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

LEANDRO TAKESHI HATTORI

INFERÊNCIA DE REDES GÊNICAS COM ALGORITMO GENÉTICO E MODELO DE ILHAS

TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO

2013

LEANDRO TAKESHI HATTORI

INFERÊNCIA DE REDES GÊNICAS COM ALGORITMO GENÉTICO E MODELO DE ILHAS

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

Orientador: Prof Dr. Fabrício Martins Lopes

Co-orientador: Prof Henrique Yoshikazu Shishido

CORNÉLIO PROCÓPIO

2013

AGRADECIMENTOS

Agradeço a meus pais Wilson e Lúcia e a minha irmã Viviane pelo inestimável aprendizado que eles me proporcionaram ao longo da vida e pelo amor e apoio cultivados em nossa família.

Ao professor Fabrício M. Lopes, pela desmedida atenção, esforço e idéias dedicadas à colaboração para a realização do trabalho aqui apresentado.

Ao professor Henrique Y. Shishido, que apresentou boas idéias, inclusive a importante menção de sua disposição em dar apoio a pesquisas sobre a área de estudo em questão, o que me despertou para a possibilidade de realização do presente trabalho.

A todos os professores do curso, que foram tão importantes na minha vida acadêmica e no desenvolvimento deste trabalho.

Aos amigos e colegas, pelo incentivo e pelo apoio constantes.

RESUMO

HATTORI, Leandro. INFERÊNCIA DE REDES GÊNICAS COM ALGORITMO GENÉTICO E MODELO DE ILHAS. 49 f. Trabalho de Conclusão de Curso – Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2013.

Uma massiva quantidade de dados de expressões gênicas vem sendo produzidas devido ao desenvolvimento de técnicas de extração de informações moleculares como, por exemplo, a técnica de RNA-Seq. Este desenvolvimento tem como base o conceito do dogma central da biologia, em que o funcionamento de um organismo é baseado nas expressões de seus genes. Saber como é formado a estrutura de uma regulação gênica (GRN) pode contribuir para diversas aplicações como entender o funcionamento de determinadas doenças, análise de doenças genéticas e desenvolvimento de terapias e drogas mais eficientes. Então, técnicas computacionais estão sendo desenvolvidas para realizar a inferência destas redes de GRNs, buscando recuperar redes com alta precisão. A inferência de GRNs é um problema desafiador dado a grande quantidade de característica (milhares de genes) e poucas amostras (dados biológicos). Existem diversos métodos propostos na literatura para tal inferência, este trabalho aborda um método de seleção de características. A seleção de características é composta basicamente por uma função critério e algoritmo de busca. A função critério abordada neste trabalho é baseada na entropia, a qual tem o objetivo de avaliar os possíveis resultados de um determinado problema. O algoritmo genético e o modelo de ilhas foram as estratégias utilizadas para realizar as buscas dos possíveis candidatos para todos os gene da rede, sendo estes componentes o alvo de avaliação deste trabalho. Para inferir e validar as redes foram utilizadas Redes Gênicas Artificiais (AGNs), pois redes são passíveis de avaliação dado o conhecimento da estrutura, que permitem medir a eficiência dos métodos abordados. Os resultados experimentais baseados no desempenho dos algoritmos de buscas utilizando o modelo de ilhas obtiveram melhores resultados quando comparados ao algoritmo genético, entretanto o tempo computacional gerado pelo modelo de ilhas é superior ao tempo de execução do algoritmo genético.

Palavras-chave: Rede Gênica, Inferência, Algoritmo Genético, Modelo de Ilhas, Entropia, Redes Complexas, Reconhecimento de Padrões, Bioinformática.

ABSTRACT

HATTORI, Leandro. GENETIC ALGORITHM WITH ISLAND MODEL FOR GENETIC NETWORK INFERENCE . 49 f. Trabalho de Conclusão de Curso – Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2013.

A massive amount of data gene expression has been produced due to the development of techniques for the extraction of molecular information, for example, the technique of RNA Seq. This development is based on the concept as the concept of central dogma of biology, in which the operation of a body is based on the expression of their genes. Knowing how is formed the structure of a regulatory gene (GRN) may contribute to a variety of applications such as understanding the operation of certain diseases, analysis of genetic diseases and to develop therapies and more effective drugs. So, computational techniques are being developed to make the inference of these networks GRNs, seeking to recover networks with high accuracy. The inference of GRNs is a challenging problem given the large amount of features (thousands of genes) and few samples (biological data). There are several methods proposed in the literature for such an inference, this paper discusses a method of feature selection. Feature selection is basically composed by a criterion function and search algorithm. The criterion function addressed in this work is based on the entropy, which aims to evaluate the possible outcomes of a given problem. And the genetic algorithm and genetic algorithm with model islands were the strategies used to perform the search of possible candidates for all gene network, and these components the target evaluation of this work. To infer and validate networks were used genetic networks Artificial (AGNs), such networks are assessable given the knowledge of the structure, and measure the effectiveness of the methods discussed. Experimental results based on the performance of the algorithms search using the model of islands obtained better results when compared to the genetic algorithm, but the computational time generated by the model of islands is higher than the runtime of the genetic algorithm.

Keywords: Gene Network, Inference, Genetic Algorithm, Island Model, Entropy, Complex Network, Pattern Recognition, Bioinformatics.

LISTA DE FIGURAS

FIGURA 1 – Fluxo de processamento do Algoritmo Genético 22
FIGURA 2 – Estrutura de um indivíduo: cromossomo (binário) e seu respectivo *fitness* . 22
FIGURA 3 – Modelo de seleção por roleta 26
FIGURA 4 – Modelo de crossover apresentado em (a) é o modelo de 1-ponto e o (b) é o modelo de N-pontos 27
FIGURA 5 – Mutação de um cromossomo binário com uma representação binária 28
FIGURA 6 – Processo migratório entre as ilhas 30
FIGURA 7 – Exemplos de algumas topologias para o modelo de ilhas (a) topologia em anel, (b) topologia em estrela e (c) topologia e rede 30
FIGURA 8 – Medida de PPV obtida pela inferência de redes utilizando a estratégias de AG e AG com MI, aplicando 2,3 e 5 ilhas para o MI. Os valores da média do PPV obtido representa a execução de 10 experimentos. 38
FIGURA 9 – Média do PPV das redes inferidas pela topoloiga BA 39
FIGURA 10 – Média do PPV das redes inferidas pela topoloiga ER 39
FIGURA 11 – Média do PPV das redes inferidas pela topoloiga WS 40
FIGURA 12 – Média do tempo de todas as topologias dos algoritmo genético e do modelo de ilhas 41
FIGURA 13 – Média do tempo computacional gerado pela inferência da topologia BA pelos métodos de algoritmo genético e do modelo de ilhas 41
FIGURA 14 – Média do tempo computacional gerado pela inferência da topologia WS pelos métodos de algoritmo genético e do modelo de ilhas 42
FIGURA 15 – Média do tempo computacional gerado pela inferência da topologia ER pelos métodos de algoritmo genético e do modelo de ilhas 42

LISTA DE SIGLAS

RNA	Ácido Ribonucleico (*Ribonucleic Acid*)
GRN	Rede de Regulação de Gênica (*Gene Regulatory Network*)
DREAM	Diálogo para Avaliações de Métodos de Engenharia Reversa (*Dialogue for Reverse Engineering Assessment and methods*)
BNs	Rede Booleana (*Boolean Networks*)
AGNs	Rede Gênica Artificial (*Artificial Genetic Networks*)
AG	Algoritmo Genético
ER	Modelo de redes aleatórias (*uniformly-random*) de Erdös-Rényi
WS	Modelo de redes mundo pequeno (*small-world*) de Watts-Strogatz.
BA	Modelo de redes livre de escala (*scale-free*) de Barabási-Albert
TP	*True Positive*
FP	*False Positive*
FN	*False Negative*
TN	*True Negative*
PPV	*Positive Predictive Value*

SUMÁRIO

1 INTRODUÇÃO 13
1.1 OBJETIVOS 14
1.1.1 Objetivo Geral 14
1.1.2 Objetivos Específicos 14
1.2 JUSTIFICATIVA 15
1.3 ORGANIZAÇÃO 16
2 REVISÃO BIBLIOGRÁFICA 17
2.1 INFERÊNCIA DE REDES GÊNICAS 17
2.2 REDES COMPLEXAS 18
2.3 ENTROPIA E INFORMAÇÃO MÚTUA 19
2.4 ALGORITMO GENÉTICO 20
2.4.1 Componentes de um algoritmo genético 22
2.4.1.1 Indivíduo 22
2.4.1.2 População 23
2.4.2 Operadores 24
2.4.2.1 População Inicial 24
2.4.2.2 Função *Fitness* 24
2.4.2.3 Seleção 25
2.4.2.4 *Crossover* 27
2.4.2.5 Mutação 27
2.4.2.6 Critério de Parada 28
2.4.2.7 Elitismo 28
2.5 MODELO DE ILHAS 28
2.5.0.8 Migração 29
2.5.0.9 Topologia de Migração 30
3 MATERIAIS E MÉTODOS 32
3.1 APLICAÇÃO DO ALGORITMO GENÉTICO E MODELO DE ILHAS 33
3.1.1 Framework para Algoritmo Genético e Modelo de Ilhas 33
3.1.1.1 Framework Watchmaker 33
3.1.2 Configuração dos Operadores Genéticos 34
3.2 VALIDAÇÃO E ANÁLISE 35
3.2.1 Configuração do Ambiente de Execução 36
4 RESULTADOS E DISCUSSÃO 37
4.1 ALGORITMO DE BUSCA: AG E MI 37
5 CONCLUSÃO 43
5.1 TRABALHOS FUTUROS 43
REFERÊNCIAS 45

1 INTRODUÇÃO

Um organismo pode ser estudado com um aglomerado de reações bioquímicas. Uma complexa rede é formada pelo envio e recebimento de mensagens efetuadas por estas reações. Estas redes vêm sendo alvo de diversos estudos com o objetivo de entender os mecanismos de controle celular com base em entidades biológicas como, por exemplo, genes e RNA. Entretanto, ainda existe muito a ser descoberto sobre os mecanismos de controle celular.

Um modo de entender os mecanismos de controle celular é observando os dados temporais dos níveis de expressão dos genes. Com a evolução de técnicas de extração de informação molecular se tornou possível analisar grandes quantidades de genes e seus níveis de expressões como, por exemplo, a técnica de RNA-Seq (WANG et al., 2009). Um dos grandes desafios é conseguir recuperar uma rede de regulação de gênica (GRN) com base nos dados de expressão, devido uma grande quantidade de variáveis (genes) e poucos experimentos (amostras) produzidos. Então, métodos computacionais e estatísticos têm sido desenvolvidos buscando inferir GRNs com maior grau de similaridade possível a rede biológica.

A inferência de GRNs é fundamentada no dogma central da biologia molecular, o qual se baseia no estado funcional de um organismo a partir da expressão de seus genes (D'HAESELEER et al., 1999). Portanto, ao termos o conhecimento de uma GRN é possível analisar informações sobre seu funcionamento e comportamento celular. Como, por exemplo, o funcionamento de diversas vias regulatórias, ciclo celular, bem com o mapeamento de alterações provocadas por estímulos. Dado aos argumentos, tal problema é representado como um dos grandes desafios da bioinformática e alvo de pesquisas como o projeto DREAM (*Dialogue for Reverse Engineering Assessment and methods*) (STOLOVITZKY et al., 2007).

Com uma GRN definida é possível entender diversas características biológicas como suas interações moleculares e suas funções biológicas. Então, tais redes podem ser utilizadas para estudos de doenças e gerar prognóstico mais específicos, bem como medicamentos mais eficazes (CHANDRA; PADIADPU, 2013), e realizar estudos mais aprofundados sobre doenças como, por exemplo, o câncer.

Para a representação de uma GRN é possível utilizar genes binários e funções booleanas que definem a dinâmica de uma GRN por meio de um circuito lógico (KAUFFMAN, 1969). Ou seja, cada gene possui um conjunto de genes preditores, estas ligações formam uma rede ou também denominada na literatura de rede booleana (BNs). Estes modelos são de simples representação, e apesar desta propriedade tem apresentado bons resultados. Em simulações de BNs como *Drosophila melanogaster* (SANCHEZ; THIEFFRY, 2001), ciclo celular da levedura (LI et al., 2004), e entre outras pesquisas tem sido aplicados com êxito. Então, a partir destas BNs é possível gerar Redes de Genes Artificiais (AGNs, do inglês *Artificial Genetic Networks*) e então, aplicar e validar métodos de inferência de GRNs, dado que toda a estrutura da rede passa a ser conhecida.

Visto que as AGNs foram desenvolvidas e os dados dos perfis de expressão simulados podem ser gerados, é possível realizar o processo de inferência e validação das GRNs. E utilizar como base para a inferência de AGNs o método de reconhecimento de padrões como, por exemplo, o método seleção de características (JAIN et al., 2000). A seleção de característica é constituída de duas partes, um algoritmo de busca e a função critério. São exemplos de pesquisas sobre a inferência de GRNs utilizando seleção de característica (LIANG et al., 1998; HASHIMOTO et al., 2004; ZHOU et al., 2004; DOUGHERTY et al., 2008). O algoritmo de busca e a função critério utilizada neste trabalho são respectivamente o algoritmo genético (AG) (HOLLAND, 1975) com modelo de ilhas (PETTEY et al., 1987) e a função baseada na entropia de condicional média (LOPES et al., 2008), os quais serão apresentados nas Seções 2.4 e 2.3 respectivamente.

1.1 OBJETIVOS

1.1.1 OBJETIVO GERAL

O objetivo deste trabalho é apresentar a inferência de redes gênicas com algoritmo genético utilizando o modelo de ilhas. Neste estudo é apresentada uma análise comparativa entre a estratégia de busca com o algoritmo genético e o algoritmo genético com modelo de ilhas. Espera-se com este trabalho obter um resultado satisfatório em termos de qualidade da rede inferida baseado em algoritmo genético e algoritmo genético com modelo de ilhas.

1.1.2 OBJETIVOS ESPECÍFICOS

- Aplicar o algoritmo genético para inferência de redes gênicas.
- Aplicar o algoritmo genético com modelo de ilhas a para inferência de redes gênicas.

- Comparar o desempenho obtido pelos dois métodos adotados neste trabalho.

1.2 JUSTIFICATIVA

Devido ao aumento da complexidade dos modelos e a massiva quantidade de dados, a aplicação de algoritmos de buscas ótimos se tornaram infactível em razão ao seu extenso tempo de execução. Então, métodos de buscas computacionais foram desenvolvidos em que aplicam uma meta-heurística, no qual buscam apresentar o melhor resultado alcançado, mas não percorrendo todo o espaço de busca, a fim de reduzir o tempo computacional (LOPES, 2011).

Uma das classes de algoritmos buscas que utilizam heurísticas amplamente aplicado são os algoritmos evolutivos, que são baseados na Teoria da Evolução (DARWIN; BYNUM, 2009). O algoritmo genético é um dos algoritmos evolutivos que vem sendo aplicado em problemas de busca como, por exemplo, em (BRAGA, 1998; DAVIS, 1991; MICHALEWICZ, 1996; De Jong, 1975; BäCK; SCHWEFEL, 1993; da Costa Filho; POPPI, 1999). Sua flexibilidade a problemas é dada pela sua estrutura, na qual necessita basicamente de uma forma de representação dos possíveis resultados e uma função critério ou função *fitness*. Outra propriedade dos algoritmos genéticos é seu grau de paralelização, em razão dos indivíduos serem independentes uns dos outros e de serem avaliados individualmente (LUCAS, 2002).

O modelo de ilhas é uma técnica aplicada aos algoritmos evolutivos que utiliza a programação concorrente, evoluindo N populações simultaneamente e independentemente. Esta técnica busca melhorar o desempenho dos algoritmos de buscas dado ao contexto da crescente dimensionalidade dos problemas (ALBA, 1999; da Costa Filho; POPPI, 1999; CANTú-PAZ, 1998). Tal técnica tem apresentado sucesso devido ao aumento de desempenho e exploração de espaço de soluções (RUCIńSKI et al., 2010). Devido as vantagens obtidas ao aplicar a estratégia de modelo de ilhas diversos problemas vêm abordando esta estratégia (BäCK, 1994; BETHKE, 1976; BACK et al., 1997; URSEM, 2002).

No âmbito da área de bioinformática um dos grandes desafios é a inferência de redes gênicas tal área está sendo alvo de diversas pesquisas e novos modelos computacionais vêm sendo desenvolvidos e aplicados com sucesso. A importância destas redes se deve ao fato de que a partir do conhecimento destas estruturas é possível observar e realizar diversas análises como do comportamento das expressões de doenças, bem como gerar drogas mais eficazes. Técnicas como a seleção de características (LOPES et al., 2008) vem sendo propostas para modelar estas redes, esta técnica é composta por uma função critério e uma algoritmo de busca

(LOPES, 2011).

1.3 ORGANIZAÇÃO

Este trabalho segue com a contextualização da área através da revisão bibliográfica, na qual pode ser encontrada na seção 2. Os materiais e métodos serão apresentados na seção 3, na seção 4 será apresentado os resultados obtidos das metodologias aplicadas. Sequencialmente será apresentado na seção 5 a conclusão deste trabalho.

2 REVISÃO BIBLIOGRÁFICA

2.1 INFERÊNCIA DE REDES GÊNICAS

Os dados de expressões gênicas podem trazer de efeitos moleculares ou funções específicas de cada gene dado uma rede biológica específica. A inferência de genes ou também denominada engenharia reversa é um problema computacional desafiador devido a grande quantidade de variáveis (gene) e a baixa quantidade de experimentos (medidas). Tal problema tem sido alvo de investigação por diversos pesquisadores.

Um dos modos para se inferir uma rede de genes é utilizar dados temporais de expressão. Tal rede inferida poderá identificar vias regulatórias, ciclo celular e passível de observação caso seja aplicados estímulos.

A inferência de genes tem o objetivo de montar os relacionamentos entre os genes a partir de suas regulações e, principalmente, apresentar redes importantes da área biológica utilizando seus dados de expressão.

Na literatura existem basicamente três tipos de funções critério utilizadas para inferir GRNs. A correlação de Pearson é um dos tipos de função critério. Esta técnica define a ligação entre os genes baseados em uma limiar definido pelos perfis de expressão gênicas (STUART et al., 2003). Os tipos de métodos baseados em correlação consideram apenas o relacionamento 1 para 1, não considerando que um determinado gene alvo pode ser regulado por diversos outros genes.

O segundo método utilizado como função critério são as funções baseados em erro Bayesiano. Esse critério é amplamente utilizado para inferir GRNs, com esta função e o coeficiente de determinação (HASHIMOTO et al., 2004; LOPES et al., 2008). Este método permite a avaliação de relacionamentos 1 para N, ou seja, permite relacionar um gene alvo a um conjunto de genes preditores.

A função critério baseada na teoria da informação é bastante difundida na aplicação de inferência de GRNs (ver seção 2.3). Tal função permite determinar os relacionamentos tanto de

1 para 1, como na função baseada na correlação de Pearson, quanto de 1 para N (LIANG et al., 1998). Neste método é utiliza a uniformidade das distribuições de probabilidade condicional de um gene alvo (LOPES, 2011).

A inferência de GRNs vem sido alvo de diversos estudos e novos estudos vêm sendo desenvolvidos a cada dia, indicando a importância de sua aplicação na inferência de GRNs. Em um estudo recente apresentado por (CHANDRA; PADIADPU, 2013), o leque de descobertas que pode ser gerado a partir de GRNs inferidas, em áreas da farmacologia para estudos de desenvolvimento de drogas. A abordagem das modelagens é importante em diversos trabalhos como (D'HAESELEER et al., 2000; LOPES, 2011; De Jong, 2002; STYCZYNSKI; STEPHANOPOULOS, 2005; MARBACH et al., 2010).

2.2 REDES COMPLEXAS

As redes complexas são uma extensão da teoria dos grafos proposto por Leonard Euler (COSTA et al., 2007). O primeiro modelo de redes complexas desenvolvidos foi proposto por Paul Erdös e Alfréd Rényi (ER) em 1959 (ERDöS; ALFRéD, 1959). Posteriormente outros modelos foram desenvolvidos com o objetivo de representar sistemas reais como, por exemplo, o modelo mundo pequeno mundo-pequeno ou *small-word* (WS) (WATTS; STROGATZ, 1998) e livre de escala ou *scale-free* (BA) (BARABáSI; ALBERT, 1999).

Estas redes têm o objetivo de simular sistemas reais como, por exemplo, representar um sistema biológico (KAUFFMAN, 1971). Cada modelo de rede complexa apresenta distintas topologias e propriedades bem definidas, neste contexto as GRNs podem ser bem representadas.

Uma rede complexa é representada por um grafo que possui V_m vértices ligados por A_n arestas. Para gerar uma rede complexa dois parâmetros são definido o número de vértices (genes) e o grau médio $< k >$ de arestas.

A topologia ER desenvolvida pelos pesquisadores Erdös e Rényi (ERDöS; ALFRéD, 1959) é formada por ligações entre vértices de forma aleatória e com distribuição uniforme. Esta topologia tenta não realizar a ligação entre mesmos vértices e não gerar muitas conexões em um único vértice.

A topologia desenvolvida pelos pesquisadores Watts e Strogatz (WS) (WATTS; STROGATZ, 1998) não é realizada totalmente aleatória. Esta topologia é baseada ao fenômeno mundo pequeno (MILGRAM, 1967), o qual é baseado no conceito de que a média das distâncias de uma pessoa a qualquer outra é aproximadamente 6. Neste contexto, os vértices da topologia WS são ligados aos vértices mais próximos.

O modelo desenvolvido pelos pesquisadores Barabási e Albert (BA) (BARABáSI; ALBERT, 1999) é formada por muitos vértices pouco conectados a outros poucos vértices muito conectados. A formação da topologia dividida em duas etapas: o crescimento e a preferência linear do crescimento.

2.3 ENTROPIA E INFORMAÇÃO MÚTUA

A entropia da termodinâmica foi apresentada por Rudolf Clausius considerando apenas apresentações macroscópicas (CLAUSIUS, 1879). Em 1877, Ludiwing Boltzmann apresentou que a entropia de Clausius poderia ser representada em probabilidade ligada a configuração de sistema microscópico (BOLTZMANN et al., 1974), tal entropia ficou conhecida como entropia de Boltzmann-Gibbs. A forma discreta da entropia de Boltzmann é apresentada a seguir (TSALLIS, 2009):

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

sendo K a constante de Boltzmann e possuindo valor igual a 1 em áreas distinta da área da física (TSALLIS, 2009), e as probabilidades de p_i são equivalentes as W configurações microscópicas possíveis, logo:

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

A entropia, posteriormente, foi aplicada na Teoria da Informação pelo pesquisador Claude Shannon (SHANNON, 2001). A entropia desenvolvida permite indicar a quantidade de informação contida em uma determinada fonte, bem como possibilita graduar a desordem de um conjunto de dados (BISHOP, 1995). Dado uma variável aleatória X que pode assumir valores booleanos como, por exemplo, 0 e 1. A entropia de Shannon assim como a entropia de Boltzmann permite determinar em termos probabilísticos as possíveis ocorrências destas variáveis aleatórias ($P(x)$):

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

tal que

$$\sum_{x\in X} P(x) = 1. \tag{4}$$

Ou seja, a entropia de Shannon apresenta como resultado a medida da incerteza dada

uma determinada variável, então quanto maior o resultado da função, consequentemente maior será a incerteza de predizer tal variável. Fazendo uso de duas variáveis (X e Y) em conjunto, a entropia conjunta é definida:

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

no qual as varáveis aleatórias X e Y em conjunto é representado pela probabilidade de $P(x,y)$.

A entropia condicional é representada por $H(Y|x)$, tal entropia calcula a incerteza de uma variável aleatória Y dado o valor de uma instância da variável aleatória x conhecida. ou seja, quanto menor o resultado da entropia condicional maior serão as chances da variável Y predizer a variável x (KELEMEN et al., 2008). A entropia condicional é definida sequentemente:

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

A entropia condicional média é definida pela média ponderadas das estropias condicionais de todas as instâncias $x \in X$ (JUNIOR, 2008). A entropia condicional média é definida como:

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

no qual $H(Y|x)$ representa a entropia condicional e $H(Y|X)$ representa um valor no qual quanto menor o valor, maior será a informação de Y pela observação de X.

2.4 ALGORITMO GENÉTICO

Os algoritmos genéticos (AG) são algoritmos heurísticos utilizados para resolver problemas de busca e otimização. Tal algoritmo possui o objetivo de resolver problemas de buscas não-triviais, no qual algoritmos convencionais não seriam capazes de resolver em tempo acessível. Este algoritmo está classificado dentro da classe de algoritmos evolutivos, assim como os algoritmos de Programação Evolutiva (FOGEL, 2009) e Estratégia Evolutiva (RECHENBERG, 1978). O AG foi apresentado por John Holland (HOLLAND, 1975) e desenvolvido pelo pesquisador Goldberg (GOLDBERG, 1989).

Os AGs são inspirados nos princípios da Teoria da Evolução de Charles Robert Darwin (DARWIN; BYNUM, 2009). Aplicando os princípios de seleção e sobrevivência dos indivíduos

mais adaptados ao meio e consequentemente, maior probabilidade deste indivíduo gerar mais descendentes, bem como o conceito de hereditariedade e mutação genética. Ou seja, realizar a busca do melhor resultado (otimização) selecionando os indivíduos (soluções candidatas), com base em seus graus de adaptabilidade (qualidade dos resultados), sendo que os descendentes herdarão as características de seus progenitores que foram selecionados, que serão passíveis a mutações genéticas (aplicação de métodos computacionais inspiradas em operadores naturais).

O AG possui a propriedade de ser amplamente adaptável a problemas, visto que apenas é necessária uma forma de representação do problema e uma função de avaliação dos possíveis resultados. Devido a esta propriedade os AGs foram aplicados em diversos problemas de busca e otimização com sucesso (BRAGA, 1998; DAVIS, 1991; MICHALEWICZ, 1996; De Jong, 1975; BäCK; SCHWEFEL, 1993; SANCHES et al., 2008; LOPES, 1999).

Outra propriedade que o AG possui é um alto grau de paralelização, devido aos dados serem altamente independestes. O argumento apresentado pode ser explicado dado ao fato dos indivíduos de uma população ser avaliados de forma independente. E da mesma forma que na natureza existem diversas subpopulações evoluindo concorrentemente, igualmente pode acontecer em um AG (LUCAS, 2002), esta técnica é conhecida com Modelo de Ilhas (ver seção 2.5).

Os AGs realizam operações de seleção natural atuando sobre os componentes (indivíduos e população). O indivíduo é uma estrutura de dados que armazena a codificação de um possível resultado (ver seção 2.4.1.1). A população é composta por um conjunto de indivíduos (ver seção 2.4.1.2), na qual é subordinada a aplicação de operadores baseados na seleção natural, conforme apresentado na Figura 1.

Os operadores realizados por um AG são: a criação de uma população inicial e em seguida são aplicados um conjunto de operadores que se repetirão N vezes até atender a um determinado critério de parada. A cada iteração deste conjunto de operadores é realizada uma geração, tal conjunto de operadores pode ser observado na Figura 1. A cada geração são aplicados aos indivíduos da população os operadores do calculo da função *fitness*, seleção, *crossover* e mutação.

Tais componentes e operadores apresentados são descritos respectivamente nas seções 2.4.1 e 2.4.2. Entretanto, ainda existem alguns outros operadores não apresentados na Figura 1, como o operador de elitismo (ver seção 2.4.2.7), que pode ser adicionado ao AG.

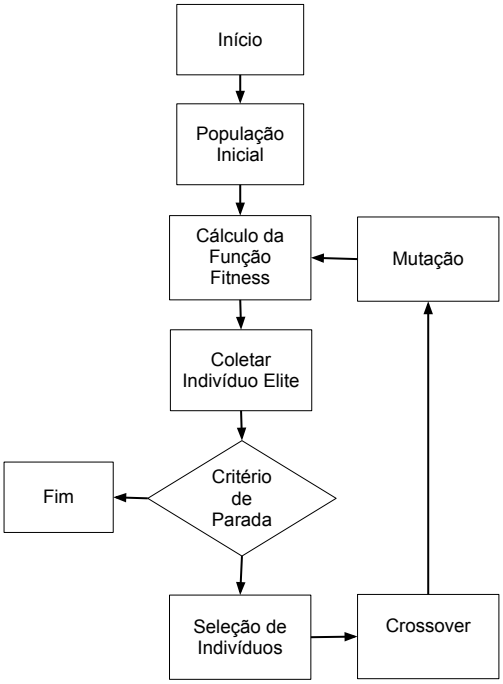

Figura 1: Fluxo de processamento do Algoritmo Genético

2.4.1 COMPONENTES DE UM ALGORITMO GENÉTICO

2.4.1.1 INDIVÍDUO

Um indivíduo é uma estrutura que armazena um cromossomo e seu respectivo *fitness*, conforme pode ser observado na Figura 2.

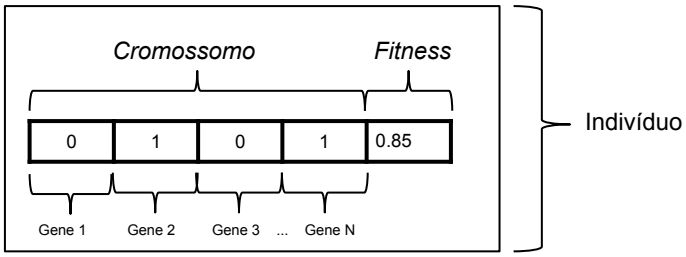

Figura 2: Estrutura de um indivíduo: cromossomo (binário) e seu respectivo *fitness*
Fonte: (ALBA, 1999)

O cromossomo ou genótipo é composto por um conjunto de genes, formando uma cadeia de string sequenciais, representada por valores binários (0 e 1), ou por letras do alfabeto (A - Z) como observado no campo cromossomo da Tabela 1. Entretanto, ainda existem outras representações de um cromossomo como número reais, permutação de símbolos repetidos

(PACHECO, 1999).

O campo fenótipo da Tabela 1 contém os dados observáveis referentes a representação dos dados do cromossomo (binário, decimal, entre outros), portanto passível de avaliação. No qual, será utilizado como parâmetro da função fitness. Ou seja, o cromossomo representado por 000110 é convertido de binário para decimal gerando o fenótipo 6, que seFrá passado como parâmetro para a avaliação do indivíduo (função *fitness*). Entretanto, o mesmo cromossomo 000110 pode ser divido em três partes e gerar três fenótipos, gerando três entradas como parâmetro de uma determinada função, tal abordagem vai depender da necessidade do problema. Existem casos em que os dados contidos no cromossomo são os mesmos do fenótipo, e não precisam ser convertidos, como no caso do problema do cacheiro viajante.

Os indivíduos também armazenam seu *fitness*, que representa quantitativamente o quão bom o resultado que o determinado indivíduo representa para o determinado problema. Este valor é calculado no processo de operação da função *fitness* no qual é referenciado na seção 2.4.2.2. Tal valor é utilizado para gerar uma probabilidade do indivíduo a ser selecionado para reproduzir, e propagar seus genes (ver seção 2.4.2.3). O *fitness* também é utilizado para o indivíduo ter a possibilidade de sobreviver em outras gerações, por meio do operador de elitismo (ver seção 2.4.2.7).

Tabela 1: Variação dos tipos de cromossomos, seus respectivos fenótipos e problema abordado

Cromossomo	Fenótipo	Problema
000110	6	Busca de um número inteiro
000110	0, 1 e 2	Busca de três número inteiros
AGDEF	percorre nodo A, G, D, E e F	Problema do cacheiro viajante

Fonte: (LUCAS, 2002)

2.4.1.2 POPULAÇÃO

A população é composta por um conjunto de indivíduos. Os indivíduos da população inicial são gerados por um determinado método a partir do operador da população inicial (seção 2.4.2.1). E após N gerações, a população será evoluída até atender a um determinado critério de parada (ver seção 2.4.2.6).

Mesmo com a evolução da população, a quantidade de indivíduos se mantém a mesma. A cada geração seus indivíduos são selecionados, cruzados e mutados gerando descendentes no qual serão substituídos pelos indivíduos anteriores. Os descendentes por sua vez serão avaliados da mesma forma que seus ancestrais e assim sucessivamente até ser atendido o critério de parada.

Após uma determinada quantidade de gerações executadas, uma população tende a um grau de convergência. Este grau representa a variação da média do *fitness* de uma população dada a média da população anterior. O objetivo do AG é convergir para um resultado ótimo global e não um ótimo local, entretanto não existe garantia que isso ocorrerá (WHITLEY, 1994; GOLDBERG, 1989; LUCAS, 2002; SRINIVAS; PATNAIK, 1994). Ou seja, dada a evolução da população, a convergência prematura não deve ocorrer, devido ao espaço de busca percorrido ser pequeno e as chances de achar um resultado ruim aumentar (RUDOLPH, 1994).

O objetivo do AG é convergir para um valor ótimo (máximo global) e evitar a convergência prematura da população (mínimo local). O aumento da diversidade genética (variedade de fenótipos gerados) aumenta a amplitude da busca e por consequente maiores as expectativas de convergir para um ótimo global. Para evitar a convergência prematura é importante manter a diversidade genética da população. A técnica de mutação auxilia a manter a variabilidade genética da população (apresentado na seção 2.4.2.5) e utiliza um método que busque equilibrar a seleção dos melhores indivíduos e que mantenha a diversidade genética, no qual o artigo (THIERENS; GOLDBERG, 1994) aborda a avaliação de alguns métodos de seleção.

2.4.2 OPERADORES

2.4.2.1 POPULAÇÃO INICIAL

A população inicial representa o primeiro operador do algoritmo genético sendo executada apenas uma vez. A população inicial tem como objetivo criar os primeiros indivíduos que irão compor a população. A criação dos indivíduos pode ocorrer de forma aleatória, ou se basear em algum método específico de inicialização dos indivíduos, dado o conhecimento a priori de bons cromossomos (PACHECO, 1999).

2.4.2.2 FUNÇÃO *FITNESS*

A função *fitness* ou função objetivo é o operador responsável pela avaliação dos indivíduos da população (resposta para o problema) a cada geração executada. A avaliação consiste em expressar a qualidade de um problema em forma de uma função matemática (LUCAS, 2002), passando como parâmetro o fenótipo do indivíduo (ver seção 2.4.1.1) e obtendo com retorno da função a qualidade do indivíduo (PACHECO, 1999). A função é aplicada a todos os indivíduos da população, para que o AG possa manipular estes dados em outros operadores.

Tal operador de avaliação do *fitness* do indivíduo é fundamental para o AG, visto que o *fitness* é utilizado como medida de adaptabilidade do indivíduo, ou seja, aumentando ou diminuindo suas chances de reprodução e sobrevivência (PACHECO, 1999).

Para que o operador funcione de forma eficiente a função *fitness* deve ser a mais representativa possível. Pois, a partir de uma melhor representação é possível ser mais expressivo nos operadores do AG e, por consequente, chegar a um resultado de forma mais rápida e precisa. Entretanto, para alguns tipos de problemas não é possível calcular com exatidão o grau de aptidão de um indivíduo, como problemas de predição de genes para inferência de

2.4.2.3 SELEÇÃO

O operador de seleção é responsável pela seletividade dos indivíduos da população para a reprodução (PACHECO, 1999), utilizando um determinado método probabilístico baseando nos *fitness* dos indivíduos da população. A seleção é executada imediatamente após o operador da função critério. Este operador possui diversos métodos que podem ser implementados (GOLDBERG; DEB, 1991). Independentemente do método utilizado, os indivíduos com melhor *fitness* terão mais chances de serem selecionados. Entretanto, é importante observar que no processo de seleção não sejam contemplados apenas os melhores indivíduos, mas também indivíduos com *fitness* variados. A seleção apenas dos indivíduos melhores adaptados pode gerar uma convergência prematura dos resultados (THIERENS; GOLDBERG, 1994). Alguns dos métodos existentes são resumidamente descritos a seguir (DYER, 2010).

O método *Stochastic Universal Sampling* é o método de roleta de forma mais elaborada. Este método garante que um individuo que possui *fitness* de melhor qualidade seja selecionado proporcionalmente.

No método *Tournament Selection* são selecionados N indivíduos aleatoriamente formando subgrupos de tamanho N. O indivíduo que possuir melhor *fitness* dentro do subconjunto será selecionado.

O *Truncation Selection* é o método de seleção mais simples, tem o objetivo selecionar os uma quantidade de melhores indivíduos da população e a partir deste subconjunto gerar os indivíduos da próxima população.

O método de *Rank Selection* também é baseado em uma roleta como no método de seleção *Roulette Wheell Selection* e *Stochastic Universal Sampling*. A construção da roleta é organizada por faixas uniformes e posicionamento ordenado. Ou seja, todos os indivíduos da

população possuem o mesmo tamanho das faixas e sendo ordenados do melhor para o pior *fitness*.

Embora os métodos apresentados a pouco tenham sua importância, neste trabalho foi adotado o método *Roulette Wheel Selection*, o qual é apresentado a seguir com informações mais detalhadas que os demais. Neste método os indivíduos estão contidos em uma faixa da roleta, sendo o tamanho da faixa relacionada proporcionalmente ao seu *fitness*. Essa seleção proporcional torna possível tanto a escolha de indivíduos melhores adaptados quanto daqueles não tão bem adaptados. Para cada indivíduo da população é calculada a seguinte equação (GUIMAR; CAMPELO, 2011):

$$ps = \frac{f(x_i)}{\sum_{j=1}^{n} if(x_j)}, \tag{8}$$

na qual $f(x_i)$ representa o valor do *fitness* do indivíduo x_i que está sendo avaliado e n_i representa o valor total da população. Ou seja, indivíduos com *fitness* melhores terão uma porção maior da roleta, e consequentemente, maior probabilidade de serem selecionados, como na Figura 3 em que o indivíduo 1 terá maior probabilidade de ser selecionado, devido seu *fitness* ser de maior qualidade que dos demais.

Após a construção da roleta é gerado um número aleatório, então é verificado em qual faixa da roleta em que este número se encontra e a partir da verificação é selecionado o indivíduo pertencente a tal faixa da roleta. Na Figura 3, apesar do indivíduo 2 possuir um faixa menor que outros indivíduos o mesmo foi selecionado. Ou seja, este modelo possibilita a seleção dos indivíduos com diversas qualidade de *fitness*. Este método requer um nível de processamento elevado, em razão da montagem da roleta ser realizada diversas vezes. Devido, a necessidade de percorrer todos os indivíduos da população para realizar os cálculos da proporção de cada individuo da roleta.

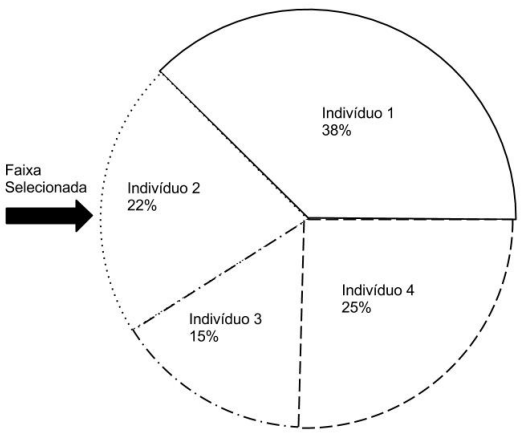

Figura 3: Modelo de seleção por roleta

2.4.2.4 *CROSSOVER*

Depois de selecionado os indivíduos da população em casais (pares de indivíduos selecionados, denominados pais), tais pares irão passar pelo operador de *crossover* e gerar dois novos indivíduos, denominados filhos. O termo *crossover* é uma analogia utilizada para representar o cruzamento do cromossomo entre dois indivíduos, que representam biologicamente a troca de material genético (PACHECO, 1999). Existem diversos métodos de *crossover*, nos quais a diferença basicamente está na variação da quantidade e nos locais dos pontos de cruzamento (Eiben, A.E. and Smith, 2003).

Existem diversas técnicas desenvolvidas para a realização do *crossover* como o *crossover* de um ponto, representado pelo exemplo 'a' na Figura 4. Primeiramente, tal técnica inicia pareando os cromossomos selecionados, então é escolhido uma posição que representará o ponto de quebra do cromossomo, as duas partes serão trocadas (HOLLAND, 2000).

Outra técnica é o *crossover* de N-pontos apresentada na Figura 4 pelo exemplo 'b'. Tal técnica seleciona duas ou mais posições do cromossomo, gerando a quebra do cromossomo nestas posições. Então, os pedaços de cromossomos gerados serão trocados, ao final do processo dois novos cromossomos serão formados (Eiben, A.E. and Smith, 2003).

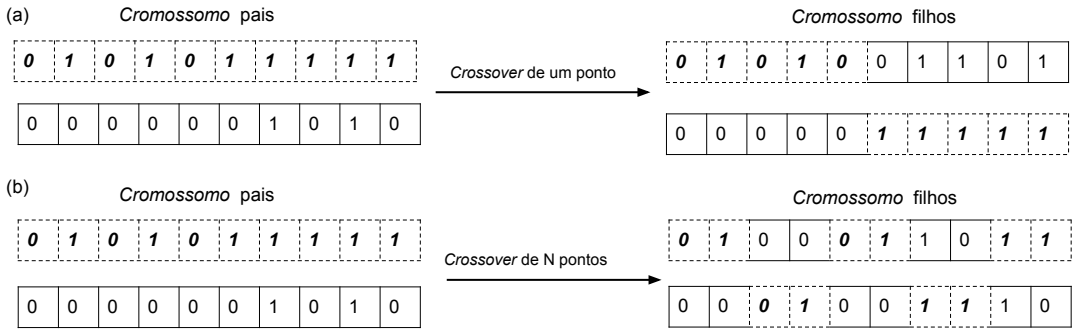

Figura 4: Modelo de crossover apresentado em (a) é o modelo de 1-ponto e o (b) é o modelo de N-pontos

2.4.2.5 MUTAÇÃO

O operador de mutação é aplicado no cromossomo dos indivíduos filhos, tal operador muta uma determinada posição do cromossomo dada certa probabilidade (DYER, 2010). Na Figura 5 para cada posição do cromossomo filho, é gerado um número aleatório que determinará se a posição do cromossomo vai ser mutado ou não, dado uma probabilidade definida. A Figura 5 apresenta o processo de mutação do cromossomo dada um cromossomo binário, mutando

a primeira posição para seu valor inverso, de zero para um. O operador agrega uma boa característica ao processo do AG, a qual impede que uma população converja para um mínimo local (RUDOLPH, 1994).

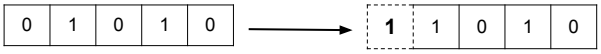

Figura 5: Mutação de um cromossomo binário com uma representação binária

2.4.2.6 CRITÉRIO DE PARADA

O critério de parada é definido por meio de um operador condicional que pode representar o término das iterações da execução do AG, ou a execução de uma próxima geração, dado a uma determinada condição. O critério de parada também pode ser definido como o controlador do processo iterativo de um AG (PACHECO, 1999). Como, por exemplo, caso encontre o melhor *fitness* caso encontre uma resposta com *fitness* aceitável ou executar G gerações pré estabelecidas, mesmo que tal busca não apresente uma resposta aceitável (HOLLAND, 2000).

2.4.2.7 ELITISMO

O operador de elitismo é muito utilizado nas implementações do AG (HOLLAND, 2000). Este operador permite que uma determinada quantidade de indivíduos de uma geração possa ser alocada na próxima geração (MAJUMDAR; BHUNIA, 2007). Esta técnica é útil, devido ao conjunto dos melhores individuos *fitness* continuarem na próxima geração, logo a qualidade do *fitness* não irá decair. Entretanto, um cuidado é necessário em grandes quantidades de indivíduos elite, em razão da minimização da variedade genética da população, aumentando as chances da convergência prematura.

2.5 MODELO DE ILHAS

Um dos problemas desafiadores que vem surgindo nas pesquisas atuais é o aumento expressivo da dimensionalidade e da complexidade dos problemas (ALBA, 1999; da Costa Filho; POPPI, 1999; CANTú-PAZ, 1998). Algoritmos como os AGs necessitam de alto processamento computacional. Então, estratégias de paralelizarão foram propostas para aumentar o desempenho destes algoritmos de buscas e otimização.

O modelo de ilhas (MI) é um modelo de evolução de multipopulações (ilhas) que ocorre simultaneamente, a fim de melhorar o desempenho dos algoritmos evolutivos. A teoria base para a inspirarão do MI é pertencente a Teoria da Evolução do Equilíbrio Pontuado (COHOON et al., 1987). As primeiras pesquisas sobre os MIs foram desenvolvidas por Jettey e colaboradores aplicadas para os AGs (PETTEY et al., 1987). Entretanto, os MIs também podem ser aplicados a outros algoritmos evolutivos como a evolução diferencial (ED) (GUIMAR; CAMPELO, 2011). Assim como o AG, o MI é bastante flexível e consegue bordar diversos tipos de problemas (BäCK, 1994; BETHKE, 1976; BACK et al., 1997; URSEM, 2002).

A possibilidade de evolução paralela entre as populações se deve a propriedade dos algoritmos evolutivos serem fracamente acoplados, possibilitando sua decomposição (ver seção 2.4). Neste modelo as multipopulações evoluem independentemente e paralelamente, sendo que em determinadas gerações, as populações realizam trocas de indivíduos (COHOON et al., 1987). Ou seja, a evolução é independente das multipopulações e gera uma competição para encontrar o melhor indivíduo, entretanto existe uma cooperação entre as multipopulações, que ocorre por meio do processo de migração (GUIMAR; CAMPELO, 2011).

2.5.0.8 MIGRAÇÃO

A migração é o operador utilizado no modelo de ilhas que orienta a troca de indivíduos entre as ilhas a cada geração (GUIMAR; CAMPELO, 2011). A migração permite que indivíduos sejam transferidos de uma população para outra, assim como apresentado na Figura 6. A transferência de indivíduos de uma ilha para outra pode aumentar a diversidade genética da população, assim como o operador de mutação (ver seção 2.4.2.5), diminuindo as possibilidades de que a busca não converta para um mínimo local (LOPES et al., 2012).

A taxa de migração do MI indica o número de indivíduos que será migrado entre as ilhas. Indivíduos são selecionados de uma ilha de forma aleatória e transferidos para uma outra ilha, dado uma determinada topologia de migração (ver seção 2.5.0.9).

As altas taxas de migrações podem conduzir a boa propriedade gerada ter um efeito contrário (TANESE, 1989). Outra variável que interfere no desempenho de um MI é a quantidade de migrações que devem ocorrer. Ou seja, a cada quantas gerações o processo de migração deve ocorrer, este problema é acentuado principalmente em MI que são síncronos. A maioria dos MI desenvolvidos realizam a troca de indivíduos de forma síncrona (ALBA, 1999). Ou seja, para realizar o processo de migração de uma ilha para outra, o processo de evolução de todas as ilhas é sincronizado, e então todas as ilhas realizam a troca indivíduos. Gerando ao final do processo um *overhead* de tempo e ao aumento do tempo de execução do algoritmo.

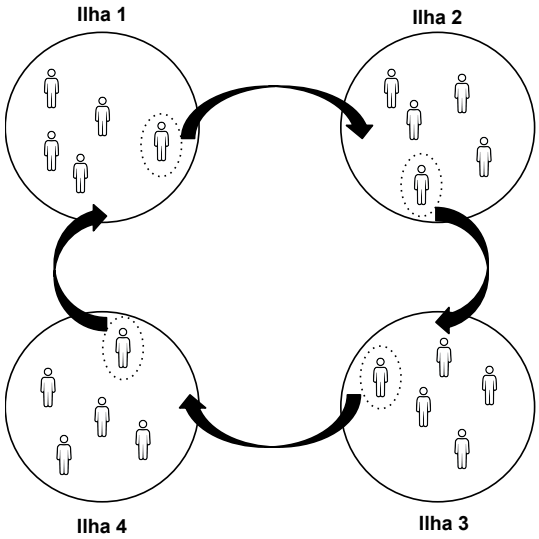

Figura 6: Processo migratório entre as ilhas
Fonte: (GUIMAR; CAMPELO, 2011)

2.5.0.9 TOPOLOGIA DE MIGRAÇÃO

A topologia do MI é a estrutura formada pelo fluxo migratório entre as ilhas (LOPES et al., 2012). A estrutura é formada por arestas (fluxo migratório) que interligam nodos (ilhas). Esta estrutura define um conjunto de nodos vizinhos como apresentado na Figura 6, na qual apresenta três topologias distintas.

A topologia de migração não muda durante o processo evolutivo é denominada topologia estática (TANG et al., 2004). Três topologias comumente utilizadas são apresentadas na Figura 7 no exemplo (a) representa a topologia em anel, na qual a migração é realizada de forma cíclica sendo a topologia utilizada neste trabalho. No exemplo (b) representa a topologia em estrela, na qual as ilhas se comunicam por meio de um nó central. A última topologia (c) é denominada topologia de rede, em que todas as ilhas comunicam se comunicam entre si.

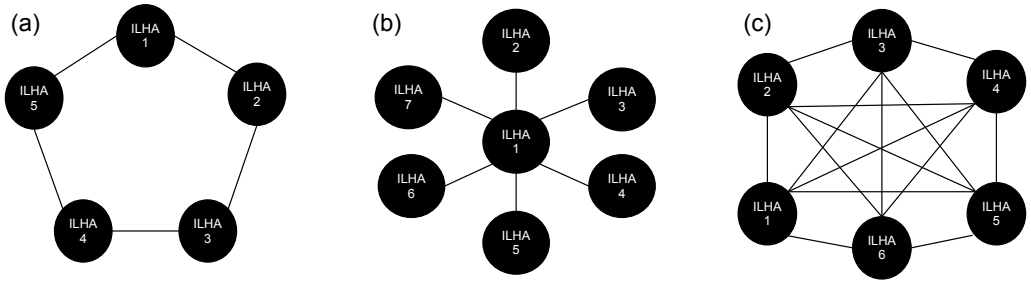

Figura 7: Exemplos de algumas topologias para o modelo de ilhas (a) topologia em anel, (b) topologia em estrela e (c) topologia e rede

A escolha de uma determinada topologia pode interferir na aplicação de um MI (RUCIńSKI et al., 2010) aumentando ou diminuindo o desempenho do algoritmo utilizado. Existem diversos tipos de topologias descritas na literatura (TANG et al., 2004; BERNTSSON; TANG, 2005; LARDEUX; GOëFFON, 2010; LOPES et al., 2012).

3 MATERIAIS E MÉTODOS

A inferência de GRNs é um problema desafiador na área de pesquisa de bioinformática, devido a enorme quantidade de genes e um número reduzido de experimentos biológicos realizada. Esse contexto leva a um problema conhecido como maldição da dimensionalidade (BISHOP, 1995; JAIN et al., 2000), que consiste na necessidade de grandes quantidade de dados para uma classificação consistente. Ou seja, é necessário uma grande quantidade de observações como, por exemplo, de dados biológicos para representar uma GRNs de qualidade. Neste contexto, a redução da dimensionalidade se torna um fator crucial.

Para a aplicação da redução de dimensionalidade, podem ser adotadas as técnicas de extração de características e a seleção de características (WEBB, 2003; CAMPOS, 2001). A abordagem de extração de características classifica uma determinada característica é fornecida com entrada a partir da coleta de dados mais relevantes (DEVIJVER; KITTLER, 1982). A seleção de características utiliza uma função critério e um algoritmo de busca que possibilite a classificação dos objetos (LOPES, 2011). Este trabalho apresenta um estudo especifico do método de seleção de características.

A seleção de características é baseada em uma função critério e um algoritmo de busca. A função critério é uma função que possibilita representar um conjunto de características por um valor de qualidade. Neste método, o algoritmo de busca a ser adotado poderá ter solução ótima ou sub-ótima.

Os algoritmos de buscas ótimos são algoritmos que percorrem todo o espaço de busca e retornam a solução ótima. Entretanto, estes métodos demandam por um alto poder computacional e por alto tempo de processamento. Então, para problemas de grande dimensionalidade com a inferência de GRNs, tais métodos se tornam inviáveis. Logo, os métodos de buscas sub-ótimos são os modelos mais adequados para resolver tais problemas de grande dimensionalidade. Estes métodos não garantem a melhor solução global, entretanto as buscas utilizando heurísticas vem apresentando bons resultados a um custo computacional razoável.

Este trabalho apresenta uma abordagem de otimização do método de inferência de redes gênicas aplicando uma heurística de busca baseado no AG (ver seção 2.4) e paralelizada pelo modelo de ilhas (ver seção 2.5). Para realizar a validação foram utilizadas as AGNs (ver seção 3.2), as quais possibilitam a validação dos métodos de inferência de GRNs. A utilização deste método de busca pode gerar muitos questionamentos como: O AG e o MI conseguem apresentar um resultado com o número de acerto desejável? Utilizando o MI a qualidade e o tempo para inferir uma rede gênica melhora em relação ao AG?

3.1 APLICAÇÃO DO ALGORITMO GENÉTICO E MODELO DE ILHAS

Dado o contexto de que a dimensionalidade e a complexidade dos problemas vêm aumentando significativamente, métodos tradicionais que percorrem todo o espaço de busca vem perdendo força, devido ao grande esforço computacional para realizar seu objetivo. Neste contexto, pesquisas vêm sendo desenvolvidas para encontrar métodos heurísticos que apresentem bons resultados com razoável tempo computacional de forma a viabilizar pesquisas.

Inspirado neste cenário, este trabalho propõe para a inferência de GRNs um método de busca inspirado na Teoria Evolucionista de Darwin (DARWIN; BYNUM, 2009) denominado algoritmo genético (AG) (HOLLAND, 1975) e aplicado a programação concorrente, fundamentado na teoria da evolução denominada Equilíbrio Pontuado (COHOON et al., 1987) com modelo de ilhas (PETTEY et al., 1987).

3.1.1 FRAMEWORK PARA ALGORITMO GENÉTICO E MODELO DE ILHAS

A escolha do *framework* é uma das etapas fundamentais para o processo de desenvolvimento do projeto. Existem *frameworks* em diversas linguagens para aplicar tal algoritmo e escolher um *framework* que atenda as necessidades do desenvolvimento é fundamental como (WALL, 1996), (MEFFERT et al., 2011), (DYER, 2010) e entre outros. No problema abordado foi necessário um *framework* que atendesse a alguns critérios de desenvolvimento, com ser em java e fornecesse o recurso do modelo de ilhas, o *framework* que abordava estes requisitos foi o *framework watchmaker* (DYER, 2010), o qual foi adotado para a execução.

3.1.1.1 FRAMEWORK WATCHMAKER

O *Watchmaker* é um *framework* que permite a abstração do desenvolvimento do algoritmo genético e do modelo de ilhas sendo implementada na linguagem Java (DYER,

2010). Também é apresentado diversos outros recursos como métodos de seleção como *Roulette Wheel Selection* , *Tournament Selection*, entre outros, os quais são apresentados na seção 2.4.2.3. Outros recursos disponíveis são os operadores de elilistimo (ver seção 2.4.2.7), critério de parada (ver 2.4.2.6) e facilidade para inserir a função *fitness* do problema abordado. Documentação e *feedback* de qualidade são outros recursos importantes, facilitando o processo de aprendizagem e aplicação do *framework*.

3.1.2 CONFIGURAÇÃO DOS OPERADORES GENÉTICOS

- **Indivíduos**: para a representação do cromossomo foi adotado valores discretos e gerados 5 fenótipos de cada cromossomo (veja seção 2.4.1.1), Nos quais representam os possíveis genes preditores.
- **População**: o tamanho da população para o AG é de 250 indivíduos e para cada ilha do MI é definida uma quantidade de 50 indivíduos por ilha, e ilhas de tamanho 2, 3 e 5.
- **População Inicial**: o método aleatório foi adotado para gerar os indivíduos da população inicial.
- **Função *Fitness***: a função utilizada para os testes do AG e MI é baseada na entropia condicional média descrito na Seção 2.3, na qual foi desenvolvida por (LOPES, 2011).
- **Seleção**: para realizar a seleção dos indivíduos foi utilizado o método *Roulette Wheel Selection* (ver seção 2.4.2.3).
  - **Elitismo**: o método de elitismo também foi aplicado ao processo evolutivo da população (ver seção 2.4.2.7). Sendo adotada uma taxa de 2% dos melhores indivíduos da população.
- ***Crossover***: o método de *crossover* aplicado é o *crossover* de um ponto descrito na seção 2.4.2.4.
- **Mutação**: o operador de mutação é definido como mutado de acordo com um alfabeto binário (0 e 1) com probabilidade de 5% de cada gene do cromossomo ser mutado.
- **Critério de Parada**: o critério de parada utilizada neste trabalho é baseado na quantidade de gerações executadas (ver seção 2.4.2.6). A quantidade para o AG e MI foi definida 200 e 70 gerações para a busca dos preditores para cada gene.
- **Modelo de ilhas**: a abordagem de topologia utilizada neste trabalho é a topologia de migração em anel (ver seção 2.5). Sendo definida que a taxa de migração de indivíduos

de uma população para outra será de 2% a cada 20 gerações (veja seção 2.5.0.8). E utilizado o algoritmo evolutivo nas ilhas o próprio AG.

3.2 VALIDAÇÃO E ANÁLISE

Com o desenvolvimento de AGNs possibilitou que métodos de inferência de GRNs sejam avaliados. A AGN utilizada neste trabalho foi desenvolvido por (LOPES et al., 2011). Uma AGN é constituída por vértices e arestas, no qual os vértices representam os genes da rede e as arestas representam as ligação regulatória entre os genes.

A AGN é representada por uma matriz de adjacência, em que a aresta do vértice v_i para o vértice v_j é representada $M(i, j)$. A posição $M(i, j)$ pode assumir valor 0 ou 1, no qual vai indicar se os vértices v_i e v_j estão ligados ou não.

Para a avaliação será utilizada uma matriz de adjacências gerada por uma AGN (LOPES, 2011) e a matriz de adjacências inferida pelo AG e MI abordado neste trabalho, para verificar a acurácia é proposto o uso da similaridade entre as redes. As medidas de avaliação entre as duas redes é descrito em (DEVIJVER; KITTLER, 1982), as quais são baseadas na matriz de adjacência (WEBB, 2003), como exibe o Quadro 1.

Neste trabalho as possíveis variáveis para a matriz de adjacência são: TP(*True Positive*), são as arestas que foram inferidas e existem na rede real (AGN), o FP(*False Positive*), são as arestas que foram inferidas pelos modelos propostos, entretanto não existem na rede original, o FN(*False Negative*), são as arestas que não foram inferidas e que existem na rede real e o TN(*True Negative*), representam as arestas que não foram inferidas e que não existem na rede real.

Quadro 1: Possíveis valores para qualificar a similaridade de uma rede (a) AGN com uma outra rede (b) rede inferida pelo algoritmo de busca, sendo TP *True Positive*, FP *False Positive*, FN *False Negative* e TN *True Negative*.)

Aresta \ Conexão	**Rede B**	**Rede B**
Rede A	TP	FN
Rede A	FP	TN

Fonte: (LOPES, 2011)

A medida utilizada para medir a qualidade das redes inferidas é apresentadas a seguir:

$$PPV = \frac{TP}{(TP+FP)}, \tag{9}$$

na qual PPV é altamente utilizado para medir a qualidade das AGN inferidas, tal equação

também conhecida como precisão, quanto maior seu valor maior será a previsão.

Para a validação dos experimentos referente ao desempenho dos AG e o AG com MI também é utilizado como critério o tempo de execução da inferência dos mesmos algoritmos. Sendo que todos os experimentos realizados nos mesmos componentes de Hardware e Software, as especificações são apresentadas na Seção 3.2.1.

3.2.1 CONFIGURAÇÃO DO AMBIENTE DE EXECUÇÃO

Para a execução dos experimentos foi utilizado um hardware com configuração de processamento que permita a execução de múltiplas *threads*, nos quais são representados pelas a execução das ilhas em paralelo pelo MI (ver seção 2.5). Tal tecnologia é oferecida e desenvolvida pelo *framework watchmaker* (ver seção 3.1.1.1). A configuração do hardware utilizado para os testes foi adotado a seguinte especificação do hardware:

- Modelo do Processador: Intel(R) Core(TM) i7-3820 CPU @ 3.60GHz
- Tamanho da Memória *Cache*: 10240 KB
- Qtd. de Núcleos de Processamento (Físico e Lógico): 8
- Tamanho da Memória Principal (RAM): 8Gb
- Sistema Operacional: Distribuição Linux Kubuntu 13.03

4 RESULTADOS E DISCUSSÃO

Neste capítulo é apresentada a avaliação de duas metodologias de buscas para os preditores de todos os genes de uma determinada rede. Neste contexto, os algoritmos propostos foram o AG e AG com MI (ver seção 3.1). Os resultados de tais experimentos são apresentados a seguir.

4.1 ALGORITMO DE BUSCA: AG E MI

Nesta seção são apresentados os resultados obtidos a partir da aplicação do algoritmo genético e do algoritmo genético com modelo de ilhas, tendo o objetivo inferir redes de regulação gênica. Para a inferência de redes foram utilizados as AGNs e a função critério descritos em (LOPES et al., 2011).

Para todos os experimentos foram utilizados redes contendo 100 genes, o tamanho do sinal, ou instâncias de tempo observado, igual a 100. O grau médio das ligações entre os genes $< k >$ igual a 2. As topologias aplicadas foram Barabási-Albert (BA), Erdös-Rényi (ER), Watts-Strogatz (WS) (veja seção 2.2). No modelo de ilhas, a quantidade de ilhas adotadas foram 2, 3 e 5 ilhas (ver seção 3.1.2), para cada experimento de inferência de GRNs foram geradas diferentes AGNs.

Para medir o grau de similaridade entre as redes inferidas pelos modelos e pelos AGNs, foi adotado o PPV (ver seção 3.2) e para obter o desempenho dos algoritmos foram coletados os tempos de execução de todos os experimentos, dado o hardware apresentado na seção 3.2.1. Para o cálculo do valor do PPV e do tempo foram executados 10 simulações.

O resultado do primeiro experimento executado para comparar as duas estratégias de busca: AG e AG com MI é apresentada pela Figura 8. Tal Figura apresenta os resultados da variação da quantidade de ilhas e a execução do AG, pela média do PPV de todos os experimentos, considerando as variações das topologias BA, ER e WS. É possível observar que utilizando o algoritmo com MI com 2 e 3 ilhas, a média melhores respostas comparado ao

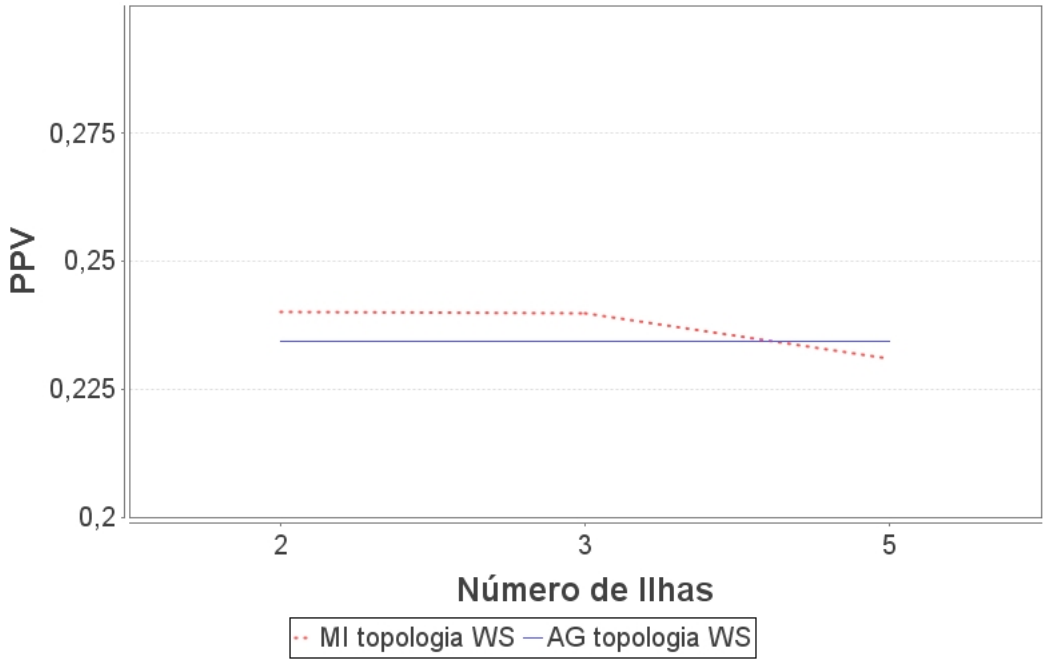

Figura 8: Medida de PPV obtida pela inferência de redes utilizando a estratégias de AG e AG com MI, aplicando 2,3 e 5 ilhas para o MI. Os valores da média do PPV obtido representa a execução de 10 experimentos.

AG. Entretanto, ao executar o MI com 5 ilhas, a médoa do PPV da rede inferida mostra menor valor.

As Figuras 9, 10 e 11 apresentam os respectivos resultados das médias do PPV das topologias BA, ER e WS, pelas variações das quantidades de ilhas. A topologia BA apresentou os melhores resultados utilizando as estratégias de AG e MI, quando comparado com as demais topologias. Entretanto, um comportamento decrescente de forma aproximadamente linear do PPV pode ser observado, quando são adicionados mais ilhas no MI. Na topologia ER o modelo de ilhas apresentou melhor PPV em comparação ao AG, podendo ser observado uma acentuada melhora do PVV no MI utilizando 3 ilhas. Os resultados da topologia WS tiveram os menores valores de PPV utilizando AG e MI, também é observado que o MI com 5 ilhas obteve uma ligeira melhoria do PPV.

De maneira geral, as variações dos resultados se mostraram diferentes para cada topologia abordada. Não foi possível identificar os elementos que causaram tais variações. Neste sentido, é necessário a execução de um número maior de experimentos, a fim de identificar as potenciais causas das variações que foram detectadas neste trabalho. Inclusive mais parâmetros podem ser considerados como a variação do número de ligações, tamanho do sinal e quantidade de genes.

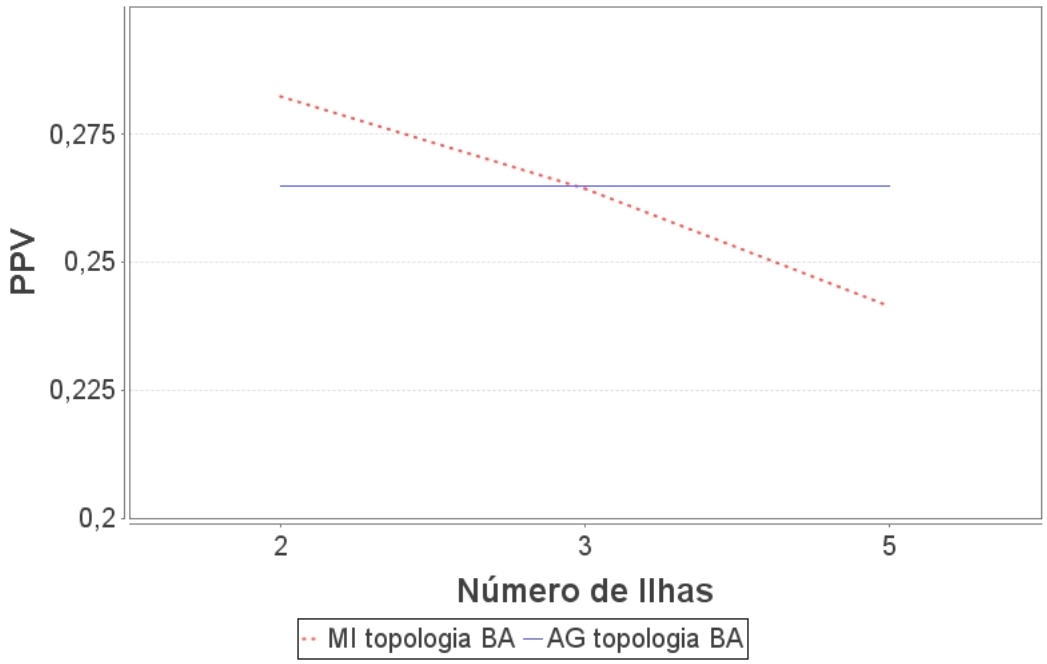

Figura 9: Média do PPV das redes inferidas pela topoloiga BA

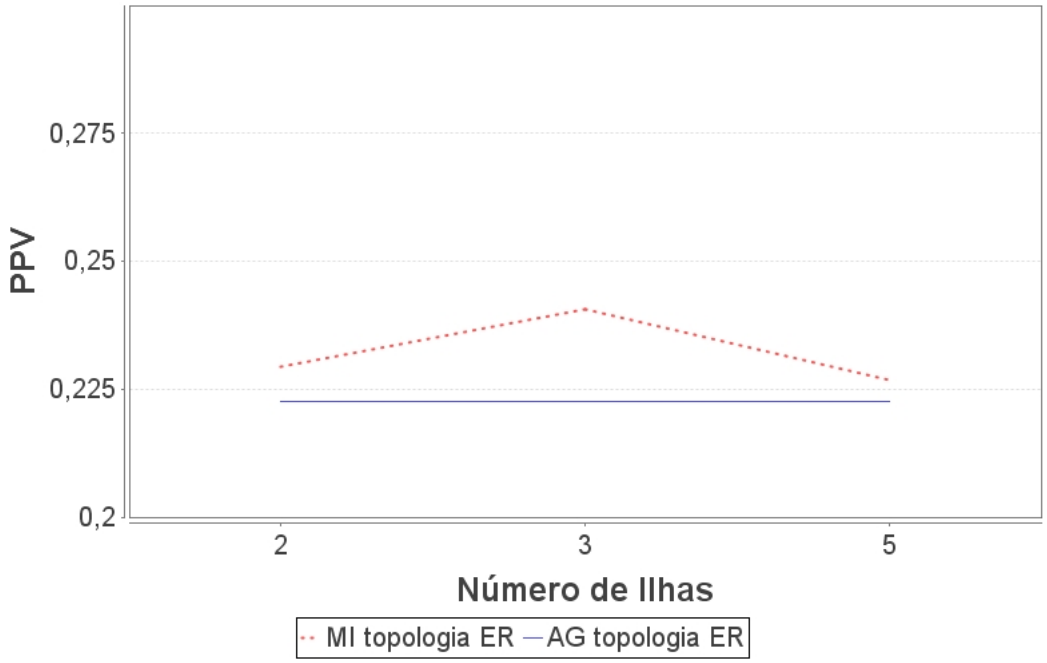

Figura 10: Média do PPV das redes inferidas pela topoloiga ER

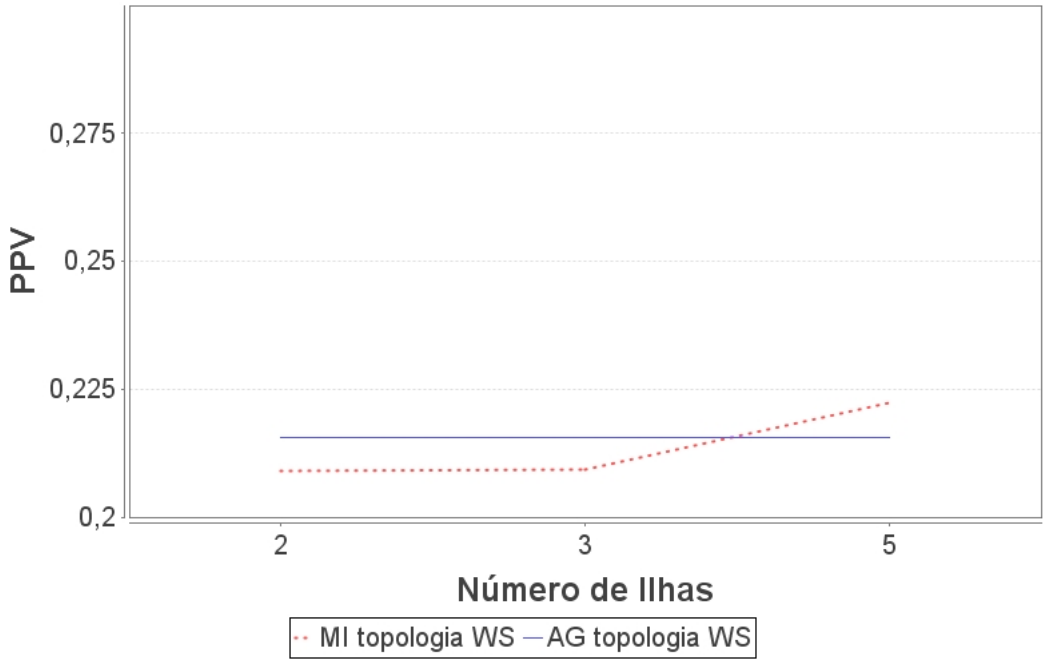

Figura 11: Média do PPV das redes inferidas pela topoloiga WS

No segundo experimento realizado, o objetivo foi avaliar o tempo computacional entre as estratégias de busca. Apresentando a mesma estratégia utilizada para o PPV, a Figura 12 apresenta o gráfico do tempo dado a variação das ilhas pelo AG com MI e o AG. É possível observar que todas as variações de ilhas (2, 3 e 5) do MI apresentaram tempo de execução superior ao AG. Este comportamento pode ser explicado pelo processo de sincronização do MI (ver seção 2.5.0.8). Ou seja, para a execução do operador de migração, todas as ilhas devem estar sincronizadas e após este processo os indivíduos serão migrados. Este processo gera um *overhead* em razão do tempo de espera da sincronização, o que é agravado pela quantidade de genes a serem inferidos, consequentemente, esta estratégia consome muito tempo, comprometendo o desempenho do algoritmo.

Nas Figuras 13, 14 e 15 são apresentados as médias dos tempos computacionais gerados pela inferência das topologias BA, WS e ER respectivamente. É possível observar o grau semelhança entre os tempos computacionais gerados nas topologia pelos dois métodos de busca (AG e MI), o que sugere que a variação de topologia não é tão relevante no desempenho computacional para a inferência das redes .

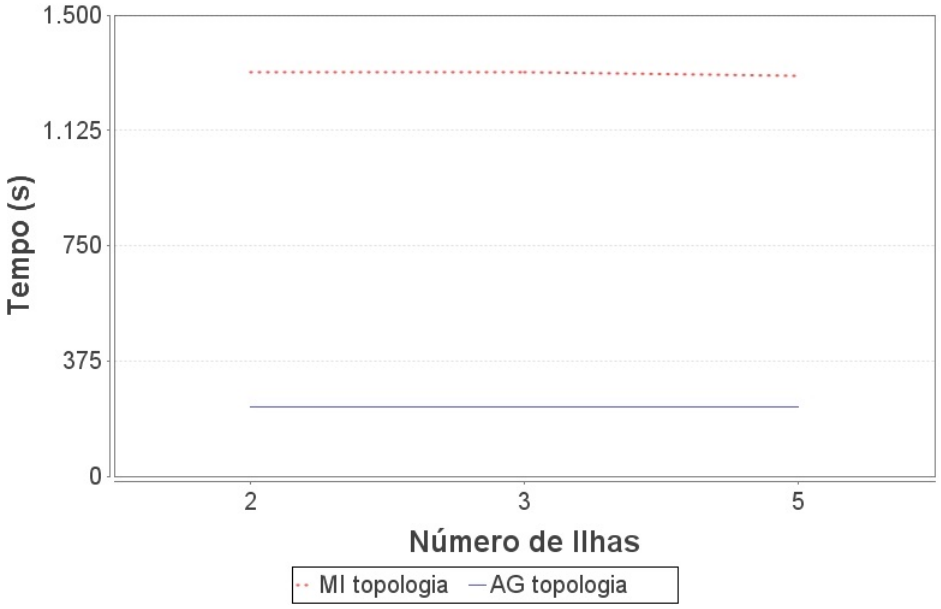

Figura 12: Média do tempo de todas as topologias dos algoritmo genético e do modelo de ilhas

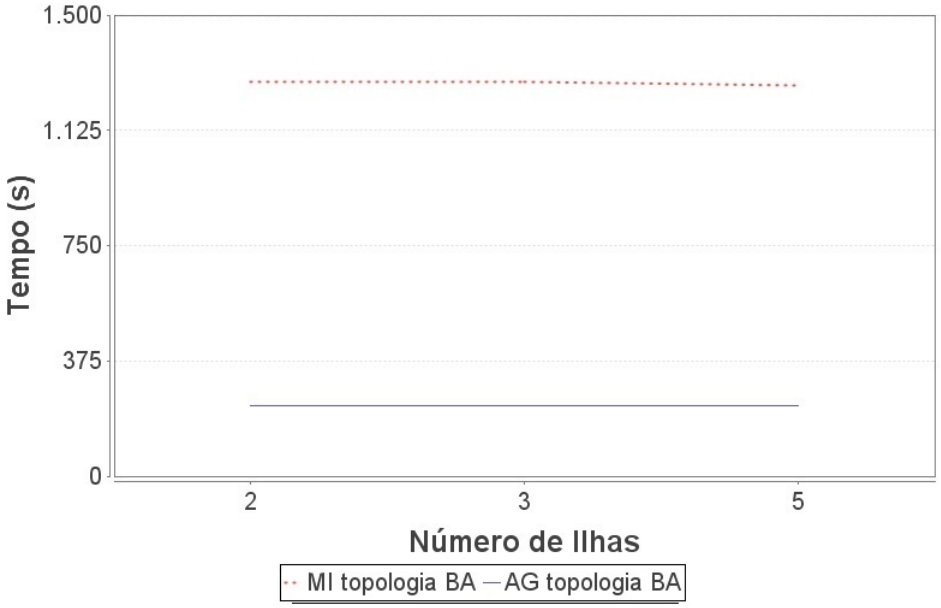

Figura 13: Média do tempo computacional gerado pela inferência da topologia BA pelos métodos de algoritmo genético e do modelo de ilhas

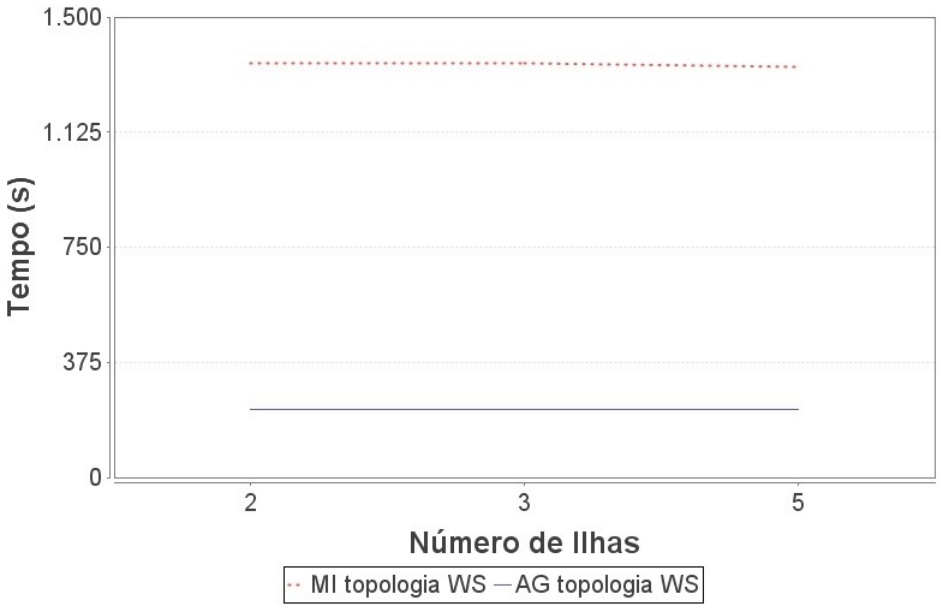

Figura 14: Média do tempo computacional gerado pela inferência da topologia WS pelos métodos de algoritmo genético e do modelo de ilhas

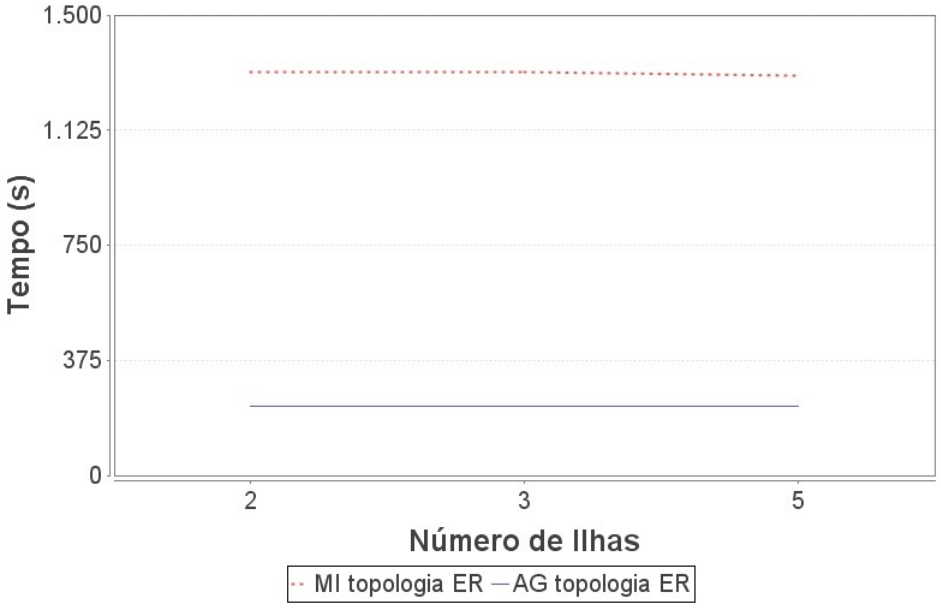

Figura 15: Média do tempo computacional gerado pela inferência da topologia ER pelos métodos de algoritmo genético e do modelo de ilhas

5 CONCLUSÃO

A partir de determinadas mudanças do ambiente, um organismo biológico pode responder de uma determinada forma, ajustando a expressão de seus genes. A conexão entre os genes formam uma grande rede complexa, na qual um gene pode regular muitos outros genes. Muitos estudos vêm sendo aplicados a tal problema, com o objetivo de compreender muitos problemas de áreas da biologia, física, psicologia, e pesquisas para desenvolvimento de remédios. Entretanto, ainda existe muito a ser descoberto sobre este processo biológico. Este trabalho aborda a inferência de GRNs a partir de um método de seleção de características, o qual é constituído por duas partes, uma função critério e um algoritmo de busca. A função critério abordada neste trabalho foi desenvolvida por (LOPES, 2011), no qual é baseada na Entropia de Condicional Média (LOPES et al., 2008). O algoritmo de busca abordado neste trabalho é baseado no princípio da teoria evolucionista de Charles Robert Darwin (DARWIN; BYNUM, 2009), denominado algoritmo genético, o qual é apresentado na seção 2.4 e um algoritmo paralelo denominado modelo de ilhas, baseado na Teoria do Equilíbrio Pontuado (COHOON et al., 1987) (ver seção 2.5). Os resultados mostraram que o MI apresentam melhores valores médios do PPV com 2 e 3 ilhas e menor valor médio de PPV para 5 ilhas. O AG apresentou um melhor desempenho computacional comparado ao MI, uma possível resposta é a sincronização necessária no MI para realizar o operador de migração, gerando um *overhead* prejudicial para seu desempenho. Não foi possível identificar os elementos que causaram as variações do resultado do PPV nas diferentes topologias adotadas neste trabalho.

5.1 TRABALHOS FUTUROS

Como um possível trabalho futuros novos experimentos contendo um número maior de simulações e também considerando mais variações de parâmetros como, por exemplo, quantidade de gene, variação da conexão média, quantidade de ilhas e tamanho do sinal podem ser desenvolvidos para uma análise mais aprofundada. Outros trabalhos que podem ser abordados são técnicas computacionais semelhantes ao AG como programação evolutiva e evolução diferencial para a inferência de GRNs. E como trabalhos futuros relacionado ao

desempenho computacional uma possível proposta é paralelizar o AG para inferência de redes gênicas em multithreads, GPU e desenvolver um MI assíncrono.

REFERÊNCIAS

ALBA, E. A Survey of Parallel Distributed Genetic Algorithms. **Complexity**, v. 4, p. 31—-52, 1999.

BäCK, T. **Evolutionary algorithms in theory and practice**. [S.l.]: T. Bäck, 1994.

BACK, T.; FOGEL, D. B.; MICHALEWICZ, Z. **Handbook of evolutionary computation**. [S.l.]: IOP Publishing Ltd., 1997. ISBN 0750303921.

BäCK, T.; SCHWEFEL, H.-P. An overview of evolutionary algorithms for parameter optimization. **Evolutionary computation**, MIT Press, v. 1, n. 1, p. 1–23, 1993.

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. ISSN 0036-8075.

BERNTSSON, J.; TANG, M. Dynamic optimization of migration topology in internet-based distributed genetic algorithms. In: **Proceedings of 2005 Genetic and Evolutionary Computation Conference**. [S.l.]: ACM, 2005. p. 1579–1580.

BETHKE, A. D. Comparison of genetic algorithms and gradient-based optimizers on parallel processors: Efficiency of use of processing capacity. 1976.

BISHOP, C. M. **Neural networks for pattern recognition**. [S.l.]: Oxford university press, 1995. ISBN 0198538642.

BOLTZMANN, L.; MCGUINNESS, B.; FOULKES, P. **Theoretical physics and philosophical problems: Selected writings**. [S.l.]: Reidel Publishing Company, 1974.

BRAGA, C. G. **O uso de Algoritmos Genéticos para aplicação de Otimização de Sistemas Mecânicos**. [S.l.]: Dissertação de mestrado, Universidade Federal de Uberlândia, 1998.

CAMPOS, T. E. de. **Técnicas de seleçao de caracterısticas com aplicaçoes em reconhecimento de faces**. [S.l.]: Master's thesis, Universidade de Sa Paulo, 2001.

CANTú-PAZ, E. A survey of parallel genetic algorithms. **Calculateurs paralleles, reseaux et systems repartis**, v. 10, n. 2, p. 141–171, 1998.

CHANDRA, N.; PADIADPU, J. Network approaches to drug discovery. **Expert opinion on drug discovery**, v. 8, n. 1, p. 7–20, jan. 2013. ISSN 1746-045X. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23140510>.

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

COHOON, J. P. et al. Punctuated equilibria: a parallel genetic algorithm. In: **Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application**. [S.l.]: L. Erlbaum Associates Inc., 1987. p. 148–154. ISBN 0805801588.

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. ISSN 0001-8732.

da Costa Filho, P. A.; POPPI, R. J. Algoritmo genético em química. SciELO Brasil, v. 22, p. 405, 1999.

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. ISBN 014040001X.

DAVIS, L. Handbook of genetic algorithms. Van Nostrand Reinhold, 1991.

De Jong, H. Modeling and simulation of genetic regulatory systems: a literature review. **Journal of computational biology**, Mary Ann Liebert, Inc., v. 9, n. 1, p. 67–103, 2002. ISSN 1066-5277.

De Jong, K. A. Analysis of the behavior of a class of genetic adaptive systems. 1975.

DEVIJVER, P. A.; KITTLER, J. **Pattern recognition: A statistical approach**. [S.l.]: Prentice/Hall International Englewood Cliffs, NJ, 1982. ISBN 0136542360.

D'HAESELEER, P.; LIANG, S.; SOMOGYI, R. Gene expression data analysis and modeling. In: **Pacific symposium on biocomputing**. [S.l.: s.n.], 1999. v. 99.

D'HAESELEER, P.; LIANG, S.; SOMOGYI, R. Genetic network inference: from co-expression clustering to reverse engineering. **Bioinformatics**, Oxford Univ Press, v. 16, n. 8, p. 707–726, 2000. ISSN 1367-4803.

DOUGHERTY, J.; TABUS, I.; ASTOLA, J. Inference of gene regulatory networks based on a universal minimum description length. **EURASIP Journal on Bioinformatics and Systems Biology**, Hindawi Publishing Corp., v. 2008, p. 5, 2008. ISSN 1687-4145.

DYER, D. W. **Watchmaker framework**. 2010. Disponível em: <Dyer, Daniel W>.

Eiben, A.E. and Smith, J. **Introduction to Evolutionary Computing**. Springer Berlin Heidelberg, 2003. 1–14 p. ISBN 978-3-642-07285-7. Disponível em: <http://dx.doi.org/10.1007/978-3-662-05094-1_1>.

ERDöS, P.; ALFRéD, R. On random graphs. **Publicationes Mathematicae Debrecen**, v. 6, p. 290–297, 1959.

FOGEL, D. **Artificial intelligence through simulated evolution**. [S.l.]: Wiley-IEEE Press, 2009. ISBN 0470544600.

GOLDBERG, D. E. Genetic algorithms in search, optimization, and machine learning. 1989. Disponível em: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0201157675>.

GOLDBERG, D. E.; DEB, K. A comparative analysis of selection schemes used in genetic algorithms. **Urbana**, v. 51, p. 61801–62996, 1991.

GUIMAR, F. G.; CAMPELO, F. **Topologias Dinâmicas para Modelo em Ilhas usando Evolução Diferencial**. Tese (Doutorado) — Universidade Federal de Minas Gerais, 2011.

HASHIMOTO, R. F. et al. Growing genetic regulatory networks from seed genes. **Bioinformatics**, Oxford Univ Press, v. 20, n. 8, p. 1241–1247, 2004. ISSN 1367-4803.

HOLLAND, C. Algoritmos Genéticos Adaptativos : Um estudo comparativo. 2000.

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. ISBN 0472084607.

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. ISSN 0162-8828.

JUNIOR, D. C. M. **Seleçao de caracterısticas e prediçao intrinsecamente multivariada em identificaçao de redes de regulaçao gênica**. [S.l.]: Ph. D. thesis, Universidade de Sao Paulo, Sao Paulo, SP, 2008.

KAUFFMAN, S. Gene regulation networks: A theory for their global structure and behaviors. **Current topics in developmental biology**, Elsevier, v. 6, p. 145–182, 1971. ISSN 0070-2153.

KAUFFMAN, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. **Journal of theoretical biology**, Elsevier, v. 22, n. 3, p. 437–467, 1969. ISSN 0022-5193.

KELEMEN, A.; ABRAHAM, A.; CHEN, Y. **Computational intelligence in bioinformatics**. [S.l.]: Springer, 2008. ISBN 3540768025.

LARDEUX, F.; GOëFFON, A. A dynamic island-based genetic algorithms framework. In: **Simulated Evolution and Learning**. [S.l.]: Springer, 2010. p. 156–165. ISBN 3642172970.

LI, F. et al. The yeast cell-cycle network is robustly designed. **Proceedings of the National Academy of Sciences of the United States of America**, National Acad Sciences, v. 101, n. 14, p. 4781–4786, 2004. ISSN 0027-8424.

LIANG, S.; FUHRMAN, S.; SOMOGYI, R. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In: **Pacific symposium on biocomputing, architectures.pdf:pdf**. [S.l.: s.n.], 1998. v. 3, n. 18-29, p. 2.

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

LOPES, F. M.; Cesar Jr, R. M.; 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. ISSN 1066-5277.

LOPES, F. M.; MARTINS, D. C.; CESAR, R. M. Feature selection environment for genomic applications. **BMC bioinformatics**, BioMed Central Ltd, v. 9, n. 1, p. 451, 2008. ISSN 1471-2105.

LOPES, H. S. Algoritmos genéticos em projetos de engenharia: aplicações e perspectivas futuras. **Anais do IV Simpósio Brasileiro de Automaçao Inteligente**, p. 64–74, 1999.

LOPES, R. A. et al. A Multi-Agent Approach to the Adaptation of Migration Topology in Island Model Evolutionary Algorithms. In: **Neural Networks (SBRN), 2012 Brazilian Symposium on**. [S.l.]: IEEE, 2012. p. 160–165. ISBN 1467326410.

LUCAS, D. C. Algoritmos Genéticos: uma Introduçao1. 2002.

MAJUMDAR, J.; BHUNIA, A. K. Elitist genetic algorithm for assignment problem with imprecise goal. v. 177, p. 684–692, 2007.

MARBACH, D. et al. Revealing strengths and weaknesses of methods for gene network inference. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 107, n. 14, p. 6286–6291, 2010. ISSN 0027-8424.

MEFFERT, K. et al. Jgap-java genetic algorithms and genetic programming package. **URL: http://jgap. sf. net**, 2011.

MICHALEWICZ, Z. **Genetic algorithms+ data structures= evolution programs**. [S.l.]: springer, 1996. ISBN 3540606769.

MILGRAM, S. The small world problem. **Psychology today**, New York, v. 2, n. 1, p. 60–67, 1967.

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.

PETTEY, C. B.; LEUZE, M. R.; GREFENSTETTE, J. J. A parallel genetic algorithm. In: **Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application**. [S.l.]: L. Erlbaum Associates Inc., 1987. p. 155–161. ISBN 0805801588.

RECHENBERG, I. **Evolutionsstrategien**. [S.l.]: Springer, 1978. ISBN 3540090509.

RUCIńSKI, M.; IZZO, D.; BISCANI, F. On the impact of the migration topology on the Island Model. **Parallel Computing**, Elsevier, v. 36, n. 10, p. 555–571, 2010. ISSN 0167-8191.

RUDOLPH, G. Convergence analysis of canonical genetic algorithms. **Neural Networks, IEEE Transactions on**, IEEE, v. 5, n. 1, p. 96–101, 1994. ISSN 1045-9227.

SANCHES, D. S. et al. Modeling Strategy by Adaptive Genetic Algorithm for Production Reactive Scheduling with Simultaneous Use of Machines and AGVs. In: **Computational Intelligence for Modelling Control & Automation, 2008 International Conference on**. [S.l.]: IEEE, 2008. p. 249–254. ISBN 0769535143.

SANCHEZ; THIEFFRY, D. A Logical Analysis of the< i> Drosophila</i> Gap-gene System. **Journal of theoretical Biology**, Elsevier, v. 211, n. 2, p. 115–141, 2001. ISSN 0022-5193.

SHANNON, C. E. A mathematical theory of communication. **ACM SIGMOBILE Mobile Computing and Communications Review**, ACM, v. 5, n. 1, p. 3–55, 2001. ISSN 1559-1662.

SRINIVAS, M.; PATNAIK, L. M. Adaptive probabilities of crossover and mutation in genetic algorithms. **Systems, Man and Cybernetics, IEEE Transactions on**, IEEE, v. 24, n. 4, p. 656–667, 1994. ISSN 0018-9472.

STOLOVITZKY, G.; MONROE, D. O. N.; CALIFANO, A. Dialogue on Reverse Engineering Assessment and Methods. **Annals of the New York Academy of Sciences**, Wiley Online Library, v. 1115, n. 1, p. 1–22, 2007. ISSN 1749-6632.

STUART, J. M. et al. A gene-coexpression network for global discovery of conserved genetic modules. **Science**, American Association for the Advancement of Science, v. 302, n. 5643, p. 249–255, 2003. ISSN 0036-8075.

STYCZYNSKI, M. P.; STEPHANOPOULOS, G. Overview of computational methods for the inference of gene regulatory networks. **Computers & Chemical Engineering**, Elsevier, v. 29, n. 3, p. 519–534, 2005. ISSN 0098-1354.

TANESE, R. Distributed genetic algorithms. In: **Proceedings of the third international conference on Genetic algorithms**. [S.l.]: Morgan Kaufmann Publishers Inc., 1989. p. 434–439.

TANG, J. et al. Study of migration topology in island model parallel hybrid-GA for large scale quadratic assignment problems. In: **Control, Automation, Robotics and Vision Conference, 2004. ICARCV 2004 8th**. [S.l.]: IEEE, 2004. v. 3, p. 2286–2291. ISBN 0780386531.

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. ISBN 3540584846.

TSALLIS, C. Nonadditive entropy: the concept and its use. **The European Physical Journal A**, Springer, v. 40, n. 3, p. 257–266, 2009. ISSN 1434-6001.

URSEM, R. K. Diversity-guided evolutionary algorithms. Springer, p. 462–471, 2002.

WALL, M. GAlib: A C++ Library of Genetic Algorithm Components. n. August, 1996.

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. ISSN 1471-0056.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. ISSN 0028-0836.

WEBB, A. R. **Statistical pattern recognition**. [S.l.]: Wiley. com, 2003. ISBN 0470854782.

WHITLEY, D. A genetic algorithm tutorial. **Statistics and computing**, Springer, v. 4, n. 2, p. 65–85, 1994. ISSN 0960-3174.

ZHOU, X. et al. A Bayesian connectivity-based approach to constructing probabilistic gene regulatory networks. **Bioinformatics**, Oxford Univ Press, v. 20, n. 17, p. 2918–2927, 2004. ISSN 1367-4803.