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

Dissertação de Mestrado

Ray Dueñas Jimenez

ALGORITMOS GENÉTICOS EM INFERÊNCIA DE REDES GÊNICAS

Santo André - SP
Março de 2014

Pós-Graduação em Ciência da Computação

Dissertação de Mestrado

Ray Dueñas Jimenez

ALGORITMOS GENÉTICOS EM INFERÊNCIA DE REDES GÊNICAS

Trabalho apresentado como requisito parcial para obtenção do título de Mestre em Ciência da Computação no Pós-Graduação em Ciência da Computação da Universidade Federal do ABC, sob orientação do David Correa Martins Junior.

Santo André - SP

Março de 2014

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

Santo André, _____ de _____ de 20_____.

Assinatura do autor: _____

Assinatura do orientador: _____

Ray Dueñas Jimenez

ALGORITMOS GENÉTICOS EM INFERÊNCIA DE REDES GÊNICAS

Essa Dissertação foi julgada e aprovada para a obtenção do grau de Mestre em Ciência da Computação no curso de Pós-Graduação em Ciência da Computação da Universidade Federal do ABC.

Santo André - SP - Março de 2014

Ronaldo Cristiano Prati

Coordenador do Curso

BANCA EXAMINADORA

David Correa Martins Junior

Orientador e Presidente

Ana Carolina Lorena
(ICT-UNIFESP)

Carlos da Silva dos Santos
(CMCC-UFABC)

Ronaldo Fumio Hashimoto
(IME-USP)

Luiz Carlos da Silva Rozante
(CMCC-UFABC)

Agradecimentos

Primeiramente, quero agradecer ao bom amigo, guia e professor David Correa Martins Jr, por ter orientado o meu mestrado, ajudado em todo o processo do mestrado e confiado no meu trabalho.

Ao professor Carlos da Silva dos Santos, pelo apoio em tirar duvidas no processo de desenvolvimento do trabalho.

Ao professor Ronaldo Prati, pela ajuda à distancia nos trâmites de documentação quando eu ainda estava em Cusco e na chegada ao Brasil, e pelas disciplinas Mineração de Dados e Metodologia de Pesquisa em Computação oferecidas por ele, as quais foram bastante úteis durante o desenvolvimento do mestrado.

Gostaria de agradecer especialmente ao meu amigo Carlos Fernando Montoya Cubas, pela ajuda durante todo o mestrado, seja nos estudos para as disciplinas do curso, seja desenvolvendo ferramentas para construção de redes artificiais, as quais serviram para gerar os dados de entrada para os nossos métodos desenvolvidos.

À UFABC e à CAPES pelo apoio financeiro concedido ao mestrado no Brasil.

À Rede Vision/eScience do IME-USP que, por meio do auxílio da FAPESP "processo # 2011/50761-2", do CNPq, da CAPES e da NAP eScience - PRP - USP, disponibilizou sua rede de computadores para execução dos experimentos.

Finalmente, sou eternamente grato aos meus pais Josefina e Ramiro, meus irmãos Redy, Red e Ted, meus sobrinhos Joaquin, José, Aurea e Blue, minhas cunhadas Yndira e Ana, e toda minha familia em geral, pelo apoio moral à distancia (alguns já nas estrelas). À minha amiga, cúmplice e namorada Shandira, quem esteve sempre no meu lado e a todos meus amigos no Perú e amigos de todas as nacionalidades que conheci no Brasil.

Resumo

A inferência de redes de regulação gênica (do inglês: *gene regulatory networks* - GRN) a partir de dados de expressão gênica é um problema importante na área da biologia sistêmica (do inglês: *systems biology*), na qual o principal objetivo é compreender os mecanismos moleculares subjacentes a doenças para o desenvolvimento de tratamentos médicos e de drogas. Tal problema envolve a estimação das dependências e das funções lógicas que governam essas interações para prover um modelo que explique o conjunto de dados (usualmente obtido de sinais de expressão gênica) sobre o qual a estimação se baseia. Neste trabalho, é proposto um método baseado em algoritmos genéticos para inferir redes gênicas, cuja idéia principal consiste em aplicar um algoritmo genético para cada gene de maneira independente, ao invés de aplicar um único algoritmo genético global para determinar a rede como normalmente é feito na literatura. Além disso, é proposta a aplicação de um método de inferência de redes para gerar as populações iniciais de modo a servirem como pontos de partida mais promissores para os algoritmos genéticos do que populações aleatórias como é feito normalmente. Para orientar os algoritmos genéticos, propõe-se o uso do critério de informação Akaike (AIC) como função de aptidão. Resultados obtidos da inferência de redes Booleanas artificiais simuladas mostram que o AIC é muito bem correlacionado com métricas de similaridades topológicas usuais mesmo em casos onde há um número pequeno de amostras. Além disso, o benefício de aplicar um algoritmo genético por gene partindo de populações iniciais definidas por uma técnica de inferência de redes é evidente de acordo com os resultados.

Abstract

Gene regulatory networks (GRN) inference from gene expression data is an important problem in systems biology field, in which the main goal is to comprehend the global molecular mechanisms underlying diseases for the development of medical treatments and drugs. This problem involves the estimation of the gene dependencies and the logic functions governing these interactions to provide a model that explains the dataset (usually obtained from gene expression data) on which the estimation relies. In this work, it is proposed a method based on genetic algorithms to infer gene networks, whose main idea consists in applying one genetic algorithm for each gene independently, instead of applying a unique genetic algorithm to determine the network as usually done in the literature. Besides, it is proposed the application of a network inference method to generate the initial populations to serve as more promising starting points for the genetic algorithms than random populations as usually done. To guide the genetic algorithms, we propose the use of Akaike information criterion (AIC) as fitness function. Results obtained from inference of artificial Boolean networks show that AIC correlates very well with popular topological similarity metrics even in cases with small number of samples. Besides, the benefit of applying one genetic algorithm per gene starting from initial populations defined by a network inference technique is evident according to the results.

Sumário

Lista de abreviaturas e termos

SAGE	Análise Serial de Expressão Gênica (*Serial Analysis of Gene Expression*)
GRN	Rede Gênica Regulatória (*Gene Regulatory Networks*)
BN	Rede Booleana (*Boolean Network*)
PBN	Rede Booleana Probabilística (*Probabilistic Boolean Network*)
PGN	Redes Gênicas Probabilísticas (*Probabilistic Gene Network*)
BEIM	Busca Exaustiva por Informação Mútua
ER	Modelo de redes aleatórias de Erdös e Rényi (Erdös-Rényi)
BA	Modelo de redes livres de escala (*scale-free*) de Barabási e Albert (Barabási-Albert)
AG	Algoritmos Genéticos
AIC	Critério de informação Akaike (*Akaike Information Criterion*)
PA	População inicial obtida Aleatoriamente (População Aleatória)
PB	População inicial obtida através de Busca Exaustiva por Informação Mútua (População Busca)
ERBN	Redes do tipo Erdös-Rényi (ER) e Booleana (BN)
ERPBN	Redes do tipo Erdös-Rényi (ER) e Booleana probabilística (PBN)
BABN	Redes do tipo Barabási-Albert (BA) e Booleana (BN)
BAPBN	Redes do tipo Barabási-Albert (BA) e Booleana probabilística (PBN)
PPV	Valor preditivo positivo (*Positive Predictive Value*)
SIM	Similaridade topológica (*similarity*)
TP	Número de verdadeiros positivos (*True Positives*)
FP	Número de falsos positivos (*False Positives*)
TN	Número de verdadeiros negativos (*True Negatives*)
FN	Número de falsos negativos (*False Negatives*)

Lista de símbolos

t	Instante de tempo
$\psi(\cdot)$	Classificador
$\mathbf{X}$	Vetor de características
$\mathbf{x}$	Padrão ou instância observada de $\mathbf{X}$
$H(\cdot)$	Entropia
$P(\cdot)$	Probabilidade
Y	Variável aleatória correspondente aos rótulos das classes
y	Rótulo da classe; valor da variável aleatória Y
c	Número de classes (rótulos)
$I(\cdot)$	Informação mútua
B	Uma rede booleana
n	Número de genes
V	Conjunto de nós (genes) de uma rede (grafo)
$\mathbf{F}$	Vetor de funções booleanas
i, j	Índices
G_i	i-ésimo gene
g_i	Valor do gene G_i
$\mathbf{g}(t)$	Estado da rede de genes no instante de tempo t
k	Grau de um nó
$\langle k \rangle$	Grau médio dos nós
γ	Constante de decaimento para redes de livres de escala (*scale-free*)
$decimal(\cdot)$	Número decimal de um número binário
P_i	Conjunto de preditores do gene i
C	Número de cromossomos de uma população em um algoritmo genético
$AIC(\cdot)$	AIC (*Akaike Information Criterion*) de um gene alvo e um conjunto de preditores
$K(\cdot)$	Fator de penalização K do AIC
$L(\cdot)$	Função de máxima verossimilhança de um gene alvo e um conjunto de preditores
M	Matriz de expressão gênica (conjunto de amostras)

max_{mut}	Número máximo de cromossomos mutados
it	Iteração (geração) atual do algoritmo genético
r	Número de repetições do melhor AIC para atingir o critério de convergência
Bin	Distribuição binomial

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Contextualização

A biologia sistêmica é um campo de pesquisa interdisciplinar que tem como foco principal de estudo as interações complexas que existem em organismos vivos [Snoep and Westerhoff, 2005]. Tais pesquisas estudam os processos que ocorrem nos sistemas biológicos, tais como os ciclos celulares e as condições para a origem de determinadas doenças. As pesquisas nessa área vem recebendo forte atenção de diversos pesquisadores do mundo todo com o objetivo de entender o genoma humano e o funcionamento dos processos celulares como um todo para auxiliar no desenvolvimento de novos tratamentos e medicamentos contra doenças, técnicas de produção de bioenergia, dentre outras aplicações.

O genoma de um organismo desempenha um papel central no controle de processos celulares, tais como a resposta de uma célula a sinais ambientais, a diferenciação de células em seus respectivos grupos funcionais, e a replicação do DNA para divisão celular. Proteínas sintetizadas a partir de genes podem funcionar como fatores de transcrição de ligação a sítios reguladores de outros genes, como enzimas que catalisam reações metabólicas, ou como componentes de vias de transdução de sinal. Com poucas exceções, as células de um organismo contêm o mesmo material genético. Isto implica que, com a finalidade de compreender como os genes estão envolvidos no controle de processos intra e intercelulares, o âmbito deve ser ampliado a partir de sequências de nucleotídeos que codificam para proteínas aos sistemas reguladores que determinam quais genes são expressos, quando, onde, e em que medida. Podemos entender um organismo vivo como uma complexa rede de moléculas conectadas por reações químicas. Explicar a forma com que esta rede se autorregula, por meio de envio e recepção de sinais, é atualmente um dos principais focos de pesquisa em biologia sistêmica [Snoep and Westerhoff, 2005].

Uma maneira de entender melhor esses mecanismos de controle regulatório é con-

siderar 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], e mais recentemente o RNA-Seq [Wang et al., 2009], têm possibilitado estimar o nível de expressão de milhares de genes simultaneamente e em múltiplos instantes de tempo.

As vias metabólicas das células são reguladas pela interação dos genes, os quais formam redes complexas de comunicação entre eles. Os genes se comunicam através da transcrição de segmentos de DNA na forma de RNA mensageiro (mRNA). Os mRNAs são transportados através dos orifícios do núcleo celular para o citoplasma, onde são traduzidos em sequências de aminoácidos que constituem as proteínas. As proteínas atuam em grande parte dos processos celulares, sendo que muitas delas catalisam reações metabólicas (enzimas) ou voltam ao núcleo para interagir com o DNA e regular a síntese de mRNA [Crick, 1970, D'haeseleer et al., 1999].

Os organismos biológicos dependem do funcionamento das diversas vias metabólicas que são reguladas por redes de expressão gênica. Uns dos maiores desafios existentes em bioinformática é analisar dados de expressão gênica, já que normalmente o número de amostras experimentais disponíveis é muito pequeno (da ordem de dezenas), ao mesmo tempo em que apresentam uma dimensionalidade muito grande (da ordem de milhares de genes). Por isso, é necessário desenvolver técnicas computacionais e estatísticas que reduzam o erro de estimação cometido na presença de um pequeno número de amostras com grande dimensionalidade. Inferir o relacionamento entre os genes com o objetivo de obter redes de regulação gênica (do inglês: *Gene Regulatory Networks* - GRN) é um problema em aberto [Shmulevich and Dougherty, 2007, Kelemen et al., 2008, Hecker et al., 2009, Marbach et al., 2010].

Atualmente, as pesquisas envolvendo inferência de GRNs estão em crescente evidência devido ao massivo volume de dados de expressão gênica que vem sendo produzido para as mais diversas espécies de organismos e condições específicas. Outros fatores que tornam essa tarefa desafiadora estão associados à dificuldade de obter informações precisas sobre as expressões dos genes, ocasionando um ruído experimental significativo nas medidas de expressão. Além disso, a alta complexidade dessas redes de inter-relacionamento aliada à falta de conhecimento *a priori* sobre o organismo biológico de interesse tornam essas pesquisas ainda mais desafiadoras [Angeletti et al., 2001].

O problema da inferência de redes gênicas é mal posto (*ill-posed*), já que para um mesmo conjunto de dados de expressão, podem existir diversas (ou mesmo infinitas) redes que geram esses dados. Esse problema é agravado pelo número limitado de amostras e a presença de ruído. Por isso, muitas metodologias têm sido propostas para auxiliar o processo de inferência, as quais são baseadas em diversas áreas, tais como reconhecimento

de padrões, otimização combinatória, inteligência artificial, otimização combinatória, teoria da informação, teoria de controle, inferência estatística, sistemas dinâmicos, redes complexas, dentre outras.

Em particular, os algoritmos genéticos constituem um importante arcabouço para otimização combinatória que tem sido empregado em diversas situações. Muitos processos biológicos possuem um comportamento otimizado e uma grande capacidade de processamento de informação, confirmado pelo processo de evolução que as espécies tiveram ao longo de milênios e pela capacidade do cérebro humano em armazenar informações de diversas formas e inferir raciocínios lógicos. Um exemplo de algoritmo bioinspirado foi desenvolvido por J. H. Holland. Ele baseou-se no conceito de evolução das espécies segundo Charles Darwin, que afirmou que as espécies evoluem através da competição entre indivíduos pela sobrevivência e reprodução, e os indivíduos mais aptos a sobreviver terão maiores probabilidades de transferir seu material genético para as futuras gerações. Na reprodução sexuada, ocorre o cruzamento (*crossover*), que é a combinação de partes do material genético dos pais. O resultado da mistura do material genético resulta em um novo individuo com características de ambos, e seu material genético tem possibilidade de ser melhor do que de seus antecessores em virtude da pressão seletiva [Holland, 1992].

Os algoritmos genéticos procuram uma solução favorável para problemas que apresentam inúmeras possíveis soluções de maneira inspirada pela seleção natural. Geralmente consistem de quatro etapas principais: criação da população inicial, seleção, cruzamento e mutação [Haupt and Haupt, 2004]. Durante a criação, uma serie de potenciais soluções são geradas, dando origem a uma população inicial de soluções. Durante a seleção, as melhores soluções são escolhidas de acordo com uma função de aptidão ou *fitness*. A partir daí, a nova população é criada da combinação das melhores soluções da população anterior por meio de cruzamento. Finalmente, durante a mutação, uma parcela da população é alterada aleatoriamente de algum modo. Esse processo se repete por um certo número de iterações. Cada repetição é conhecida como "geração". Quando as estratégias adequadas são empregadas, o segmento da população com o melhor *fitness* vai crescendo ao longo de várias gerações. O maior *fitness* na população também deve crescer de modo a obter uma solução satisfatória [Goldberg, 1989, Haupt and Haupt, 2004].

Os algoritmos genéticos normalmente requerem personalização para produzir resultados melhores mais rapidamente para um problema específico. Cada passo de um algoritmo genético oferece seus próprios meios de personalização. Durante a criação, o tamanho de uma população, os meios de representação de soluções, e os meios de criação de soluções podem variar. Durante a seleção, a porcentagem selecionada, o método de cálculo de aptidão e o método de seleção também podem variar. Finalmente, durante o cruzamento e mutação, os métodos para a manipulação de soluções devem ser especificados [Goldberg, 1989, Haupt and Haupt, 2004].

1.2 Objetivos

Esta dissertação propõe o estudo e desenvolvimento de estratégias baseadas em algoritmos genéticos para inferência de redes gênicas. A estratégia principal proposta para isso foi a aplicação de um algoritmo genético por gene de maneira independente para tentar obter o melhor conjunto de genes preditores para cada gene alvo. Tal estratégia é diferente daquela que os métodos existentes na literatura empregam para o mesmo fim, os quais aplicam um único algoritmo genético para inferir a rede toda.

Outra proposta consiste na geração da população inicial com base em métodos de inferência de GRNs já existentes (por exemplo, o método de redes gênicas probabilísticas proposto em [Barrera et al., 2007]) para obter pontos de partida mais promissores do que aquelas providas por populações iniciais geradas aleatoriamente, como usualmente é feito pelos algoritmos genéticos propostos na literatura. Além disso, propomos a utilização do critério de informação Akaike (*Akaike Information Criterion*), o qual baseia-se na verossimilhança do conjunto de dados dada a rede, além de embutir um fator que penaliza a complexidade do modelo (no caso das redes gênicas, a penalização cresce em relação à dimensionalidade dos subconjuntos de preditores) [Akaike, 1974, Burnham and Anderson, 2002].

Com o objetivo de analisar os resultados do método proposto, experimentos envolvendo redes Booleanas artificiais geradas por modelos de redes complexas, tais como o aleatório de Erdôs-Rényi (ER) [Erdös and Rényi, 1959] e o livre de escala (*scale-free*) de Barabási-Albert (BA) [Barabási and Albert, 1999], foram realizados. Além disso, o método proposto foi comparado com a técnica proposta recentemente por Mendoza *et al*, que também é baseada em algoritmo genético para inferência de redes gênicas [Mendoza et al., 2012].

1.3 Justificativa

O presente trabalho 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:* a técnica desenvolvida aqui é útil na modelagem e identificação de redes de interação gênica, um problema importante no contexto da biologia sistêmica.

Entender o comportamento dinâmico dos genes que atuam como agentes de controle, regulando grande parte dos fenômenos que ocorrem nos organismos é um dos objetivos

mais importantes das pesquisas em biologia sistêmica. Diversas pesquisas interdisciplinares tem sido realizadas nesse sentido dada a complexidade inerente aos processos celulares.

A biologia sistêmica, a qual é considerada um dos últimos estágios das pesquisas em bioinformática, consiste em entender o funcionamento global dos organismos biológicos a partir do entendimento do funcionamento das suas partes e de como elas se integram. Os resultados das pesquisas nessa área têm potencial para impulsionar, por exemplo, as pesquisas envolvendo biocombustíveis, sendo a cana de açúcar um dos principais objetos de estudo. Tais pesquisas são consideradas estratégicas ao país, devido ao ramo do agronegócio que movimenta uma parcela significativa da economia brasileira. O foco do programa BIOEN da FAPESP, por exemplo, tem como objetivo justamente o fomento às pesquisas em fontes de biocombustível. Além disso, o combate às doenças tropicais humanas como a dengue e a malária também é um dos principais interesses da pesquisa nacional. Finalmente, do ponto de vista internacional, é essencial o desenvolvimento e aprimoramento de aplicações médicas, tais como o entendimento do câncer e o tratamento de doenças como a AIDS, doenças neurodegenerativas e do neurodesenvolvimento, dentre outras. Esta dissertação de mestrado oferece uma parcela de contribuição para o avanço da pesquisa nesses setores.

Por se inserir no contexto da biologia sistêmica, este trabalho possui um caráter interdisciplinar, já que as pesquisas nessa área envolvem diversas áreas do conhecimento, como mencionado anteriormente.

1.4 Contribuições

As principais contribuições deste trabalho, além de uma revisão bibliográfica do estado da arte na área, são:

- Desenvolvimento de um método baseado em algoritmos genéticos para inferência de redes gênicas modeladas por redes Booleanas e redes Booleanas probabilísticas a partir de dados de expressão gênica. As novidades do método proposto em relação ao que existe na literatura sobre algoritmos genéticos para inferência de redes gênicas consistem da aplicação de um algoritmo genético por gene de maneira independente; o uso do AIC para orientar o algoritmo genético; e a proposta de gerar as populações iniciais pela busca exaustiva por subconjuntos de grau fixo (Busca Exaustiva por Informação Mútua).

- Desenvolvimento de um software de código fonte aberto em Java disponibilizando todo o processo do algoritmo genético proposto. O código fonte estará disponível

publicamente em breve.

- Análise e validação dos resultados obtidos pela aplicação do método sobre dados gerados por modelos de redes complexas. A avaliação dos resultados foi realizada com base em critérios topológicos, tendo incluido também uma análise comparativa com o algoritmo genético proposto por Mendoza *et al* [Mendoza et al., 2012].

1.5 Organização do texto

O Capítulo 2 apresenta uma revisão bibliográfica sobre modelagem e inferência de redes gênicas, incluindo conceitos sobre reconhecimento de padrões, seleção de características e algoritmos genéticos. O Capítulo 3 descreve o método baseado em algoritmos genéticos proposto para inferência de redes gênicas modeladas como redes Booleanas. O Capítulo 4 apresenta e discute os resultados obtidos pela aplicação do método proposto. Finalmente, o Capítulo 5 encerra o texto com as considerações finais e as perspectivas futuras abertas por este trabalho.

Capítulo 2

Revisão e conceitos básicos

2.1 Sinais de expressão gênica

As células eucariotas definem seus comportamentos específicos segundo os genes que ela expressa. A transcrição é o primeiro passo para converter a informação armazenada como DNA do organismo em proteínas. O processo de transcrição é regulado por uma rede de controle que coordena a atividade celular [Shmulevich and Dougherty, 2007]. Os genes expressam uma determinada concentração de RNAs mensageiros (mRNA), as quais controlam a produção de proteínas. Esse processo é o meio primário de regulação da atividade celular. Em seguida, o mRNA é transportado para fora do núcleo, até os ribossomos presentes no citoplasma. Nos ribossomos, o mRNA é traduzido em sequências de aminoácidos que formam a base para a construção das proteínas. Algumas dessas proteínas são enzimas que catalisam reações metabólicas vitais para a manutenção e a sinalização celular. Outras 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 celulares.

A tecnologia de *microarray* [Shalon et al., 1996] é uma das mais utilizadas para extração de sinais de transcrição genética. Essa técnica consiste em um processo bioquímico que mede os níveis de expressão de milhares de genes simultaneamente. A Figura 2.2 mostra uma imagem de *microarray* do tipo *two-color*. A matriz de cores mostrada serve como dados para uma série de análises, desde classificação e caracterização de fenômenos biológicos até identificação de redes de regulação gênica. Mais recentemente, tecnologias de sequenciamento de nova geração têm possibilitado o desenvolvimento de técnicas mais avançadas de medição de expressão gênica, como por exemplo o RNA-Seq [Wang et al., 2009].

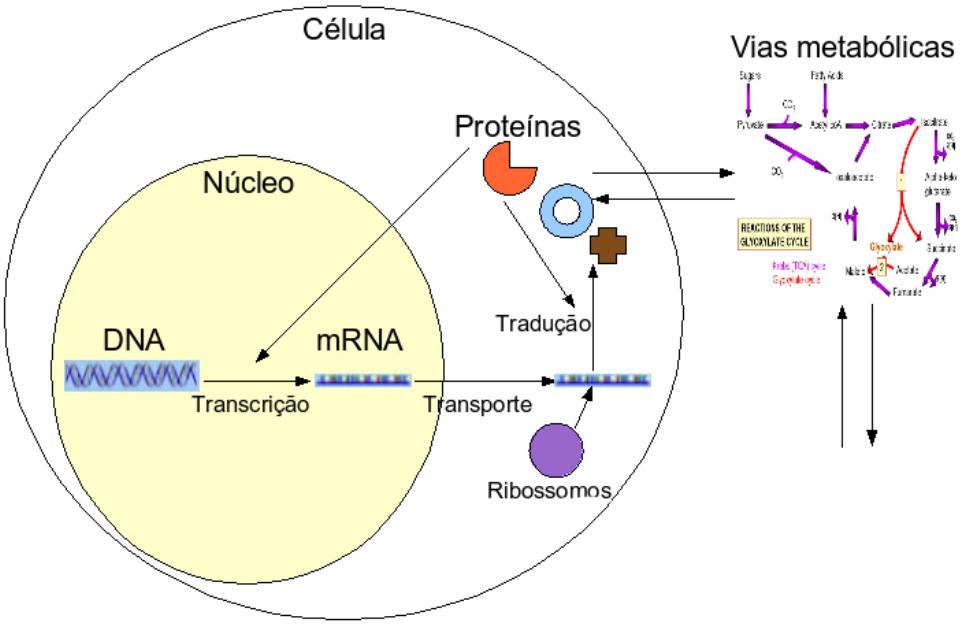


Figura 2.1: Esquema simplificado da dinâmica celular (Fonte: [Martins-Jr., 2008]).

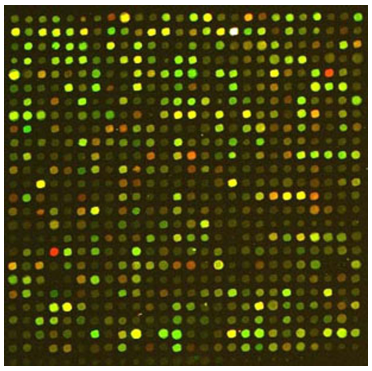


Figura 2.2: Exemplo de imagem de *microarray*.

2.2 Modelagem de redes gênicas

Para modelar as redes complexas de interações gênicas, existem duas abordagens genéricas [Shmulevich and Dougherty, 2007]. Uma dessas abordagens consiste em considerar os valores de expressão gênica no domínio continuo (números reais). Nesse domínio, as interações entre os genes podem ser modeladas por meio de equações diferenciais. As equações diferenciais permitem projetar modelos quantitativos detalhados das redes bioquímicas com funções celulares de interesse [Jong, 2002]. A outra abordagem considera os valores de expressão gênica no domínio discreto. Exemplos de modelos desse tipo são: redes Booleanas [Kauffman, 1969], redes Booleanas probabilísticas [Shmulevich et al., 2002] e redes gênicas probabilísticas [Barrera et al., 2007].

A abordagem continua oferece um entendimento quantitativo detalhado do sistema

em questão, mas em geral necessita de um conjunto considerável de amostras experimentais, além de informações sobre determinadas características das reações bioquímicas [Karlebach and Shamir, 2008], o que a torna apropriada apenas em situações muito específicas. Por outro lado, as abordagens discretas podem ser facilmente modeladas no computador, além de possibilitar um entendimento qualitativo global dos sistemas biológicos. Essas abordagens têm sido empregadas com sucesso na modelagem e simulação de redes e processos biológicos de diversas espécies, 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], *Saccharomyces cerevisiae* [Li and Lu, 2005], ciclo celular de mamíferos [Faure et al., 2006], *Plasmodium falciparum* [Barrera et al., 2007], dentre outros.

As redes Bayesianas são amplamente utilizadas para representar redes gênicas [Friedman et al., 2000, Kelemen et al., 2008], constituindo um modelo probabilístico capaz de representar uma rede gênica causal, sendo relativamente robustas a ruído e sensíveis a relações multivariadas não-lineares. Tal modelo utiliza distribuição 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.

No contexto dos modelos discretos, as redes Booleanas (do inglês: *Boolean Networks* - BNs) são 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 das medidas de expressão [Kauffman, 1969]. Elas consistem em um modelo discreto de Redes Bayesianas no qual as variáveis assumem apenas dois valores possíveis (0 - subexpresso, e 1 - superexpresso). As BNs são muito úteis em diversos casos, mas possuem a limitação de serem um modelo determinístico, que faz a suposição de um ambiente sem incerteza. É necessário, porém, levar em conta que a célula é um sistema aberto propenso a receber estímulos externos. Dependendo das condições externas em um dado instante de tempo, a célula pode alterar sua dinâmica [Shmulevich and Dougherty, 2007].

Para fazer com que as BNs tenham um caráter estocástico, existe o modelo de redes Booleanas probabilisticas (*Probabilistic Boolean Networks - PBNs*), que além de considerar genes com valores binários, associa a cada um deles um conjunto de funções Booleanas preditoras, atribuindo uma probabilidade específica a cada função [Shmulevich et al., 2002]. Essa abordagem também tem desvantagens importantes, a principal delas referente à perda de informação decorrente da discretização dos dados. Só que isso faz os mode-

los Booleanos mais simples e mais fáceis de serem tratados e modelados computacionalmente [Styczynski and Stephanopoulos, 2005]. Uma discussão a respeito disso pode ser vista em [Ivanov and Dougherty, 2006]. Este trabalho se concentra nos modelos de redes Booleanas e de redes Booleanas probabilísticas.

2.3 Reconhecimento de padrões

O reconhecimento de padrões é uma área na qual o objetivo é a classificação de objetos de interesse em um número de categorias ou classes [Theodoridis and Koutroumbas, 1999]. Em bioinformática há uma grande variedade de problemas que envolve o reconhecimento de padrões. Por exemplo, 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]. Nesse caso, os sinais de expressão são usados para projetar um classificador ψ que receba como entrada os níveis de expressão gênica sendo representadas por um vetor de características $\mathbf{X} = (x_1, x_2, ..., x_n)$ (sendo x_i o nível de expressão do gene X_i) e devolva um rótulo ou classe $Y = \{0, 1, ..., c-1\}$ à qual o vetor considerado pertence.

Os 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 e/ou observado em diferentes instantes de tempo (*e.g.* diferentes estágios do ciclo celular de uma determinada espécie [Li et al., 2004, Zhang et al., 2006, Barrera et al., 2007]).

2.3.1 Seleção de características e o problema da dimensionalidade

O problema de seleção de características consiste em selecionar um subconjunto de características que represente adequadamente os objetos em estudo. Os métodos de seleção de características consistem basicamente de duas partes principais: um algoritmo de otimização e uma função critério a ser otimizada. Em análise de expressão gênica, as características são os genes, cujos valores são dados pela expressão gênica. 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 dados de expressão gênica, o número de características (genes) usualmente é da ordem de milhares. Uma das razões para reduzir o numero de características (redução de dimensionalidade) é a complexidade computacional. Um importante passo para o projeto de um sistema de classificação é a avaliação do desempenho de um classifi-

cador, 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 2.3) [Jain and Zongker, 1997]. 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.

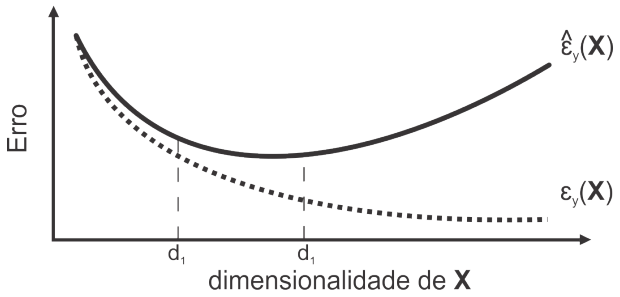


Figura 2.3: Gráfico das taxas de erro em função da dimensionalidade com número fixo de amostras ilustrando o problema da dimensionalidade (Fonte: [Martins-Jr., 2008]).

O número de amostras necessárias para que um classificador tenha um desempenho satisfatório é exponencial com relação à dimensão do vetor de características [Jain and Zongker, 1997]. 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 proibitivo face ao tamanho do conjunto de amostras de treinamento.

2.3.2 Algoritmos de busca

Algoritmos de seleção de características percorrem parte do conjunto de todas os possíveis vetores de características em busca de um vetor 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]. Sendo assim, diversos algoritmos têm sido propostos na literatura. A Figura 2.4 apresenta a taxonomia dos principais métodos de seleção de características utilizados em reconhecimento de padrões. Tais algoritmos são categorizados de 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), e única solução *versus* várias soluções.

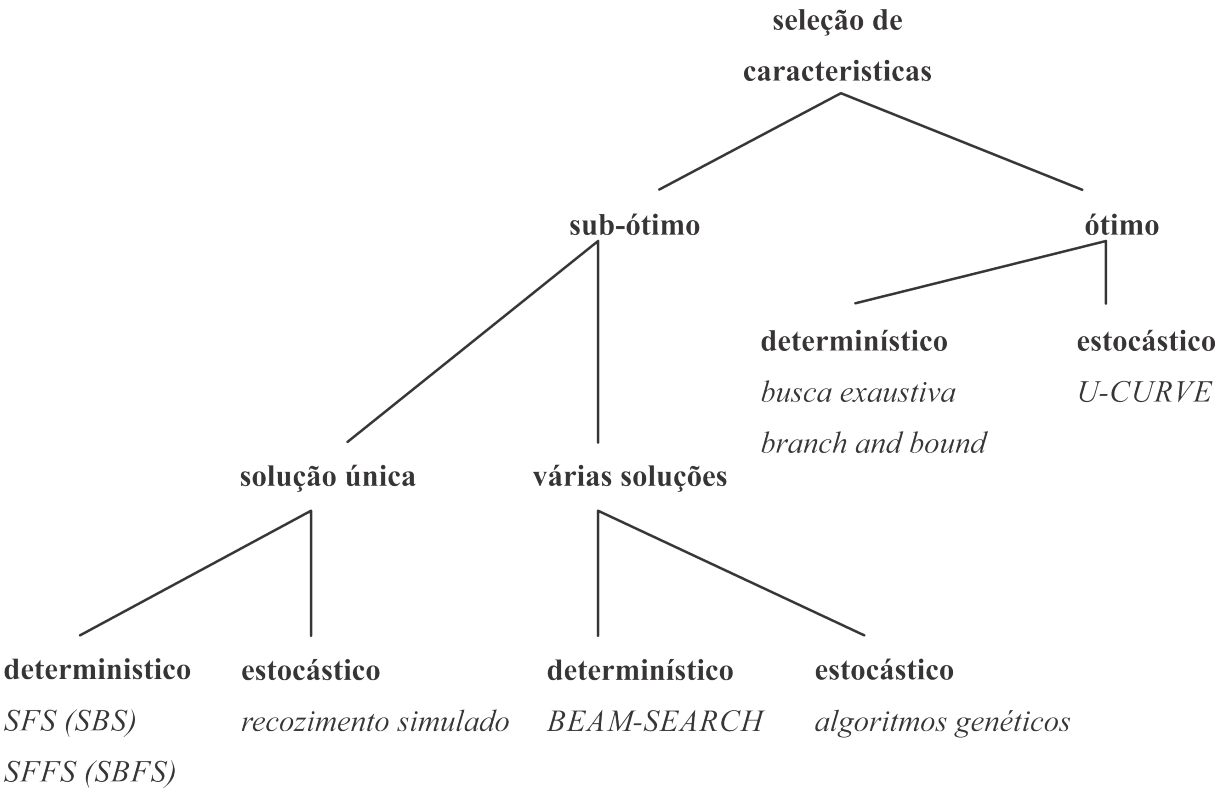


Figura 2.4: Categorização dos algoritmos de seleção de características comumente empregados em reconhecimento de padrões (Fonte: [Reis, 2012]).

Os algoritmos genéticos são categorizados como métodos sub-ótimos, estocásticos de múltiplas soluções, sendo o foco deste trabalho. Na Seção 2.7 são discutidos os conceitos básicos de um algoritmo genético.

2.3.3 Informação mútua

A informação mútua, baseada na entropia de Shannon, 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., 2008a, Lopes et al., 2009, Lopes et al., 2010, Martins-Jr et al., 2010, Lopes et al., 2011]. Tal função é utilizada na Busca Exaustiva por Informação Mútua (BEIM), método utilizado no trabalho de Barrera *et al* para inferência de redes gênicas pela abordagem PGN [Barrera et al., 2007]. Nossa proposta consiste em utilizar esse método para a geração das populações iniciais dos algoritmos genéticos (ver Seção 3.3). Além disso, o método BEIM será comparado com o método proposto baseado em algoritmos genéticos (Capítulo 4).

A entropia mede o grau de desordem de uma variável, ou seja, quanto maior a entropia de uma variável, mais caótico é o seu comportamento. A entropia de uma variável Y é

definida por:

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

Similarmente, a entropia condicional de uma variável Y dada uma instância $\mathbf{x} \in \mathbf{X}$ é definida por:

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

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

A entropia condicional média diz o quanto as instâncias do vetor de características $\mathbf{X}$ consegue predizer o comportamento da variável Y em média. Quanto menor a entropia condicional média de Y dado $\mathbf{X}$, melhor será a predição média de Y através de $\mathbf{X}$. Ela é definida como a média ponderada das entropias condicionais de todas as possíveis instâncias $\mathbf{x} \in \mathbf{X}$. Sua equação é dada por:

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

em que $H(Y|\mathbf{x})$ é a entropia condicional dada pela Equação 2.2. E finalmente, a informação mútua entre $\mathbf{X}$ e Y é a diferença entre a entropia de Y (entropia *a priori* de Y) pela entropia condicional de Y dado $\mathbf{X}$:

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

Ou seja, quanto menor o valor da entropia condicional média $H(Y|\mathbf{X})$, maior será o ganho de informação sobre Y através do conjunto de características $\mathbf{X}$.

2.4 Modelo de Redes Booleanas

Uma Rede Booleana (*Boolean Network* - BN) $B = (V, \mathbf{F})$ de n variáveis (genes) é definida por um conjunto de n nós $V = \{g_1, ..., g_n\}, g_i \in \{0, 1\}$, e um vetor de n funções booleanas $\mathbf{F} = (f_1, f_2, ..., f_n), f_i : \{0, 1\}^n \rightarrow \{0, 1\}$. Cada nó g_i representa o estado ou expressão do gene i e cada função f_i é a função preditora de g_i. Os estados de todos os genes em $B = (V, \mathbf{F})$ são sincronamente atualizados em cada passo (instante de tempo t) de acordo com as suas funções preditoras, *i.e.*, $g_i(t+1) = f_i(g_1(t), ..., g_n(t)) = f_i(\mathbf{g}(t))$. Em outras palavras, o próximo estado $\mathbf{g}(t+1)$ é obtido pela aplicação do vetor $\mathbf{F}$ de n funções para

o estado atual $\mathbf{g}(t)$. O vetor $\mathbf{F}$ é a função de transição da rede booleana $B = (V, \mathbf{F})$.

Como cada componente f_i de $\mathbf{F}$ é uma função que pode depender de todas as n variáveis no máximo, o numero de possíveis funções para cada gene i é 2^{2^n}, uma vez que o número de possíveis valores de entrada é de 2^n onde cada entrada conduz a uma das duas saídas possíveis (0 ou 1). Assim, o numero de possíveis funções de transição $\mathbf{F}$ (i.e., o numero de diferentes redes booleanas de tamanho n) é $2^{n(2^n)}$, que é o espaço total de busca do problema de inferência. É claro que, na maioria dos cenários, as topologias das redes são escassas, o que significa que cada variável depende apenas de uma pequena fração de outras variáveis. Os sistemas biológicos, tais como as GRNs, compõem um desses cenários [Husmeier, 2003].

Assim, a principal preocupação em relação a inferência de BNs é encontrar a topologia da rede correta com base nos dados experimentais. Mesmo considerando que cada gene depende de um pequeno número fixo de genes (preditores), *e.g.* 2, o número de possíveis topologias é de $\binom{n}{2}^n$, que ainda é enorme, mesmo para $n \leq 10$. Além disso, embora o número médio de preditores por gene seja pequeno, tal número varia significativamente de gene para gene. De fato, alguns genes podem atuar como *hubs*, possuindo um grande número de preditores [Barabasi et al., 2011]. Essa característica faz com que o problema de inferência de GRNs seja ainda mais desafiador, exigindo métodos muito bem projetados para enfrentá-lo.

Redes Booleanas Probabilísticas

As Redes Booleanas Probabilisticas, as quais incluem as Redes Gênicas Probabilísticas (Seção 2.5) [Barrera et al., 2007], são tipos específicos de Redes Bayesianas, onde cada gene tem um valor de expressão entre 0 e 1 em um determinado instante de tempo. A expressão de um gene é determinada por um conjunto de funções Booleanas que recebem os valores de expressão de determinados conjuntos de genes preditores no instante de tempo anterior. Para cada função, é atribuída uma probabilidade de aplicação [Shmulevich et al., 2002].

Para modelar o quasi-determinismo inerente em GRNs, bastar impor que uma das funções tenha probabilidade bastante próxima de 1, enquanto as funções de menor probabilidade geram perturbações ou mudanças de contexto biológico [Brun et al., 2005] [Dougherty et al., 2007].

2.5 Redes gênicas probabilísticas

O perfil de expressão de um dado gene alvo em uma rede de regulação gênica normalmente é determinado pelo perfil de expressão de um subconjunto de genes chamados preditores. Para encontrar o subconjunto de genes (preditores), pode-se empregar métodos de seleção de características para selecionar o subconjunto com maior conteúdo informativo sobre os valores do gene alvo. Em particular, a abordagem de redes gênicas probabilísficas (*Probabilistic Gene Networks* - PGN) [Barrera et al., 2007, Lopes et al., 2008b, Lopes et al., 2009, Lopes et al., 2014] usa o princípio de seleção de características: para cada gene alvo, realiza-se uma busca pelo subconjunto de preditores que melhor descreve o comportamento do alvo de acordo com e uma função critério que avalia a qualidade da predição com base nos sinais de expressão gênica. 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]. Cada gene alvo em um dado instante de tempo depende apenas dos valores dos seus preditores no tempo anterior, já que a abordagem asume que as amostras temporais seguem uma cadeia de Markov de primeira ordem. 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 transitar no próximo instante de tempo) e condicionalmente independente (ou seja, o valor de um determinado gene é dependente apenas dos valores de seus preditores). Devido ao pequeno número de amostras tipicamente disponíveis em dados reais, essas suposições são simplificações importantes.

O método para inferência de redes gênicas por meio de algoritmos genéticos proposto neste projeto (Capítulo 3) segue os pressupostos da abordagem PGN.

2.6 Topologias de redes complexas

A necessidade de ter medidas e métodos fundamentados em propriedades reais de um sistema fez com que a teoria de redes complexas estenda o formalismo da teoria dos grafos [Costa et al., 2007, Lopes, 2011]. As redes gênicas podem ser caracterizadas por modelos de redes complexas, já que estas provêem diversas topologias com propriedades bem definidas. Deste modo a teoria de redes complexas permite a caracterização, análise e representação dos mais variados sistemas complexos, como por exemplo os 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, Narasimhan et al., 2009] [Barabasi et al., 2011].

O modelo de redes aleatórias proposto por Paul Erdös e Alfréd Rényi

em 1959 [Erdös and Rényi, 1959], foi o primeiro modelo de redes complexas. Desde então, foram propostos outros modelos de redes complexas 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]. Neste trabalho, são apresentados apenas os modelos aleatório e livres de escala, os quais foram usados na geração das redes gabarito simuladas nos experimentos do Capítulo 4.

2.6.1 Redes aleatórias

No modelo de redes aleatórias de Erdös-Rényi (ER) [Erdös and Rényi, 1959] todas as possíveis arestas entre os vertices têm uma probabilidade p de estar na rede. Assim, o grau médio de cada nó é dado por $\langle k \rangle = p \times n$, sendo n o número de vértices do grafo, levando em conta a possibilidade de um nó estar ligado por uma aresta consigo mesmo (auto-ligações ou *self-loops*).

Em um grafo gerado pelo modelo aleatório a distribuição do número de conexões por vértices aproxima-se por uma distribuição de Poisson [Costa et al., 2007, Lopes, 2011]. Este modelo 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 $\langle k \rangle = 3$.

2.6.2 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:

$$P(k) \sim k^{-\gamma} \tag{2.5}$$

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 et al., 2011]. Com relação a 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, sendo que poucos nós são altamente conectados a outros nós da rede, enquanto um grande número de nós possuem poucas conexões [Costa et al., 2007, Lopes, 2011]. 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 nó 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 nó v_j já existente na rede ser conectado ao novo nó v_i é linearmente proporcional ao grau k_j do nó v_j tal que:

$$P(v_i \leftrightarrow v_j) = \frac{k_j}{\sum_u k_u}, \forall v_u \in V \tag{2.6}$$

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". Parte dos experimentos do Capítulo 4 foram gerados a partir de redes BA com $\langle k \rangle = 3$ e $\gamma = 2,5$.

2.7 Algoritmos genéticos

A computação evolutiva é um ramo da inteligência artificial onde são propostos modelos computacionais do processo natural da evolução para resolver problemas de otimização combinatória. Neste contexto, nos algoritmos genéticos são aplicados conceitos da teoria da seleção natural como herança, seleção, cruzamento e mutação para obter iterativamente soluções parciais de um problema combinatório até soluções aproximadamente ótimas [Mitchell, 1996].

O arcabouço dos algoritmos genéticos (AG) foi desenvolvido por John Holand (1975), mas foi popularizado por um aluno dele, David Goldberg, quem fez uso dos algoritmos genéticos para o difícil problema de controle de transmissão em gasodutos [Goldberg, 1989]. Desde então, diversas contribuições foram realizadas na área. Algumas das vantagens dos AGs são:

- Otimização com variáveis discretas ou continuas;
- Trata um numero grande de variáveis;
- São adequados para computadores paralelos;

- Otimizam variáveis com superfícies de custo extremamente complexas (que podem saltar de um mínimo local);
- Fornece uma lista de soluções, e não apenas uma solução;
- Funciona com dados gerados numericamente, dados experimentais, ou funções analíticas.

Estas vantagens são intrigantes e produzem resultados impressionantes se comparados a abordagens tradicionais de otimização [Randy L. Haupt, 2004].

O esquema geral de um algoritmo genético esta descrito na Figura 2.5, consistindo das seguintes etapas principais:

1. Inicialização da população dos cromossomos;
2. Avaliar cada cromossomo da população usando a função de aptidão adotada;
3. Selecionar os indivíduos (pais) que gerarão a nova população,
4. Aplicar o cruzamento dos pais selecionados,
5. Mutar uma pequena quantidade dos novos cromossomos,
6. Apagar os velhos membros da população e manter os novos,
7. Caso não seja satisfeito o critério de parada (convergência ou numero de iterações), voltar para a linha 2, caso contrário finalizar o algoritmo.

Para esclarecer alguns aspectos de cada etapa do algoritmo genético, as próximas seções consideram o problema que consiste em descobrir os valores das variáveis $0 \leq x \leq 10$ e $0 \leq y \leq 10$ que maximizam a função:

$$f(x, y) = x \cos y + 10 \sin x + \frac{y}{2} - 2x \tag{2.7}$$

2.7.1 Codificação cromossômica

A representação cromossomal é fundamental para o algoritmo genético, a qual consiste em uma maneira de traduzir a informação de nosso problema em uma maneira viável de ser tratada pelo computador. Cada pedaço indivisível dessa representação é chamado de gene, por analogia com as parte que compõem um cromossomo biológico. É importante notar que a representação cromossomal é completamente arbitrária, ficando sua definição a cargo do projetista, que deve conhecer o domínio do problema [Linden, 2012]. Essa codificação

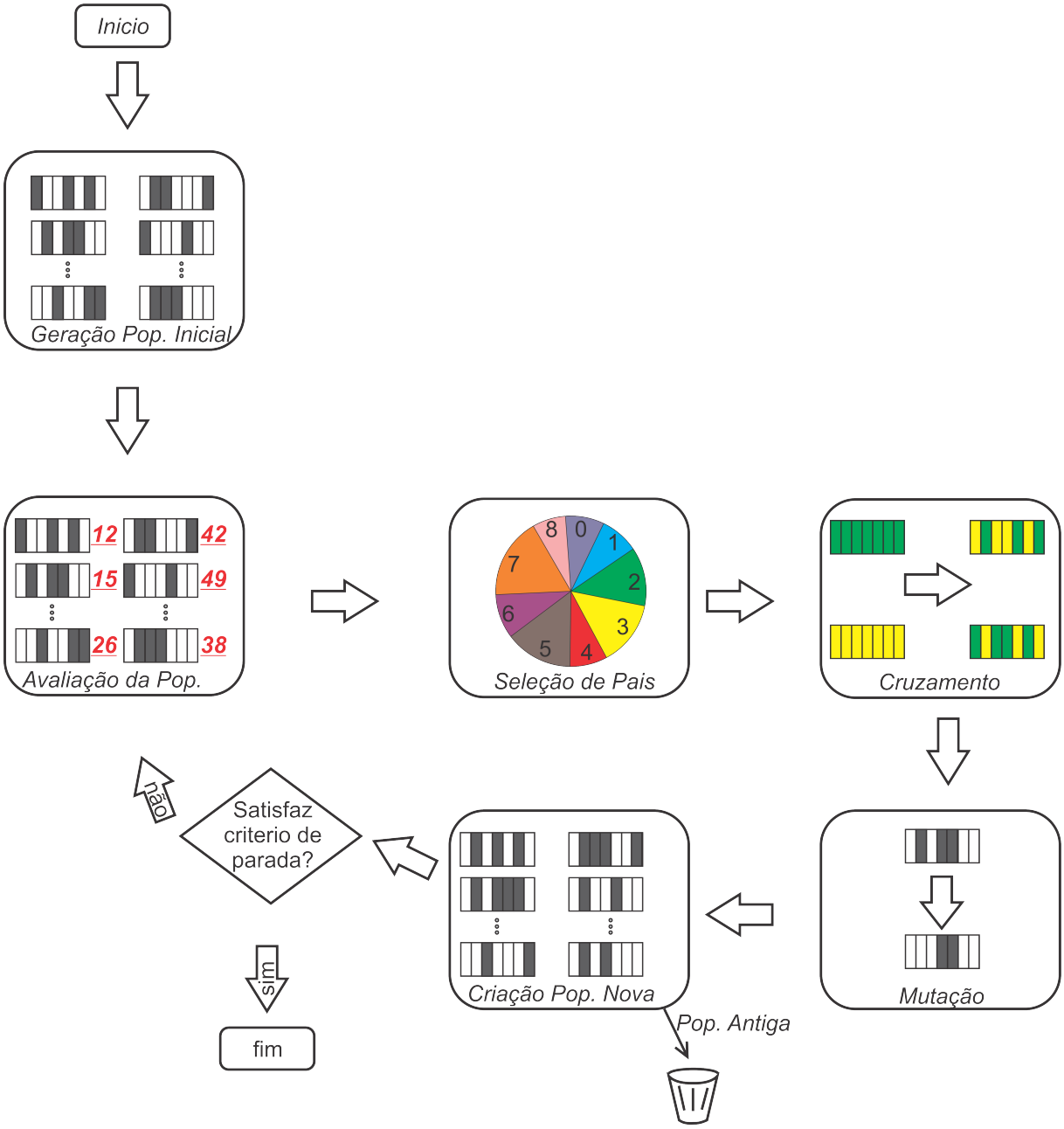


Figura 2.5: Esquema geral de um algoritmo genético, adaptado de [Linden, 2012]

deve permitir avaliar todos os possíveis estados de um cromossomo. A representação mais comumente adotada é a binária, isto é, um cromossomo sendo uma seqüencia de bits, sendo um bit por gene. O que cada bit significa é inerente ao problema[Linden, 2012]. Uma vez que os valores das variáveis são representados em binário, deve haver um meio de conversão de valores inteiros ou contínuos em binário, e vice-versa [Randy L. Haupt, 2004]. Os cromossomos contêm uma ou mais variáveis, sendo cada variável representadada por um determinado subconjunto de bits.

Por exemplo, para representar uma variável continua $0 \leq x \leq 10$ em binário, é preciso

selecionar uma quantidade suficiente de bits. Quanto maior a quantidade de bits, mais precisa será a sua representação numérica no computador. Porém, se tal quantidade for elevada demais, o algoritmo genético consome mais memória e suas operações ficam mais custosas. A Figura 2.6 mostra algumas codificações binárias para o numero x.

4-bits		5-bits		10-bits		20-bits	
deci(x_{bin})	x	deci(x_{bin})	x	deci(x_{bin})	x	deci(x_{bin})	x
0	0	0	0	0	0	0	0
1	0.67	2	0.65	1	0.010	1	0.00001
2	1.33	4	1.29	2	0.020	2	0.00002
3	2.00	6	1.94	3	0.029	3	0.00003
4	2.67	8	2.58	4	0.039	4	0.00004
5	3.33	10	3.23	5	0.049	5	0.00005
6	4.00	12	3.87	101	0.987	103465	0.98672
7	4.67	14	4.52	201	1.965	234161	2.23314
8	5.33	16	5.16	301	2.942	345261	3.29267
9	6.00	18	5.81	401	3.920	451612	4.30691
10	6.67	20	6.45	506	4.946	512345	4.88611
11	7.33	22	7.10	614	6.002	676543	6.45202
12	8.00	24	7.74	756	7.390	718273	6.84999
13	8.67	26	8.39	834	8.152	876541	8.35935
14	9.33	28	9.03	901	8.807	972641	9.27584
15	10	31	10	1023	10	1048575	10

Figura 2.6: Possíveis codificações binárias para uma variável $0 \leq x \leq 10$

Quando precisa-se usar o valor que está representado em binário, é necessário fazer uma decodificação para a variável. Para o problema de maximização da Equação 2.7 e codificação adotada de 20 bits, a forma de obter o valor da variável em binário é:

$x = \frac{10}{2^{20}-1} \times decimal(x_{bin})$

Para representar os dois números x, y é necessário um cromossomo de 40 bits. Por exemplo, seja 0100100001000100000100100001000001000100 o valor do cromossomo. Assim, a decodificação dos valores de x, y é dada por:

$x = \frac{10}{2^{20}-1} \times decimal(01001000010001000001) = 10/1048575 \times 296001 = 2.82289$

e

$y = \frac{10}{2^{20}-1} \times decimal(00100001000001000100) = 10/1048575 \times 135236 = 1.28971$

2.7.2 População

A população de um algoritmo genético é um conjunto de cromossomos que geralmente são inicializados aleatoriamente, e sobre a qual o AG vai tentando aperfeiçoar os resultados obtidos em cada geração. A forma de como se passa de uma população a outra é variável. Às vezes toda a população da geração anterior é substituída por uma população nova.

Pode-se também manter uma proporção de cromossomos da geração anterior (pais) que possuem uma boa avaliação na população nova (elitismo).

2.7.3 Função de aptidão

A função de avaliação é a maneira utilizada pelos AGs para determinar a qualidade de um indivíduo como solução do problema em questão [Linden, 2012].A função de aptidão gera um resultado de saída para um conjunto de variáveis de entrada (cromossomo). Tal função pode ser matemática, um experimento, ou um jogo [Randy L. Haupt, 2004]. Esta função pode ser discreta ou continua, dependendo do objeto real, matemático ou experimental que se tente modelar. Na maioria das situações essa função $f_1(x)$ é usada para achar um valor máximo. Caso o objetivo seja minimizar, basta criar outra função $f_2(x) = 1/f_1(x)$, a qual pode ser maximizada. As avaliações da função de aptidão servem para o processo de seleção descrito em 2.7.4. No caso do problema de maximizar a função dada na Equação 2.7, a função de aptidão é a própria equação usando como entrada as decodificações (os valores reais de x, y) de cada cromossomo a ser avaliado.

2.7.4 Cruzamento

Esta é uma das operações fundamentais de um algoritmo genético, que consiste em simular o mecanismo de seleção natural que atua sobre os seres vivos, onde os pais mais aptos têm um chance maior de gerar descendentes [Linden, 2012]. O cruzamento consiste em duas partes: a seleção e a recombinação dos cromossomos.

Seleção

Sabendo qual é o tamanho da população de filhos a serem criados, seleciona-se a mesma quantidade de pais para o cruzamento, já que para cada par de pais, gera-se um par de filhos. A seleção dos cromossomos (pais) que vão produzir filhos na próxima geração pode ser feita de diversas maneiras. Para refletir o comportamento natural dos seres vivos, a seleção deve privilegiar indivíduos mais aptos de acordo com a função de aptidão.

O método da roleta, consiste em atribuir probabilidades de seleção proporcionais à aptidão do indivíduo, ou seja, quanto maior a aptidão de um indivíduo, maior a chance dele ser selecionado para participar da etapa de recombinação. Uma desvantagem desse método é que cromossomos com avaliações muito baixas dificilmente são escolhidos, sendo que alguns deles poderiam produzir filhos com melhores avaliações.

A vantagem do método da roleta é que ela pode ser calibrada para atribuir uma chance um pouco maior aos indíviduos menos aptos de criar filhos. Por exemplo, a Figura 2.7

ilustra duas roletas. Uma delas adota o próprio valor de $f(x, y)$ diretamente para atribuir a proporção que o cromossomo (x, y) irá obter na roleta. Já a outra roleta aplica o logaritmo ln sobre a função para obter a proporção. No segundo caso, a distribuição de probabilidades acaba sendo um pouco mais próxima da uniforme do que no primeiro caso, fazendo com que a diferença entre as probabilidades de seleção dos indivíduos mais aptos para os menos aptos diminua. A Tabela 2.1 apresenta esses dois tipos de roletas aplicados para 6 valores de (x, y) para o problema da maximização da Equação 2.7.

Tabela 2.1: Valores de $f(x, y)$ e $\ln(f(x, y) + 1)$ para uma população de 6 cromossomos do problema de maximização da Equação 2.7 para obtenção de dois tipos de roleta.

Cromossomo	x	y	$f(x, y)$	$\ln(f(x, y) + 1)$	Roleta 1	Roleta 2
00000110...01010001	1.18	0.07	8.09	2.09	24.79%	22.24%
00000110...01010001	1.18	7.45	11.07	2.40	33.92%	25.10%
10110101...00111101	8.36	6.93	2.19	0.78	6.71%	11.69%
00000110...10111000	1.17	3.93	8.02	2.08	24.58%	22.17%
00000001...01101101	0.22	0.85	2.32	0.84	7.11%	12.09%
00000000...01001100	0.50	2.01	0.94	0.66	2.89%	6.70%

Recombinação

A recombinação consiste na criação de descendentes a partir dos pais selecionados através de um processo de emparelhamento [Randy L. Haupt, 2004]. Uma vez que os cromossomos são selecionados, eles criarão os filhos da nova população através da troca de código genético. A idéia básica para criar um novo cromossomo (filho) é misturar os genes de dois pais selecionados. A maneira de como misturar os códigos genéticos pode variar. A Figura 2.8 ilustra três maneiras usuais de misturar os códigos genéticos.

2.7.5 Mutação

Com o objetivo de evitar soluções de mínimos locais, a mutação é uma importante operação de algoritmo genético para mudar algumas características pontuais dos indivíduos. A mutação altera uma certa porcentagem dos bits (genes) de um cromossomo [Randy L. Haupt, 2004]. Essa porcentagem costuma ser bem pequena (usualmente da ordem de 1% ou menos). O processo da mutação é bem simples já que uma vez selecionado o gene a ser mutado, ele muda seu valor de 0 para 1 ou de 1 para 0.

A probabilidade de mutação pode ser mantida ou mudar segundo o número de iterações transcorridas do algoritmo genético. na Figura 2.9. pode-se observar 3 formas de mutação: mutação invariável (constante), mutação como uma função linear decrescente, e mutação

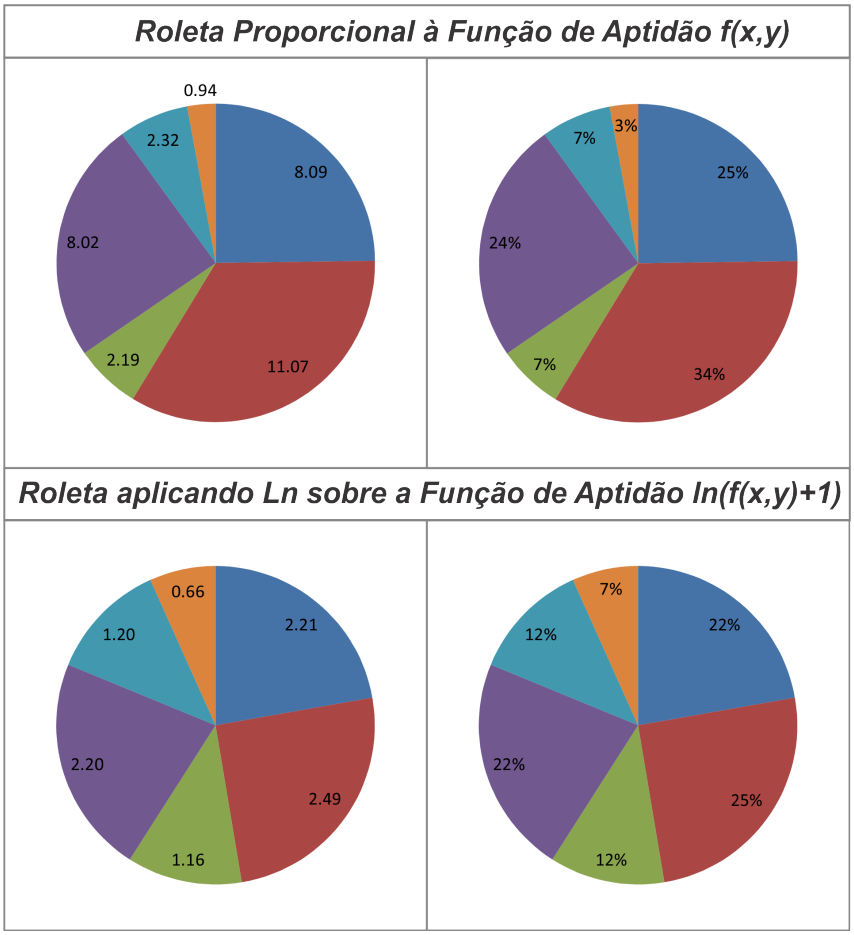


Figura 2.7: Os gráficos de cima representam as proporções da roleta baseadas nos valores da função de aptidão $f(x, y)$, enquanto os gráficos de baixo representam as proporções da roleta baseadas nos valores do logaritmo da função aptidão $\ln(f(x, y) + 1)$. Esta última atribui uma porcentagem um pouco maior aos cromossomos com avaliações baixas.

como uma função senoidal. Todas elas são funções do numero da iteração (geração) do algoritmo genético.

A idéia da mutação linear decrescente é que a mutação seja mais preponderante no início, já que as populações iniciais geradas aleatoriamente dificilmente apresentam resultados próximos do ideal. Após um certo número de gerações, espera-se que as populações se aproximem da solução ideal, então aplica-se uma mutação mais branda justamente para não modificar significativamente essas populações.

Já a idéia que está por trás da mutação senoidal é a de tentar evitar com que as populações convirjam para soluções de máximos locais, as quais podem estar relativamente distantes das soluções ótimas. Assim, a mutação senoidal provê temporadas de forte mutação (crises) intercaladas com temporadas de fraca mutação (calmarias).

A Figura 2.10 mostra a superfície formada pela função dada na Equação 2.7. Note a presença de máximos locais, cujos valores podem estar relativamente distantes do máximo

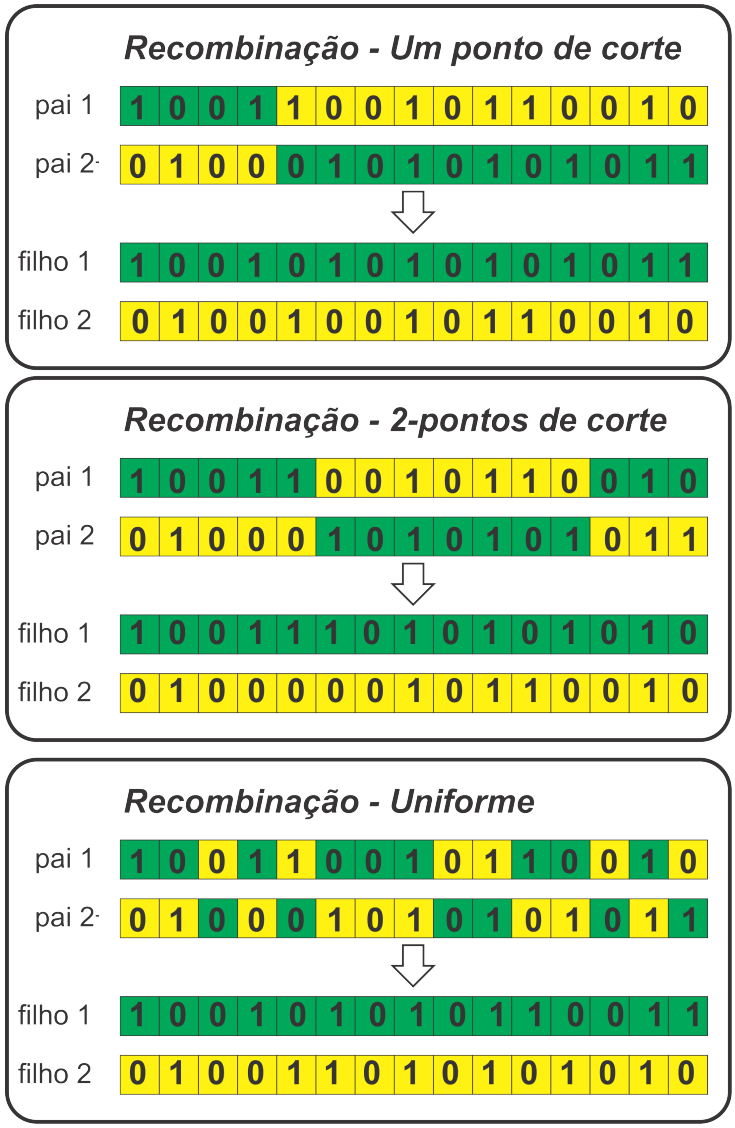


Figura 2.8: Nos três tipos de recombinação, as partes verdes dos cromossomos pais criam o primeiro filho, e as partes amarelas criam o segundo filho. A primeira forma de recombinação sorteia um ponto de corte aleatoriamente, o qual divide cada cromossomo pai em duas partes. A segunda forma sorteia dois pontos de corte aleatoriamente e, de maneira similar, divide os cromossomos pais em três partes. No caso da recombinação uniforme, percorre-se cada gene do primeiro pai e sorteia-se dois valores (0 ou 1). Caso o valor sorteiado seja 0, o gene do primeiro pai passa para o primeiro filho e o gene do segundo pai passa para o segundo filho. Caso contrario, o gene do primeiro pai passa para o segundo filho, enquanto o gene do segundo pai passa para o primeiro filho.

global, o que motiva a aplicação de uma estratégia de mutação apropriada para que o algoritmo genético tente escapar desses máximos locais.

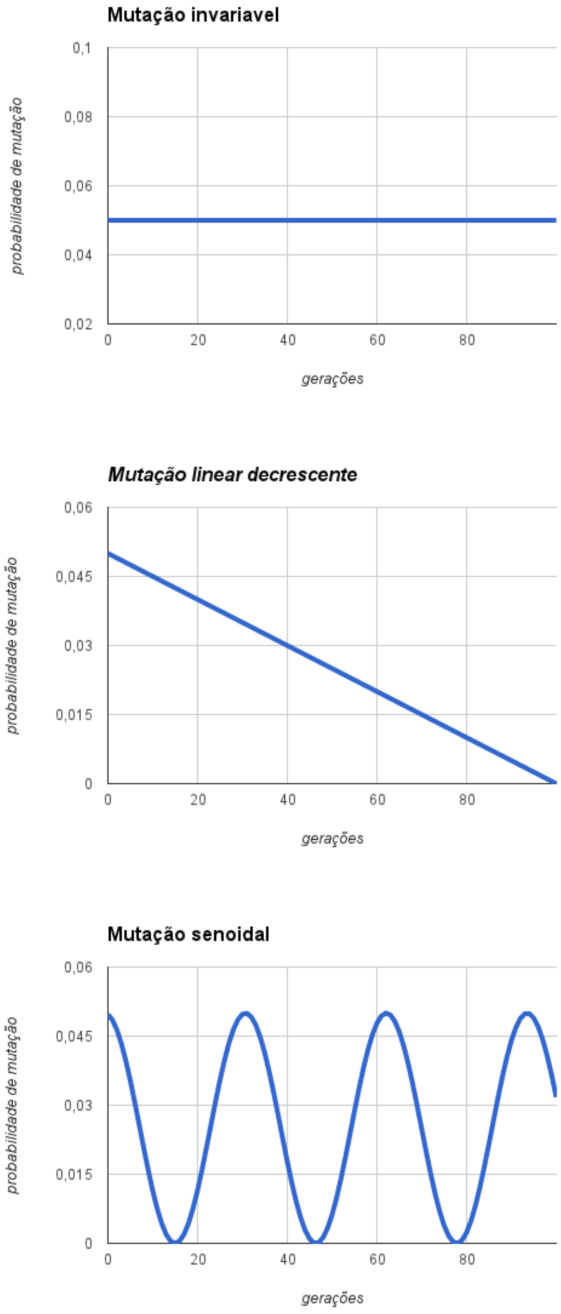


Figura 2.9: Três tipos de mutações.

2.8 Algoritmos genéticos para inferência de redes gênicas

Os algoritmos genéticos vêm sendo bastante empregados para inferir redes de interação gênica. Alguns métodos baseados em algoritmos genéticos para inferência de redes gênicas podem ser encontrados na literatura [Larranaga et al., 1996, Shin and Iba, 2003,

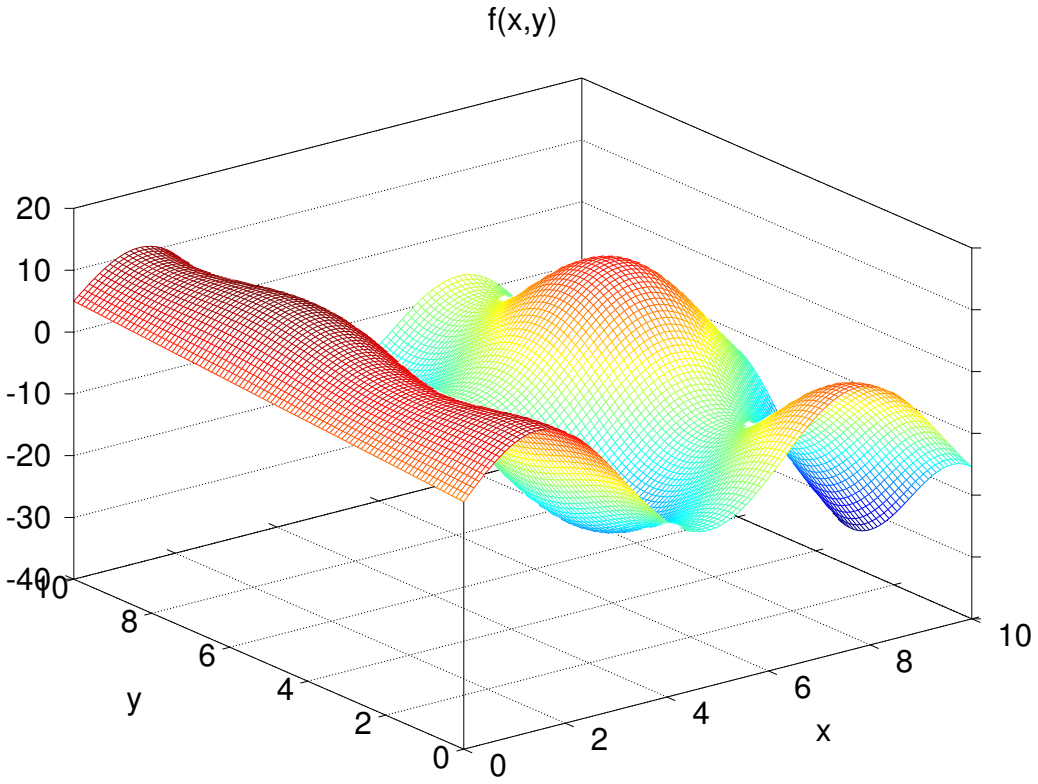


Figura 2.10: Superfície da função dada pela Equação 2.7. Nessa função o máximo global é obtido para o cromossomo correpondendo aos valores $x = 1.457492311$ e $y = 6.633354791$ e $f(x, y) = 11.706616$.

Mamakou et al., 2005, Swain et al., 2005, Knabe et al., 2010, Davidson, 2010] [Mendoza et al., 2012]. Em particular, Davidson *et al* [Davidson, 2010] apresenta uma discussão sobre algoritmos genéticos para inferência de redes Bayesianas através da ordenação topológica dos genes, tendo desenvolvido um algoritmo genético híbrido personalizado para tal fim. Já no trabalho de Mendoza *et al* [Mendoza et al., 2012], propõe-se um algoritmo genético para inferência de redes gênicas modeladas por redes Booleanas a partir de dados experimentais. Nesse trabalho, as topologias são codificadas por inteiros representando os índices dos preditores para cada gene em que é necessário determinar previamente um número máximo de preditores por gene, a população inicial é gerada aleatoriamente com a restrição de um preditor por gene, e a função de aptidão é baseada na entropia de Tsallis com um fator de penalidade que precisa ser apropriadamente determinada para obter redes com um bom balanço entre complexidade (número de arestas) e consistência com o dados.

De maneira geral, os algoritmos genéticos desenvolvidos para inferência de redes gêni-

cas codificam a estrutura da rede como um todo, ou seja, cada cromossomo representa uma rede, sendo que a população inicial consiste de redes geradas aleatoriamente. Já o método desenvolvido neste trabalho, o qual é descrito no próximo capítulo (Capítulo 3), sugere que seja aplicado um algoritmo genético por gene, de modo a buscar o melhor subconjunto de preditores para cada gene. Sendo assim, um cromossomo codifica um subconjunto de genes preditores específico de um determinado gene. Além disso, os algoritmos genéticos partem de populações iniciais pré-inferidas por um método de busca exaustiva por grau fixo utilizando a informação mútua como função critério para orientar essa busca (Busca Exaustiva por Informação Mútua - BEIM). Os resultados experimentais obtidos da aplicação do método proposto são discutidos no Capítulo 4.

Capítulo 3

Método proposto

3.1 Visão geral

Sendo n o número de genes presentes na rede, o problema da busca pela topologia ideal possui um espaço de busca super-exponencial ($2^{n \times n}$, já que cada célula da matriz de adjacência é um valor binário, indicando presença ou ausência de aresta entre dois genes). Uma das vantagens de utilizar o modelo de redes Booleanas (vide Seção 2.4) consiste no fato de ser um tipo de rede Bayesiana na qual cada gene depende unicamente dos preditores (pais) dele. Assim, pode-se dividir o problema em n subproblemas de seleção de características (um subproblema para cada gene), sendo que o espaço de busca de cada subproblema é exponencial (2^n conjuntos de preditores candidatos possíveis). Dessa forma, como a busca exaustiva ainda é impraticável mesmo para um número moderado de genes, algoritmos de aproximação tais como os algoritmos genéticos se fazem necessários.

Neste capítulo é apresentado o método proposto para inferir redes de interação gênica por meio de algoritmos genéticos. O método toma como entrada dados de expressão gênica e devolve uma topologia de rede que tenta descrever de maneira satisfatória os dados de expressão. Uma vez obtida a topologia da rede, as regras das dependências entre os genes podem ser obtidas a partir dos próprios dados de expressão. Neste trabalho foi adotado o modelo de redes Booleanas, seguindo os axiomas das redes gênicas probabilísticas (PGN) como apresentados na Seção 2.5.

O procedimento geral do método proposto é esquematizado na Figura 3.1. A proposta consiste em dividir a inferência da rede gênica em n algoritmos genéticos (uma para cada gene), onde n é o numero de genes da rede. Primeiramente, uma população de redes inicial é inferida através da aplicação de seleção de características por busca exaustiva orientada pela informação mútua como função critério (daqui para frente esse método será denotado por Busca Exaustiva por Informação Mútua - BEIM), seguindo os axiomas do modelo de redes gênicas probabilísticas (PGN) [Barrera et al., 2007] (Seção 3.3). Em seguida, a par-

tir dessas redes, a população inicial de conjuntos de preditores de cada gene é extraída e codificada, servindo como entrada para os algoritmos genéticos (Seção 3.2). Esses algoritmos são guiados pela função de aptidão *Akaike Information Criterion* (AIC), a qual avalia o conjunto de preditores de cada gene com base nos dados de entrada (Seção 3.4). Finalmente, operadores genéticos de cruzamento e mutação são aplicados aos conjuntos de preditores dos genes das redes para produzir as próximas populações (gerações), uma nova população para cada gene (Seções 3.5 e 3.6). Cada geração é submetida a uma nova rodada de avaliação, cruzamento e mutação iterativamente até que um critério de parada seja satisfeito, obtendo assim as populações finais. Em cada geração, realiza-se a comparação do melhor cromossomo (cromossomo com menor AIC) da geração atual com o melhor obtido até o momento. Se o primeiro for melhor, então este passa a ser o cromossomo de referência de comparação para as próximas gerações. Assim o melhor indivíduo (subconjunto de preditores) obtido para cada gene compõe a rede final resultante. As próximas seções detalham os aspectos do método proposto.

3.2 Codificação cromossômica

Os cromossomos são representações das possíveis soluções do problema de busca. Uma solução para o problema de inferência de redes gênicas pode ser representada por um grafo de dependência entre os genes e as lógicas de dependência entre eles. Segundo a abordagem proposta por Mendoza *et al* [Mendoza et al., 2012], as redes gênicas são modeladas como redes de operadores Booleanos, para os quais cada gene é expresso ou não (1 ou 0) baseado na função Booleana que toma como entrada valores de um subconjunto dos genes (que também pode incluir a si mesmo, permitindo auto-laços). As lógicas de dependência não precisam ser codificadas nos cromossomos, sendo suficiente apenas a informação topológica do grafo, que pode ser codificada como uma matriz de adjacências, como um vetor de adjacências ou como conjuntos de preditores. Uma vez definida a topologia, as lógicas de dependência podem ser estimadas a partir dos dados, sendo determinadas pela minimização do erro de predição (maximização da verossimilhança).

Porém, diferentemente do método proposto em [Mendoza et al., 2012], a técnica proposta aqui consiste na aplicação de algoritmos geneticos isolados (um para cada gene). Para isso, é preciso escolher uma representação que contenha o gene alvo e seus respectivos genes preditores. Assim, o conjunto de preditores para cada gene é proposto como uma maneira natural de codificar as dependências entre os genes, Um cromossomo codifica então um conjunto de genes preditores para um determinado gene alvo. Seja P_i o conjunto de preditores do gene i. Se o elemento $j \in P_i$ então o gene i depende do gene j. Caso contrário ($j \notin P_i$), então o gene i não depende do gene j. Como consequência desta representação, não existe restrição sobre o numero de preditores que podem estar

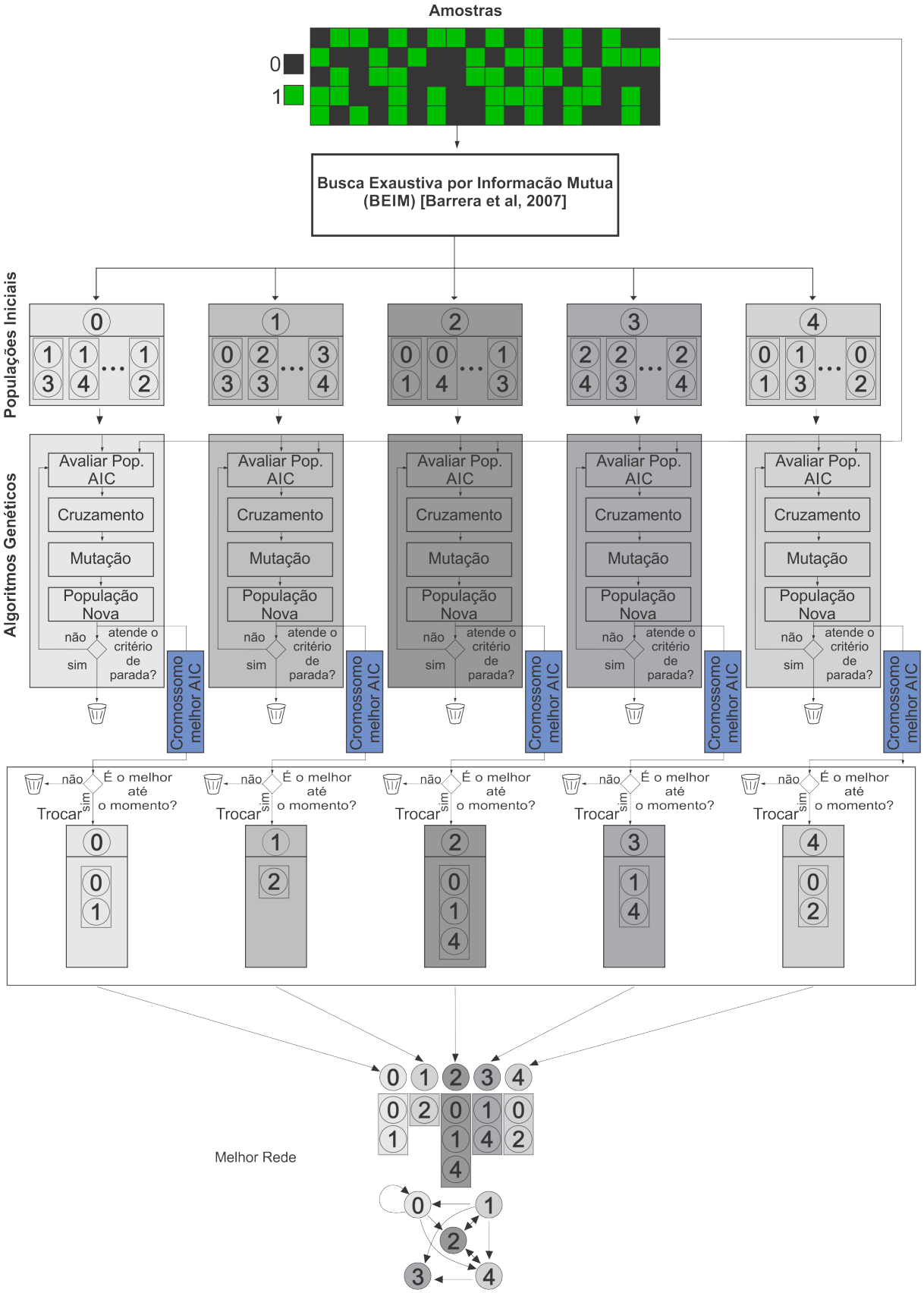


Figura 3.1: Esquema geral do metodo proposto para inferência de redes gênicas.

associados a um determinado gene. A Figura 3.2 ilustra um exemplo de 5 genes com seus respectivos conjuntos de preditores.

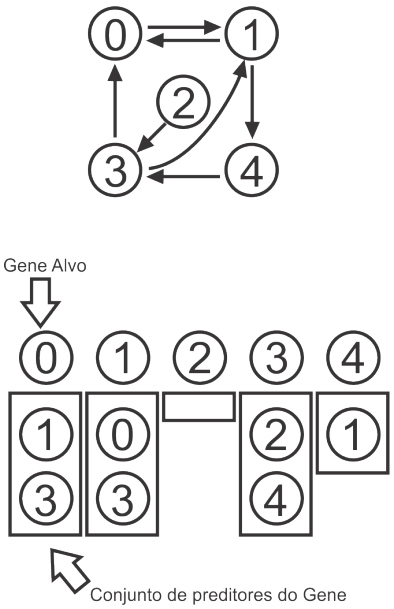


Figura 3.2: Exemplo de uma rede com 5 genes, onde cada gene tem os seus preditores representados por um conjunto (cromossomo).

3.3 Geração da população inicial

Usualmente um algoritmo genético começa partindo de uma população inicial aleatória. No trabalho de Mendoza *et al* [Mendoza et al., 2012], por exemplo, cada gene inicia somente com um preditor escolhido aleatoriamente. Aqui, a proposta consiste em aplicar a abordagem de redes gênicas probabilísticas (PGN) [Barrera et al., 2007], gerando uma população inicial (lista de subconjuntos de preditores) para cada gene. Conforme visto na Seção 2.5, essa abordagem consiste na aplicação de métodos de seleção de características para encontrar os subconjuntos de genes (preditores) que apresentam o maior conteúdo informativo sobre os valores de um determinado gene. Com o objetivo de criar uma população inicial para um determinado gene, gera-se todos os subconjuntos de preditores de tamanho fixo k pré-definido (busca exaustiva por subconjuntos de grau k), os quais são ordenados posteriormente segundo a informação mútua (Equação 2.4) a respeito do gene alvo (esse método será denotado de agora em diante por Busca Exaustiva por Informação Mútua - BEIM). Após a ordenação, os C melhores subconjuntos de preditores compõem a população inicial do gene alvo (cromossomos). Este processo é repetido para todos os genes, criando então as populações iniciais dos cromossomos de todos os algoritmos genéticos (um algoritmo genético por gene) para dar início ao procedimento da inferência da rede gênica.

3.4 Função de aptidão

Uma função de aptidão (*fitness*) é necessária para avaliar as soluções atuais e auxiliar os passos seguintes (cruzamento e mutação) na busca pelas melhores soluções [Linden, 2012]. Para cada solução, é atribuído um valor de aptidão que indica a capacidade da solução sobreviver, reproduzir e manter suas características genéticas nas próximas gerações. Nesse caso cada conjunto de preditores precisa ser avaliado baseado em quão bem ele explica o gene nos dados de expressão gênica em questão. Adotamos o criterio de informação Akaike (do inglês: *Akaike Information Criterion* - AIC) [Akaike, 1974] como medida de avaliação (fitness). O AIC estima a probabilidade dos valores de um gene ser gerado através de um determinado conjunto de preditores com base nos dados, adicionando um fator para penalizar conjuntos de maior dimensionalidade (o número de parâmetros a serem estimados aumenta com o número de preditores do conjunto). O AIC oferece um compromisso entre a complexidade e a aptidão do modelo. Formalmente, o AIC para um gene alvo G_i e o conjunto de preditores $\mathbf{X_i}$ é definido pela Equação 3.1:

$$AIC(G_i, \mathbf{X_i}) = 2K(\mathbf{X_i}) - 2ln(L(G_i, \mathbf{X_i})) \tag{3.1}$$

em que K é o numero de parametros no modelo estatistico, e $L(G_i, \mathbf{X_i})$ é a função de máxima verossimilhança para o gene G_i e o conjunto de preditores $\mathbf{X_i}$. Dadas m amostras e o gene alvo $G_i = \{0, 1\}$ que depende do subconjunto de preditores $\mathbf{X_i} = \{0, 1\}^{k_i}$, onde k_i é o numero de preditores do gene G_i, L pode ser estimado a partir dos dados como segue na Equação 3.2:

$$L(G_i, \mathbf{X_i}) = \prod_{\mathbf{x_i} \in \{0,1\}^{k_i}} \prod_{g_i \in \{0,1\}} P(G_i = g_i | \mathbf{X_i} = \mathbf{x_i})^{mP(\mathbf{X_i} = \mathbf{x_i}, G_i = g_i)} \tag{3.2}$$

em que $P(G_i = g_i | \mathbf{X_i} = \mathbf{x_i})$ é a probabilidade condicional de $G_i = g_i$ dado $\mathbf{X_i} = \mathbf{x_i}$, e $P(\mathbf{X_i} = \mathbf{x_i}, G_i = g_i)$ é a probabilidade conjunta de $\mathbf{X_i} = \mathbf{x_i}$ e $G_i = g_i$. Essas probabilidades são estimadas a partir dos dados.

Além disso, $K(\mathbf{X_i})$ é o número de parâmetros estatísticos a serem estimados para o gene G_i, dado pela Equação 3.3:

$$K(\mathbf{X_i}) = 2^{1+k_i} \tag{3.3}$$

ou seja, dado que um gene G_i depende de k_i preditores binarios, existem 2^{k_i} possíveis valores para esses preditores. Como o gene G_i é binário, deve-se estimar duas probabilidades condicionais para cada um dos possíveis valores que os preditores podem assumir, sendo uma para o caso do gene alvo ser igual a zero e outra para o caso do gene alvo

ser igual a um. Logo, o numero total de parâmetros estatísticos a serem estimados é de $2 \times 2^{k_i} = 2^{1+k_i}$.

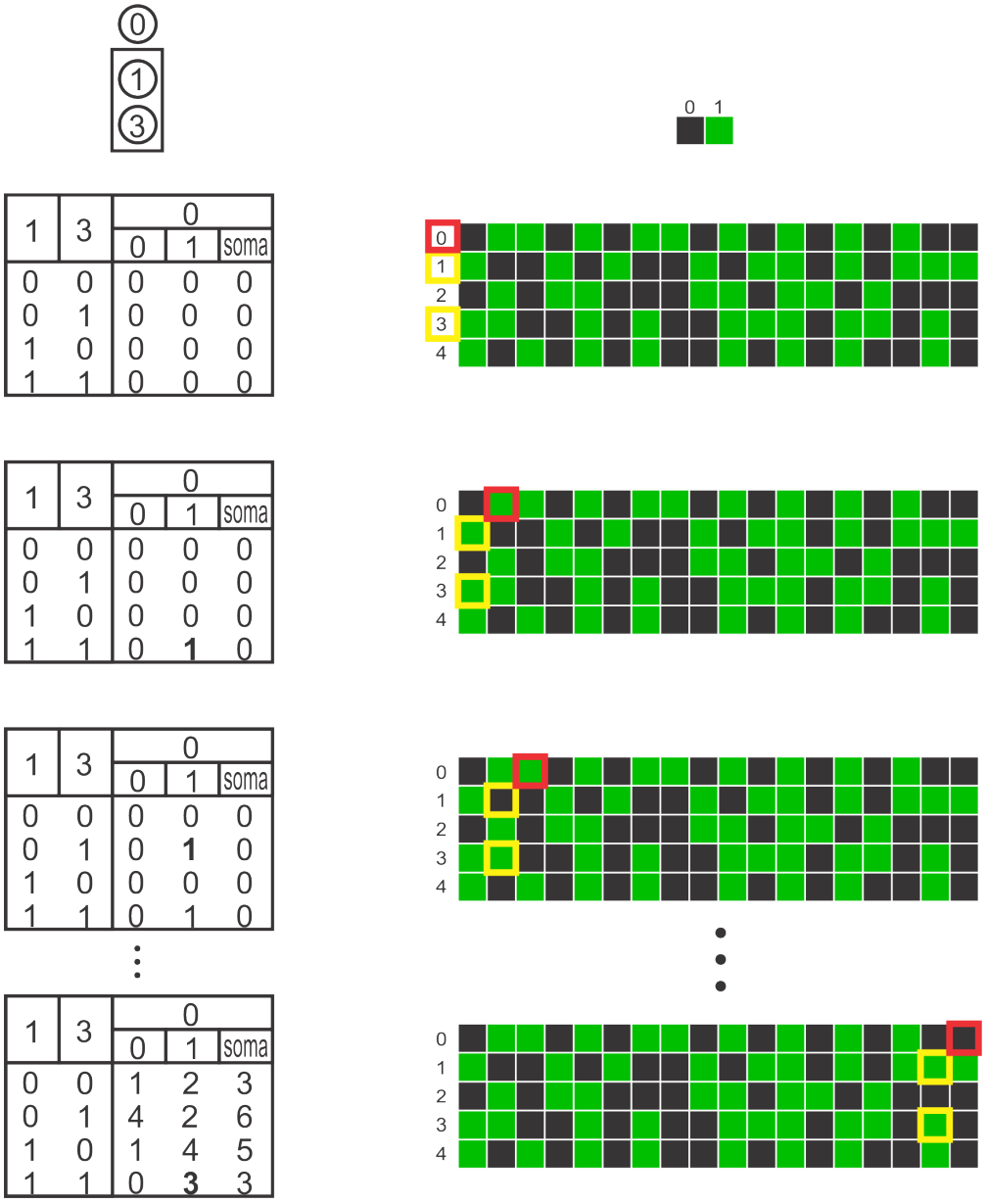
Os melhores subconjuntos de características tendem a ter menores valores de AIC, pois um maior numero de preditores implica em um maior valor de K (aumenta o risco das redes estarem super-ajustadas aos dados: *overfitting*), enquanto maiores valores de L indicam uma melhor explicação dos dados pelo conjunto de preditores.

A Figura 3.3 ilustra um exemplo do processo para o calculo do AIC do gene G_0 tendo como conjunto de preditores $\{G_1, G_3\}$ e o conjunto de amostras (*microarray*) $M_{n\times m}$ onde n é o número de genes e m o número de amostras. Para o calculo de $L(G_0, \{G_1, G_3\})$ cria-se uma tabela de contagem que é preenchida segundo os valores do gene alvo G_0 e do conjunto de preditores $\{G_1, G_3\}$. O índice t $(1 \leq t \leq n-1)$ indica a amostra temporal atual dos preditores. Aqui vale notar que a qualidade da predição do valor do gene alvo no próximo instante de tempo $(t+1)$ está sendo avaliada com base nos valores dos preditores no instante de tempo atual (t), conforme os axiomas do modelo PGN. Uma vez preenchida a tabela de contagens através da observação de todas as amostras $(1 \leq t \leq n-1)$, as contagens são transformadas em probabilidades condicionais (dividindo-se cada contagem pela soma das contagens de sua respectiva linha). Uma vez determinadas todas as probabilidades condicionais, procede-se aos cálculos da verossimilhança L e do fator de penalização K conforme definidos nas Equações 3.2 e 3.3 respectivamente. Uma observação importante é que para os casos onde a contagem resulta em zero, ela simplesmente não afeta o valor do produtório (não entra na conta). Ou seja, $0^0 = 1$.

Como a rede inferida é obtida pela composição do melhor subconjunto (cromossomo) de cada gene obtido até o momento, o AIC de uma rede pode ser obtido através da soma $\sum_i AIC(G_i, \mathbf{X_i}*)$, sendo $\mathbf{X_i}*$ o melhor subconjunto obtido para o gene G_i até a geração atual.

3.5 Cruzamento

A fase de cruzamento deve simular o mecanismo de seleção natural, no qual os pais mais adaptados de acordo com a função de aptidão escolhida tendem a gerar mais filhos, mas permitindo que os pais menos aptos também possam gerar descendentes. Isso porque os indivíduos com baixa aptidão podem ter características genéticas que sejam favoráveis à geração de um indivíduo que seja a melhor solução para um dado problema, sendo que tais características podem não estar presentes em nenhum outro cromossomo da população [Linden, 2012]. Se apenas os melhores indivíduos se reproduzirem, a população tenderá a ser composta por indivíduos cada vez mais semelhantes, e faltará diversidade a esta população, impedindo que a evolução seja satisfatória, causando o efeito da convergência



$K(G_0, \{G_1, G_3\}) = 2^{1+2} = 2^3 = 8$

$L(G_0, \{G_1, G_3\}) = (\frac{1}{3})^1 \times (\frac{2}{3})^2 \times (\frac{4}{6})^4 \times (\frac{2}{6})^2 \times (\frac{1}{4})^1 \times (\frac{3}{4})^3 \times (\frac{0}{3})^0 \times (\frac{3}{3})^3 = 0.000257202$

$AIC(G_0, \{G_1, G_3\}) = 2 \times K(G_0, \{G_1, G_3\}) - 2 \times L(G_0, \{G_1, G_3\}) =$
$2 \times 8 - 2 \times 0.000257202 = 32.53130033$

Figura 3.3: Calculo do AIC do gene G_0 com base no seu conjunto de preditores $\{G_1, G_3\}$.

genética.

As duas etapas principais do cruzamento são: seleção e recombinação. Na técnica proposta para a inferência de redes gênicas, adotamos o método da roleta para seleção dos indivíduos, no qual cada indivíduo (cromossomo) recebe uma fatia proporcional à sua aptidão. Uma vez que os indivíduos a serem cruzados foram selecionados, eles trocam

partes de seus cromossomos entre si através de um operador de recombinação.

3.5.1 Seleção

Na fase de seleção, um subconjunto dos indivíduos de uma dada população é escolhido para reproduzir e gerar a próxima população de indivíduos. Foi aplicado o método da roleta, em que a probabilidade de um conjunto ser escolhido é inversamente proporcional ao seu AIC. Desta forma, os melhores conjuntos tendem a compor a maior parte dos indivíduos a serem recombinados. O método da roleta requer a soma dos inversos dos AICs dos indivíduos e usa este valor para determinar a probabilidade de cada individuo ser escolhido para ser submetido à recombinação, conforme segue na Equação 3.4:

$$Roleta(i) = \frac{\frac{1}{AIC_i}}{\sum_{i=1}^{n} \frac{1}{AIC_i}} \tag{3.4}$$

Na Tabela 3.1 observa-se um exemplo de uma população de 5 indivíduos, seus respectivos valores de AIC, os inversos dos AICs e suas probabilidades (pedaços da roleta) de serem escolhidos pela roleta. A roleta para esse exemplo está ilustrada na Figura 3.4.

Tabela 3.1: Roleta para 5 indivíduos

AIC	1/AIC	Pedaço da Roleta (%)
40	0.0250	16.22875
35	0.0286	18.54714
37.5	0.0267	17.31066
20	0.0500	32.45750
42	0.0238	15.45595

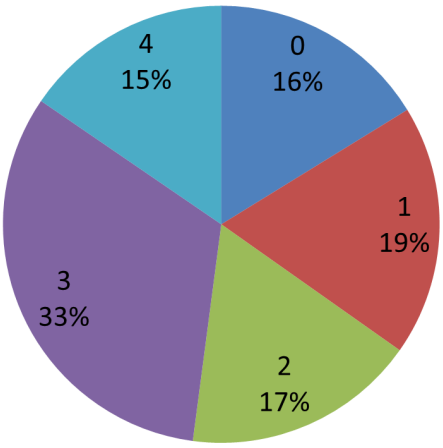


Figura 3.4: Divisão da roleta entre os indivíduos da população para o exemplo da Tabela 3.1.

Para selecionar um dos indíviduos, escolhe-se aleatoriamente um valor do intervalo de 0 até o somatório dos inversos dos AICs (denominador da Equação 3.4) e, em seguida, seleciona-se o primeiro indivíduo cuja soma acumulada nele superar o valor sorteado. A Figura 3.5 ilustra as somas acumuladas de cada indivíduo para o exemplo da Tabela 3.1. Por exemplo, se o valor escolhido for 0,04, o indivíduo 1 será selecionado.

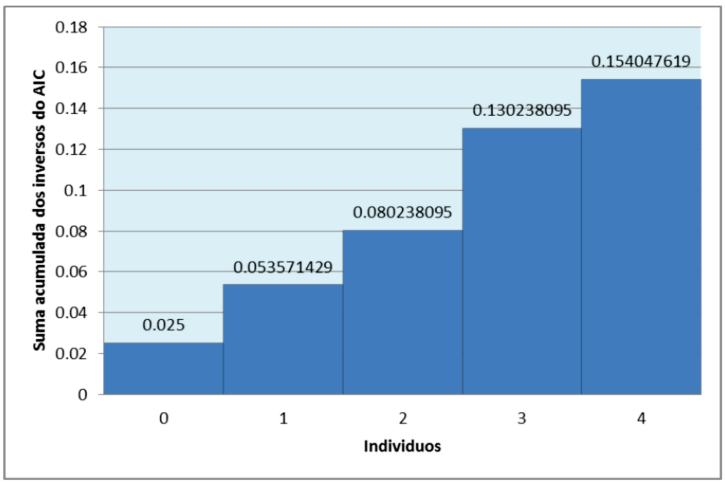


Figura 3.5: Avaliaçoes acumuladas, para selecionar um dos indivíduos pelo método da roleta. Por exemplo, se o valor sorteado for 0,04, o indivíduo 1 será o escolhido.

3.5.2 Recombinação

Na fase de recombinação (cruzamento), os indivíduos selecionados formam casais (pares) aleatórios. Cada par recombina sua informação genética para gerar dois filhos que irão compor a próxima geração de indivíduos. No presente trabalho, o processo de recombinação de um determinado par de indivíduos (pais) inicia-se a partir da geração de um novo conjunto resultante da união com repetição dos seus preditores. Os elementos dessa união são embaralhados, gerando um vetor contendo esses elementos em uma ordem arbitrária. Finalmente, um dos índices desse vetor é sorteado como ponto de corte, dividindo o vetor em dois vetores, um para cada filho. Havendo elementos (preditores) em duplicata para um determinado filho, um desses preditores é transferido para o outro filho. A Figura 3.6 ilustra o processo da recombinação.

3.6 Mutação

Com o objetivo de evitar soluções de mínimos locais, a mutação é uma importante operação de algoritmo genético para mudar algumas características pontuais dos indivíduos. Já que o método proposto aplica um algoritmo para cada gene alvo, sendo que cada gene

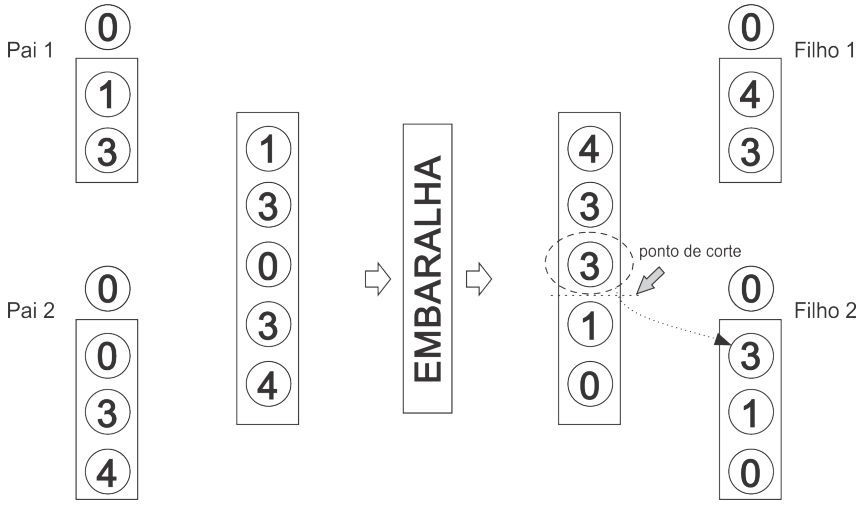


Figura 3.6: Cruzamento dos conjuntos de preditores $\{G_1, G_3\}, \{G_0, G_3, G_4\}$ do gene alvo G_0.

possui uma população composta por conjuntos de preditores (cromossomos), um percentual desses cromossomos deverá ser mutado. A quantidade de cromossomos a serem mutados é calculada em cada iteração segundo uma função linear decrescente que vai desde o valor max_{mut} (número máximo de cromossomos mutados) até 0 em 100 iterações, sendo que a variável it diz respeito a iteração atual, a qual é incrementada de 1 em 1 para cada iteração. Havendo r repetições do melhor AIC (convergência), it é reiniciado a 0, fazendo com que $CromoMutados$ seja reiniciado ao seu valor máximo (max_{mut}). A quantidade de cromossomos a serem mutados é dada pela Equação 3.5:

$$CromoMutados = \max\{0, \lceil \frac{max_{mut}(100 - it)}{100} \rceil\} \tag{3.5}$$

A Figura 3.7 mostra uma situação hipotética da evolução da função linear decrescente (em azul no grafico, dada pela Equação 3.5 sem a operação teto) e o número de cromossomos mutados a cada iteração (em vermelho no gráfico, dado pela Equação 3.5) para $max_{mut} = 10$. Note que o número de cromossomos a serem mutados é reiniciado para o valor de max_{mut} toda vez que há convergência do menor valor de AIC da população (após r repetições, sendo $r = 10$ para esse exemplo hipotético).

Uma vez determinado o número de cromossomos a serem mutados em uma determinada iteração, os cromossomos a serem mutados são sorteados com distribuição uniforme. Sendo C o conjunto de preditores de um cromossomo a ser mutado, em uma mutação tradicional, um gene G_i qualquer da rede seria sorteado com distribuição uniforme, seja para ser incluido no conjunto C caso $G_i \notin C$, ou para ser excluído de C caso contrário ($G_i \in C$). O problema desse esquema de mutação é que o número de preditores de um cromossomo normalmente é muito menor que o número total de genes da rede, fazendo com que a maioria das mutações acabe incluindo o gene sorteado ao invés de excluí-lo,

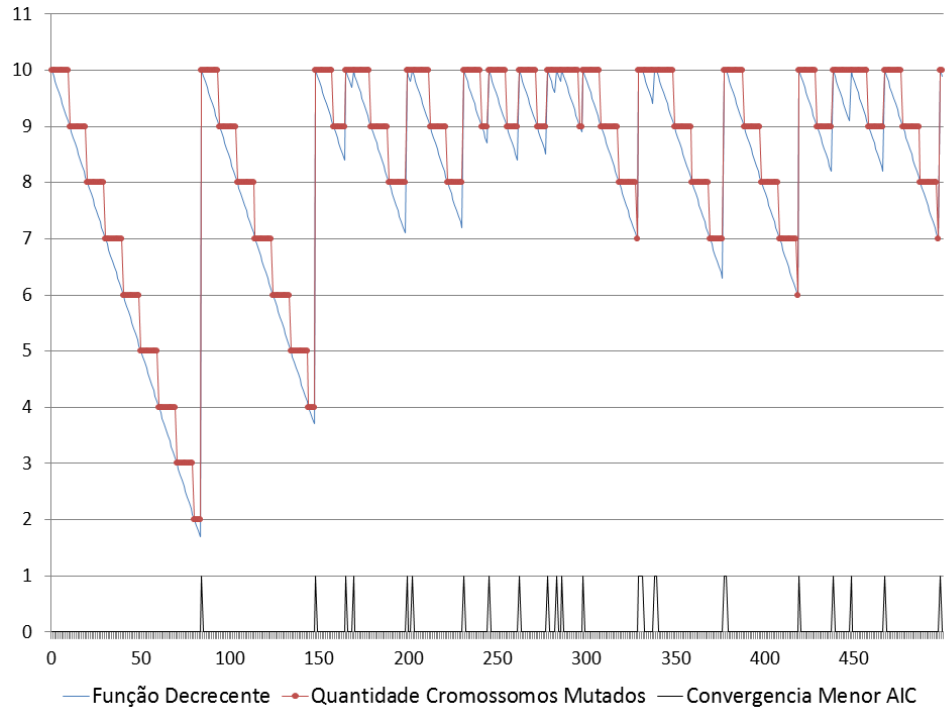
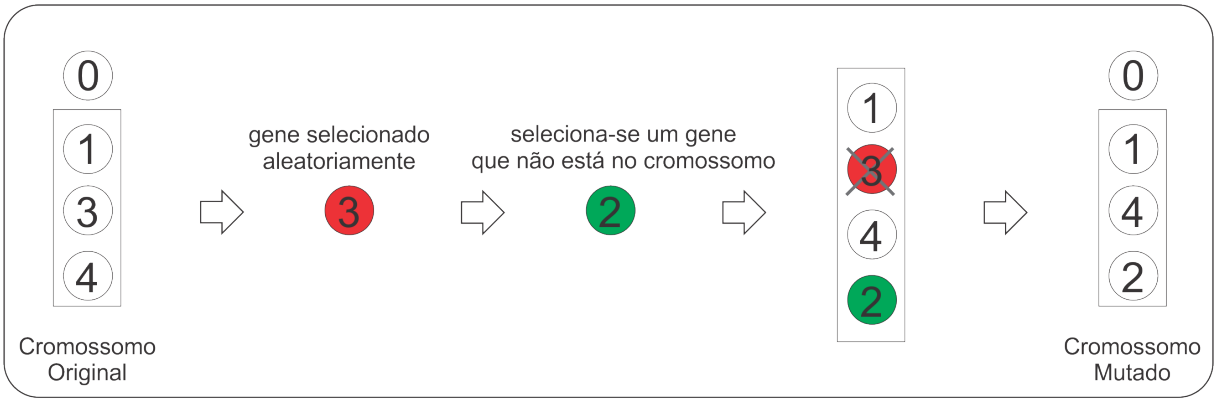


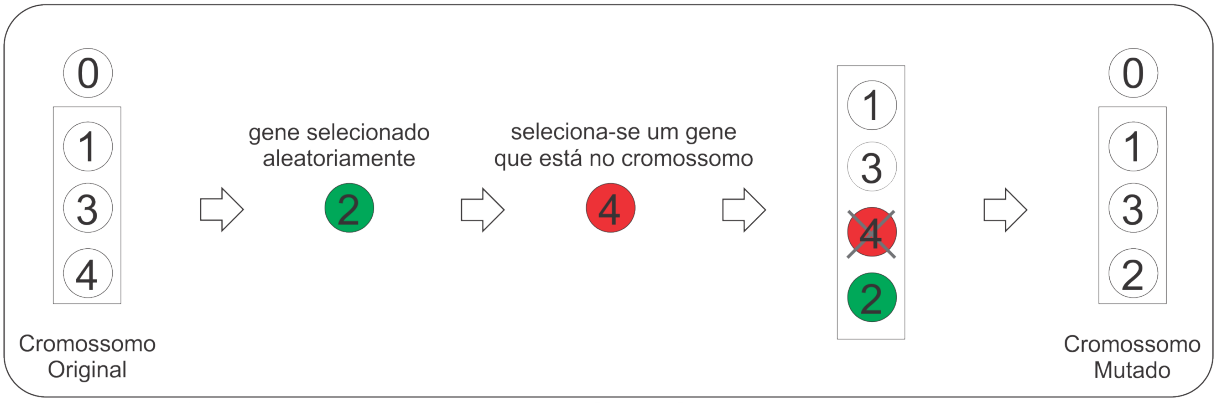
Figura 3.7: Evolução da quantidade de genes mutados nas iterações.

induzindo a um aumento do grau de quase todos os cromossomos. Para evitar que isso aconteça, propomos um esquema de mutação que se baseia em troca de genes. Ou seja, se $G_i \in C$ então ele dará lugar a um outro gene $G_j \notin C$ também sorteado com distribuição uniforme. Caso contrário, se $G_i \notin C$, ele será incluído no lugar de outro gene $G_j \in C$ sorteado com distribuição uniforme. A Figura 3.8 ilustra esse esquema de mutação

No próximo capítulo (Capítulo 4) serão apresentados alguns resultados desse método aplicado a dados simulados de expressão gênica.



Caso 1: o gene a mutar está incluído no cromossomo



Caso 2: o gene a mutar não está incluído no cromossomo

Figura 3.8: Processo de mutação de um cromossomo

Capítulo 4

Resultados experimentais

4.1 Considerações preliminares

4.1.1 Descrição dos experimentos

Neste capitulo, serão apresentados quatro experimentos:

1. O primeiro experimento mostra o potencial do AIC como função de aptidão para orientar os algoritmos genéticos em busca de redes que tenham um bom balanço entre complexidade e explicação do conjunto de dados disponíveis.

2. O segundo experimento mostra que há um benefício ao aplicar algoritmos genéticos partindo de uma população inicial gerada pela busca exaustiva por informação mútua (BEIM) (Seção 3.3) ao invés de gerar a população inicial aleatoriamente. A esses métodos, denotamos pelas siglas PB (de População inicial obtida por Busca exaustiva - "População Busca"), e PA (de População inicial obtida Aleatoriamente - "População Aleatória") respectivamente.

3. O terceiro experimento compara as inferências usando apenas BEIM com o método proposto (PB).

4. O quarto experimento envolve uma comparação entre PB, BEIM e o método de algoritmo genético proposto no trabalho de Mendoza *et al* [Mendoza et al., 2012].

4.1.2 Busca Exaustiva por Informação Mútua – BEIM

Conforme a Seção 3.3, o método BEIM é aplicado para gerar a população inicial para o método proposto PB. Além disso, o BEIM por si só será comparado com os métodos de algoritmos genéticos. É importante destacar que, embora o método BEIM seja de

natureza determinística de solução única, em muitas situações há diversos subconjuntos de preditores empatados (de acordo com a informação mútua) em primeiro lugar. Assim, nos experimentos, fizemos com que o BEIM sorteie um dos subconjuntos empatados, tornando-o estocástico. Sendo assim, execuções diferentes do BEIM para um mesmo conjunto de dados poderão resultar em redes distintas.

4.1.3 Geração das populações iniciais aleatórias

Para a geração das populações iniciais aleatórias como pontos de partida do método PA, cada cromossomo é constituído de um certo subconjunto de genes obtido do seguinte procedimento. Para cada gene da rede, sorteia-se um número do intervalo $[0; 1]$ com distribuição uniforme. Caso esse número seja menor ou igual a $\frac{\langle k \rangle}{n}$, sendo $\langle k \rangle$ o número médio de genes pertencentes aos cromossomos (grau médio) definido previamente e n o número de genes da rede, então o gene em questão pertence ao cromossomo. Caso contrário, o gene em questão não pertence ao cromossomo. Dessa forma, a distribuição de probabilidades de graus dos cromossomos segue uma distribuição binomial $Bin(n, \frac{\langle k \rangle}{n})$.

4.1.4 Geração das redes gabarito

Em todos os experimentos, as redes gabarito (*groundtruths*) foram geradas a partir dos modelos de redes aleatórias Erdös-Rényi (ER) [Erdös and Rényi, 1959] e livres de escala Barabási-Albert (BA) [Barabási, 2009] com $\langle k \rangle = 3$ (ver Seção 2.6). No caso de redes livres de escala, foi adotado o parâmetro $\gamma = 2.5$ como fator de decaimento da lei de potência, tendo em vista que a distribuição de graus dos elementos das topologias de redes biológicas normalmente seguem uma lei de potência com $2 < \gamma < 3$ (ver Seção 2.6.2). As funções de dependência (regras lógicas) entre os genes são modeladas por redes Booleanas (BN) e redes Booleanas probabilísticas (PBN) (ver Seção 2.4). Tais regras lógicas são escolhidas aleatoriamente com distribuição uniforme. Para cada gene da rede gabarito, o algoritmo de Quine–McCluskey [McCluskey, 1956] é aplicado para checar se a função Booleana de uma regra lógica de predição selecionada pode ou não ser reduzida, tendo em vista que algumas regras lógicas podem não depender necessariamente de todos os preditores estipulados (*e.g.*, tautologia e contradição são regras lógicas constantes que não dependem de qualquer variável). Apenas regras lógicas que dependem de todas as variáveis são definidas para as funções preditoras.

Para todos os experimentos, foram gerados então quatro tipos de redes gabarito:

- ERBN (topologia Erdös-Rényi e Rede Booleana),
- ERPBN (topologia Erdös-Rényi e Rede Booleana Probabilística),

- BABN (topologia Barabasi-Albert e Rede Booleana) e
- BAPBN (topologia Barabasi-Albert e Rede Booleana Probabilística).

Nas redes Booleanas probabilísticas (ERPBN,BAPBN), todo gene terá 3 funções Booleanas preditoras distintas, sendo que a tripla de probabilidades está indicada como quarto parâmetro na Tabela 4.1 (Seção 4.1.7). Ou seja, uma das funções é aplicada quase sempre (probabilidade $0,98$) enquanto as outras duas são aplicadas com probabilidade $0,01$ cada, simulando assim o quasi-determinismo inerente em sistemas biológicos.

4.1.5 Geração dos dados de expressão gênica

Uma vez tendo a topologia e as regras lógicas de predição da rede gabarito, é possível gerar dados de expressão gênica simulados a partir de um determinado estado inicial, bastando para isso aplicar essas regras para determinar o valor do gene no próximo instante de tempo. Para isso, fixando-se o número de instantes de tempo m (número de amostras), sorteia-se com distribuição uniforme um estado inicial s_0 dos 2^n possíveis estados. A partir daí a evolução dos estados do sistema evolui de s_0 a s_{m-1} através de aplicações sucessivas das funções preditoras. Em caso de algum estado ser repetido no processo da simulação para uma rede Booleana (na qual há apenas uma única função lógica possível para cada gene), isso significa que o sistema entrou em um ciclo atrator, o que é um caso indesejável já que subsequências amostrais se repetem, reduzindo o número de amostras distintas. Nesse caso, os dados são descartados e uma nova simulação é iniciada a partir de um outro estado inicial s_0 sorteado. Esse processo se repete enquanto houver estados idênticos. Assim, garante-se que cada simulação gerará m estados diferentes.

4.1.6 Métricas de avaliação topológica

Para avaliar os resultados, foram comparadas as redes inferidas com as redes gabaritos usando duas métricas de similaridades topológicas com base nos números de verdadeiros/falsos positivos e verdadeiros/falsos negativos: o valor de predição positiva (do inglês: *Positive Predictive Value* - PPV) e a similaridade (SIM) [Dougherty, 2011]. Uma aresta verdadeira-positiva é uma aresta dirigida presente tanto no gabarito quanto na rede inferida, e uma aresta verdadeira-negativa é uma aresta dirigida que não está presente em ambas as redes. Uma aresta falso-positiva é uma aresta presente na rede inferida que não está presente na rede gabarito, enquanto uma aresta falso-negativa está presente na rede gabarito e ausente na rede inferida. Sendo TP, TN, FP e FN as quantidades de verdadeiros-positivos, verdadeiros-negativos, falsos-positivos e falsos-negativos respectiva-

mente, PPV é definido por:

$$PPV = \frac{TP}{TP + FP} \tag{4.1}$$

e SIM é definido por

$$SIM = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{4.2}$$

4.1.7 Configuração dos parâmetros

Para os três primeiros experimentos, os valores dos parâmetros adotados são apresentados na Tabela 4.1. O objetivo consistiu na análise de situações nas quais o número de amostras de dados (instantes de tempo) seja muito limitado, uma vez que tais situações são típicas em conjuntos de dados reais.

Tabela 4.1: Parâmetros utilizados nos experimentos.

Parâmetro	Valores
Tamanho da rede gabarito (numero de genes n)	100
Grau médio da rede gabarito $\langle k_{gab} \rangle$	3
Número de amostras (tamanho do sinal m)	{30,60}
Probabilidades das funções Booleanas das PBNs	$(0,98; 0,01; 0,01)$
Número de gerações (iterações dos algoritmos genéticos it)	1000
Tamanho da população (número de cromossomos C por gene)	100
Probabilidade mínima de cada cromossomo ser mutado	0
Probabilidade máxima de cada cromossomo ser mutado	0.1
Número de repetições do melhor AIC como critério de convergência (r)	10
Grau médio dos cromossomos na população inicial $\langle k_{cro} \rangle$	3
Grau k da busca exaustiva por informação mútua (BEIM)	3

Para cada tipo de rede, foram geradas 10 redes, sendo que para cada rede foram gerados 10 conjuntos de dados de expressão gênica. Finalmente, pelo fato dos algoritmos genéticos serem estocásticos, podendo gerar um resultado diferente a cada execução, os método propostos foram executado 10 vezes para cada conjunto de dados de expressão gênica. A mesma observação se aplica ao método BEIM, embora ele seja de natureza determinística, é possível torná-lo estocástico devido a possibilidade de empates (ver Seção 4.1.2). Sendo assim, cada tipo de rede resulta em 1000 redes inferidas, as quais são avaliadas pelas métricas PPV e SIM citadas anteriormente.

4.2 AIC *versus* PPV e SIM

Aqui serão analisadas as correlações do AIC em relação ao PPV e do AIC em relação ao SIM para os quatro tipos de rede mencionados na Seção 4.1.4 (ERBN, ERPBN, BABN,BAPBN), com os valores de parâmetros dados na Tabela 4.1 (variando o número de amostras m em 30 e 60), com o objetivo de avaliar o potencial do AIC em conduzir o algoritmo genético para resultados satisfatórios. As populações iniciais dos cromossomos foram obtidas aleatoriamente conforme descrição na Seção 4.1.3 (método PA).

As Figuras 4.1, 4.3, 4.5 e 4.7 mostram gráficos para ERBN, ERPBN BABN e BAPBN respectivamente, onde o AIC é representado no eixo x e as métricas de similaridade e o valor de predição positiva são representados no eixo y. Cada figura contém 4 gráficos dispostos numa matriz 2 por 2, sendo que a métrica de similaridade é variada nas colunas (SIM na primeira coluna e PPV na segunda coluna), enquanto o número de amostras é variado nas linhas (30 na primeira linha, e 60 na segunda linha). Os pontos dos gráficos representam os valores (AIC, SIM) e (AIC, PPV) alcançados pela melhor rede inferida (em termos de AIC) até uma determinada geração. A rede inferida de uma determinada geração é composta pelo melhor cromossomo (de acordo com o AIC) de cada gene obtido até então. As cores dos pontos indicam os intervalos de gerações, em que foram obtidas as redes correspondentes: gerações 1-200 (azul), 201-400 (verde escuro), 401-600 (verde claro), 601-800 (marrom), 801-1000 (vermelho). As correlações de Pearson correspondentes entre AIC e SIM e entre AIC e PPV estão indicadas nos títulos dos gráficos. Apesar de 1000 experimentos envolvendo diferentes gabaritos e conjuntos de dados terem sido gerados para cada configuração de parâmetros, os resultados são mostrados apenas para um experimento, Os gráficos e correlações foram semelhantes para a maioria dos experimentos, conforme as distribuições dos valores das correlações obtidas para os 1000 experimentos ilustradas nos *boxplots* das Figuras 4.2, 4.4, 4.6 e 4.8. Os sumários dessas distribuições com as respectivas médias e desvios padrões são apresentadas na Tabela 4.2. Vale notar que um maior número de amostras implica em médias maiores dos valores absolutos das correlações.

É importante ressaltar que todas as correlações foram altamente negativas (valores absolutos maiores do que $0,90$) para todas os experimentos. Além disso, as cores dos pontos indicam que as últimas gerações tendem a alcançar os menores valores de AIC, como esperado. Finalmente, vale notar que a convergência para os melhores resultados é obtida após um número pequeno de iterações, já que a variação dos valores de AIC, SIM e PPV é muito maior para as redes obtidas nas 200 primeiras gerações (todos os pontos azuis) do que para as redes obtidas nas 800 últimas (todos os pontos, exceto os azuis). Como o objetivo é minimizar o AIC, esses resultados indicam que o AIC tem um grande potencial para conduzir os algoritmos genéticos na obtenção de redes que descrevem bem

os dados.

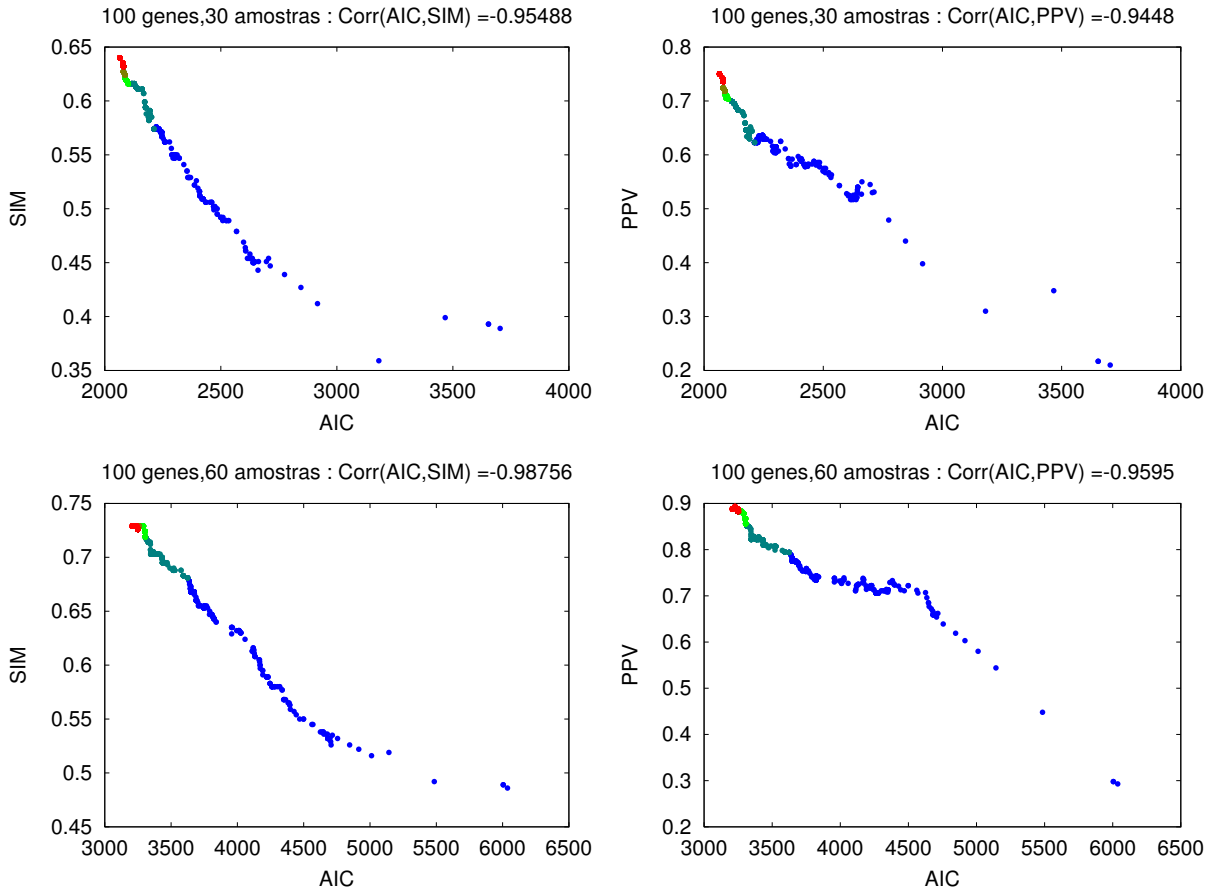


Figura 4.1: Valores de (AIC, SIM) e (AIC, PPV) obtidos em cada geração para uma execução do método proposto partindo de populações iniciais aleatórias. Tipo de rede: ERBN. As cores dos pontos indicam o intervalo de gerações, em que foram obtidas as redes correspondentes: gerações 1-200 (azul), 201-400 (verde escuro), 401-600 (verde claro), 601-800 (marrom), 801-1000 (vermelho).

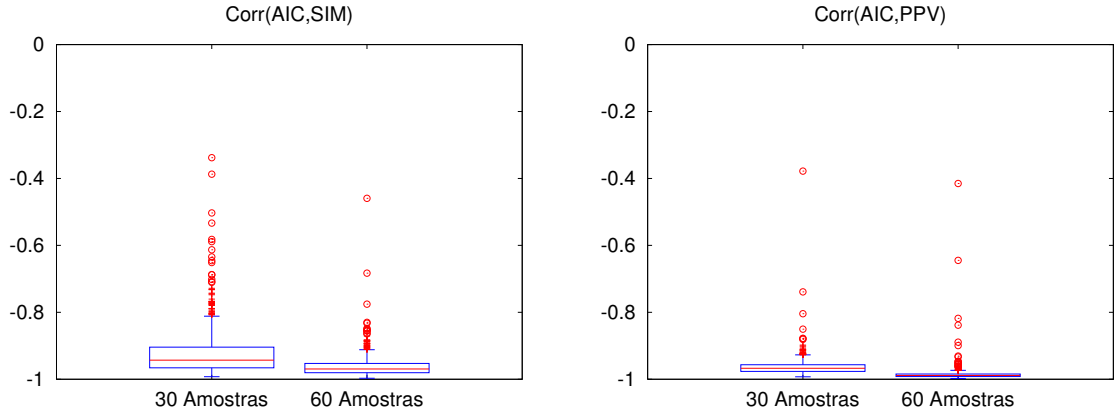


Figura 4.2: *Boxplots* das correlações entre (AIC, SIM) e (AIC, PPV) obtidas sobre 1000 experimentos envolvendo redes do tipo ERBN.

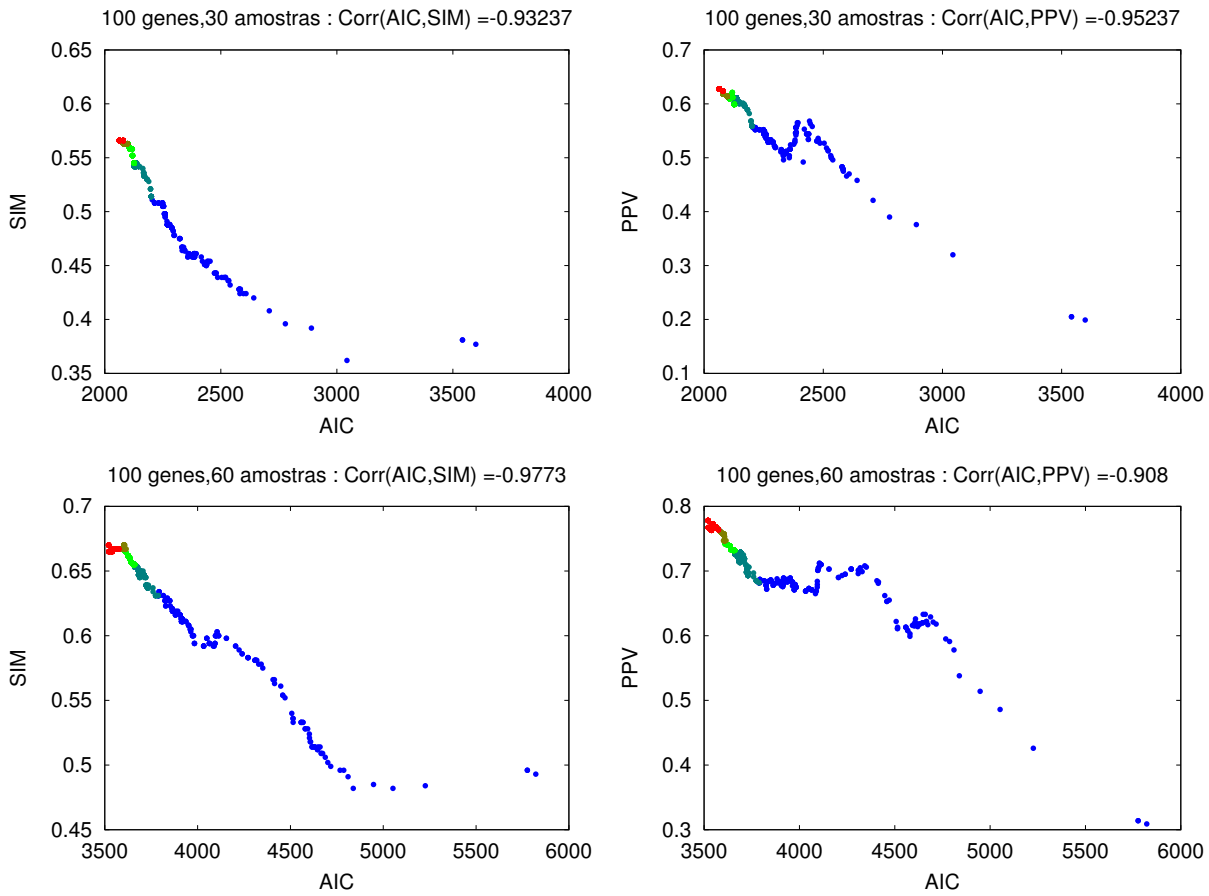


Figura 4.3: Valores de (AIC, SIM) e (AIC, PPV) obtidos em cada geração para uma execução do método proposto partindo de populações iniciais aleatórias. Tipo de rede: ERPBN. As cores dos pontos indicam o intervalo de gerações, em que foram obtidas as redes correspondentes: gerações 1-200 (azul), 201-400 (verde escuro), 401-600 (verde claro), 601-800 (marrom), 801-1000 (vermelho).

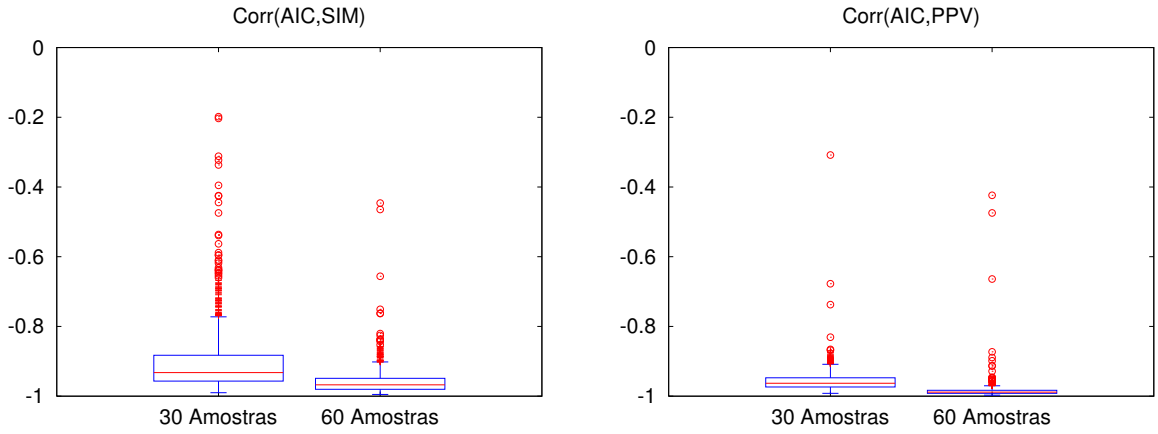


Figura 4.4: *Boxplots* das correlações entre (AIC, SIM) e (AIC, PPV) obtidas sobre 1000 experimentos envolvendo redes do tipo ERPBN.

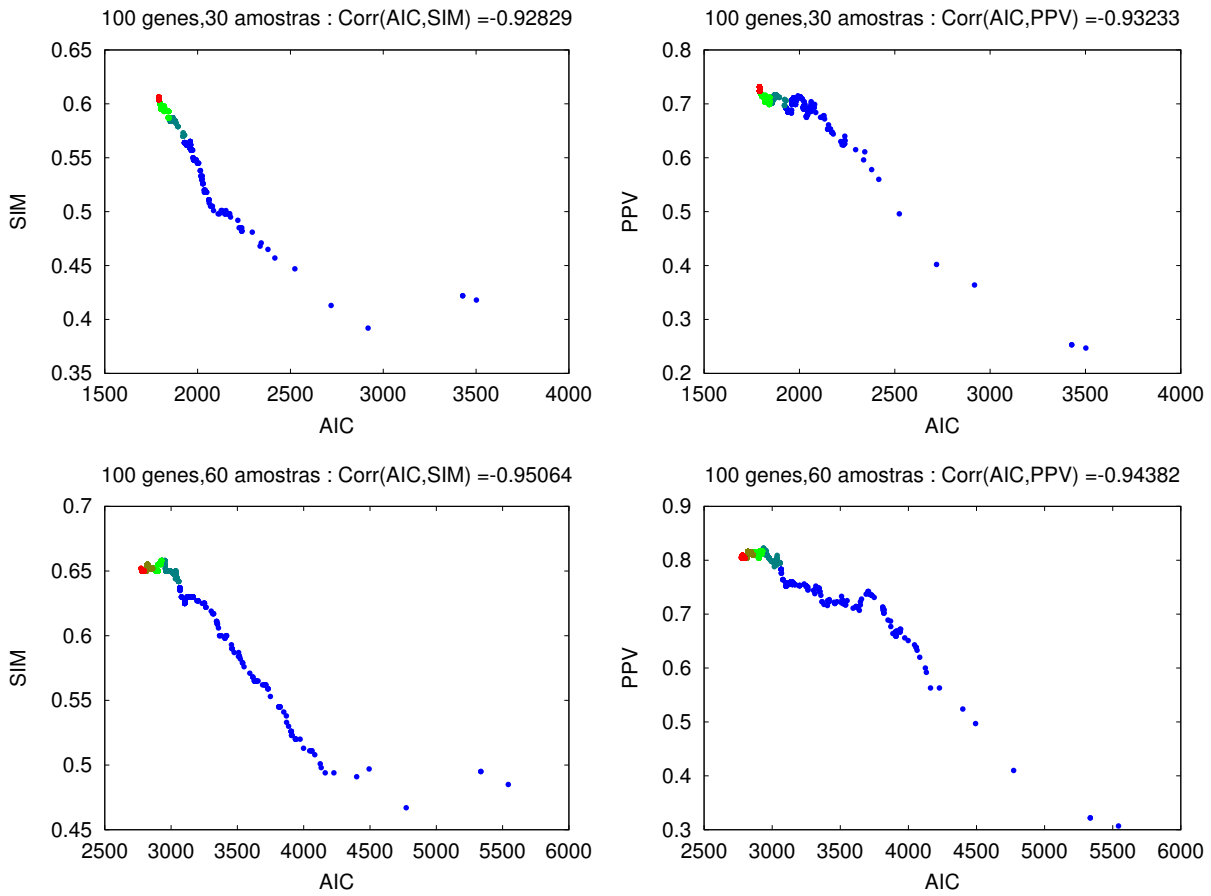


Figura 4.5: Valores de (AIC, SIM) e (AIC, PPV) obtidos em cada geração para uma execução do método proposto partindo de populações iniciais aleatórias. Tipo de rede: BABN. As cores dos pontos indicam o intervalo de gerações, em que foram obtidas as redes correspondentes: gerações 1-200 (azul), 201-400 (verde escuro), 401-600 (verde claro), 601-800 (marrom), 801-1000 (vermelho).

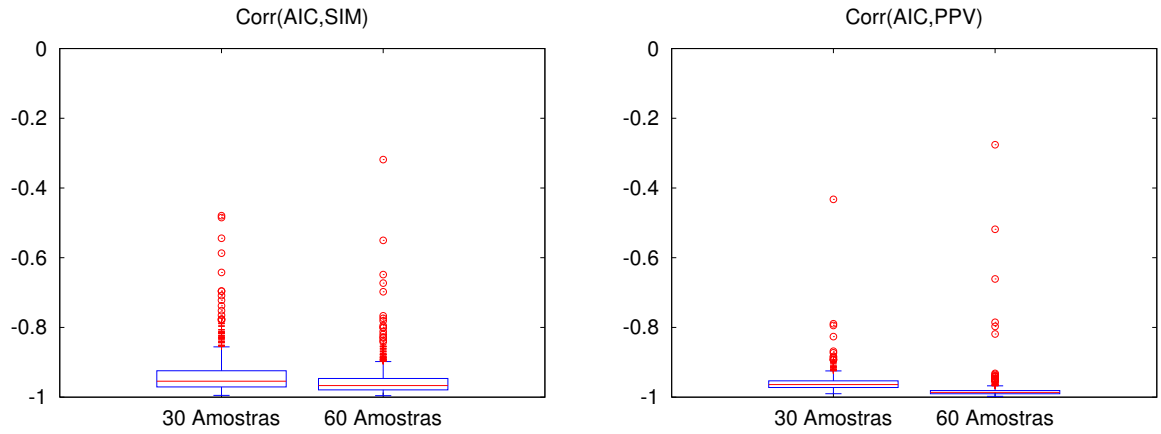


Figura 4.6: *Boxplots* das correlações entre (AIC, SIM) e (AIC, PPV) obtidas sobre 1000 experimentos envolvendo redes do tipo BABN.

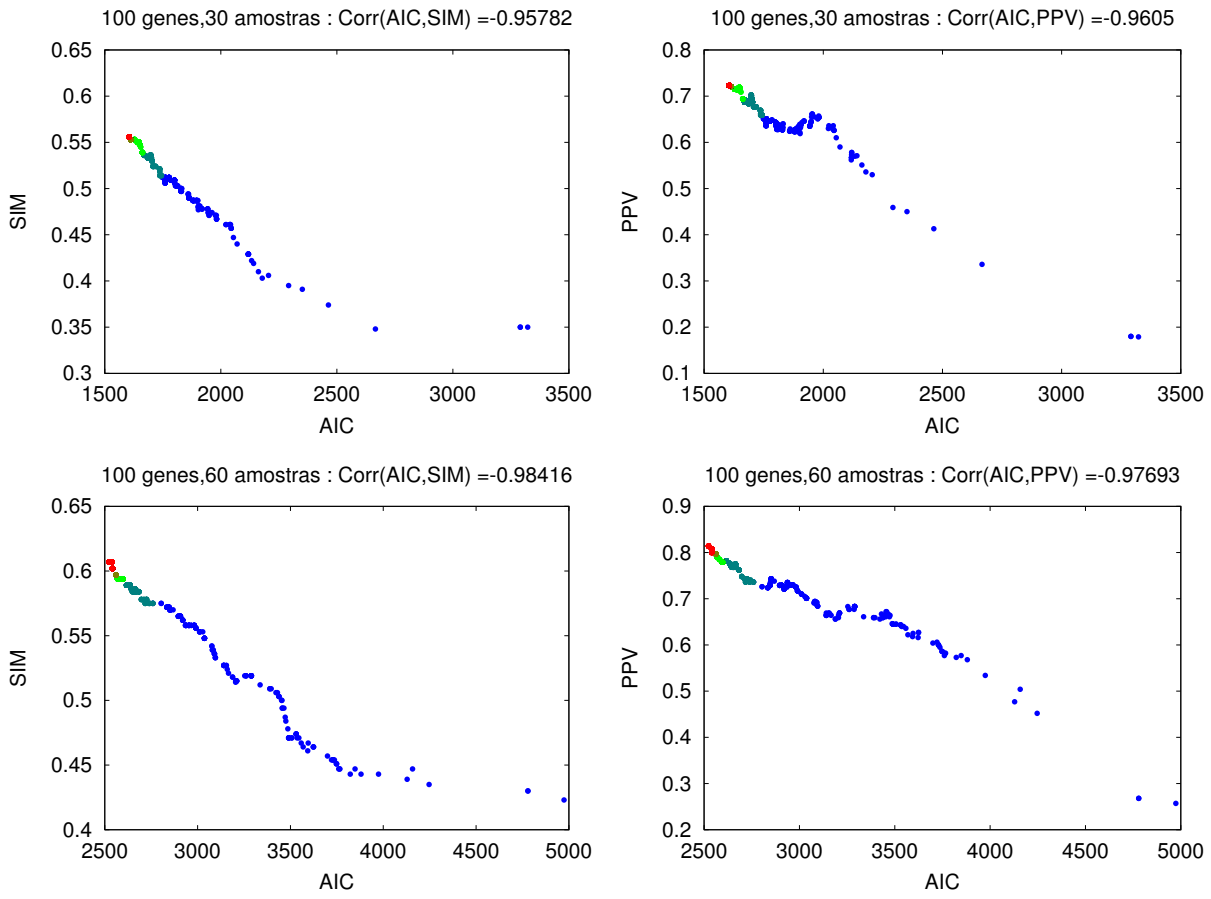


Figura 4.7: Valores de (AIC, SIM) e (AIC, PPV) obtidos em cada geração para uma execução do método proposto partindo de populações iniciais aleatórias. Tipo de rede: BAPBN. As cores dos pontos indicam o intervalo de gerações, em que foram obtidas as redes correspondentes: gerações 1-200 (azul), 201-400 (verde escuro), 401-600 (verde claro), 601-800 (marrom), 801-1000 (vermelho).

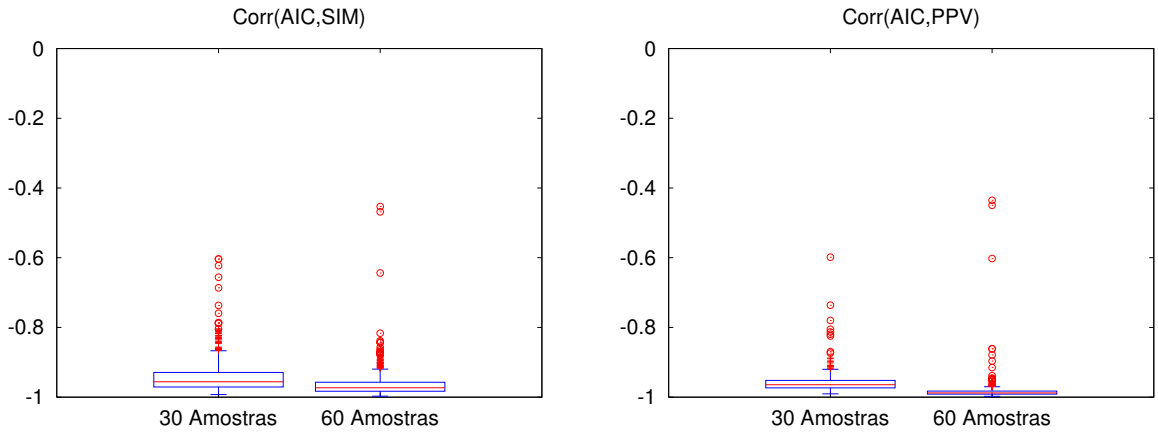


Figura 4.8: *Boxplots* das correlações entre (AIC, SIM) e (AIC, PPV) obtidas sobre 1000 experimentos envolvendo redes do tipo BAPBN.

Tabela 4.2: Médias e desvios padrões dos *boxplots* apresentados nas Figuras 4.2, 4.4, 4.6 e 4.8.

		SIM		PPV	
		30 amostras	60 amostras	30 amostras	60 amostras
ERBN	média	-0.96	-0.984	-0.943	-0.965
	desvio padrão	0.024	0.029	0.043	0.035
ERPBN	média	-0.96	-0.982	-0.939	-0.956
	desvio padrão	0.025	0.032	0.051	0.043
BABN	média	-0.956	-0.985	-0.903	-0.959
	desvio padrão	0.032	0.028	0.092	0.039
BAPBN	média	-0.964	-0.986	-0.925	-0.962
	desvio padrão	0.027	0.024	0.065	0.031

4.3 PA *versus* PB

Nesta seção, mostraremos que os resultados da inferência podem ser significativamente melhorados quando a população inicial é criada através de um método de inferência, tal como o algoritmo BEIM discutida na Seção 3.3.

Primeiramente, a análise será feita sobre os resultados de 50 execuções do PA e do PB sobre um único conjunto de amostras gerado para cada tipo de rede (ERBN, ERPBN, BABN, BAPBN) para mostrar que, embora esses métodos sejam estocásticos, os valores obtidos em diferentes execuções sobre um mesmo conjunto de dados são similares. A Figura 4.9 mostra os *boxplots* obtidos para as métricas PPV e SIM. Cada gráfico possui quatro *boxplots* denotados por PA.30, PB.30, PA.60, PB.60:

- PA.30: população inicial aleatória, conjunto de 30 amostras ($m = 30$)
- PB.30: população inicial inferida por BEIM, conjunto de 30 amostras ($m = 30$)
- PA.60: população inicial aleatória, conjunto de 60 amostras ($m = 60$)
- PB.60: população inicial inferida por BEIM, conjunto de 60 amostras ($m = 60$)

A Tabela 4.3 mostra as médias e os desvios padrões para os *boxplots* da Figura 4.9. É possível notar com base nesses resultados que os valores de SIM e PPV possuem uma variação pequena.

Sabendo-se que diferentes execuções do método para um mesmo conjunto de dados produzem resultados bastante similares em termos qualitativos, seja partindo-se de populações iniciais aleatórias ou de populações iniciais obtidas por BEIM, a idéia então é ampliar a análise levando em conta diversos conjunto de amostras geradas por diferentes redes gabarito para cada tipo de rede, com o objetivo de verificar se o método proposto é beneficiado quando parte de uma população inicial inferida por BEIM.

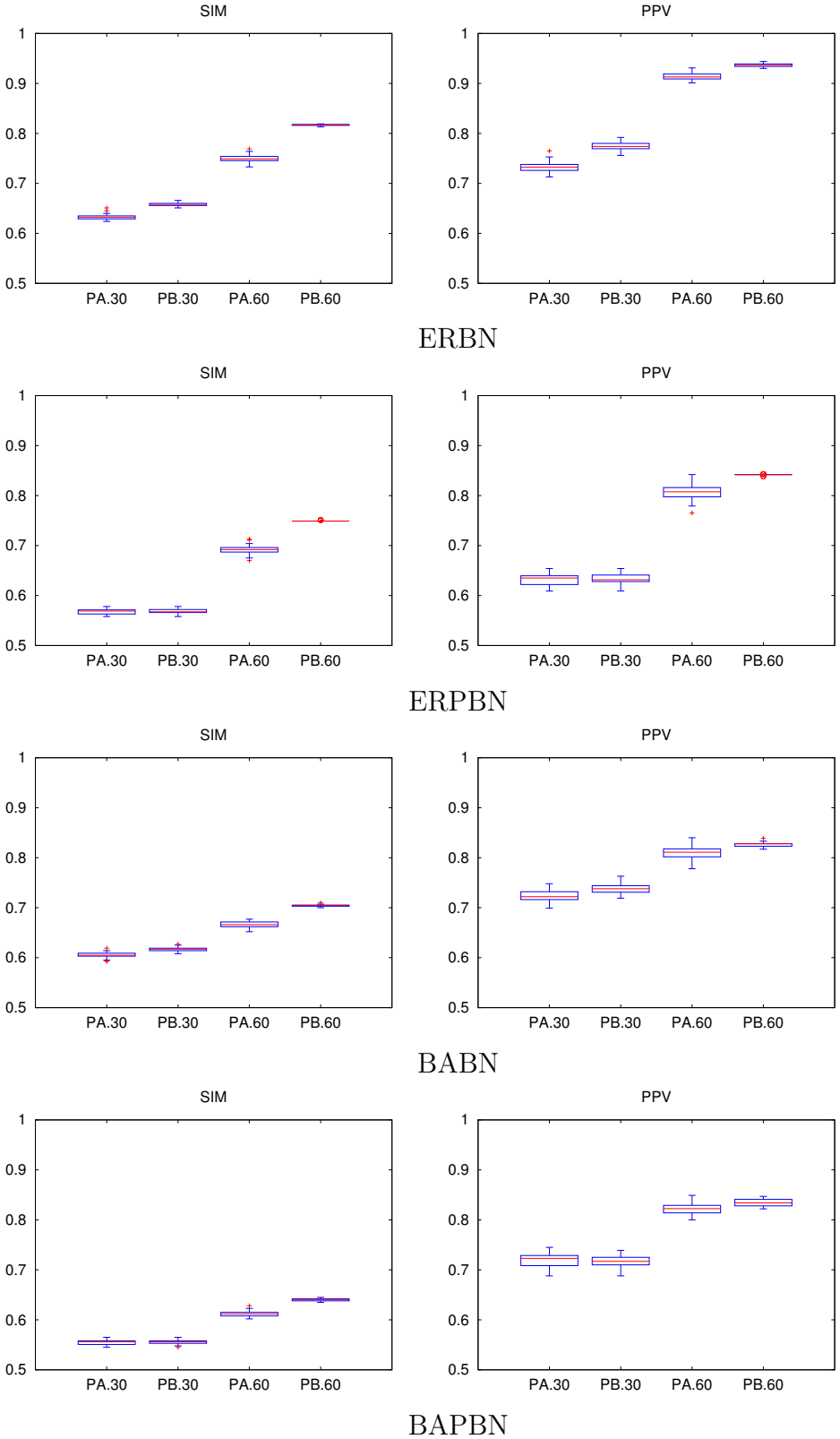


Figura 4.9: *Boxplots* das métricas SIM e PPV das 50 execuções do método proposto com população inicial aleatória (P.A.) e com população inicial obtida por BEIM (P.B.) sobre um único conjunto de amostras para cada configuração envolvendo redes gabaritos dos tipos {ERBN, ERPBN, BABN, BAPBN}, variando o número de amostras entre 30 e 60 ($m = \{30, 60\}$).

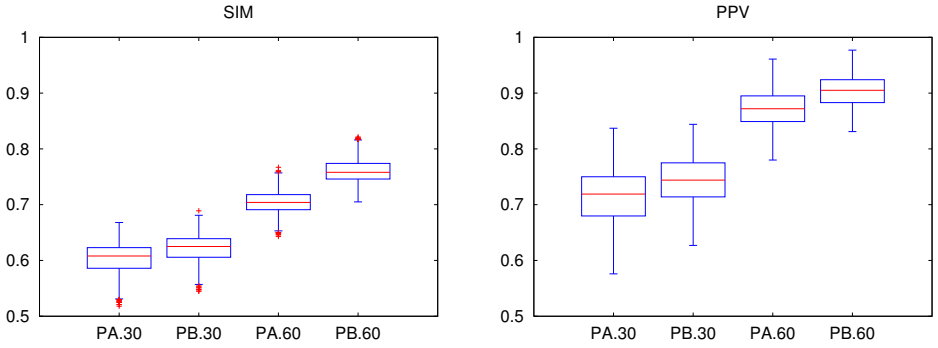
Tabela 4.3: Médias e desvios padrões dos *boxplots* apresentados na Figura 4.9.

		SIM				PPV			
		30 amostras		60 amostras		30 amostras		60 amostras	
		PA	PB	PA	PB	PA	PB	PA	PB
ERBN	média	0.632	0.658	0.75	0.817	0.733	0.775	0.914	0.936
	desvio padrão	0.006	0.003	0.007	0.002	0.011	0.007	0.007	0.004
ERPBN	média	0.567	0.568	0.692	0.749	0.63	0.633	0.806	0.842
	desvio padrão	0.005	0.005	0.009	0.001	0.012	0.011	0.016	0.001
BABN	média	0.605	0.617	0.666	0.704	0.723	0.739	0.811	0.826
	desvio padrão	0.005	0.004	0.007	0.002	0.011	0.009	0.013	0.005
BAPBN	média	0.555	0.556	0.612	0.64	0.72	0.716	0.822	0.835
	desvio padrão	0.005	0.005	0.006	0.002	0.014	0.012	0.011	0.006

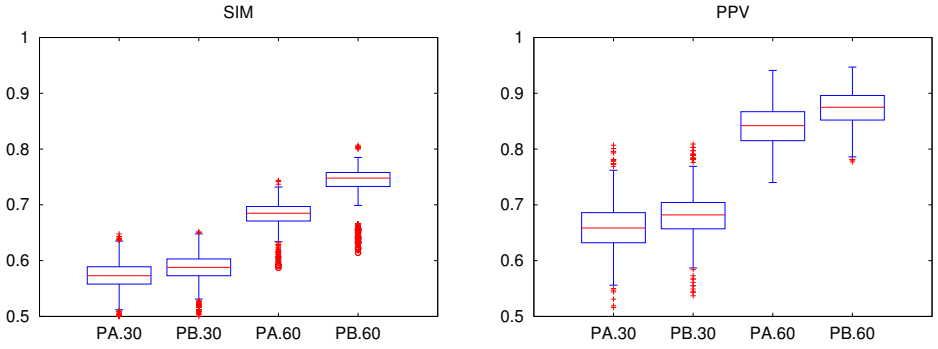
A Figura 4.10 contêm *boxplots* dos valores de SIM e PPV, onde cada *boxplot* refere-se a 1000 redes inferidas a partir de 10 redes gabaritos, 10 conjuntos de amostras simulados a partir de cada rede gabarito, e 10 execuções dos métodos PA e PB para cada conjunto de amostras. A Tabela 4.4 mostra as médias e os desvios padrões dos *boxplots* apresentados na Figura 4.10. Pode-se observar que, em todos os casos, os *boxplots* obtidos pelo método proposto partindo de populações iniciais inferidas por BEIM (PB) estão acima dos respectivos *boxplots* obtidos pelo método proposto partindo de populações iniciais aleatórias (PA). Conclui-se então que há um benefício em aplicar algoritmos genéticos partindo de populações iniciais pré-inferidas por algum método de inferência (neste caso particular, o método BEIM).

Tabela 4.4: Médias e desvios padrões dos *boxplots* apresentados na Figura 4.10.

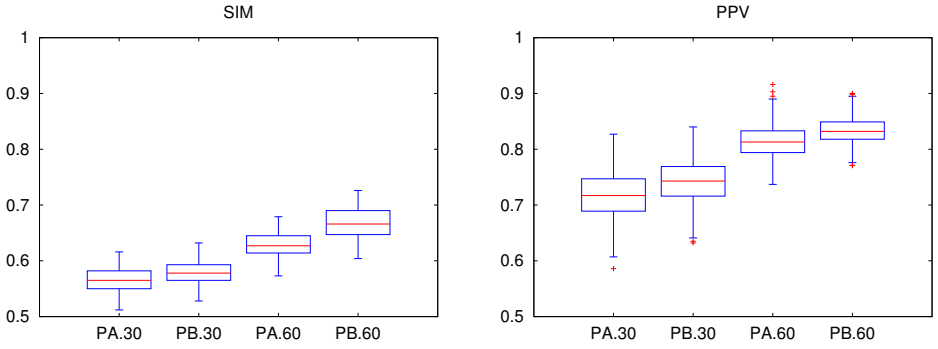
		SIM				PPV			
		30 amostras		60 amostras		30 amostras		60 amostras	
		PA	PB	PA	PB	PA	PB	PA	PB
ERBN	média	0.604	0.621	0.704	0.762	0.715	0.742	0.872	0.903
	desvio padrão	0.027	0.025	0.023	0.023	0.048	0.043	0.032	0.029
ERPBN	média	0.572	0.585	0.68	0.747	0.661	0.682	0.841	0.873
	desvio padrão	0.029	0.029	0.028	0.034	0.043	0.04	0.036	0.029
BABN	média	0.566	0.579	0.628	0.668	0.719	0.743	0.814	0.834
	desvio padrão	0.021	0.021	0.021	0.026	0.041	0.038	0.029	0.024
BAPBN	média	0.57	0.583	0.633	0.669	0.705	0.727	0.801	0.819
	desvio padrão	0.025	0.027	0.027	0.032	0.038	0.038	0.037	0.039



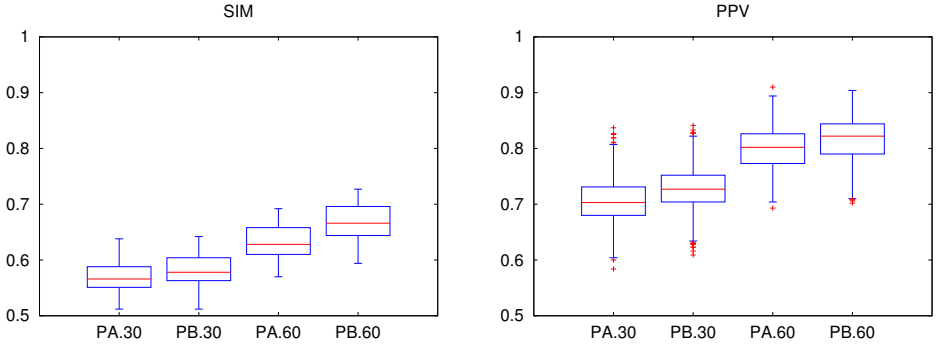
ERBN



ERPBN



BABN



BAPBN

Figura 4.10: *Boxplots* referentes a 1000 redes inferidas a partir de 10 redes gabaritos, 10 conjuntos de amostras simuladas a partir de cada rede gabarito, e 10 execuções do método para cada conjunto de amostras. Redes gabaritos dos tipos {ERBN, ERPBN, BABN, BAPBN}. Número de amostras variando entre 30 e 60 ($m = \{30, 60\}$).

4.4 PB vs BEIM

Nesta seção, é apresentada uma análise comparativa entre o método BEIM e a aplicação do método proposto partindo de uma população inicial inferida por BEIM (PB). Os gráficos dos *boxplots* das figuras desta seção seguem a mesma estrutura dos gráficos da seção anterior, substituindo-se PA por BEIM.

A Figura 4.11 contém *boxplots* dos valores de SIM e PPV, onde cada *boxplot* obtido pelos métodos refere-se a 1000 redes inferidas a partir de 10 redes gabaritos, 10 conjuntos de amostras simulados a partir de cada rede gabarito e 10 execuções dos métodos BEIM e PB para cada conjunto de amostras. A Tabela 4.5 mostra as médias e os desvios padrões dos *boxplots* apresentados na Figura 4.11. Pode-se observar que, embora os valores de SIM tenham sido ligeiramente superiores para o método BEIM (médias com diferenças inferiores a $0,04$ considerando-se o mesmo número de amostras), os valores de PPV foram consideravelmente superiores para o método PB (médias com diferenças superiores a $0,28$ considerando-se o mesmo número de amostras) em todos os casos. Assim, pode-se concluir que o método PB possui uma qualidade geral de inferência superior ao BEIM, já que o primeiro apresenta valores de PPV muito superiores enquanto mantém valores de SIM relativamente equiparáveis.

Tabela 4.5: Médias e desvios padrões dos *boxplots* apresentados na Figura 4.11.

		SIM				PPV			
		30 amostras		60 amostras		30 amostras		60 amostras	
		BEIM	PB	BEIM	PB	BEIM	PB	BEIM	PB
ERBN	média	0.656	0.621	0.785	0.762	0.432	0.742	0.614	0.903
	desvio padrão	0.023	0.025	0.024	0.023	0.03	0.043	0.048	0.029
ERPBN	média	0.618	0.585	0.759	0.74	0.396	0.68	0.593	0.874
	desvio padrão	0.03	0.029	0.033	0.034	0.035	0.04	0.043	0.029
BABN	média	0.605	0.579	0.677	0.668	0.399	0.743	0.498	0.834
	desvio padrão	0.021	0.021	0.025	0.026	0.026	0.038	0.033	0.024
BAPBN	média	0.6	0.583	0.677	0.669	0.392	0.727	0.498	0.819
	desvio padrão	0.026	0.027	0.033	0.032	0.033	0.038	0.047	0.039

4.5 Comparação envolvendo o método de Mendoza *et al*

Esta seção apresenta uma comparação dos métodos PA, PB e BEIM com o método descrito em [Mendoza et al., 2012], o qual também é baseado em algoritmos genéticos para inferência de de redes de regulação gênica. O protocolo experimental desta seção segue

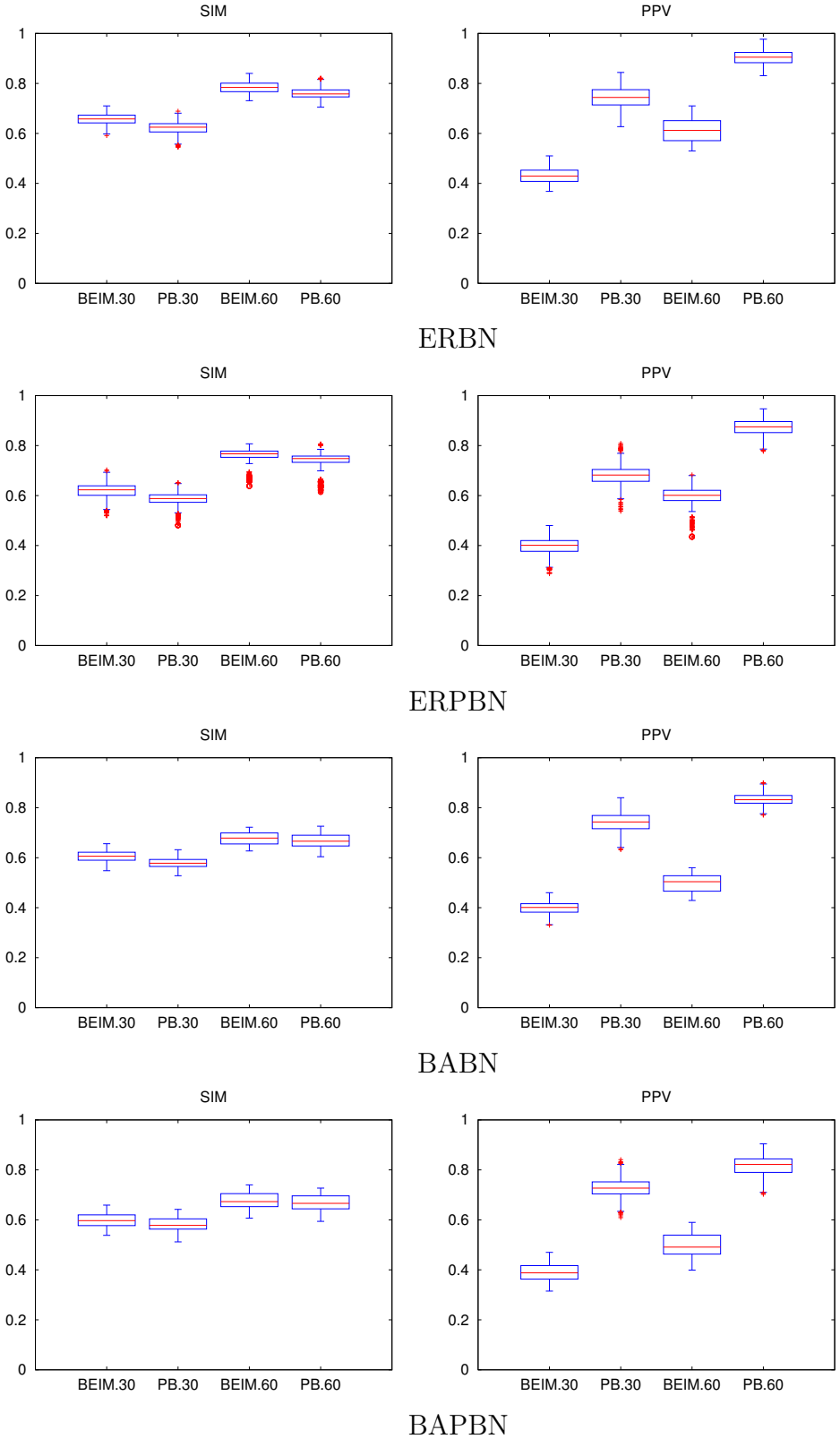


Figura 4.11: *Boxplots* referentes a 1000 redes inferidas a partir de 10 redes gabaritos, 10 conjuntos de amostras simuladas a partir de cada rede gabarito, e 10 execuções dos métodos para cada conjunto de amostras. Redes gabaritos dos tipos {ERBN, ERPBN, BABN, BAPBN}. Número de amostras variando entre 30 e 60 ($m = \{30, 60\}$).

o descrito na Tabela 4.1, exceto pelo número de amostras de expressão gênica (tamanho do sinal), que aqui é 300 sendo 10 concatenações de 30 amostras, conforme feito em [Mendoza et al., 2012]. A geração de cada trecho de 30 amostras segue o mesmo procedimento descrito na Seção 4.1.5.

No método de Mendoza *et al*, para um determinado conjunto de amostras executa-se o algoritmo genético 30 vezes para obter uma única rede consenso pelas arestas que mais aparecerem ao longo dessas redes, seguindo o princípio "sabedoria das multidões" (*wisdom of crowds*) [Mendoza et al., 2012]. Tais redes consensos foram avaliadas apenas em termos de similaridade SIM. Uma restrição importante desse método é que o grau máximo k_{max} dos genes da rede inferida deve ser definido *a priori*. Tal método foi avaliado com os valores $k_{max} = \{2, 3\}$. Os resultados desse métdodo serão denotados por "Mendoza".

A Figura 4.12 mostra os *boxplots* dos resultados de SIM obtidos para os métodos PA, PB, BEIM e Mendoza. No caso do método Mendoza, há apenas dois pontos ilustrados por $*$ e $+$, correspondendo aos resultados obtidos para $k_{max} = 2$ e $k_{max} = 3$ respectivamente. Já no caso dos métdos PA, PB e BEIM, os boxplots representam 10 redes gabarito, cada uma resultando em 10 conjuntos de amostras, sendo 10 execuções por conjunto de amostras. Sendo assim, os *boxplots* dos métodos PA, PB e BEIM representam a distribuição de 1000 valores de SIM. A Tabela 4.6 mostra as médias e os desvios padrões dos *boxplots* apresentados nas Figura 4.12. O melhor método em todos os casos foi o método proposto com população inicial inferida por BEIM (PB), enquanto o método de Mendoza acabou tendo similaridade inferior em todos os casos. Aqui é interessante notar também que, diferentemente do que ocorre nos resultados da Seção 4.4, o PB supera o BEIM em termos de similaridade.

Tabela 4.6: Médias e desvios padrões dos *boxplots* apresentados na Figura 4.12.

		SIM				
		300 amostras (concatenação de 10 conjuntos de 30)				
		PA	BEIM	PB	Mendoza $k_{max} = 2$	Mendoza $k_{max} = 3$
ERBN	média	0.799	0.863	0.886	0.611	0.564
	desvio padrão	0.019	0.018	0.019		
ERPBN	média	0.794	0.86	0.886	0.598	0.576
	desvio padrão	0.021	0.02	0.016		
BABN	média	0.689	0.728	0.741	0.612	0.561
	desvio padrão	0.015	0.012	0.014		
BAPBN	média	0.691	0.731	0.746	0.61	0.5
	desvio padrão	0.016	0.019	0.02		

Embora o artigo [Mendoza et al., 2012] não tenha apresentado os valores de PPV, é interessante observar o que acontece com os valores de PPV para os métodos BEIM, PA e PB em um cenário com um volume maior de amostras (neste caso, 300 amostras). A

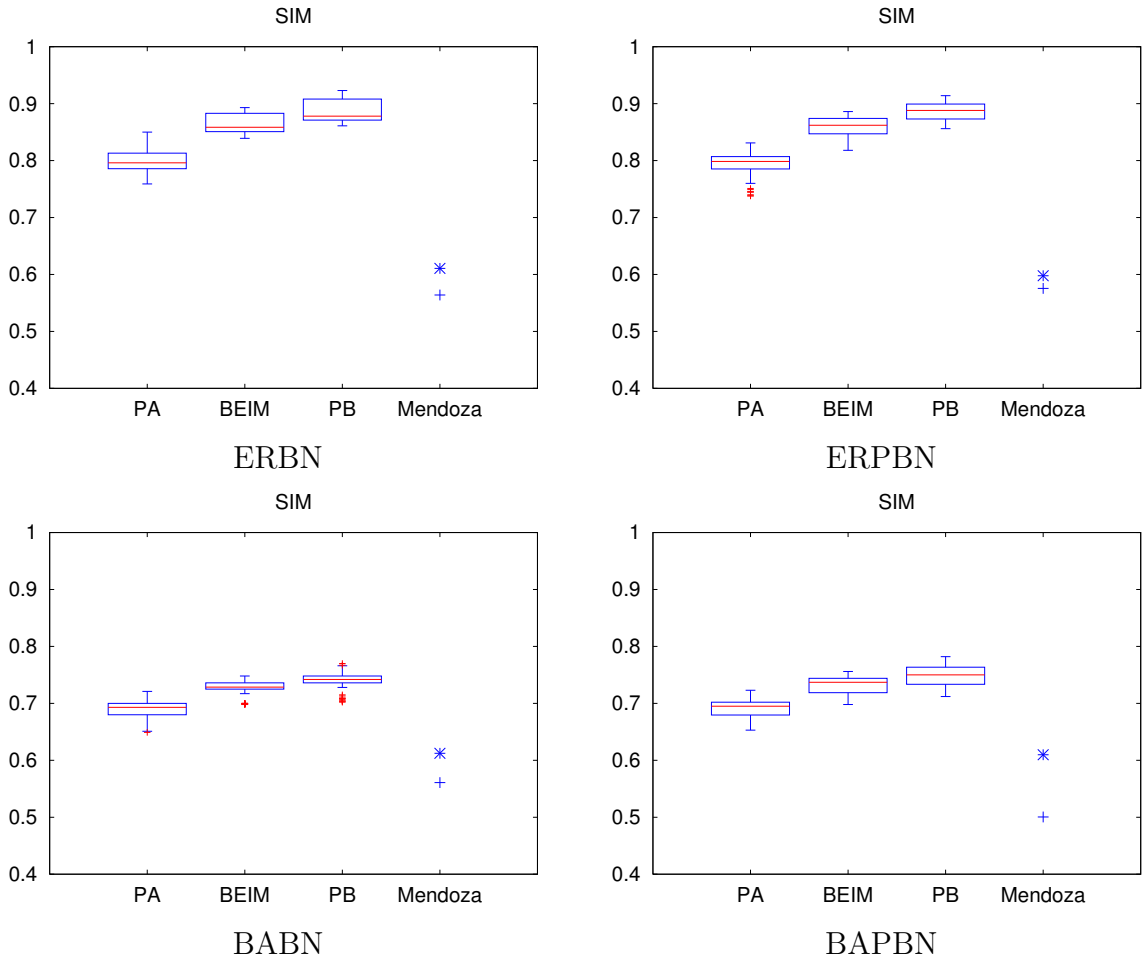


Figura 4.12: *Boxplots* correspondentes a valores de SIM obtidos para os métodos PA,PB,BEIM e Mendoza aplicados sobre os tipos de rede ERBN, ERPBN, BABN e BAPBN. Os *boxplots* de PA, PB e BEIM contêm 1000 valores de SIM (10 redes gabarito gerando 10 conjuntos de amostras, sendo 10 execuções por conjunto). No caso do método Mendoza, há apenas dois pontos ilustrados por $*$ e $+$, correspondendo aos resultados obtidos para $k_{max} = 2$ e $k_{max} = 3$ respectivamente.

Figura 4.13 mostra os *boxplots* dos resultados de PPV obtidos para os métodos PA, PB e BEIM. Os boxplots representam 10 redes gabarito, cada uma resultando em 10 conjuntos de amostras, sendo 10 execuções por conjunto de amostras (total de 1000 valores de PPV). A Tabela 4.7 mostra as médias e os desvios padrões dos *boxplots* apresentados nas Figura 4.13. O melhor método em todos os casos foi o método proposto com população inicial inferida por BEIM (PB), enquanto o pior método foi o BEIM. Assim, para um maior número de amostras, o método PB supera os métodos BEIM e PA tanto em termos de similaridade como em termos de PPV.

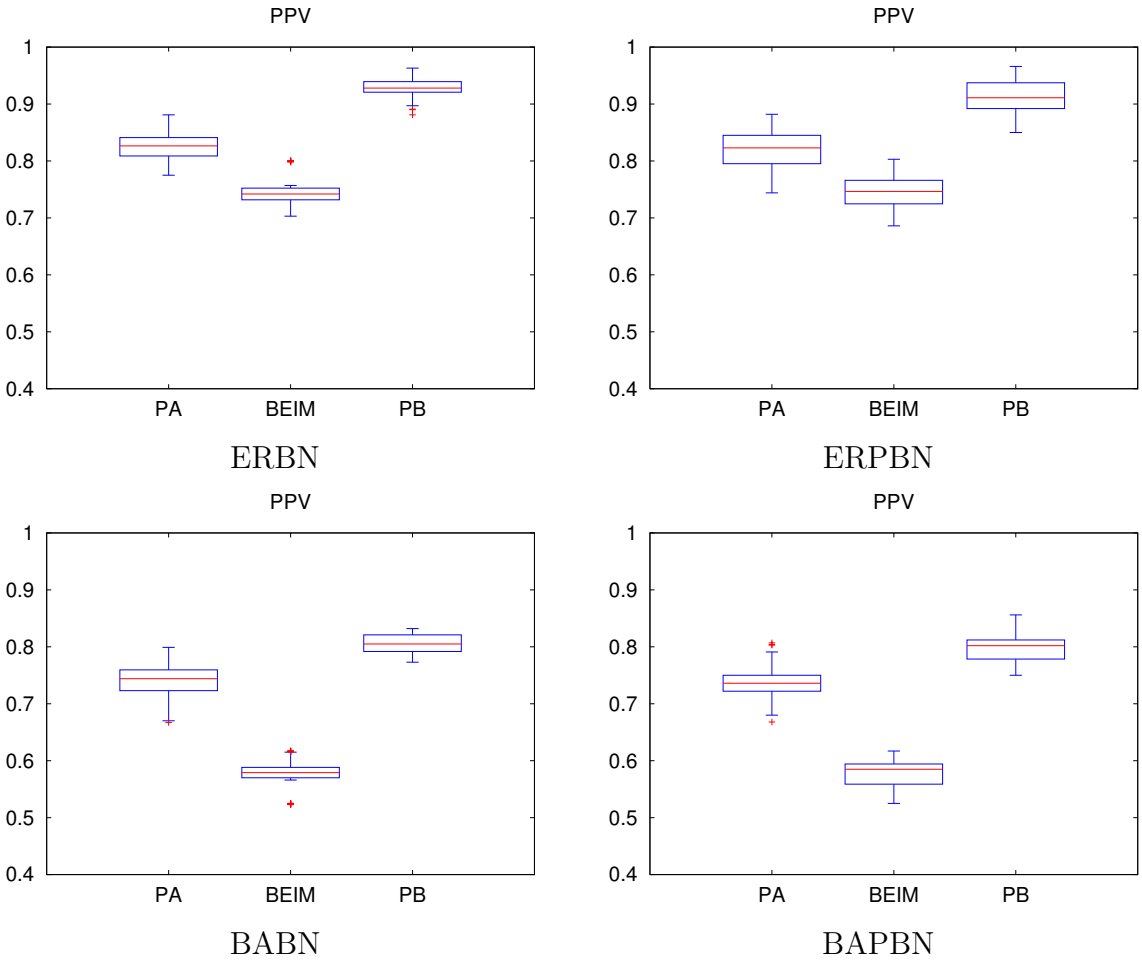


Figura 4.13: *Boxplots* correspondentes a valores de *PPV* obtidos para os métodos PA, PB e BEIM aplicados sobre os tipos de rede ERBN, ERPBN, BABN e BAPBN. Os *boxplots* de PA, PB e BEIM contêm 1000 valores de *PPV* (10 redes gabarito gerando 10 conjuntos de amostras, sendo 10 execuções por conjunto).

Tabela 4.7: Médias e desvios padrões dos *boxplots* apresentados na Figura 4.13.

		PPV		
		300 amostras (concatenação de 10 conjuntos de 30)		
		PA	BEIM	PB
ERBN	média	0.825	0.743	0.928
	desvio padrão	0.023	0.024	0.016
ERPBN	média	0.818	0.746	0.913
	desvio padrão	0.038	0.034	0.029
BABN	média	0.74	0.578	0.805
	desvio padrão	0.03	0.023	0.016
BAPBN	média	0.736	0.576	0.799
	desvio padrão	0.028	0.028	0.026

Capítulo 5

Conclusão

Neste trabalho, foi proposta uma abordagem baseada em algoritmos genéticos para inferir redes de interação gênica modeladas por redes Boolenas e redes Booleanas probabilísticas. As principais vantagens dessa proposta em comparação com técnicas similares para o mesmo propósito são: (i) aplicação de diversos algoritmos genéticos independentes, sendo um para cada gene alvo; (ii) a geração das populações iniciais pré-inferidas com base em um método de inferência de redes gênicas como a busca exaustiva por subconjuntos de tamanho (grau) fixo orientada pela informação mútua [Barrera et al., 2007] (BEIM); (iii) o uso do critério de informação Akaike (AIC) como função de aptidão para orientar o algoritmo, o qual basicamente oferece uma medida da probabilidade das amostras de expressão gênica a serem geradas pela rede de acordo com sua topologia e regras lógicas, incluindo um parâmetro que penaliza topologias com um número excessivo de arestas (o que implicaria em um aumento expressivo do número de parâmetros estatísticos a serem estimados) dadas as amostras disponíveis, evitando um super-ajuste aos dados (*overfitting*).

Os experimentos realizados com base em dados gerados por redes gênicas artificais construídas pelos modelos de redes complexas de Erdös-Rényi (aleatório) e Barabási-Albert (livres de escala) indicam que as três estratégias apontadas anteriormente funcionando em conjunto produzem resultados superiores, em termos de similaridade, aos obtidos pelo método de algoritmo genético recém-publicado por Mendoza *et al* [Mendoza et al., 2012]. Os resultados se mantiveram superiores mesmo sem a aplicação da segunda estratégia (geração aleatória das populações iniciais, sem aplicar um método de inferência). Além disso, a geração das populações iniciais pelo método de inferência BEIM apresentou benefícios em relação à geração aleatória, especialmente para um volume maior de amostras.

Com relação à função de aptidão adotada, os resultados experimentais mostraram que o AIC exibe um bom balanço entre a captura da complexidade e a qualidade de ajuste em situações com um número muito limitado de amostras, efetivamente orientando os

algoritmos genéticos em busca de redes que apresentam boas similaridades topológicas com o gabarito. Em todos os cenários apresentados, aplicando os algoritmos genéticos a partir de populações iniciais aleatórias, as correlações absolutas médias entre o AIC e as métricas topológicas de similaridade (SIM) e valor preditivo positivo (PPV) foram superiores a $0,9$. Além disso, os algoritmos convergem rapidamente, necessitando de cerca de 200 iterações (gerações) para obter as redes com os menores valores de (AIC) e, consequentemente, os maiores valores de SIM e PPV.

O AIC ainda pode ser adaptado para que seja possível controlar sua penalização de modo que os algoritmos possam incluir um número maior o menor de arestas nas redes inferidas. Uma maneira de fazer isso é conferir uma constante multiplicativa ao fator K, de forma que se essa constante estiver bem calibrada, resultados melhores poderiam ser alcançados.

Embora o método proposto tenha se mostrado promissor de acordo com os resultados em termos de similaridade topológica, a validação do método pode ser aprofundada levando-se em conta não apenas características topológicass, mas também através da análise de aspectos dinâmicos dos sinais produzidos pelas redes inferidas, comparando-os com o sinal produzido pela rede gabarito [Dougherty, 2011]. Assim, é possível que topologias mais simples (menor número médio de arestas) possam explicar a dinâmica das amostras produzidas pela rede gabarito de maneira quase tão desejável quanto as topologias mais complexas (maior número médio de arestas).

Os aspectos topológicos conhecidos sobre as redes biológicas podem servir como uma informação *a priori* valiosa para aperfeiçoar os métodos de inferência de redes. Por exemplo, Lopes *et al* [Lopes et al., 2014] apresentou um método de seleção de características orientado a obter redes gênicas com topologia livre de escala. No caso do método de algoritmos genéticos proposto neste trabalho, algo similar pode ser feito, por exemplo, em relação à estratégia de cruzamento. Na recombinação, ao invés de sortear um ponto de corte com distribuição de probabilidades uniforme, fazendo com que em boa parte dos casos a união dos cromossomos dos pais seja dividido próximo do ponto médio, resultando em dois cromossomos filhos de tamanhos aproximadamente iguais, pode-se atribuir uma distribuição que tenda a privilegiar os cortes nas pontas, forçando assim que muitos cromossomos tenham grau pequeno, enquanto alguns tenham grau alto. Isso pode fazer com que a rede final possua características de redes livres de escala.

Finalmente, uma vantagem computacional do método proposto é que ele pode ser facilmente paralelizável, já que cada algoritmo genético (um por gene alvo) pode ser executado de maneira independente dos demais. Assim, é possível fazer com que cada núcleo de processamento seja encarregado de obter o melhor subconjunto de preditores para um gene alvo. Dessa forma, redes de tamanho real (milhares de genes) podem ser inferidas rapidamente. Atualmente existem arquiteturas paralelas de baixo custo como as

unidades gráficas de processamento (Graphical Processing Units - GPU), que vêm sendo cada vez mais empregadas para propósito geral em aplicações científicas. Borelli *et al* implementou o método BEIM discutido aqui em GPUs, tendo obtido *speedups* da ordem de centenas quando comparadas a *multicore* CPUs convencionais [Borelli et al., 2013].

Referências Bibliográficas

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723. 18, 46

[Albert, 2005] Albert, R. (2005). Scale-free networks in cell biology. *J Cell Sci*, 118(21):4947–4957. 29, 30

[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. 23

[Angeletti et al., 2001] Angeletti, M., Culmone, R., and Merelli, E. (2001). An intelligent agents architecture for dna-microarray data integration. Technical report, U. of Camerino, Italy. 16

[Barabási, 2009] Barabási, A.-L. (2009). Scale-Free Networks: A Decade and Beyond. *Science*, 325(5939):412–413. 55

[Barabási and Albert, 1999] Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512. 18, 30, 31

[Barabasi et al., 2011] Barabasi, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nat Rev Genet*, 12(1):56–68. 28, 29, 30

[Barrera et al., 2007] Barrera, J., Cesar-Jr, R. M., Martins-Jr, D. C., Vencio, R. Z. N., Merino, E. F., Yamamoto, M. M., Leonardi, F. G., Pereira, C. A. 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*, chapter 2, pages 11–26. Springer. 18, 22, 23, 24, 26, 28, 29, 42, 45, 72

[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). 74

[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. 28

[Burnham and Anderson, 2002] Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, 2nd edition. 18

[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*. 18

[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. 29, 30

[Costa et al., 2007] Costa, L. F., Rodrigues, F. A., Travieso, G., and Villas-Boas, P. R. (2007). Characterization of complex networks: a survey of measurements. *Advances in Physics*, 56(1):167–242. 29, 30, 31

[Crick, 1970] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563. 16

[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. 23

[Davidson, 2010] Davidson, C. (2010). Identifying gene regulatory networks using evolutionary algorithms. *Journal of Computing Sciences in Colleges*, 25(5):231–237. 40

[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. 16

[Dougherty, 2011] Dougherty, E. R. (2011). Validation of gene regulatory networks: scientific and inferential. *Briefings in Bioinformatics*, 12(3):245–252. 56, 73

[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*. 28

[Erdös and Rényi, 1959] Erdös, P. and Rényi, A. (1959). On random graphs. *Publ. Math. Debrecen*, 6:290–297. 18, 30, 55

[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. 23

[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. 29, 30

[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. 23

[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. 23

[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. 30

[Goldberg, 1989] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley. 17, 31

[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. 29, 30

[Haupt and Haupt, 2004] Haupt, R. L. and Haupt, S. E. (2004). *Practical Genetic Algorithms*. John Wiley & Sons, 2nd edition. 17

[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. 16

[Holland, 1992] Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems*. MIT Press. 17

[Husmeier, 2003] Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, 19:2271–2282. 28

[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. 24

[Jain and Zongker, 1997] Jain, A. K. and Zongker, D. (1997). Feature-selection: Evaluation, application, and small sample performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(2):152–157. 25

[Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 467:651–654. 29, 30

[Jong, 2002] Jong, H. D. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103. 22

[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. 23

[Kauffman, 1969] Kauffman, S. A. (1969). Homeostasis and differentiation in random genetic control networks. *Nature*, 224(215):177–178. 22, 23

[Kauffman, 1993] Kauffman, S. A. (1993). *The Origins of Order*. Oxford University Press. 29

[Kelemen et al., 2008] Kelemen, A., Abraham, A., and Chen, Y. (2008). *Computational Intelligence in Bioinformatics*. Springer. 16, 23

[Knabe et al., 2010] Knabe, J. F., Wegner, K., Nehaniv, C. L., and Schilstra, M. J. (2010). Genetic algorithms and their application to in silico evolution of genetic regulatory networks. *Methods Mol. Biol.*, 673:297–321. 40

[Larranaga et al., 1996] Larranaga, P., Poza, M., Yurramendi, Y., Murga, R. H., and Kuijpers, C. M. H. (1996). Structure learning of bayesian networks by genetic algorithms: A performance analysis of control parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):912–926. 40

[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. 23, 24

[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. 23

[Linden, 2012] Linden, R. (2012). *Algoritmos Genéticos*. Editora Ciência Moderna, 3rd edition. 10, 32, 33, 35, 46, 47

[Lopes, 2011] Lopes, F. M. (2011). *Redes complexas de expressão gênica: síntese, identificação, análise e aplicações*. PhD thesis, Instituto de Matemática e Estatística - Universidade de São Paulo, Rua do Matão, 1010. 29, 30, 31

[Lopes et al., 2010] Lopes, F. M., Martins-Jr, D. C., Barrera, J., and Cesar-Jr, R. M. (2010). SFFS-MR: a floating search strategy for grns inference. In *Pattern Recognition in Bioinformatics, Proceedings*, volume 6282 of *Lecture Notes in Computer Science*, pages 407–418, Nijmegen, Netherlands. Springer Berlin / Heidelberg. 26

[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*. (in press). 29, 30, 73

[Lopes et al., 2008a] Lopes, F. M., Martins-Jr, D. C., and Cesar-Jr, R. M. (2008a). Feature selection environment for genomic applications. *BMC Bioinformatics*, 9(451). 26

[Lopes et al., 2008b] Lopes, F. M., Martins-Jr, D. C., and Cesar-Jr, R. M. (2008b). Feature selection environment for genomic applications. *BMC Bioinformatics*, 9(451). 29

[Lopes et al., 2009] Lopes, F. M., Martins-Jr, D. C., and Cesar-Jr, R. M. (2009). Comparative study of grn's inference methods based on feature selection by mutual information. In *IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Minneapolis,MN,USA. 26, 29

[Lopes et al., 2011] Lopes, F. M., Oliveira, E. A., and Cesar-Jr, R. M. (2011). Inference of gene regulatory networks from time series by tsallis entropy. *BMC Systems Biology*, 5:61. 26

[Mamakou et al., 2005] Mamakou, M. E., Sirakoulis, G. C., Andreadis, I., and Karafyllidis, I. (2005). Adaptive reverse engineering of gene regulatory networks using genetic algorithms. In *IEEE International Conference on EUROCON*, pages 401–404, Belgrade. 40

[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. 16

[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. 10, 22, 25

[Martins-Jr et al., 2010] Martins-Jr, D. C., Oliveira, E. A., Louzada, V. H., and Hashimoto, R. F. (2010). Inference of restricted stochastic boolean grn's by bayesian error and entropy based criteria. In *15th Iberoamerican Congress on Pattern Recognition (CIARP)*, volume 6419 of *Lecture Notes in Computer Science*, pages 144–152. Springer-Verlag. 26

[McCluskey, 1956] McCluskey, E. J. (1956). Minimization of boolean functions. *Bell Syst Tech, J*, 35(5):1417–1444. 55

[Mendoza et al., 2012] Mendoza, M. R., Lopes, F. M., and Bazzan, A. L. C. (2012). Reverse engineering of grns: An evolutionary approach based on the tsallis entropy. In *Proceedings of the 14th international conference on Genetic and evolutionary computation (GECCO)*, pages 185–192, Philadelphia. 18, 20, 40, 43, 45, 54, 67, 69, 72

[Mitchell, 1996] Mitchell, M. (1996). *An introduction to Genetic Algorithms*. MIT Press, Cambridge, MA. 31

[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. 25

[Narasimhan et al., 2009] Narasimhan, 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. 29

[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. 24

[Przulj et al., 2004] Przulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515. 29, 30

[Pudil et al., 1994] Pudil, P., NovovicovÃ¡, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125. 25

[Randy L. Haupt, 2004] Randy L. Haupt, S. E. H. (2004). *Practical genetic algorithms 2nd ed.* Wiley-Interscience, 111 River Street, Hooboken,NJ 07030. 32, 33, 35, 36

[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. 10, 26

[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. 23

[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. 16, 21

[Shin and Iba, 2003] Shin, A. and Iba, H. (2003). Construction of genetic network using evolutionary algorithm and combined fitness function. *Genome Informatics*, 14:94–103. 40

[Shmulevich and Dougherty, 2007] Shmulevich, I. and Dougherty, E. R. (2007). *Genomic Signal Processing*. Princeton University Press, New Jersey. 16, 21, 22, 23

[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. 22, 23, 28

[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. 15

[Somol et al., 1999] Somol, P., Pudil, P., NovovicovÃ¡, J., and PaclÃk, P. (1999). Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, 20:1157–1163. 25

[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. 24

[Swain et al., 2005] Swain, M., Hunniford, T., Dubitsky, W., Mandel, J., and Palfreyman, N. (2005). Reverse-engineering gene-regulatory networks using evolutionary algorithms and grid computing. *Journal of Clinical Monitoring and Computing*, 19:329–337. 40

[Theodoridis and Koutroumbas, 1999] Theodoridis, S. and Koutroumbas, K. (1999). *Pattern Recognition*. Academic Press, USA, 1st edition. 24

[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. 16

[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. 16, 21

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393:440–442. 30

[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. 23, 24