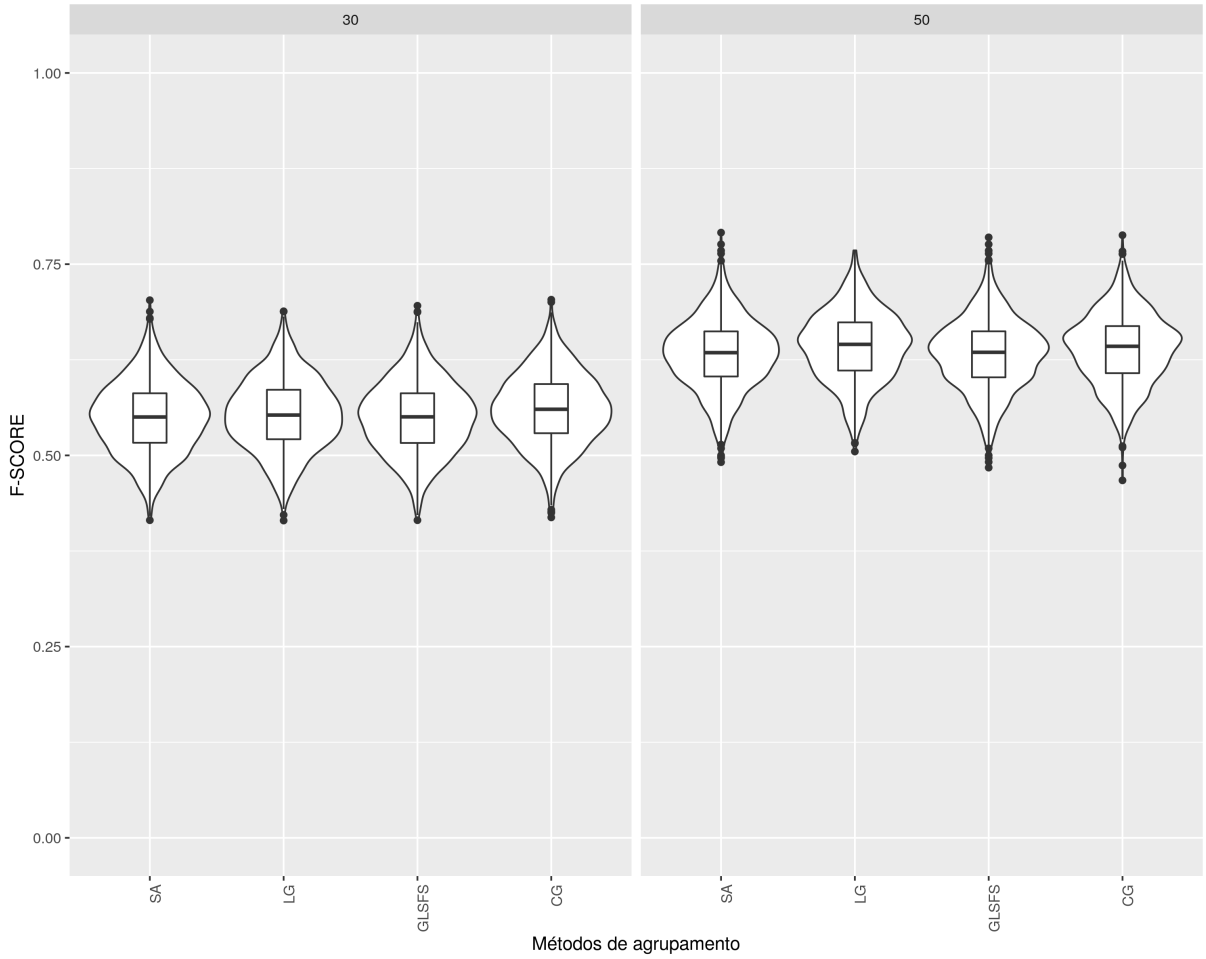


Figura 4.6: *Violin plots* dos valores de F-Score para redes inferidas, para 30 amostras (à esquerda) e 50 amostras (à direita), com funções gabarito canalizadoras. Cada gráfico contém 4 *Violin plots*, um para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, e CG: agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de valores de F-Score de 1000 redes inferidas.

Tabela 4.6: Sumário dos valores de F-Score para as redes inferidas, para 30 amostras e 50 amostras, considerando redes gabaritos compostos apenas por funções canalizadoras. Cada dado corresponde a média, o desvio padrão, o valor mínimo e o valor máximo dos valores de F-Score de 1000 redes inferidas, para cada método de inferência.

Método	30 amostras				50 amostras			
	Media	DP	Mín	Máx	Média	DP	Mín	Máx
SA	0.5495	0.0470	0.4152	0.7029	0.6324	0.0462	0.4911	0.7913
LG	0.5533	0.0470	0.4148	0.6884	0.6424	0.0460	0.5051	0.7680
GLSFS	0.5495	0.0470	0.4152	0.6957	0.6321	0.0464	0.4840	0.7850
CG	0.5603	0.0465	0.4189	0.7036	0.6388	0.0464	0.4674	0.7879

número de amostras em $\{30, 50\}$. As barras correspondentes ao grau 5 representam o acumulado das frequências dos graus 5 e superior. A Tabela 4.7 apresenta o grau médio e o erro quadrático médio de cada método em comparação ao gabarito.

A diferença em relação ao caso de gabaritos com funções aleatórias é que neste cenário a superestimação do grau ocorre a partir de graus menores, apresentando a maior concentração de graus nos graus menores com picos no grau 3 para $M = 30$ amostras, e no grau 4 para $M = 50$ amostras. Porém, o método de agrupamento por canalização (CG) apresenta uma melhor distribuição para ambos os casos. Neste cenário o erro quadrático tanto para 30 como para 50 amostras foi maior para todos os métodos em comparação ao primeiro cenário (redes com funções aleatórias). Entretanto, a diferença dos erros quadráticos entre os métodos não foi significativa, embora o método de agrupamento por canalização (CG) apresentou erros ligeiramente superiores aos erros obtidos pelos demais métodos.

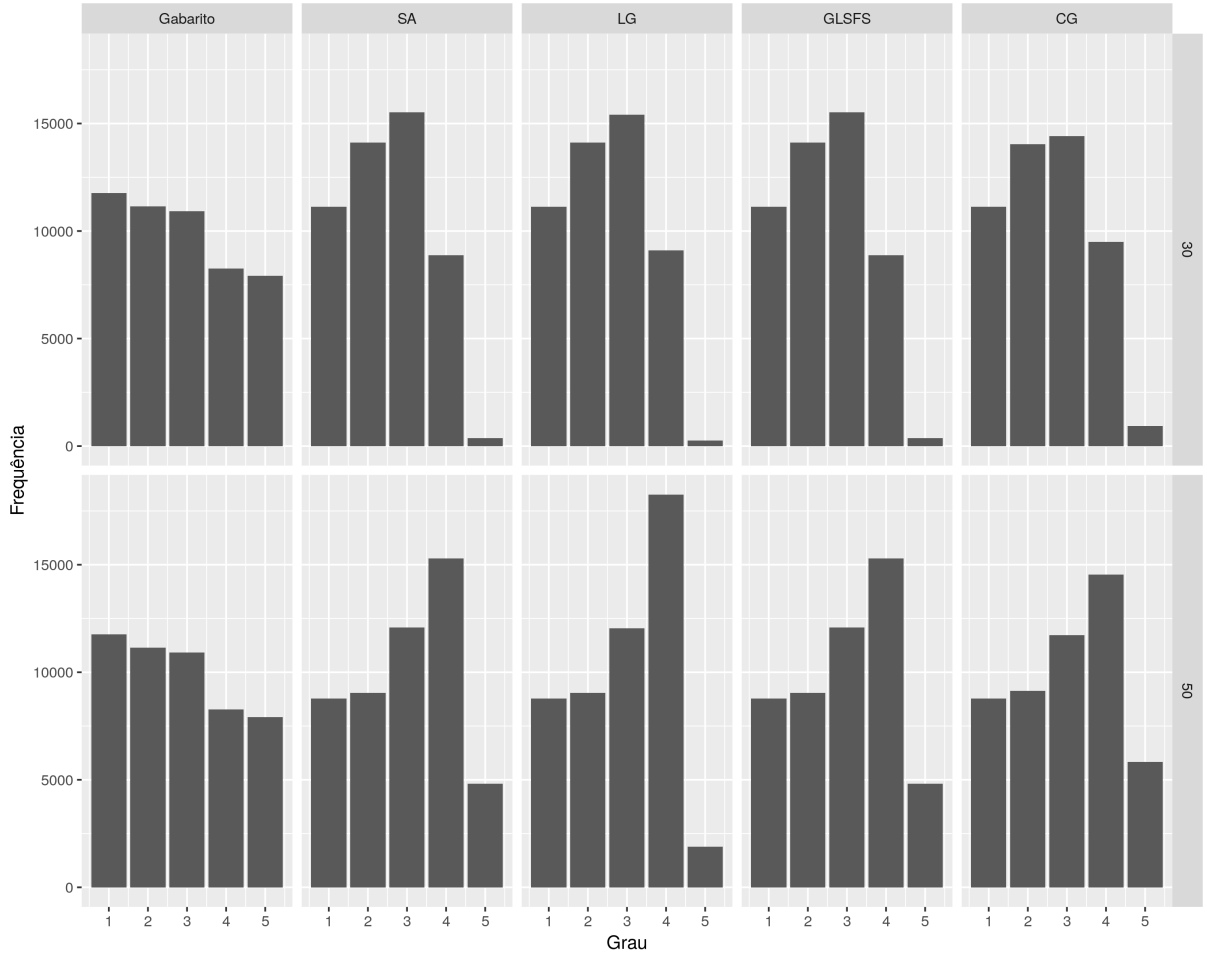


Figura 4.7: Histogramas de graus das redes gabaritos compostas apenas por funções canalizadoras (à esquerda) e das redes inferidas pelos 4 métodos considerados com base em conjuntos de 30 amostras (topo) e 50 amostras (embaixo). SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

A Figura 4.8 apresenta os heatmaps da matriz de confusão dos graus. Assim, é possível perceber células escuras tanto abaixo como acima da diagonal principal, evidenciando

Tabela 4.7: Grau médio (GM) e erro quadrático médio (EQM) das redes inferidas pelos 4 métodos considerados com base em conjuntos de 30 amostras e 50 amostras, considerando redes gabarito compostas por funções canalizadoras. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

Método	GM.30	EQM.30	GM.50	EQM.50
Gabarito	2.789	—	2.789	—
SA	2.465	1.436	2.967	1.466
LG	2.465	1.439	2.909	1.430
GLSFS	2.465	1.436	2.967	1.466
CG	2.502	1.458	2.990	1.521

uma maior dificuldade dos métodos em atingir o grau correto quando as funções são canalizadoras. Tal observação é corroborada pela comparação entre os erros quadráticos médios das Tabelas 4.3 e 4.7. Ainda em relação ao cenário de redes com funções de canalização, para $M = 30$ amostras observa-se que todos os métodos apresentam células mais escuras acima da diagonal, o que mostra uma tendência de subestimar o grau, sendo que nenhum método consegue identificar graus maiores do que 4. Já para $M = 50$ amostras, há mais células escuras embaixo da diagonal, o que mostra a tendência dos métodos a superestimar o grau para um maior número de amostras. O único método que apresenta um comportamento diferente é o agrupamento linear (LG), apresentando a coluna 4 mais escura do que os demais métodos, indicando uma subestimação para graus 5 e superiores. Estes resultados sugerem uma dificuldade maior dos métodos em atingir o grau certo para redes gabaritos com funções canalizadoras em comparação com gabaritos com funções aleatórias.

Em linhas gerais, o método de agrupamento por canalização (CG) apresenta uma melhora significativa para inferir a topologia das redes gabaritos compostas exclusivamente por funções de canalização em comparação aos seus próprios resultados obtidos para gabaritos sem restrição na geração das funções. Entretanto esta melhora ainda não foi suficiente para superar de modo significativo os resultados topológicos de todos os métodos.

Avaliação da dinâmica gerada pelas redes inferidas

A Figura 4.9 mostra os *Violin plots* correspondentes às taxas de acerto das dinâmicas obtidas para 1000 redes inferidas pelos métodos comparados. Cada uma das redes inferidas gerou o próximo estado a partir de 1000 estados iniciais sorteados. A Tabela 4.8 apresenta um sumário dos resultados apresentados na Figura 4.9.

Embora o método de agrupamento por canalização (CG) tenha apresentado resultados topológicos apenas equiparáveis aos demais métodos, conforme observado na Seção 4.2.2,

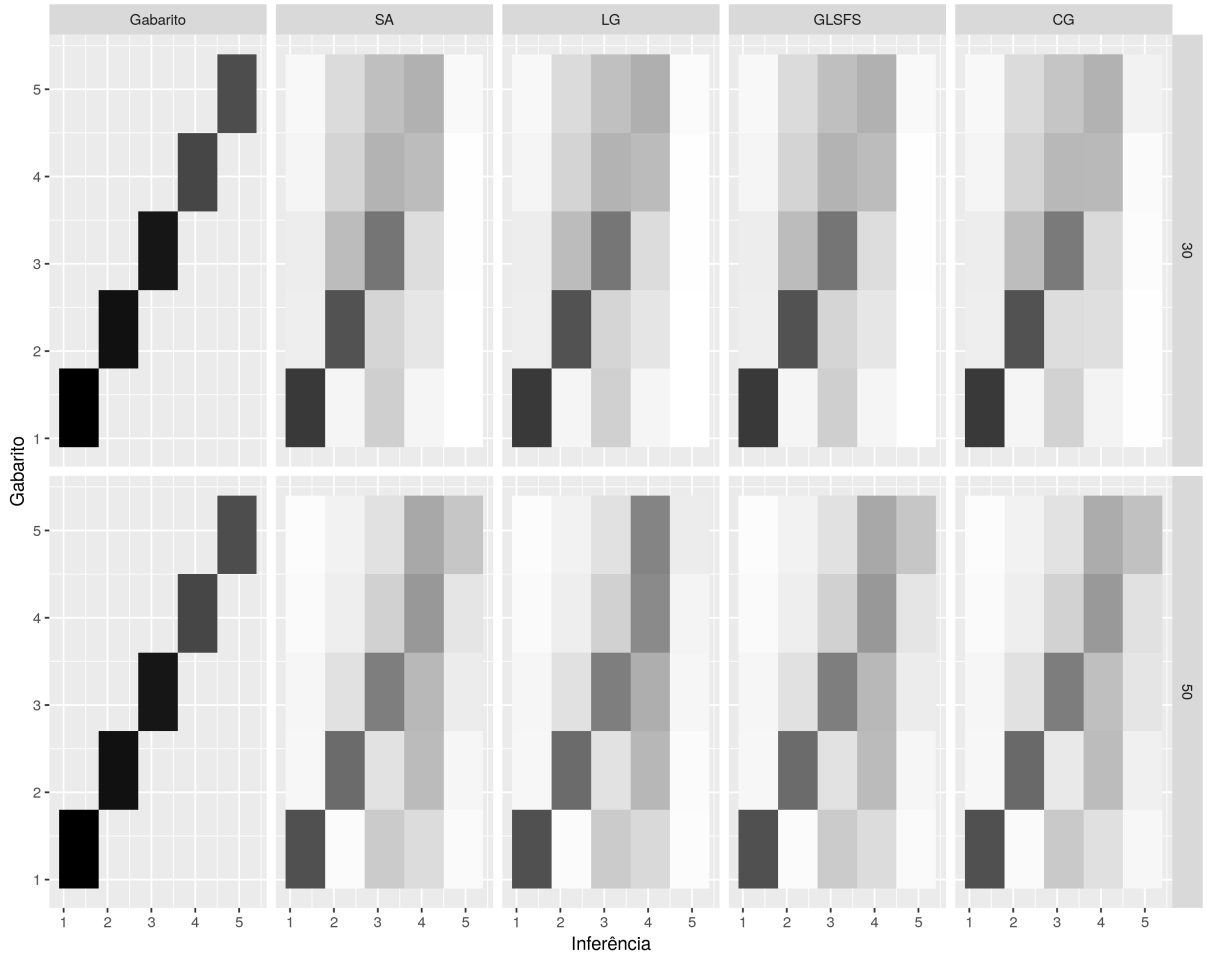


Figura 4.8: Heatmaps nas quais cada célula (i, j) representa a quantidade de vezes que foram inferidos conjuntos de preditores de dimensão i (linhas) para genes alvos que possuem conjuntos de preditores de dimensão j (colunas) nas redes gabaritos compostas apenas por funções canalizadoras. Os heatmaps indicados por "Gabarito" representam os heatmaps ideais. Quanto mais escuro o tom de preto, maior é a proporção. Número de amostras $M = \{30, 50\}$. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

Tabela 4.8: Sumário dos valores de taxa de acerto das dinâmicas geradas pelas redes inferidas, para 30 amostras e 50 amostras, considerando redes gabaritos compostos apenas por funções canalizadoras. Cada dado corresponde a média, o desvio padrão, o valor mínimo e o valor máximo dos valores de taxa de acerto de 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado a partir de 1000 estados iniciais sorteados.

Método	30 amostras				50 amostras			
	Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	0.8234	0.0231	0.7388	0.9004	0.8515	0.0224	0.7771	0.9189
LG	0.8161	0.0234	0.7417	0.8901	0.8469	0.0216	0.7736	0.9148
GLSFS	0.8247	0.0231	0.7304	0.9034	0.8517	0.0233	0.7733	0.9327
CG	0.8459	0.0212	0.7660	0.9163	0.8765	0.0199	0.8150	0.9370

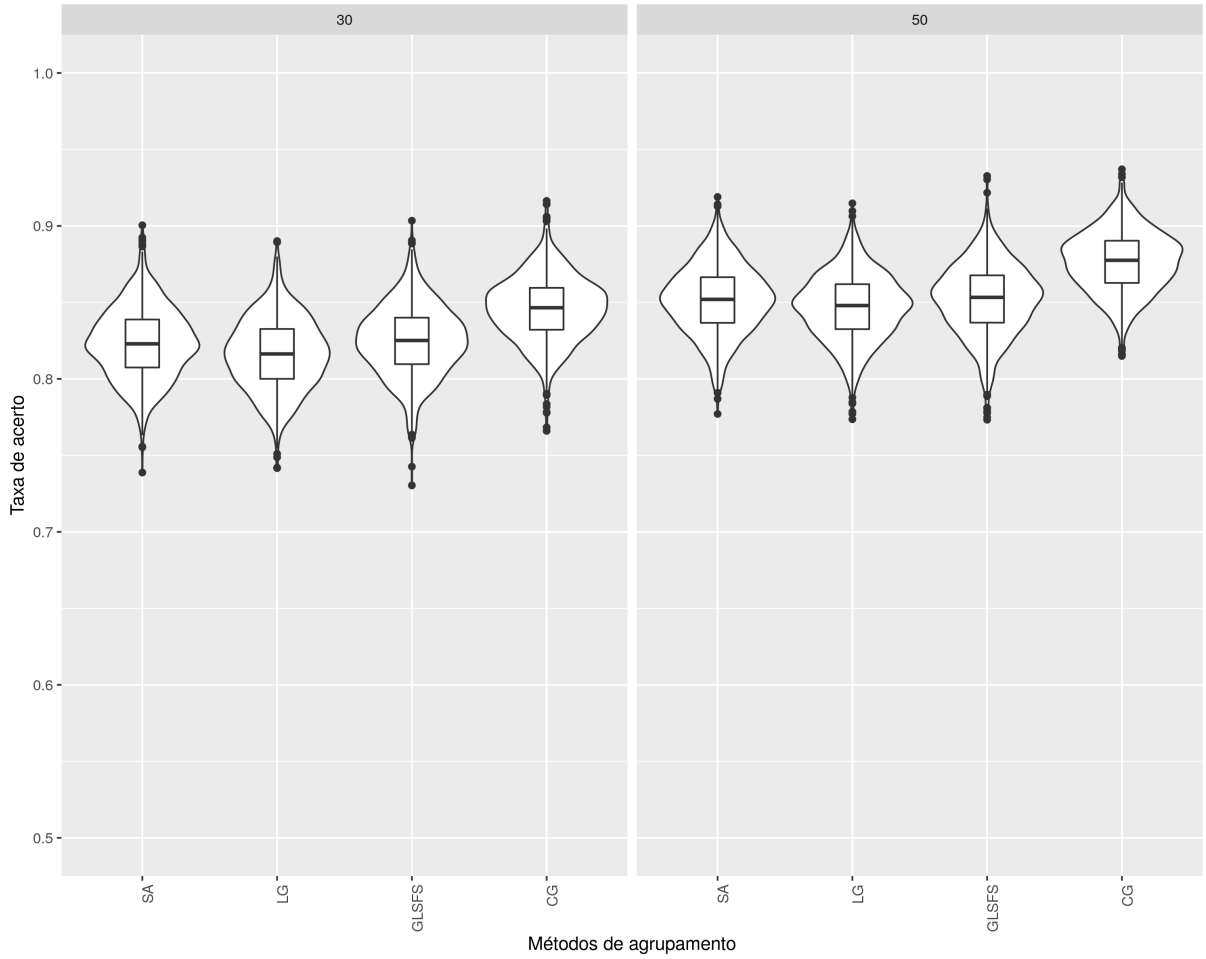


Figura 4.9: *Violin plots* dos valores de taxa de acertos das dinâmicas geradas pelas redes inferidas para 30 amostras (à esquerda) e 50 amostras (à direita), considerando gabaritos compostos exclusivamente por funções canalizadoras. Cada gráfico contém 4 *Violin plots*, um para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de valores de taxa de acerto para 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado de 1000 estados iniciais sorteados.

a situação muda substancialmente quando a taxa de acerto da dinâmica é considerada, pois o método CG apresenta nitidamente os melhores desempenhos tanto para $M = 30$ como para $M = 50$ amostras. Esse resultado reforça a tese de que embutir informações a priori sobre a topologia e as regras lógicas nos métodos de inferência de redes gênicas tende a levar a resultados melhores, conforme já observado em [Lopes et al., 2014, Montoya-Cubas et al., 2015, Martins-Jr et al., 2016]. Além disso, é importante notar que qualidade da inferência do ponto de vista topológico não necessariamente implica em qualidade da inferência do ponto de vista da geração da dinâmica gerada. Ou seja, apesar do método CG não ter se sobressaído em termos de qualidade topológica, a inferência resultante deste método foi capaz de explicar melhor a dinâmica do sistema.

Com o propósito de entender melhor os resultados da dinâmica, conforme realizado no cenário de gabaritos com funções aleatórias, aqui também foi analisada a proporção de instâncias não observadas exigidas na predição dos genes alvos. A Figura 4.10, apresenta os *Violin plots* referentes às distribuições das proporções de instâncias não observadas exigidas para 1000 conjuntos de amostras (250 gabaritos × 4 conjuntos de amostras), enquanto a Tabela 4.9 apresenta um sumário destas distribuições. Observa-se que, de modo geral, todos os métodos mantêm os mesmos comportamentos apresentados no cenário de redes gabaritos compostos por funções aleatórias. Contudo, vale notar que no caso do método de agrupamento por canalização (CG), as proporções de instâncias observadas foram menores para o cenário de redes com funções canalizadoras. Ao comparar as médias obtidas para o cenário anterior (redes gabaritos com funções aleatórias, Tabela 4.5) com o cenário de redes compostas exclusivamente por funções canalizadoras (Tabela 4.9), observa-se que a média de instâncias não observadas exigidas diminuiu em 41% para 30 amostras e em 27% para 50 amostras. Esse resultado indica que o método CG melhorou seu poder de generalização para inferir redes compostas exclusivamente por funções canalizadoras, conforme esperado.

Tabela 4.9: Sumário das proporções de instâncias não observadas exigidas na predição dos valores dos genes alvos, para 30 amostras e 50 amostras, considerando redes gabaritos compostos exclusivamente por funções canalizadoras. Cada dado corresponde a média, o desvio padrão, o valor mínimo e o valor máximo das proporções de 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado a partir de 1000 estados iniciais sorteados.

Método	30 amostras				50 amostras			
	Media	DP	Mín	Máx	Media	DP	Mín	Máx
SA	0.0811	0.0187	0.0293	0.1508	0.1264	0.0240	0.0504	0.1973
LG	0.0039	0.0031	0.0000	0.0194	0.0081	0.0037	0.0000	0.0203
GLSFS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CG	0.0171	0.0062	0.0023	0.0393	0.0362	0.0101	0.0107	0.0652

4.2.3 Resultados para redes com funções linearmente separáveis

Neste cenário foi seguido o mesmo protocolo experimental descrito na Seção 4.1, com a diferença que a caraterística biológica testada nesta seção diz respeito a redes gabarito compostas exclusivamente por funções linearmente separáveis. Tanto a importância como as características biológicas destas funções foram discutidas na Seção 2.2.4.

Avaliação das Topologias das Redes Inferidas

A Figura 4.11 apresenta os *Violin plots* para o F-Score obtido para 1000 redes inferidas pelos métodos comparados para 30 e 50 amostras, enquanto a Tabela 4.10 apresenta um

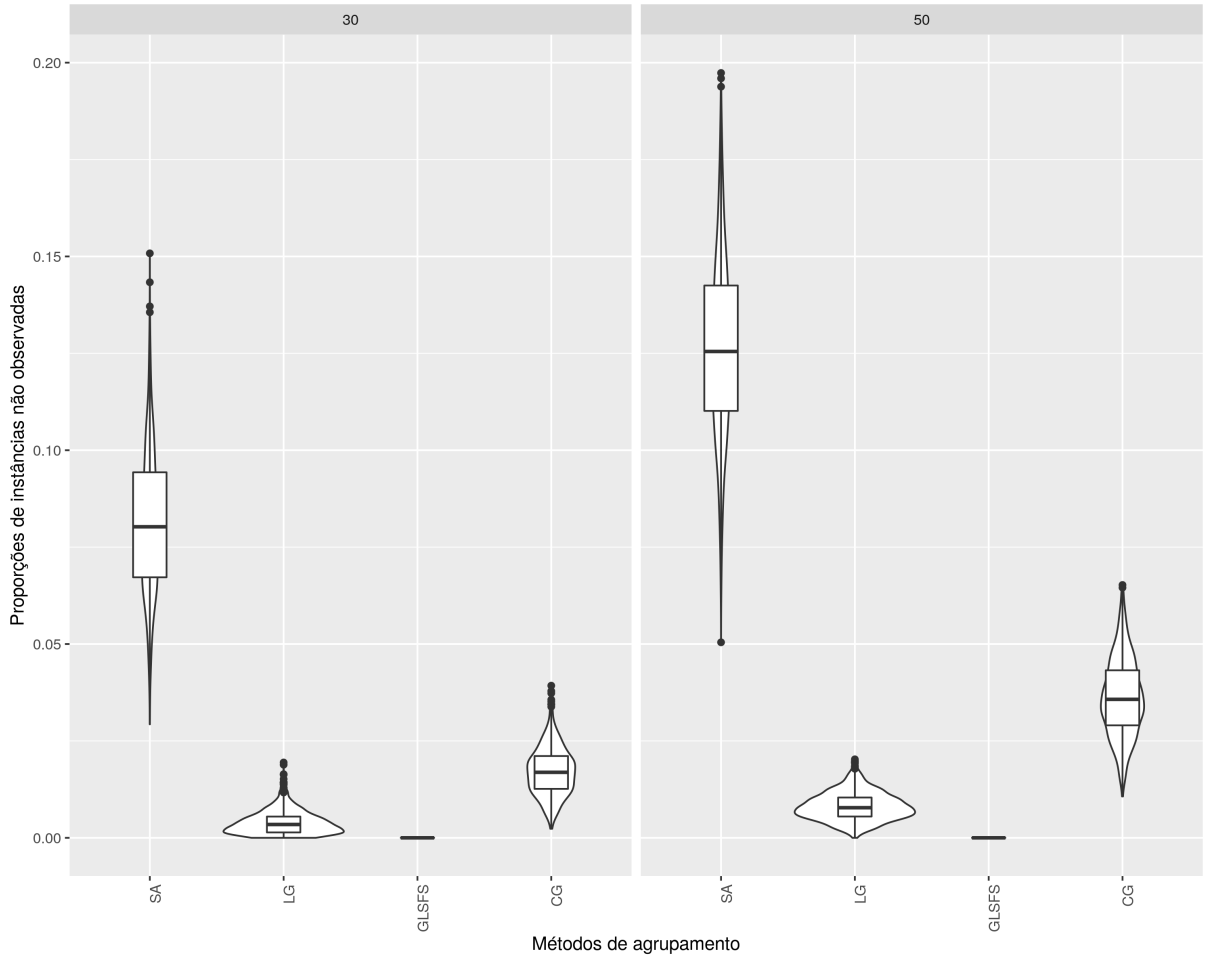


Figura 4.10: *Violin plots* das proporções de instâncias não observadas exigidas na predição dos valores dos genes alvos para redes inferidas para 30 amostras (à esquerda) e 50 amostras (à direita). Cada gráfico contém 4 *violin plots*, um para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização). Cada *violin plot* corresponde a distribuição de 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado de 1000 estados iniciais sorteados. Nesse caso, as redes gabaritos são compostas exclusivamente por funções canalizadoras.

sumário desses resultados.

Primeiramente, observa-se que o método de agrupamento linear (LG) tanto para $M = 30$ e $M = 50$ amostras, apresenta resultados similares aos outros métodos, com um pequeno destaque no caso de $M = 50$ amostras. Esse resultado já era esperado, já que este método têm apresentado resultados competitivos tanto para redes com funções aleatórias (Seção 4.2.1), como nas redes compostas exclusivamente por funções canalizadoras (Seção 4.2.2). Ainda assim, o método de agrupamento linear (LG) não obteve um destaque considerável em relação aos outros métodos em termos topológicos (F-SCORE). Este resultado pode ser explicado pelo fato de que o método LG também consegue inferir um conjunto considerável de funções que não são necessariamente linearmente separáveis,

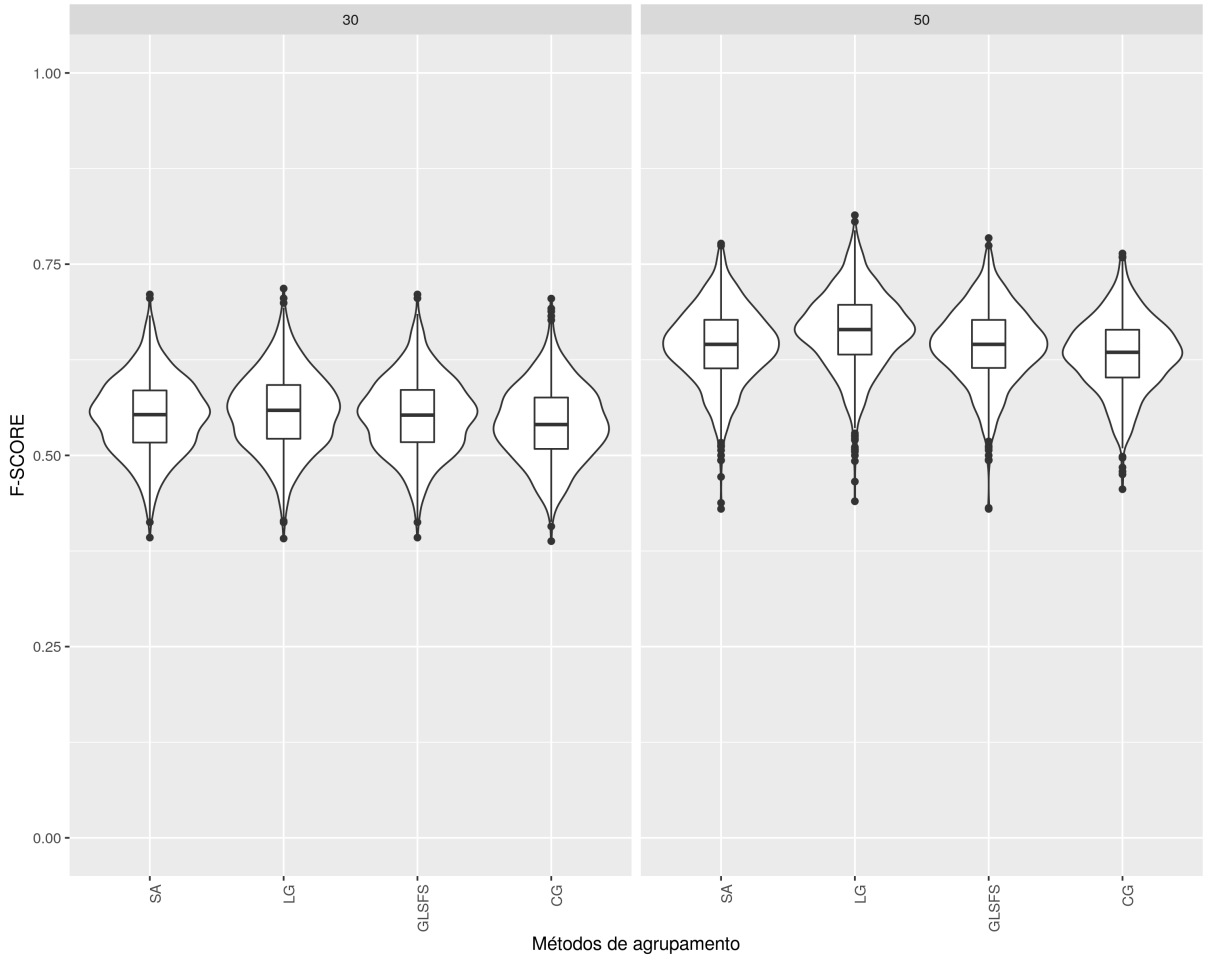


Figura 4.11: *Violin plots* dos valores de F-Score para redes inferidas, para 30 amostras (à esquerda) e 50 amostras (à direita), com funções gabarito linearmente separáveis. Cada gráfico contém 4 *Violin plot*, um para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições e CG agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de valores de F-Score de 1000 redes inferidas.

o que o faz um método não completamente especializado neste tipo de funções.

Por outro lado, é preciso pontuar que o método de agrupamento por canalização (CG), mesmo obtendo resultados ligeiramente inferiores aos outros métodos, a diferença não é tão significativa quanto no caso de redes com funções aleatórias. Isso ocorre porque muitas funções linearmente separáveis também são canalizadoras.

Para o problema de identificar a dimensão ideal do conjunto de preditores, a Figura 4.12 apresenta as comparações dos histogramas de graus das redes gabaritos com os histogramas de graus das redes inferidas pelos 4 métodos considerados, variando-se o número de amostras em $\{30, 50\}$. As barras correspondentes ao grau 5 representam o acúmulo das frequências de graus 5 e superior. Nota-se que os perfis de distribuição dos graus são semelhantes aos perfis obtidos para redes gabaritos compostas exclusivamente

Tabela 4.10: Sumário dos valores de F-Score para redes inferidas para 30 amostras e 50 amostras geradas por redes gabaritos compostas exclusivamente por funções linearmente separáveis. Cada dado corresponde a média, ao desvio padrão, ao valor mínimo e ao valor máximo dos valores de F-Score de 1000 redes inferidas para cada método de inferência.

Método	30 amostras				50 amostras			
	Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	0.5515	0.0497	0.3925	0.7104	0.6437	0.0500	0.4300	0.7770
LG	0.5573	0.0509	0.3913	0.7181	0.6630	0.0514	0.4400	0.8140
GLSFS	0.5517	0.0495	0.3925	0.7104	0.6436	0.0499	0.4300	0.7842
CG	0.5416	0.0495	0.3879	0.7050	0.6320	0.0481	0.4558	0.7640

por funções canalizadoras, sendo novamente o método de agrupamento por canalização (CG) aquele que apresenta uma melhor distribuição para ambos os casos.

A Tabela 4.11 apresenta o grau médio e o erro quadrático médio de cada método em comparação com o gabarito. Os resultados são todos muito próximos entre si, com um erro ligeiramente superior no caso do método CG. Este resultado está de acordo com os resultados anteriores nos quais o método CG também foi o método que apresentou maior dificuldade para identificar os grau corretos.

Tabela 4.11: Grau médio (GM) e erro quadrático médio (EQM) das redes inferidas pelos 4 métodos com base em conjuntos de 30 e 50 amostras. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização. Nesse caso as redes gabaritos são compostas exclusivamente por funções linearmente separáveis.

Método	GM.30	EQM.30	GM.50	EQM.50
Gabarito	2.878	—	2.878	—
SA	2.564	1.286	3.066	1.345
LG	2.566	1.285	2.996	1.309
GLSFS	2.564	1.286	3.066	1.345
CG	2.637	1.300	3.102	1.416

A Figura 4.13 apresenta as matrizes de confusão dos graus em mapas de calor (*heatmap*), indicando que o comportamento também é similar aos cenários anteriores. O problema da superestimação do grau está presente com mais nitidez no caso de 50 amostras, como é possível observar nas células mais escuras abaixo da diagonal para cada método.

Em linhas gerais, o método de agrupamento linear (LG) apresenta uma pequena melhora em termos de F-Score em comparação aos resultados para gabaritos compostos por funções aleatórias e gabaritos compostos exclusivamente por funções de canalização. Já em relação a distribuição de graus, o comportamento é similar ao cenário de gabaritos exclusivamente compostos por funções canalizadoras.

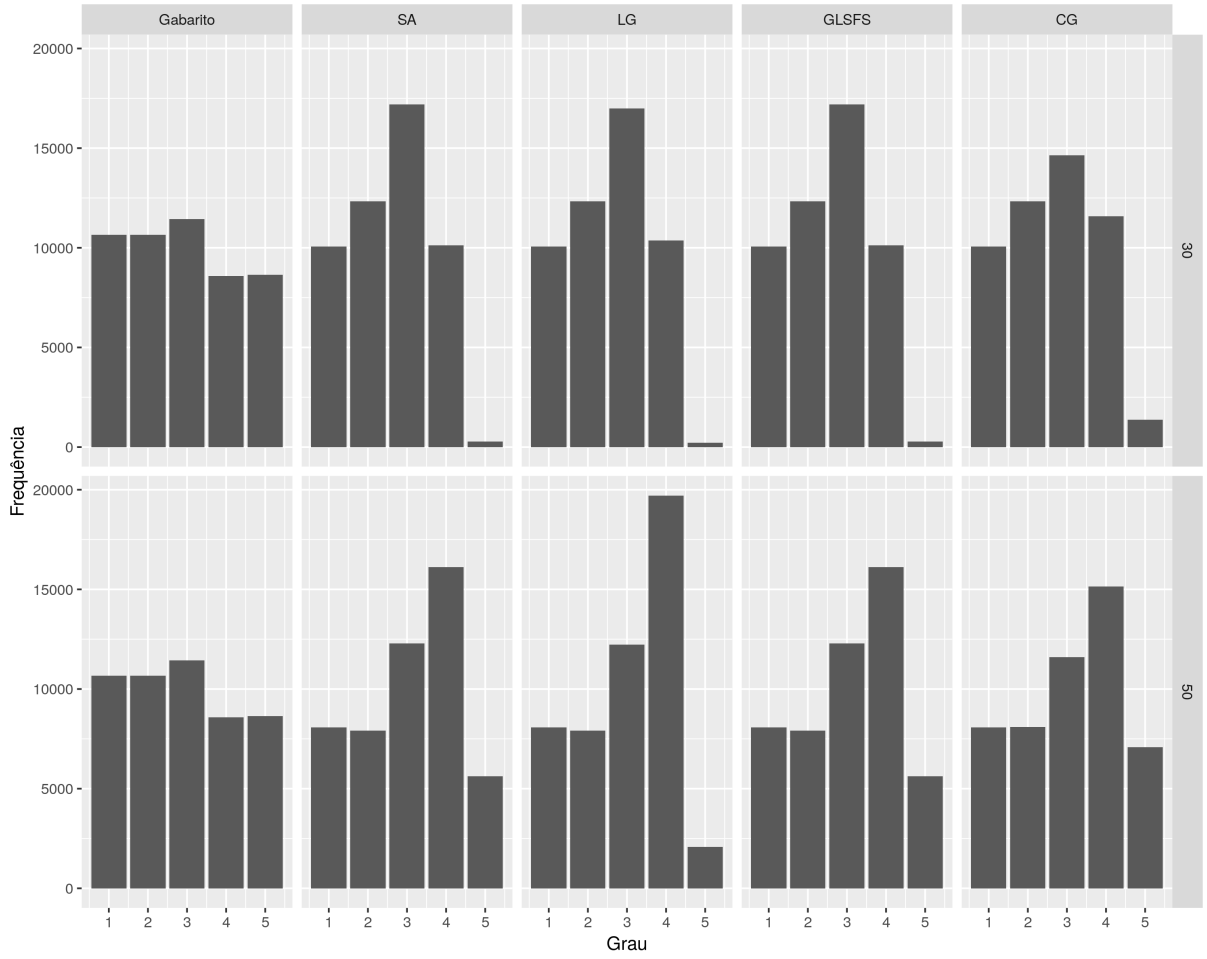


Figura 4.12: Histogramas de graus das redes gabaritos compostas exclusivamente por funções linearmente separáveis (à esquerda) e das redes inferidas pelos 4 métodos considerados com base em conjuntos de 30 amostras (topo) e 50 amostras (embaixo). SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização. A barra correspondente ao grau 5 representa o acúmulo das frequências de graus 5 e superior.

Avaliação da dinâmica gerada pelas redes inferidas

A Figura 4.14 apresenta os *Violin plots* correspondentes às taxas de acerto das dinâmicas geradas pelas 1000 redes inferidas pelos métodos comparados. Cada uma das redes inferidas gerou o próximo estado a partir de 1000 estados iniciais sorteados. A Tabela 4.12 apresenta um sumário desses resultados.

Neste cenário, os resultados da dinâmica possuem constatações semelhantes aos dos resultados apresentados na avaliação topológica. O método de agrupamento linear (LG) apresenta desempenho comparável ao método sem agrupamento (SA), mas pior que os métodos GLSFS e CG para conjuntos de poucas amostras ($M = 30$). Já para conjuntos com mais amostras ($M = 50$), o método de agrupamento linear (LG) se destaca dos

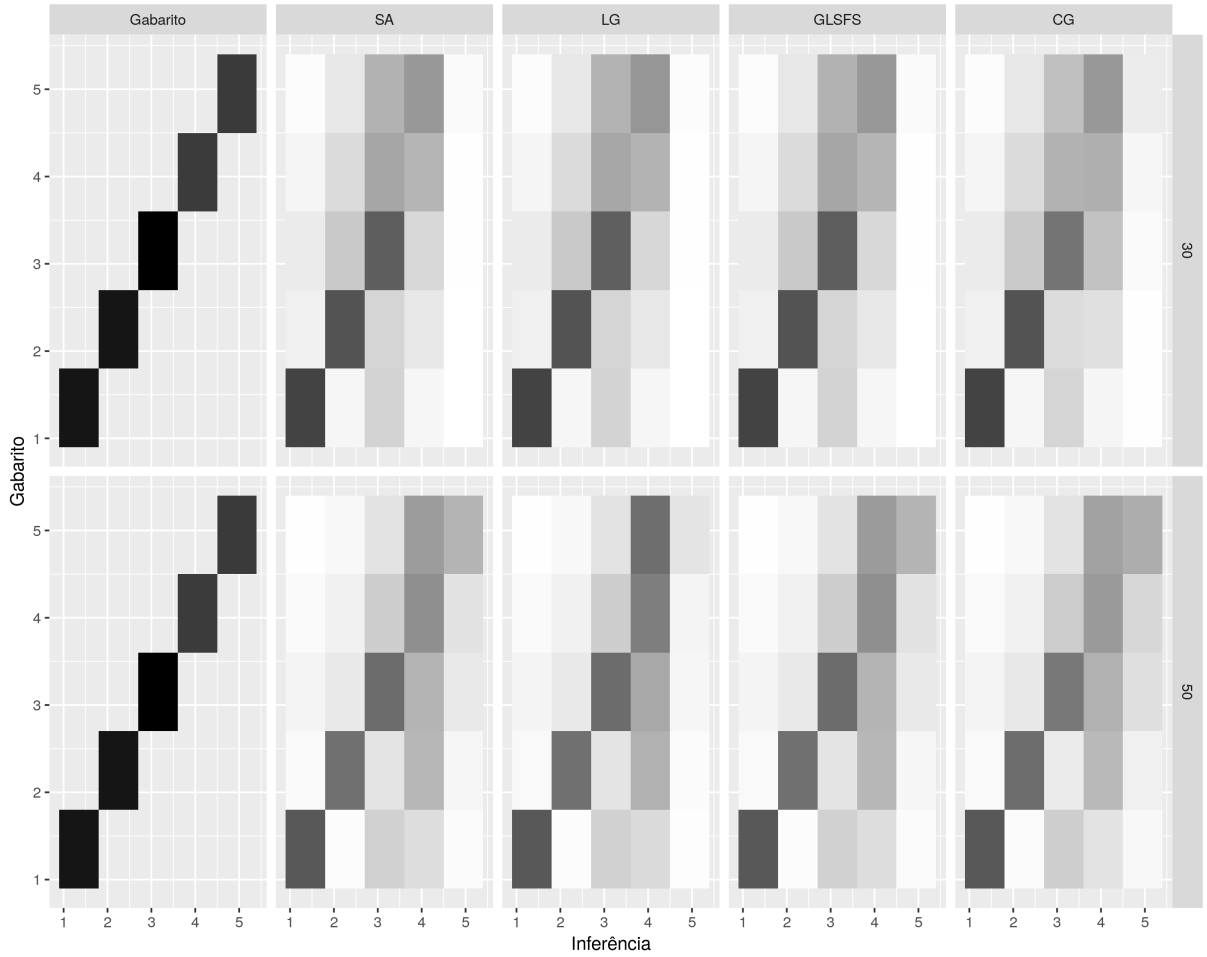


Figura 4.13: Mapas de calor (*heatmaps*) nos quais cada célula (i, j) representa a quantidade de vezes que foram inferidos conjuntos de preditores de dimensão i (linhas) para genes alvos que possuem conjuntos de preditores de dimensão j (colunas) nas redes gabaritos compostas exclusivamente por funções linearmente separáveis. Os *heatmaps* indicados por "Gabarito" representam os heatmaps ideais. Quanto mais escuro, maior é a proporção. Número de amostras $= \{30, 50\}$. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

demais. Essa constatação está de acordo com os resultados de avaliação topológica, para o a qual o método LG também obteve o melhor resultado.

Uma observação interessante é que tanto para as dinâmicas geradas por redes gabaritos compostas exclusivamente por funções canalizadoras (Seção 4.2.2) quanto para as dinâmicas geradas por gabaritos compostas exclusivamente por funções linearmente separáveis, apenas os métodos que fazem algum tipo de agrupamento melhoram consideravelmente seu poder de estimação. Esses resultados apontam que a abordagem de agrupamentos de instâncias aumenta o poder de estimação especialmente para redes compostas majoritariamente por funções que possuem algum significado biológico.

Ao avaliar as proporções de instâncias não observadas exigidas na predição dos genes al-

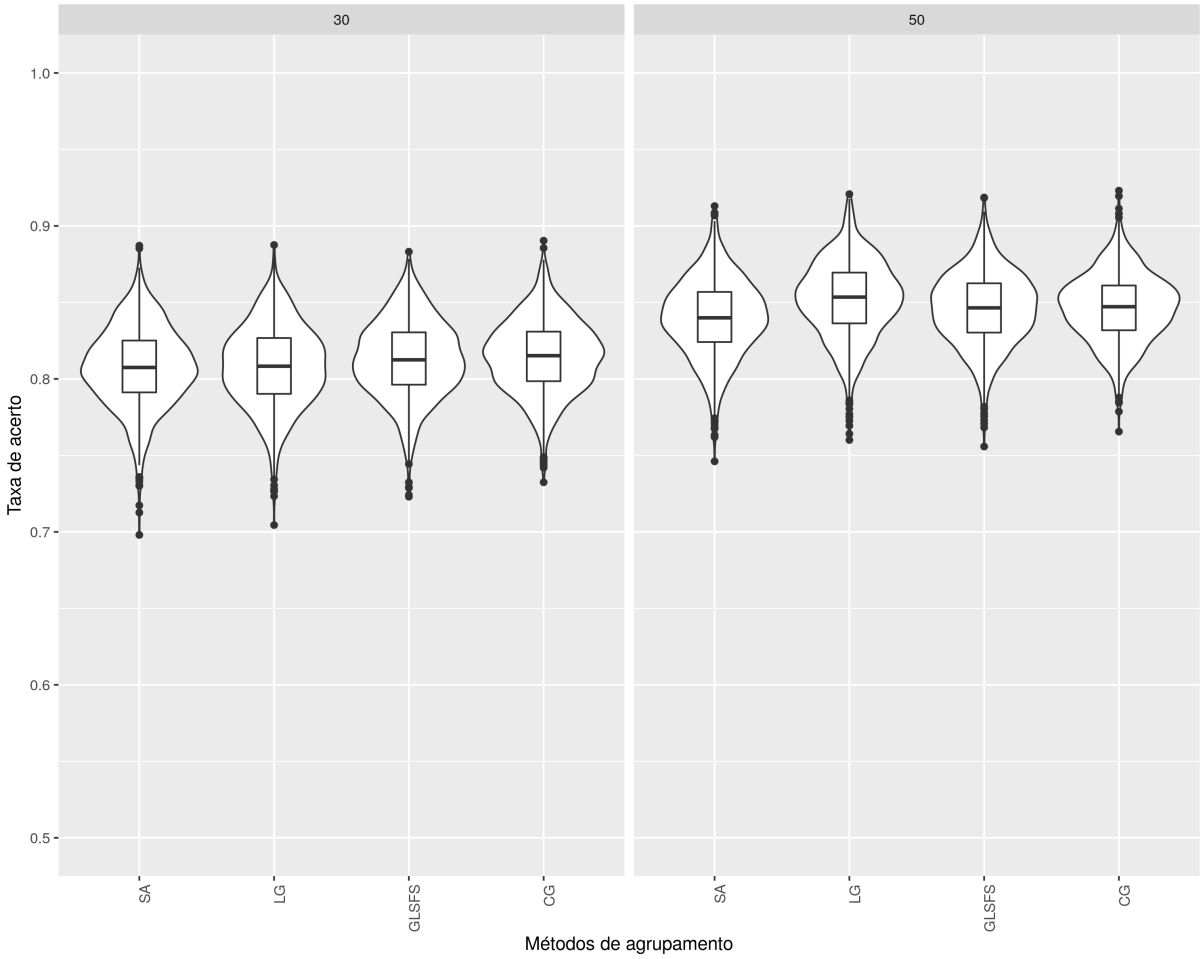


Figura 4.14: *Violin plots* dos valores de taxa de acerto para redes inferidas para 30 amostras (à esquerda) e 50 amostras (à direita), considerando redes gabaritos compostas exclusivamente por funções linearmente separáveis. Cada gráfico contém 4 *Violin plots*, um para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de valores de taxa de acerto para 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado a partir de 1000 estados iniciais sorteados.

Tabela 4.12: Sumário dos valores de taxa de acerto das dinâmicas geradas pelas redes inferidas, para 30 e 50 amostras, considerando gabaritos compostos exclusivamente por funções linearmente separáveis. Cada dado corresponde a média, desvio padrão, mínimo e máximo dos valores de taxa de acerto de 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado a partir de 1000 estados iniciais sorteados.

Método	30 amostras				50 amostras			
	Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	0.8076	0.0263	0.6980	0.8871	0.8397	0.0254	0.7461	0.9130
LG	0.8080	0.0266	0.7044	0.8875	0.8520	0.0254	0.7601	0.9208
GLSFS	0.8128	0.0254	0.7230	0.8832	0.8458	0.0240	0.7558	0.9186
CG	0.8147	0.0240	0.7324	0.8904	0.8465	0.0223	0.7655	0.9231

vos, as constatações são similares ao cenário de redes gabaritos compostas exclusivamente por funções canalizadoras, conforme pode ser observado na Figura 4.15 e na Tabela 4.13.

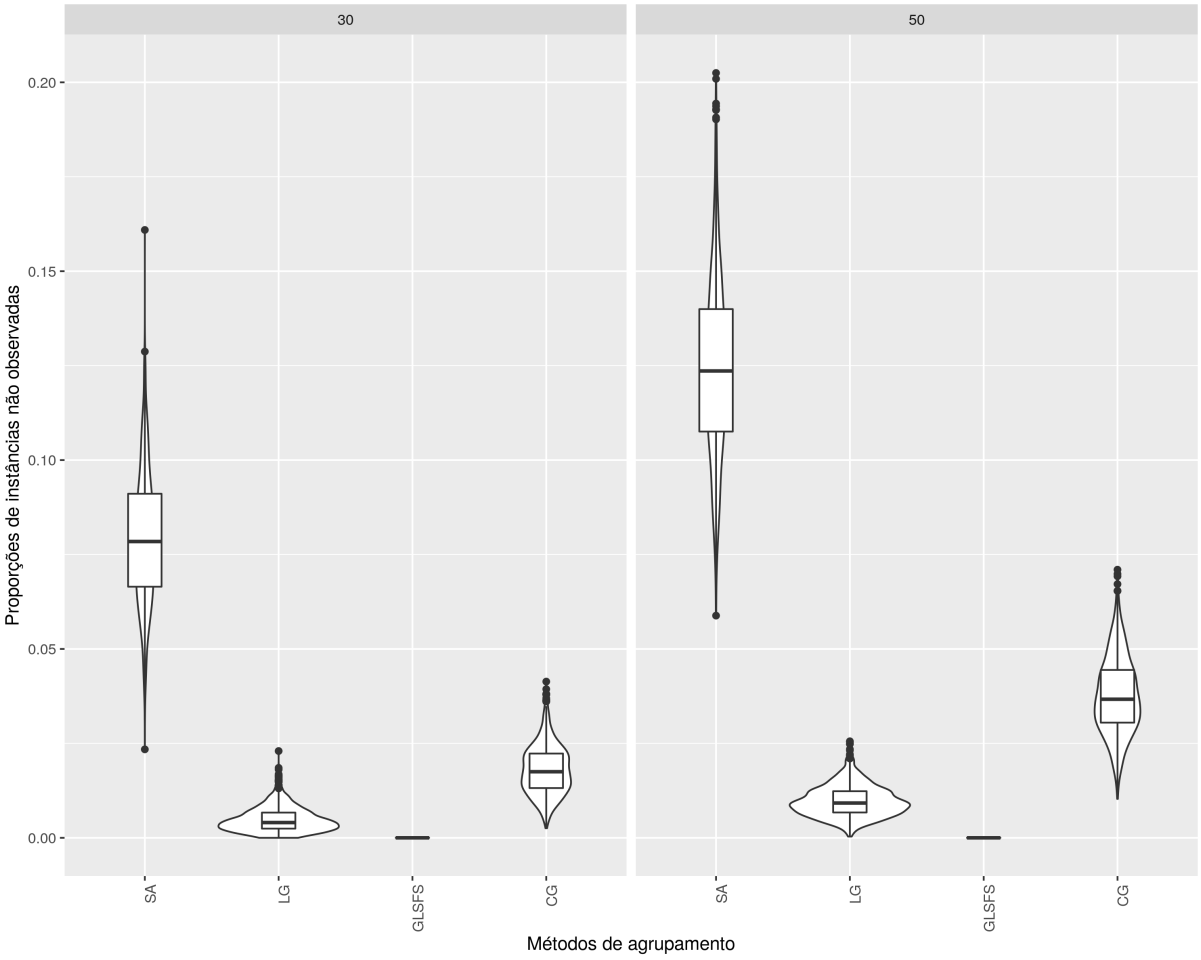


Figura 4.15: *Violin plots* das proporções de instâncias não observadas exigidas na predição dos valores dos genes alvos para redes inferidas para 30 amostras (à esquerda) e 50 amostras (à direita). Cada gráfico contém 4 *Violin plots*, um para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado de 1000 estados iniciais sorteados. As redes gabaritos são compostas exclusivamente por funções linearmente separáveis.

Dado que muitas das instâncias não observadas exigidas para a geração da dinâmica pode ser resultante de uma superestimação do grau, o próximo capítulo (Capítulo 5) discorrerá sobre os resultados da aplicação do método de aprendizado supervisionado dos graus ideais a partir dos perfis de evolução da função critério (Seção 3.5), bem como o impacto positivo que isso tende a causar tanto na topologia como na dinâmica das redes inferidas por todos os métodos considerados.

Tabela 4.13: Sumário das proporções de instâncias não observadas exigidas na predição dos valores dos genes alvos, para 30 amostras e 50 amostras, considerando redes gabaritos compostas exclusivamente por funções linearmente separáveis. Cada dado corresponde a média, o desvio padrão, mínimo e máximo dos proporções de 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado a partir de 1000 estados iniciais sorteados.

Método	30 amostras				50 amostras			
	Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	0.0793	0.0179	0.0234	0.1609	0.1244	0.0243	0.0588	0.2025
LG	0.0048	0.0033	0.0000	0.0230	0.0097	0.0041	0.0003	0.0256
GLSFS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CG	0.0180	0.0066	0.0025	0.0414	0.0376	0.0103	0.0103	0.0710

Capítulo 5

Resultados experimentais para o aprendizado dos graus

Os resultados do capítulo anterior (Capítulo 4) apresentaram o desafio de se obter a dimensão ideal do conjunto de preditores, especialmente a superestimação para os graus menores (1, 2, 3), além da subestimação para os graus maiores (4 em diante). Embora os resultados do F-SCORE melhorem quando o número de amostras aumenta (conforme Tabelas 4.2, 4.6, 4.10), o erro quadrático médio dos graus das redes inferidas também aumentou, sugerindo que o aumento de amostras por um lado melhora a estimação estatística das probabilidades condicionais, mas por outro lado tende a implicar em uma superestimação dos graus (conforme Tabelas 4.3, 4.7 e 4.11),

Para lidar com esse problema, projetamos um algoritmo que tenta aprender o grau correto com base na evolução da informação mútua normalizada com a dimensão imediatamente menor (Seção 2.3.6). A Seção 3.5 apresenta uma descrição desse método. Neste capítulo são apresentados os resultados obtidos da aplicação desse método.

5.1 Protocolo experimental

Os seguintes experimentos foram realizados seguindo o mesmo protocolo experimental do capítulo anterior (Seção 4.1). Com o objetivo de obter conjuntos de dados de treinamento independentes dos dados de validação foi necessário um novo conjunto de dados com as mesmas caraterísticas dos dados de validação (Seção 4.1.5). Foi aplicado sobre este conjunto de dados o método de treinamento apresentado na Seção 3.5 adotando o algoritmo de classificação supervisionada k-vizinhos mais próximos (k-nn) conforme descrito na Seção 2.3.7. Em seguida, o método foi treinado sobre o conjunto de treinamento contendo as mesmas redes do Capítulo 4, com o objetivo de comparar a eficácia da aplicação dessa

abordagem com a ausência dela.

5.2 Resultados para redes com funções aleatórias

Nesta seção foram comparados os resultados apresentados na Seção 4.2.1, com os resultados do aprendizado de grau por k-nn, usando as mesmas redes gabaritos e os mesmos dados de expressão.

5.2.1 Avaliação das topologias das redes inferidas

A Figura 5.1 apresenta a comparação entre os resultados do F-SCORE para todos os métodos de inferência tanto para os resultados aplicando a informação mútua (IM), como os resultados aplicando o aprendizado do grau por k-nn (KNN). Já a Tabela 5.1 apresenta um sumário da média, desvio padrão, mínimo e máximo das medidas de F-SCORE para todos os métodos.

Tabela 5.1: Sumário da comparação dos métodos sem aprendizado de grau (critério de parada baseado apenas na evolução da informação mútua: IM) e com aprendizado de grau (KNN) dos valores de F-Score para 30 e 50 amostras. Cada dado corresponde a média, o desvio padrão, o mínimo e o máximo dos valores de F-Score de 1000 redes inferidas, para cada método de inferência.

Método	KNN - IM	30 amostras				50 amostras			
		Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	IM	0.562	0.054	0.371	0.721	0.693	0.050	0.537	0.844
	KNN	0.574	0.056	0.375	0.7433	0.756	0.054	0.594	0.919
	Dif.	0.013	0.003	0.004	0.022	0.063	0.005	0.057	0.074
LG	IM	0.551	0.057	0.373	0.723	0.705	0.049	0.561	0.847
	KNN	0.563	0.059	0.386	0.759	0.752	0.054	0.604	0.913
	Dif.	0.012	0.002	0.013	0.035	0.047	0.004	0.043	0.066
GLSFS	IM	0.562	0.054	0.371	0.721	0.693	0.050	0.537	0.844
	KNN	0.575	0.057	0.378	0.749	0.756	0.054	0.592	0.913
	Dif.	0.013	0.003	0.007	0.028	0.064	0.004	0.055	0.068
CG	IM	0.506	0.053	0.332	0.654	0.634	0.048	0.476	0.778
	KNN	0.520	0.056	0.347	0.698	0.691	0.053	0.540	0.843
	Dif.	0.014	0.003	0.015	0.044	0.057	0.005	0.064	0.065

Como pode-se observar, todos os métodos apresentam melhoras no F-SCORE ao aplicar o aprendizado por KNN, sendo esta melhora mais substancial no caso de 50 amostras. A diferença entre KNN e IM para cada método, apresenta melhoras para todos os métodos, considerando a média, o mínimo e o máximo. Para 30 amostras a melhora da média varia entre 0.012 para o método LG e 0.014 para o método CG. Já para 50 amostras, a melhora na média varia entre 0.047 para o método LG e 0.064 para GLSFS. Isso se deve a

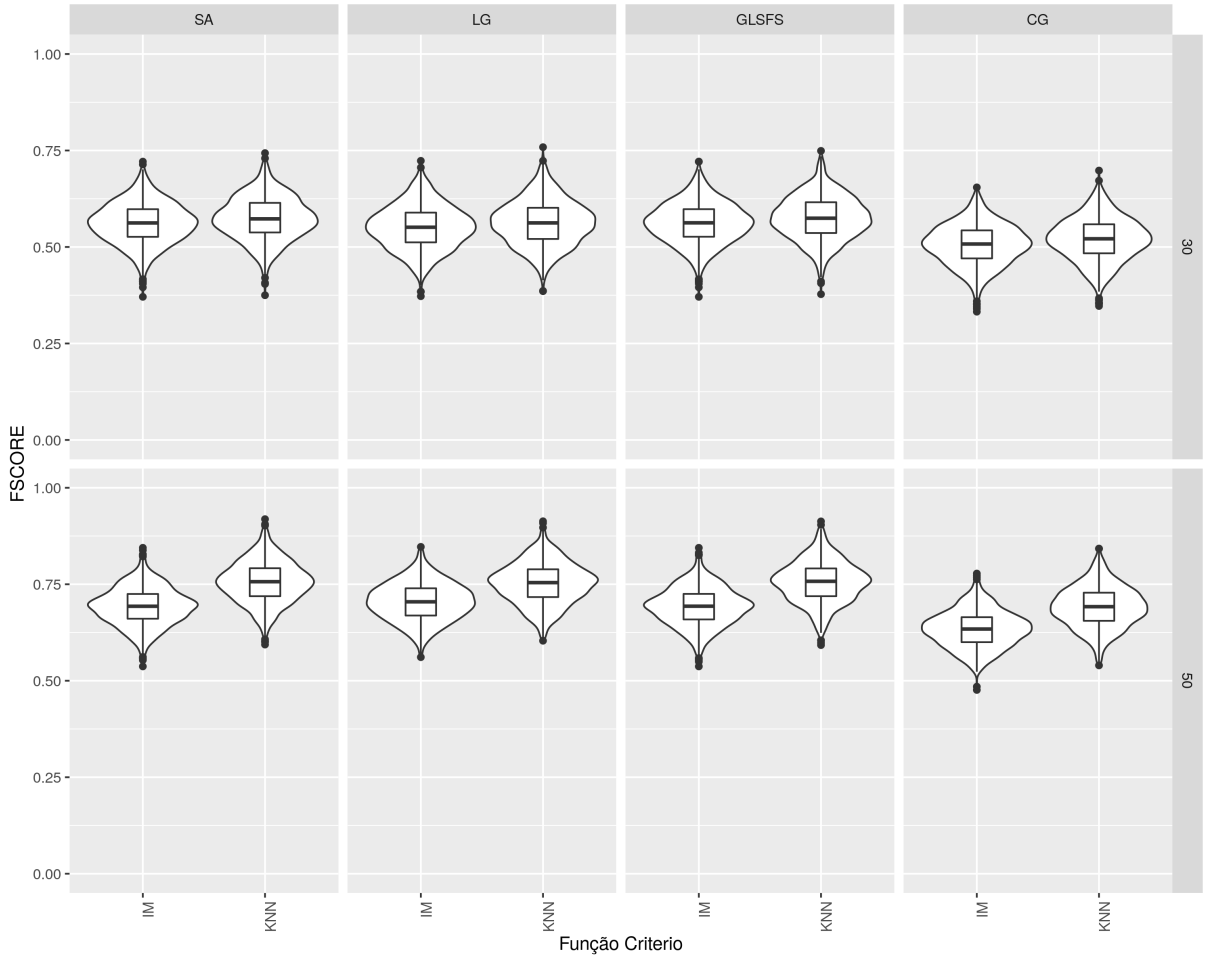


Figura 5.1: *Violin plots* da comparação dos métodos sem aprendizado de grau (critério de parada baseado apenas na evolução da informação mútua: IM) e com aprendizado de grau (KNN) dos valores de F-Score para 30 amostras (topo) e 50 amostras (embaixo). Cada gráfico contém 8 *Violin plots*, agrupados dois a dois (IM - KNN) para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de valores de F-Score de 1000 redes inferidas.

uma diminuição dos falsos positivos nas redes inferidas, visto que os métodos diminuem o tamanho dos conjuntos de seus preditores (aliviam a superestimação do grau), ao mesmo tempo que o F-SCORE apresenta melhoras em todos os casos.

Seguindo com a análise topológica das redes inferidas, as Figuras 5.2 e 5.3 apresentam os perfis das distribuições dos graus para 30 e 50 amostras, respectivamente, comparando IM com KNN. Em todos os casos pode-se observar uma correção do viés de superestimação dos graus pelos métodos, já que os perfis de distribuições estão mais parecidos com os perfis dos gabaritos usando o KNN. Essa constatação é confirmada pelos *headmaps* das matrizes de confusão apresentados nas Figuras 5.4 (30 amostras) e 5.5 (50 amostras), já que o KNN implica em redução de células escuras abaixo da diagonal principal, ou seja, corrige os

graus que haviam sido superestimados. Tal correção é mais evidenciada no cenário com 50 amostras.

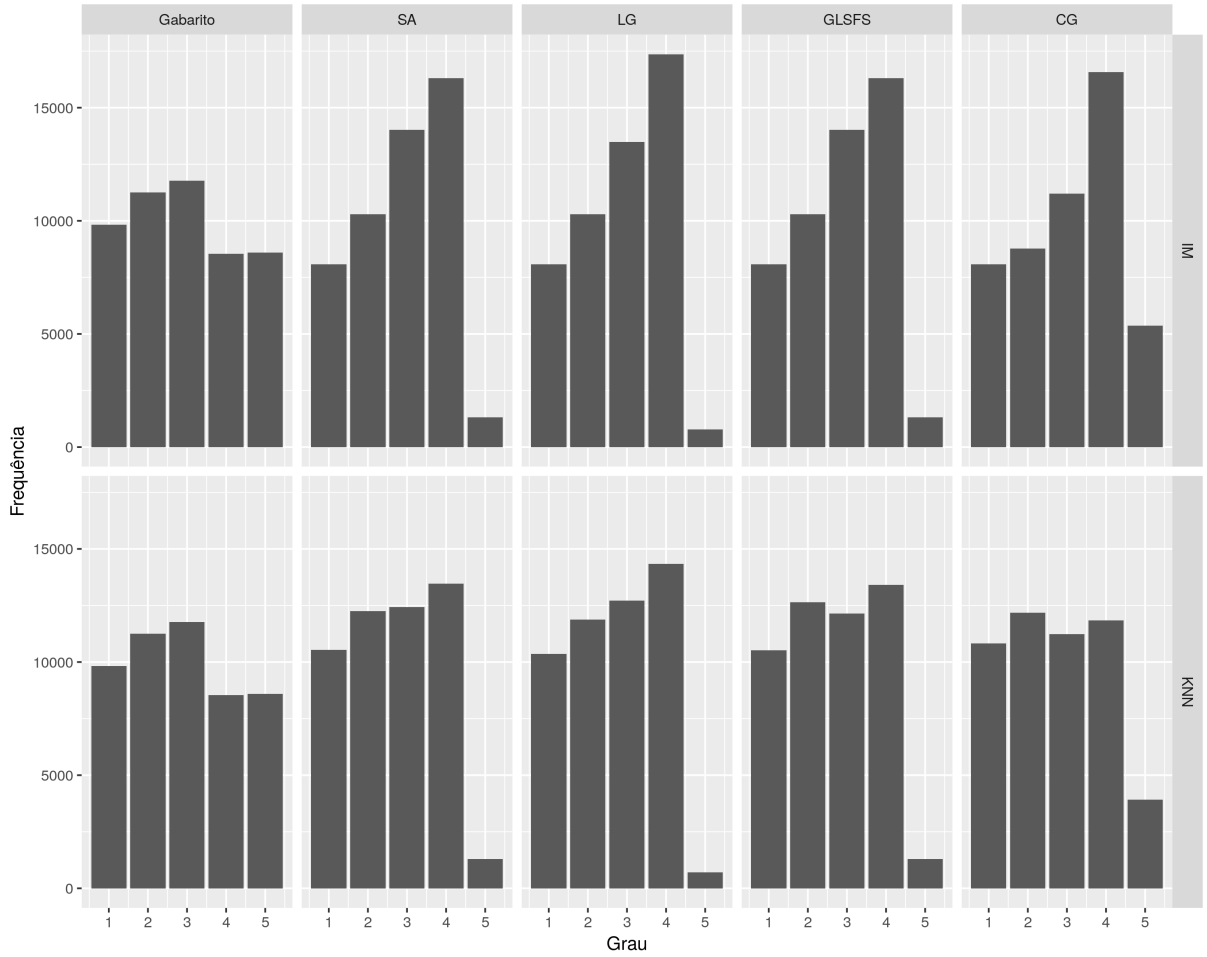


Figura 5.2: Histogramas de graus das redes gabaritos (à esquerda) e das redes inferidas pelos 4 métodos considerados com base em conjuntos de 30 amostras, para IM (topo) e KNN (em-baixo). SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

Na Tabela 5.2 pode-se observar tanto os graus médios, como os erros quadráticos médios para o IM e KNN. É possível notar que para 30 amostras, mesmo o valor do grau médio ficando mais distante do grau médio do gabarito ao aplicar o KNN, reduzindo o grau médio para todos os métodos, o erro quadrático acabou sendo menor para todos os casos com o KNN. Isso mostra que o KNN apresentou uma melhora na estimativa do grau, o que sugere que na maioria dos casos os graus diminuídos são os graus superestimados. Já no cenário de 50 amostras, a melhora é mais significativa, pois os erros quadráticos médios foram ainda menores com o uso do KNN. Isso mostra que o aumento de amostras é benéfico para o uso do KNN, evitando a superestimação dos graus que ocorre sem o seu uso.

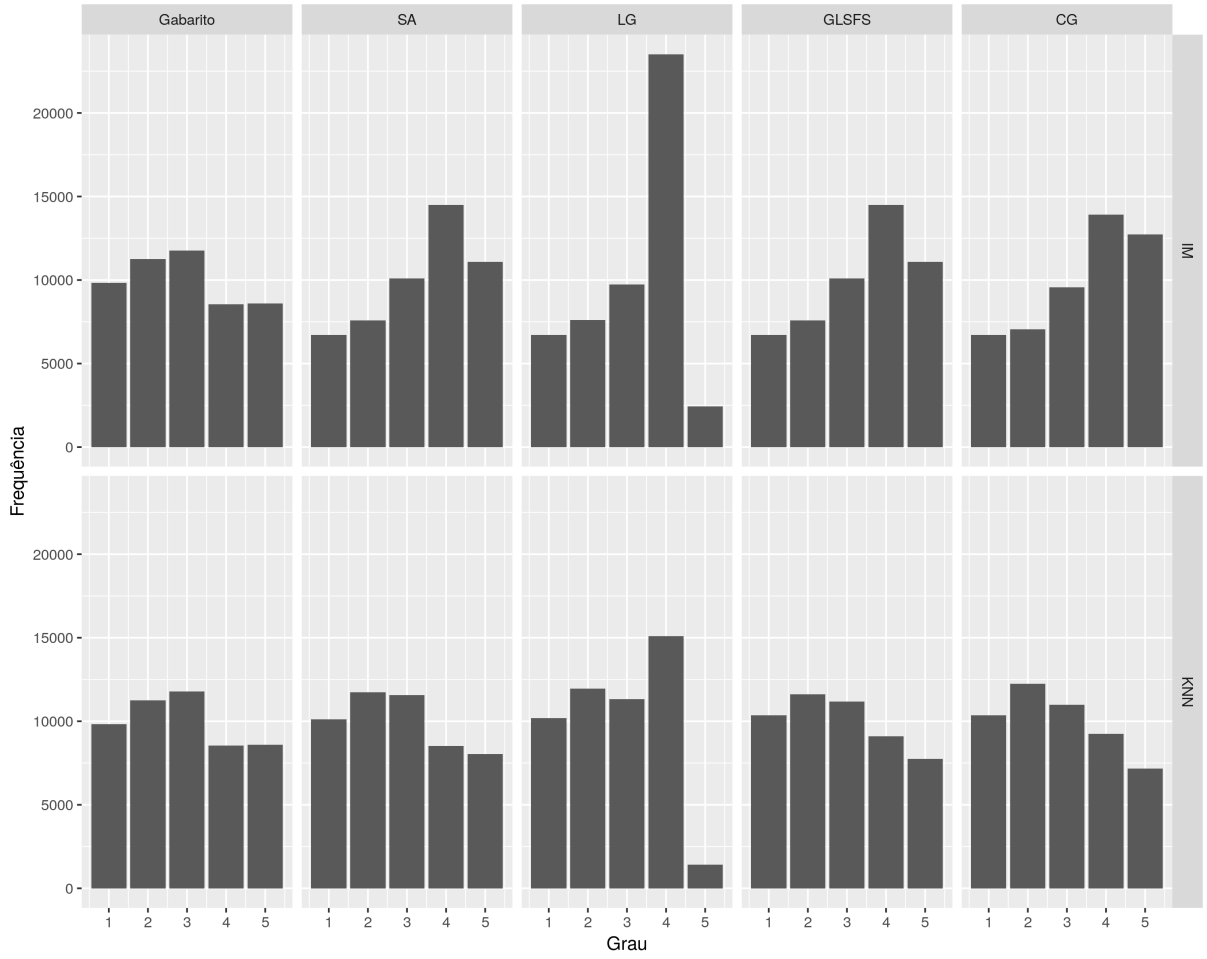


Figura 5.3: Histogramas de graus das redes gabaritos (à esquerda) e das redes inferidas pelos 4 métodos considerados com base em conjuntos de 50 amostras, para IM (topo) e KNN (embaixo). SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

5.2.2 Avaliação das dinâmicas geradas pelas redes inferidas

A Figura 5.6 apresenta a comparação das taxas de acerto das dinâmicas geradas pelas redes inferidas pelos 4 métodos de inferência, tanto para o IM como para o KNN. Um sumário desses resultados é apresentado na Tabela 5.3. Conforme pode ser observado, o aprendizado por KNN implicou em resultados melhores para todos os cenários, sendo essa melhora mais evidenciada para um maior numero de amostras. Estes resultados estão de acordo com os resultados topológicos, confirmando que o KNN melhora tanto as topologias das redes inferidas como também a geração das dinâmicas por essas redes. O método SA foi o mais beneficiado pela aplicação do KNN para o cenário com 30 amostras, com uma melhora da média em 0.01. Já para o cenário de 50 amostras, o GLSFS foi o mais beneficiado, com uma melhora da média em 0.03.

Em relação a proporção de instâncias não observadas exigidas na geração das dinâmicas

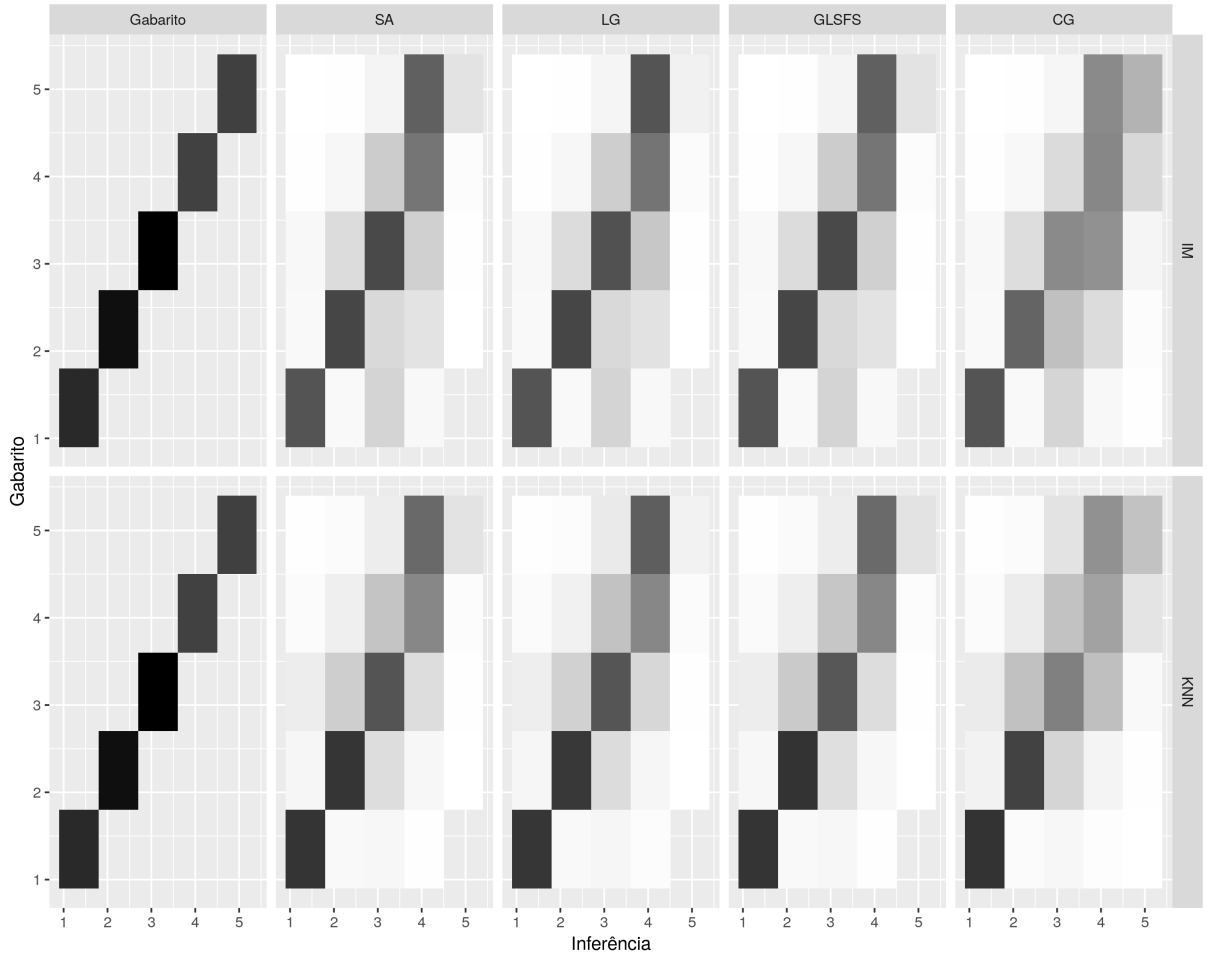


Figura 5.4: *Heatmaps* onde cada célula (i, j) representa a proporção de vezes que foram inferidos conjuntos de preditores de dimensão i (linhas) para genes alvos que possuem conjuntos de preditores de dimensão j (colunas) nas redes gabaritos. Os *heatmaps* indicados por "Gabarito" representam os *heatmaps* ideais. Quanto mais escuro o tom de cinza, maior é a proporção. Número de amostras = 30. IM (topo) e KNN (embaixo). SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

pelas redes inferidas, a Figura 5.7 apresenta os *violin plots* desses resultados, enquanto a Tabela 5.4 apresenta um sumário desses resultados. Pode-se observar que o KNN diminui a proporção de instâncias não observadas para todos os métodos de inferência. Ainda assim o método SA apresenta as maiores proporções de instâncias não observadas em comparação aos métodos de agrupamento. Como esperado, o método SA permanece como o que possui a maior dificuldade de generalização, visto que a ausência de agrupamento implica em tabelas de frequências mais rarefeitas, com um menor poder de estimação das probabilidades condicionais.

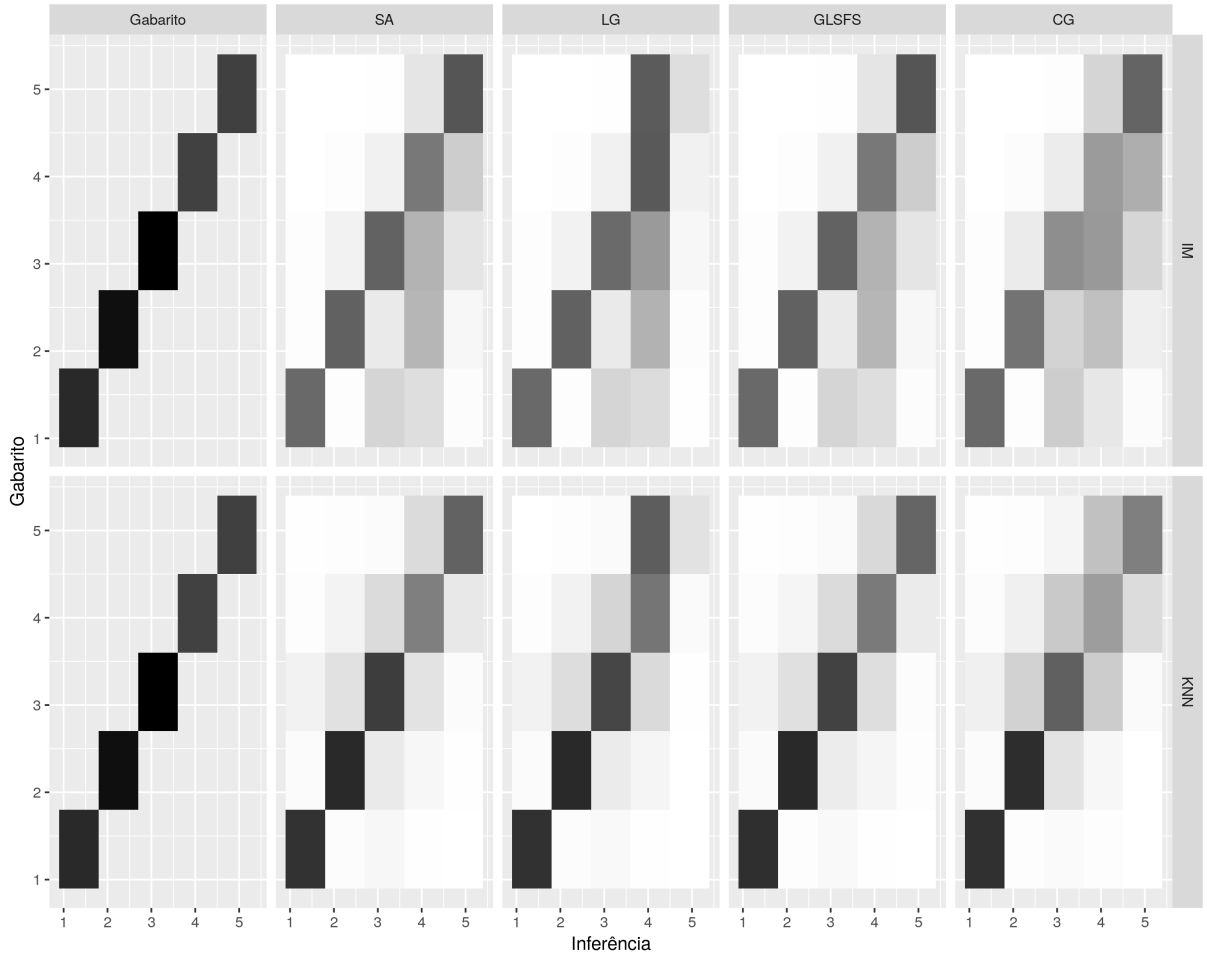


Figura 5.5: *Heatmaps* onde cada célula (i, j) representa a quantidade de vezes que foram inferidos conjuntos de preditores de dimensão i (linhas) para genes alvos que possuem conjuntos de preditores de dimensão j (colunas) nas redes gabaritos. Os *heatmaps* indicados por "Gabarito" representam os *heatmaps* ideais. Quanto mais escuro o tom de cinza, maior é a proporção. Número de amostras = 50. IM (topo) e KNN (embaixo). SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

5.3 Resultados para redes com funções canalizadoras

Nesta seção será examinado o desempenho do aprendizado dos graus por KNN para redes gabaritos compostas exclusivamente por funções canalizadoras.

5.3.1 Avaliação das topologias das redes inferidas

A Figura 5.8 apresenta a comparação entre os resultados do F-SCORE para todos os métodos de inferência tanto para os resultados aplicando a informação mútua (IM), como os resultados aplicando o aprendizado do grau por k-nn (KNN). Já a Tabela 5.5 apresenta

Tabela 5.2: Grau Médio (GM) e Erro Quadrático Médio (EQM) das redes gabaritos e das redes inferidas, pelos 4 métodos considerados com base em conjuntos de 30 amostras e 50 amostras, comparando IM com KNN para cada método. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

Método	KNN - IM	GM.30	EQM.30	GM.50	EQM.50
Gabarito		2.897	—	2.897	—
SA	IM	2.850	0.676	3.313	0.995
SA	KNN	2.655	0.604	2.852	0.370
LG	IM	2.850	0.695	3.147	1.026
LG	KNN	2.663	0.614	2.712	0.471
GLSFS	IM	2.850	0.676	3.313	0.995
GLSFS	KNN	2.647	0.611	2.846	0.362
CG	IM	3.048	0.805	3.378	1.149
CG	KNN	2.717	0.724	2.812	0.480

Tabela 5.3: Sumário dos valores de taxas de acerto das dinâmicas geradas pelas redes inferidas, para 30 e 50 amostras. Cada dado corresponde a média, desvio padrão, mínimo e máximo dos valores de taxa de acerto de 1000 redes inferidas, para cada método de inferência, comparando IM e KNN.

Metodo		30 amostras				50 amostras			
		Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	IM	0.7984	0.0263	0.7113	0.8822	0.8464	0.0235	0.7710	0.9245
	KNN	0.8088	0.0274	0.7228	0.8894	0.8763	0.0240	0.7888	0.9531
	Dif.	0.0103	0.0011	0.0115	0.0072	0.0299	0.0006	0.0177	0.0286
LG	IM	0.7918	0.0266	0.7009	0.8790	0.8405	0.0232	0.7728	0.9172
	KNN	0.8014	0.0271	0.7166	0.8838	0.8644	0.0238	0.7924	0.9414
	Dif.	0.0096	0.0005	0.0157	0.0048	0.0239	0.0006	0.0196	0.0242
GLSFS	IM	0.7991	0.0267	0.7195	0.8762	0.8463	0.0243	0.7671	0.9303
	KNN	0.8089	0.0274	0.7084	0.8874	0.8765	0.0240	0.7936	0.9465
	Dif.	0.0098	0.0007	-0.0111	0.0112	0.0302	-0.0002	0.0265	0.0162
CG	IM	0.7653	0.0247	0.6820	0.8332	0.8044	0.0224	0.7139	0.8683
	KNN	0.7732	0.0256	0.6825	0.8444	0.8237	0.0228	0.7502	0.8955
	Dif.	0.0079	0.0009	0.0005	0.0111	0.0193	0.0004	0.0364	0.0272

um sumário da média, desvio padrão, mínimo e máximo das medidas de F-SCORE para todos os métodos.

Pode-se observar que, embora todos os métodos apresentem melhoras com o uso do KNN, o maior beneficiado foi o método CG tanto para 30 como para 50 amostras. De fato, as diferenças entre KNN e IM aumentaram cerca de 3 a 4 vezes ao aumentar o número de amostras de 30 para 50. É importante notar também que para 50 amostras o método com melhor desempenho sem a aplicação do KNN foi o LG, enquanto o método CG obteve o melhor desempenho com a aplicação do KNN. Isso indica uma conexão do aprendizado dos graus com o tipo de funções considerado nas redes gabaritos.

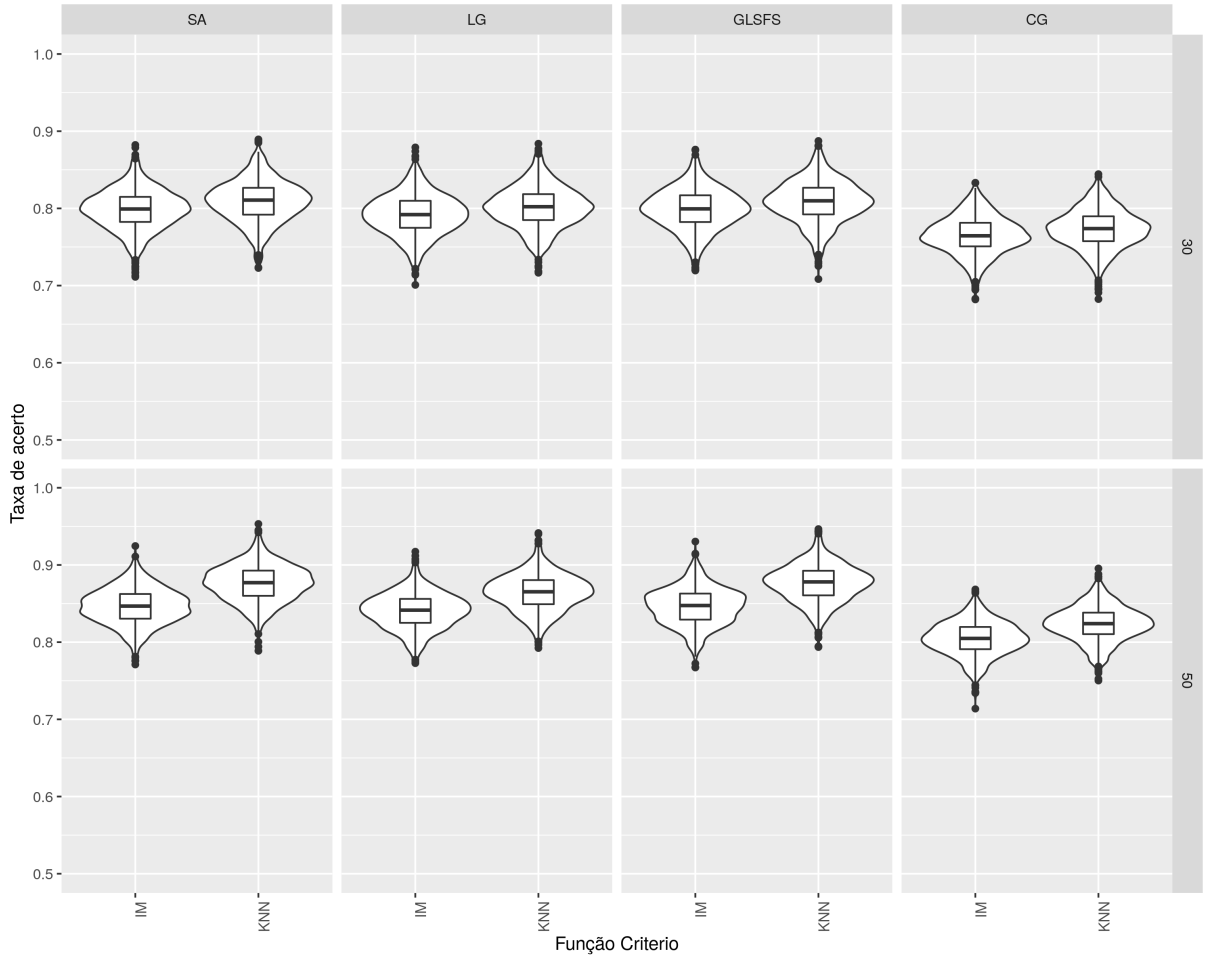


Figura 5.6: *Violin plots* dos valores de taxas de acerto das dinâmicas geradas pelas redes inferidas para 30 amostras (topo) e 50 amostras (embaixo). Cada gráfico contém 8 *Violin plots*, agrupados dois a dois (IM - KNN) para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições e CG agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de valores de taxa de acerto para 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado de 1000 estados iniciais sorteados.

Para observar a capacidade dos métodos em obter os graus corretos com e sem o uso do KNN, as Figuras 5.9 e 5.10 apresentam uma comparação entre os histogramas das redes gabaritos para 30 e 50 amostras, respectivamente. Já a Tabela 5.6 apresenta as respectivas médias e erros quadráticos médios (EQM) dos graus das redes gabaritos e das redes inferidas, com e sem o uso do KNN (KNN e IM, respectivamente). De modo geral pode-se observar um comportamento mais conservador com o uso do KNN reduzindo substancialmente a superestimação do grau, ao custo de subestimar o grau em alguns casos. Como esperado, os histogramas dos métodos estão mais similares aos do gabarito para 50 amostras. De fato, é possível notar que um maior número de amostras (50) implica em diminuição do EQM para todos os métodos. Além disso, como o F-SCORE aumenta em todos os casos conforme observado anteriormente, conclui-se que o KNN efetivamente

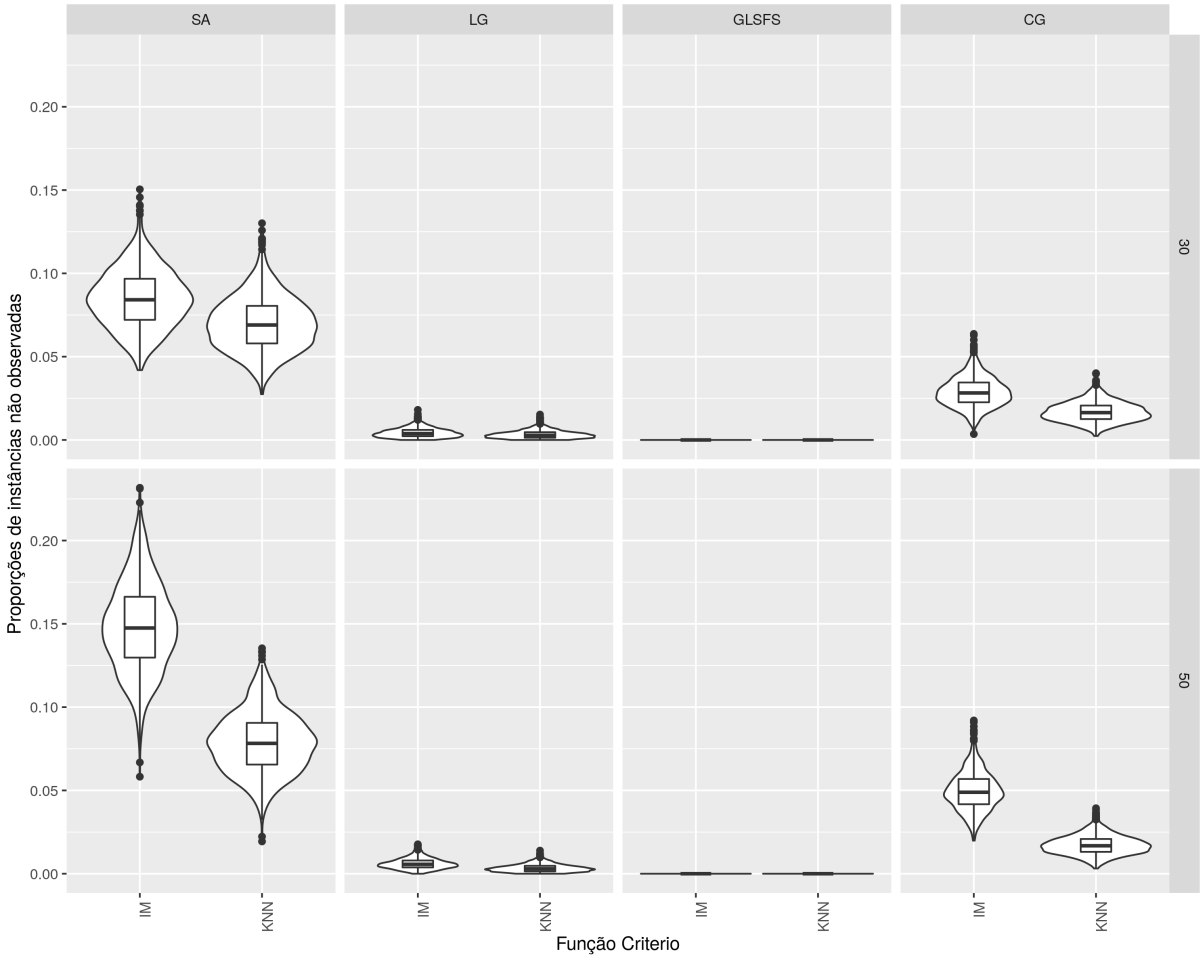


Figura 5.7: *Violin plots* das proporções de instâncias não observadas exigidas na predição dos valores dos genes alvos para redes inferidas para 30 amostras (topo) e 50 amostras (embaixo). Cada gráfico contém 2 *violin plots* por método de modo a comparar o efeito da aplicação do KNN e com a ausência dessa aplicação (IM). SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização. Cada *violin plot* corresponde a distribuição de 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado de 1000 estados iniciais sorteados.

aliviou o problema da superestimação e eliminou falsos positivos.

As Figuras 5.11 e 5.12 apresentam as matrizes de confusão (*heatmaps*) para 30 e 50 amostras, respectivamente. É possível ver em ambos os casos a redução de células escuras embaixo da diagonal principal para todos os métodos com o uso do KNN. Entretanto, o KNN induziu a um ligeiro escurecimento das células acima da diagonal, indicando um pequeno viés de subestimação dos graus.

Tabela 5.4: Sumário das proporções de instâncias não observadas exigidas pelas dinâmicas geradas pelas redes inferidas, para 30 amostras e 50 amostras. Cada dado corresponde a média, desvio padrão, mínimo e máximo das proporções de instâncias não observadas de 1000 redes inferidas para cada método de inferência.

Metodo		30 amostras				50 amostras			
		Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	IM	0.0848	0.0178	0.0419	0.1505	0.1482	0.0267	0.0582	0.2317
	KNN	0.0698	0.0163	0.0275	0.1301	0.0786	0.0182	0.0194	0.1353
	Dif.	-0.0151	-0.0015	-0.0144	-0.0204	-0.0695	-0.0086	-0.0389	-0.0963
LG	IM	0.0043	0.0029	0.0000	0.0181	0.0060	0.0032	0.0000	0.0177
	KNN	0.0033	0.0025	0.0000	0.0153	0.0034	0.0024	0.0000	0.0139
	Dif.	-0.0010	-0.0004	0.0000	-0.0028	-0.0026	-0.0008	0.0000	-0.0038
GLSFS	IM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	KNN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Dif.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CG	IM	0.0288	0.0091	0.0035	0.0637	0.0496	0.0117	0.0199	0.0920
	KNN	0.0168	0.0060	0.0024	0.0401	0.0173	0.0059	0.0031	0.0393
	Dif.	-0.0120	-0.0031	-0.0011	-0.0236	-0.0323	-0.0058	-0.0167	-0.0527

Tabela 5.5: Sumário dos valores de F-Score para redes inferidas, para 30 e 50 amostras, considerando redes gabaritos compostas exclusivamente por funções canalizadoras. Cada dado corresponde a média, desvio padrão, mínimo e máximo dos valores de F-Score de 1000 redes inferidas, para cada método de inferência.

Método		30 amostras				50 amostras			
		Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	IM	0.5495	0.0470	0.4152	0.7029	0.6324	0.0462	0.4911	0.7913
	KNN	0.5616	0.0494	0.4113	0.7204	0.6869	0.0506	0.5313	0.8475
	Dif.	0.0121	0.0025	-0.0039	0.0175	0.0546	0.0043	0.0401	0.0562
LG	IM	0.5533	0.0470	0.4148	0.6884	0.6424	0.0460	0.5051	0.7680
	KNN	0.5665	0.0496	0.4231	0.7177	0.6913	0.0495	0.5534	0.8608
	Dif.	0.0132	0.0027	0.0083	0.0293	0.0489	0.0035	0.0483	0.0928
GLSFS	IM	0.5495	0.0470	0.4152	0.6957	0.6321	0.0464	0.4840	0.7850
	KNN	0.5629	0.0502	0.4130	0.7143	0.6874	0.0500	0.5190	0.8333
	Dif.	0.0134	0.0032	-0.0022	0.0186	0.0553	0.0036	0.0350	0.0483
CG	IM	0.5603	0.0465	0.4189	0.7036	0.6388	0.0464	0.4674	0.7879
	KNN	0.5749	0.0497	0.4326	0.7464	0.6943	0.0496	0.5382	0.8438
	Dif.	0.0146	0.0032	0.0137	0.0429	0.0555	0.0032	0.0708	0.0559

5.3.2 Avaliação das dinâmicas geradas pelas redes inferidas

A Figura 5.13 apresenta os *Violin plots* dos valores de taxa de acerto das dinâmicas geradas pelas redes inferidas pelos 4 métodos considerados, com e sem o uso do KNN (IM, KNN), enquanto o sumário correspondente contendo a média, o desvio padrão, o mínimo e o máximo dessas taxas de acerto encontra-se na Tabela 5.7. Pode-se observar que todos os métodos apresentaram melhoras com o uso do KNN tanto para 30 amostras

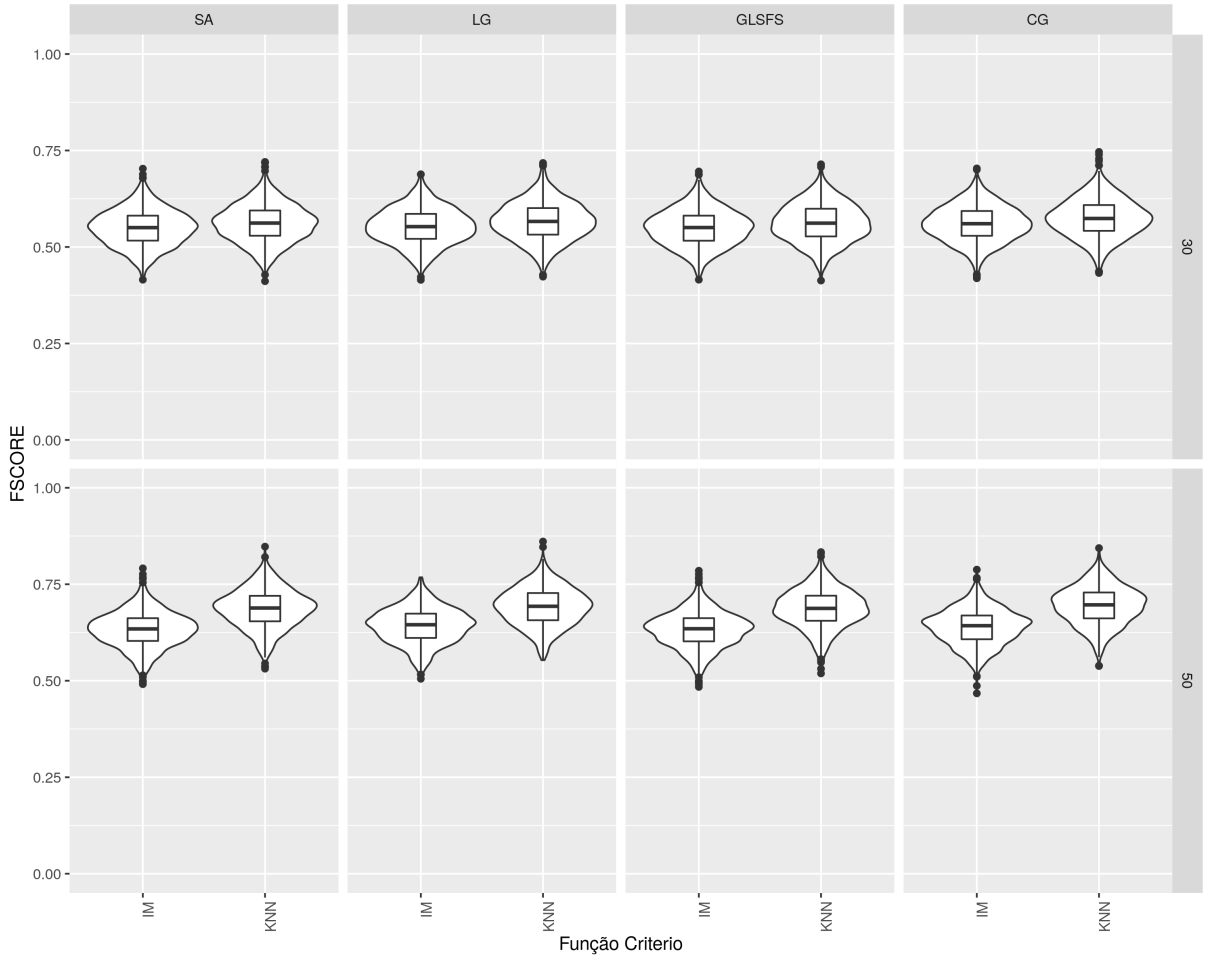


Figura 5.8: *Violin plots* da comparação dos métodos sem aprendizado de grau (critério de parada baseado apenas na evolução da informação mútua: IM) e com aprendizado de grau (KNN) dos valores de F-Score para 30 amostras (topo) e 50 amostras (embaixo), considerando redes gabaritos compostas exclusivamente por funções de canalização. Cada gráfico contém 8 *Violin plots*, agrupados dois a dois (IM - KNN) para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de valores de F-Score de 1000 redes inferidas.

quanto para 50 amostras, sendo essas melhoras mais evidentes no cenário de 50 amostras. O método LG foi o mais beneficiado no cenário de 30 amostras, enquanto o método SA foi o mais beneficiado no cenário de 50 amostras. Ainda assim o método CG foi o que obteve o melhor desempenho em ambos os cenários, como esperado. A melhora com o uso do KNN para 30 amostras foi de aproximadamente 1% para todos os métodos, e entre 2% e 3% no cenário de 50 amostras.

Com o objetivo de avaliar a capacidade de generalização dos métodos, a Figura 5.14 apresenta os *Violin plots* das proporções de instâncias não observadas exigidas na predição dos valores dos genes alvos, enquanto a Tabela 5.8 apresenta o sumário correspondente a

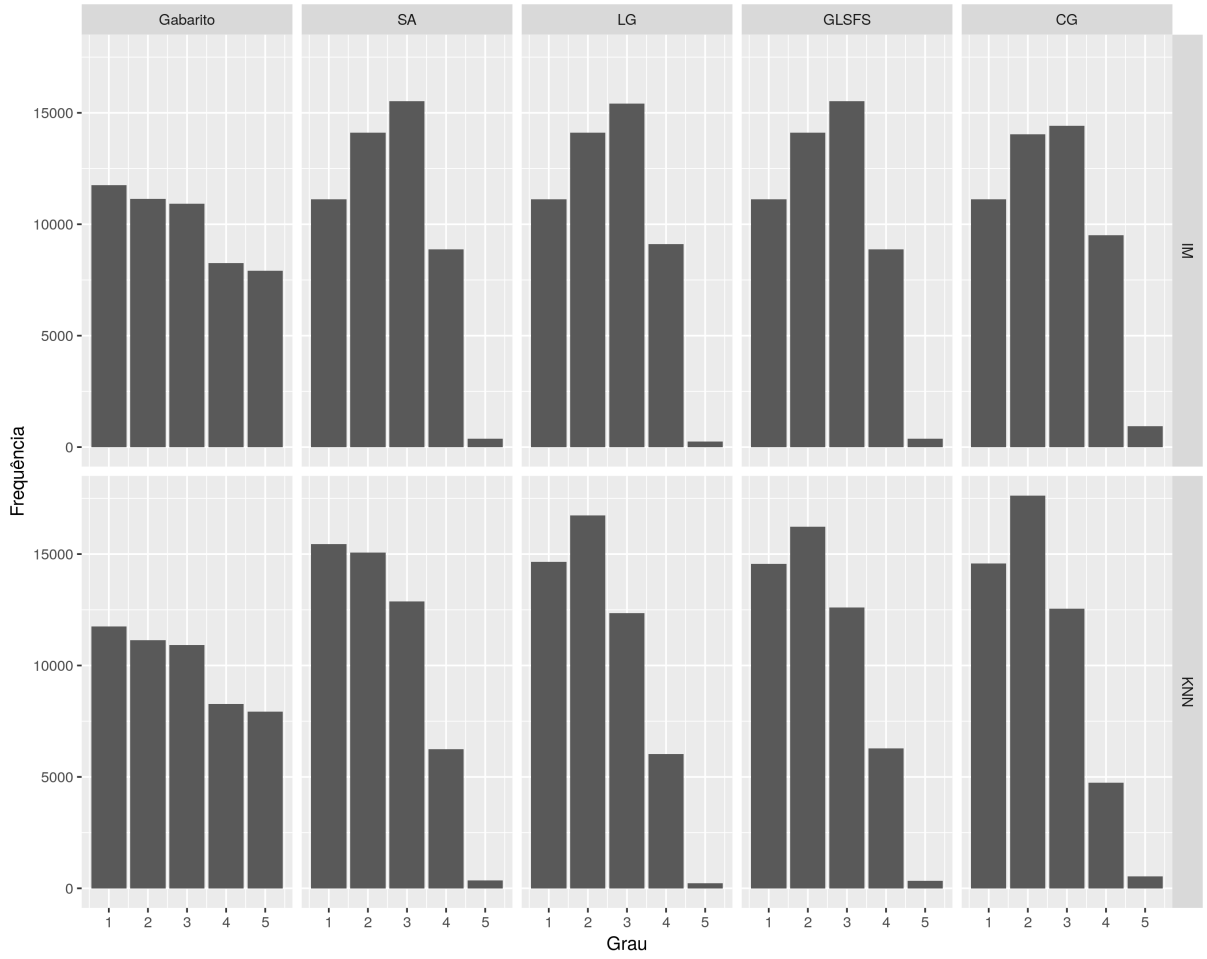


Figura 5.9: Histogramas de graus das redes gabaritos compostas exclusivamente por funções de canalização (à esquerda) e das redes inferidas pelos 4 métodos considerados com base em conjuntos de 30 amostras, para IM (topo) e KNN (embaixo). SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

esses valores. Podemos observar que o método SA apresenta uma melhora substancial com o uso do KNN, embora ele continue sendo o método com pior capacidade de generalização.

5.4 Resultados para redes com funções linearmente separáveis

Nesta seção examinamos o desempenho dos métodos de inferência com o aprendizado dos graus por KNN para lidar com amostras geradas por redes gabaritos compostas exclusivamente por funções linearmente separáveis, visando entender como este tipo de função influencia no desempenho dos métodos de inferência. Os experimentos foram realizados sobre os mesmos dados gerados na Seção 4.2.3 para promover uma comparação entre os

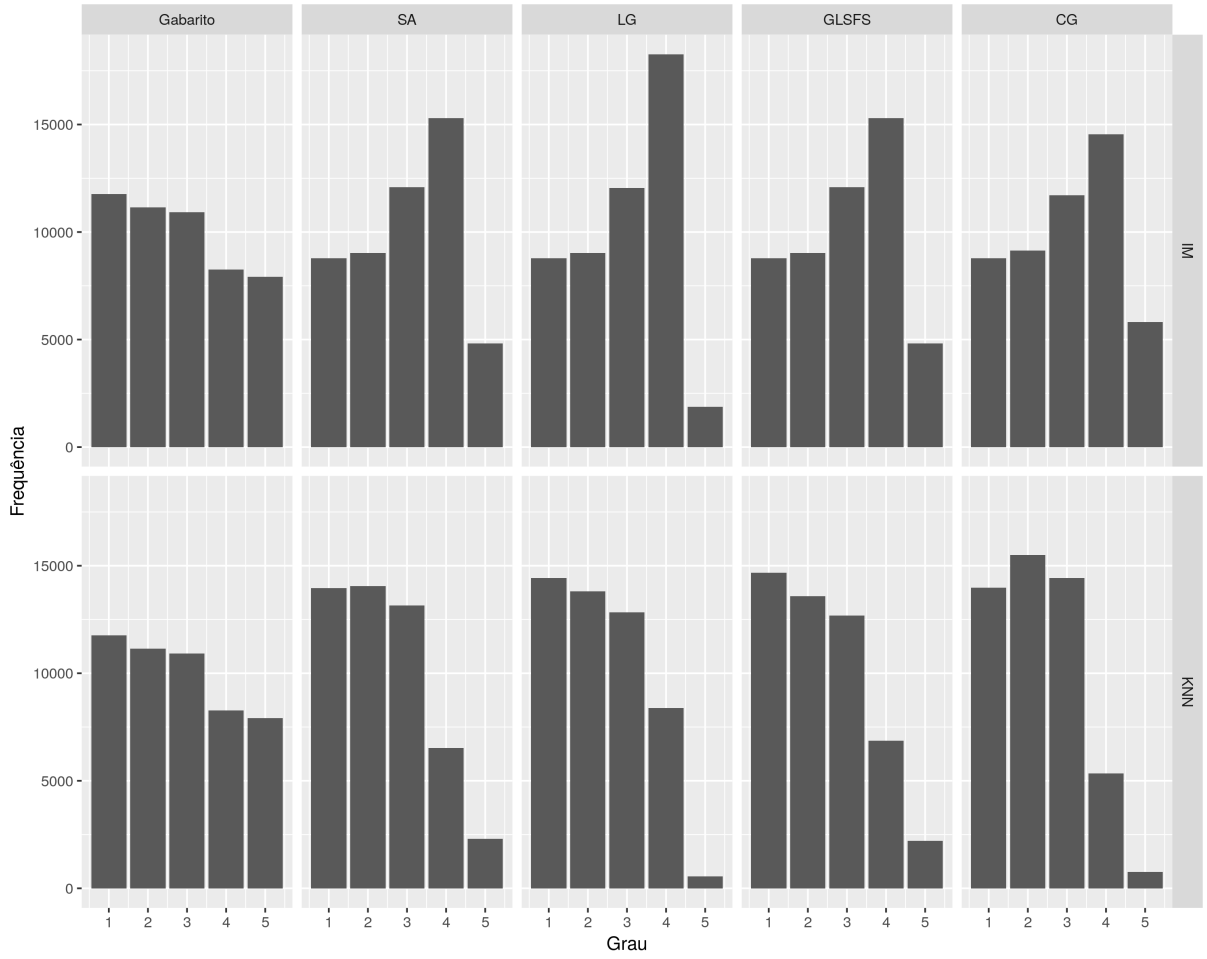


Figura 5.10: Histogramas de graus das redes gabaritos compostas exclusivamente por funções de canalização (à esquerda) e das redes inferidas pelos 4 métodos considerados com base em conjuntos de 50 amostras, para IM (topo) e KNN (embaixo). SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

o desempenho dos métodos com o aprendizado por KNN, e sem esse aprendizado (IM).

5.4.1 Avaliação das topologias das redes inferidas

A Figura 5.15 apresenta a comparação entre os resultados do F-Score para todos os métodos de inferência tanto para os resultados aplicando a informação mútua (IM), como os resultados aplicando o aprendizado do grau por k-nn (KNN). Já a Tabela 5.9 apresenta um sumário da média, desvio padrão, mínimo e máximo das medidas de F-SCORE para todos os métodos. É possível observar que o método LG apresenta o melhor desempenho tanto no cenário com 30 amostras quanto no caso de 50 amostras, embora as diferenças sejam pequenas entre os diferentes métodos. Além disso, o método CG foi o mais beneficiado pelo aprendizado pelo KNN, muito embora apresente o pior desempenho

Tabela 5.6: Graus médios (GM) e erros quadráticos médios (EQM) das redes gabaritos compostas exclusivamente por funções canalizadoras e das redes inferidas pelos 4 métodos considerados com base em conjuntos de 30 amostras e 50 amostras. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

Método	KNN - IM	GM.30	EQM.30	GM.50	EQM.50
Gabarito		2.7888	—	2.7888	—
SA	IM	2.4653	1.4356	2.9671	1.4657
SA	KNN	2.2197	1.6015	2.3825	1.2850
LG	IM	2.4653	1.4393	2.9089	1.4298
LG	KNN	2.2091	1.5868	2.3366	1.3305
GLSFS	IM	2.4653	1.4356	2.9670	1.4657
GLSFS	KNN	2.2322	1.5664	2.3673	1.3205
CG	IM	2.5015	1.4582	2.9901	1.5211
CG	KNN	2.1805	1.6264	2.2681	1.3651

Tabela 5.7: Sumário dos valores de taxa de acerto das dinâmicas geradas pelas redes inferidas, para 30 amostras (acima) e 50 amostras (abaixo), considerando redes gabaritos compostas exclusivamente por funções canalizadoras. Cada dado corresponde a média, o desvio padrão, o valor mínimo e o valor máximo dos valores de taxa de acerto de 1000 redes inferidas, para cada método de inferência.

Método		30 amostras				50 amostras			
		Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	IM	0.8234	0.0231	0.7388	0.9004	0.8515	0.0224	0.7771	0.9189
	KNN	0.8348	0.0240	0.7603	0.9152	0.8830	0.0221	0.8004	0.9641
	Dif.	0.0114	0.0009	0.0215	0.0147	0.0315	-0.0002	0.0233	0.0452
LG	IM	0.8161	0.0234	0.7417	0.8901	0.8469	0.0216	0.7736	0.9148
	KNN	0.8296	0.0244	0.7494	0.9088	0.8764	0.0222	0.8056	0.9631
	Dif.	0.0135	0.0009	0.0077	0.0187	0.0295	0.0006	0.0320	0.0484
GLSFS	IM	0.8247	0.0231	0.7304	0.9034	0.8517	0.0233	0.7733	0.9327
	KNN	0.8356	0.0237	0.7503	0.9235	0.8830	0.0222	0.8078	0.9715
	Dif.	0.0109	0.0006	0.0199	0.0201	0.0313	-0.0011	0.0345	0.0388
CG	IM	0.8459	0.0212	0.7660	0.9163	0.8765	0.0199	0.8150	0.9370
	KNN	0.8534	0.0222	0.7657	0.9271	0.8971	0.0198	0.8317	0.9768
	Dif.	0.0075	0.0010	-0.0003	0.0107	0.0207	-0.0001	0.0168	0.0398

topológico em todos os cenários. Esses resultados apresentam constatações similares aos resultados obtidos para redes gabaritos compostas exclusivamente por funções canalizadoras (Seção 5.3.1), exceto que desta vez o método LG (ao invés do CG) obteve desempenho ligeiramente superior em comparação com os outros métodos.

Para observar a capacidade dos métodos em obter os graus corretos com e sem o aprendizado por KNN, as Figuras 5.16 e 5.17 apresentam uma comparação entre os histogramas das redes gabaritos para 30 e 50 amostras, respectivamente. Já a Tabela 5.10 apresenta as respectivas médias e erros quadráticos médios (EQM) dos graus das redes gabaritos

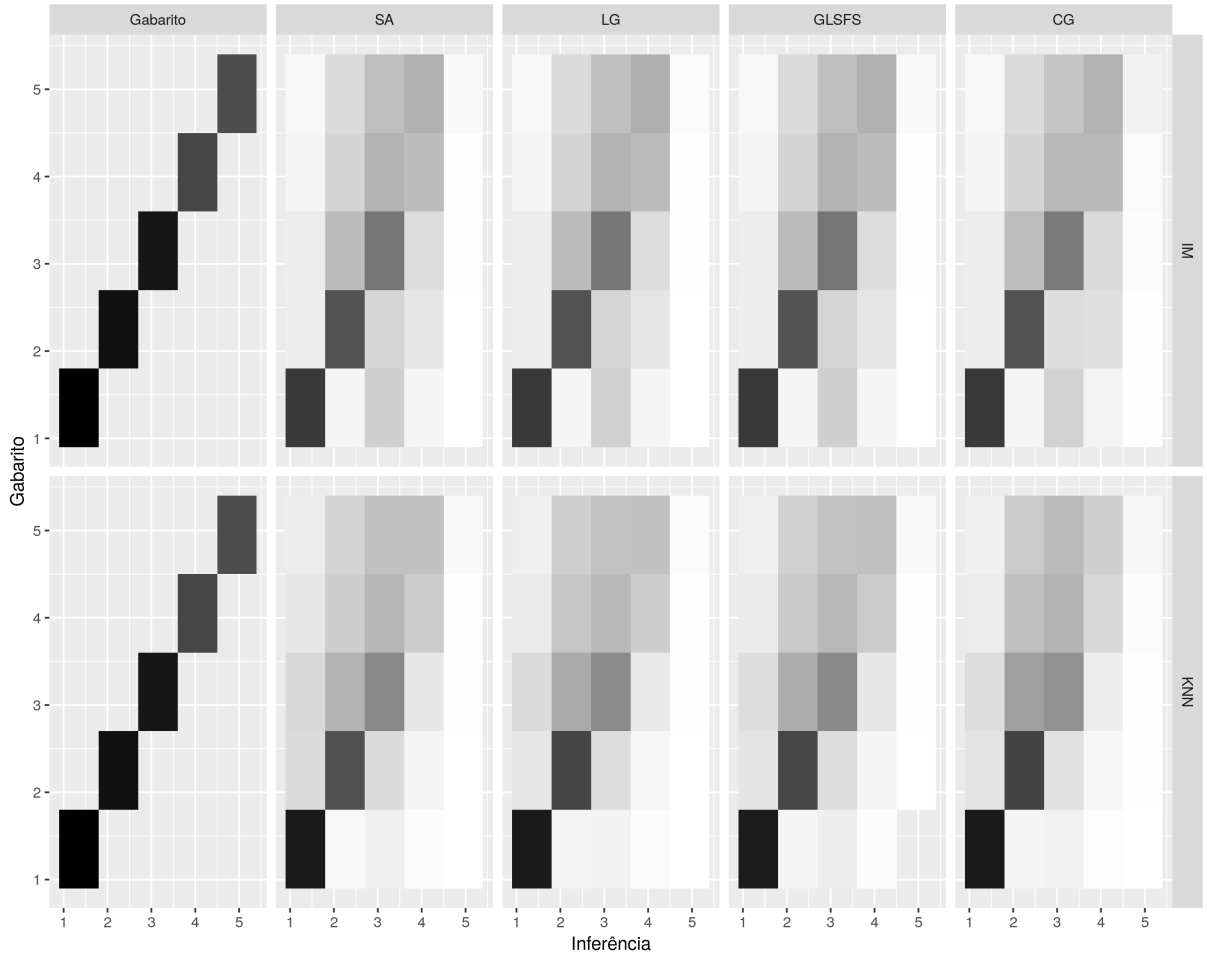


Figura 5.11: *Heatmaps* onde cada célula (i, j) representa a quantidade de vezes que foram inferidos conjuntos de preditores de dimensão i (linhas) para genes alvos que possuem conjuntos de preditores de dimensão j (colunas) nas redes gabaritos compostas exclusivamente por funções canalizadoras. Os *heatmaps* indicados por "Gabarito" representam os *heatmaps* ideais. Quanto mais escuro o tom de cinza, maior é a proporção. Número de amostras $M = 30$. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

e das redes inferidas, com e sem o uso do KNN (KNN e IM, respectivamente). Para o cenário de 30 amostras pode-se observar um comportamento similar ao observado para redes gabaritos compostas exclusivamente por funções canalizadoras. É possível verificar ainda que a subestimação do grau aumenta para todos os métodos, mas como ao mesmo os F-Scores melhoram, isso indica que os genes removidos dos conjuntos de preditores são em sua maioria falsos positivos. O mesmo comportamento ocorre no cenário de 50 amostras, para o qual a subestimação ainda ocorre, porém as distribuições dos graus são mais similares às distribuições dos graus das redes gabaritos. Tal constatação é confirmada pelos valores de EQM, que são menores para o cenário de 50 amostras. O método que apresentou os menores valores de EQM foi o GLSFS.

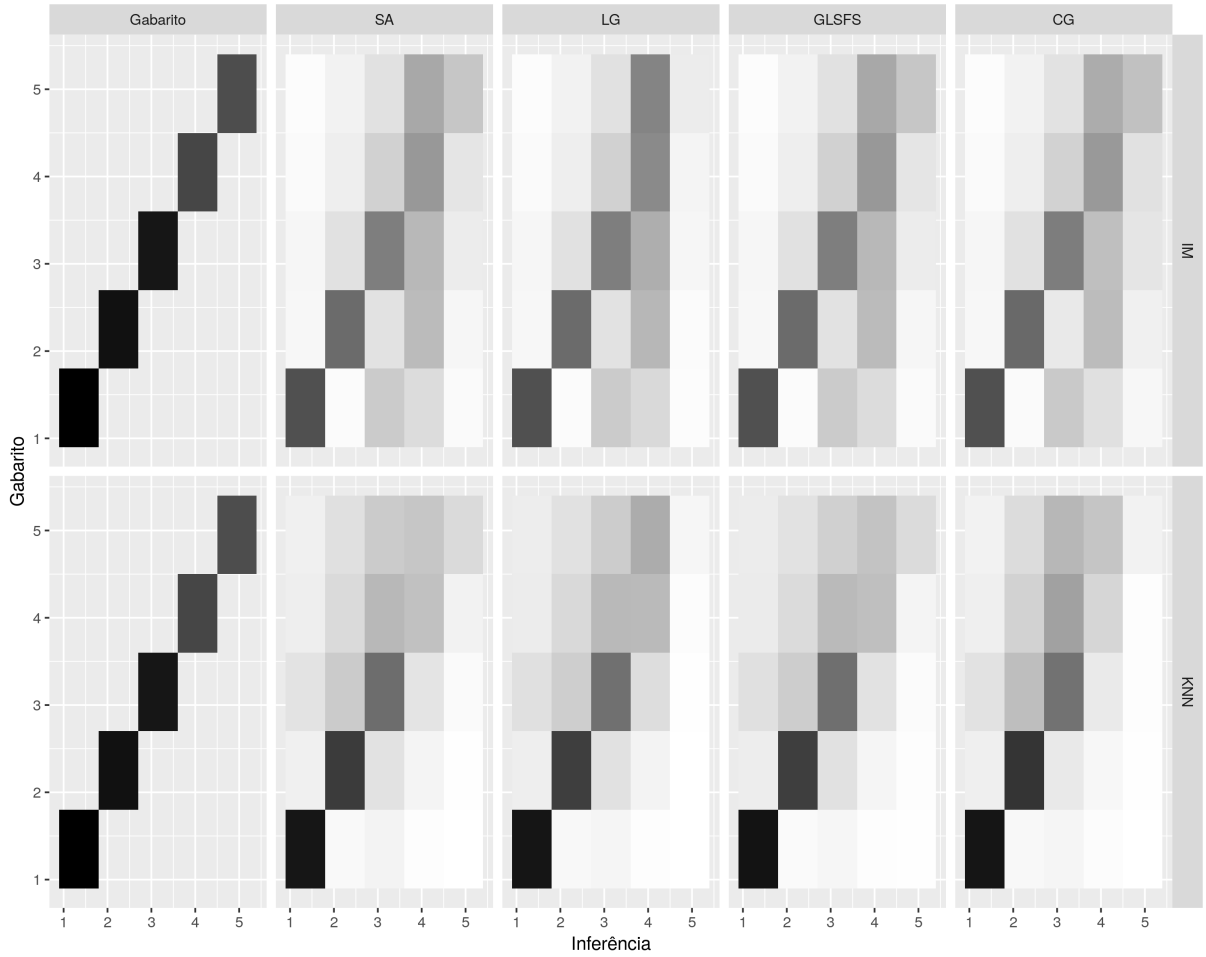


Figura 5.12: *Heatmaps* onde cada célula (i, j) representa a quantidade de vezes que foram inferidos conjuntos de preditores de dimensão i (linhas) para genes alvos que possuem conjuntos de preditores de dimensão j (colunas) nas redes gabaritos compostas exclusivamente por funções canalizadoras. Os *heatmaps* indicados por "Gabarito" representam os *heatmaps* ideais. Quanto mais escuro o tom de cinza, maior é a proporção. Número de amostras $M = 50$. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

As Figuras 5.18 e 5.19 apresentam as matrizes de confusão (*heatmaps*) tanto para 30 quanto para 50 amostras. Pode-se observar claramente como as células embaixo da diagonal se tornam mais claras para todos os métodos através do aprendizado dos graus por KNN. Estes resultados são similares aos já apresentados para redes gabaritos compostas por funções aleatórias e exclusivamente por funções canalizadoras.

5.4.2 Avaliação das dinâmicas geradas pelas redes inferidas

A Figura 5.20 apresenta os *violin plots* dos valores de taxa de acerto das dinâmicas geradas pelas redes inferidas pelos 4 métodos considerados, com e sem o uso do KNN (IM, KNN),

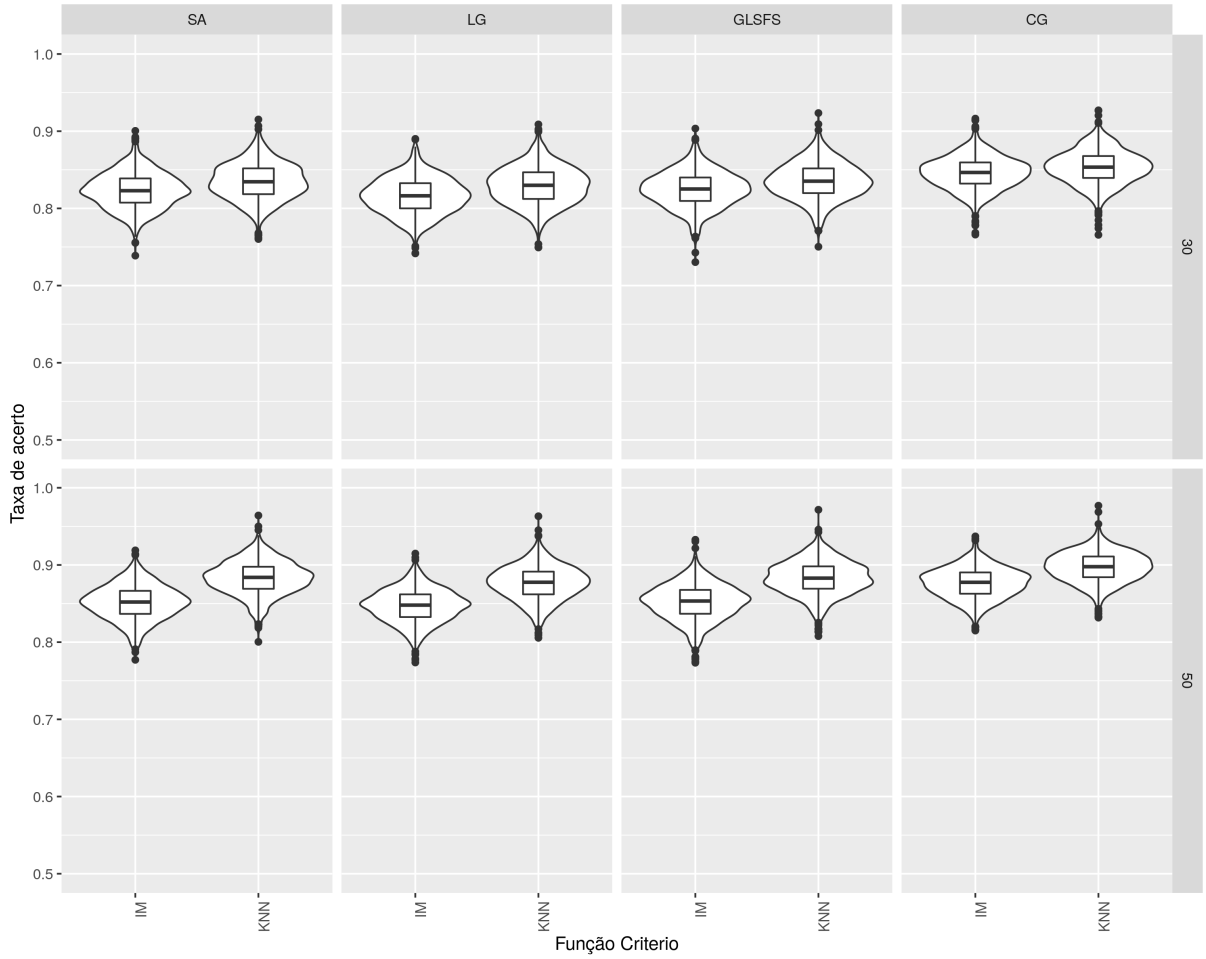


Figura 5.13: *Violin plots* dos valores de taxas de acerto das dinâmicas geradas pelas redes inferidas para 30 amostras (topo) e 50 amostras (embaixo), considerando redes gabaritos compostas exclusivamente por funções canalizadoras. Cada gráfico contém 8 *Violin plots*, agrupados dois a dois (IM - KNN) para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições e CG agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de valores de taxa de acerto para 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado de 1000 estados iniciais sorteados.

enquanto o sumário correspondente contendo a média, o desvio padrão, o mínimo e o máximo dessas taxas de acerto encontra-se na Tabela 5.11. Pode-se observar que todos os métodos apresentaram melhoras com o uso do KNN tanto para 30 amostras quanto para 50 amostras. O método SA foi o maior beneficiado pelo aprendizado dos graus por KNN. Porém, o método CG obteve o melhor desempenho para 30 amostras, enquanto o método LG alcançou o melhor desempenho para 50 amostras.

Esses resultados sugerem que, embora o aprendizado dos graus pelo KNN beneficie todos os métodos, o tipo de funções preditoras continua sendo um fator decisivo no desempenho dos métodos, visto que o método CG apresentou o melhor valor para 30 amostras.

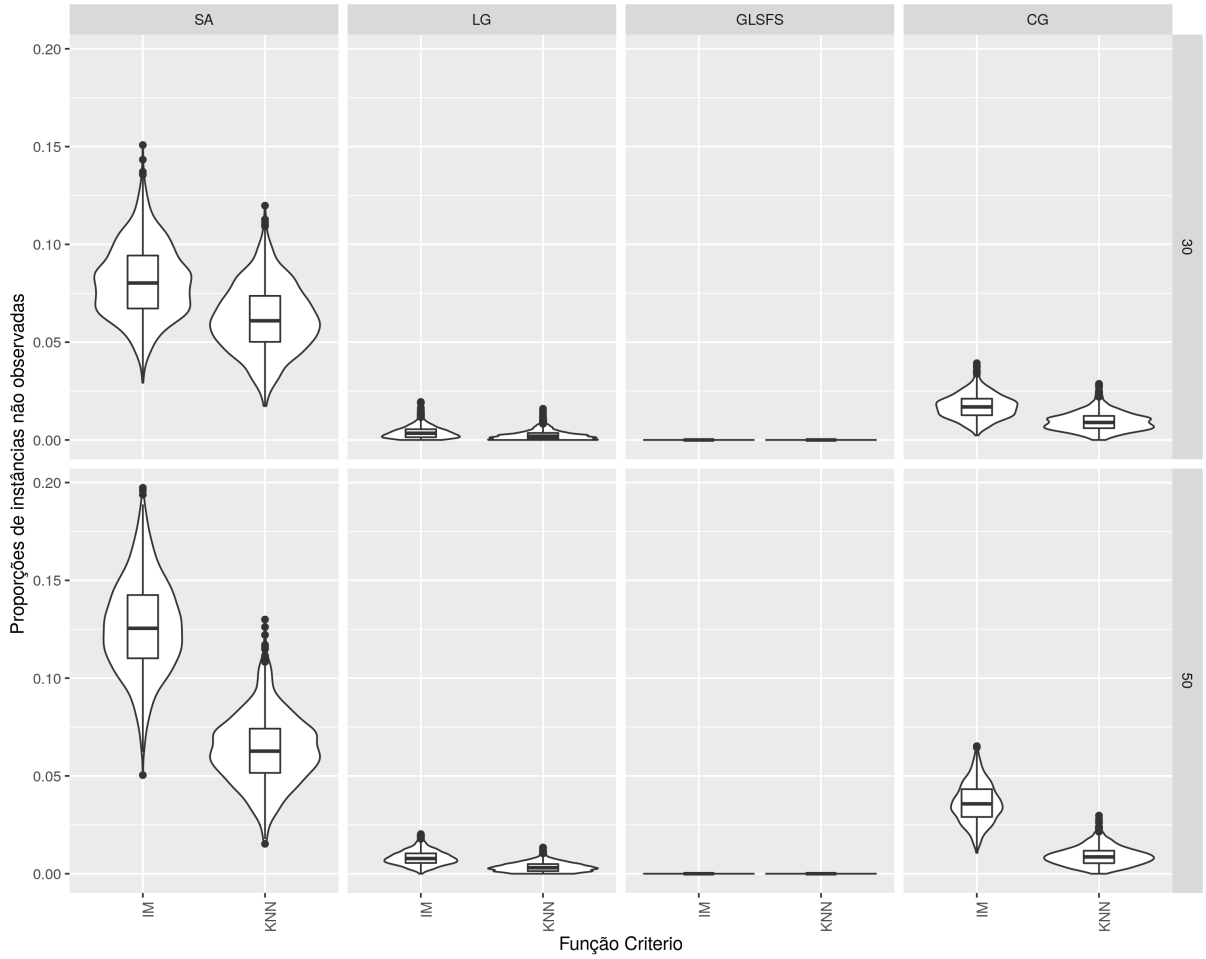


Figura 5.14: *Violin plots* das proporções de instâncias não observadas exigidas na predição dos valores dos genes alvos para redes inferidas para 30 amostras (à esquerda) e 50 amostras (à direita), considerando redes gabaritos compostas exclusivamente por funções canalizadoras. Cada gráfico contém 2 *violin plots* para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização), sendo um sem o uso do KNN (IM) e o outro com o uso do KNN (KNN). Cada *Violin plot* corresponde a distribuição de 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado a partir de 1000 estados iniciais sorteados.

Isso pode ser explicado pelo fato de muitas funções linearmente separáveis serem também canalizadoras, especialmente para graus baixos. Já um maior número de amostras ($M = 50$) faz com que o método LG apresente o melhor desempenho nesse cenário, como esperado.

Com o objetivo de avaliar a capacidade de generalização dos métodos, a Figura 5.21 apresenta os *violin plots* das proporções de instâncias não observadas exigidas na predição dos valores dos genes alvos, enquanto a Tabela 5.12 apresenta um sumário correspondente a esses valores. Podemos observar que os resultados são similares aos apresentados para redes gabaritos compostas exclusivamente por funções canalizadoras.

Tabela 5.8: Sumário dos valores de proporções de instâncias não observadas para as redes inferidas, para 30 e 50 amostras, considerando redes gabaritos compostas exclusivamente por funções canalizadoras. Cada dado corresponde a média o desvio padrão, o valor mínimo e o valor máximo dos valores de proporções de instâncias não observadas de 1000 redes inferidas.

Método		30 amostras				50 amostras			
		Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	IM	0.0811	0.0187	0.0293	0.1508	0.1264	0.0240	0.0504	0.1973
	KNN	0.0619	0.0172	0.0173	0.1198	0.0635	0.0178	0.0153	0.1300
	Dif.	-0.0192	-0.0015	-0.0119	-0.0310	-0.0630	-0.0061	-0.0352	-0.0674
LG	IM	0.0039	0.0031	0.0000	0.0194	0.0081	0.0037	0.0000	0.0203
	KNN	0.0024	0.0024	0.0000	0.0160	0.0034	0.0024	0.0000	0.0134
	Dif.	-0.0015	-0.0006	0.0000	-0.0035	-0.0047	-0.0013	0.0000	-0.0068
GLSFS	IM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	KNN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Dif.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CG	IM	0.0171	0.0062	0.0023	0.0393	0.0362	0.0101	0.0107	0.0652
	KNN	0.0095	0.0049	0.0000	0.0287	0.0090	0.0048	0.0000	0.0297
	Dif.	-0.0076	-0.0013	-0.0023	-0.0105	-0.0272	-0.0053	-0.0107	-0.0355

Tabela 5.9: Sumário dos valores de F-Score para redes inferidas, para 30 e 50 amostras, considerando redes gabaritos compostas exclusivamente por funções linearmente separáveis. Cada dado corresponde a média, desvio padrão, mínimo e máximo dos valores de F-Score de 1000 redes inferidas, para cada método de inferência.

Método		30 amostras				50 amostras			
		Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	IM	0.5515	0.0497	0.3925	0.7104	0.6437	0.0500	0.4300	0.7770
	KNN	0.5621	0.0515	0.3948	0.7200	0.6869	0.0531	0.4545	0.8321
	Dif.	0.0106	0.0017	0.0023	0.0096	0.0432	0.0031	0.0245	0.0551
LG	IM	0.5573	0.0509	0.3913	0.7181	0.6630	0.0514	0.4400	0.8140
	KNN	0.5674	0.0523	0.4078	0.7368	0.7005	0.0552	0.4676	0.8496
	Dif.	0.0101	0.0013	0.0165	0.0187	0.0375	0.0039	0.0276	0.0356
GLSFS	IM	0.5517	0.0495	0.3925	0.7104	0.6436	0.0499	0.4300	0.7842
	KNN	0.5613	0.0510	0.3949	0.7234	0.6872	0.0533	0.4714	0.8425
	Dif.	0.0096	0.0015	0.0024	0.0130	0.0436	0.0034	0.0414	0.0583
CG	IM	0.5416	0.0495	0.3879	0.7050	0.6320	0.0481	0.4558	0.7640
	KNN	0.5552	0.0516	0.4089	0.7200	0.6779	0.0507	0.4982	0.8397
	Dif.	0.0136	0.0020	0.0211	0.0150	0.0459	0.0026	0.0424	0.0757

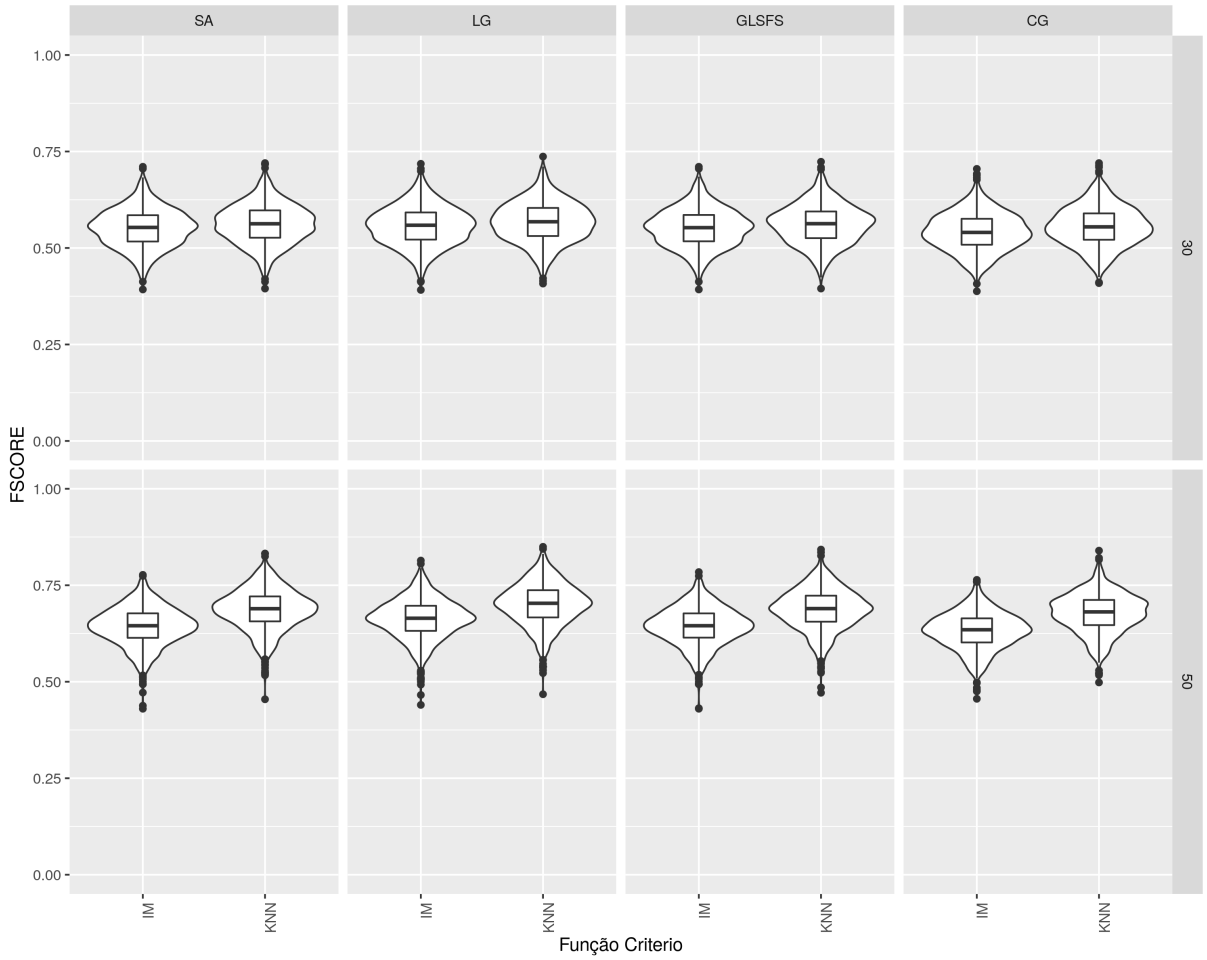


Figura 5.15: *Violin plots* da comparação dos métodos sem aprendizado de grau (critério de parada baseado apenas na evolução da informação mútua: IM) e com aprendizado de grau (KNN) dos valores de F-Score para 30 amostras (topo) e 50 amostras (embaixo), considerando redes gabaritos compostas exclusivamente por funções linearmente separáveis. Cada gráfico contém 8 *Violin plots*, agrupados dois a dois (IM - KNN) para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de valores de F-Score de 1000 redes inferidas.

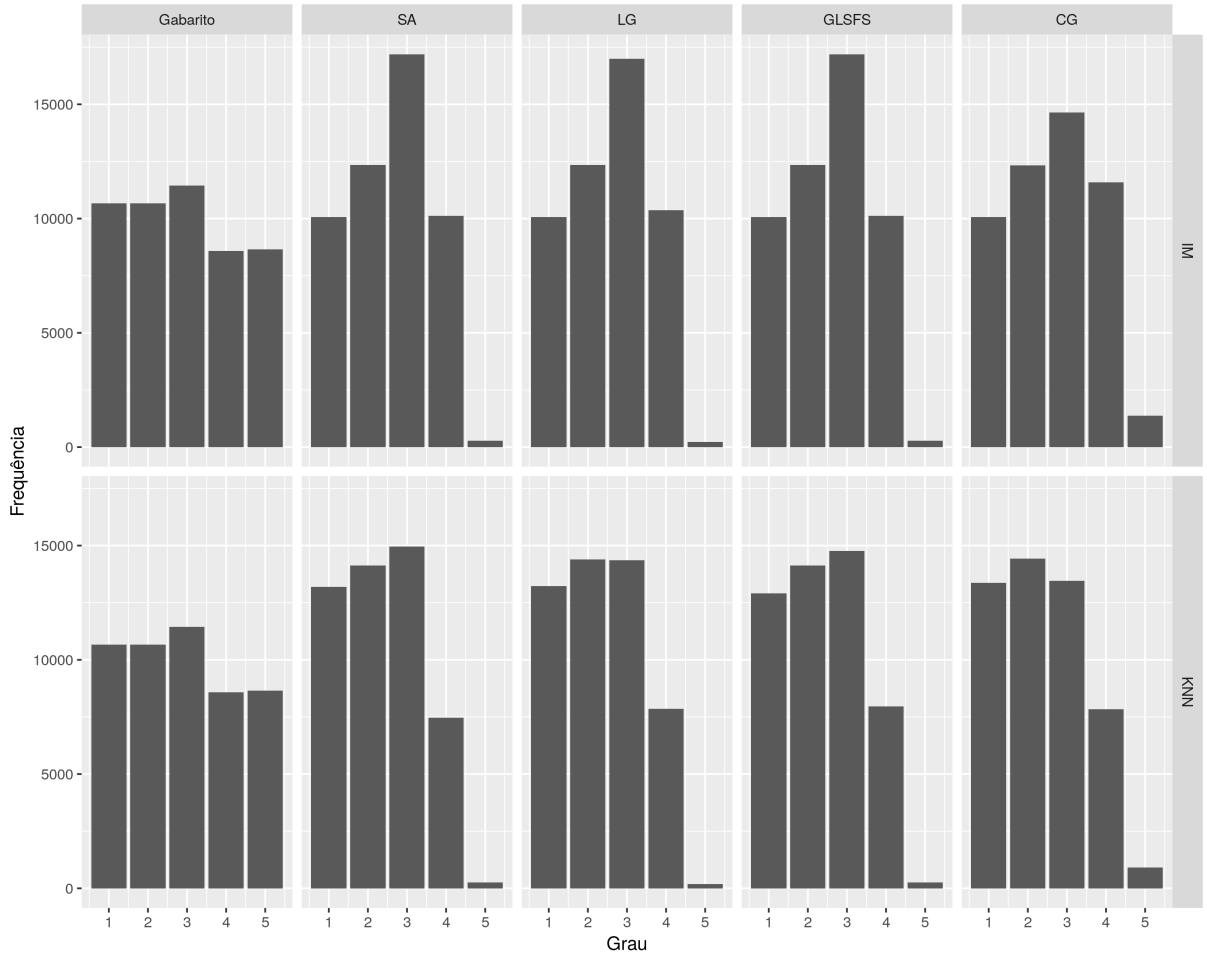


Figura 5.16: Histogramas de graus das redes gabaritos compostas exclusivamente por funções linearmente separáveis (à esquerda) e das redes inferidas pelos 4 métodos considerados com base em conjuntos de 30 amostras, para IM (topo) e KNN (embaixo). SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

Tabela 5.10: Graus médios (GM) e erros quadráticos médios (EQM) das redes gabaritos compostas exclusivamente por funções linearmente separáveis e das redes inferidas pelos 4 métodos considerados com base em conjuntos de 30 amostras e 50 amostras. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

Método	KNN - IM	GM.30	EQM.30	GM.50	EQM.50
Gabarito		2.8781	—	2.8781	—
SA	IM	2.5639	1.2860	3.0658	1.3449
SA	KNN	2.3489	1.4462	2.5964	1.1452
LG	IM	2.5664	1.2853	2.9959	1.3091
LG	KNN	2.3471	1.4448	2.5311	1.1788
GLSFS	IM	2.5639	1.2860	3.0658	1.3449
GLSFS	KNN	2.3707	1.4190	2.6137	1.1069
CG	IM	2.6372	1.2996	3.1016	1.4163
CG	KNN	2.3694	1.4641	2.5623	1.2073

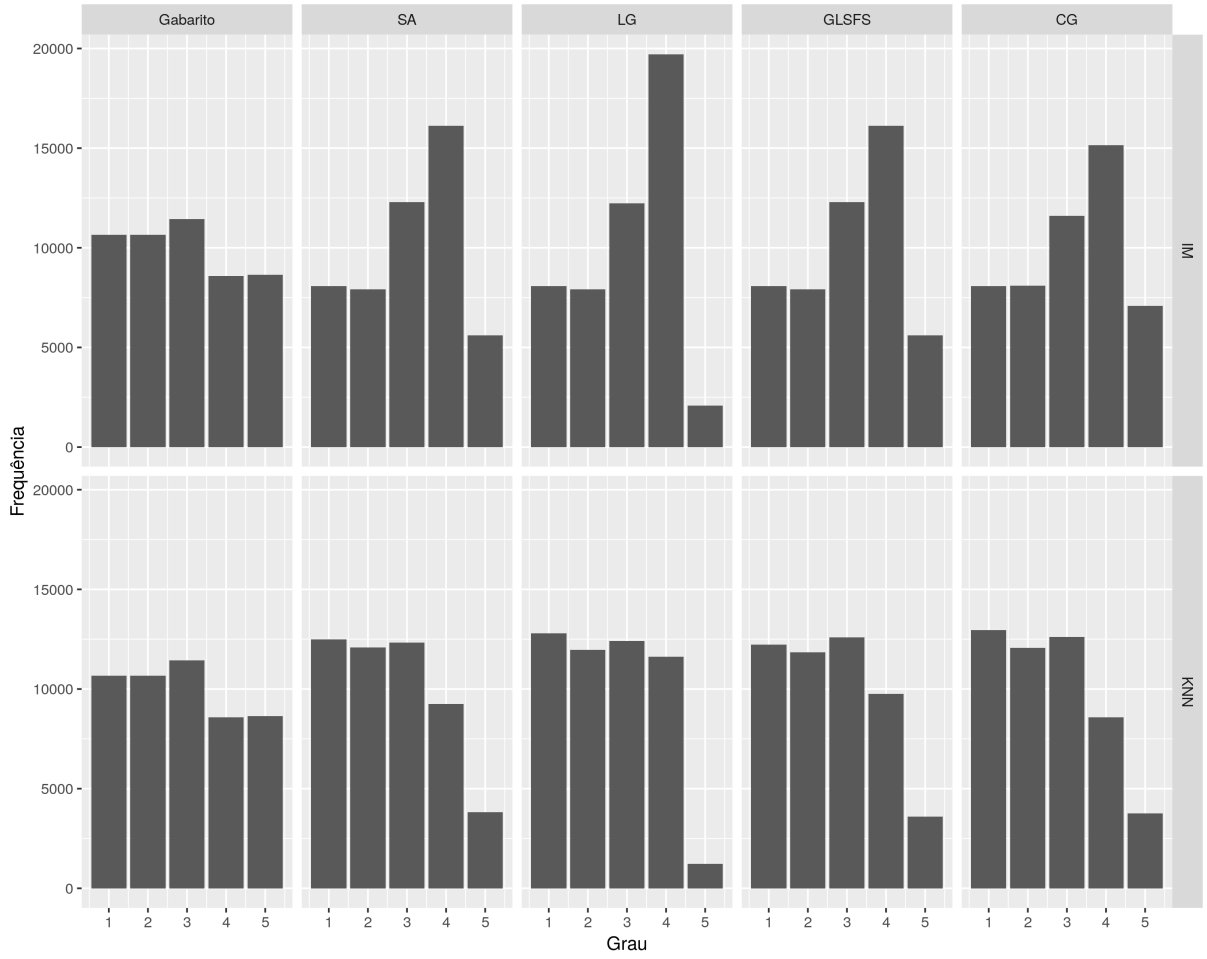


Figura 5.17: Histogramas de graus das redes gabaritos compostas exclusivamente por funções linearmente separáveis (à esquerda) e das redes inferidas pelos 4 métodos considerados com base em conjuntos de 50 amostras, para IM (topo) e KNN (embaixo). SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

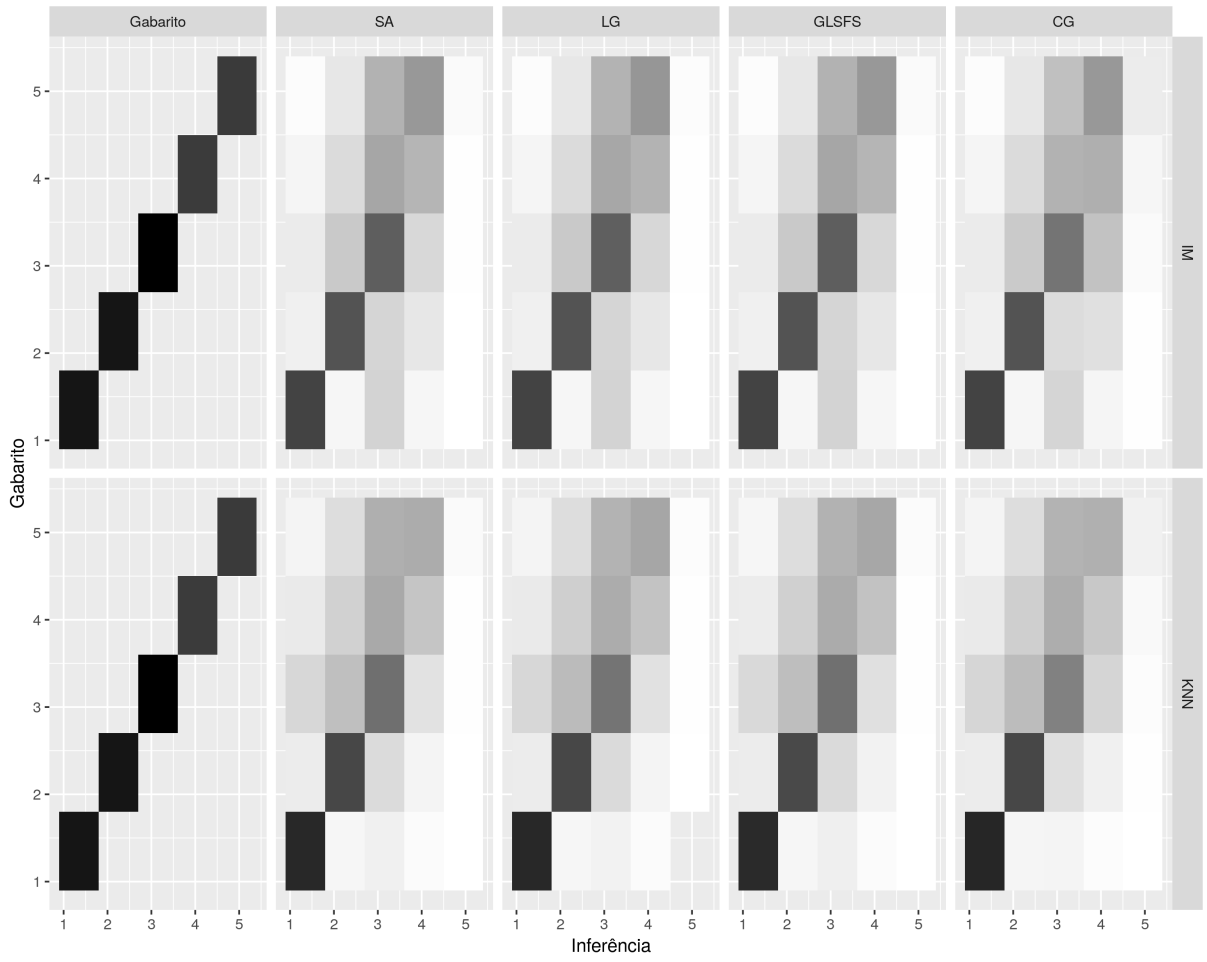


Figura 5.18: *Heatmaps* onde cada célula (i, j) representa a quantidade de vezes que foram inferidos conjuntos de preditores de dimensão i (linhas) para genes alvos que possuem conjuntos de preditores de dimensão j (colunas) nas redes gabaritos compostas exclusivamente por funções linearmente separáveis. Os *heatmaps* indicados por "Gabarito" representam os *heatmaps* ideais. Quanto mais escuro o tom de cinza, maior é a proporção. Número de amostras $M = 30$. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

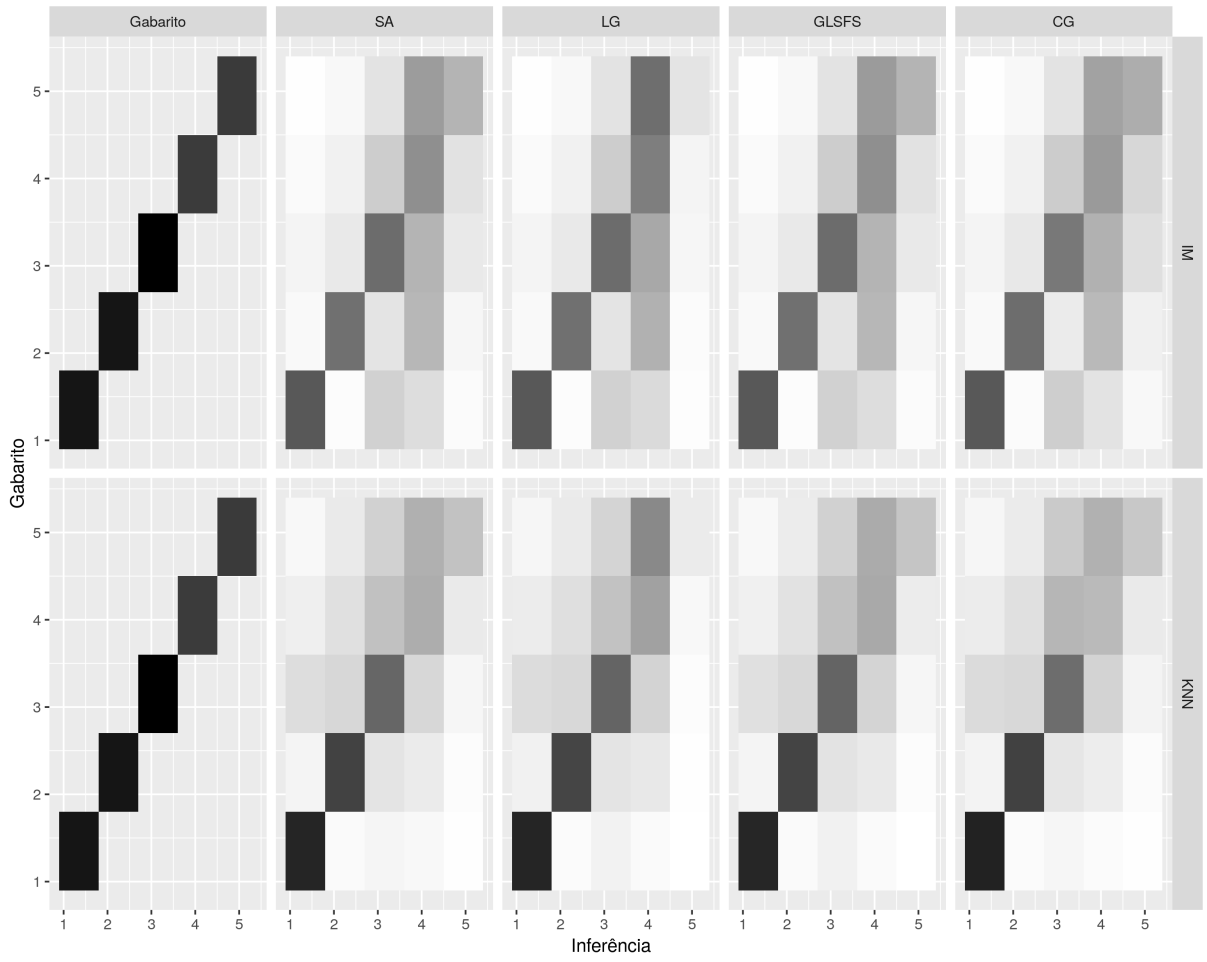


Figura 5.19: *Heatmaps* onde cada célula (i, j) representa a quantidade de vezes que foram inferidos conjuntos de preditores de dimensão i (linhas) para genes alvos que possuem conjuntos de preditores de dimensão j (colunas) nas redes gabaritos compostas exclusivamente por funções linearmente separáveis. Os *heatmaps* indicados por "Gabarito" representam os *heatmaps* ideais. Quanto mais escuro o tom de cinza, maior é a proporção. Número de amostras $M = 50$. SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização.

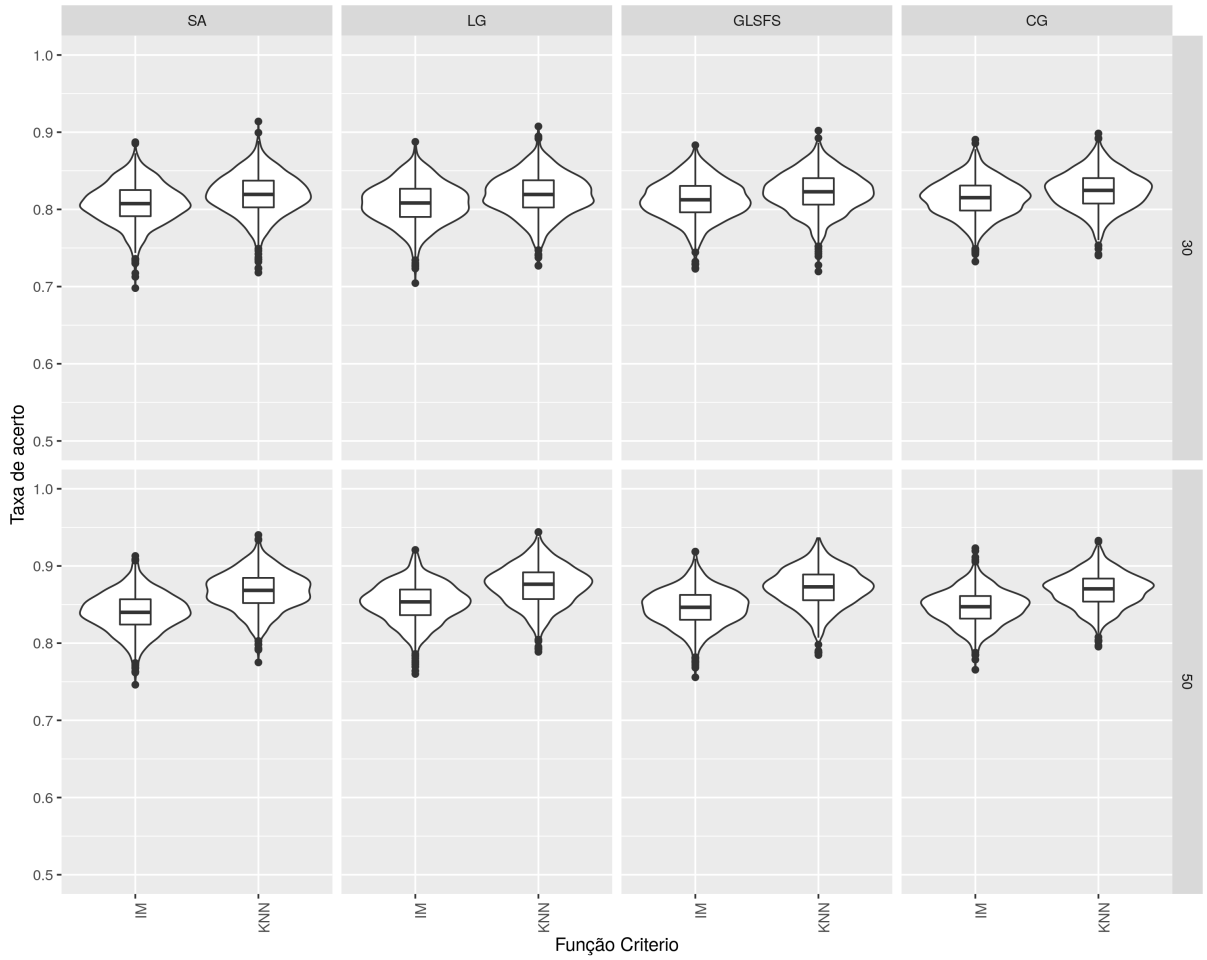


Figura 5.20: *Violin plots* dos valores de taxas de acerto das dinâmicas geradas pelas redes inferidas para 30 amostras (topo) e 50 amostras (embaixo), considerando redes gabaritos compostas exclusivamente por funções linearmente separáveis. Cada gráfico contém 8 *Violin plots*, agrupados dois a dois (IM - KNN) para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições e CG agrupamento por canalização). Cada *Violin plot* corresponde a distribuição de valores de taxa de acerto para 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado de 1000 estados iniciais sorteados.

Tabela 5.11: Sumário dos valores de taxa de acerto das dinâmicas geradas pelas redes inferidas, para 30 amostras (acima) e 50 amostras (abaixo), considerando redes gabaritos compostas exclusivamente por funções linearmente separáveis. Cada dado corresponde a média, o desvio padrão, o mínimo e o máximo dos valores de taxa de acerto de 1000 redes inferidas, para cada método de inferência.

Método		30 amostras				50 amostras			
		Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	IM	0.8076	0.0263	0.6980	0.8871	0.8397	0.0254	0.7461	0.9130
	KNN	0.8192	0.0269	0.7181	0.9139	0.8675	0.0250	0.7750	0.9402
	Dif.	0.0116	0.0006	0.0201	0.0268	0.0278	-0.0005	0.0289	0.0272
LG	IM	0.8080	0.0266	0.7044	0.8875	0.8520	0.0254	0.7601	0.9208
	KNN	0.8189	0.0269	0.7269	0.9075	0.8744	0.0258	0.7887	0.9440
	Dif.	0.0108	0.0003	0.0225	0.0200	0.0224	0.0004	0.0287	0.0232
GLSFS	IM	0.8128	0.0254	0.7230	0.8832	0.8458	0.0240	0.7558	0.9186
	KNN	0.8223	0.0258	0.7196	0.9020	0.8717	0.0248	0.7846	0.9360
	Dif.	0.0095	0.0004	-0.0034	0.0189	0.0259	0.0008	0.0289	0.0174
CG	IM	0.8147	0.0240	0.7324	0.8904	0.8465	0.0223	0.7655	0.9231
	KNN	0.8243	0.0245	0.7402	0.8983	0.8685	0.0222	0.7955	0.9328
	Dif.	0.0096	0.0005	0.0078	0.0080	0.0220	-0.0001	0.0299	0.0097

Tabela 5.12: Sumário dos valores de proporções de instâncias não observadas para as redes inferidas, para 30 e 50 amostras, considerando redes gabaritos compostas exclusivamente por funções linearmente separáveis. Cada dado corresponde a média o desvio padrão, o mínimo e o máximo dos valores de proporções de instâncias não observadas de 1000 redes inferidas.

Método		30 amostras				50 amostras			
		Média	DP	Mín	Máx	Média	DP	Mín	Máx
SA	IM	0.0793	0.0179	0.0234	0.1609	0.1244	0.0243	0.0588	0.2025
	KNN	0.0615	0.0167	0.0153	0.1253	0.0757	0.0196	0.0272	0.1682
	Dif.	-0.0178	-0.0013	-0.0081	-0.0356	-0.0487	-0.0047	-0.0316	-0.0343
LG	IM	0.0048	0.0033	0.0000	0.0230	0.0097	0.0041	0.0003	0.0256
	KNN	0.0036	0.0029	0.0000	0.0228	0.0059	0.0034	0.0000	0.0202
	Dif.	-0.0013	-0.0004	0.0000	-0.0002	-0.0039	-0.0008	-0.0003	-0.0053
GLSFS	IM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	KNN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Dif.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CG	IM	0.0180	0.0066	0.0025	0.0414	0.0376	0.0103	0.0103	0.0710
	KNN	0.0120	0.0053	0.0000	0.0324	0.0173	0.0064	0.0010	0.0390
	Dif.	-0.0059	-0.0013	-0.0025	-0.0089	-0.0203	-0.0039	-0.0093	-0.0320

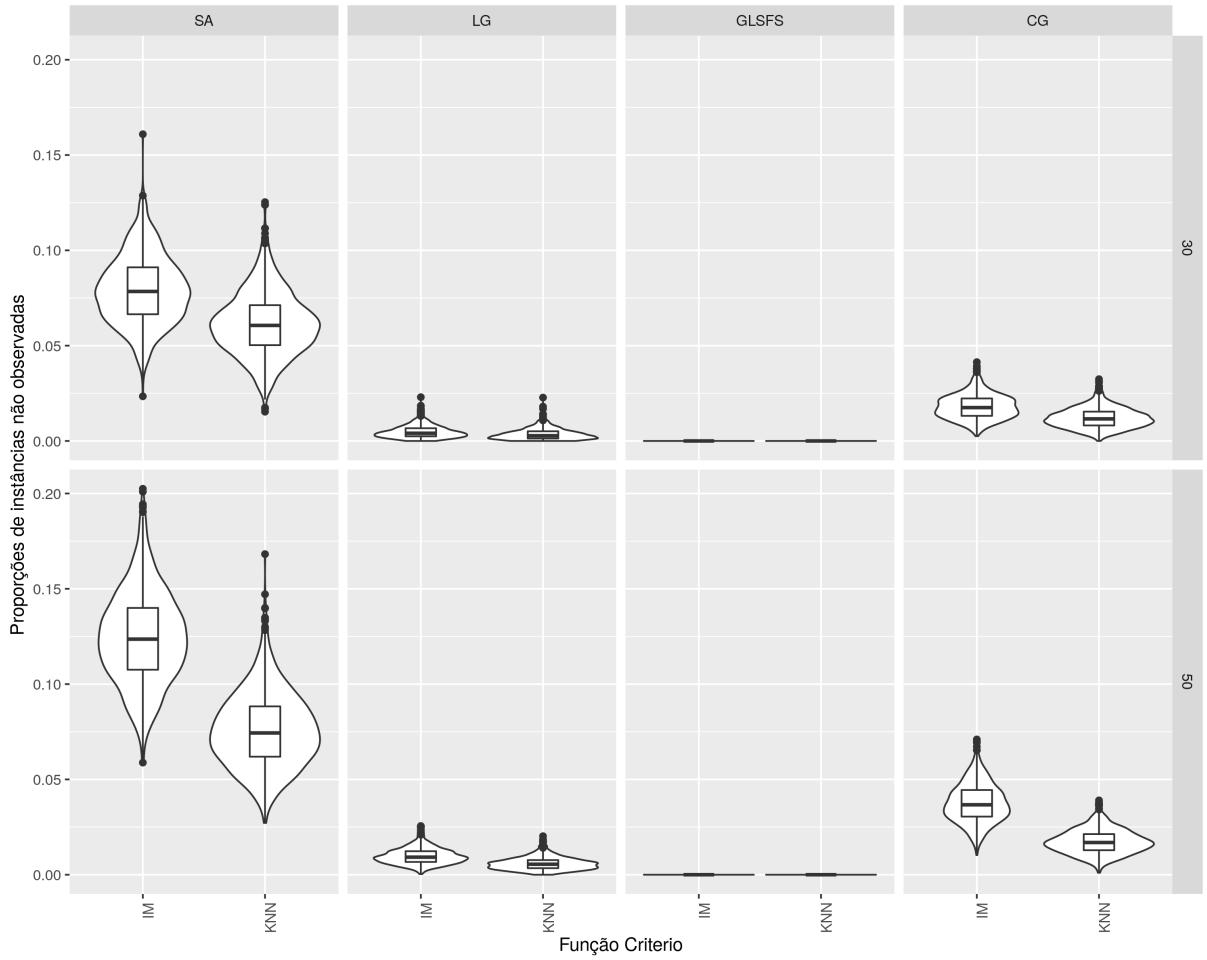


Figura 5.21: *Violin plots* das proporções de instâncias não observadas exigidas na predição dos valores dos genes alvos para redes inferidas para 30 amostras (à esquerda) e 50 amostras (à direita), considerando redes gabaritos compostas exclusivamente por funções linearmente separáveis. Cada gráfico contém 2 *violin plots* para cada método (SA: sem agrupamento, LG: agrupamento linear, GLSFS: agrupamento por busca SFS no reticulado de partições, CG: agrupamento por canalização), sendo um sem o uso do KNN (IM) e o outro com o uso do KNN (KNN). Cada *Violin plot* corresponde a distribuição de 1000 redes inferidas, sendo que cada rede inferida gerou o próximo estado a partir de 1000 estados iniciais sorteados.

Capítulo 6

Resultados experimentais para dados de *microarray*

Nesta seção apresentaremos uma análise dos métodos desenvolvidos aplicados em dados reais de microarray do *Plasmodium falciparum*, um parasita agente causador da malária bastante estudado na literatura [Bozdech et al., 2003]. Adotamos o *Plasmodium falciparum* como estudo de caso pelo fato de já haver estudos de inferência de redes booleanas sobre dados de expressão gênica dele [Barrera et al., 2007, Montoya-Cubas et al., 2015, Jacomini et al., 2017].

Barrera e colegas [Barrera et al., 2007] aplicaram exatamente o mesmo método denominado aqui como "sem agrupamento" (SA) para analisar a rede glicolítica do parasita. Eles adotaram 10 genes conhecidos por codificar enzimas pertencentes à via glicolítica como sementes para testar a capacidade de inferência do modelo PGN proposto por eles. Uma rede foi gerada através da seleção dos 20 melhores preditores individuais para cada alvo considerado, interconectando 9 dos 10 genes alvos no mesmo módulo. Esse resultado foi considerado satisfatório, já que confirmou a hipótese de que os genes sementes da glicólise devem formar um módulo no qual eles fiquem próximos um do outro. Entretanto a mesma modularidade não foi observada para a busca exaustiva pelos melhores pares de preditores por semente. Além disso, eles realizaram o mesmo experimento para sementes do apicoplasto (plastídeo) como controle negativo, tendo como hipótese que as redes em torno da glicólise e do apicoplasto apresentariam alta modularidade interna, porém poucas conexões entre essas redes.

Assim, nesse capítulo replicamos o mesmo experimento conduzido por Barrera e colegas para os quatro métodos considerados e a tranferência de aprendizado supervisionado pelo uso do KNN, de modo a avaliar as topologias das redes inferidas, bem como as dinâmicas geradas por essas redes.

6.1 Dados de expressão de *Plasmodium falciparum*

Adotamos dados de expressão gênica do transcriptoma do ciclo de desenvolvimento intra-eritrocítico (IDC) do *Plasmodium falciparum* (um agente da malária) para avaliar os métodos de agrupamento propostos. Esse transcriptoma foi gerado por medições relativas dos níveis de abundância de mRNA de amostras coletadas de uma cepa chamada HB3, que é bem caracterizada e originada de Honduras [Bozdech et al., 2003]. O conjunto de dados de controle de qualidade (chamado *QC Dataset*), contendo 48 amostras com 5080 genes, foi utilizado nos experimentos. Essas 48 amostras foram extraídas de hora em hora, correspondendo às 48 horas (instantes de tempo) do IDC, As amostras correspondentes às 23^a e 29^a horas foram descartadas devido à má qualidade (portanto, consideramos a amostra 22 como predecessora da amostra 24 e a amostra 28 como antecessora da amostra 30), o que levou a $M = 46$ amostras temporais.

Como os valores de expressão são contínuos, aplicamos um processo de quantização em dois níveis (para os métodos de inferência propostos, onde 0 representa subexpressão do gene e 1 representa superexpressão) e em três níveis (-1 é subexpressão, 0 é expressão normal e 1 é superexpressão). A quantização em três níveis foi aplicada para ajudar no desempate dos conjuntos de preditores empatados em primeiro lugar para um dado gene alvo. Mais precisamente, os procedimentos de quantização em dois e em três níveis estão descritos a seguir [Barrera et al., 2007]:

1. Aplicação do logaritmo base 2 a todas as expressões originais, resultando na matriz $\mathbf{G}$;

2. Os sinais de $\mathbf{G}$ foram normalizados por uma transformação normal dada por, para cada gene e a cada instante t, $g(t) \in \mathbf{G}$, $\eta[g(t)] = \frac{g(t)-\mu[g(t)]}{\sigma[g(t)]}$, onde $\mu[g(t)]$ e $\sigma[g(t)]$ são média e desvio padrão de $g(t)$, respectivamente;

3. Seja $g'(t) = \eta[g(t)]$. A quantização em dois níveis de um gene g' é realizada por um mapeamento definido do seguinte modo:

$$g''(t) = \begin{cases} 0, & \text{if } g'(t) \leq 0 \\ 1 & \text{if } g'(t) > 0 \end{cases} \tag{6.1}$$

4. Seja $g'(t) \in \eta[g(t)]$. A quantização de um gene g' em três níveis é realizada por um mapeamento definido, para cada t, por:

$$g'''(t) = \begin{cases} -1, & \text{if } g'(t) < l \\ 0, & \text{if } l \leq g'(t) \leq h \\ 1 & \text{if } g'(t) > h \end{cases} \tag{6.2}$$

em que l é a média dos valores negativos de $g'(t)$ e h é a média dos valores positivos de $g'(t)$.

6.2 Avaliação das topologias das vizinhanças ao redor dos genes sementes

Para avaliar o desempenho do método em dados reais realizamos a inferência de redes dos 5080 genes da malaria, através das 46 amostras disponíveis, aplicando a busca exaustiva até a dimensão 2, seguida da busca sequencial para frente (SFS) até a dimensão 6. Caso haja empates nos conjuntos de preditores, isto é, múltiplos conjuntos de preditores empatados com o melhor valor de função critério, o desempate se deu através da aplicação da mesma função critério para dados ternários. Para validar o método as redes foram construídas adotando como sementes (genes alvos) os genes pertencentes aos módulos funcionais da via glicolítica e do plastídeo (apicoplasto) [Barrera et al., 2007].

Para observar a topologia da rede formada por uma vizinhança ao redor dos genes sementes, analisamos um recorte da rede completa considerando apenas os preditores dos genes sementes (”pais”), os preditores dos preditores dos genes sementes (”avôs”), e os genes que possuem os genes sementes como preditores (”filhos”).

As Figuras 6.1,6.2,6.3 e 6.4 apresentam a vizinhança ao redor dos genes sementes obtida para cada método, sendo que os nós amarelos representam os genes sementes da glicólise, enquanto os nós verdes representam os genes sementes do apicoplasto. Em todos os casos percebe-se modularidade entre as sementes de cada componente embora nos casos do SA e GLSFS existem um número relativamente alto de conexões entre as componentes da glicólise e do apicoplasto, não havendo uma clara separação entre eles (Figuras 6.1 e 6.3). Já para os métodos LG e CG, embora também existam conexões entre as componentes, a separação entre as componentes é mais perceptível (Figura 6.2 e 6.4). Isso indica que os métodos LG e CG obtiveram um melhor desempenho topológico, possivelmente devido a tendência de redes biológicas possuírem uma proporção maior de funções linearmente separáveis e de canalização do que o esperado para redes geradas de forma aleatória [Waddington, 1942, Kauffman, 1969, Layne et al., 2012, Li et al., 2013, Li et al., 2004, Davidich and Bornholdt, 2008].

Como discutido nos experimentos com dados simulados (Capítulo 5), a inferência de redes gênicas através de um pequeno número de amostras e com presença de ruido, como é o caso dos dados de *microarray* em questão, tende a implicar em uma estimação problemática da dimensão (grau) do conjunto de genes preditores de um gene alvo determinado. Dessa forma, muitas arestas das Figuras 6.1,6.2,6.3 e 6.4 podem ser resultantes desse problema, o que faz necessária uma estrategia de transferência de aprendizado supervisionado

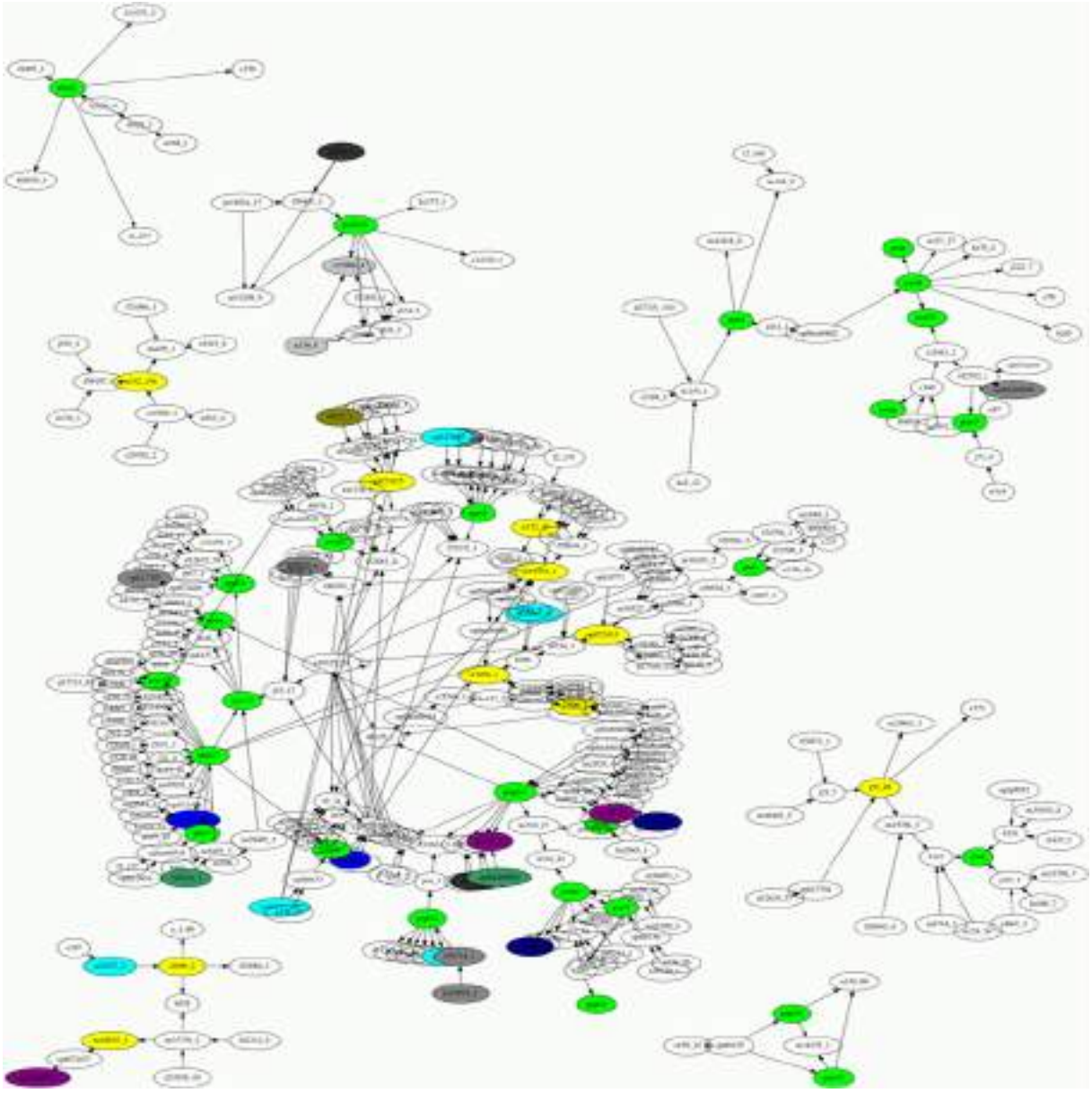


Figura 6.1: Vizinhanças ao redor dos genes sementes da glicólise (nós amarelos) e do apicoplasto (nós verdes) obtidas pelo método Sem Agrupamento (SA)

dos graus a partir de redes simuladas.

6.2.1 Transferência de aprendizado dos graus via KNN

Nesta seção, serão apresentados experimentos de transferência de aprendizado dos graus a partir de redes simuladas para o contexto dos dados reais de *microarray* de *Plasmodium falciparum*. Tal estratégia foi apresentada na Seção 3.5, sendo que os resultados aplicados sobre redes simuladas foram apresentados no Capítulo 5 e se mostraram promissores. Do mesmo modo que foi feito para os experimentos com redes simuladas, aqui também foi

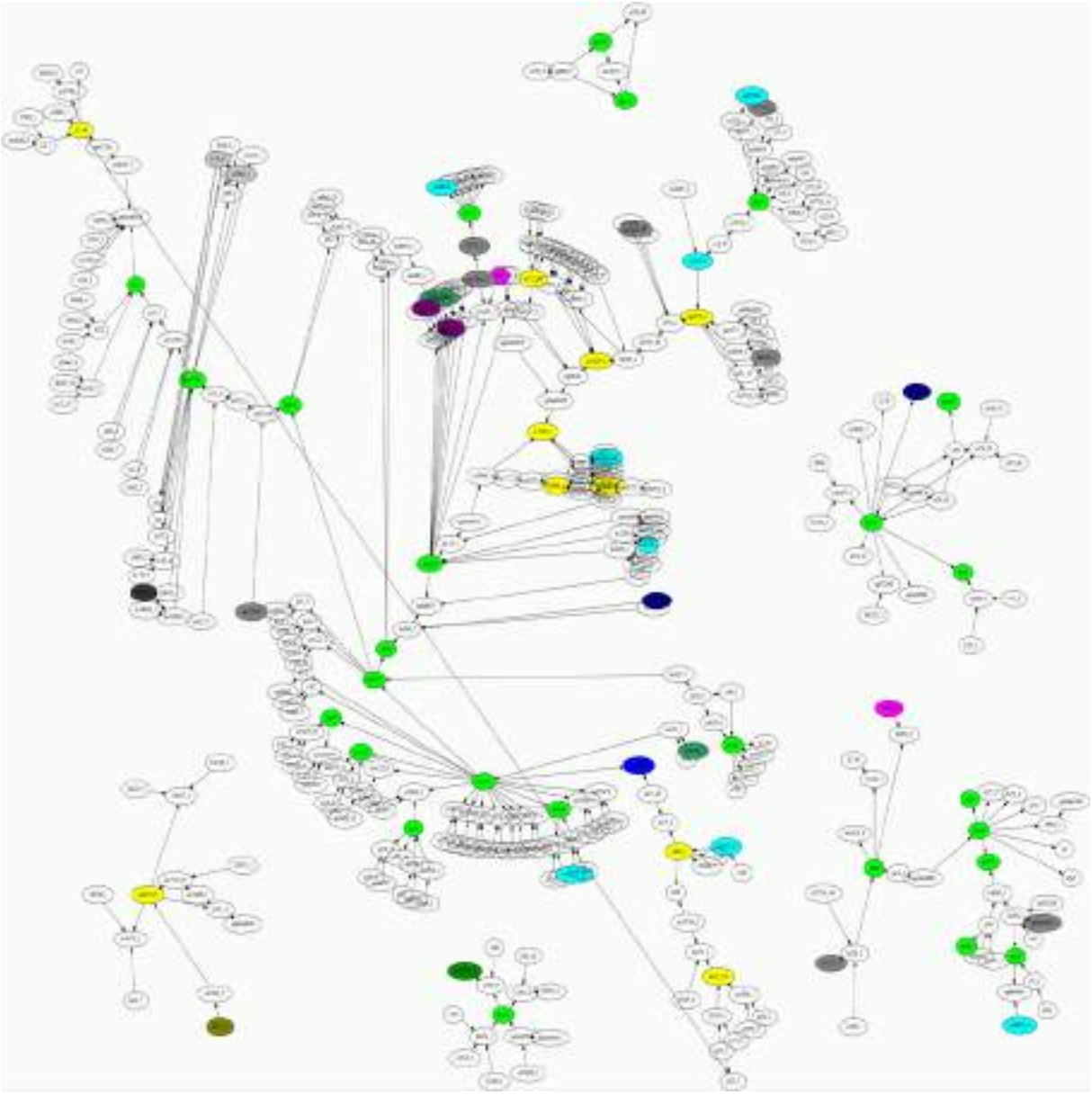


Figura 6.2: Vizinhanças ao redor dos genes sementes da glicólise (nós amarelos) e do apicoplasto (nós verdes) obtidas pelo método Agrupamento Linear (LG)

adotado o algoritmo de classificação supervisionada por vizinhos mais próximos (KNN). Como no caso do *Plasmodium falciparum* a rede gabarito (padrão ouro) não é conhecida, o aprendizado supervisionado por KNN foi treinado sobre conjuntos de 50 amostras geradas por redes simuladas compostas por funções aleatórias do mesmo modo como apresentado no Capítulo 5. Em seguida, esse aprendizado foi transferido para o domínio dos dados reais de *microarray* em questão. De fato, trata-se de domínios diferentes embora relacionados, porque não são conhecidas a priori a proporção dos tipos de funções existentes, a topologia exata, e outras características próprias da rede da malária.

As Figuras 6.5,6.6,6.7 e 6.8 apresentam as redes em torno da vizinhança dos genes

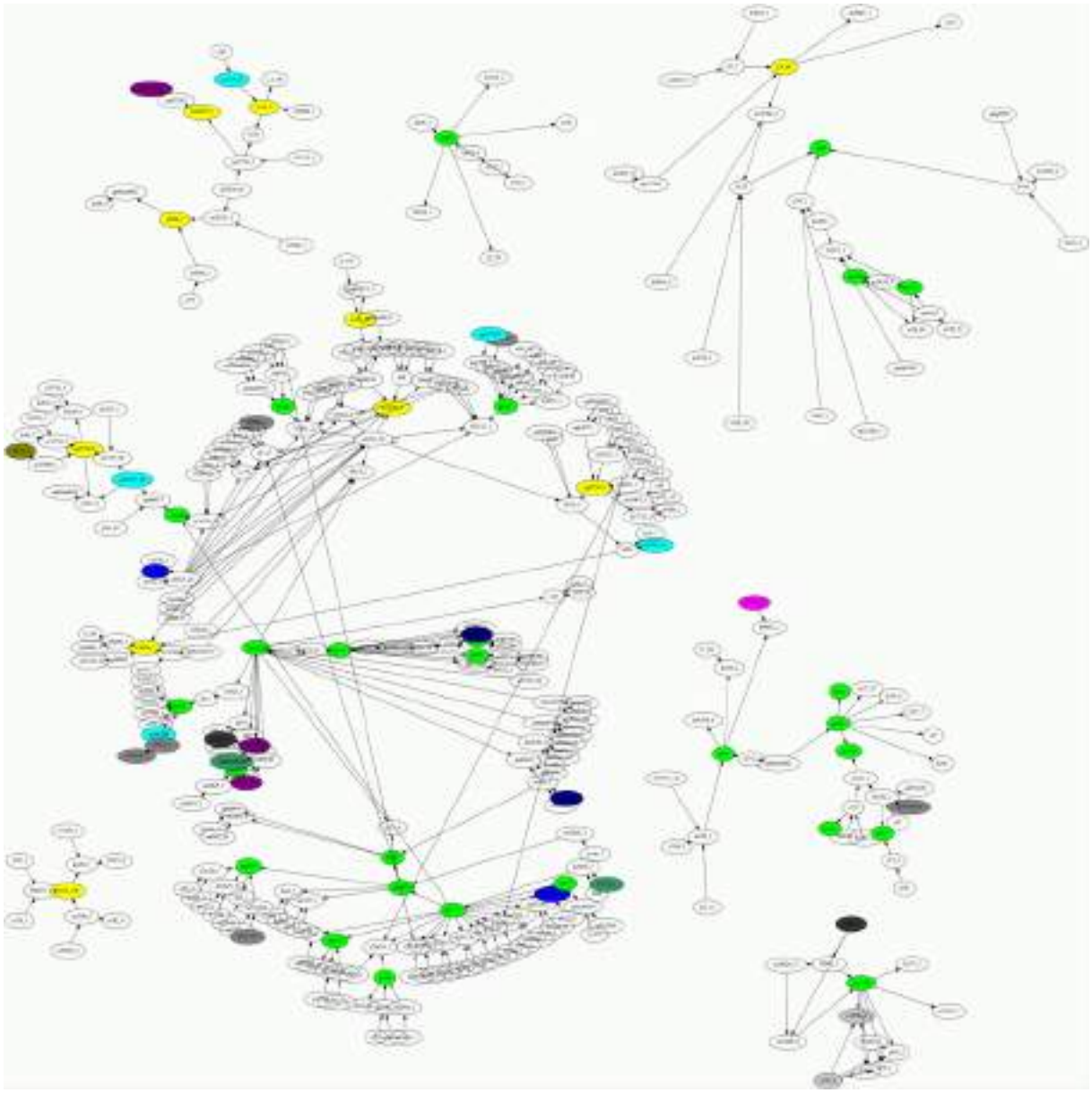


Figura 6.3: Vizinhanças ao redor dos genes sementes da glicólise (nós amarelos) e do apicoplasto (nós verdes) obtidas pelo método Agrupamento por GLSFS

sementes da glicólise e do apicoplasto, para os quatro métodos de inferência. É possível notar que a conectividade intramodular (entre genes da mesma componente) foi mantida, porém foram reduzidas as conexões entre os diferentes componentes foram apagadas. Essa observação é mais marcante para os métodos SA, LG e GLSFS, para os quais são eliminadas por parte das arestas que conectam diferentes componente. Embora no caso do método CG também foi observada essa mesma tendência, a diferença não foi tão nítida porque o resultado sem a transferência de aprendizado por KNN já apresentavam um número relativamente pequeno de conexões entre genes de diferentes componentes. Tal resultado é satisfatório em todos os cenários porque filtra as arestas potencialmente

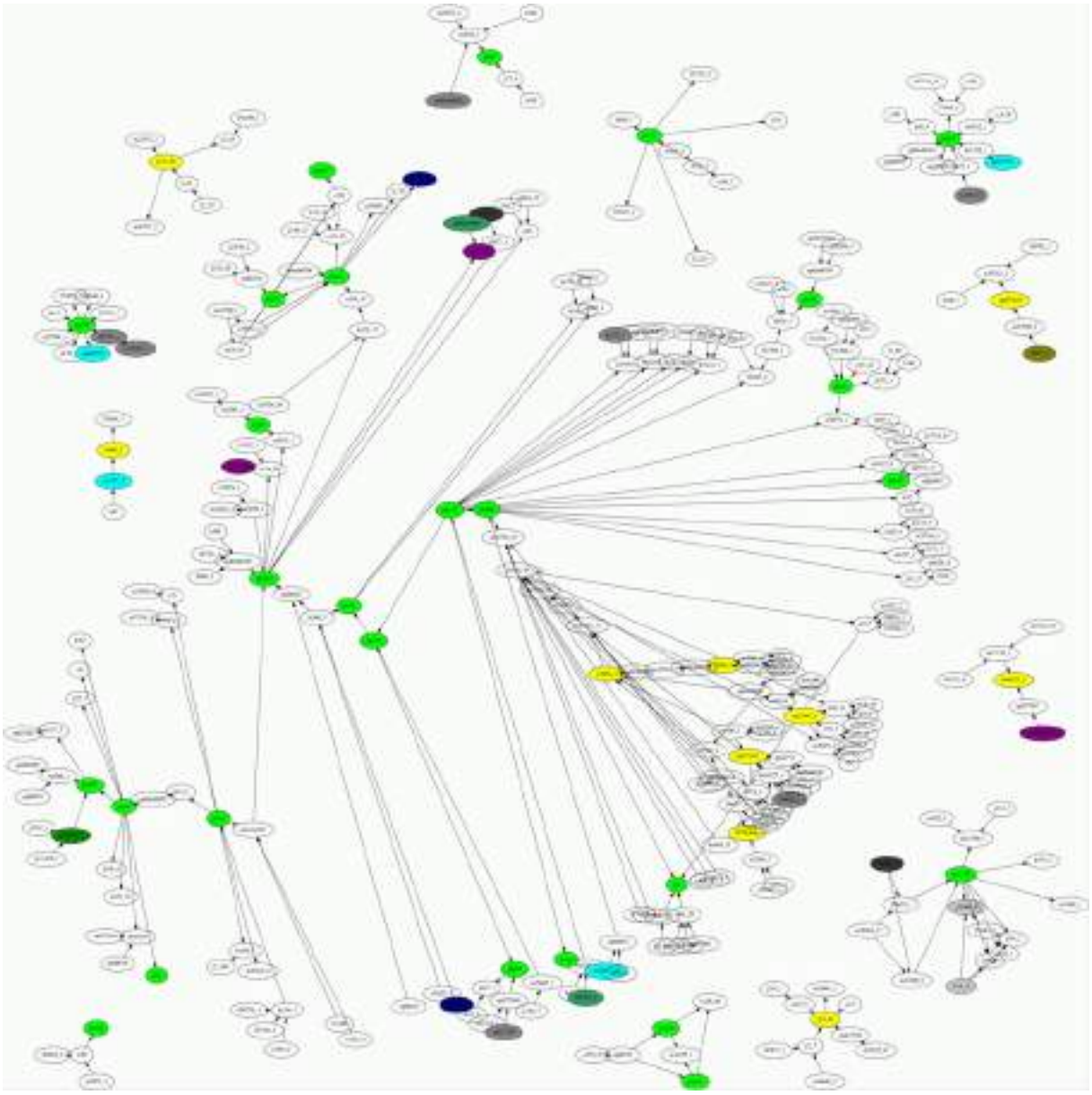


Figura 6.4: Vizinhanças ao redor dos genes sementes da glicólise (nós amarelos) e do apicoplasto (nós verdes) obtidas pelo método Agrupamento por Canalização (CG)

problemáticas (que conectam componentes), mantendo a intramodularidade (conexões internas dos componentes).

6.3 Avaliação da dinâmica gerada pelas redes inferidas

Como feito para dados simulados (Capítulos 4 e 5), aqui o objetivo é avaliar a precisão da predição da dinâmica gerada pelas redes inferidas quando comparada à dinâmica re-

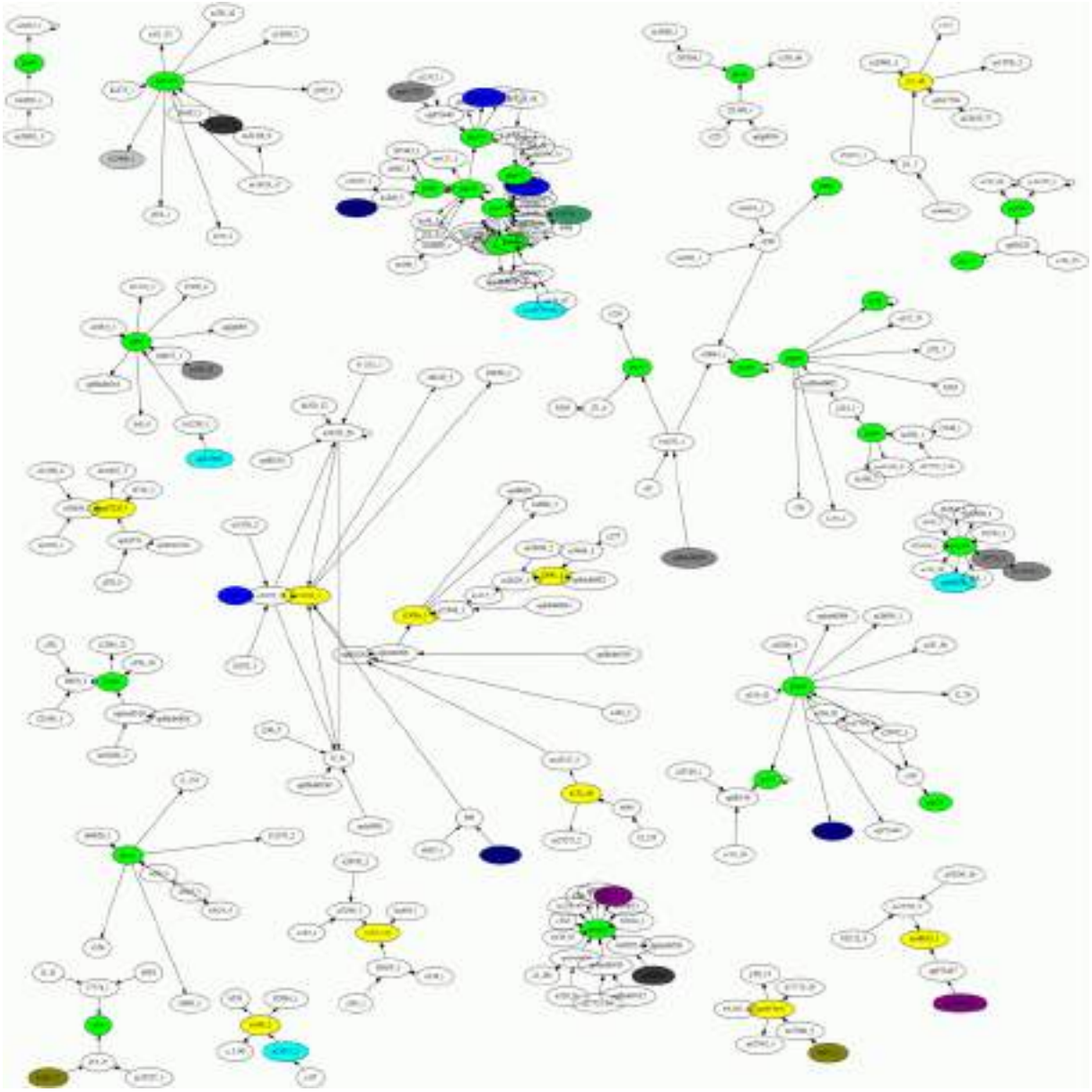


Figura 6.5: Vizinhanças ao redor dos genes sementes da glicólise (nós amarelos) e do apicoplasto (nós verdes) obtidas pelo método Sem agrupamento(SA), com a transferência de aprendizado do grau via KNN em redes simuladas.

sultante das próprias amostras de *microarray* (adotadas como padrão ouro). Para isso foi realizada uma analise de validação cruzada dos dados tomando um particionamento de 38 amostras de treinamento e 8 amostras de teste, dentre as 46 amostras disponíveis. O conjunto de amostras é dividido em 5 particionamentos com 8 amostras de teste e um particionamento com apenas 6 amostras de teste para completar as 46 amostras, conforme ilustrado na Figura 6.9. Aqui vale ressaltar que, pelo fato dos dados serem provenientes do ciclo intraeritrocítico da malária, suas amostras são circulares, ou seja, a última amostra pode ser usada para predizer a primeira amostra (tempo 0).

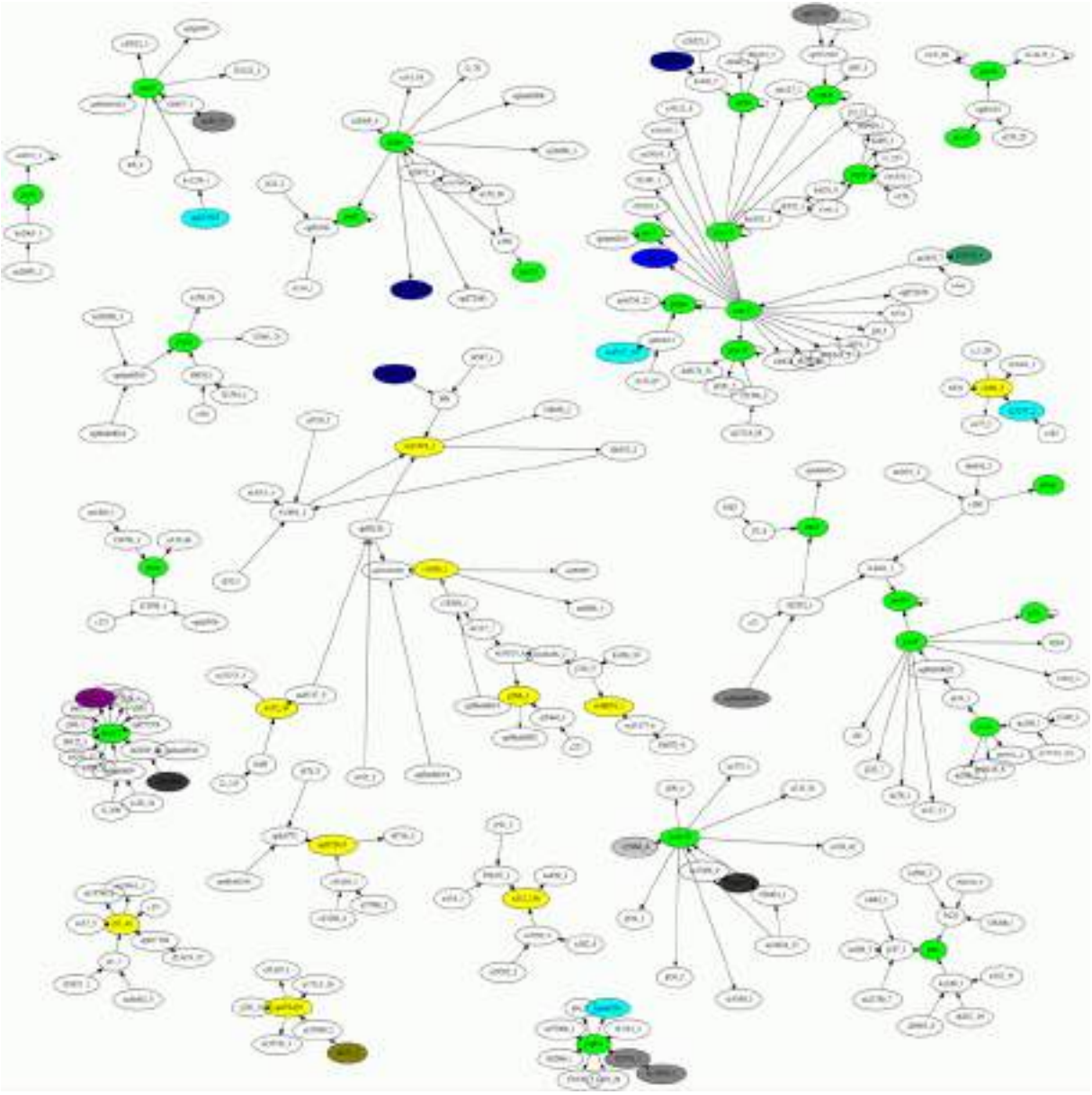


Figura 6.6: Vizinhanças ao redor dos genes sementes da glicólise (nós amarelos) e do apicoplasto (nós verdes) obtidas pelo método Agrupamento linear (LG), com a transferência de aprendizado do grau via KNN em redes simuladas.

Para cada conjunto de teste, foi determinada a precisão com que a rede inferida consegue gerar os dados de teste, isto é, se os dados de treinamento vão de $t_{i \mod M}$ até $t_{j \mod M}$, os dados de teste serão compostos pelas amostras desde $t_{(j+1) \mod M}$ até $t_{(i-1) \mod M}$. Dessa forma, infere-se uma rede com as amostras desde $t_{i \mod M}$ até $t_{j \mod M}$ e, a partir dessa rede inferida, os dados são gerados desde o tempo $t_{(j+1) \mod M}$ tomando como base o tempo anterior $t_{j \mod M}$, até o tempo $t_{(i-1) \mod M}$. Esses dados gerados são comparados com os dados de teste através da taxa de acerto (número de bits de diferença).

Durante a geração dos dados a partir da rede inferida, espera-se que os erros (bits

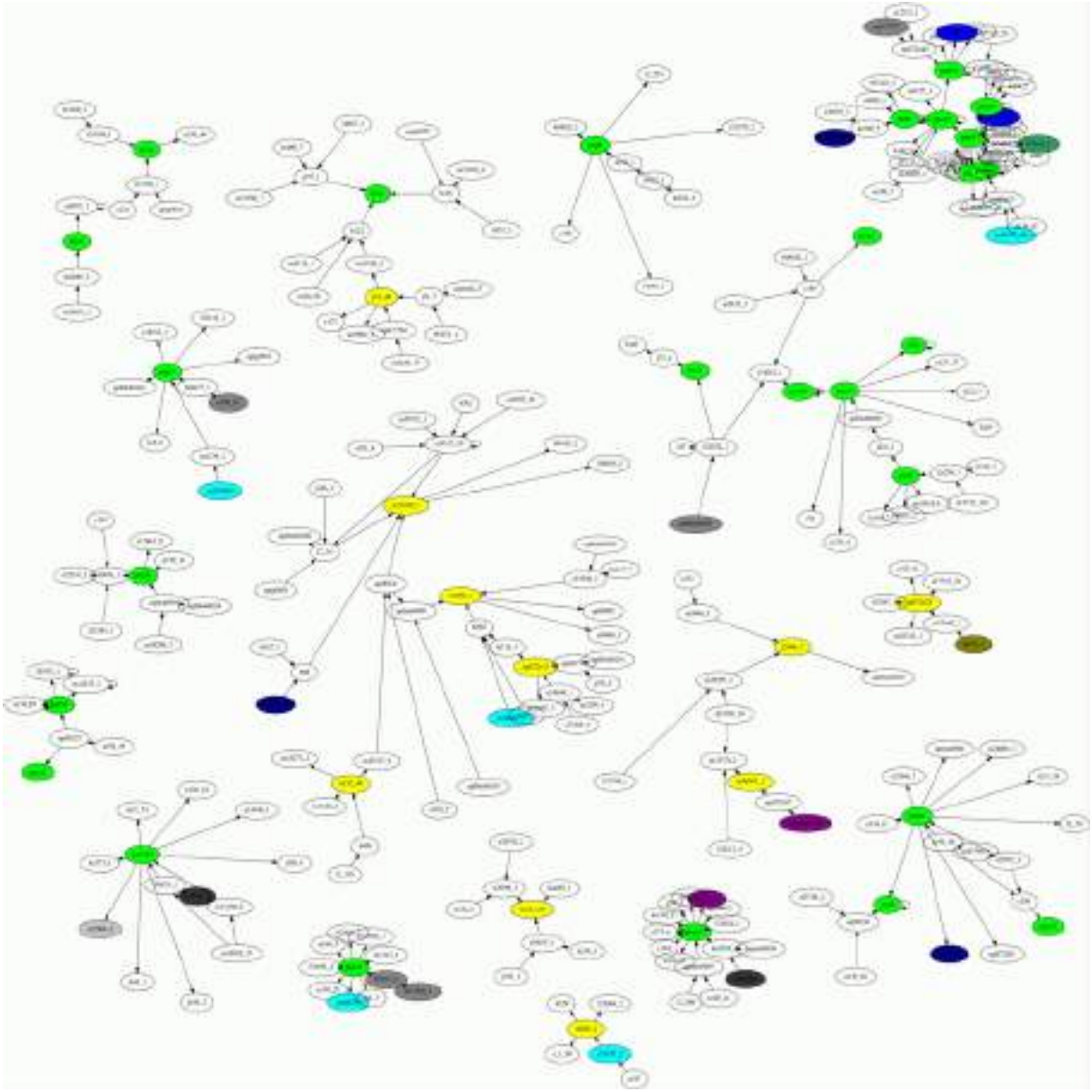


Figura 6.7: Vizinhanças ao redor dos genes sementes da glicólise (nós amarelos) e do apicoplasto (nós verdes) obtidao método Agrupamento por GLSFS, com a transferência de aprendizado do grau via KNN em redes simuladas.

de diferença) vão sendo acumulados ao longo do tempo, tendo em vista que a primeira amostra será gerada com base em uma das amostras dos dados originais, mas a segunda amostra em diante serão geradas a partir do amostra gerada anteriormente, possivelmente contendo alguns erros. Sendo assim, as taxas de acerto tendem a diminuir ao longo das amostras geradas. A Figura 6.10 e a Tabela 6.1 apresentam os resultados das médias e desvios padrões das taxas de acerto de séries de 8 amostras temporais geradas pelas redes inferidas pelos 4 métodos considerados. Como esperado, devido ao acúmulo de erros que ocorrem ao longo do tempo, a taxa de acerto tem comportamento decrescente em função do tempo para todos os métodos. Considerando apenas o primeiro instante

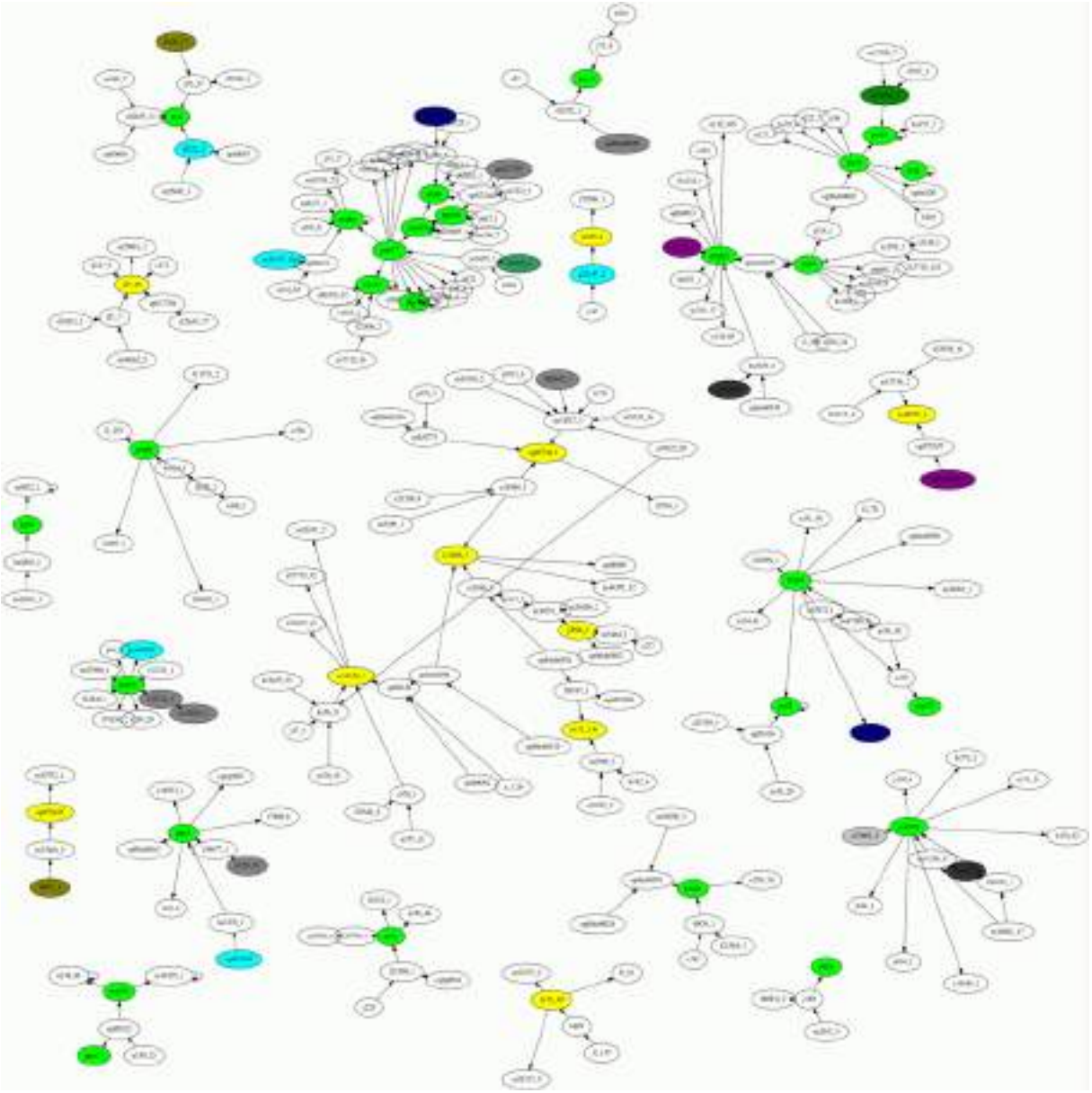


Figura 6.8: Vizinhanças ao redor dos genes sementes da glicólise (nós amarelos) e do apicoplasto (nós verdes) obtidas pelo método Agrupamento por canalização (CG), com a transferência de aprendizado do grau via KNN em redes simuladas.

de tempo (tempo 1), os 4 métodos apresentam desempenhos similares, mas a medida que o tempo avança, a diferença de desempenho entre os métodos passam a ser cada vez mais marcantes, com destaque para o método de agrupamento por canalização (CG), chegando a ter uma taxa de acerto 0.024 maior do que o obtido pelo GLSFS, que obteve o segundo melhor desempenho. Por outro lado, pode-se observar que o desvio padrão tende a aumentar ao longo do tempo, o que deve-se à quantidade de empates dos genes alvos que, junto ao erro acumulado, fazem com que a variação dos resultados seja maior no decorrer do tempo.

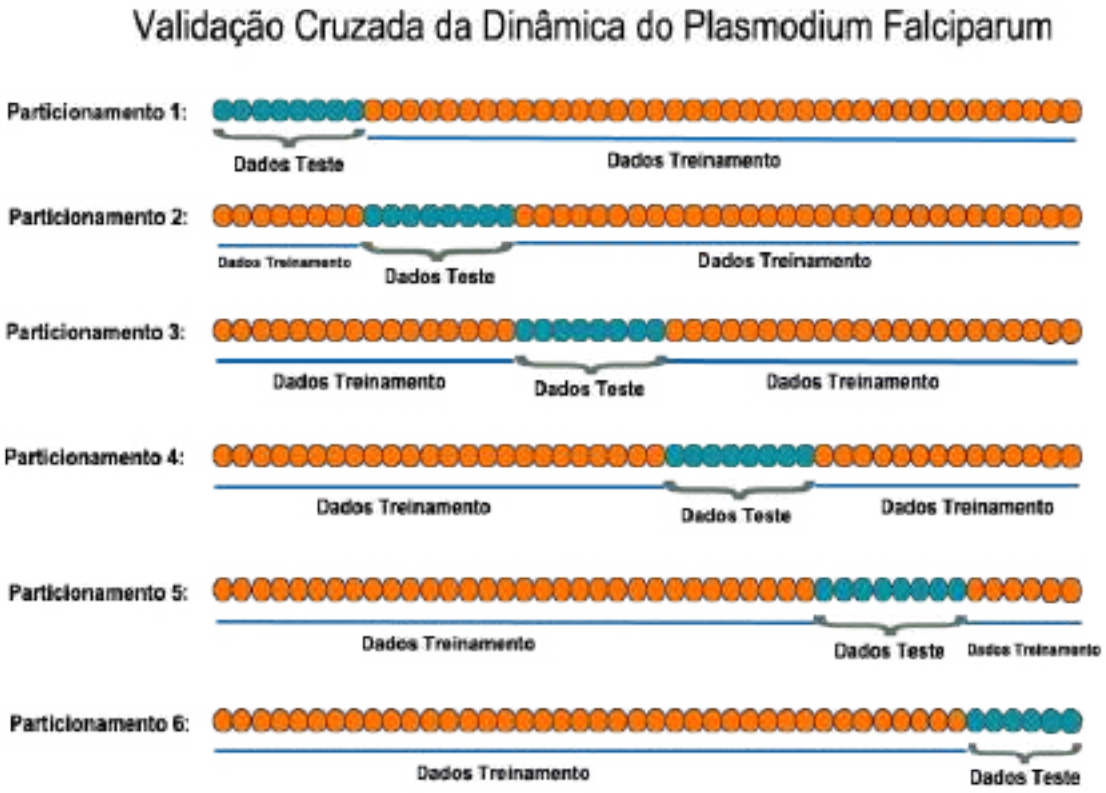


Figura 6.9: Esquema de validação cruzada com 6 particionamentos em conjunto de treinamento e conjunto de teste. Todos os particionamentos tem 46 amostras ao todo, sendo 5 particionamentos com 8 amostras de teste e um particionamento com 6 amostras de teste.

Tabela 6.1: Médias e desvios padrões correspondentes aos valores ilustrados na Figura 6.10.

		Tempo							
Método		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
CG	**Média**	0.839	0.818	0.798	0.771	0.762	0.747	0.728	0.725
	DP	0.034	0.032	0.032	0.045	0.047	0.056	0.055	0.066
GLSFS	**Média**	0.832	0.806	0.780	0.755	0.746	0.732	0.705	0.701
	DP	0.034	0.031	0.033	0.047	0.048	0.056	0.054	0.066
LG	**Média**	0.833	0.808	0.785	0.755	0.745	0.729	0.702	0.688
	DP	0.034	0.032	0.037	0.044	0.045	0.054	0.053	0.067
SA	**Média**	0.832	0.805	0.780	0.753	0.745	0.733	0.703	0.700
	DP	0.034	0.033	0.032	0.049	0.049	0.058	0.051	0.065

6.3.1 Transferência de aprendizado dos graus via KNN

Em relação a transferência de aprendizado dos graus por meio do KNN, de modo geral as taxas de acerto permaneceram as mesmas do cenário sem esse aprendizado, embora tenha havido uma ligeira melhora para os métodos CG e SA, com maior destaque para

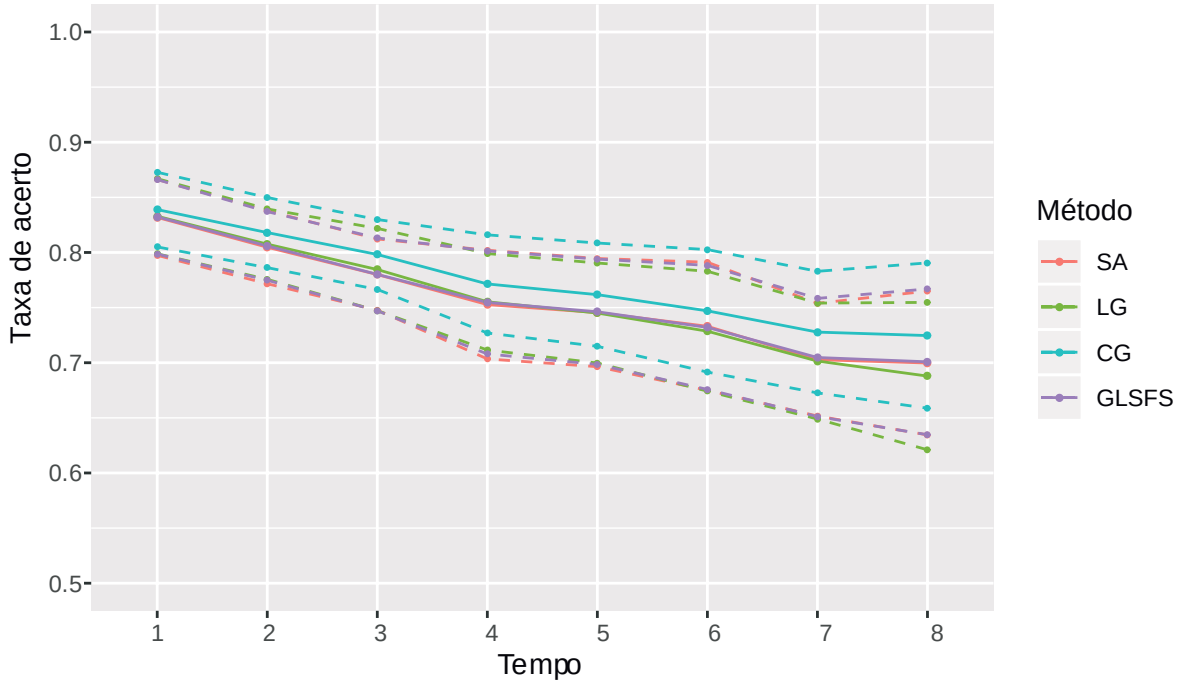


Figura 6.10: Evolução das taxas de acerto médias das dinâmicas geradas pelos 4 métodos para séries de 8 amostras temporais consecutivas que ficaram de fora dos conjuntos de treinamento, via validação cruzada. As linhas sólidas correspondem às evoluções das médias, enquanto os linhas tracejadas correspondem aos respectivos desvios padrões acima e abaixo das médias.

o SA, conforme pode ser observado na Figura 6.11. Este resultado é consequência dos resultados sobre as topologias, já que a correção das conexões na topologia, especialmente em relação a eliminação de potenciais falsos positivos (amenização da superestimação dos graus), refletem em melhoras na estimação da dinâmica, especialmente em relação ao método SA. Isso se deve ao fato da ausência de agrupamento implicar em um número considerável de instâncias mal observadas para graus maiores, resultando em uma pior generalização. Com a transferência de aprendizado, houve uma diminuição na média dos graus, reduzindo o problema da generalização. Este beneficio foi observado também no cenário de dados simulados (Capítulo 5). Embora a melhora tenha sido pequena, ela se mantém ao longo do tempo conforme pode ser observado na Figura 6.11, sugerindo que a aplicação da transferência de aprendizado dos graus por KNN tende a melhorar o poder de estimação especialmente do método sem agrupamento (SA).

Já em relação aos métodos LG e GLSFS os resultados praticamente não se alteraram ao longo do tempo com a transferência de aprendizado, conforme observado na Tabela 6.2. Cabe ressaltar que o aprendizado dos graus por KNN foi realizado com treinamento baseado em redes gabaritos com funções aleatórias, o que pode explicar porque esses métodos acabaram não sendo beneficiados em relação à geração de suas dinâmicas.

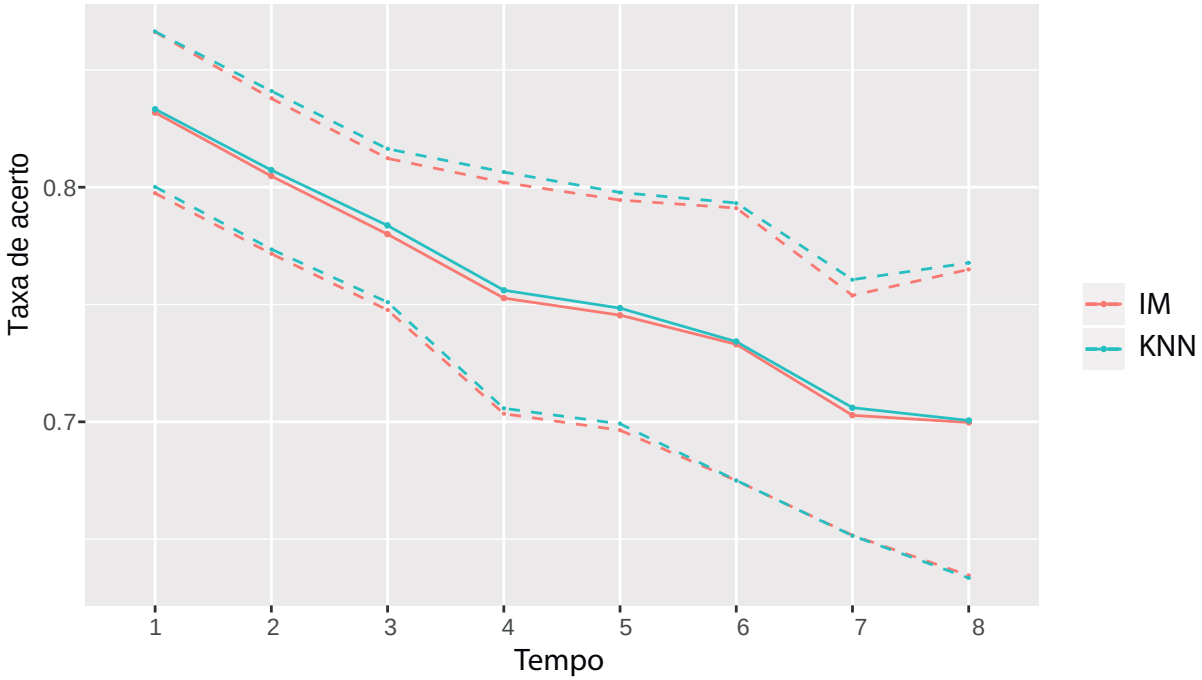


Figura 6.11: Comparação das médias das taxas de acerto das dinâmicas geradas pelas redes inferidas pelo método sem agrupamento (SA) ao longo do tempo, com e sem a transferência de aprendizado dos graus por KNN (respectivamente IM e KNN). As linhas sólidas correspondem às médias das taxas de acerto, enquanto as linhas tracejadas correspondem aos respectivos desvios padrões.

Tabela 6.2: Médias e desvios padrões das taxas de acerto das dinâmicas geradas pelos 4 métodos considerados, com e sem a transferência de aprendizado dos graus por KNN (IM e KNN, respectivamente).

		Tempo							
Método		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
CG	**IM**	0.839	0.818	0.798	0.771	0.762	0.747	0.728	0.725
	KNN	0.840	0.819	0.799	0.771	0.762	0.747	0.729	0.728
GLSFS	**IM**	0.832	0.806	0.780	0.755	0.746	0.732	0.705	0.701
	KNN	0.833	0.807	0.782	0.757	0.749	0.734	0.704	0.697
LG	**IM**	0.833	0.808	0.785	0.755	0.745	0.729	0.702	0.688
	KNN	0.833	0.807	0.784	0.755	0.744	0.728	0.701	0.687
SA	**IM**	0.832	0.805	0.780	0.753	0.745	0.733	0.703	0.700
	KNN	0.833	0.807	0.784	0.756	0.748	0.734	0.706	0.701

Capítulo 7

Conclusão

7.1 Considerações finais

A proposta desta tese consistiu no desenvolvimento de técnicas de seleção de características para inferir redes gênicas modeladas por modelos discretos, tais como as redes Booleanas, cujo princípio se baseia em reduzir o número de parâmetros de estimação (instâncias) dos valores dos preditores através de agrupamentos. Trata-se de uma extensão das pesquisas realizadas no mestrado, cujo foco foi o desenvolvimento de um método e variantes dele que consistem em agrupar em uma mesma classe de equivalência configurações que levem a um mesmo valor de combinação linear de acordo com os coeficientes lineares que otimizem uma função critério adotada [Montoya-Cubas et al., 2014, Montoya-Cubas, 2014, Montoya-Cubas et al., 2015]. Esses agrupamentos também podem ser vistos geometricamente através de cortes do reticulado Booleano por hiperplanos paralelos. Dois desses hiperplanos necessariamente interceptam um único vértice cada, sendo que esses vértices possuem distância de Hamming[1] máxima. Esses métodos efetivamente amenizam o problema da dimensionalidade, já que o número de classes de equivalência cresce linearmente com a dimensão (número de preditores), ao invés de crescer exponencialmente como é o caso das configurações originais. Isso é feito ao custo de alguma perda de informação dado que o conjunto original de configurações é mapeado para um conjunto menor (processo semelhante a uma quantização).

No entanto, a contribuição central dessa tese de doutorado foi reformular o problema de agrupamento de instâncias como um problema de busca no espaço de partições e desenvolver métodos para busca nesse espaço. Tal espaço cresce de forma superexponencial em relação a dimensão do conjunto de genes preditores para um dado gene alvo, portanto uma busca exaustiva é impensável para conjuntos com 4 ou mais preditores. Foi verificado que o espaço de partições pode ser estruturado em um reticulado de partições com ordem

[1] Número de bits de diferença entre duas cadeias binárias

parcial, de modo que uma partição é vizinha à outra se a primeira é o resultado da união de dois subconjuntos da segunda ou, de modo dual, se a primeira é o resultado da divisão de um de seus subconjuntos em dois. Isso é importante porque permite o desenvolvimento de técnicas de busca incrementais nesse reticulado que sejam ao mesmo tempo eficientes computacionalmente e que examinem um número de partições substancialmente maior do que os método baseado em agrupamento linear desenvolvidos no mestrado.

Foi desenvolvida uma primeira tentativa de busca incremental no reticulado de partições: a adaptação da busca sequencial para frente (*Sequential Forward Search* [Pudil et al., 1994]), um método guloso clássico para seleção de características, para a busca no reticulado de partições (denominado pela sigla GLSFS). A função critério que propomos é bastante simples, procurando sempre unir instâncias não observadas a outras já observadas previamente (análise multiresolução) e ao mesmo tempo procurando maximizar a informação mútua das tabelas de probabilidades condicionais do alvo dados seus candidatos a preditores. Esse método não impõe qualquer restrição em relação às instâncias que podem estar no mesmo grupo, desde que elas maximizem a informação mútua. No geral, tanto os resultados topológicos das redes inferidas como os resultados das dinâmicas geradas por essas redes se mostraram competitivos frente aos demais métodos. Entretanto uma desvantagem desse método é que ele não embute qualquer informação a priori sobre o significado biológico dos agrupamentos formados.

Na mesma linha de embutir informação a priori sobre tipos de funções que sejam frequentes em redes gênicas reais, como as funções linearmente separáveis que motivou o desenvolvimento do método de agrupamento linear, também foi desenvolvido durante o doutorado a estrategia de agrupamento por canalização, admitindo que as funções canalizadoras possuem papel fundamental na adaptação e homeostase dos sistemas biológicos [Waddington, 1942, Kauffman et al., 2004, Just et al., 2004, Martins-Jr et al., 2008, Layne et al., 2012, Li et al., 2013], Esta estratégia apresentou os melhores resultados em comparação aos outros métodos considerados em diversos cenários, tanto do ponto de vista topológico, como do ponto de vista da dinâmica gerada pelas redes inferidas. Em particular os cenários nos quais esse método obteve um certo destaque foram para redes simuladas compostas exclusivamente por funções lineares ou por funções de canalização, e também para dados de *microarray* do *Plasmodium falciparum*, um agente causador da malária. Os resultados encorajadores do método de agrupamento por canalização para esse cenário sugerem que redes gênicas reais como a do *Plasmodium falciparum* de fato possuem uma frequência considerável de funções canalizadoras, como já observado em redes gênicas reais por estudos anteriores [Waddington, 1942, Kauffman et al., 2004, Just et al., 2004, Martins-Jr et al., 2008, Layne et al., 2012, Li et al., 2013].

Ao longo dessa pesquisa, foi observado que um problema inerente à inferência de redes gênicas está relacionada à estimação da dimensão (grau) dos preditores para um

dado gene alvo. Como observado em [Montoya-Cubas et al., 2014, Montoya-Cubas, 2014, Montoya-Cubas et al., 2015] para os método de agrupamento linear e suas variantes, todos os métodos de agrupamento desenvolvidos aqui também tenderam a superestimar os graus dos genes, introduzindo falsos positivos que impactam negativamente os resultados topológicos e as dinâmicas geradas pelas redes inferidas. Essa superestimação do grau ocorrida especialmente para os métodos de agrupamento é natural, tendo em vista que eles tendem a obter tabelas de probabilidades condicionais mais enxutas, abrindo caminho para conjuntos de preditores candidatos de dimensões maiores aumentarem suas chances de compor a rede.

Para lidar com o problema de estimação dos graus, foi desenvolvido adicionalmente um método para a estimação da dimensão do subconjunto de preditores de um determinado gene alvo. Essa estimação é obtida através de aprendizado supervisionado no qual a geração aleatória de redes gabaritos fornece os rótulos (graus corretos dos conjuntos de preditores dos genes alvos), enquanto a inferência das redes fornece perfis (padrões) de como a função critério evolui à medida que a dimensão aumenta. Esta estratégia induziu a uma melhora da qualidade das redes inferidas para todos os métodos em cenários com redes geradas artificialmente. Para o caso das redes reais de *Plasmodium falciparum*, foi realizada uma transferência do aprendizado obtido com redes artificiais para esse novo cenário, e a avaliação se deu através da inferência de uma vizinhança em torno de dois conjuntos de genes sementes, um para a via glicolítica (glicólise), e outro para o apicoplasto (plastídeo). Do ponto de vista topológico, essa transferência de aprendizado levou a redes um pouco mais intramodulares (mais conexões dentro de um mesmo módulo) e menos conexões conectando os módulos distintos, sugerindo um melhor desempenho. Já do ponto de vista da dinâmica gerada pelas redes inferidas, apenas os métodos original (sem agrupamento) e o de agrupamento por canalização obtiveram alguma melhora perceptível nas taxas de acerto com a transferência do aprendizado, com um destaque para o método sem agrupamento. Para os métodos de agrupamento linear e de agrupamento por busca sequencial para frente no reticulado de partições, praticamente não houve alteração nas taxas de acerto. Uma possível explicação para isso seria devido ao fato das redes gabarito que serviram de base para o aprendizado dos graus terem sido geradas pelo modelo de redes aleatórias (Erdös-Renyi [Erdös and Rényi, 1959]), embora redes gênicas possuam topologias que não são puramente aleatórias, mas sim tendem a apresentar características topológicas de redes livres de escala (Barabasi-Albert [Barabási and Albert, 1999]) e mundo pequeno (*small world* [Watts and Strogatz, 1998]) [van Noort et al., 2004, Barabasi, 2009, Lopes et al., 2014]. Além disso, o número de genes adotados para gerar as redes artificiais foi muito inferior ao número de genes presentes nos dados reais de *Plasmodium falciparum* (50 contra 5080), de modo a privilegiar a geração de grandes quantidades de amostras para o aprendizado.

Finalmente, o código fonte e os dados deste trabalho serão disponibilizados em breve no repositório Github para fins de reprodutibilidade de todos os resultados apresentados no presente trabalho.

7.2 Trabalhos futuros

Dados os desempenhos encorajadores dos métodos de agrupamento biologicamente inspirados (linear e por canalização) propostos neste trabalho, uma possível continuação seria explorar outros tipos de agrupamentos bioinspirados como, por exemplo, restringir a exploração de funções de canalização a funções de canalização aninhadas ($n-$canalizadoras) [Layne et al., 2012, Li et al., 2013], Por exemplo, para $n = 2$, dado um gene preditor e seu correspondente valor canalizador (denotado aqui por $Z_c = a \in \{0, 1\}$), o subconjunto dos preditores restantes também formam uma função canalizadora para $Z_c \neq a$, desde que $Z_c \neq a$ não seja um valor canalizador. Nesse caso, teríamos uma função no mínimo 2-canalizadora. Entretanto, vale notar que pelo fato da canalização aninhada se tratar de uma definição recursiva, a profundidade da canalização aninhada pode ser de no máximo o número de variáveis (k), desde que a definição recursiva seja satisfeita para todas as variáveis.

Em relação ao método de agrupamento linear, nos experimentos realizados até então fixamos os pesos da combinação linear em $\{-1, 0, +1\}$. Entretanto, o método proposto é flexível, permitindo um conjunto maior de pesos possíveis (por exemplo $\{-2, -1, 0, +1, +2\}$). Portanto, um possível caminho a partir daí é testar o método para outros conjuntos de pesos e, eventualmente, desenvolver um método que combine vários conjuntos de pesos e defina aquele que obtiver o melhor desempenho. Entretanto, deve-se ter em mente que esses conjuntos de pesos devem ter um tamanho limitado, tendo em vista que o método deve permanecer computacionalmente eficiente.

Ainda na linha de desenvolvimento de métodos de agrupamento biologicamente inspirados, pode-se projetar um método hibrido, que aplique um tipo de agrupamento sob demanda, dependendo do contexto. Isso poderia ser realizado projetando um classificador que tente obter a priori o tipo de função mais provável dado o perfil de expressão de um determinado gene alvo a partir de geração de redes artificiais.

Já para o método de busca sequencial para frente no espaço de partições (denominado aqui por GLSFS), o qual não assume qualquer informação a prori sobre o tipo de funções que a rede contém majoritariamente, também há espaço para aperfeiçoamentos. Por exemplo, a função critério que orienta esse método foi muito pouco explorada até então. Uma possível melhora nesse sentido seria atribuir uma restrição em quais instâncias podem ou não ser agrupadas com base nos seus valores. A forma como a função critério

foi projetada permite que, por exemplo, instâncias com valores completamente opostos (distância de Hamming máxima) sejam agrupadas. Por exemplo, uma possível restrição seria agrupar apenas instâncias com 1 bit de diferença (distância de Hamming 1). Essa restrição poderia ser relaxada ao considerar também agrupamentos lineares, que permitem que instâncias sejam agrupadas na mesma classe de equivalência mesmo possuindo mais de 1 bit de diferença, desde que possuam o mesmo valor de combinação linear.

Ainda em relação ao método GLSFS, embora a busca sequencial para frente (SFS), uma estratégia puramente gulosa, tenha sido adotada para percorrer o espaço de partições, outros algoritmos podem ser facilmente explorados para percorrer um volume maior do reticulado sem abrir mão da eficiência computacional, como por exemplo, a busca sequencial flutuante para frente (*Sequential Floating Forward Search* - SFFS [Pudil et al., 1994]), o que pode também dividir subgrupos em dois (retroceder na busca), ao invés de apenas unir dois subgrupos em um como o SFS faz. Uma outra alternativa de melhorar o algoritmo GLSFS é realizar uma busca U-Curve [Ris et al., 2010, Reis et al., 2019] sobre o reticulado de partições procurando obter o melhor agrupamento. Além disso, realizar uma estimação da dimensão-VC [Vapnik and Chervonenkis, 2015] com base nas características dos dados pode auxiliar a restringir a busca a uma ou poucas camadas do reticulado de partições, de modo a permitir uma busca mais aprofundada nessas camadas e, consequentemente, obter agrupamentos mais próximos da solução ótima.

Em relação à abordagem de transferência de aprendizado dos graus dos preditores, um grande leque de perspectivas futuras foi aberta a partir desta tese. Uma dessas perspectivas seria projetar um algoritmo de sintonização automática de parâmetros do algoritmo de transferência de aprendizado. Existem também diversos algoritmos de aprendizado supervisionado que podem ser analisados nesse contexto. Também pode-se aplicar outros modelos de aprendizado, como por exemplo o aprendizado por reforço, algoritmos evolutivos (e.g. genéticos), meta-heurísticas, dentre outros. Adicionalmente, o modelo de geração das redes gabaritos que servem de base para gerar os conjuntos de treinamento poderia ter características topológicas mais condizentes com o esperado em redes biológicas, tais como o modelo de redes livres de escala (Barabási-Albert [Barabási and Albert, 1999]), o modelo de redes mundo pequeno (*small world*) [Watts and Strogatz, 1998], ou mesmo um misto de ambos os modelos. A hipótese é que modelos mais realistas de geração das redes gabaritos combinados com os melhores algoritmos de classificação para esse contexto poderiam beneficiar ainda mais todos os métodos em questão, incluindo o método original sem agrupamento. Outro ponto importante é o tamanho das redes artificiais geradas para o treinamento nos experimentos realizados, com dimensão cerca de 100 vezes inferior ao dos dados reais. Encontrar um melhor balanço entre o tamanho das redes artificiais geradas e quantidade de amostras geradas para o treinamento também pode induzir a desempenhos melhores. Finalmente, um afinamento desse aprendizado po-

deria ser realizado baseando-se também no numero de amostras disponíveis e no número de genes presentes nos dados.

Outro aprimoramento em relação à transferência de aprendizado seria o desenvolvimento de uma estrategia de retroalimentação dos resultados obtidos para dados reais de modo a reajustar os parâmetros do modelo de geração das redes sintéticas e do algoritmo de aprendizagem. Essa retroalimentação seria orientada pela avaliação das topologias e dinâmicas inferidas a partir dos dados reais.

Os resultados obtidos para os dados reais de *microarray* de *Plasmodium falciparum*, embora encorajadores, necessitam de uma validação biológica dos genes encontrados nas vizinhanças dos módulos de interesse, checando se de fato eles estão relacionados com as vias do apicoplasto (plastídeo) e da glicólise. Tal validação pode atestar a utilidade do método para realizar anotação funcional de alguns desses genes cujas funções ainda sejam desconhecidas.

Ainda em relação a análise dos dados do *Plasmodium falciparum*, pode-se considerar uma separação dos dados de expressão por fases do parasita, para obter diferentes redes, uma para cada fase, e testar o poder de inferência dos métodos para cada fase. Esta validação poderia seguir o mesmo protocolo apresentado na Seção 6.3, em que cada particionamento seria constituído de amostras de treinamento que representem a maior parte de uma das fases do parasita, enquanto uma pequena porção dessa fase seja composta por amostras de teste. Um desafio natural é o número de amostras presentes em cada uma das 3 fases (cerca de 16 em média), o que pode evidenciar ainda mais a utilidade dos métodos de agrupamento desenvolvidos neste trabalho em situações que contam com um limite severo no número de amostras temporais disponíveis.

Este trabalho assumiu uma das hipóteses simplificadoras em que os dados de expressão gênica seguem uma cadeia de Markov de primeira ordem (um dos axiomas do modelo PGN [Barrera et al., 2007]). Entretanto, como os métodos de agrupamento de instâncias possuem um maior poder de estimação das probabilidades condicionais, pode-se avaliá-los assumindo cadeias de Markov de segunda ordem ou superior.

Em relação aos critérios de validação dos resultados das dinâmicas geradas pelas redes inferidas, uma possibilidade seria construir redes gabaritos relativamente pequenas (cerca de 20 a 30 genes) e realizar una avaliação da dinâmica completa que eles geram (ao invés de avaliar a dinâmica a partir de uma amostragem aleatória de estados iniciais como feito neste trabalho). Dessa forma, pode-se realizar uma análise comparativa dos atratores e das bacias de atração geradas pelas redes gabaritos e pelas redes inferidas, com potencial para desenvolver um critério de validação da dinâmica mais robusto.

Outra perspectiva aberta seria realizar a comparação dos métodos de agrupamento de instâncias desenvolvidos aqui com outros algoritmos que já foram analisados para

os dados de expressão gênica do *Plasmodium falciparum*, como por exemplo o método GeNICE [Jacomini et al., 2017]. Eventualmente estratégias híbridas que combinem esses métodos podem efetivamente melhorar a qualidade das redes gênicas inferidas.

Embora o foco desta tese tenha sido no problema de inferência de redes gênicas, vale ressaltar que boa parte de suas contribuições são relevantes para as áreas de reconhecimento de padrões e aprendizado de máquina de modo geral, podendo ser aplicadas em diversos outros contextos, especialmente para conjuntos de dados de alta dimensionalidade (contendo pelo menos da ordem de centenas de variáveis) e um número limitado de amostras (no máximo poucas dezenas).

Referências Bibliográficas

[Albert, 2005] Albert, R. (2005). Statistical mechanics of complex networks. *J Cell Sci*, 118(21):4947–4957. 26, 27

[Albert and Othmer, 2003] Albert, R. and Othmer, H. G. (2003). The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster. *Journal of Theoretical Biology*, 223(1):1–18. 12

[Anastassiou, 2007] Anastassiou, D. (2007). Computational analysis of the synergy among multiple interacting genes. *Molecular Systems Biology*, 3(83). 2

[Banf and Rhee, 2017] Banf, M. and Rhee, S. Y. (2017). Computational inference of gene regulatory networks: Approaches, limitations and opportunities. *Biochimica et Biophysica Acta (BBA) – Gene Reglatory Mechanisms*, 1860(1):41–52. 2, 4

[Barabasi, 2009] Barabasi, A. L. (2009). Scale-Free Networks: A Decade and Beyond. *Science*, 325(5939):412–413. 26, 27, 114

[Barabási and Albert, 1999] Barabási, A. L. and Albert, R. (1999). Emergence of scaling in Random Networks. *Science*, 286(5439):509–512. 26, 27, 114, 116

[Barrera et al., 2007] Barrera, J., Cesar, R. M., Martins, D. C., Vêncio, R. Z., Merino, E. F., Yamamoto, M. M., Leonardi, F. G., Pereira, C. A. d. B., and del Portillo, H. A. (2007). Constructing probabilistic genetic networks of plasmodium falciparum from dynamical expression signals of the intraerythrocytic development cycle. In *Methods of Microarray Data Analysis V*, pages 11–26. Springer. 2, 4, 12, 13, 15, 17, 18, 23, 45, 98, 99, 100, 117

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. 18, 30

[Borelli et al., 2013] Borelli, F. F., de Camargo, R. Y., Martins-Jr, D. C., and Rozante, L. C. S. (2013). Gene regulatory networks inference using a multi-gpu exhaustive search algorithm. *BMC Bioinformatics*, 14(S5). 18

[Bozdech et al., 2003] Bozdech, Z., Llinás, M., Pulliam, B. L., Wong, E. D., Zhu, J., and DeRisi, J. L. (2003). The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *Plos Biology*, 1(1). 98, 99

[Brun et al., 2005] Brun, M., Dougherty, E. R., and Shmulevich, I. (2005). Steady-state probabilities for attractors in probabilistic boolean networks. *Signal Processing*, 85(10):1993–2013. 15

[Butte and Kohane, 2000] Butte, A. and Kohane, I. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Pacific Symposium on Biocomputing*, pages 418–429. 18

[Carastan-Santos et al., 2017] Carastan-Santos, D., Camargo, R. Y., Martins-Jr, D. C., Song, S. W., and Rozante, L. C. S. (2017). Finding exact hitting set solutions for systems biology applications using heterogeneous gpu clusters. *Future Generation Computer Systems*, 67:418–429. 4

[Carvalho, 2006] Carvalho, A. C. P. L. F. (2006). Grandes desafios da pesquisa em computação no brasil 2006 –2016. *Relatório do Seminário realizado pela SBC*. 7

[Commons, 2014] Commons, W. (2014). File:set partitions 4; hasse; circles.svg — wikimedia commons, the free media repository. [Online; accessed 26-October-2019]. 33

[Costa et al., 2008] Costa, L. F., Rodrigues, F. A., and Cristino, A. S. (2008). Complex networks: the key to systems biology. *Genetics and Molecular Biology*, 31(3):591–601. 26, 27

[Cover and Hart, 1967] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27. 25

[Crick, 1970] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563. 2

[Davidich and Bornholdt, 2008] Davidich, M. I. and Bornholdt, S. (2008). Boolean network model predicts cell cycle sequence of fission yeast. *PLoS ONE*, 3(2):e1672. 12, 16, 100

[De-Smet and Marchal, 2010] De-Smet, R. and Marchal, K. (2010). Advantages and Limitations of Current Network Inference Methods. *Nature Reviews Microbiology*, 8(10):717–729. 4

[Delgado and Gómez-Vela, 2019] Delgado, F. M. and Gómez-Vela, F. (2019). Computational methods for gene regulatory networks reconstruction and analysis: A review. *Artificial Intelligence in Medicine*, 95:133–145. 2, 4

[D'haeseleer et al., 1999] D'haeseleer, P., Liang, S., and Somgyi, R. (1999). Tutorial: Gene expression data analysis and modeling. In *Pacific Symposium on Biocomputing*, Hawaii. 2, 14

[Dougherty, 2011] Dougherty, E. R. (2011). Validation of gene regulatory networks: scientific and inferential. *Briefings in Bioinformatics*, 12(3):245–252. 28

[Dougherty et al., 2001] Dougherty, E. R., Barrera, J., Mozelle, G., Kim, S., and Brun, M. (2001). Multiresolution analysis for optimal binary filters. *J. Math. Imaging Vis.*, 14(1):53–72. 5, 8

[Dougherty et al., 2007] Dougherty, E. R., Brun, M., Trent, J., and Bittner, M. L. (2007). A conditioning-based model of contextual regulation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 15

[Dougherty et al., 2000] Dougherty, E. R., Kim, S., and Chen, Y. (2000). Coefficient of determination in nonlinear signal processing. *EURASIP Journal on Signal Processing*, 80(10):2219–2235. 22

[Dougherty et al., 2008] Dougherty, J., Tabus, I., and Astola, J. (2008). Inference of gene regulatory networks based on a universal minimum description length. *EURASIP Journal on Bioinformatics and Systems Biology*, 2008:1–11. 18

[Duda et al., 2000] Duda, R. O., Hart, P. E., and Stork, D. (2000). *Pattern Classification*. Wiley-Interscience, NY. 22

[Eberwine et al., 2014] Eberwine, J., Sul, J., Bartfai, T., and Kim, J. (2014). The promise of single-cell sequencing. *Nature Methods*, 11:25–27. 1, 11

[Erdös and Rényi, 1959] Erdös, P. and Rényi, A. (1959). On random graphs. *Publ. Math. Debrecen*, 6:290–297. 26, 27, 114

[Espinosa-Soto et al., 2004] Espinosa-Soto, C., Padilla-Longoria, P., and Alvarez-Buylla, E. R. (2004). A gene regulatory network model for cell-fate determination during arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. *Plant Cell*, 16(11):2923–2939. 12

[Faith et al., 2007] Faith, J., Hayete, B., Thaden, J., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J., and Gardner, T. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1):259–265. 18

[Farkas et al., 2003] Farkas, I. J., Jeong, H., Vicsek, T., Barabási, A.-L., and Oltvai, Z. N. (2003). The topology of the transcription regulatory network in the yeast, saccharomyces cerevisiae. *Physica A: Statistical Mechanics and its Applications*, 318(3-4):601–612. 26, 27

[Faure et al., 2006] Faure, A., Naldi, A., Chaouiya, C., and Thieffry, D. (2006). Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 22(14):e124–131. 12

[Friedman, 2004] Friedman, N. (2004). Inferring Cellular Networks Using Probabilistic Graphical Models. *Science*, 303(5659):799–805. 2, 12

[Friedman et al., 2000] Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620. 2, 12, 13

[Gastner and Newman, 2006] Gastner, M. T. and Newman, M. E. J. (2006). The spatial structure of networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 49:247–252. 26

[Guelzim et al., 2002] Guelzim, N., Bottani, S., Bourgine, P., and Képès, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, 31:60–63. 26, 27

[Hashimoto et al., 2004] Hashimoto, R. F., Kim, S., Shmulevich, I., Zhang, W., Bittner, M. L., and Dougherty, E. R. (2004). Growing genetic regulatory networks from seed genes. *Bioinformatics*, 20(8):1241–1247. 18

[Hecker et al., 2009] Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models-a review. *Biosystems*, 96:86–103. 2, 3, 4

[Huang et al., 2009] Huang, S., Ernberg, I., and Kauffman, S. (2009). Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective. *Seminars in Cell & Developmental Biology*, 20(7):869–876. 15

[Huynh-Thu and Sanguinetti, 2019] Huynh-Thu, V. A. and Sanguinetti, G. (2019). Gene regulatory network inference: An introductory survey. In *Methods in Molecular Biology*, volume 1883. Humana Press, New York, NY. 2, 4

[Ivanov and Dougherty, 2006] Ivanov, I. and Dougherty, E. R. (2006). Modeling genetic regulatory networks: continuous or discrete? *Journal of Biological Systems*, 14(2):219–229. 3, 13

[Jacomini et al., 2017] Jacomini, R. S., Martins-Jr, D. C., Silva, F. L., and Costa, A. H. R. (2017). Genice: A novel framework for gene network inference by clustering, exhaustive search, and multivariate analysis. *Journal of Computational Biology*, 24(8). 4, 13, 18, 23, 98, 118

[Jain et al., 2000] Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37. 4, 18, 20

[Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*. 26, 27

[Jong, 2002] Jong, H. D. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103. 2, 12

[Just et al., 2004] Just, W., Shmulevich, I., and Konvalina, J. (2004). The number and probability of canalizing functions. *Physica D: Nonlinear Phenomena*, 197(3):211 – 221. 16, 113

[Karlebach and Shamir, 2008] Karlebach, G. and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol*, 9(10):770–780. 2, 12

[Kauffman et al., 2004] Kauffman, S., Peterson, C., Samuelsson, B., and Troein, C. (2004). Genetic networks with canalyzing boolean rules are always stable. *Proceedings of the National Academy of Sciences*, 101(49):17102–17107. 15, 113

[Kauffman, 1969] Kauffman, S. A. (1969). Homeostasis and differentiation in random genetic control networks. *Nature*, 224(215):177–178. 2, 3, 12, 13, 14, 100

[Kauffman, 1993] Kauffman, S. A. (1993). *The Origins of Order*. Oxford University Press. 26

[Kelemen et al., 2008] Kelemen, A., Abraham, A., and Chen, Y. (2008). *Computational Intelligence in Bioinformatics*. Springer. 13

[Konidaris and Barto, 2007] Konidaris, G. and Barto, A. G. (2007). Building portable options: Skill transfer in reinforcement learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 895–900. 41

[Kotiang and Eslami, 2020] Kotiang, S. and Eslami, A. (2020). A probabilistic graphical model for system-wide analysis of gene regulatory networks. *Bioinformatics*, 36(10):3192–3199. 2, 4

[Lähdesmäki et al., 2006] Lähdesmäki, H., Hautaniemi, S., Shmulevich, I., and Yli-Harjaa, O. (2006). Relationships between probabilistic boolean networks and dynamic bayesian networks as models of gene regulatory networks. *Signal Processing*, 86(4):814–834. 13

[Layne et al., 2012] Layne, L., Dimitrova, E., and Macauley, M. (2012). Nested canalyzing depth and network stability. *Bulletin of mathematical biology*, 74(2):422–433. 15, 100, 113, 115

[Li et al., 2004] Li, F., Long, T., Lu, Y., Ouyang, Q., and Tang, C. (2004). The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci. USA*, 101(14):4781–4786. 12, 15, 16, 100

[Li and Lu, 2005] Li, L. M. and Lu, H. H. S. (2005). Explore biological pathways from noisy array data by directed acyclic boolean networks. *Journal of Computational Biology*, 12(2):170–185. 12

[Li et al., 2006] Li, S., Assmann, S. M., and Albert, R. (2006). Predicting essential components of signal transduction networks: A dynamic model of guard cell abscisic acid signaling. *PLoS Biology*, 4(10):e312. 12

[Li et al., 2013] Li, Y., Adeyeye, J. O., Murrugarra, D., Aguilar, B., and Laubenbacher, R. (2013). Boolean nested canalizing functions: A comprehensive analysis. *Theoretical Computer Science*, 481:24–36. 15, 100, 113, 115

[Liang et al., 1998] Liang, S., Fuhrman, S., and Somogyi, R. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Simposium on Biocomputing*, volume 3, pages 18–29. 4, 18

[Lin, 1991] Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151. 22

[Lopes et al., 2011] Lopes, F. M., Cesar-Jr, R. M., and Costa, L. F. (2011). Gene expression complex networks: synthesis, identification and analysis. *Journal of Computational Biology*, 18:1353–1367. 23

[Lopes et al., 2014] Lopes, F. M., Martins-Jr., D. C., Barrera, J., and Cesar-Jr, R. M. (2014). A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks. *Information Sciences*, 272:1–15. 4, 18, 23, 27, 55, 60, 114

[Lopes et al., 2008] Lopes, F. M., Martins-Jr, D. C., and Cesar-Jr, R. M. (2008). Feature selection environment for genomic applications. *BMC Bioinformatics*, 9(451). 4, 13, 17, 18, 23

[Ma et al., 2020] Ma, B., Fang, M., and Jiao, X. (2020). Inference of gene regulatory networks based on nonlinear ordinary differential equations. *Bioinformatics*, btaa032. 2, 12

[Marbach et al., 2012] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., , Prill, R. J., Camacho, D. M., and Zimmer, R. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, 9:796–814. 4

[Marbach et al., 2010] Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceeings of the National Academy of Sciences*, 107(14):6286–6291. 2

[Marco et al., 2019] Marco, A. G., Gazziro, M. A., and Martins-Jr, D. C. (2019). High performance computing architectures analysis for gene networks inference. In *Anais Principais do XX Simpósio em Sistemas Computacionais de Alto Desempenho (WSCAD)*, pages 49–60, Campo Grande. SBC. 4

[Margolin et al., 2006] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla-Favera, R., and Califano, A. (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1). 18

[Markowetz and Spang, 2007] Markowetz, F. and Spang, R. (2007). Inferring Cellular Networks – A Review. *BMC Bioinformatics*, 8(Suppl 6):S5. 4

[Martins-Jr., 2008] Martins-Jr., D. C. (2008). *Seleção de características e predição intrinsecamente multivariada em identificação de redes de regulação gênica*. PhD thesis, Instituto de Matemática e Estatística - Universidade de São Paulo, Rua do Matão, 1010. xiii, 12, 23

[Martins-Jr et al., 2008] Martins-Jr, D. C., Braga-Neto, U., Hashimoto, R. F., Dougherty, E. R., and Bittner, M. L. (2008). Intrinsically multivariate predictive genes. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):424–439. 2, 113

[Martins-Jr et al., 2006] Martins-Jr, D. C., Cesar-Jr, R. M., and Barrera, J. (2006). W-operator window design by minimization of mean conditional entropy. *Pattern Analysis & Applications*, 9:139–153. 22

[Martins-Jr et al., 2016] Martins-Jr, D. C., Lopes, F. M., and Ray, S. S. (2016). Inference of gene regulatory networks by topological prior information and data integration. In *Emerging Research in the Analysis and Modeling of Gene Regulatory Networks*, chapter 1, pages 1–51. IGI Global. 2, 4, 18, 23, 55, 60

[McCluskey, 1956] McCluskey, E. J. (1956). Minimization of boolean functions. *Bell Syst Tech, J*, 35(5):1417–1444. 42

[Mestl et al., 1995] Mestl, T., Plahte, E., and Omholt, S. W. (1995). A Mathematical Framework for Describing and Analysing Gene Regulatory Networks. *Journal of Theoretical Biology*, 176:291–300. 2, 12

[Montagna et al., 2020] Montagna, S., Braccini, M., and Roli, A. (2020). The impact of self-loops on boolean networks attractor landscape and implications for cell differentiation modelling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2, 13, 14

[Montoya-Cubas, 2014] Montoya-Cubas, C. F. (2014). Seleção de características em inferência de redes de interação gênica a partir de conjuntos reduzidos de amostras. Master's thesis, UFABC. 5, 112, 114

[Montoya-Cubas et al., 2014] Montoya-Cubas, C. F., Martins-Jr, D. C., Santos, C. S., and Barrera, J. (2014). Gene networks inference through linear grouping of variables. In *14th IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 243–250, Boca Raton, FL. 5, 112, 114

[Montoya-Cubas et al., 2015] Montoya-Cubas, C. F., Martins-Jr, D. C., Santos, C. S., and Barrera, J. (2015). Linear grouping of predictor instances to infer gene networks. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 4:34. 4, 5, 8, 13, 18, 22, 23, 39, 47, 55, 60, 98, 112, 114

[Nakariyakul and Casasent, 2009] Nakariyakul, S. and Casasent, D. P. (2009). An improvement on floating search algorithms for feature subset selection. *Pattern Recognition*, 42(9):1932–1940. 20

[Narasinham et al., 2009] Narasinham, S., Rengaswamy, R., and Vadigepalli, R. (2009). Structural properties of gene regulatory networks: Definitions and connections. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 6(1):158–170. 26

[Peng et al., 2005] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE TPAMI*, 27(8):1226–1238. 18

[Porter et al., 2001] Porter, D. A., Krop, I. E., Nasser, S., Sgroi, D., Kaelin, C. M., Marks, J. R., Riggins, G., and Polyak, K. (2001). A sage (serial analysis of gene expression) view of breast tumor progression. *Cancer Research*, 61(15):5697–5702. 17

[Przulj et al., 2004] Przulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515. 26

[Pudil et al., 1994] Pudil, P., Novovicová, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125. 20, 21, 36, 113, 116

[Rani et al., 2018] Rani, S. S., Nagendra Rao, D., and Vatsal, S. (2018). Review on neural networks associative memory models. *Int. J. Pure Appl. Math*, 120(6):3143–3154. 16

[Reis, 2012] Reis, M. S. (2012). *Minimização de curvas decomponíveis em curvas em U definidas sobre cadeias de posets - algoritmos e aplicações*. PhD thesis, Instituto de Matemática e Estatística - Universidade de São Paulo, Rua do Matão, 1010. xiii, 21

[Reis et al., 2019] Reis, M. S., Estrela, G., Ferreira, C. E., and Barrera, J. (2019). Optimal boolean lattice-based algorithms for the u-curve optimization problem. *Information Sciences*, 471:97–114. 20, 116

[Ris et al., 2010] Ris, M., Martins-Jr, D. C., and Barrera, J. (2010). U-curve: A branch-and-bound optimization algorithm for u-shaped cost functions on boolean lattices applied to the feature selection problem. *Pattern Recognition*, 43(3):557–568. 20, 116

[Ristevski, 2013] Ristevski, B. (2013). A survey of models for inference of gene regulatory networks. nonlinear analysis: Modelling and control. *Nonlinear Analysis: Modelling and Control*, 18(4):444–465. 2, 4

[Sánchez and Thieffry, 2001] Sánchez, L. and Thieffry, D. (2001). A logical analysis of the drosophila gap-gene system. *Journal of Theoretical Biology*, 211(2):115–141. 12

[Shalon et al., 1996] Shalon, D., Smith, S. J., and Brown, P. O. (1996). A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. *Genome Res*, pages 639–45. 1, 11

[Shi et al., 2020] Shi, N., Zhu, Z., Tang, K., Parker, D., and He, S. (2020). ATEN: And/Or tree ensemble for inferring accurate Boolean network topology and dynamics. *Bioinformatics*, 36(2):578–585. 2, 4

[Shmulevich and Dougherty, 2014] Shmulevich, I. and Dougherty, E. R. (2014). *Genomic Signal Processing*. Princeton University Press. 3, 4, 11, 12, 13, 14, 15

[Shmulevich et al., 2002] Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W. (2002). Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274. 2, 3, 12, 13, 15

[Snoep and Westerhoff, 2005] Snoep, J. L. and Westerhoff, H. V. (2005). From isolation to integration, a systems biology approach for building the silicon cell. *Topics in Current Genetics*, 13:13–30. 1

[Somol and Pudil, 2004] Somol, P. and Pudil, P. (2004). Fast branch & bound algorithms for optimal feature selection. *Pattern Analysis and Machine Intelligence*, 26(7):900–912. 20

[Somol et al., 1999] Somol, P., Pudil, P., Novovicová, J., and Paclík, P. (1999). Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, 20:1157–1163. 20

[Song et al., 2009] Song, M. J., Lewis, C. K., Lance, E. R., Chesler, E. J., Yordanova, R. K., Langston, M. A., Lodowski, K. H., and Bergeson, S. E. (2009). Reconstructing generalized logical networks of transcriptional regulation in mouse brain from temporal gene expression data. *EURASIP J Bioinform Syst Biol.*, 2009(1):545176. 2

[Styczynski and Stephanopoulos, 2005] Styczynski, M. P. and Stephanopoulos, G. (2005). Overview of computational methods for the inference of gene regulatory networks. *Computers & Chemical Engineering*, 29(3):519–534. 3, 13

[Taylor et al., 2007] Taylor, M. E., Whiteson, S., and Stone, P. (2007). Transfer via inter-task mappings in policy search reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 41

[Theodoridis and Koutroumbas, 1999] Theodoridis, S. and Koutroumbas, K. (1999). *Pattern Recognition*. Academic Press, USA, 1st edition. 18, 22

[Thomas, 1991] Thomas, R. (1991). Regulatory Networks seen as Asynchronous Automata: A Logical Description. *Journal of Theoretical Biology*, 153:01–23. 2

[Tovar et al., 2019] Tovar, C. R. P., Araujo, E., Carastan-Santos, D., Martins-Jr, D. C., and Rozante, L. C. S. (2019). Finding attractors in biological models based on boolean dynamical systems using hitting set. In *19th IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*, Athens. 2

[Tran et al., 2013] Tran, V., McCall, M., McMurray, H., and Almudevar, A. (2013). On the underlying assumptions of threshold boolean networks as a model for genetic regulatory network behavior. *Frontiers in Genetics*, 4:263. 32

[van Noort et al., 2004] van Noort, V., Snel, B., and Huynen, M. A. (2004). The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.*, 5(3):280–284. 114

[Vapnik and Chervonenkis, 2015] Vapnik, V. N. and Chervonenkis, A. Y. (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer. 116

[Velculescu et al., 1995] Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270:484–487. 1, 11

[Waddington, 1942] Waddington, C. H. (1942). Canalization of development and the inheritance of acquired characters. *Nature*, pages 563–565. 15, 100, 113

[Wang et al., 2009] Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63. 1, 11

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440—-442. 26, 114, 116

[Wicik et al., 2020] Wicik, Z., Eyileten, C., Jakubik, D., ao, R. P., Siller-Matula, J. M., and Postula, M. (2020). ACE2 interaction networks in COVID-19: a physiological framework for prediction of outcome in patients with cardiovascular risk factors. *bioRxiv*. 8

[Wilf, 1994] Wilf, H. S. (1994). *Generatingfunctionology*. Academic Press, Boston, MA. 31

[Zanudo et al., 2011] Zanudo, J. G., Aldana, M., and Martínez-Mekler, G. (2011). Boolean threshold networks: Virtues and limitations for biological modeling. In *Information Processing and Biological Systems*, pages 113–151. Springer. 16

[Zhang et al., 2006] Zhang, Y., Qian, M., Ouyang, Q., Deng, M., Li, F., and Tang, C. (2006). Stochastic model of yeast cell-cycle network. *Physica D*, 219(1):35–39. 12, 15

[Zhao et al., 2008] Zhao, W., Serpedin, E., and Dougherty, E. R. (2008). Inferring connectivity of genetic regulatory networks using information-theoretic criteria. *IEEE/ACM TCBB*, 5(2):262–274. 18