UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

MARCELO MASSASHI KURODA

UMA ABORDAGEM BASEADA EM COMPUTAÇÃO EVOLUTIVA APLICADO NA INFERÊNCIA DE REDES DE REGULAÇÃO GÊNICA

TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO

2015

MARCELO MASSASHI KURODA

UMA ABORDAGEM BASEADA EM COMPUTAÇÃO EVOLUTIVA APLICADO NA INFERÊNCIA DE REDES DE REGULAÇÃO GÊNICA

Trabalho de Conclusão de Curso apresentada ao Curso Superior de Tecnologia em Analise e Desenvolvimento de Sistemas da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do grau de Tecnólogo em Análise e Desenvolvimento de Sistemas

Orientador: Prof. Dr. Danilo Sipoli Sanches

CORNÉLIO PROCÓPIO

2015

RESUMO

KURODA, Marcelo M.. UMA ABORDAGEM BASEADA EM COMPUTAÇÃO EVOLUTIVA APLICADO NA INFERÊNCIA DE REDES DE REGULAÇÃO GÊNICA. 45 f. Trabalho de Conclusão de Curso – Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2015.

Um organismo vivo pode ser visto como uma rede de moléculas conectadas por reações bioquímicas que possui um complexo sistema de envio e recebimento de sinais que realizam o controle celular. Estudos vêm sendo realizados para entender os mecanismos de controle e os relacionamentos destas reações. Estes mecanismos podem ser entendidos observando a evolução temporal dos níveis de expressão gênica, que são estimados utilizando métodos de extração de informação molecular. A partir destes dados de expressão gênica é possível inferir redes de regulação gênica. Existem muitos métodos para inferir redes de regulação gênica na literatura, o método adotado neste trabalho é o método de seleção de características, composto por um algoritmo de busca e uma função critério. A função critério utilizada é baseada na entropia condicional média e o adotado é o algoritmo genético. Devido ao grande número de genes e os poucos experimentos produzidos, a inferência de redes de regulação gênica é um desafio em aberto na bioinformática. A descoberta dos relacionamentos entre os genes permite entender e analisar doenças, contribuindo na produção de medicamentos mais eficazes contra doenças e em tratamentos. Para a inferência de redes de regulação gênica foram desenvolvidos dois algoritmos genéticos, um com representação dos cromossomos por genes e o outro com representação dos cromossomos por redes. Para a validação das redes inferidas foi utilizado redes gênicas artificiais. O algoritmo genético com representação dos cromossomos por redes apresentou um maior similaridade entre as redes inferidas e as redes gênicas artificiais, porém o algoritmo genético com representação dos cromossomos por genes obteve um tempo computacional muito melhor, dessa forma o algoritmo genético com representação do cromossomo por genes possui um melhor custo-benefício, já que a diferença na similaridade entre as redes inferidas pelos dois algoritmos foi pequena.

Palavras-chave: Inferência, Redes de Regulação Gênica, Entropia, Algoritmo Genético, Bioinformática.

ABSTRACT

KURODA, Marcelo M.. AN APPROACH BASED ON EVOLUTIONARY COMPUTATION APPLIED TO THE INFERENCE OF GENE REGULATORY NETWORKS. 45 f. Trabalho de Conclusão de Curso – Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2015.

A living organism can be seen as a network of molecules connected by biochemical reactions that has a complex system of sending and receiving signals that perform cellular control. Studies are being conducted to understand the mechanisms of control and the relationships of these reactions. They can be understood by observing the time course gene expression levels, which are estimated using methods of molecular information extraction. From these gene expression data it is possible to infer genetic regulatory networks. There are many methods to infer genetic regulatory networks in the literature, the method adopted in this work is the method of selection of features, composed of a search algorithm and a criterion function. The function criterion used is based on the average conditional entropy and the adopted is the genetic algorithm. Due to the large number of genes and the few experiments produced, the inference of genetic regulatory networks is an open challenge on bioinformatics. The discovery of relationships between the genes allows you to understand and analyze diseases, contributing to the production of more effective drugs against diseases and treatments. For inference of genetic regulatory networks have been developed two genetic algorithms, a chromosome representation for genes and the other with representation of chromosomes by networks. To validate the inferred networks used artificial gene networks. Genetic algorithm with representation of chromosomes by networks presented a greater similarity between the inferred networks and artificial gene networks, but the genetic algorithm with representation of chromosomes for genes obtained a much better computational time, thus the genetic algorithm representation of chromosome for genes has a better cost-benefit, since the difference in similarity between networks inferred by the two algorithms has been small.

Keywords: Inference, Gene Regulatory Networks, Entropy, Genetic Algorithm, Bioinformatics.

LISTA DE FIGURAS

LISTA DE SIGLAS

RNA	Ácido Ribonucleico (*Ribonucleic Acid*)
GRNs	Redes de Regulação Gênica (*Gene Regulatory Networks*)
AGNs	Redes Gênicas Artificais (*Artificial Genetic Networks*)
BNs	Redes Booleanas (*Boolean Networks*)
AG	Algoritmo Genético
ER	Erdös e Rényi
SW	Mundo Pequeno (*Small-World*)
TP	Verdadeiro Positivo (*True Positive*)
FP	Falso Positivo (*False Positive*)
FN	Falso Negativo (*False Negative*)
TN	Verdadeiro Negativo (*True Negative*)
PPV	Valor Preditivo Positivo (*Positive Predictive Value*)

LISTA DE SÍMBOLOS

k	Constante de Boltzman
$p\imath$	Probabilidade
W	Número de Configurações Microscópicas
$P(x)$	Probabilidade de Ocorrência das Variáveis Aleatórias
$P(x,y)$	Probabilidade de X e Y
$H(Y \mid x)$	Entropia Condicional
$x \in X$	x pertence a X
$H(Y \mid X)$	Entropia Condicional Média
$M(i,j)$	Aresta

SUMÁRIO

1 INTRODUÇÃO

Um organismo pode ser entendido como uma complexa rede de moléculas conectadas por reações bioquímicas, na qual ocorre por meio do envio e recebimento de sinais. Um dos desafios da biologia sistêmica é entender os mecanismos de controle de regulação celular a partir de entidades biológicas como, por exemplo, genes e Ácido Ribonucleico (RNA, do inglês *Ribonucleic Acid*) (SNOEP; WESTERHOFF, 2005). Uma forma para entender melhor estes mecanismos de controle celular é considerando a evolução temporal dos níveis de expressão gênica. Estes níveis podem ser estimados simultaneamente em múltiplos instantes de tempo através da extração de informação molecular como, por exemplo, RNA-seq (WANG et al., 2009).

A inferência ou engenharia reversa, de Rede de Regulação Gênica (GRNs, do inglês *Gene Regulatory Networks* (SHMULEVICH; DOUGHERTY, 2014) a partir de dados de expressão gênica, segundo (NELSON et al., 2008) é fundamentada no dogma central da biologia, onde o funcionamento de um organismo é determinado pela sua expressão gênica. Neste contexto, inferir as GRNs a partir dos dados de expressão gênica é um problema em aberto, devido ao grande número de variáveis (genes) e o pequeno número de experimentos (amostras) existente. Outros fatores que também dificultam a inferência é a falta de informação sobre o organismo de interesse, a alta complexidade das redes e o ruído existente nas medidas de expressão. Desta forma, para inferir estas redes é necessário um grande esforço no desenvolvimento de novas técnicas computacionais e estatísticas.

Por meio de uma GRN inferida, é possível entender características biológicas de organismos como as interações moleculares e funções biológicas. Dessa forma, a inferência destas redes podem, ser empregados em estudos de doenças e prognósticos específicos, assim como medicamentos mais eficazes.

Para validar as redes inferidas é preciso ter conhecimento real sobre as conexões e os relacionamentos funcionais dos genes (DOUGHERTY, 2007). Este conhecimento real segundo LOPES (2011) pode ser obtido por meio das Redes Gênicas Artificais (AGNs do inglês *Artificial*

Genetic Networks), inferidas com base no método de Redes Booleanas (BNs, do inglês *Boolean Networks*).

Neste contexto, são utilizados neste trabalho para inferir GRNs, o método de seleção de características (JAIN et al., 2000), constituído por duas partes: um algoritmo de busca e uma função critério, o algoritmo de busca utilizado é o Algoritmo Genético (AG) e a função critério utilizada é baseada na entropia condicional média. Assim, a partir de dados de expressão gênica simulados são inferidas GRNs utilizando o AG, logo, para validar estas redes inferidas pelo AG, é utilizado uma AGN.

1.1 PROBLEMA

Um dos principais problemas encontrados é a identificação dos relacionamentos entre os genes, considerando o grande volume de genes e os poucos experimentos na área, conhecido como a maldição da dimensionalidade (JAIN et al., 2000) (BISHOP et al., 1995). Alguns fatores dificultam a identificação desses relacionamentos, são eles: a complexidade que uma rede possui, a falta de conhecimento sobre o organismo de interesse e o ruído intrínseco das medidas de expressão.

Técnicas computacionais e estatísticas estão sendo aplicadas na inferência de GRNs, visando uma maior eficiência no tempo computacional e na acurácia das redes inferidas.

1.2 JUSTIFICATIVA

A área da bioinformática tem sido alvo de diversos estudos, visto a dificuldade na identificação de redes gênicas, devido ao grande número de genes e os poucos experimentos produzidos, ou seja, dezenas de experimentos com milhares de genes (alta dimensionalidade). Para inferir estas redes com maior precisão, métodos computacionais e estatísticos veem sendo desenvolvidos que buscam encontrar o melhor resultado, sem percorrer todo o espaço de busca, com a finalidade diminuir o tempo de computacional (LOPES, 2011).

Neste contexto, exitem basicamente duas abordagens utilizadas para a redução da dimensionalidade, o método de extração de características, onde características são criadas a partir de transformações e combinações de características; e o método de seleção de características, onde através da função critério escolhida, procuram por um subconjunto de características que levam a uma boa representação, classificação ou predição (CAMPOS, 2001) (WEBB, 2003).

Os algoritmos de busca são divididos basicamente em duas categorias: ótimos e sub-

ótimos. Os algoritmos ótimos são aqueles que realizam buscas exaustivas, retornando a melhor solução, no entanto, possui um custo computacional elevado, principalmente em problemas que possui alta dimensionalidade como no caso da inferência de GRNs. Os algoritmos sub-ótimos são mais adequados para resolver este tipo problema, mesmo não trazendo soluções ótimas, eles apresenta um bom custo-benefício em relação ao desempenho computacional e a qualidade da solução apresentada (LOPES, 2011).

Neste trabalho é utilizado para a inferência de GRNs, algoritmos de seleção de características sub-ótimos, onde tem sido aplicado em vários trabalhos como (BRAGA, 1998) (FILHO; POPPI, 1999) (JONG, 1975) (MICHALEWICZ, 1996), neste trabalho foram utilizados duas abordagens baseadas na computação evolutiva, uma com representação dos cromossomos por redes, e outra com representação dos cromossomos por genes.

A importância de se inferir GRNs, está no conhecimento que pode ser obtido a partir da estrutura de uma rede inferida, ou seja, o conhecimento dos genes que regulam um determinado gene alvo, e dos genes que são regulado por este mesmo gene alvo, onde este gene alvo pode ser um possível causador de uma doença. Dessa maneira, este conhecimento adquirido, pode ser aplicado em diversas análises como, do comportamento das expressões de doenças, logo, realizar estudos para o desenvolvimento de drogas mais eficazes e para tratamentos de doenças(LOPES, 2011).

1.3 OBJETIVOS

1.3.1 OBJETIVO GERAL

O objetivo geral deste trabalho consiste no desenvolvimento e análise de duas abordagens baseadas na computação evolutiva aplicados na inferência de GRNs.

1.3.2 OBJETIVOS ESPECÍFICOS

Para que o objetivo geral fosse realizado foram necessários alguns objetivos específicos:

- Desenvolvimento dos algoritmos evolutivos (AG) para inferência de GRNs.
- Inferência de GRNs a partir dos dados de expressão gênica, aplicando os algoritmos genéticos desenvolvidos.
- Validação do tempo computacional e da acurácia dos dois algoritmos genéticos, utilizando o tempo de execução de cada algoritmo e a AGN respectivamente.

- Comparação dos resultados de tempo computacional e acurácia das redes inferidas pelos dois algoritmos genéticos.

1.4 ORGANIZAÇÃO DO TEXTO

Este trabalho é composto de seis capítulos, neste primeiro capítulo é apresentado á introdução, o problema, os objetivos e a justificativa do trabalho. No capítulo 2 é apresentada a fundamentação teórica do trabalho. No capítulo 3 são apresentados os materiais e métodos utilizados neste trabalho. No capítulo 4 é apresentada os resultados. No capítulo 5 é apresentado a conclusão e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

O gene abrange a região da molécula de DNA, e seu segmento é composto por uma instrução gênica codificada através de bases nitrogenada como, por exemplo, adenina, guanina, citosina e timina, onde por meio da expressão transcrita, ou seja, formação de moléculas de RNA, coordena de maneira indireta a síntese de uma proteína (DANTAS, 2015).

Um conjunto de genes forma uma unidade cromossômica, na qual realiza atividade funcional quando o material encontra-se na forma filamentosa, ou seja, não compactada no momento em que sua expressão regulada por outros genes é liberada . Dessa maneira, quando a célula está se organizando para a divisão celular, e atingem o grau máximo de compactação, o material genético é denominado como cromossomo (DANTAS, 2015).

2.1 REDES DE REGULAÇÃO GÊNICA

O mecanismo de controle de regulação do ciclo celular é um sistema complexo de envio e recebimento de sinais, onde muitas vezes ocorre por meio de ligações de realimentação. Estes sinais são proteínas produzidas pela expressão gênica (após a transcrição e a tradução da Figura 1) que formam complexos multi-protéicos que controlam a atividade celular (vias metabólicas) através da interação com outros complexos multi-protéicos, tanto internos como externos à célula. Os complexos multi-protéicos recebem sinais de volta das vias metabólicas controladas (Figura 1 setas 3 e 4), e enviam sinais de realimentação para os níveis de transcrição e tradução (Figura 1 setas 1 e 2). Estes sinais de realimentação modificam os padrões futuros de expressão dos genes. Desta maneira, os genes e os produtos gerados pela sua expressão formam uma rede de sinalização responsável pelo controle das funções celular, pelo ciclo de divisão celular e pela morte celular programada. A Figura 1 representa esquematicamente este tipo de sistema, onde normalmente é chamado de Rede de Regulação Gênica (TREPODE, 2007).

O nível de expressão dos genes de uma GRN depende de estímulos externos e dos valores de expressão próprios e de outros genes, em momentos anteriores de tempo (TREPODE, 2007). Estes níveis de expressão podem ser medidos utilizando métodos de extração molecular

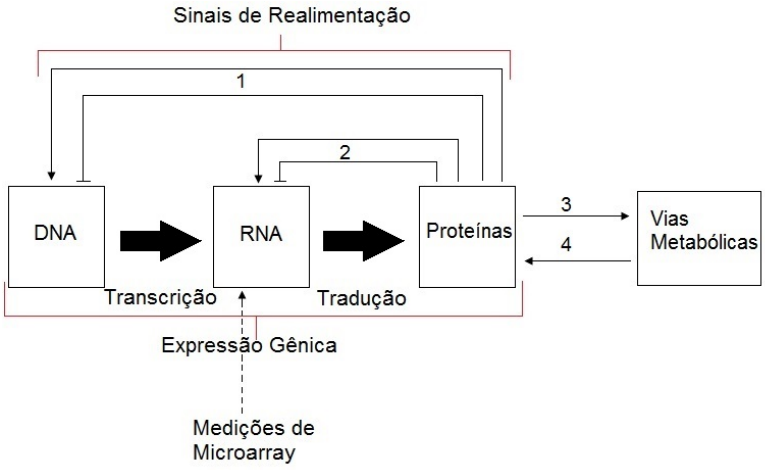


Figura 1: Rede de Regulação Gênica.

Fonte: (TREPODE, 2007)

como o RNA-seq citado anteriormente. Desta forma, a partir dos dados de níveis de expressão gênica, é possível modelar as GRNs utilizando métodos de inferência.

2.2 INFERÊNCIA DE REDES DE REGULAÇÃO GÊNICA

O objetivo de inferir GRNs utilizando os dados de expressão gênica é encontrar interações entre genes e apresentar redes importantes para a área da bioinformática. Devido ao grande número de variáveis (genes) e as poucas amostras (medidas) existentes, um dos desafios da bioinformática é a inferência de GRNs utilizando dados de expressão gênica, também conhecido como engenharia reversa. A inferência destas redes gênicas possibilita indicar diferentes vias regulatórias, ciclo celular e o mapeamento de alterações provocadas por estímulos.

Para inferir GRNs existem basicamente três funções critério que podem ser utilizadas (LOPES, 2011), são elas:

- Correlação de Pearson, onde os genes possuem relacionamentos 1-para-1. Se a correlação entre os perfis de expressão for maior que o limiar os genes são relacionados.
- Função critério baseada no erro Bayesiano, onde utiliza o coeficiente de determinação como critério para inferir GRNs. Os relacionamentos entre os genes são N-para-1.
- Função critério baseada na teoria da informação. Esta função define relacionamentos tanto 1-para-1 quanto N-para-1, nas quais as distribuições da probabilidade condicional de um gene são uniformes.

Métodos computacionais e estatísticos veem sendo desenvolvido e aplicado para inferir GRNs, buscando maior acurácia e melhor desempenho computacional.

2.3 TAXONOMIA DOS ALGORITMOS DE SELEÇÃO DE CARACTERÍSTICAS

A taxonomia dos algoritmos de seleção de características disponíveis é apresentada na Figura 2. Inicialmente os métodos são separados naqueles baseados em técnicas de reconhecimento de padrões estatísticos e os que utilizam redes neurais artificiais. A categoria de reconhecimento de padrões estatísticos é então dividido em aqueles para encontrar a solução ideal e aqueles que podem resultar em um conjunto de recursos abaixo do ideal. Os métodos abaixo do ideal são divididos em aqueles que armazenam apenas um subconjunto de características, contra aqueles que possui uma população de subconjuntos de características. Outra é feita distinção entre algoritmos que são deterministas, produzindo o mesmo subconjunto de um determinado problema cada vez, e aqueles que têm um elemento aleatório que pode produzir diferentes subconjuntos em cada interação (JAIN; ZONGKER, 1997).

O AG está classificado como um método estocástico de múltiplas soluções, ou seja, a cada interação do algoritmo é gerado subconjuntos diferentes com mais de uma única solução (JAIN; ZONGKER, 1997), o AG é descrito mais detalhadamente na seção (2.3.1).

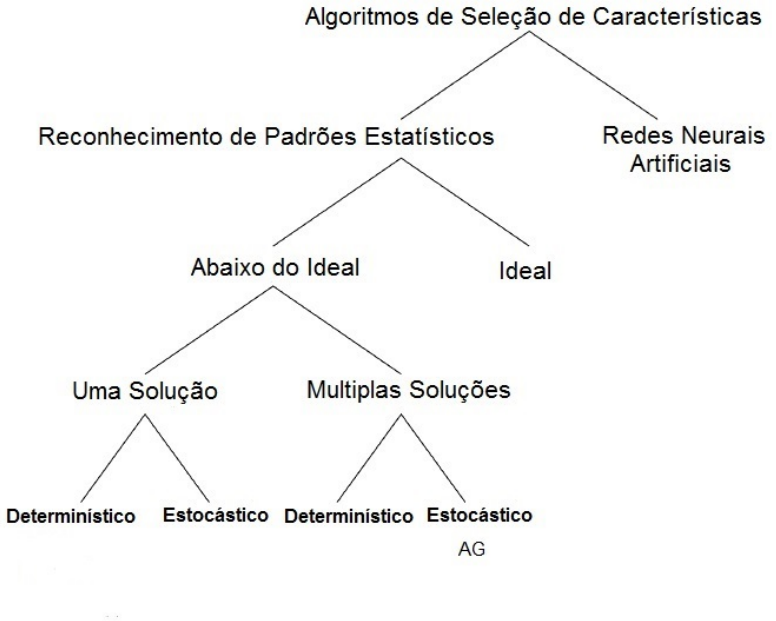


Figura 2: Taxonomia dos algoritmos de seleção de características
Fonte: (JAIN; ZONGKER, 1997)

2.3.1 ALGORITMO GENÉTICO

O AG foi criado por John Holland (1975) e popularizados por David Goldberg (GOLBERG, 1989). O AG é fundamentado no princípio da seleção natural e sobrevivência do mais apto, do naturalista e fisiologista Charles Darwin (DARWIN; BYNUM, 2009), onde o indivíduo que melhor se adaptar em seu meio ambiente, terá maior chance de sobreviver e gerar descendentes. Já aqueles menos aptos irão desaparecer.

O AG é um método de busca e otimização, que consiste em encontrar a melhor solução para um determinado problema, ou seja, tentar várias soluções e a partir da informação obtida destas soluções encontrar soluções cada vez melhor. Geralmente os métodos de busca e otimização apresentam um espaço de busca e uma função objetivo, que são respectivamente, as possíveis soluções para um determinado problema e um método de avaliação, que atribui uma nota as soluções produzidas. Estes métodos permitem realizar buscas de soluções em diferentes regiões do espaço de busca utilizando os indivíduos adequados (LACERDA; CARVALHO, 1999).

Sabendo que AG é um método de busca e otimização, sua população a cada evolução (geração) ela é atualizada, ou seja, o processo do AG ocorre de forma evolutiva. Este processo evolutivo do AG pode ser visto na Figura 3, onde uma população inicial é gerada de forma aleatória e cada indivíduo desta população é avaliado associando uma nota que representa sua adaptabilidade em um determinado ambiente. Em seguida, é realizada a seleção dos indivíduos mais aptos, e após essa seleção, os indivíduos selecionados podem sofrer alterações, através dos operadores de cruzamento (*Crossover*) e mutação gerando descendentes para a próxima geração. Enquanto a solução satisfatória não for encontrada este processo é realizado.

2.3.1.1 INDIVÍDUO

O indivíduo ou cromossomo é uma possível solução para um dado problema a ser otimizado, normalmente sua estrutura é representada por cadeias de valores binários com tamanho fixo. Este conjunto de parâmetros que o cromossomo representa pode ter como resposta da função de avaliação um resultado maximizado ou minimizado. A variação dos parâmetros que um cromossomo assume representa o seu espaço de busca. Na biologia, as características de um indivíduo são chamadas de fenótipo e elas codificadas são chamadas de genótipo (SANCHES, 2010).

A representação utilizando números reais apresenta bom desempenho e a representação binária facilita o uso dos operadores de *Crossover* e mutação. Entretanto, independente da

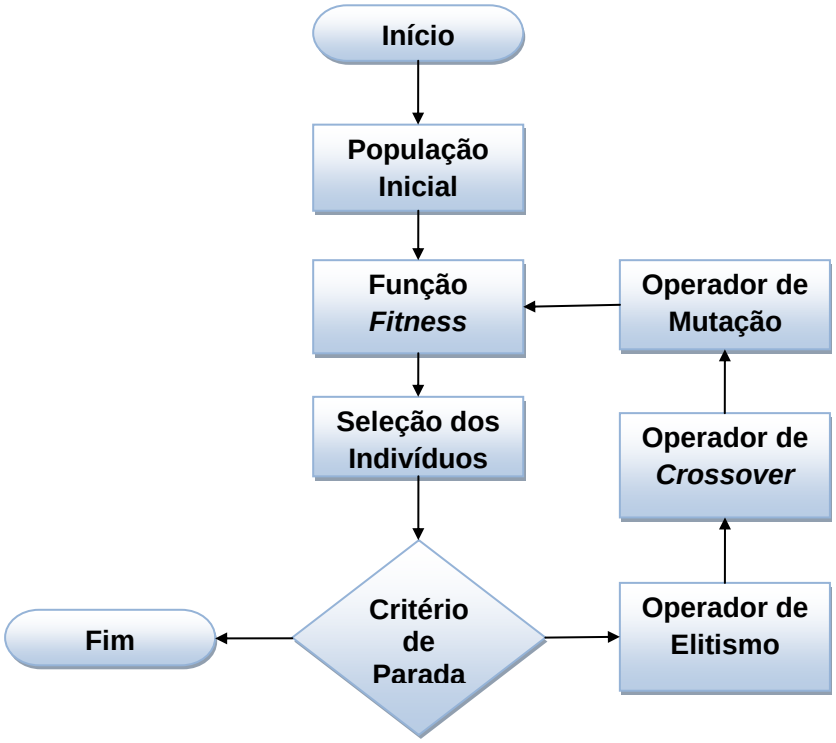


Figura 3: Fluxo de processamento do Algoritmo Genético.

representação adotada, o cromossomo deve ser capaz de representar o espaço de busca que se pretende pesquisar (PACHECO, 1999), um exemplo de indivíduo é apresentado na Figura 4.

0	0	1	0	1	1	1	0	0	0

Figura 4: Estrutura de um indivíduo (binário).

2.3.1.2 POPULAÇÃO

A população é um conjunto de indivíduos candidatos para a solução de um problema, onde serão selecionados os melhores indivíduos para gerar uma nova população de indivíduos.

Segundo (POZO et al., 2005), o número de indivíduos de uma população pode prejudicar o desempenho global e a eficiência dos AGs, ou seja, populações muito pequenas perdem facilmente sua diversidade, pois o espaço de busca é pequeno, não convergindo para uma boa solução, já em populações muito grandes, o algoritmo perde sua eficiência, pois ele terá que percorrer toda população avaliando cada indivíduo a cada interação, até encontrar uma boa solução.

Após um número de gerações, a população tende a um grau de convergência, este grau

de convergência é a variação do *Fitness* de uma população com a média da população anterior, ou seja, é quando cromossomos com aptidão alta surgem em uma população, onde ainda existem cromossomos com aptidão ótima (LACERDA; CARVALHO, 1999). Com isto, o AG busca convergir para um valor ótimo global e evitar que a população tenha uma convergência prematura (mínimo ou máximo local).

2.3.1.3 POPULAÇÃO INICIAL

A geração da população inicial é realizada uma única vez e tem como objetivo gerar os primeiros indivíduos que representam possíveis soluções do problema. Este processo pode ser feito de diversas maneiras, uma delas e gerar aleatoriamente a população inicial (1999).

2.3.1.4 SELEÇÃO

O método de seleção tem como objetivo selecionar os indivíduos para reprodução de descendentes que irão compor a nova população. Os indivíduos mais bem avaliados possuem uma vantagem maior de ser selecionado para reprodução independente do método de seleção escolhido(POZO et al., 2005). No entanto, é necessário variar o *Fitness* dos indivíduos selecionados (THIERENS; GOLDBERG, 1994), não escolhendo apenas os melhores, pois a escolha dos melhores pode acarretar numa convergência prematura dos resultados. Entre os vários métodos de seleção existente (GOLDBERG; DEB, 1991), dois métodos de seleção são descritos a seguir.

Na seleção por Roda da Roleta (*Roulette Wheel Selection*) os indivíduos da população atual são selecionados por meio de uma roleta, onde cada indivíduo da população possui uma aptidão tem maior probabilidade de serem selecionados. A Figura 5 mostra uma roleta onde o tamanho da faixa foi distribuído de acordo com a aptidão do indivíduo. Para selecionar N indivíduos é necessário rodar N vezes a roleta, sendo N o tamanho da população inicial. Em cada giro da roleta o cromossomo que for selecionado pelo marcador será copiado para a população seguinte. A seleção por Roda da Roleta não trabalha com números negativos e pode gerar problemas de convergência prematura (CARVALHO et al., 2003).

Na seleção por Torneio (*Tournament Selection*) Figura 6, a partir da população atual, são selecionados aleatoriamente e com mesma probabilidade um número n de cromossomos que irão compor uma subpopulação. O cromossomo com melhor aptidão é selecionado entre os cromossomos da subpopulação e inserido na população de descendentes. Este processo ocorre enquanto a população de descendentes não estiver completo (SANCHES, 2010).

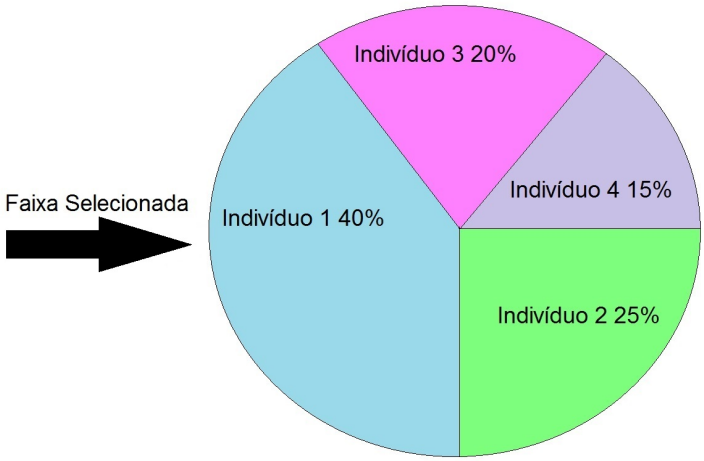


Figura 5: Modelo de seleção por roleta.

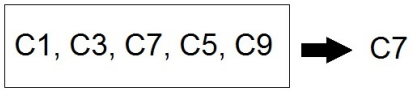


Figura 6: Modelo de seleção por torneio.

2.3.1.5 *CROSSOVER*

O *Crossover* é o operador responsável pelo cruzamento das características dos pais durante a reprodução, permitindo que os descendentes gerados herdem suas características genéticas.

Existem muitas técnicas de *Crossover*, entre elas estão o cruzamento de um ponto e o cruzamento de N-pontos. No *Crossover* de um ponto (LACERDA; CARVALHO, 1999) Figura 7 (a) é sorteado aleatoriamente um ponto de corte para cada casal de cromossomos pais dividindo-os em duas partes, cabeça e cauda. As cabeças dos dois cromossomos são mantidas e as caudas delas são trocadas entre si, gerando assim dois novos filhos. Já a técnica de N-pontos Figura 7 (b), cada casal de pais são divididos em mais de um único ponto, assim, os pedaços dos cromossomos gerados são tocados trocados entre si, gerando dois novos filhos (EIBEN; SMITH, 2003).

O *Crossover* é aplicado utilizando uma probabilidade em cada casal, chamada de taxa de *Crossover* que varia entre 0.6 e 0.9. Caso os números sorteados aleatoriamente em um intervalo de zero e um, for menor que a taxa de *Crossover* determinada, este operador é plicado no casal(LACERDA; CARVALHO, 1999).

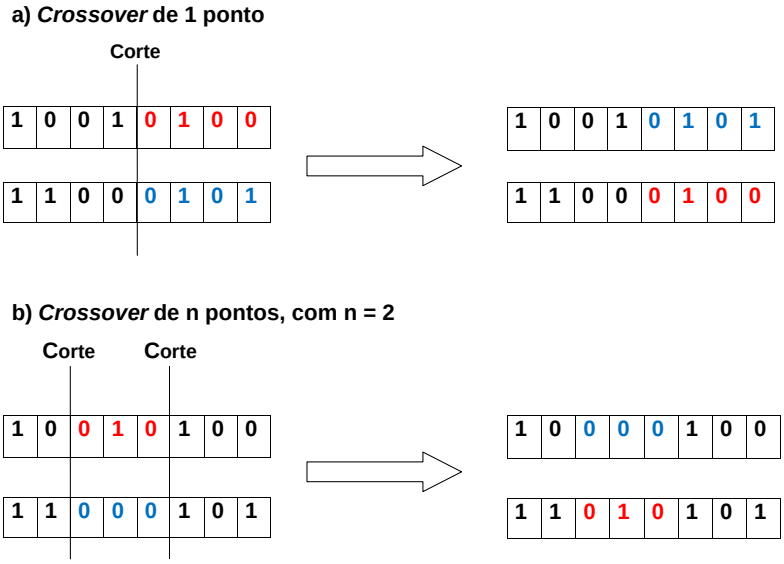


Figura 7: Modelo de crossover de 1-ponto em (a) e de N-pontos em (b).

2.3.1.6 MUTAÇÃO

A mutação é um operador responsável por introduzir e manter a diversidade genética da população, alterando de forma arbitrária uma pequena parte da estrutura do cromossomo selecionado, Figura 8. A mutação, assim como o cruzamento possui a taxa de mutação (POZO et al., 2005), no caso, se a taxa de mutação for alta de mais o algoritmo realizará de forma aleatória a busca pela solução ótima, no contrário, se a taxa for baixa de mais o processo da busca da solução ótima será lento.

A taxa de mutação varia entre 0.001 e 0.1, desta maneira, a mutação só é aplicado nos descendentes, caso os números sorteados aleatoriamente em um intervalo de zero e um, for menor que a taxa de mutação definida (CARVALHO et al., 2003).

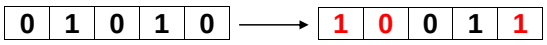


Figura 8: Mutação de um cromossomo com representação binária.

2.3.1.7 CRITÉRIO DE PARADA

O critério de parada tem como objetivo cessar o laço de repetição do AG caso tenha alcançado o ponto de parada pré-estabelecido (LUCAS, 2002). Os critérios de parada são

vários, entre eles estão, quando AG atingir um número de gerações ou quando algoritmo encontrar um *Fitness* aceitável (LACERDA; CARVALHO, 1999).

2.3.1.8 ELITISMO

O elitismo é um método muito utilizado que permite o AG convergir para uma boa solução. Este operador seleciona a cada população gerada, os melhores indivíduos (elite), que serão diretamente alocados na população seguinte. Se a elite for muito grande, a diversidade genética da população diminui, implicando na convergência prematura. Assim, devem ser selecionados a cada 50 indivíduos de uma população um ou dois dos melhores indivíduos para compor a elite (GUIMARÃES; RAMALHO, 2001).

2.3.1.9 FUNÇÃO *FITNESS*

A função *Fitness* ou função aptidão avalia todos os indivíduos da população de todas as gerações atribuindo uma nota a cada um. Esta função é representada na maioria das vezes por uma expressão matemática que mede a solução verificando se está longe ou perto da solução ótima. A função aptidão tem como entrada os valores do gene do cromossomo e fornece como saída a qualidade deste indivíduo, a aptidão (SANCHES, 2010).

A função aptidão a ser utilizada neste trabalho será a entropia. Esta é descrita mais detalhadamente a seguir (seção 2.5).

2.4 REDES COMPLEXAS

A teoria de redes complexas estende o formalismo da teoria de grafo. Seus modelos de redes possuem topologias distintas e propriedades bem definidas, que podem ser utilizadas para modelar redes gênicas e, além disso, caracteriza-las em termos de medidas de redes complexas (COSTA et al., 2007). Assim, utilizando redes complexas é possível caracterizar, analisar e representar diversos sistemas complexos, como por exemplo, sistemas biológicos (KAUFFMAN, 1993).

O primeiro modelo de redes complexas foi proposto por Paul Erdös e Alfréd (ER) Rényi em 1959 (ERDOS; RENYI, 1959) chamadas de redes aleatórias. A partir desta rede, muitos outros modelos de redes complexas foram criados, entre eles estão, Mundo Pequeno (SW, do inglês *Small-World*) (WATTS; STROGATZ, 1998) e livre escala (BARABÁSI; ALBERT, 1999).

As redes aleatórias de Erdös e Rényi (1959) evitam auto-relacionamento e conexões múltiplas. Sua topologia é baseada na conexão aleatória dos vértices considerando que todos os vértices possuem a mesma probabilidade de se conectar. A Figura 9 ilustra este tipo de rede aleatória.

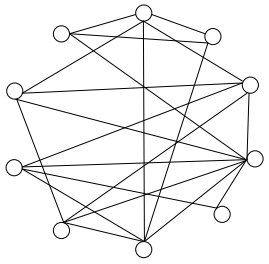


Figura 9: Topologia de rede complexa ER.

O modelo de redes SW de Watts e Strogatz (1998) tem como objetivo em criar redes que não sejam totalmente aleatórias, pelo fato de que topologias de redes biológicas, tecnológicas e sociais podem não apresentar total aleatoriedade. A Figura 10 ilustra a rede SW.

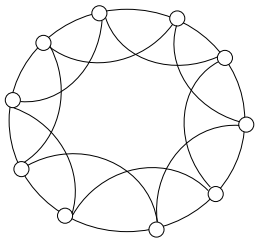


Figura 10: Topologia de rede complexa SW.

As redes *scale-free* de Barabási e Albert (1999) não apresentam uma distribuição homogênea de conexões entre seus vértices, ou seja, possui um pequeno número de vértices com muitas conexões e um grande número de vértices com poucas conexões. A Figura 11 ilustra a topologia desta rede.

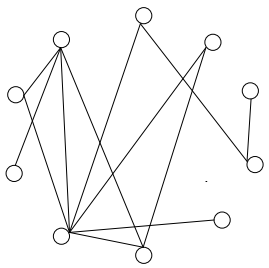


Figura 11: Topologia de rede complexa scale-free.

Os modelos de rede utilizado neste trabalho foi ER e SW.

2.5 ENTROPIA

O conceito de entropia foi introduzido em 1865 por Rudolf Clausius (CLAUSIUS, 1879). Ludwig Boltzman em 1877 mostrou que a entropia definida em termos de probabilidade pode associar-se á configuração microscópica de um sistema (BOLTZMANN, 1974), tal entropia ficou conhecida como entropia de Boltzman-Gibbs. A seguir é apresentado a entropia de Boltzman-Gibs de forma discreta:

$$H_{BG}(X) = k - \sum_{i=1}^{W} p_i log p_i \tag{1}$$

Onde k é a constante de Boltzman, e $p\imath$ é a probabilidade que corresponde as W configurações microscópicas, devendo satisfazer:

$$\sum_{i=1}^{W} p_i = 1 \tag{2}$$

Em 1948, Claude Shannon aplicou a entropia na Teoria da informação (SHANNON, 2001). Ela é utilizada para indicar o número de dados contido em determinada fonte e graduar a incerteza em um conjunto de dados (BISHOP et al., 1995). Considerando X uma variável aleatória de valor discreto, pode assumir valores booleanos 0 e 1. A entropia de Shannon determina em termos de probabilidade as ocorrências das variáveis aleatórias $P(x)$. Estas variáveis tem como resultado a incerteza, logo, quanto maior o resultado da função, maior a incerteza de predizer tal variável :

$$H(X) = -\sum_{x \in \mathbb{X}} p(x) log p(x) \tag{3}$$

Tal que:

$$\sum_{x \in \mathbb{X}} p(x) = 1 \tag{4}$$

Utilizando um conjunto de duas variáveis a entropia conjunta é determinada por:

$$H(X,Y) = -\sum_{x \in \mathbb{X}, y \in \mathbb{Y}} P(x,y) log P(x,y) \tag{5}$$

Onde o conjunto de variáveis aleatórias X e Y é representado pela probabilidade de

$P(x,y)$. Na entropia condicional representada por $H(Y \mid x)$ calcula-se a incerteza da variável aleatória Y dado o valor aleatório de x. Desta maneira, quanto maior o resultado maior será a chance da variável Y predizer x (KELEMEN et al., 2008). Esta entropia é definida por:

$$H(Y|x) = -\sum_{y \in \mathbb{Y}} P(y|x) log P(y|x) \tag{6}$$

A entropia condicional média é definida pela média ponderada das entropias condicionais de todas as instâncias $x \in X$ (JUNIOR, 2009), a entropia condicional média é representada por:

$$H(Y|X) = -\sum_{x \in \mathbb{X}} P(x) H(Y|x) \tag{7}$$

Onde $H(Y \mid x)$ representa a entropia condicional e $H(Y \mid X)$ representa uma valor, no qual quanto menor este valor, maior é a informação de Y pela observação de X .

2.6 TRABALHOS RELACIONADOS

Em (HIGA, 2011) a Inferência de Redes de Regulação Gênica Utilizando o Paradigma de Crescimento de Sementes tem como objetivo criar um algoritmo para inferência de GRNs a partir de dados temporais de expressão. Baseando-se nos modelos determinístico e estocástico, foi proposto para modelar as redes um método utilizando redes booleanas probabilísticas sensíveis ao contexto, já para a inferência destas redes foi utilizada redes booleanas com limiar e o paradigma de crescimento de semente de genes. Para validar o algoritmo foi utilizada uma rede conhecida. O Algoritmo foi aplicado sobre dados artificiais e biológicos de células HeLa, onde os resultados da validação foram satisfatórios, e o tempo de execução aceitável quando o algoritmo é aplicado sobre uma pequena quantidade de genes.

Em (MENDOZA; BAZZAN, 2011) Redes Booleanas Aleatórias Evoluindo com Algoritmos Genéticos para Reconstrução de Redes de Regulação (*Evolving Random Boolean Networks with Genetic Algorithms for Regulatory Networks Reconstruction*), o objetivo foi a inferência de GRNs modelados como redes booleanas aleatórias sem a utilização de informações biológicas e a partir de experimentos biológicos percorrer o espaço de busca utilizando o AG. Como resultado observou-se um elevado número de falsos positivo, porém o nível de precisão foi satisfatório na reconstrução da GRN.

Em (MENDOZA et al., 2012) Engenharia Revessa de GRNs: Uma Abordagem Evolu-

tiva baseada na Entropia de Tsallis (*Reverse Engineering of GRNs: An Evolutionary Approach based on the Tsallis Entropy*), o objetivo foi indicar os relacionamentos entre os genes de forma global e simples, identificando as conexões mais relevantes e construindo os modelos que descrevem os mecanismos e dinâmicas de expressão gênica e regulamentação. Para isto, foi aplicado um AG em redes baseadas nos modelos booleanas, onde a estrutura é inferida através da otimização da entropia de Tsallis. Os resultados mostraram que as redes geradas foram 50% mais precisas, comparadas com outras abordagens baseadas em booleanos.

Em (HATTORI, 2013) a Inferência de Redes Gênicas com Algoritmo Genético e Modelo de Ilhas (um modelo de evolução simultânea de multipopulações (ilhas)), tem como objetivo realizar a comparação entre o AG e o mesmo utilizando o Modelo de Ilhas na inferência de GRNs, a fim de verificar a acurácia e o tempo computacional de busca das redes. Para a validação da rede foi utilizada uma AGN e para avaliar a acurácia foi utilizado uma métrica de similaridade. Por fim, os algoritmos de buscas utilizando o Modelo de Ilhas obtiveram melhores resultados de acurácia comparados com o AG, porém, o tempo computacional gerado pelo Modelo de Ilhas é superior ao tempo de execução do AG.

Em (CUBAS, 2014) a Seleção de Características em Inferência de Redes de Interação Gênica a partir de Conjuntos Reduzidos de Amostras, que tem como objetivo desenvolver técnicas de seleção de características para diminuir o problema de estimação estatística existente ao inferir redes gênicas, dado um pequeno número de amostras. Para reduzir o problema de estatística existente foram propostos métodos de agrupamento linear para inferir BNs. Estes métodos seguem a abordagem de Redes Gênicas Probabilísticas que aplica a cada gene uma seleção de características local na inferência de redes gênicas. Desta maneira, a inferência com agrupamento linear, mesmo perdendo configurações originais dos preditores, obteve uma similaridade topológica melhor, comparada com a inferência sem agrupamento.

Este trabalho segue a mesma linha do trabalho desenvolvido por Hattori, onde foi aplicado o AG e o Modelo de Ilhas na inferência de GRNs. Neste contexto, este trabalho consiste no desenvolvimento de dois AGs, com a representação do cromossomo distintas, aplicados na inferência de GRNs, onde no primeiro AG desenvolvido, o cromossomo representa uma rede genes, e no segundo AG desenvolvido, o cromossomo representa um gene da rede. Desenvolvido os AGs, eles foram aplicados na inferência de GRNs, e após a inferência, foi realizado a validação do tempo computacional, na qual foi realizada por meio do tempo de execução dos AGs, e da acurácia na qual foi realizada através das AGNs. Por fim, foi realizado uma comparação do tempo computacional e da acurácia entre estes dois AGs.

3 DESENVOLVIMENTO DO TRABALHO

3.1 AG COM REPRESENTAÇÃO DOS CROMOSSOMOS POR REDES

Cada cromossomo deste algoritmo representa uma rede de genes. A representação da estrutura destes cromossomos é dada por uma matriz como pode ser vista na Figura 12. Sendo assim, a população deste algoritmo é composta por matrizes, onde cada matriz é uma rede de genes, e cada linha desta matriz representa um gene da rede. O tamanho das linhas possuem um tamanho fixo de 3 posições, na qual representam os genes reguladores do gene alvo, dessa forma foi adotado para este algoritmo $K = 3$.

Para a criação do cromossomo é sorteado um número aleatório que varia de 0 á 3, este número define a quantidade de reguladores do gene alvo inicialmente. Em seguida são gerados números aleatórios que variam de 0 até o número de genes que a rede possui, exceto ele mesmo para preencher os reguladores de um determinado gene alvo (caso o número de preditores sorteados para o gene alvo for menor que 3, estas posições sem preditores são preenchidas por -1, porém ao aplicar os operadores genéticos de mutação as posições com preditores podem ser alteradas para posições sem preditores e vice-versa).

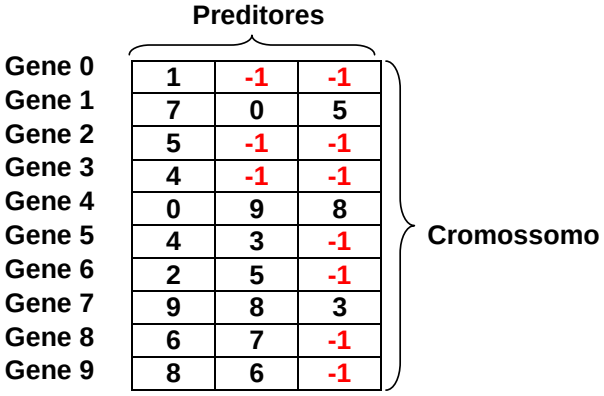


Figura 12: Exemplo de cromossomo do AG com representação dos cromossomos por redes, onde os número em preto são os reguladores dos respectivos genes alvos, e os números em vermelho são os coringas utilizados para manter o tamanho dos genes fixos.

A função adotada é baseada na a entropia condicional média descrito na seção 2.4, desenvolvido por (LOPES, 2011).

O método de seleção empregado neste AG é o método por torneio descrito na seção 2.3.4.1. Primeiramente o torneio seleciona três cromossomos aleatoriamente, e dentre estes três selecionados, o que possuir melhor *Fitness* será inserido em uma população auxiliar. O processo de seleção por meio do torneio é realizado enquanto o número de cromossomos inseridos na população auxiliar for menor que o número de cromossomos definidos para a população.

Neste AG o elitismo seleciona os melhores cromossomos após aplicar o operador de seleção e substitui os piores cromossomos após aplicar o operadores de cruzamento e mutação. Em cada aplicação do elitismo é verificado se os novos cromossomos produzidos são melhores que os cromossomos elite, assim se existir um cromossomo com melhor *Fitness* em relação aos cromossomos da elite, este cromossomo substituirá o cromossomo da elite com o pior *Fitness*.

O cruzamento adotado neste algoritmo é o de dois pontos (seção 2.3.1.5), aplicado após o método de seleção. Os cromossomos que se cruzarão são definidos de forma aleatória e divididos em pares, em seguida, é sorteado para cada par de cromossomos dois pontos de corte aleatoriamente, sendo que a primeira e a ultima posição não são sorteados, permitindo assim que os cromossomos se dividam em três partes, na qual uma destas partes é sorteada para realizar a troca de características entre o par de cromossomos. Vale observar que as características dos cromossomos são os genes alvos que compõem a rede/cromossomo. Este processo de cruzamento pode ser visto na Figura 13.

O método de mutação desenvolvido para este algoritmo possui como parâmetros a taxa de mutação sobre a população e a taxa de mutação sobre os genes que compõem os cromossomos, ou seja, a mutação sobre a população define quantos cromossomos da população serão modificados e a taxa sobre os gene define se o genes sofrerá mutação ou não. A mutação é aplicada sobre os genes preditores de um determinado gene alvo enquanto não encontrar uma melhora em seu *Fitness*, dessa maneira a mutação não retorna uma rede com o *Fitness* pior, retornando uma rede com o mesmo *Fitness* ou com o *Fitness* melhor. O processo de mutação pode ser visto na Figura 14.

O critério de parada adotado neste algoritmo é baseada na quantidade de gerações, logo, a quantidade adotada foi de 1000 gerações da população.

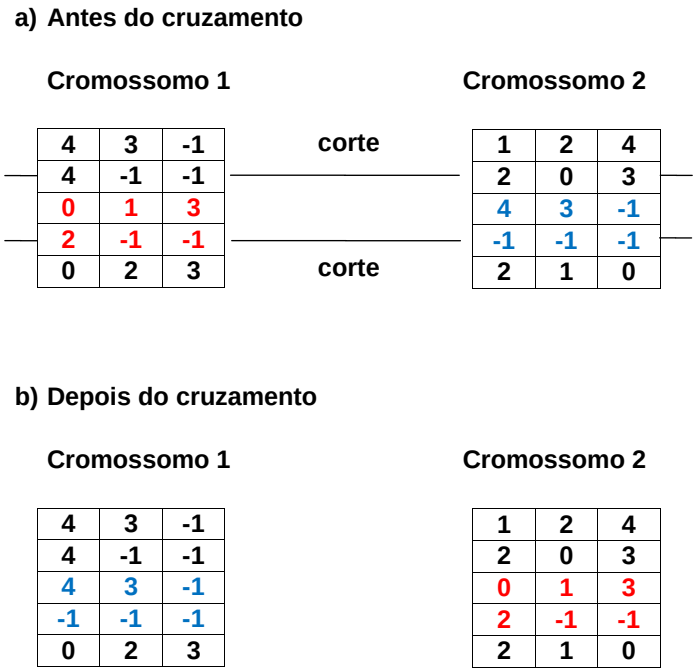


Figura 13: Ilustração do cruzamento do AG com representação dos cromossomos por redes. Em (a) cromossomos 0 e 1 antes do cruzamento das características, com pontos de corte nas linhas 1 e 3, e em (b) cromossomos 0 e 1 após o cruzamento, onde pode ser observado a troca das linhas 2 e 3 pelos cromossomos.

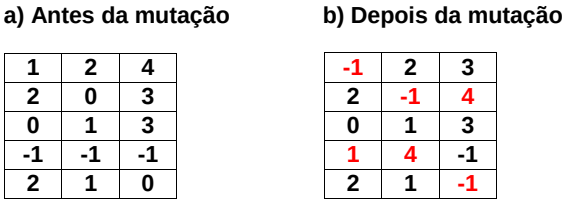


Figura 14: Ilustração da mutação do AG com representação dos cromossomos por redes. Em (a) cromossomo antes da mutação e em (b) cromossomo após a mutação, com algumas posições modificas pela mutação em vermelho.

3.2 AG COM REPRESENTAÇÃO DOS CROMOSSOMOS POR GENES

Os cromossomos da população deste genético representam conjuntos de preditores de um gene, assim, a solução retornada por este algoritmo é um conjunto de preditores de um gene da rede, onde cada conjunto possui no máximo três preditores, representados por números binários. Esta população de cromossomos pode ser observada na Figura 15, onde representa uma população de 10 cromossomos de uma rede de 31 genes, sendo que cada cromossomo é representado por uma linha desta matriz composta por três preditores, e cada preditor é representado por 5 posições.

Pelo fato da representação dos preditores serem por números binários, redes com de-

terminados números de genes são inviáveis para este AG, assim, a quantidade de genes que este AG suporta é 2^n, onde ^{n} é um número positivo e diferente de 0.

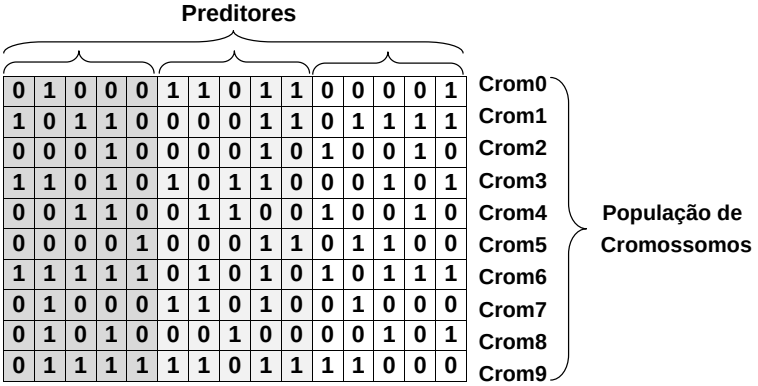


Figura 15: Exemplo de população de cromossomos do AG com representação dos cromossomos por um gene alvo, cada linha representa um cromossomo da população, e cada cromossomo é constituído por 3 preditores, sendo que todos os cromossomos representam um único gene alvo, que no caso é o gene 0.

A população inicial deste genético é gerado aleatoriamente, e os operadores de seleção, de elitismo e de cruzamento são os mesmos empregados no AG, com representação dos cromossomos por redes. Dessa forma, o que diferencia é o operador de mutação, onde este realiza modificações na estrutura do cromossomo, podendo gerar um conjunto de preditores com um *Fitness* inferior ao conjunto de preditores antes da mutação, e o critério de parada, onde este também é com base no número de gerações, porém em relação a cada gene da rede, assim o número de gerações adotado para cada gene da rede foi de 1000 gerações. O cruzamento e a mutação deste genético são ilustradas nas Figuras 16 e 17 respectivamente.

3.3 CONFIGURAÇÕES DA AGN PARA EXECUÇÃO DOS EXPERIMENTOS

Para gerar os dados de expressão gênica e as AGNs, foi utilizado o aplicativo JAGN (Figura 18) desenvolvido por (LOPES et al., 2011).

Neste trabalho, foi realizado experimentos com AGNs distintas, as configurações destas AGNs para execução dos experimentos pelos dois algoritmos genéticos foram as seguintes:

- Número de genes na rede: 31, 63 e 127, estes foram definidos com base nos números de genes que o AG com representação dos cromossomos por um gene aceita (seção 3.2);
- Número médio de conexões na rede: 3;
- Tamanho do sinal de expressão gênica: 20;

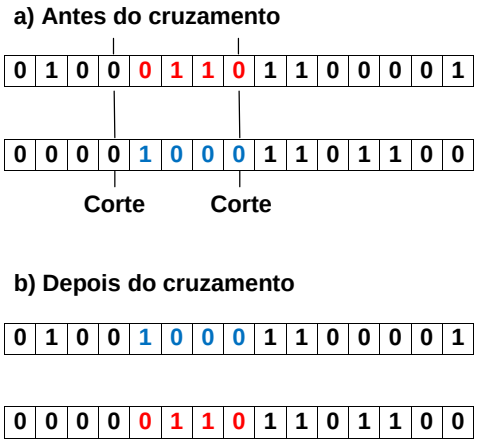


Figura 16: Ilustração de cruzamento dos cromossomos do AG com representação dos cromossomos por um gene alvo. Em (a) par de cromossomos antes do cruzamento das características, com pontos de corte nas posições 3 e 7, e em (b) par de cromossomos após o cruzamento, onde pode ser observado a troca das posições 4, 5, 6 e 7.

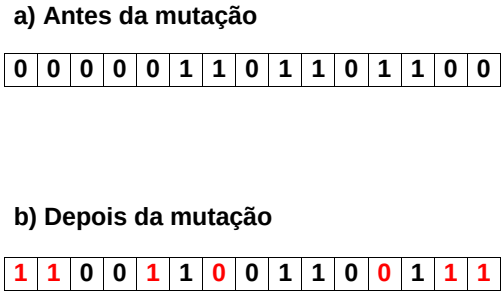


Figura 17: lustração da mutação dos cromossomos do AG com representação dos cromossomos por um gene alvo. Em (a) cromossomo antes da mutação e em (b) cromossomo após a mutação, com algumas posições modificas pela mutação em vermelho.

- Topologias das redes: Erdös-Rényi (ER) e Small World (SW).

3.4 CONFIGURAÇÕES DOS OPERADORES GENÉTICOS

- Cromossomos: para o AG com representação dos cromossomos por redes a quantidade de fenótipos é baseado no número de genes que a rede possuí, logo cada gene possui no máximo 3 preditores, já no AG com representação dos cromossomos por genes a quantidade de fenótipo é 3, nos quais representam seus respectivos preditores. Assim, em ambos os AGs, o número máximo de preditores para cada gene é 3;
- Tamanho da População: 50 e 100 cromossomos;
- População Inicial: método aleatório para gerar a população inicial;

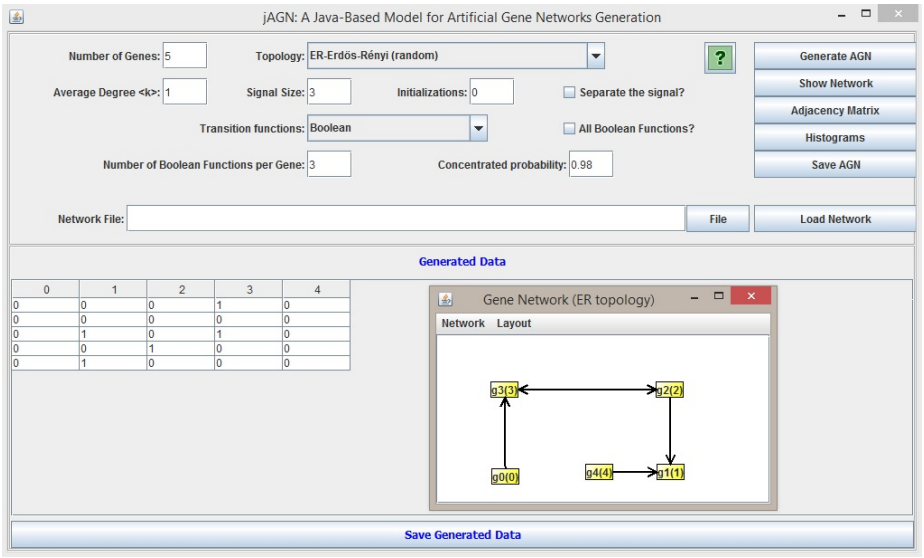


Figura 18: Aplicativo JAGN.

Fonte: (LOPES et al., 2011)

- Função *Fitness*: a função é baseada na entropia condicional média (seção 2.4), desenvolvida por (LOPES, 2011);
- Seleção: método de seleção por torneio, com 3 cromossomos por torneio, e entre estes apenas 1 é selecionado;
- Operador de Elitismo: taxa de 10% da população é selecionada para a elite;
- Operador de *Crossover*: foi adotado o cruzamento de 2 pontos com taxa de cruzamento de 80% da população;
- Operador de Mutação: taxa de mutação sobre a população de cromossomos de 80% e taxa de mutação sobre os genes dos cromossomos de 30%;
- Critério de Parada: para ambos os AGs foi adotado o critério de parada com base no número de gerações, para o AG com representação dos cromossomos por redes foi empregado 1000 gerações da população, já em relação ao AG com representação dos cromossomos por genes, foi adotado 1000 gerações para cada gene da rede;

3.5 VALIDAÇÃO E ANÁLISE

As duas abordagens desenvolvidas foram executadas 20 vezes para cada modelo de rede (ER e SW) e para cada tamanho de população 50 e 100. A avaliação do tempo computa-

cional dos algoritmos desenvolvidos neste trabalho foi realizada através do tempo de execução de cada abordagem desenvolvida, os resultados gerados são descritos no capítulo 4.

A avaliação da acurácia das GRNs inferidas pelos algoritmos de inferência foi por meio AGN, gerado pelo aplicativo JAGN, desta maneira, não serão utilizadas redes gênicas reais para avaliar a acurácia das redes inferidas. Uma AGN é composta por vértices (genes) e arestas (ligações entre genes), podendo ser representada por uma matriz de adjacências, na qual a aresta que liga o vértice v_i ao vértice v_j é representado por $M(i, j)$, podendo assumir valor 1 ou 0, sendo que 1 está conectado e 0 não.

Diante disto, a avaliação da acurácia das redes inferidas pelos AGs deste trabalho foi realizado por meio de medidas de similaridade. Estas medidas utilizam as variáveis entre a matriz de adjacências da rede inferida e a matriz de adjacência da AGN que pode ser gerada pelo aplicativo JAGN (LOPES et al., 2008; LOPES, 2011). As variáveis entre estas matrizes de adjacência são: *True Positive* (TP), arestas que foram inferidas e existem na rede original, *False Positive* (FP), arestas que foram inferidas, mas não existem na rede original, o *False Negative* (FN), arestas que não foram inferidas e que existem na rede real e o *True Negative* (TN), arestas que não foram inferidas e que não existem na rede real.

Para verificação da acurácia das redes inferidas neste trabalho foram utilizadas medidas de similaridade, são elas: Valor Preditivo Positivo (PPV, do inglês *Positive Predictive Value*) também conhecido como precisão Equação 8, a $Especificidade$ que quantificam as arrestas inferidas corretamente e incorretamente Equação 9 e a $Sensibilidade$ que quantificam as arestas da rede original que não foram inferidas Equação 10.

$$PPV = \frac{TP}{(TP+FP)} \tag{8}$$

$$Especificidade = \frac{TN}{(TN+FP)} \tag{9}$$

$$Sensibilidade = \frac{TP}{(TP+FN)} \tag{10}$$

Estas medidas não são independentes entre elas, assim, quando utilizadas em conjuntos é necessário utilizar uma média geométrica, a média adotada neste trabalho é dada por:

$$Similaridade(A,B) = \sqrt[3]{PPV * Especificidade * Sensibilidade} \tag{11}$$

Onde A representa a AGN gerada pelo aplicativo JAGN, e B a rede inferida pelo algoritmo genético desenvolvido. As taxas de arestas corretas e incorretas são consideradas pelas

medidas na Equação 11, onde o valor máximo da similaridade é obtida por valores destas medidas próximos a 1, (LOPES, 2011).

3.6 CONFIGURAÇÃO DO AMBIEMTE DE EXECUÇÃO

As configurações do *hardware* onde os testes foram executados, possuem as seguintes especificações:

- Tamanho da memória principal: 4Gb;
- Sistema Operacional: Windows 7 *Professional*;
- Processador: Intel(R) Core(TM) i5-2400S @ 3.60GHz.

4 RESULTADOS

Neste capitulo é apresentado os resultados obtidos pelos AGs desenvolvidos (seções 3.1 e 3.2) na inferência de redes usando modelos de redes aleatórias (*uniformly-random*) (ER) e modelo de redes mundo pequeno (*small-world*) (SW), considerando a quantidade de 31, 63 e 127 genes, tamanho do sinal igual a 20 e o grau médio das ligações entre os genes $k = 3$. Assim, para cada rede, foram gerados duas redes com as topologias ER e SW.

Nas Figuras 19, 20, 21 e 22, são apresentados as médias de similaridade das redes inferidas pelo AG com representação dos cromossomos por genes e pelo AG com representação dos cromossomos por redes, em redes com os tamanhos 31, 63 e 127.

Nas Figuras 19 e 20, com topologias ER e SW respectivamente, o tamanho da população utilizado pelos dois algoritmos foi 50, já nas Figuras 21 e 22, com topologias ER e SW respectivamente, o tamanho da população utilizado pelos dois algoritmos foi 100. Na topologia ER para ambos os tamanhos, a rede com 31 genes apresentou uma maior similaridade quando inferida pelo AG com representação dos cromossomos por genes, porém nas redes com tamanhos 63 e 127, o AG com representação dos cromossomos por redes obteve uma maior similaridade. Considerando a topologia SW, o AG com representação dos cromossomos por redes apresentou maior similaridade para as redes com 31, 63 e 127 genes.

Já as Figuras 23, 24, 25 e 26, são apresentados as médias dos tempos de execução dos algoritmos desenvolvidos. Em todos os experimentos para as topologias ER e SW, o AG com representação do cromossomo por genes obteve um melhor tempo computacional quando comparado com o AG com representação do cromossomo por redes.

Finalmente, as figuras 27 a 32 ilustram os boxplots obtidos para as topologias ER e SW, considerando redes com 31, 63 e 127 genes. Para as redes utilizando a topologia ER é possível observar que o AG com representação dos cromossomos por genes (AG2) apresentou maior similaridade para as redes compostas por 31 genes. Entretanto, com o aumento da quantidade de genes para 63 e 127, o AG com representação dos cromossomos por redes (AG1) obteve maior similaridade, em especial para população com 100 indivíduos. Considerando a topologia

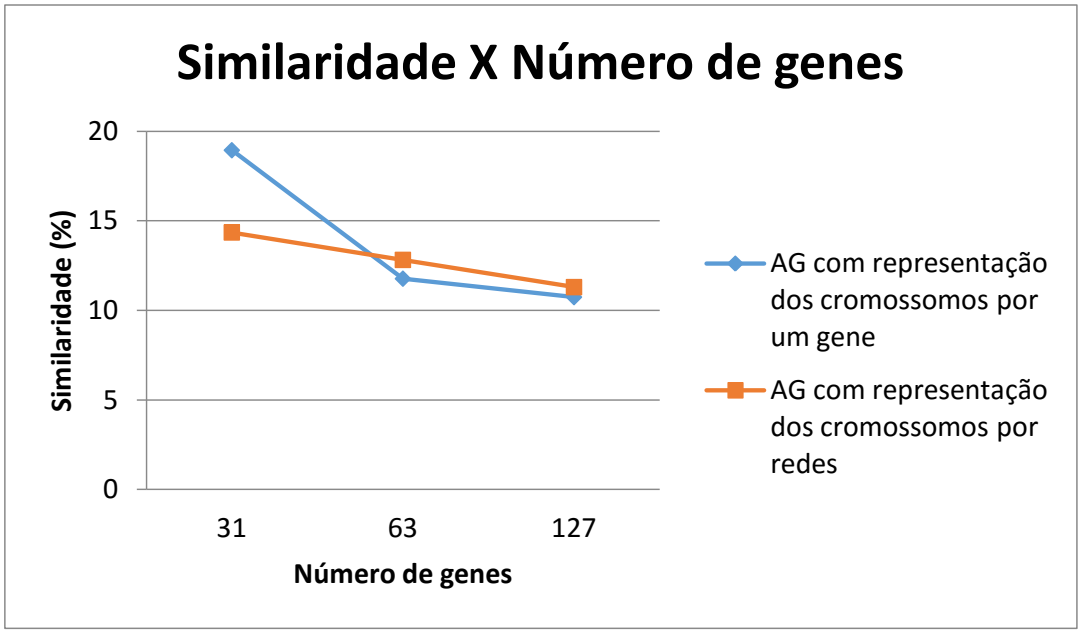


Figura 19: Média de similaridade das redes inferidas pelos AGs desenvolvidos, em redes ER com tamanho 31, 63 e 127 genes, e população com 50 cromossomos.

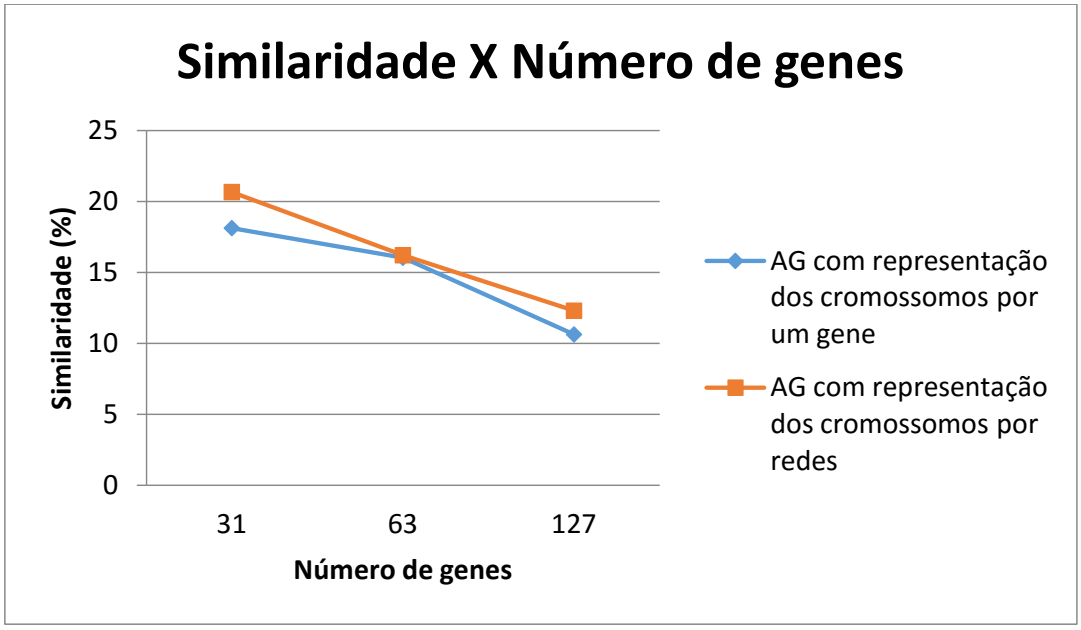


Figura 20: Média de similaridade das redes inferidas pelos AGs desenvolvidos, em redes SW com tamanho 31, 63 e 127 genes e população com 50 cromossomos.

SW, o AG1 manteve o desempenho superior obtido com a topologia ER, gerando redes com maiores similaridades em relação ao AG2.

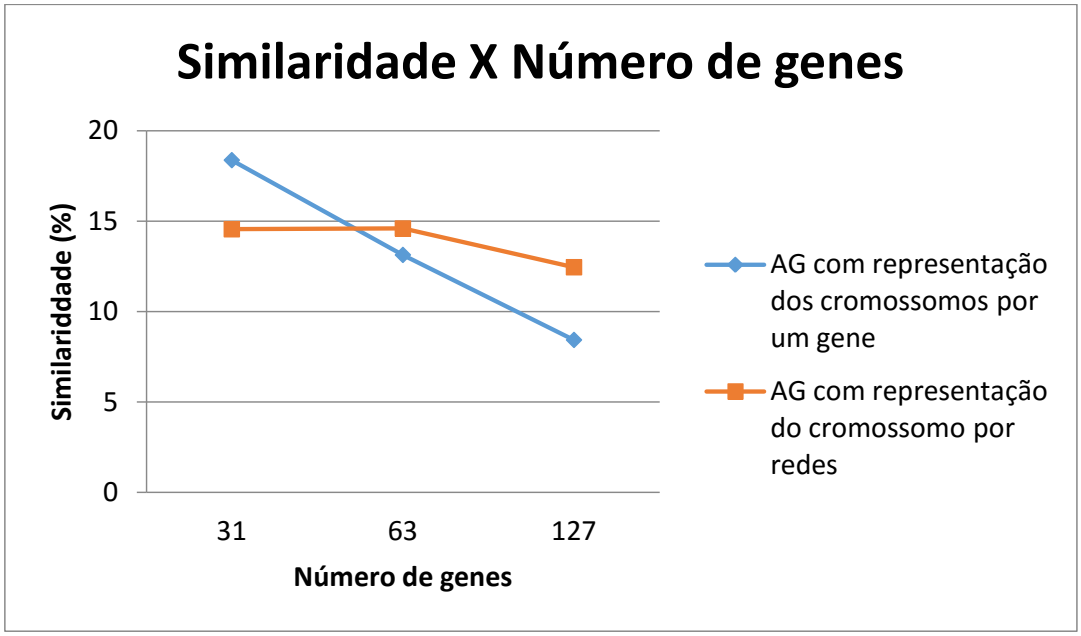


Figura 21: Média de similaridade das redes inferidas pelos AGs desenvolvidos, em redes ER com tamanho 31, 63 e 127 genes e população com 100 cromossomos.

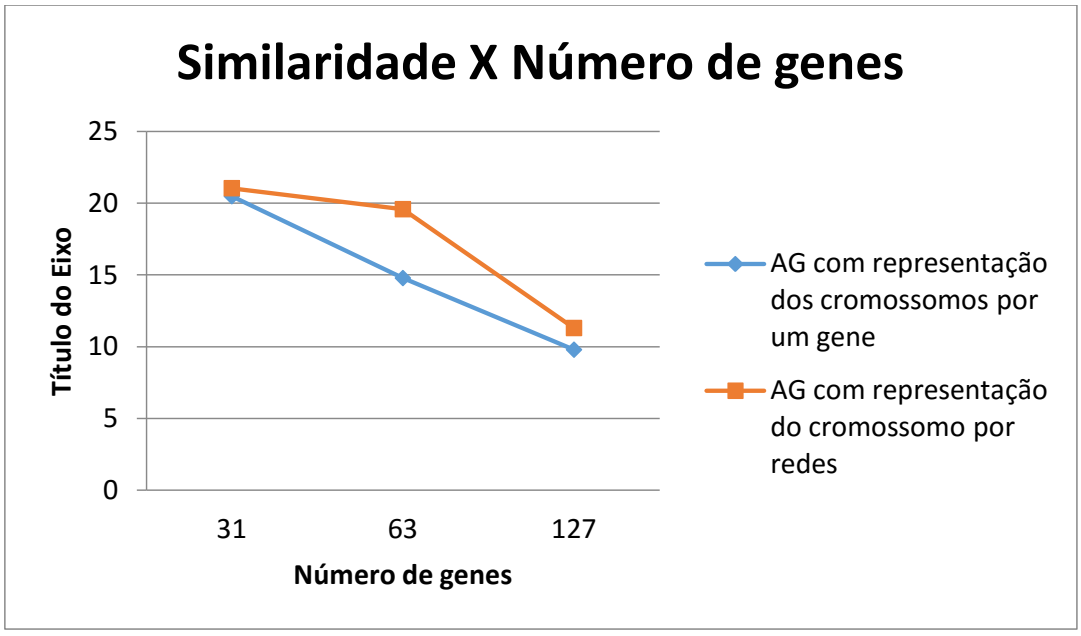


Figura 22: Média de similaridade das redes inferidas pelos AGs desenvolvidos, em redes SW com tamanho 31, 63 e 127 genes e população com 100 cromossomos.

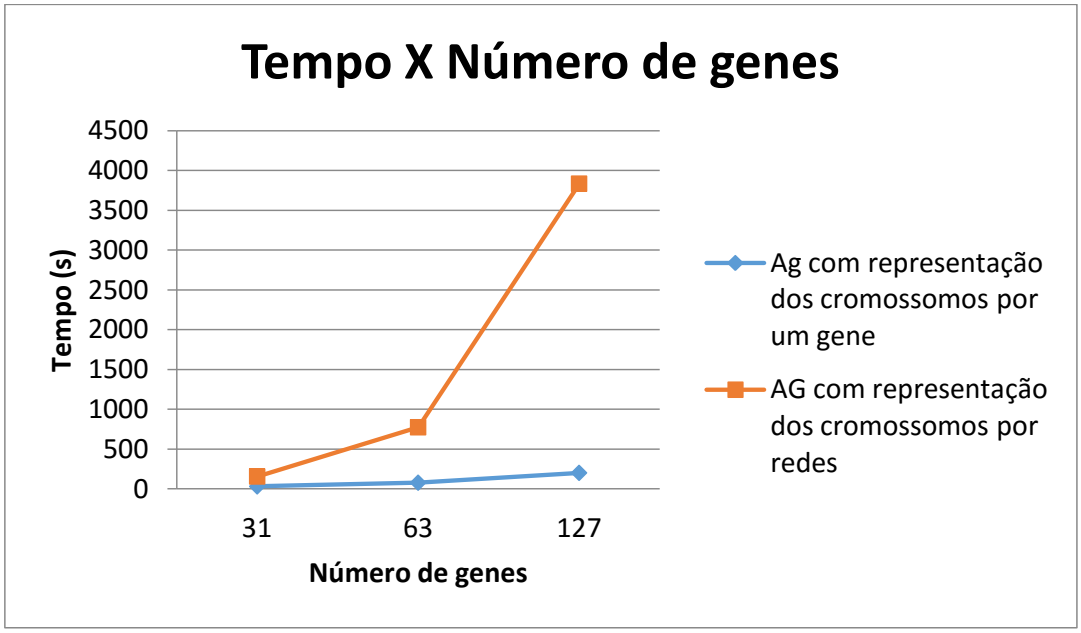


Figura 23: Média do tempo de execução dos AGs desenvolvidos, em redes ER com tamanho 31, 63 e 127 genes e população com 50 cromossomos.

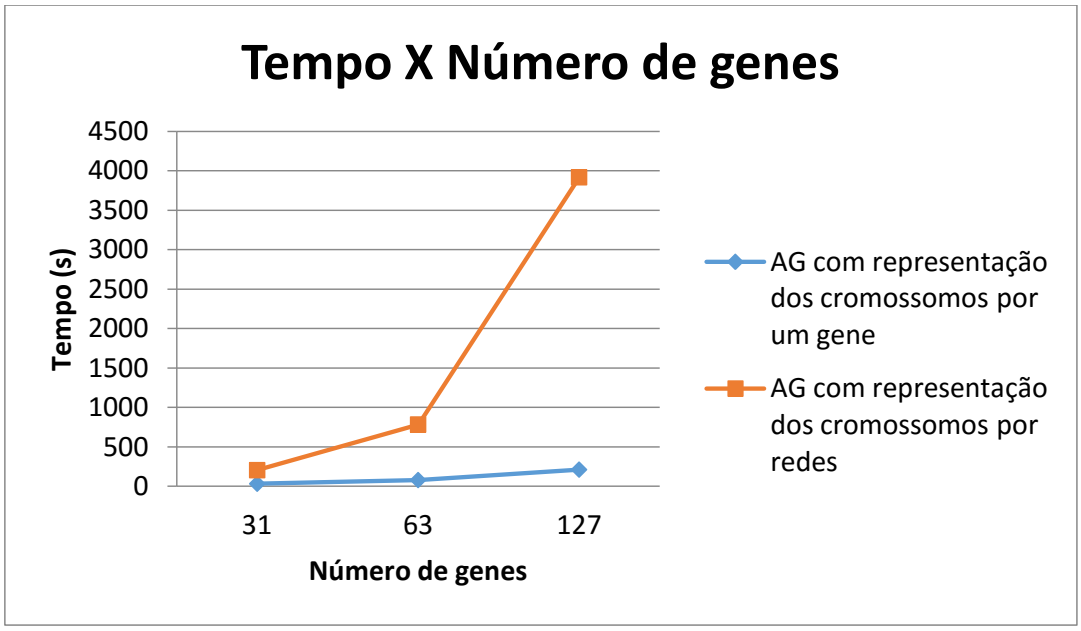


Figura 24: Média do tempo de execução dos AGs desenvolvidos, em redes SW com tamanho 31, 63 e 127 genes e população com 50 cromossomos.

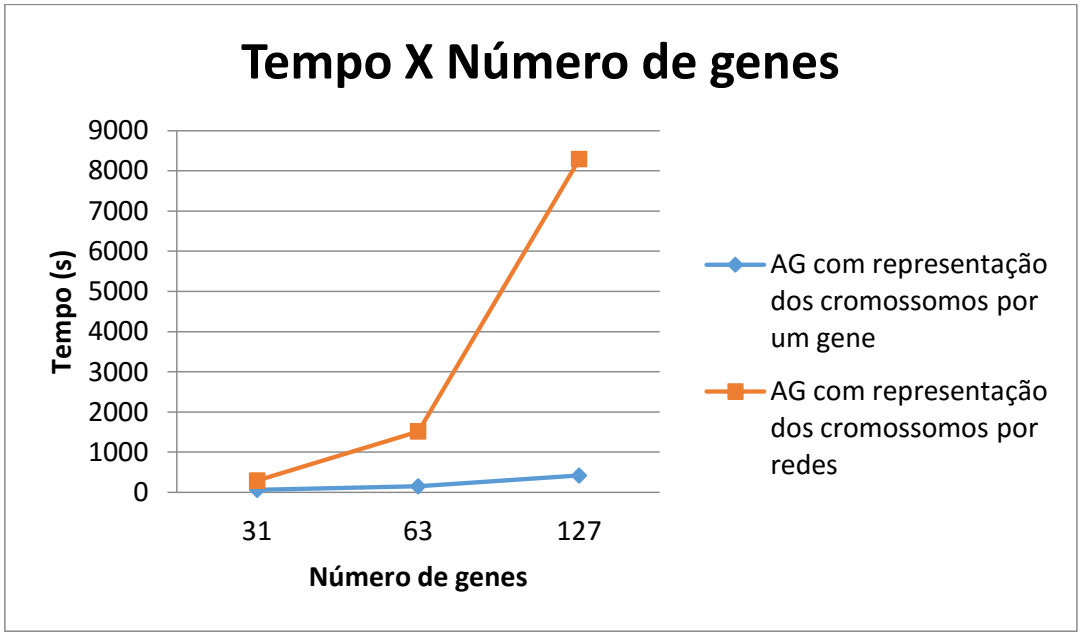


Figura 25: Média do tempo de execução dos AGs desenvolvidos, em redes ER com tamanho 31, 63 e 127 genes e população com 100 cromossomos.

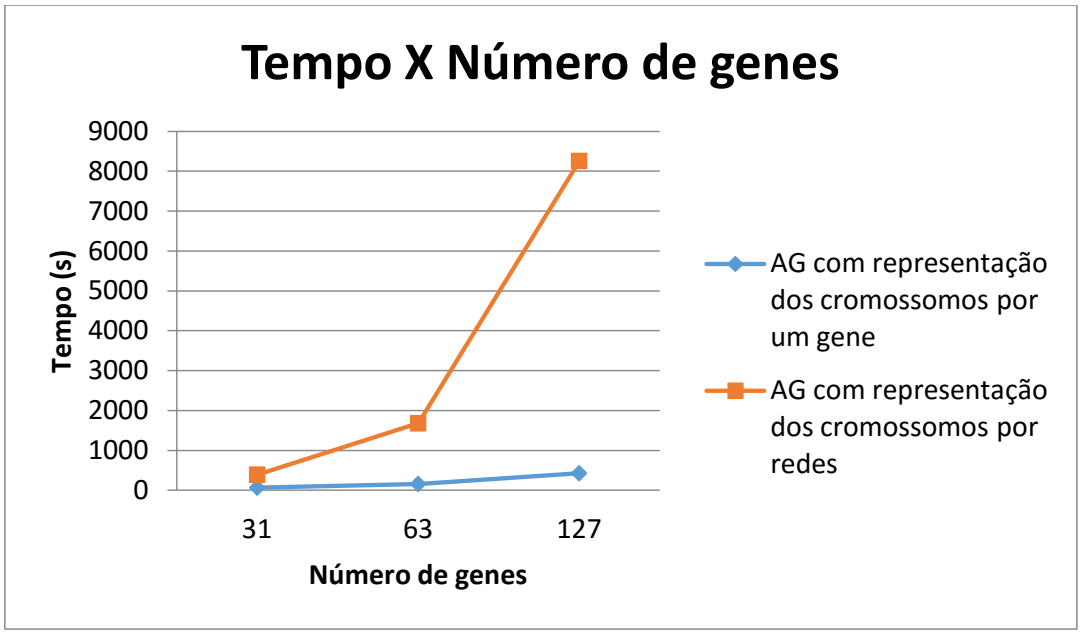
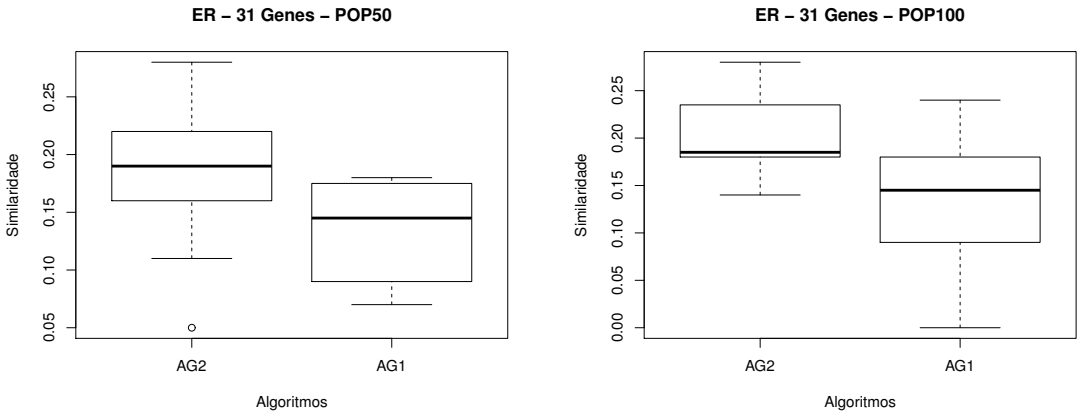
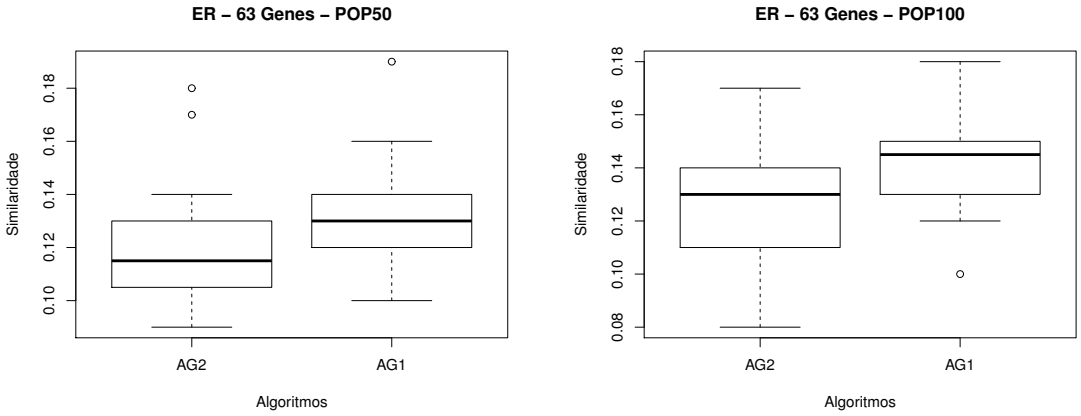


Figura 26: Média do tempo de execução dos AGs desenvolvidos, em redes SW com tamanho 31, 63 e 127 genes e população com 100 cromossomos.



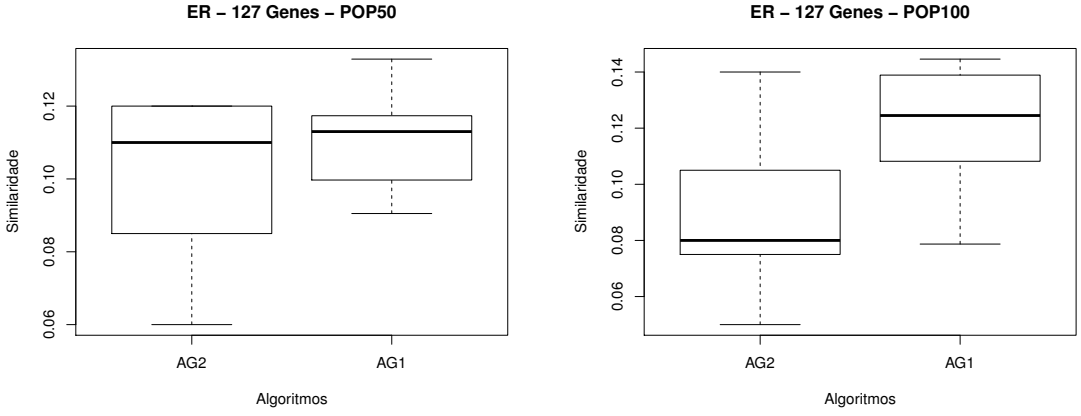
(a) POP50 (b) POP100

Figura 27: Box plot das similaridades das redes inferidas pelo modelo de rede ER com 31 genes.



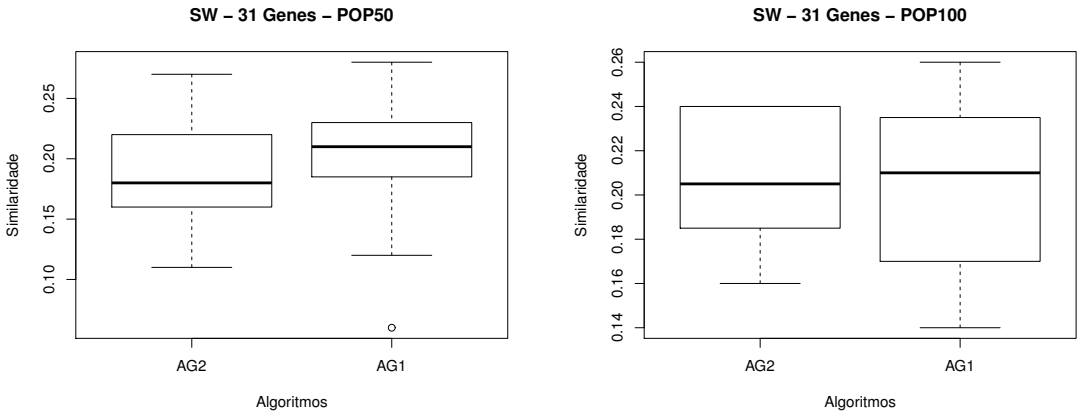
(a) POP50 (b) POP100

Figura 28: Box plot das similaridades das redes inferidas pelo modelo de rede ER com 63 genes.



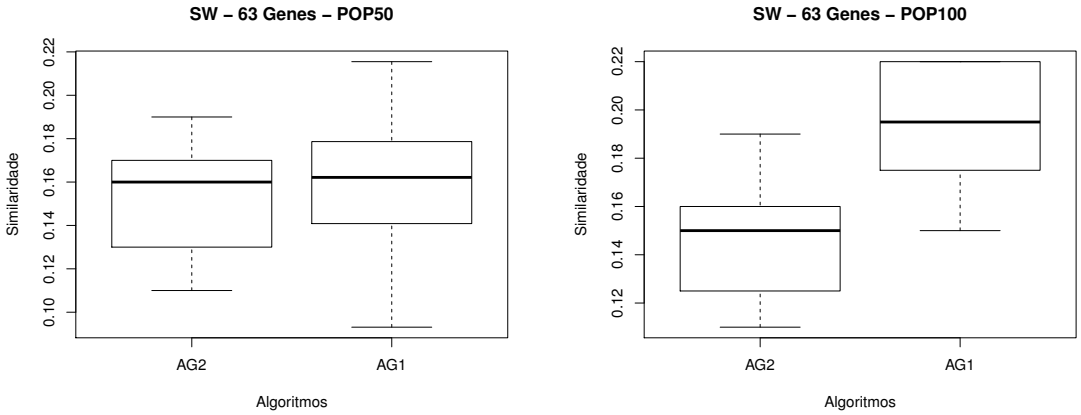
(a) POP50 (b) POP100

Figura 29: Box plot das similaridades das redes inferidas pelo modelo de rede ER com 127 genes.



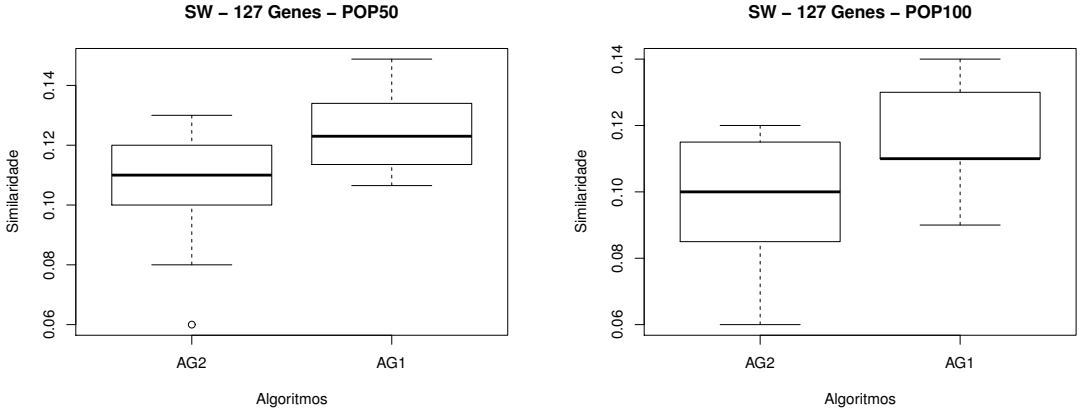
(a) POP50 (b) POP100

Figura 30: Box plot das similaridades das redes inferidas pelo modelo de rede SW com 31 genes.



(a) POP50 (b) POP100

Figura 31: Box plot das similaridades das redes inferidas pelo modelo de rede SW com 63 genes.



(a) POP50 (b) POP100

Figura 32: Box plot das similaridades das redes inferidas pelo modelo de rede SW com 127 genes.

5 CONCLUSÃO

O trabalho desenvolvido abordou a inferência de GRNs a partir de dados de expressão gênica, por meio de métodos de seleção de característica, na qual é composto por uma função critério e um algoritmo de busca. A função critério utilizada é baseada na entropia condicional média desenvolvida por (LOPES, 2011), já o algoritmo de busca adotado é o algoritmo genético, um método de busca e otimização, fundamentado no princípio da seleção natural e sobrevivência do mais apto, do naturalista e fisiologista Charles Darwin (DARWIN; BYNUM, 2009).

Neste sentido, foi proposto neste trabalho o desenvolvimento de duas abordagens baseadas em algoritmos evolutivos, com foco em diferentes formas de representação dos cromossomos. A primeira abordagem é baseada no retorno de uma rede de genes completa como solução, e a segunda abordagem é baseada no retorno de preditores de um gene alvo como solução, ou seja, na primeira um cromossomo representa uma rede de genes e na segunda um cromossomo representa um conjunto de preditores de um gene da rede.

Os resultados de similaridade para o AG com representação dos cromossomos por genes, teve maior eficácia nas redes ER com 31 genes e com o tamanho da população 50 e 100. Entretanto, para as redes ER com 63 e 127 genes e as redes SW com 31, 63 e 127 o AG com representação dos cromossomos por redes, obteve redes com maiores similaridades. Por outro lado, o AG com representação dos cromossomos por redes apresentou um baixo desempenho computacional, quando comparado o AG com representação dos cromossomos por genes.

Com base nos resultados obtidos, é possível observar que dependendo da representação utilizada é possível obter melhores direcionamentos no espaço de busca. Entretanto, um dos grandes desafios é melhorar o mecanismo de busca por regiões promissoras sem comprometer o desempenho computacional dos algoritmos.

5.1 TRABALHOS FUTUROS

Como possíveis trabalhos futuros pretende-se realizar experimentos utilizando redes com maior número de genes, maior grau de conexão, aumento do tamanho do sinal de expressão gênica e utilização de outras topologias.

Considerando os algoritmos evolutivos já desenvolvidos, pretende-se explorar variações dos parâmetros do AG como, a taxa de mutação, taxa de elitismo, taxa de *crossover*, número de interações, tamanho da população e outras formas de representação do cromossomo.

Pretende-se ainda investigar algoritmos baseados na programação dinâmica e outras classes de algoritmos evolutivos além da inclusão de *multithreads* para a inferência de redes de regulação gênica.

REFERÊNCIAS

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999.

BISHOP, C. M. et al. Neural networks for pattern recognition. Clarendon press Oxford, 1995.

BOLTZMANN, L. **Theoretical physics and philosophical problems**. [S.l.]: Reidel, 1974.

BRAGA, C. **O uso de Algoritmos Genéticos para aplicação de Otimização de Sistemas Mecânicos**. Tese (Doutorado) — Dissertação de mestrado, Universidade Federal de Uberlândia, 1998.

CAMPOS, T. E. de. **Técnicas de seleçao de caracterısticas com aplicaçoes em reconhecimento de faces**. Tese (Doutorado) — Universidade de São Paulo, 2001.

CARVALHO, A.; BRAGA, A. d. P.; LUDERMIR, T. Computação evolutiva. **Sistemas Inteligentes: Fundamentos e Aplicaçoes. Manole**, p. 225–248, 2003.

CLAUSIUS, R. **The mechanical theory of heat**. [S.l.]: MacMillan, 1879.

COSTA, L. d. F. et al. Characterization of complex networks: A survey of measurements. **Advances in Physics**, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007.

CUBAS, C. F. M. Selecao de características em inferência de redes de interacao gênica a partir de conjuntos reduzidos de amostras. 2014.

DANTAS, T. **Genes e cromossomos**. 2015. Brasil Escola. Disponível em: <http://www.brasilescola.com/biologia/genes-cromossomos.htm>. Acesso em: 20 nov. 2015.

DARWIN, C.; BYNUM, W. F. **The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life**. [S.l.]: AL Burt, 2009.

DOUGHERTY, E. R. Validation of inference procedures for gene regulatory networks. **Current Genomics**, Bentham Science Publishers, v. 8, n. 6, p. 351, 2007.

EIBEN, A. E.; SMITH, J. E. **Introduction to evolutionary computing**. [S.l.]: Springer Science & Business Media, 2003.

ERDOS, P.; RENYI, A. On random graphs. **Publicationes Mathematicae Debrecen**, v. 6, p. 290–297, 1959.

FILHO, P. A. da C.; POPPI, R. J. Algoritmo genético em química. **Química Nova**, SciELO Brasil, v. 22, n. 3, p. 405, 1999.

GOLBERG, D. E. Genetic algorithms in search, optimization, and machine learning. **Addion wesley**, v. 1989, 1989.

GOLDBERG, D. E.; DEB, K. A comparative analysis of selection schemes used in genetic algorithms. **Foundations of genetic algorithms**, v. 1, p. 69–93, 1991.

GUIMARÃES, F. G.; RAMALHO, M. C. Implementação de um algoritmo genético. **Trabalho referente à disciplina Otimização, junho de**, 2001.

HATTORI, L. T. **Inferência de Redes Gênicas com Algoritmo Genético e Modelo de Ilhas.** 2013.

HIGA, C. H. A. **Inferência de redes de regulação gênica utilizando o paradigma de crescimento de sementes**. Tese (Doutorado) — Universidade de São Paulo, 2011.

HOLLAND, J. H. **Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.** [S.l.]: U Michigan Press, 1975.

JAIN, A.; ZONGKER, D. Feature selection: Evaluation, application, and small sample performance. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, IEEE, v. 19, n. 2, p. 153–158, 1997.

JAIN, A. K.; DUIN, R. P. W.; MAO, J. Statistical pattern recognition: A review. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, IEEE, v. 22, n. 1, p. 4–37, 2000.

JONG, K. A. D. Analysis of the behavior of a class of genetic adaptive systems. 1975.

JUNIOR, D. C. M. **Seleçao de caracterısticas e prediçao intrinsecamente multivariada em identificaçao de redes de regulaçao gênica**. Tese (Doutorado) — Universidade de Sao Paulo, 2009.

KAUFFMAN, S. A. **The origins of order: Self-organization and selection in evolution**. [S.l.]: Oxford university press, 1993.

KELEMEN, A.; ABRAHAM, A.; CHEN, Y. **Computational intelligence in bioinformatics**. [S.l.]: Springer Science & Business Media, 2008.

LACERDA, E. G. de; CARVALHO, A. de. Introdução aos algoritmos genéticos. **Sistemas inteligentes: aplicaçoes a recursos hıdricos e ciências ambientais**, v. 1, p. 99–148, 1999.

LOPES, F. M. **Redes complexas de expressão gênica: síntese, identificação, análise e aplicações**. Tese (Doutorado) — Universidade de São Paulo, 2011.

LOPES, F. M.; JR, R. M. C.; COSTA, L. d. F. Agn simulation and validation model. In: **Advances in Bioinformatics and Computational Biology**. [S.l.]: Springer, 2008. p. 169–173.

LOPES, F. M.; JR, R. M. C.; COSTA, L. D. F. Gene expression complex networks: synthesis, identification, and analysis. **Journal of Computational Biology**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 18, n. 10, p. 1353–1367, 2011.

LUCAS, D. C. Algoritmos genéticos: uma introduçao1. 2002.

MENDOZA, M. R.; BAZZAN, A. L. C. Evolving random boolean networks with genetic algorithms for regulatory networks reconstruction. In: ACM. **Proceedings of the 13th annual conference on Genetic and evolutionary computation**. [S.l.], 2011. p. 291–298.

MENDOZA, M. R.; LOPES, F. M.; BAZZAN, A. L. C. Reverse engineering of grns: An evolutionary approach based on the tsallis entropy. In: ACM. **Proceedings of the 14th annual conference on Genetic and evolutionary computation**. [S.l.], 2012. p. 185–192.

MICHALEWICZ, Z. **Genetic algorithms+ data structures= evolution programs**. [S.l.]: Springer Science & Business Media, 1996.

NELSON, D. L.; LEHNINGER, A. L.; COX, M. M. **Lehninger principles of biochemistry**. [S.l.]: Macmillan, 2008.

PACHECO, M. A. C. Algoritmos genéticos: princípios e aplicações. **ICA: Laboratório de Inteligência Computacional Aplicada. Departamento de Engenharia Elétrica. Pontifícia Universidade Católica do Rio de Janeiro. Fonte desconhecida**, 1999.

POZO, A. et al. Computação evolutiva. **Grupo de Pesquisas em Computação Evolutiva. Departamento de Informática. Universidade Federal do Paraná**, 2005.

SANCHES, D. S. Estratégia de modelagem por algoritmo genético adaptativo para programação reativa da produção de produtos com uso simultâneo de máquinas e sistemas de transporte em sistemas de manufatura. Biblioteca Digital de Teses e Dissertações da Universidade Federal de São Carlos, 2010.

SHANNON, C. E. A mathematical theory of communication. **ACM SIGMOBILE Mobile Computing and Communications Review**, ACM, v. 5, n. 1, p. 3–55, 2001.

SHMULEVICH, I.; DOUGHERTY, E. R. **Genomic signal processing**. [S.l.]: Princeton University Press, 2014.

SNOEP, J. L.; WESTERHOFF, H. V. From isolation to integration, a systems biology approach for building the silicon cell. In: **Systems Biology**. [S.l.]: Springer, 2005. p. 13–30.

THIERENS, D.; GOLDBERG, D. Convergence models of genetic algorithm selection schemes. In: **Parallel problem solving from nature - PPSN III**. [S.l.]: Springer, 1994. p. 119–129.

TREPODE, N. W. **Modelagem do controle gênico do ciclo celular por redes genéticas probabilısticas**. Tese (Doutorado) — Texas A&M University, 2007.

WANG, Z.; GERSTEIN, M.; SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, Nature Publishing Group, v. 10, n. 1, p. 57–63, 2009.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of small-world networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998.

WEBB, A. R. **Statistical pattern recognition**. [S.l.]: John Wiley & Sons, 2003.