UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

MARIA FERNANDA BARBOSA WANDERLEY

Uma Abordagem Baseada em Computação Evolucionista para o Problema da Regulação Gênica

Profa. Ph.D Maria Luiza Machado Campos
Orientador

Prof. D.Sc. João Carlos Pereira da Silva
Co-orientador

Prof. D.Sc. Carlos Cristiano Hasenclever Borges
Co-orientador

Rio de Janeiro, Maio de 2009

Ficha Catalográfica

Barbosa Wanderley, Maria Fernanda

Uma Abordagem Baseada em Computação Evolucionista para o Problema da Regulação Gênica / Maria Fernanda Barbosa Wanderley. – Rio de Janeiro: UFRJ IM, 2009.

111 f.: il.

Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro. Programa de Pós-Graduação em Informática, Rio de Janeiro, BR–RJ, 2009.

Orientador: Maria Luiza Machado Campos; Co-orientador: João Carlos Pereira da Silva; Co-orientador: Carlos Cristiano Hasenclever Borges.

1. Algoritmos Genéticos. 2. Regulação Gênica. I. Machado Campos, Maria Luiza. II. Pereira da Silva, João Carlos. III. Hasenclever Borges, Carlos Cristiano. IV. Título.

Uma Abordagem Baseada em Computação Evolucionista para o Problema da Regulação Gênica

Maria Fernanda Barbosa Wanderley

Dissertação de Mestrado submetida ao Corpo Docente do Departamento de Ciência da Computação do Instituto de Matemática, e Núcleo de Computação Eletrônica da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do título de Mestre em Informática.

Aprovado por:

_____

Profa. Maria Luiza Machado Campos, Ph.D (Orientador)

_____

Prof. João Carlos Pereira da Silva, D.Sc. (Co-orientador)

_____

Prof. Carlos Cristiano Hasenclever Borges, D.Sc. (Co-orientador)

_____

Profa. Ana Tereza Ribeiro de Vasconcelos, D.Sc.

_____

Profa. Priscila Machado Vieira Lima, Ph.D.

_____

Prof. Adriano Joaquim de Oliveira Cruz, Ph.D.

_____

Prof. Josefino Cabral Melo Lima, Docteur

Rio de Janeiro, Maio de 2009

A todos que contribuíram direta e indiretamente nessa jornada.

AGRADECIMENTOS

Ao meu pai e minhas irmãs por toda paciência, compreensão e apoio nos momentos mais difíceis e madrugadas insones.

Aos meus orientadores, Professora Maria Luiza, Professor João Carlos, e Professor Carlos Cristiano, por me ajudarem com todas as dúvidas, pelas idéias para a resolução de algumas dificuldades que apareceram pelo caminho e pelas críticas construtivas sobre esse texto.

Ao meu noivo, Tiago Mota, por ter me acompanhado e incentivado nessa longa jornada, além de me ajudar a resolver alguns problemas com o LaTeX.

Aos amigos do Laboratório de Inteligência Computacional, por terem me acolhido no laboratório e pelos momentos de descontração. Em especial ao amigo Vinícius dos Santos pelas sugestões, conversas e apoio.

Às amigas Lígia Guedes, Nicole Scherer e Thaina Lima, por me ajudarem a compreender alguns conceitos biológicos.

À UFRJ e seus excelentes professores, pela minha formação acadêmica.

Ao CENPES e à CAPES pelo auxílio financeiro durante o período da minha pesquisa.

RESUMO

Expressão gênica é o processo de decodificar a informação contida em uma seqüência de DNA em uma proteína. Nesse processo, uma enzima chamada RNA-polimerase transcreve o DNA em RNA-mensageiro, que é então traduzido para uma proteína. Os fatores determinantes para decidir qual proteína pertence a qual célula e quanto dela vai ser produzido são a concentração de RNA-mensageiro e a freqüência com a qual esse RNA é traduzido. Operadores e reguladores, chamados de fatores de transcrição, controlam o processo de transcrição. Uma rede de regulação gênica consiste em determinar como e quais fatores de transcrição estão posicionados em uma dada seqüência de DNA. Nesse trabalho vamos analisar diferentes formas de codificar os fatores para serem usados com os algoritmos genéticos, algumas funções de avaliação, métodos de recombinação e taxas de mutação, além do impacto do uso de elitismo no operador de seleção, com o objetivo de identificar um conjunto de parâmetros pertinente para ser usado.

Palavras-chave: Algoritmos Genéticos, Regulação Gênica.

An Approach Based on Evolutionary Computing to the Genetic Regulation Problem

ABSTRACT

Gene expression is the process of decoding the information in a DNA sequence into a protein. In this process, an enzyme called RNA-polymerase transcribes DNA into messenger-RNA, which is translated into protein. The determinant factors to decide which protein belongs to each cell and how much of it will be produced are the concentration of mRNA and the frequency mRNA is translated. Operators and regulators, called transcription factors, control the transcription process. The gene regulation network consists in determining how and which transcription factors are positioned in some DNA sequence. In this work we analize different ways of encoding the factors to be used with genetic algorithms, some evaluation functions, methods of recombination and mutation rates, besides the impact of using elitism at the selection operator, trying to identify the pertinent set of parameters to be used.

Keywords: Genetic Algorithms, Genetic Regulation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AG	Algoritmos Genéticos
AIBSCS	*Automated Inference Based on Similarity to CONSENSUS sequences*
BPP	*Binding of Purified Proteins*
DNA	Ácido Desoxirribonucléico
FT	Fator de Transcrição
GEA	*Gene Expression Analysis*
HIBSCS	*Human Inference Based on Similarity to CONSENSUS sequences*
mRNA	Ácido Ribonucléico Mensageiro
RNA	Ácido Ribonucléico
rRNA	Ácido Ribonucléico Ribossomal
TRACTOR_DB	*Transcription Factors' Predicted Binding Sites in Prokariotic Genomes*
tRNA	Ácido Ribonucléico Transportador
UT	Unidade de Transcrição
UIG	Unidade de Informação Genética

SUMÁRIO

1 INTRODUÇÃO

Um dos principais conceitos de Biologia Molecular é o que diz que as ações e propriedades de cada célula são determinadas pelas proteínas que ela contém (ALBERTS et al., 1999). Expressão gênica é o processo de decodificar a informação contida em uma determinada seqüência de DNA em uma proteína. Nesse processo, uma enzima chamada RNA polimerase transcreve DNA em RNA mensageiro, que é posteriormente traduzido em proteína (LODISH et al., 2000). Os fatores determinantes para decidir que proteína pertence a cada célula e quanto dela será produzido são a concentração de RNA mensageiro, a freqüência com que ele é traduzido e quão estável é a proteína. É importante observar que a síntese de proteínas afeta diretamente a de DNA e RNA, uma vez que existem proteínas especiais que catalizam sua síntese.

Existem três tipos de genes no DNA (ALBERTS et al., 1999): os estruturais, que codificam proteínas, os operadores, que controlam os genes estruturais e os reguladores, que controlam os genes operadores. Os genes operadores e reguladores(que podem ser repressores ou ativadores), são chamados de fatores de transcrição devido ao seu papel no controle da transcrição do DNA em RNA. Os repressores controlam a transcrição codificando uma proteína que se conecta ao gene operador, prevenindo a RNA polimerase de se ligar à região promotora, que indica onde a transcrição deve

ser iniciada. Os ativadores trabalham de maneira semelhante, se ligando ao gene operador para estimular o processo de transcrição.

As redes de regulação gênica (LODISH et al., 2000) consistem em determinar como e quais fatores de transcrição estão posicionados numa seqüência de DNA. Esses fatores definem se uma proteína será ou não produzida, em qual quantidade e quando encerrar a produção. A configuração desses fatores é chamada unidade de informação genética (UIG). A determinação da localização desses sítios na seqüência de DNA é uma importante tarefa para a compreensão do processo de regulação gênica e é um problema ainda em aberto.

1.1 Objetivos

Neste trabalho, propomos o uso de algoritmos genéticos (AG) para determinar possíveis candidatos a unidade de informação genética em uma seqüência de DNA. Analisamos diferentes modos de codificar as UIG para serem usadas com algoritmos genéticos, algumas funções de avaliação, taxas de recombinação e mutação e o impacto do uso do elitismo no operador de seleção, tentando identificar um conjunto pertinente de parâmetros a serem usados nas predições. Como estudo de caso utilizamos promotores sigma 70 da *E.coli*, por ser um organismo bastante estudado e conhecido.

A idéia de fazer uso de algoritmos genéticos para resolução desse problema surgiu devido à capacidade de tal algoritmo realizar buscas por espaços de solução pouco conhecidos ou irregulares e, por isso, parecendo bastante adequado para o problema supracitado. Buscamos, nesse trabalho, criar uma codificação e medição adequadas para que um AG padrão seja eficiente.

1.2 Detalhamento

Temos como hipótese, que é possível gerar esses fatores de transcrição satisfatoriamente utilizando algoritmos genéticos e a informação gerada pelo programa PATSER. Para medir a qualidade dos resultados gerados, utilizaremos as informações contidas nos bancos de dados Tractor_DB (GONZÁLEZ et al., 2005) e RegulonDB (SALGADO et al., 2006).

A metodologia utilizada no presente trabalho divide-se, resumidamente nas seguintes etapas:

- Geração das pontuações para a seqüência a ser estudada utilizando o programa PATSER
- Uso de quatro estratégias de codificação dos indivíduos e função de avaliação do algoritmo genético para gerar candidatos a UIG
- Avaliação dos resultados obtidos

Na primeira etapa do trabalho, utilizamos uma seqüência de DNA (de 450 pares de base) da região regulatória do operon *fixA-fixB-fixC-fixX-yaaU* do genoma da *E. coli K12* e 35 matrizes de alinhamento, correspondentes aos fatores de transcrição a serem posicionados na seqüência. O programa PATSER (HERTZ; STORMO, 1999) gera pontuações que medem a similaridade entre a seqüência e cada uma das matrizes e, posteriormente, tais pontuações serão utilizadas para calcular a função de avaliação do algoritmo genético.

Na segunda etapa, quatro algoritmos genéticos distintos realizam uma busca no espaço de posições e fatores de transcrição. Cada uma das quatro buscas é realizada

utilizando seis conjuntos diferentes de parâmetros do algoritmo genético, sendo eles, o tamanho da população, o número de gerações, a probabilidade de recombinação e a probabilidade de mutação. Além disso, também foram feitos testes com e sem o operador de elitismo ativado.

Por fim, na fase de avaliação, os 480 arquivos gerados durante os testes foram analisados sob diversos aspectos, como o impacto do uso do operador de elitismo, a influência da probabilidade de recombinação e a qualidade dos resultados gerados com relação ao Tractor_DB (GONZÁLEZ et al., 2005) e RegulonDB (SALGADO et al., 2006).

1.3 Organização da Dissertação

O presente trabalho está organizado da seguinte forma: nos Capítulos 2 e 3 são apresentados os conceitos de biologia molecular e algoritmos genéticos, necessários para uma melhor compreensão do problema tratado e da solução proposta na dissertação. No Capítulo 4 mostramos os trabalhos relacionados utilizados como referência. Nos Capítulos 5 e 6 apresentamos a metodologia proposta e os resultados encontrados, analisados do ponto de vista dos algoritmos genéticos e em relação aos dados armazenados no TRACTOR_DB e REGULON_DB. Por fim, no Capítulo 7 são apresentadas as conclusões e os trabalhos futuros.

2 CONCEITOS DE BIOLOGIA MOLECULAR

2.1 Introdução

Nesse capítulo serão apresentados os conceitos básicos de biologia molecular (LODISH et al., 2000) para melhor compreensão desse trabalho. Além dos conceitos apresentados aqui, pode-se utilizar como fonte de consulta (SOARES, 1995), (SOARES, 1994) e (ALBERTS et al., 1999).

Na seção 2.2 os conceitos de DNA e RNA serão introduzidos. Na seção 2.3 será apresentada como ocorre a síntese de proteínas nas células e na seção 2.4, como é controlada a produção dessas proteínas. Por fim, nas seções 2.5 e 2.6 apresentamos o programa PATSER que gera pontuações que indicam qual a probabilidade de um dado fator de transcrição se ligar a determinadas posições da seqüência de DNA e fazemos as considerações finais do capítulo.

2.2 Conceitos Básicos

2.2.1 DNA

O **ácido desoxirribonucléico (DNA)** é a biblioteca celular que contém as informações necessárias para construção de todas as células e tecidos de um organismo (LODISH et al., 2000). Essas informações são arranjadas em unidades chamadas genes.

O DNA é um ácido nucléico, e como tal é formado por um grande número de unidades menores chamadas nucleotídeos. Os nucleotídeos do DNA são formados por um radical fosfato, uma pentose, nesse caso uma desoxirribose (figura 2.1) e uma base nitrogenada, que pode ser a adenina (A), timina (T), citosina (C) ou guanina (G). Cada molécula de DNA tem duas cadeias polinucleotídicas enroladas

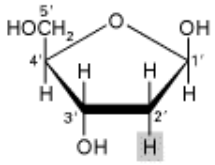


Figura 2.1: Desoxirribose (LODISH et al., 2000)

em trajetória helicoidal (figura 2.2).

2.2.2 RNA

O **ácido ribonucléico (RNA)** desempenha três diferentes papéis na síntese de proteínas. O **RNA mensageiro (mRNA)** carrega a informação copiada do DNA. A informação está codificada na forma de trincas de bases nitrogenadas, chamadas códons. Cada um desses códons especifica um aminoácido que formará a proteína.

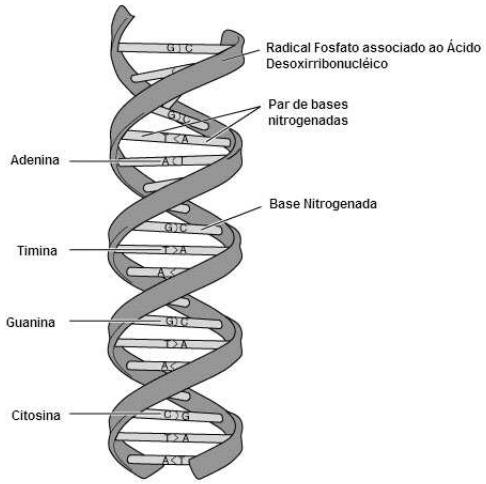


Figura 2.2: Segmento de uma molécula de DNA (BIOLOGY, 2007)

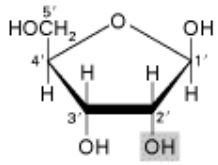


Figura 2.3: Ribose (LODISH et al., 2000)

O **RNA transportador (tRNA)** é responsável por transportar o aminoácido que irá formar a proteína. Por fim, o **RNA ribossomal (rRNA)** participa da formação dos ribossomos no citoplasma (LODISH et al., 2000).

O RNA, independentemente da função que virá a desempenhar, é formado por nucleotídeos, assim como o DNA. Os nucleotídeos do RNA são formados por um radical fosfato, uma pentose e uma base nitrogenada. Porém, no RNA, a pentose é uma ribose (figura 2.3) e as suas bases nitrogenadas podem ser a adenina (A), uracila (U), citosina (C) ou guanina (G).

Ao contrário do DNA, cada molécula de RNA tem apenas uma cadeia polinucleotídica (figura 2.4).

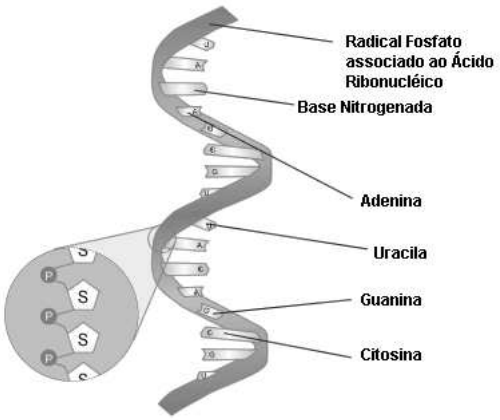


Figura 2.4: Segmento de uma molécula de RNA (MAKING THE MODERN WORLD, 2004)

2.3 Síntese de Proteínas

A síntese proteica tem início com a transcrição da informação contida no DNA para o mRNA e, a partir deste, acontece a tradução do mRNA para proteína.

2.3.1 A Transcrição do DNA

A transcrição do DNA é indireta, ou seja, os nucleotídeos do mRNA são complementares aos do DNA. A correspondência entre os nucleotídeos é realizada utilizando os padrões de combinação A - U (adenina corresponde à uracila) e C - G (a citosina corresponde à guanina).

O processo de transcrição inicia-se com a união da enzima RNA polimerase (RNAp) com o promotor. O promotor é uma região específica do DNA localizada na parte

anterior (*upstream*) ao ponto onde há o início da transcrição. Sua função é indicar a partir de qual nucleotídeo o DNA será transcrito. Além do promotor, outra região importante para a transcrição é a que contém o regulador, pois este determina quando e quantas vezes o gene deve ser transcrito (SOARES, 1994).

Após a união do promotor com a RNAp esta promove a separação de parte das hélices do DNA e o mRNA começa a se parear com o DNA seguindo os padrões citados anteriormente. Conforme a RNAp vai se movendo pela cadeia de DNA as regiões anteriormente separadas voltam a se juntar e as que estão a frente vão se separando. O movimento da RNAp continua até que esta encontre o sítio de terminação, se separe da fita de DNA e libere o mRNA transcrito (figura 2.5).

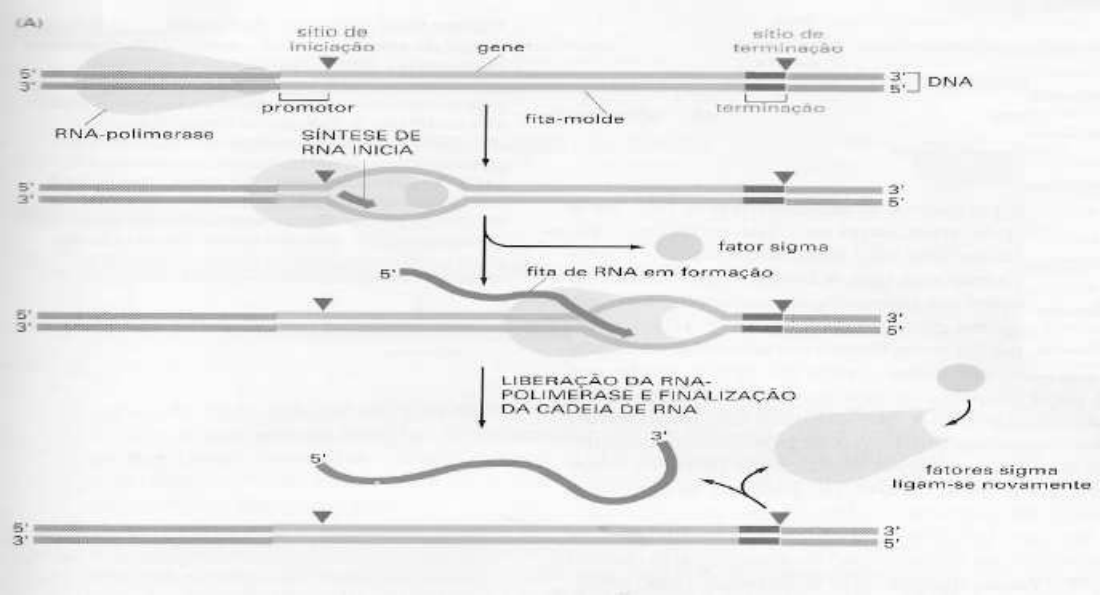


Figura 2.5: Processo de transcrição de uma molécula de DNA (ALBERTS et al., 1999)

2.3.2 A Tradução do mRNA em Proteína

O processo de tradução do mRNA em proteína pode ser dividido em três estágios: iniciação, alongamento e terminação. A figura 2.6 mostra todo o processo da tradução.

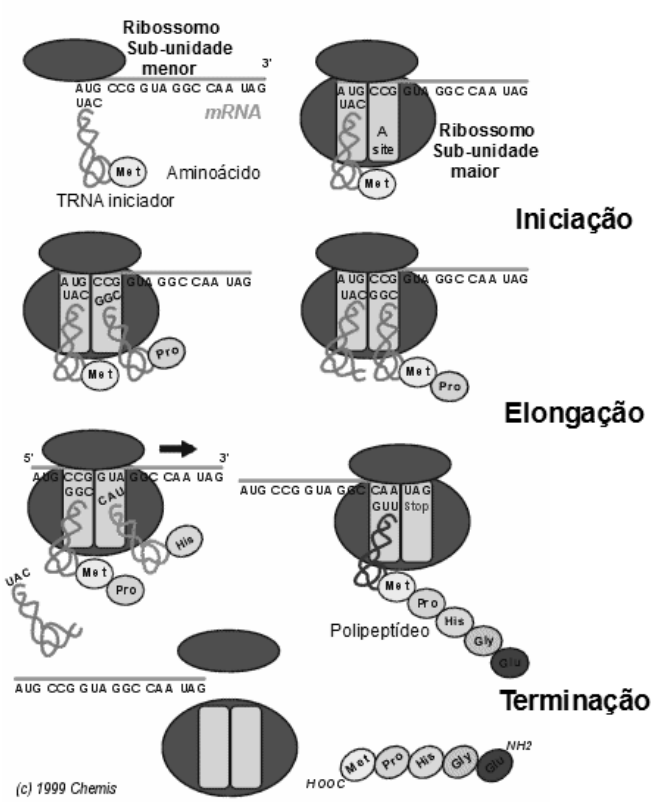


Figura 2.6: Processo de tradução do mRNA para proteína (COLLOMB, 2007)

2.3.2.1 Iniciação

Na fase de iniciação, primeiramente, a unidade ribossomal menor junta-se ao códon iniciador do mRNA (geralmente AUG) que indica em que posição se iniciará a tradução. Após essa união, o tRNA, que possui o anti-códon complementar ao

códon iniciador e carrega o aminoácido metionina (MET), liga-se ao mRNA.

2.3.2.2 Alongamento

A unidade ribossomal maior junta-se ao complexo formado na fase anterior, dando origem ao Complexo de Iniciação para dar continuidade à formação da proteína. Um novo tRNA aproxima-se do complexo e, se possuir o anti-códon correto, liga-se ao mRNA e ocorre a ligação entre o aminoácido que já estava presente e o novo. Com isso, o tRNA que carregava o primeiro aminoácido fica livre e abandona o complexo, dando espaço para um novo tRNA se ligar. Por fim, o complexo formado pelos ribossomos movimenta-se na cadeia de mRNA para ler o próximo códon. A tradução continua até que um dos códons de terminação (UAA, UAG, UGA) seja encontrado.

2.3.2.3 Terminação

Quando um dos códons de terminação é encontrado uma proteína chamada **fator de liberação** liga-se a esse códon ao invés de um tRNA. A ligação dessa proteína causa a liberação da proteína formada e a separação das subunidades do ribossomo, finalizando, assim, a tradução do mRNA.

2.4 Regulação Gênica

Um dos princípios básicos da biologia molecular diz que as ações e propriedades de cada célula são determinadas pelas proteínas que ela contém (LODISH et al., 2000). Com isso algumas perguntas surgem: o que determina o tipo e as quantidades de proteínas das células? O que permite que um organismo unicelular responda às modificações ambientais? Os fatores determinantes para responder essas perguntas

são a concentração do mRNA correspondente à proteína, a freqüência com que o mRNA é traduzido e a estabilidade da proteína. O principal mecanismo de regulação gênica é o que trata da concentração de mRNA, que é determinada por quais genes são transcritos e quanto desses genes são transcritos numa determinada célula.

O modelo de regulação gênica dos procariontes é o mais largamente conhecido e é apresentado a seguir, como um modelo definido por Jacob-Monod (JACOB; MONOD, 1961) para explicar a utilização da lactose pela bactéria *E. Coli*.

2.4.1 O Modelo de Jacob-Monod

Nas bactérias, o controle genético serve principalmente para permitir à célula que se ajuste às mudanças no seu ambiente nutricional, de modo que seu crescimento e divisão possam ser otimizados.

A bactéria *Escherichia coli* produz a enzima β-galactosidase para digerir a lactose. Quando não há presença de lactose o gene que codifica a enzima para a digestão da mesma precisa ser desligado. Isso garante que não haverá desperdício de energia na produção de uma enzima que não é necessária.

Existem três categorias de genes no DNA, os estruturais (que codificam uma proteína), os operadores (que controlam os genes estruturais) e os reguladores (que controlam os genes operadores). Os genes reguladores (repressores e ativadores) e os genes operadores são chamados de fatores de transcrição, por realizarem um controle sobre a transcrição do DNA em RNA. Os conjuntos de um ou mais genes que são co-transcritos são chamados unidades de transcrição (UT). Por exemplo, na figura 2.7 os genes *lac Z*, *lac Y* e *lac A* compõem uma UT. Além das UT também estão presentes no DNA os operons, que contém o gene promotor, o gene operador e um ou mais genes estruturais que são transcritos em uma molécula de mRNA(que

codifica para uma ou mais proteínas). O operon também pode conter um ou mais genes reguladores mas isso não ocorre necessariamente.

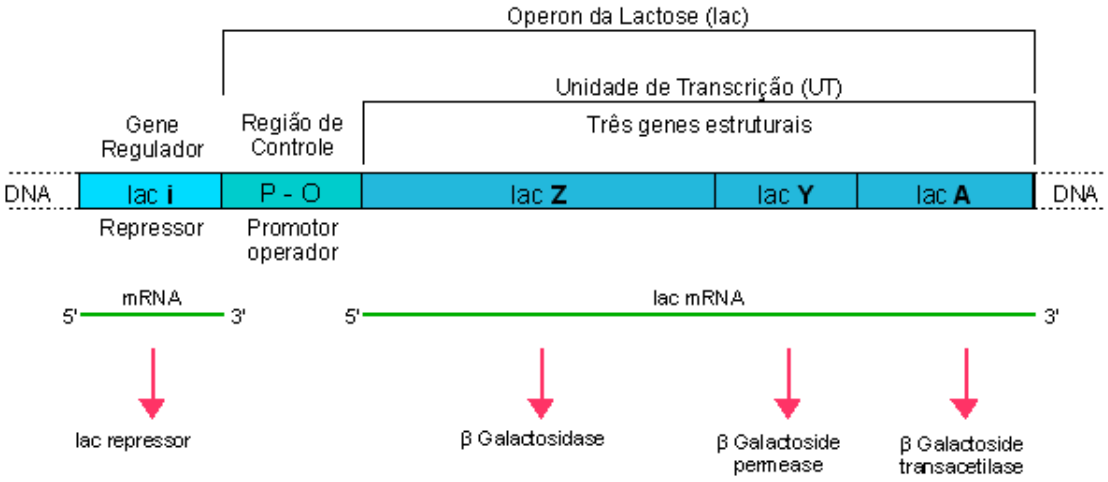


Figura 2.7: Primeiro estágio do modelo de Jacob-Monod (PASSARGE, 1995)

O gene regulador *lac i* codifica uma proteína chamada *lac* repressor, que se liga ao gene operador (seqüência de poucos nucleotídeos contida na região promotora). Essa proteína impede que a RNA polimerase se ligue à região promotora e, por conseqüência, impede a transcrição dos genes estruturais que codificam as enzimas β-galactosidase, β-galactoside permease e β-galactoside transacetilase. (figura 2.8).

Quando moléculas de lactose são inseridas no meio, estas se ligam ao repressor, liberando, assim, o gene operador e a região promotora. Quando isso ocorre a RNA polimerase pode se ligar ao sítio promotor e os genes estruturais podem ser codificados (figura 2.9). A lactose é chamada de indutor, pois induz a síntese da enzima β-galactosidase.

Quando a β-galactosidase passa a ser sintetizada ela inicia a digestão das moléculas

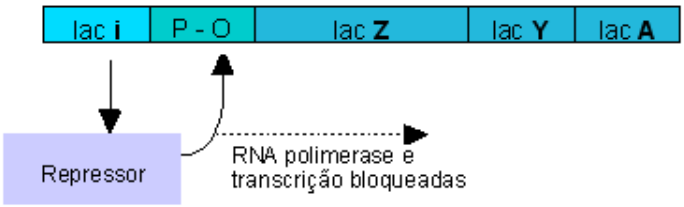


Figura 2.8: Segundo estágio do modelo de Jacob-Monod (PASSARGE, 1995): O gene é inativado pelo repressor

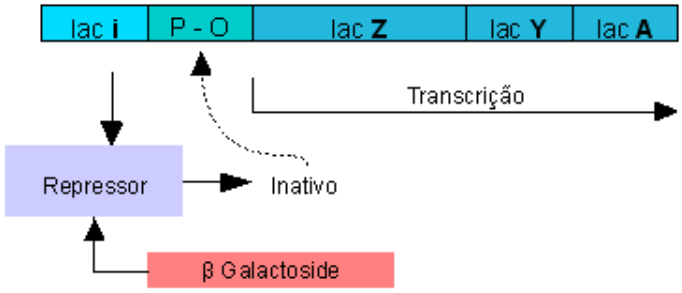


Figura 2.9: Terceiro estágio do modelo de Jacob-Monod (PASSARGE, 1995): Inativação do repressor devido a ligação da moléculo da lactose (β-galactoside)

de lactose presentes no meio, até que não reste nenhuma. Com a ausência de lactose o repressor volta a se ligar no sítio promotor e a produção da enzima é interrompida.

Em oposição às proteínas repressoras existem as proteínas ativadoras. Também codificadas a partir de um gene regulador, elas agem ligando os genes estruturais. Nesse caso, as proteínas ativadoras ligam-se à região promotora para estimular a iniciação da transcrição.

2.5 O Programa PATSER

Em nosso estudo utilizamos uma seqüência de DNA (de 450 pares de base) da região regulatória do operon *fixA-fixB-fixC-fixX-yaaU* do genoma da *E. coli K12*. Temos como objetivo determinar quais e em que posições os fatores de transcrição aparecem nesta seqüência. Em geral, a posição dos fatores de transcrição é dada com relação ao ponto do início da transcrição. Para o nosso caso de estudo a posição absoluta 358 da seqüência corresponde à posição +1, que marca o início do sítio de transcrição (figura 2.10).

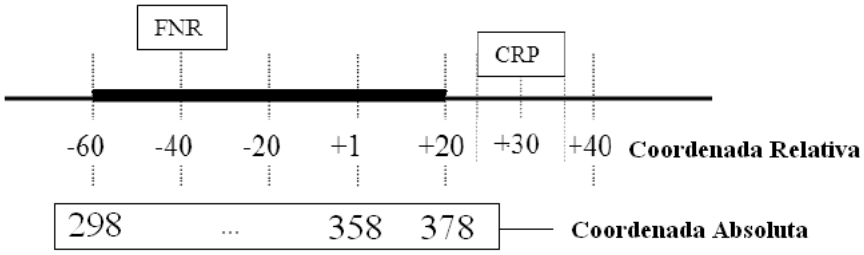


Figura 2.10: Representação da região regulatória utilizada no trabalho. A posição relativa +1, que marca o início da transcrição, corresponde a posição absoluta 358.

A lista dos 35 fatores de transcrição utilizados no presente trabalho é apresentada na tabela 2.1. O tamanho do espaço de busca das UIG candidatas, considerando-se a seqüência de 450 pares de base e os 35 fatores de transcrição utilizados, é da ordem

de 6.8×10^{16}.

Tabela 2.1: Fatores de transcrição utilizados

Fator de Transcrição	Tamanho da Matriz de alinhamento	Fator de Transcrição	Tamanho da Matriz de alinhamento
AraC	26	Lrp	19
ArcA	17	MalT	16
ArgR	25	MarA	23
CaiF	29	MelR	18
CpxR	24	MetJ	18
CRP	30	ModE	24
CysB	22	NarL	16
CytR	22	NtrC	29
DnaA	20	OmpR	16
FadR	29	OxyR	17
FIS	23	PhoB	25
FNR	16	PurR	16
FruR	16	Rob	23
Fur	25	SoxS	19
GadX	16	TorR	18
GlpR	18	TyrR	18
IHF	25	XylR	29
LexA	16		

Para gerar as pontuações a serem utilizadas na função de avaliação do algoritmo genético utilizamos o PATSER (HERTZ; STORMO, 1999), programa criado por Gerald Z. Hertz que gera pontuações para subseqüências de tamanho L de uma dada seqüência em relação a uma determinada matriz de alinhamento de um fator de transcrição. Essa pontuação indica o quão relacionadas são a subseqüência e a matriz de alinhamento, indicando a probabilidade de um certo fator de transcrição se ligar a determinadas posições da seqüência em estudo.

Os elementos de uma matriz de alinhamento $A_{i,j}$ são obtidos contando-se o número de vezes que um nucleotídeo i aparece na posição j das seqüências de alinhamento.

Na figura 2.11, para a posição $j = 6$, o nucleotídeo A aparece duas vezes, o C e o T aparecem uma vez e o G nenhuma vez. Por isso, temos $A_{A,6} = 2$ e $A_{C,6} = A_{T,6} = 1$ e $A_{G,6} = 0$.

Matriz de Alinhamento

	A	A	G	G	T	A
	A	T	G	G	C	C
	A	T	T	A	G	A
	A	C	G	G	G	T

A	4	1	0	1	0	2
C	0	1	0	0	1	1
G	0	0	3	3	2	0
T	0	2	1	0	1	1

Figura 2.11: Matriz de Alinhamento (HERTZ; STORMO, 1999)

Cada elemento da matriz de pesos é obtido utilizando o elemento correspondente na matriz de alinhamento do seguinte modo: primeiro, adiciona-se a probabilidade *a priori* do nucleotídeo ao número de vezes que ele ocorre no elemento da matriz de alinhamento. Em seguida, divide-se o resultado pelo número total de seqüências utilizadas na construção da matriz de alinhamento mais um. A freqüência resultante é então normalizada pela probabilidade *a priori* do nucleotídeo em questão. O resultado final é convertido em um elemento da matriz de pesos calculando-se seu logaritmo neperiano. Formalmente, o peso $W_{i,j}$ correspondente ao nucleotídeo i na posição j é definido por:

$$W_{i,j} = \ln \frac{(n_{i,j} + p_i)\,/(N+1)}{p_i} \tag{2.1}$$

onde $n_{i,j}$ representa o número de vezes que o nucleotídeo$i$ é observado na posição j do alinhamento, p_i representa a probabilidade *a priori* do nucleotídeo i e N representa o número total de seqüências usadas no alinhamento.

Na figura 2.12 é mostrada a matriz de pesos resultante da matriz de alinhamento da figura 2.11. O elemento $W_{C,1}$, por exemplo, é obtido utilizando-se $n_{C,1} = 0$, $p_i = 0.25$ (a probabilidade *a priori* é a mesma para todos os nucleotídeos) e $N = 4$. Com isso, temos $W_{C,1} = -1.6$.

Matriz de Pesos

A	1.2	0	-1.6	0	-1.6	0.59
C	-1.6	0	-1.6	-1.6	0	0
G	-1.6	-1.6	0.96	0.96	0.59	-1.6
T	-1.6	0.59	0	-1.6	0	0

Figura 2.12: Matriz de pesos obtida a partir matriz de alinhamento da figura 2.11.

O PATSER percorre a região de regulação com a matriz de peso de tamanho$m$ correspondente a um determinado fator de transcrição (no presente trabalho o tamanho das matrizes de peso pode ser visto na tabela 2.1) e calcula os $450 - m + 1$ alinhamentos correspondentes a avançar a matriz uma base por vez. A cada um desses alinhamentos é atribuída uma pontuação, que mede quão similar é a sub-seqüência com a matriz. A pontuação S_i para a sub-seqüência que começa na posição i é calculada por:

$$S_i = \sum_{j=i}^{i+m} W_{X(i+j-1),j} \tag{2.2}$$

onde $X(i + j - 1)$ é o nucleotídeo da posição $i + j - 1$ da sub-seqüência.

Nas figuras 2.13 e 2.14 é mostrado o cálculo da pontuação para as posições 1 e 2 do alinhamento.

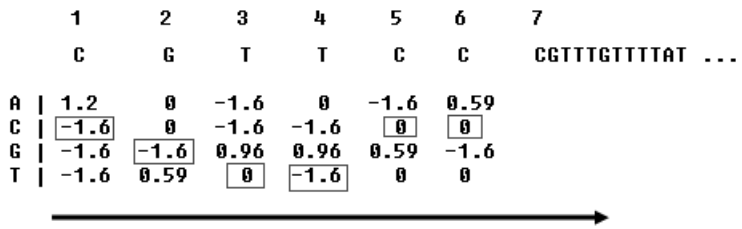


Figura 2.13: Seqüência sendo percorrida pela matriz de pesos. Para a posição 1 do alinhamento, a pontuação S_1 calculada é o somatório dos valores destacados, ou seja, $S_1 = -1.6 + -1.6 + 0 + -1.6 + 0 + 0 = -4.8$.

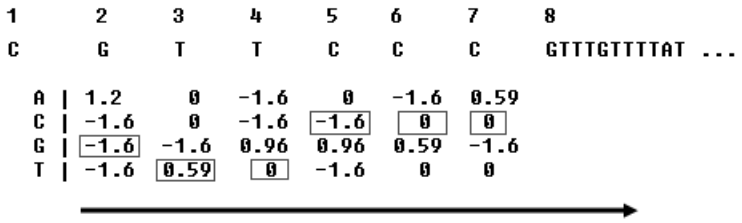


Figura 2.14: Seqüência sendo percorrida pela matriz de pesos. Para a posição 2 do alinhamento, a pontuação S_2 calculada é o somatório dos valores destacados, ou seja, $S_1 = -1.6 + 0.59 + 0 + -1.6 + 0 + 0 = -2.61$.

2.6 Considerações

Nesse capítulo foram apresentados os fundamentos da biologia molecular necessários para a compreensão desse trabalho. Inicialmente apresentamos os conceitos de DNA e RNA, bem como o processo de síntese de proteínas (seções 2.2 e 2.3), explicando o processo de transcrição do DNA e tradução do mRNA.

Em seguida (seção 2.4) mostramos o processo de regulação gênica, que controla a produção de proteínas e o modelo de Jacob-Monod, utilizado na dissertação para

ilustrar o funcionamento do mecanismo regulatório e os fatores envolvidos no processo. Foram também apresentados os conceitos de operon e unidades de transcrição, que serão utilizados no capítulo 4 na explicação dos métodos utilizados pelos bancos TRACTOR_DB (GONZÁLEZ et al., 2005) e RegulonDB (GAMA-CASTRO et al., 2008).

Por fim, na seção 2.5 apresentamos as matrizes de alinhamento dos fatores de transcrição a serem posicionados na região regulatória do operon *fixA-fixB-fixC-fixX-yaaU* da *E.coli K12*. Tais matrizes de alinhamento são utilizadas pelo programa PATSER para calcular as pontuações para subseqüências da região regulatória, que posteriormente serão usadas na função de avaliação do algoritmo genético (capítulo 3).

3 ALGORITMOS GENÉTICOS

Algoritmos genéticos são uma classe de algoritmos baseados na teoria de evolução de Darwin. A idéia básica é encontrar a melhor solução para um dado problema através de um processo evolutivo que seleciona a mais adaptada dentre as soluções (GOLDBERG, 1989).

Informalmente, o algoritmo genético básico pode ser descrito pelos seguintes passos:

1. Inicializar população;
2. Avaliar os indivíduos;
3. Gerar nova população aplicando as operações de seleção, recombinação e mutação.

No algoritmo, inicialmente, um conjunto de possíveis soluções para o problema é gerado aleatoriamente. Cada solução gerada é chamada de indivíduo ou cromossomo e o conjunto de indivíduos é chamado de população (MUNAKATA, 1998). Após a geração da população inicial uma série de iterações (gerações) ocorrem até que a condição de parada seja alcançada. Nessas gerações os seguintes passos ocorrem:

primeiro os melhores indíviduos são selecionados de acordo com uma função de avaliação. Após a seleção, segundo uma certa probabilidade, os indivíduos escolhidos sofrem recombinação e, em seguida, uma mutação. Ao fim da iteração uma nova população está formada e o algoritmo segue até que um critério de parada seja satisfeito (MICHALEWICZ, 1996).

A utilização de um AG depende da definição dos valores de diversos parâmetros. A probabilidade de recombinação, a probabilidade de mutação, o tamanho da população e o uso, ou não, de elitismo (ver 3.3.4) adaptam o algoritmo genético básico para cada diferente tipo de problema que se deseje resolver (DE JONG; SPEARS, 1990). Mais adiante todos os parâmetros serão apresentados e analisados.

Outro ponto muito importante do algoritmo diz respeito à codificação dos indivíduos da população de acordo com o tipo de problema que se deseja resolver. Uma codificação que não seja muito apropriada pode dificultar a convergência do algoritmo ou até mesmo gerar inconsistências durante a execução do mesmo. A codificação mais utilizada é a binária devido à sua simplicidade.

Outras referências sobre algoritmos genéticos podem ser encontradas em (GOLDBERG, 1989), (MICHALEWICZ, 1996), (WHITLEY, 1993) e (RUSSEL; NORVIG, 2004).

3.1 Algoritmo Genético Básico

Algoritmo 3.1 Algoritmo Genético Básico

Início: Gerar randomicamente uma população (P_0) com n indivíduos;
$i = 0$;
enquanto $i \leq numeroDeGeracoes$ **faça**
  avaliar a função de avaliação $f(x)$ dos n indivíduos da população;
  enquanto P_i não estiver completa **faça**
    selecionar dois indivíduos de acordo com seu $f(x)$;
    fazer a recombinação dos indivíduos (pais) de acordo com p_c;
    fazer a mutação em cada uma das posições dos filhos de acordo com p_m;
    colocar os filhos na nova população;
  fim de enquanto
  $i = i + 1$;
fim de enquanto

O algoritmo 3.1 utiliza os três operadores básicos que serão introduzidos nas seções a seguir. As entradas para o algoritmo são o número de indivíduos da população (n), a probabilidade de recombinação (p_c), a probabilidade de mutação (p_m) e uma condição de parada. Geralmente a condição usada é o número de gerações, ou seja, o número de vezes que o laço principal foi executado.

3.2 Operadores

3.2.1 Seleção

O operador seleção de um algoritmo genético atua na população de um modo muito semelhante à atuação da seleção natural em uma população de uma dada espécie, isto é, os indivíduos mais fortes, mais adaptados sobrevivem e geram descendentes. Seguindo esta idéia, os indivíduos da população inicial são selecionados probabilisticamente de acordo com a sua **função de avaliação**, cujo valor é dado de acordo com o tipo de problema a ser resolvido.

Existem várias formas de selecionar os indivíduos mais adaptados porém a mais

utilizada é a **seleção por roleta**. Nesse método de seleção uma roleta imaginária é criada e os indivíduos são dispostos nela de acordo com a sua função de avaliação. Para calcular a probabilidade de um dado indivíduo ser sorteado devemos, primeiro, calcular o valor total da função de avaliação da população dada por

$$F = \sum_{i=1}^{n} f_i \tag{3.1}$$

onde n é o número de indivíduos da população e f_i é o valor da função de avaliação de cada indivíduo. Tendo calculado o valor total podemos calcular a probabilidade de cada indivíduo ser selecionado. Essa probabilidade é dada por

$$p_i = \frac{f_i}{F} \tag{3.2}$$

onde f_i tem o mesmo significado que em (3.1). Um indivíduo é selecionado com uma dada probabilidade se o número gerado g for tal que

$$\sum_{j=1}^{i-1} p_j \leq g < \sum_{j=1}^{i} p_j \tag{3.3}$$

onde $g \in [0, 1[$. Sorteando os indivíduos dessa maneira, os que possuem uma maior probabilidade são selecionados mais vezes deixando, assim, mais representantes na próxima geração. Vale ressaltar que esse método de seleção só é válido se a função de avaliação gerar apenas valores positivos.

Exemplo 3.1: Supondo uma população de quatro indivíduos representados por números binários de três digitos. A função de avaliação usada será o número decimal codificado pelo indivíduo.

Sorteamos aleatoriamente quatro números, 0.483, 0.091, 0.652 e 0.725. Para o número 0.483, o indivíduo 3 é selecionado pois

$$\sum_{j=1}^{2} p_j = 0.428 \leq 0.483 < \sum_{j=1}^{3} p_j = 0.571$$

Tabela 3.1: Exemplo de Seleção por Roleta

i	Indivíduo i	f_i	p_i
1	101	5	$5/14 = 0.357$
2	001	1	$1/14 = 0.071$
3	010	2	$2/14 = 0.143$
4	110	6	$6/14 = 0.429$
	Total	14	1.000

Com o número 0.091, o indivíduo 1 é selecionado pois $0 < 0.091 \leq 0.357$, com o número 0.652 o indivíduo 4 é selecionado pois $0.571 < 0.652 \leq 1$ e com o número 0.725 o indivíduo 4 é novamente selecionado. Podemos ver com o exemplo que o melhor indivíduo (o número 4) foi selecionado mais vezes devido a sua alta probabilidade.

•

3.2.2 Recombinação

O operador de recombinação possibilita a troca de informação entre dois indivíduos a fim de executar uma busca em regiões codificadas em cada um deles separadamente.

O funcionamento básico da recombinação é o seguinte: randomicamente selecionam-se dois indivíduos do conjunto de cromossomos gerados através da seleção e um ponto de recombinação é sorteado. Esse ponto deve estar dentro dos limites do tamanho dos cromossomos. Por exemplo, dados cromossomos de comprimento L e um ponto de recombinação k, então k deve estar no intervalo [1, L-1]. Após a escolha do ponto ocorre a permutação dos segmentos cromossômicos.

Esse operador está relacionado com uma probabilidade, ou seja, nem todos os pares selecionados farão a permuta de segmentos. A influência do valor dessa probabilidade será discutida na sub seção 3.3.1.

Exemplo 3.2: Suponha que os indivíduos 0001100100110110 e 1101111000011110 tenham sido selecionados para recombinação. Como os cromossomos têm comprimento L = 16, temos que sortear um número no intervalo [1, 15] para ser o ponto de recombinação. Se o ponto selecionado for o 5, então teremos:

Tabela 3.2: Recombinação

Cromossomo 1	**00011**|**00100110110**
Cromossomo 2	11011|11000011110
Filho 1	**00011**|11000011110
Filho 2	11011|**00100110110**

•

3.2.3 Mutação

O próximo passo após a recombinação é a mutação. A mutação, como definida na Biologia, é uma modificação no material genético e é responsável por várias modificações que ocorrem em diversas espécies.

A atuação desse operador num indivíduo de uma população em um algoritmo genético consiste na troca de, por exemplo, um bit 1 por um bit 0 em determinado locus gênico. Essa modificação pode ser benéfica e ajudar a população a escapar de um mínimo local para o qual poderia estar convergindo ou então pode gerar um indivíduo que tenha um valor baixo na função de avaliação e seja eliminado numa próxima geração.

A ocorrência da mutação em um indivíduo também está relacionada com uma certa probabilidade cuja influência será discutida na sub seção 3.3.2

Exemplo 3.3: Suponha que os indivíduos 0001100100110110 e 1101111000011110

tenham sido selecionados para sofrerem uma mutação, que o valor da função de avaliação seja dado pelo valor decimal codificado por cada indivíduo e que desejamos maximizar essa função.

Se para o primeiro indivíduo a mutação for no primeiro bit mais à esquerda ele passará a ser **1**001100100110110 e o valor de sua função de avaliação será 39222 e antes era 6454. Se para o segundo indivíduo a mutação for no décimo segundo bit da esquerda para direita ele passará a ser 11011110000**0**1110 e o valor de sua função de avaliação será de 56846 e antes era 56862.

Como podemos ver pelo exemplo, a mutação pode melhorar ou piorar o valor da função de avaliação de um indivíduo e, conseqüentemente, o total da população. •

3.3 Parâmetros do Algoritmo

Os parâmetros do algoritmo genético são variáveis que podem ser alteradas para moldar o algoritmo a um dado problema. São eles: a probabilidade de recombinação, a probabilidade de mutação, o tamanho da população e o elitismo. Os valores sugeridos a seguir são utilizados com mais freqüência, servindo como uma referência para os testes iniciais do algoritmo. Como o valor desses parâmetros depende muito do tipo de problema a ser resolvido é necessário haver uma fase inicial de testes onde os parâmetros são ajustados empiricamente.

3.3.1 Probabilidade de Recombinação

A probabilidade de recombinação define com que freqüência o operador de recombinação será aplicado. Normalmente a probabilidade utilizada é $p_c \geq 0.6$ (DE JONG; SPEARS, 1990). Se a probabilidade de recombinação for $p_c = 1$ todos os pares de

cromossomos serão recombinados. Se a probabilidade for zero nenhum dos pares de cromossomos sofrerá recombinação. O objetivo do crossover é obter novos indivíduos que consigam uma avaliação melhor devido à combinação das características de seus pais e, por isso, a $p_c = 0$ é prejudicial ao desempenho do AG. Porém, utilizar $p_c = 1$ pode causar a perda de boas qualidades dos indivíduos da população, não sendo recomendado seu uso também.

3.3.2 Probabilidade de Mutação

A probabilidade de mutação define com que freqüência o operador de mutação será aplicado. Geralmente o valor inicial utilizado é $p_m = 0.001$. Se a probabilidade de mutação for $p_m = 0$, então nenhum bit de nenhum indivíduo será afetado e se for igual à um todos os bits de todos os indivíduos serão trocados. Percebe-se, que quando $p_m = 1$, a busca realizada pelo algoritmo torna-se randômica, uma vez que nenhuma característica dos melhores indivíduos é mantida.

3.3.3 Tamanho da População

O tamanho da população está relacionado ao espaço em que será realizada a busca. Se tivermos poucos indivíduos na população então o espaço coberto pela busca é menor e pode-se levar mais tempo para achar uma boa solução. Porém, se a população tiver muitos indivíduos, apesar de cobrir um espaço de busca maior, o tempo para montar cada geração é maior, encarecendo a computação. Normalmente a população é inicializada com $n = 50$ indivíduos.

3.3.4 Elitismo

O elitismo é baseado na idéia de que os melhores indivíduos da presente população devem sobreviver para a próxima, sem serem modificados pela recombinação ou por mutação. Assim, o melhor indíviduo (ou alguns dos melhores) é copiado diretamente para a próxima geração e, em seguida, as operações de recombinação e mutação são realizadas normalmente.

3.4 Considerações

Nesse capítulo foram apresentados o algoritmo genético básico e seus operadores. No capítulo 4 apresentaremos os trabalhos relacionados com os quais nossos resultados serão comparados e no capítulo 5 explicaremos a metodologia utilizada e como foram configurados os operadores genéticos.

4 TRABALHOS RELACIONADOS

4.1 Introdução

Nesse capítulo apresentaremos os trabalhos que nos forneceram o embasamento teórico e os dados para a validação dos resultados apresentados no capítulo 6. Na seção 4.2 mostramos as contribuições do trabalho de Collado-Vides (COLLADO-VIDES, 1993a), (COLLADO-VIDES, 1993b) utilizando gramáticas livre de contexto para a predição de UIG. Nas seções seguintes falamos sobre os dois bancos de dados que possuem os dados utilizados para a comparação com os resultados que obtivemos. Por fim, na seção 4.5 são indicados outros trabalhos realizados na área de regulação gênica.

4.2 Geração de UIGs utilizando Gramática Livre de Contexto

Collado-Vides em seu estudo (COLLADO-VIDES, 1993a), (COLLADO-VIDES, 1993b) tinha como objetivo desenvolver uma teoria de regulação gênica na forma de uma teoria lingüística. Essa busca por uma teoria lingüística adequada foi necessária devido a um trabalho do próprio Collado-Vides (COLLADO-VIDES, 1991),

que demonstrou que gramáticas livres de contexto são inadequadas para a descrição da informação sobre a regulação contida no DNA. Como o trabalho de Collado-Vides baseia-se em teoria de linguagens formais não nos aprofundaremos aqui pois foge ao escopo do trabalho proposto. Para um maior entendimento de seu trabalho sugerimos a leitura de (COLLADO-VIDES, 1993a), (COLLADO-VIDES, 1993b) e (COLLADO-VIDES, 1991).

Algumas importantes características da regulação gênica foram apresentadas nos trabalhos de Collado-Vides (COLLADO-VIDES, 1993a), (COLLADO-VIDES, 1993b). A primeira delas é a necessidade da existência de um fator no sítio proximal (região que vai de -60 a $+20$ na seqüência de DNA, com relação à posição do início da transcrição, em destaque na figura 4.1). A posição $+1$ é representada pelo nucletídeo onde começa a transcrição e a posição dos fatores de transcrição são definidas de acordo com ela. Assim, os fatores à esquerda da posição $+1$ recebem posições negativas e os à direita posições positivas.

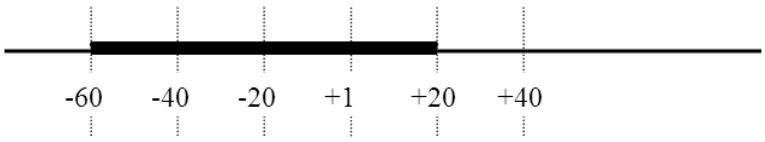


Figura 4.1: Em destaque o sítio proximal

Além disso, observou-se a existência de um padrão de posicionamento seguido pelos operadores e pelos ativadores, chamado princípio da precedência proximal. Esse princípio diz que os operadores proximais localizam-se à direita do promotor, enquanto os ativadores proximais estão posicionados à esquerda deste. Para sítios remotos (fora do intervalo -60 a $+20$) a relação de precedência entre os fatores é regida pelo princípio da precedência posicional que diz que dada a coordenada C_1

do nucleotídeo localizado no meio do sítio A, e a coordenada C_2 do nucleotídeo localizado no meio de um sítio B, A precede B se $C_1 < C_2$. Por exemplo, na figura 4.2, dizemos que o fator de transcrição FNR, cujo nucleotídeo central está na posição -40, precede o fator de transcrição CRP, cujo nucleotídeo central está na posição $+30$, pois $-40 < +30$.

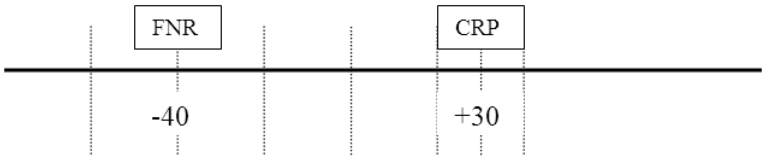


Figura 4.2: Princípio da Precedência Posicional

Além das características apresentadas acima, os trabalhos de Collado-Vides deram origem a novos trabalhos, como por exemplo (ROSENBLUETH et al., 1996), que tenta determinar a que distância os operadores e ativadores estão do promotor, baseando-se nos critérios estabelecidos por Collado-Vides.

4.3 TRACTOR_DB

TRACTOR_DB (GONZÁLEZ et al., 2005) (do inglês TRAnscription FaCTORs' predicted binding sites in prokariotic genomes) é um banco de dados relacional que contém predições computacionais para novos membros de 74 regulons (conjunto de genes regulados por um fator de transcrição) nos genomas de 17 gamma-proteobactérias. As informações contidas no banco de dados podem ser acessadas em `http://tractor.lncc.br`.

A partir dos sítios de ligação conhecidos para os fatores de trascrição da *E.coli* é construído um modelo estatístico. Esse modelo é utilizado para predição de novos

sítios candidatos tanto na própria *E.coli* quanto em outras gamma-proteobactérias relacionadas a ela.

4.3.1 Busca no TRACTOR_DB

A informação contida no TRACTOR_DB pode ser acessada de cinco modos diferentes (GONZÁLEZ et al., 2005), utilizando o botão *Browse* da página web (Figura 4.3).

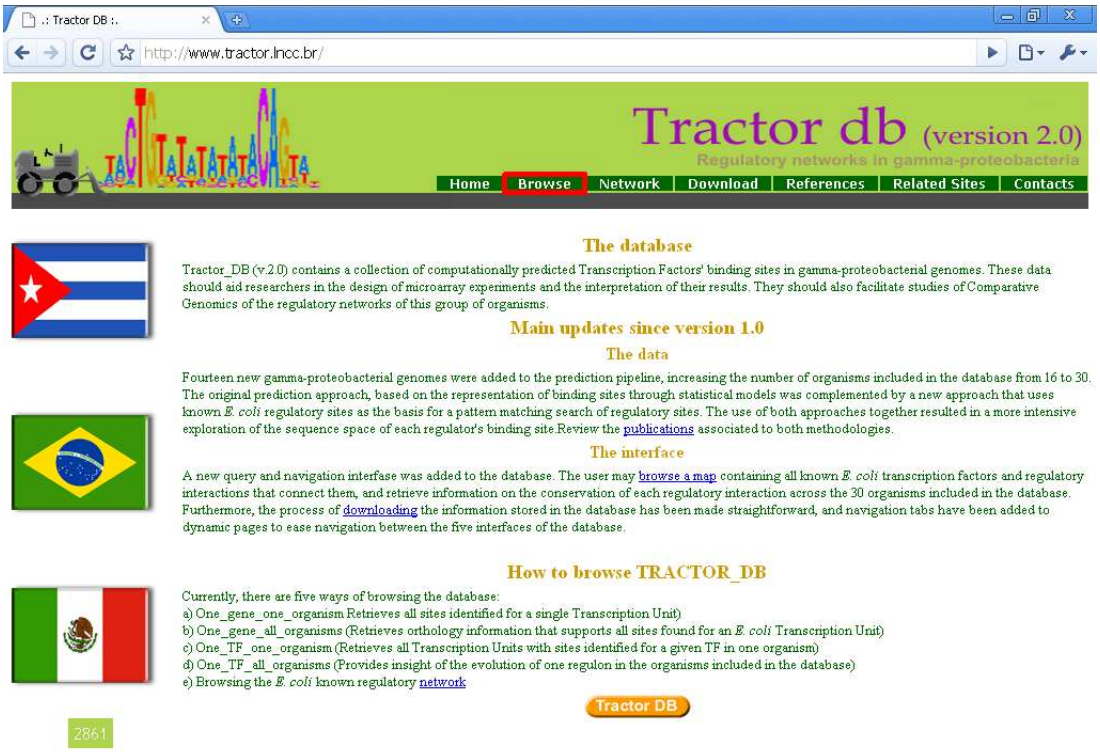


Figura 4.3: Página principal do TRACTOR_DB. Em vermelho destacamos o botão *Browse*.

A primeira maneira é chamada de "um gene, um organismo" e mostra todos os sítios encontrados para uma dada unidade de transcrição. Esse modo pode ser acessado

marcando o nome de um organismo e preenchendo o nome de um gene (nas áreas 1 e 3 da figura 4.4, respectivamente).

Figura 4.4: Página de busca do TRACTOR_DB. Destacamos em vermelho três áreas e o botão *Network*. A área 1 é utilizada para a busca por organismo e regulon, a 2 para exibição de informações sobre a conservação de um regulon, a área 3 para busca por gene e o botão *Network* utilizado para a exibição da rede regulatória da *E. coli*.

O segundo modo de busca, chamado "um gene, todos os organismos", retorna todas as informações de ortologia (informações de outros organismos que possuem ancestral comum) que confirmam os sítios preditos na região localizada à esquerda da unidade (*upstream*) de transcrição da *E. coli* na qual o gene está localizado. Para realizar a busca marca-se a opção *All organisms* (na área 2 da figura 4.4) e preenche-se o nome de um gene (na área 3 da figura 4.4).

No modo "um fator de transcrição, um organismo" são mostrados todos os membros do regulon do fator de transcrição, para um dado organismo. Para obter tal informação marca-se o nome do organismo desejado e seleciona-se o nome do fator de transcrição no menu ao lado (área 1 da figura 4.4).

O quarto modo, "um fator de transcrição, todos os organismos", mostra a informação relativa à conservação de regulons simples e complexos da *E.coli* nos outros 16 organismos do TRACTOR_DB. Essa busca é feita marcando a opção *All organisms* e selecionando o nome do fator de transcrição no menu ao lado (área 2 da figura 4.4).

Para obtermos os fatores de transcrição que regulam o operon *fixA-fixB-fixC-fixX-yaaU* da *Escherichia coli* K12 a fim de compararmos os resultados do TRACTOR_DB com o nosso trabalho, utilizamos a busca do tipo "um gene, um organismo". As entradas utilizadas na busca foram *Escherichia coli K12* na área 1 e o gene *fixA* na área 3, como pode ser visto na figura 4.5. O resultado retornado pode ser visto na figura 4.6 e na tabela 4.1. Para obter tais predições dois métodos foram utilizados, um baseado em expressões regulares e outro baseado em modelos estatísticos, apresentados nas subseções 4.3.2 e 4.3.3.

Tabela 4.1: Fatores de transcrição, suas respectivas posições previstas pelo TractorDB e o tipo de método utilizado para obter o resultado

Fator de Transcrição	Posição	Método
CRP	-280	Modelo Estatístico
CRP	-279	Modelo Estatístico
CRP	-178	Expressões Regulares
CRP	-160	Modelo Estatístico
CRP	-121	Expressões Regulares
CRP	-103	Modelo Estatístico
FNR	-278	Modelo Estatístico
FNR	-229	Modelo Estatístico
CaiF	-192	Modelo Estatístico

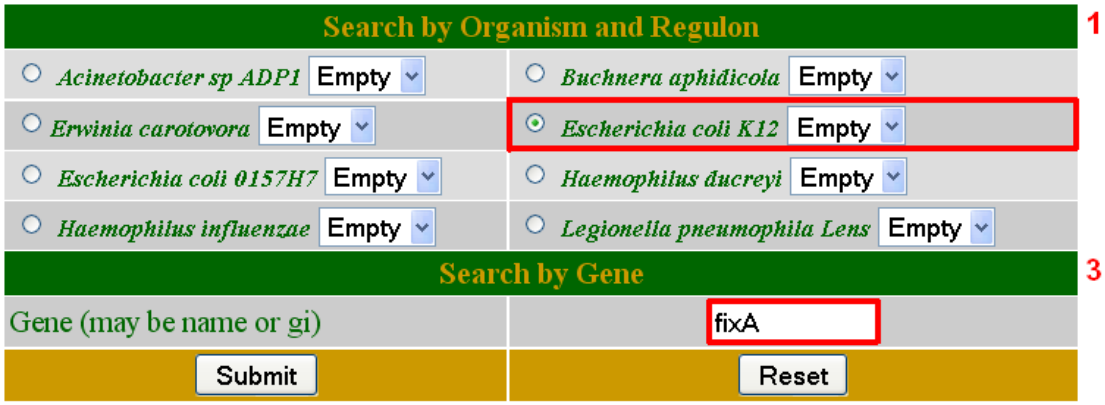


Figura 4.5: Parâmetros da busca "um gene, um organismo" no TRACTOR_DB para obter os fatores de transcrição que regulam o operon *fixA-fixB-fixC-fixX-yaaU*. Destacamos em vermelho os campos a serem preenchidos.

4.3.2 Expressões Regulares

Uma expressão regular é uma expressão que descreve, através de uma linguagem formal composta por um alfabeto e algumas operações, um determinado conjunto de cadeias (FRIEDL, 1997). Por exemplo, dados o alfabeto $A = \{A, C, T, G\}$ e o operador $O = \{*\}$, a expressão regular e, dada por

$$e = ATC * G,$$

representa o conjunto $C = \{ATG, ATCG, ATCCG, ATCCCG, ATCCCCG, \ldots\}$, pois o operador $O = \{*\}$ indica que o elemento anterior a ele pode ser repetido 0 ou mais vezes.

Tal linguagem pode ser interpretada por um programa que examina uma seqüência de caracteres e identifica as partes que casam (total ou parcialmente) com o padrão descrito pela expressão.

Nesse método cada sítio de ligação conhecido de um determinado fator de transcri-

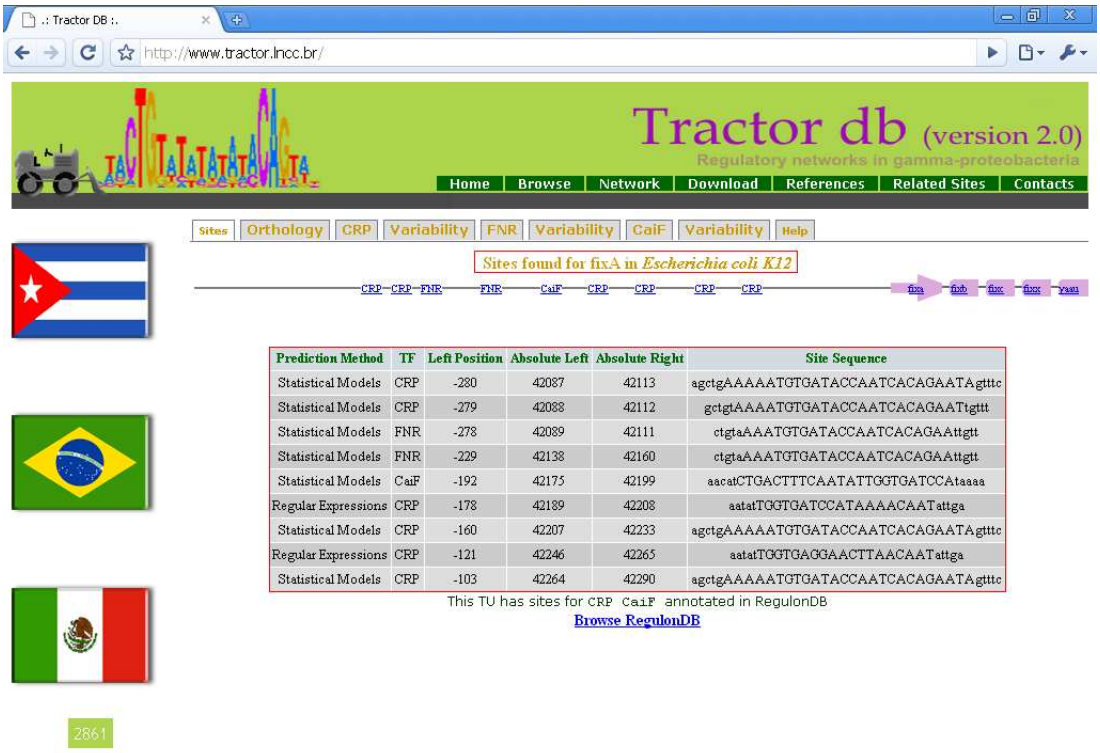


Figura 4.6: Página de resultado do TRACTOR_DB para a busca do tipo "um gene, um organismo". Destacamos em vermelho os resultados obtidos. A informação sobre a posição do fator de transcrição é fornecida de três maneiras: a posição do nucleotídeo mais a esquerda em relação ao promotor e as posições absolutas do nucleotídeo mais a esquerda e mais a direita do seqüência do fator.

ção é tratado como uma expressão regular e é verificado se esse mesmo sítio ocorre em regiões ortólogas (regiões com mesma função mas que ocorrem em espécies diferentes) não codificantes (GUÍA et al., 2005). A importância estatística de cada predição na região ortóloga é estimada utilizando a matriz de pesos dos fatores de transcrição.

A transformação dos sítios de ligação em expressões regulares segue as seguintes regras:

1. Se o comprimento do sítio for par, menor do que 14 nucleotídeos, sua seqüência é expandida nos dois extremos, até atingir 14 nucleotídeos. A posição dos dois nucleotídeos centrais do sítio original é mantida.

2. Se o comprimento do sítio for ímpar, menor do que 13 nucleotídeos, sua seqüência é expandida nos dois extremos, até atingir 13 nucleotídeos.

3. Se o comprimento do sítio for acima de 13 (no caso de tamanho ímpar) ou 14 (no caso de tamanho par) nucleotídeos então o tamanho do sítio é respeitado.

Após a conversão, é realizada uma busca das expressões regulares em regiões regulatórias ortólogas de 450 pares de base. No processo de reconhecimento de padrões é permitido um número mínimo de erros de casamento, para que a quantidade encontrada seja da mesma ordem do número de sítios conhecidos para um determinado fator de transcrição na *E. coli*. Além disso, um número de corte é selecionado para reduzir a quantidade de resultados de baixa qualidade.

Para estimar a significância estatística dos resultados são construídas matrizes de peso utilizando o programa CONSENSUS (HERTZ; HARTZELL; STORMO, 1990), para alinhar as seqüências originais da *E. coli* com os resultados encontrados com as expressões regulares.

O programa CONSENSUS utiliza um algoritmo guloso para buscar a matriz de pesos com baixa probabilidade de ocorrer ao acaso, ou seja, que possui um alto conteúdo de informação, dado por:

$$I_{seq} = \sum_{b=A}^{T} f_b * log_2(f_b/p_b)$$

onde b representa as bases A, C, G e T, f_b é a freqüência de cada base nas sub-seqüências de alinhamento (informação contida na matriz de alinhamento) ep_b é a fração de cada base no genoma.

O algoritmo inicia formando uma matriz de alinhamento (figura 4.7-B) para cada um dos L-mers (sub-seqüências de tamanho L, sendo no exemplo $L = 6$) da primeira seqüência apresentada na figura 4.7-A. Em geral, o tamanho$L$ é determinado através do conhecimento bioquímico que se tem do problema estudado ou de algum semelhante. A partir desse conhecimento são testados valores dentro de um intervalo ao redor do valor conhecido para se eleger o melhor (HERTZ; HARTZELL; STORMO, 1990).

Após a formação das matrizes de alinhamento, cada uma delas é combinada com os L-mers da segunda seqüência (figura 4.7-C). Somente a melhor matriz gerada por cada pai, ou seja, a com maior conteúdo de informação, é salva em cada ciclo. Por fim, as matrizes salvas no ciclo anterior são combinadas com cada L-mer da terceira seqüência (figura 4.7-D) e a melhor matriz desse ciclo (figura 4.7-E) é utilizada para calcular a matriz de pesos (ver seção 2.5).

4.3.3 Modelos Estatísticos

Para cada fator de transcrição da *E. coli* é montado um conjunto de treinamento contendo seqüências de ligação dos fatores de transcrição, extraídas do RegulonDB (SALGADO et al., 2006), e seqüências de regiões ortólogas aos membros do regulon da *E. coli*. Esse conjunto é filtrado duas vezes para eliminar as seqüências de ligação fracas e dois modelos estatísticos são construídos. Com o programa CONSENSUS (ver sub-seção 4.3.2) são criadas matrizes de peso e com o método Gibbs-SAMPLER (LAWRENCE et al., 1993) são criadas seqüências de alinhamento.

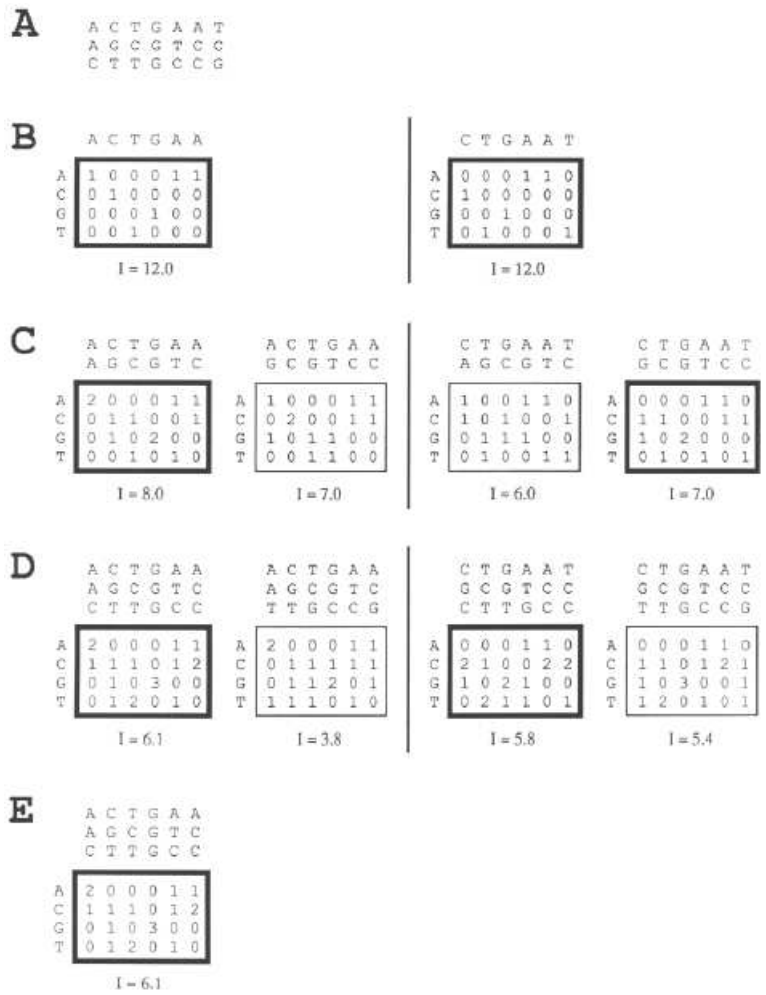


Figura 4.7: Passos do algoritmo do programa CONSENSUS (HERTZ; HARTZELL; STORMO, 1990). A - Seqüências de alinhamento utilizadas para montar a matriz; B - Matrizes dos L-mers de tamanho 6 da primeira seqüência,I é o valor do conteúdo de informação das matrizes; C e D - matrizes formadas nos ciclos 3 e 4 do programa; E - melhor matriz

O algoritmo básico do método Gibbs-SAMPLER consiste em, dado um conjunto de N seqüências $S_1, \ldots, S_N$, identificar o padrão mais provável de acontecer. Em cada uma das seqüências mencionadas anteriormente são procurados segmentos similares de tamanho W.

O algoritmo possui duas estruturas de dados, uma para descrição de padrões e outra para os alinhamentos. Os padrões são representados por um modelo probabilístico de freqüências residuais f_r, dadas por

$$f_r = f_e - f_o$$

onde f_e e f_o são a freqüência esperada e a freqüência observada, respectivamente. A descrição do padrão (dada pelas variáveis $q_{i,1}, \ldots, q_{i,20}$, com i variando de 1 até W) é complementada por uma descrição probabilística das freqüências, chamada de probabilidade secundária, cujo resíduos ocorrem em sítios não descritos pelo padrão (freqüência secundária), representada por $p_1, \ldots, p_{20}$.

A segunda estrutura de dados, que representa o alinhamento, é composta por um conjunto de posições a_k, com k variando de 1 até N, para o padrão comum encontrado nas seqüências. O padrão mais comum é obtido localizando-se o alinhamento que maximiza a razão entre a probabilidade do padrão e a probabilidade secundária. Para um descrição detalhada do método ver (LAWRENCE et al., 1993).

Em seguida, é realizada uma busca em regiões regulatórias de oito genomas (que compartilham pelo menos 30% de genes ortólogos com a *E.coli*), por candidatos a novos sítios de ligação. Para as matrizes de peso a varredura é feita com o uso do programa PATSER (ver seção 2.5) e, para as seqüências de alinhamento, com o programa DSCAN (NEUWALD; LIU; LAWRENCE, 1995).

Para um dado conjunto de tamanho T de seqüências, o DSCAN retorna o valor esperado (*e-value*), número de sítios com pontuação maior ou igual a que seria retornada se a mesma busca fosse realizada em um conjunto de tamanho S de seqüências aleatórias, onde $T = S$. Além do valor esperado, também são retornadas as probabilidades (*p-value*) de encontrar um segmento de pontuação melhor em uma seqüência aleatória de mesmo tamanho.

Os candidatos dos dois modelos (PATSER e DSCAN) são agrupados, filtrados com base nas informações de ortologia e pontuação, e separados para cada um dos genomas utilizados no processo.

Na figura 4.8 pode ser visto um fluxograma simplificado do processo descrito nesta sub-seção.

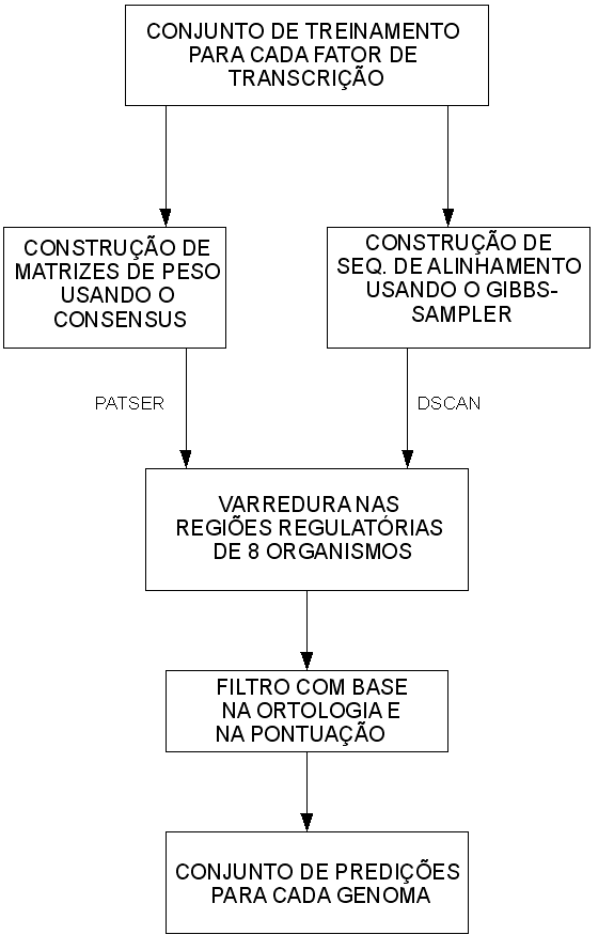


Figura 4.8: Fluxograma simplificado do processo preditivo dos métodos estatísticos utilizados pelo TRACTOR_DB.

4.4 RegulonDB

RegulonDB (SALGADO et al., 2006) é um banco de dados que integra conhecimento biológico dos mecanismos que regulam o início da transcrição da *E. coli* e sobre genes e sinais regulatórios, que se organizam como operons no cromossomo. Atualmente o banco de dados encontra-se na versão 6.2 (GAMA-CASTRO et al., 2008) e pode ser acessado em `http://regulondb.ccg.unam.mx/`.

4.4.1 Busca no RegulonDB

A busca no RegulonDB (figura 4.9) pode ser realizada de quatro maneiras diferentes: busca por gene, busca por condição de crescimento (condições experimentais nas quais uma cepa é desenvolvida em experimentos particulares para a observação de mudanças na expressão gênica), busca por regulon e busca por operon (ver subseção 2.4.1).

Para obtermos os fatores de transcrição preditos pelo RegulonDB para o operon *fixA-fixB-fixC-fixX-yaaU* a fim de compararmos seus resultados com o nosso trabalho utilizamos a busca por operon, com a entrada *fixA*. O resultado retornado pode ser visto nas figuras 4.10 e 4.11 e na tabela 4.2.

Tabela 4.2: Posições previstas pelo RegulonDB para os fatores CRP, CaiF e FNR

Fator de Transcrição	Posição	Tipos de Evidência
CRP	-137	GEA e HIBSCS
CRP	-80	GEA e HIBSCS
CaiF	-144	BPP e GEA
CaiF	-84	BPP e GEA
FNR	-204	AIBSCS E GEA

Figura 4.9: Página principal do RegulonDB. Em vermelho destacamos as quatro opções de busca.

Para obter tais predições foram usados quatro métodos, classificados como evidências fortes (cujo dados experimentais dão alto nível de certeza da existência do fator) e fracas (as demais), que serão apresentados nas subseções a seguir.

4.4.2 Análise de Expressão Gênica - GEA

A análise de expressão gênica (GEA, do inglês *gene expression analysis*) em bactérias é realizada, em geral, utilizando-se um gene repórter, sendo o gene lacZ da *E. coli* que codifica β-galactosidade o mais popular. A expressão gênica é medida através da

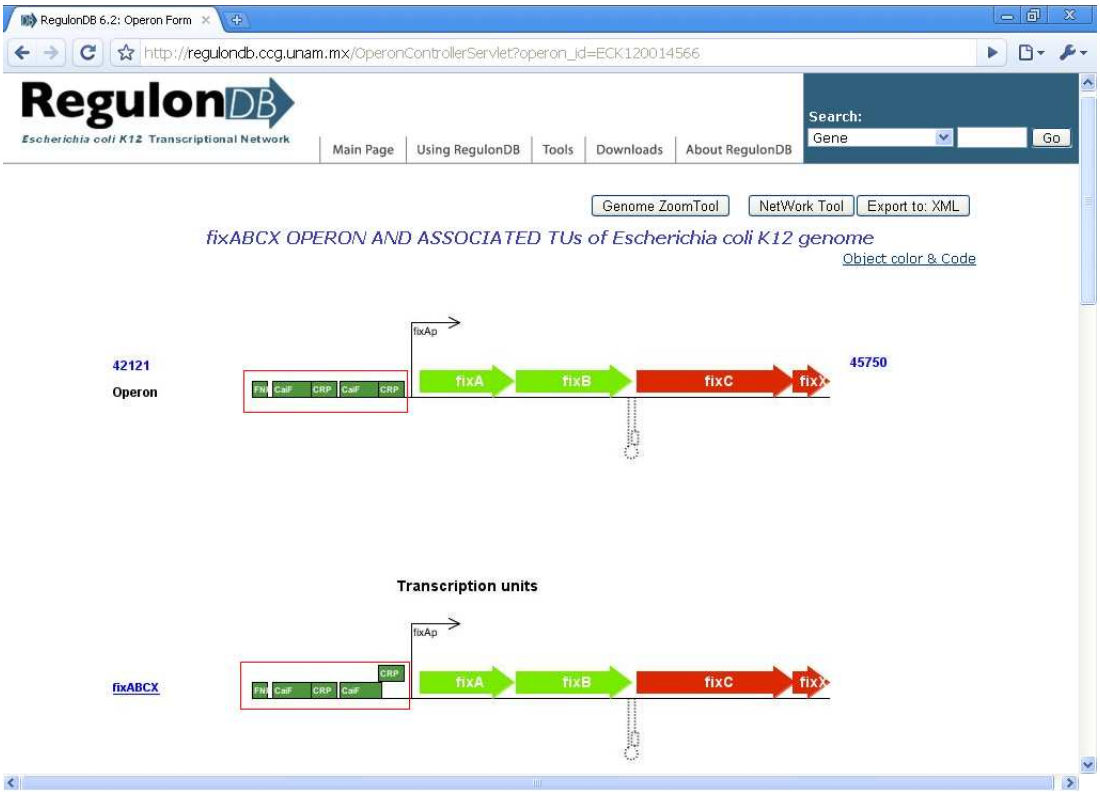


Figura 4.10: Primeira parte do resultado obtido pela busca. Destacamos os fatores de transcrição em sua representação gráfica.

fusão transcricional, onde o gene repórter (que contém seu próprio códon de iniciação da tradução) é fundido a uma seqüência de DNA que inclui apenas o promotor e os sítios reguladores. Em seguida, observa-se a quantidade de β-galactosidase expressada, que é proporcional à quantidade transcrita do gene cuja expressão se quer analisar. Em (SILHAVY; BECKWITH, 1985) podem ser vistos mais detalhes sobre o método.

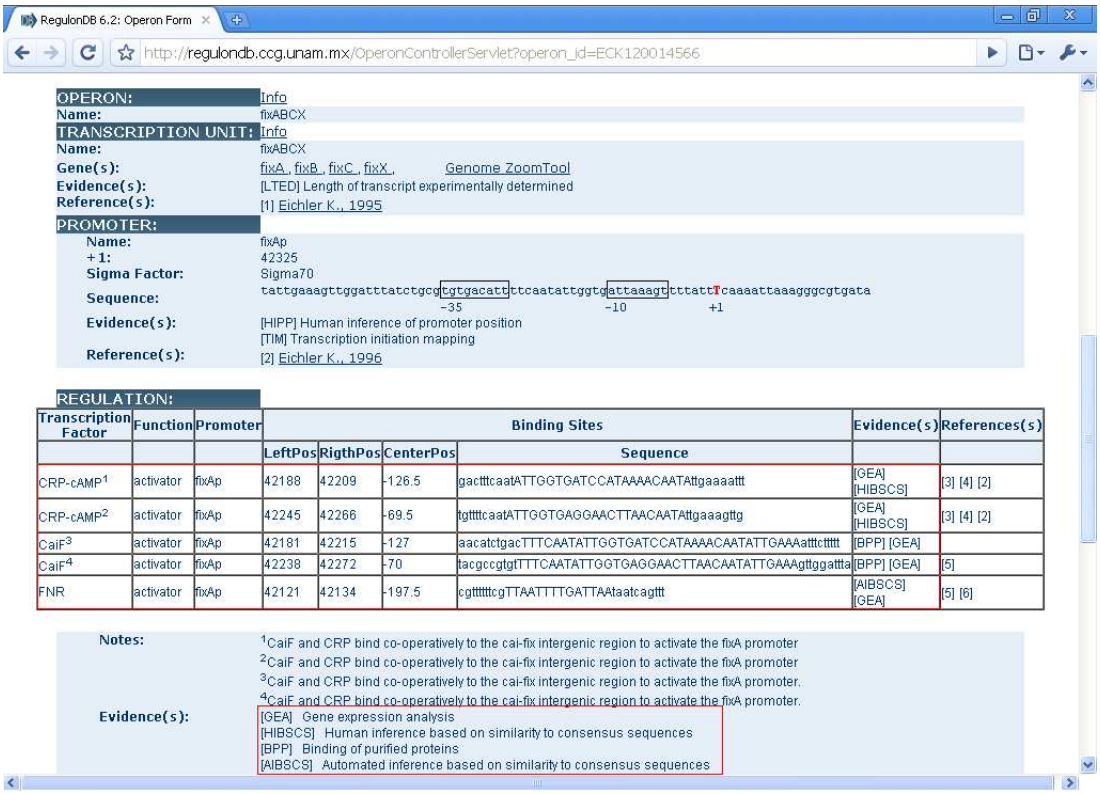


Figura 4.11: Segunda parte do resultado obtido pela busca. Em vermelho destacamos os tipos de evidências que comprovam a posição dos fatores de transcrição encontrados. A informação sobre a posição do fator de transcrição é fornecida de três maneiras: a posição do nucleotídeo central em relação ao promotor e as posições absolutas do nucleotídeo mais a esquerda e mais a direita do seqüência do fator.

4.4.3 Inferência Humana Baseada em Similaridade com Seqüências CONSENSUS - HIBSCS

A inferência humana baseada em similaridade a seqüências CONSENSUS (HIBSCS, do inglês *human inference based on similarity to CONSENSUS sequences*) é realizada por um especialista que, com o auxílio de um programa de leitura de seqüências, identifica os sítios de ligação candidatos.

4.4.4 Ligação de Proteínas Purificadas - BPP

A ligação de proteínas purificadas (BPP, do inglês *binding of purified proteins*) é um conjunto de métodos para investigação de regiões de DNA onde determinadas proteínas se ligam. Um exemplo de BPP é o *DNA footprinting* que consiste de quatro etapas:

- Primeiro é realizada uma reação de PCR (MULLIS, 1990) para amplificar e marcar a região de interesse que contém o provável sítio de ligação da proteína.
- Em seguida a proteína a ser estudada é adicionada a uma parte do DNA marcado, separando a outra parte para comparação futura.
- Uma enzima de restrição é adicionada às duas partes do DNA. A clivagem ocorre nos pontos do DNA onde a enzima reconhece um padrão determinado na seqüência de nucleotídeos, por isso o tempo de reação deve ser o suficiente para cortar cada molécula de DNA somente uma vez. A proteína ligada a uma região da fita de DNA protege a porção da cadeia a qual está ligada de sofrer um corte.
- Após a clivagem é realizada uma eletroforese em gel, técnica de separação de moléculas que envolve migração de partículas em um determinado gel durante a aplicação de uma diferença de potencial (GARNER; REVZIN, 1981). Como a separação ocorre de acordo com a massa das moléculas (as de menor massa "correm" mais rapidamente) o grupo de fitas de DNA sem a proteína estará cortado formando uma distribuição do tipo "escada", enquanto o que possui a proteína apresenta uma quebra na distribuição, na parte onde o DNA foi protegido.

Para maiores detalhes sobre o método de *DNA footprinting* consultar (GALAS; SCHMITZ, 1978).

4.4.5 Inferência Automatizada Baseada em Similaridade com Seqüências CONSENSUS - AIBSCS

A inferência automatizada baseada em similaridade com seqüências CONSENSUS (AIBSCS, do inglês *automated inference based on similarity to CONSENSUS sequences*) é um conjunto de métodos computacionais usados para identificar um sítio de ligação. Como exemplo temos o PATSER (ver seção2.5).

4.5 Outros trabalhos

Outros trabalhos podem ser citados sobre o estudo das redes de regulação gênica, tais como (ZHANG; KUO; BRUNKHORS, 2006; TOH; HORIMOTO, 2002). O primeiro trabalho utiliza redes neurais do tipo *feed-forward* para realizar predições de posicionamento do promotor em *E. coli* e o segundo utiliza análise de *clusters* e Modelo gráfico Gaussiano para tentar montar redes de regulação gênica de *Saccharomyces cerevisiae*. Optamos por não nos aprofundar nesses dois trabalhos pois não nos fornecem informações disponíveis em bancos de dados para comparamos nossos resultados.

4.6 Considerações

Nesse capítulo foram apresentados os trabalhos utilizados para embasamento e futura comparação dos resultados obtidos nessa dissertação. A partir da contribuição de Collado-Vides (seção 4.2) retiramos duas características importantes das unidades

de informação genética, a necessidade da presença de um fator no sítio proximal e a não sobreposição de fatores de transcrição. Nas seções seguintes foram apresentados os bancos de dados TRACTOR_DB e RegulonDB, além de apresentarmos como é realizada a busca nos bancos e quais métodos foram utilizados para conseguir tais dados. As informações coletadas serão utilizadas em 6 para medir a acurácia dos resultados obtidos pela metodologia proposta no capítulo 5.

5 METODOLOGIA

Nosso objetivo com a solução proposta a seguir é obter combinações de fatores de transcrição que sejam boas candidatas a serem unidades de informação genética (UIGs).

Temos como hipótese que é possível gerar essas combinações satisfatoriamente utilizando algoritmos genéticos desenvolvidos especificamente para esse problema, utilizando uma codificação e operadores genéticos adequados e a informação gerada pelo programa PATSER. Para avaliar a qualidade dos indivíduos gerados vamos comparar as posições indicadas pelo AG com as posições obtidas pelo RegulonDB e pelo TractorDB.

5.1 Etapas do Trabalho

A metodologia consiste em utilizar a capacidade dos algoritmos genéticos de realizar buscas por espaços irregulares ou que não são inteiramente conhecidos (GOLDBERG, 1989) para descobrir os melhores candidatos a UIG.

Nosso trabalho divide-se nas seguintes etapas:

- Utilização das estratégias propostas por (WANDERLEY et al., 2008) para gerar avaliação para os possíveis candidatos a UIG.
- Análise quantitativa dos resultados obtidos pelos AGs em relação aos dados armazenados no TractorDB e RegulonDB.
- Análise qualitativa dos resultados obtidos pelos AGs em relação aos dados armazenados no TractorDB e RegulonDB.
- Proposição de novos operadores para o refinamento dos resultados obtidos.

Para a geração da pontuação utilizamos o programa PATSER, uma seqüência regulatória de 450 pares de bases do operon *fixA-fixB-fixC-fixX-yaaU* da *E.coli K12* e matrizes de alinhamento para cada um dos fatores de transcrição (ver tabela 2.1).

A parte central de nossa metodologia consiste no ajuste do algoritmo genético para satisfazer as restrições impostas (necessidade de existência de um fator de transcrição no sítio proximal e não sobreposição de fatores). Por esse motivo, além da necessidade de estabelecer quais os melhores parâmetros a serem utilizados também é necessário propor operadores genéticos que preservem as características básicas de uma UIG.

5.2 Avaliação dos candidatos a UIG

Para avaliar a qualidade dos candidatos a UIG e posteriormente decidir para quais deles as análises quantitativa e qualitativa serão realizadas, utilizaremos as quatro estratégias propostas por (WANDERLEY et al., 2008).

5.2.1 Primeira estratégia

A primeira estratégia proposta teve como base os algoritmos genéticos clássicos, utilizando parâmetros simples, tais como representação binária (sem restrição do intervalo de números a serem gerados) e função de avaliação composta pelo somatório simples das pontuações retornadas pelo PATSER.

5.2.1.1 Codificação dos Indivíduos

Nessa estratégia os indivíduos são representados como cadeias binárias de 90 bits, sendo os primeiros 36 são destinados à representação dos fatores de transcrição, 6 no total, e o restante para a representação das posições desses fatores numa seqüência de DNA de 450 pares de base. O número de fatores de transcrição representados no indivíduo do AG indica a quantidade máxima de fatores que podem estar presentes simultaneamente em uma seqüência regulatória. Desse modo, os indivíduos são da forma $FT_1 \ldots FT_6\ POS_1 \ldots POS_6$, onde FT_i são os fatores de transcrição e POS_i a posição em que se encontra o fator FT_i (figura 5.1).

Figura 5.1: Representação gráfica da codificação dos indivíduos do algoritmo genético da estratégia 1.

5.2.1.2 Função de Avaliação

A função de avaliação consiste do somatório das pontuações obtidas através do programa PATSER para cada um dos fatores codificados e suas respectivas posições.

O valor da função é calculado da seguinte maneira:

- Se o valor codificado no fator não for válido: não soma valor nenhum à função de avaliação
- Se a posição sorteada para o fator não for válida ou não possuir pontuação positiva: não soma valor nenhum à função de avaliação
- Caso contrário, soma-se o valor da pontuação.

Matematicamente, a função de avaliação pode ser expressa por:

$$FA_1 \quad = \quad 1 + SCORE(FT_1) + \cdots + SCORE(FT_6)$$

para os pares (FT, POS) que possuem pontuação positiva.

5.2.2 Segunda estratégia

Na segunda estratégia foi utilizada a mesma codificação utilizada na estratégia anterior(ver 5.2.1.1) modificando somente a função de avaliação.

5.2.2.1 Função de Avaliação

Na função de avaliação apresentada em 5.2.1.2 não considerava-se o número de fatores que não eram válidos. Nessa nova função, a avaliação final do indivíduo é calculada somando as pontuações obtidas por cada um dos fatores e depois dividindo esse valor pelo número de fatores não válidos. Essa função é representada matematicamente por:

$$FA_2 \quad = \quad (1 + SCORE(FT_1) + \cdots + SCORE(FT_6))/(1 + NFT)$$

onde NFT representa o número de fatores de transcrição não ativos no indivíduo que está sendo avaliado.

5.2.3 Terceira estratégia

5.2.3.1 Codificação dos Indivíduos

A representação dos indivíduos é feita por cadeias binárias, tal como nas duas estratégias anteriores, porém o tamanho da cadeia é de 60 bits. Essa modificação no tamanho aconteceu pois são representados 4 fatores de transcrição e suas respectivas posições ao invés de 6, como usado até então. O objetivo dessa redução no número de fatores de transcrição foi conseguir restringir mais o espaço de busca, aumentando a eficiência do AG. Também foi modificada a ordem de representação dos fatores e posições, tendo como resultado indivíduos da forma $FT_1, POS_1 \ldots FT_4, POS_4$ (figura 5.2.3.1). Essa nova representação aumenta a eficiência da busca, pois o operador de recombinação torna-se menos disruptivo e com isso diminuem as chances de afetar um indivíduo que tinha uma boa avaliação.

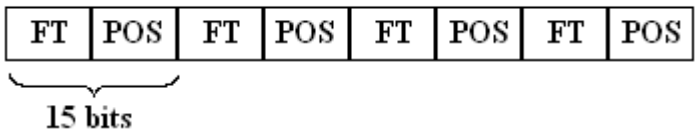


Figura 5.2: Representação gráfica da codificação dos indivíduos do algoritmo genético da estratégia 3.

5.2.4 Quarta estratégia

5.2.4.1 Codificação dos Indivíduos

Nessa nova codificação as partes dos indivíduos que representam as posições dos fatores não geram mais números de 1 a 450 e sim reais na faixa $[0, 1]$. No momento da avaliação, esse número real é mapeado no número de pontuações positivas do fator de transcrição que corresponde a tal posição. Essa codificação foi pensada para permitir um processo dinâmico de mapeamento, uma vez que o número de posições com pontuação positivo é diferente para cada um dos fatores.

Por exemplo, suponha que o fator em questão tenha POS_{SP} posições com pontuações positivas. A posição da tabela, que contém a pontuação a ser utilizada na função de avaliação, é dada por:

$$POS_T \quad = \quad 1 + round(POS * (POS_{SP} - 1))$$

onde POS representa o valor de posição codificado no indivíduo.

As figuras 5.3 e 5.4 ilustram como é feito o mapeamento entre o valor real codificado no indivíduo e a posição da tabela.

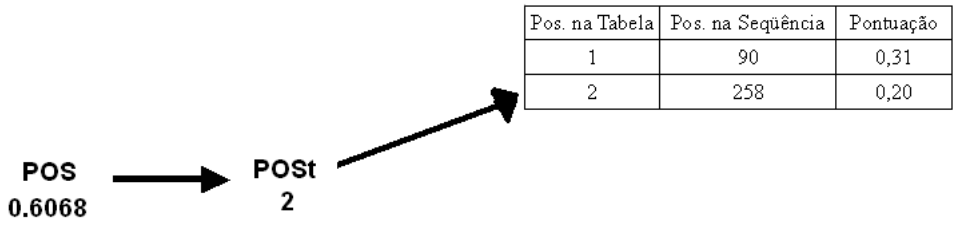
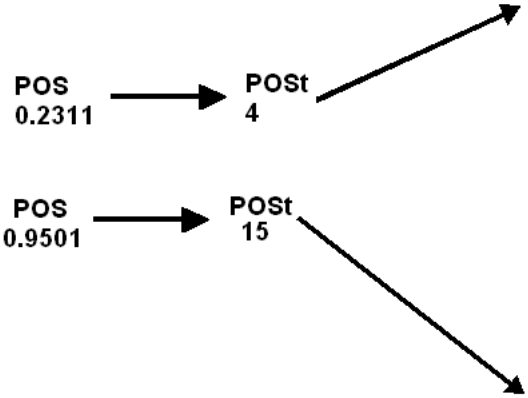


Figura 5.3: Mapeamento entre POS e POS_T para o fator de transcrição $FadR$



Pos. na Tabela	Pos. na Seqüência	Pontuação
1	1	2,92
2	145	2,26
3	149	0,61
4	177	1,73
5	181	3,18
6	194	0,48
7	203	2,74
8	238	1,70
9	258	2,66
10	260	0,15
11	261	1,96
12	281	0,77
13	285	0,77
14	305	1,45
15	308	3,03
16	340	0,91

Figura 5.4: Dois mapeamentos entre POS e POS_T para o fator de transcrição $AraC$.

5.3 Considerações

Nesse capítulo apresentamos nossa proposta de resolução do problema de encontrar UIGs que seriam boas candidatas. Na tabela 5.1 podemos ver um resumo de todas as estratégias utilizadas, com as respectivas codificações, número de bits utilizados na representação, funções de avaliação e formato do indíviduo.

No capítulo 6, mostraremos os resultados obtidos com cada uma das estratégias propostas e as análises qualitativa e quantitativa, realizadas com os resultados do TRACTOR_DB e RegulonDB.

Tabela 5.1: Resumo das estratégias utilizadas

	Codificação do Indivíduo	Número de bits	Função de Avaliação
1	$FT_1 \ \ldots \ FT_6 \ POS_1 \ \ldots \ POS_6$	90	$FA_1 = 1 + SCR(FT1) + \\ + \cdots + SCR(FT6)$
2	$FT_1 \ \ldots \ FT_6 \ POS_1 \ \ldots \ POS_6$	90	$FA_1 = (1 + SCR(FT1) + \\ + \cdots + SCR(FT6))/(1 + NFT)$
3	$FT_1 \ POS_1 \ \ldots \ FT_4 \ POS_4$	60	$FA_1 = (1 + SCR(FT1) + \\ + \cdots + SCR(FT4))/(1 + NFT)$
4	$POS_T = 1 + round(POS * (POS_{SP} - 1))$	60	$FA_2 = (1 + SCR(FT1) + \\ + \cdots + SCR(FT4))/(1 + NFT)$

6 RESULTADOS

Para gerar os resultados a serem analisados nesse capítulo executamos dez vezes os testes propostos por (WANDERLEY, 2007) com os parâmetros mostrados na tabela 6.1, para cada uma das abordagens citadas no capítulo 5.

Utilizamos como ambiente de teste o programa MATLAB [1] em um computador com processador Core 2 Duo de 1,8 GHz, com 2,0 GB de memória RAM. Cada teste levou aproximadamente 5 segundos no melhor caso e 20 segundos no pior caso.

Nos casos em que o elitismo era utilizado, optamos por preservar dois indivíduos pois testes preliminares mostraram que, com esse valor, bom resultados eram obtidos.

Nas seções seguintes serão apresentadas as análises realizadas após os testes. Na seção 6.1, todas as estratégias são analisadas do ponto de vista do desempenho do algoritmo genético, através da comparação dos gráficos Geração x Função de Avaliação. Nas seções 6.2 e 6.3 foram analisados somente os resultados da melhor estratégia.

[1] http://www.mathworks.com/

Tabela 6.1: Parâmetros utilizados nos testes

Teste	Número de Gerações	Número de Indivíduos	Probabilidade de Recombinação	Probabilidade de Mutação
1	64	64	0,8	0,05
2	64	64	0,7	0,05
3	64	64	0,6	0,05
4	128	128	0,8	0,05
5	128	128	0,7	0,05
6	128	128	0,6	0,05

6.1 Resultados do Ponto de Vista do Desempenho do Processo Evolutivo do Algoritmo Genético

Após a realização dos testes, como primeira análise optamos por verificar as variações ocorridas nos gráficos Função de Avaliação x Geração de acordo com o uso de elitismo, a variação do tamanho da população e da taxa de recombinação.

Esperamos, tendo como base os conceitos básicos de algoritmos genéticos, que os resultados para uma população maior e utilizando elitismo sejam melhores, uma vez que assim é possível explorar melhor o espaço de soluções e preservar os melhores indivíduos de cada população, respectivamente.

6.1.1 Influência do Uso de Elitismo

Em problemas onde o espaço de soluções é complexo o uso de elitismo é bastante necessário para o aumento da performance do AG, prevenindo a perda da melhor solução encontrada até o momento. Na figura 6.1 é mostrado o efeito do não uso do elitismo para a função de avaliação. Os gráficos da esquerda representam os casos nos quais o elitismo foi utilizado e os da direita os sem o operador de elitismo.

Por diversas vezes o algoritmo genético encontrou indivíduos com bons valores de função de avaliação, porém algumas gerações depois esse indivíduo é perdido, seja pela ação da recombinação ou da mutação. Além disso, o gráfico cujo teste não utiliza elitismo apresenta maior variação no valor da função de avaliação do melhor indivíduo e uma maior diferença da avaliação média da população em relação ao caso em que os indivíduos são preservados.

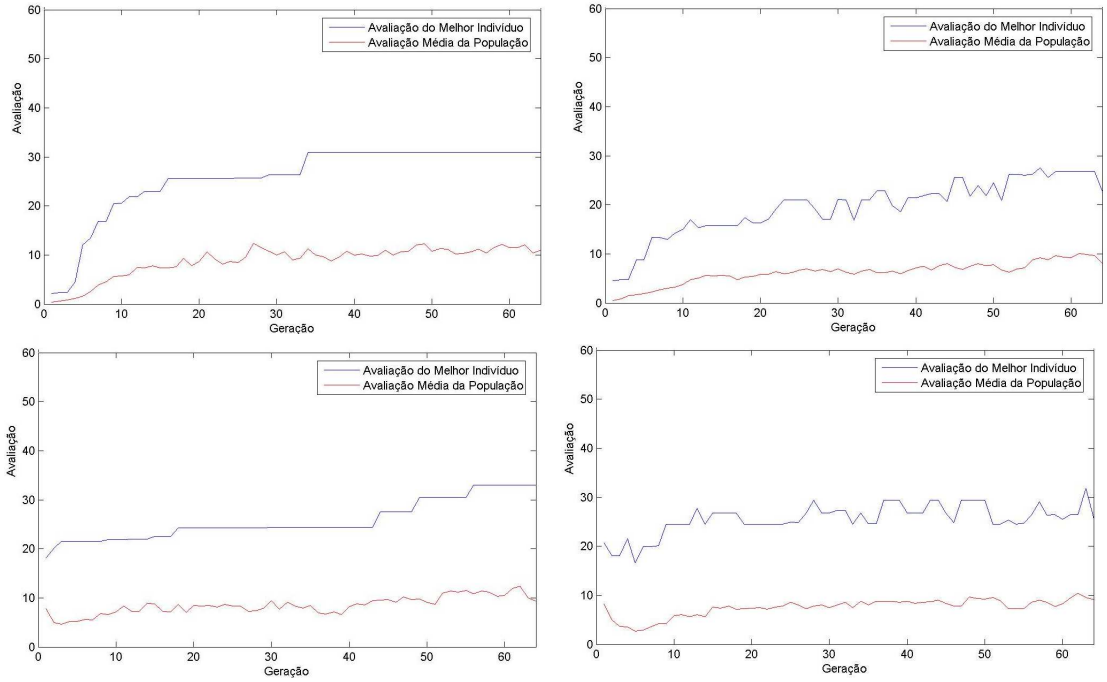


Figura 6.1: Comparação entre os gráficos do testes 5 e 6 com e sem o uso de elitismo. Os gráficos superiores representam o processo evolutivo para o teste 5 (número de gerações = 128, tamanho da população = 128, $p_c = 0,7$) com e sem elitismo, respectivamente. Os gráficos inferiores representam o processo evolutivo para o teste 6 (número de gerações = 128, tamanho da população = 128, $p_c = 0,6$) com e sem elitismo, respectivamente. Nos gráficos é possível ver que a evolução dos gráficos onde o elitismo é utilizado é superior aos que não utilizam.

Como visto, a utilização do elitismo é fundamental para este problema e, por esse motivo, as análises realizadas nas seções e subseções seguintes levarão em consideração somente os testes que utilizaram este operador genético.

6.1.2 Influência da Variação do Tamanho da População

Nesta análise iremos comparar os gráficos dos testes 1 e 4 para a estratégia 1, testes 2 e 5 para a estratégia 2, testes 3 e 6 para a estratégia 3 e testes 3 e 6 para a estratégia 4. Os demais gráficos comparativos encontram-se no apêndice A. Para os testes 1, 2 e 3 o tamanho da população é de 64 indivíduos e para os testes 4, 5 e 6 é de 128 indivíduos.

Nas figuras 6.2, 6.3, 6.4 e 6.5 são mostrados os efeitos da variação do tamanho da população para as quatro estratégias. Todos os testes que utilizavam uma população de 128 indivíduos foram mais bem sucedidos do que os com 64 indívíduos. A diferença é ainda mais acentuada no caso em que a probabilidade de recombinação p_r é mais alta, como na figura 6.2, onde $p_r = 0,8$. Isso ocorre pois os efeitos disruptivos da recombinação são diluídos numa população maior, tornando-se menos críticos (DE JONG; SPEARS, 1990).

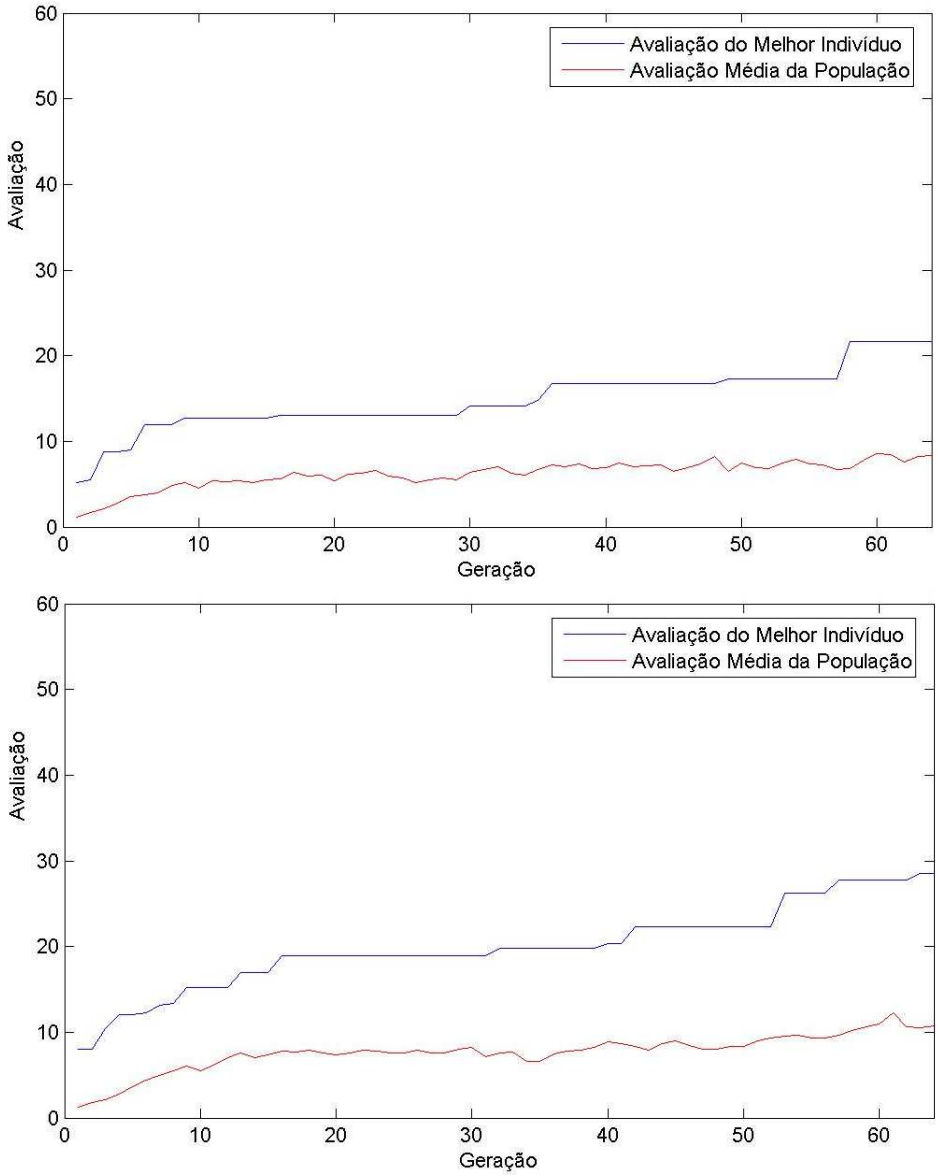


Figura 6.2: Comparação entre os testes 1 e 4, com probabilidade de recombinação $p_r = 0,8$, para a estratégia 1. No primeiro gráfico, correspondente ao teste 1 (com população de 64 indivíduos), podemos ver que a avaliação do melhor indivíduo e a média da avaliação é inferior a do segundo gráfico, correspondente ao teste 4 (com população de 128 indivíduos).

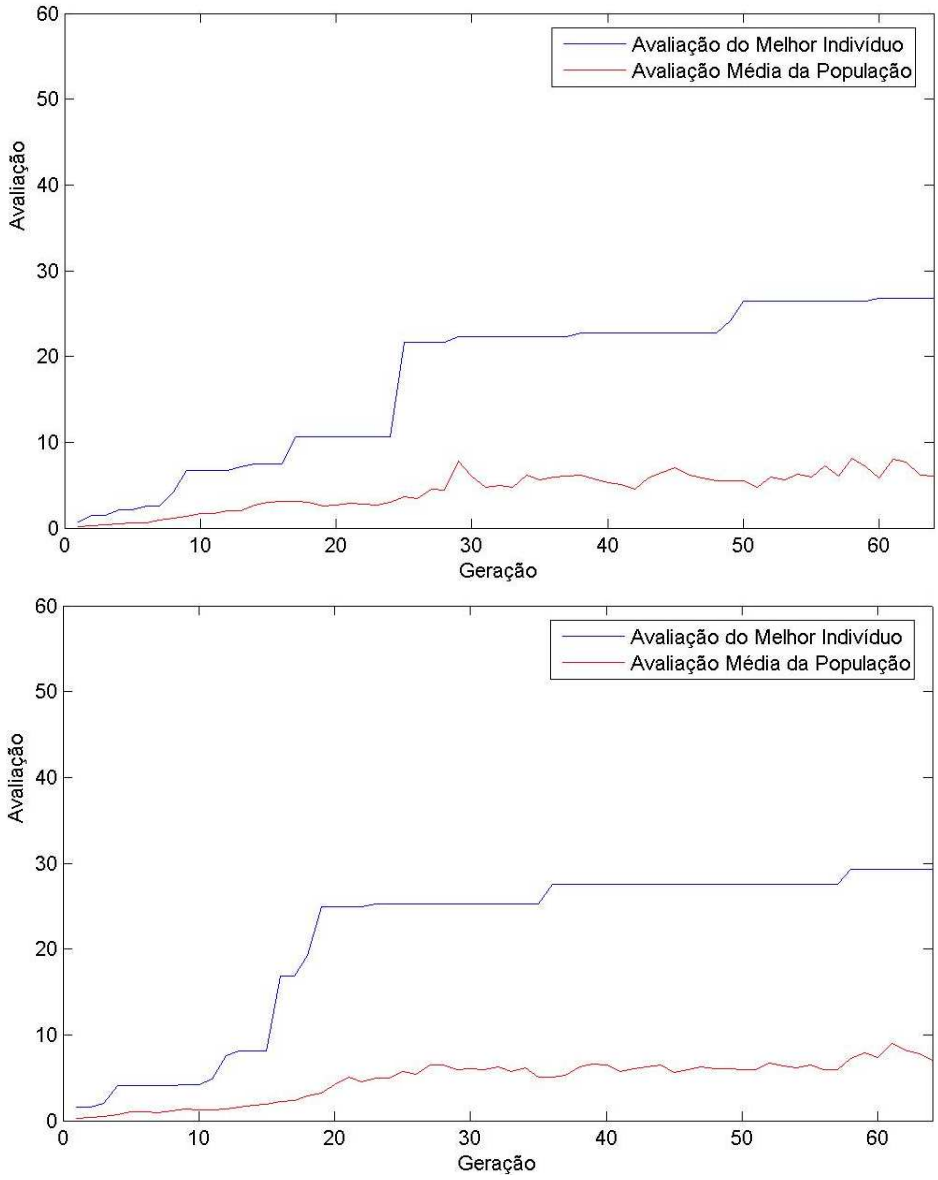


Figura 6.3: Comparação entre os testes 2 e 5, com probabilidade de recombinação $p_r = 0,7$, para a estratégia 2. No primeiro gráfico, correspondente ao teste 2 (com população de 64 indivíduos), podemos ver que a avaliação do melhor indivíduo e a média da avaliação é inferior a do segundo gráfico, correspondente ao teste 5 (com população de 128 indivíduos).

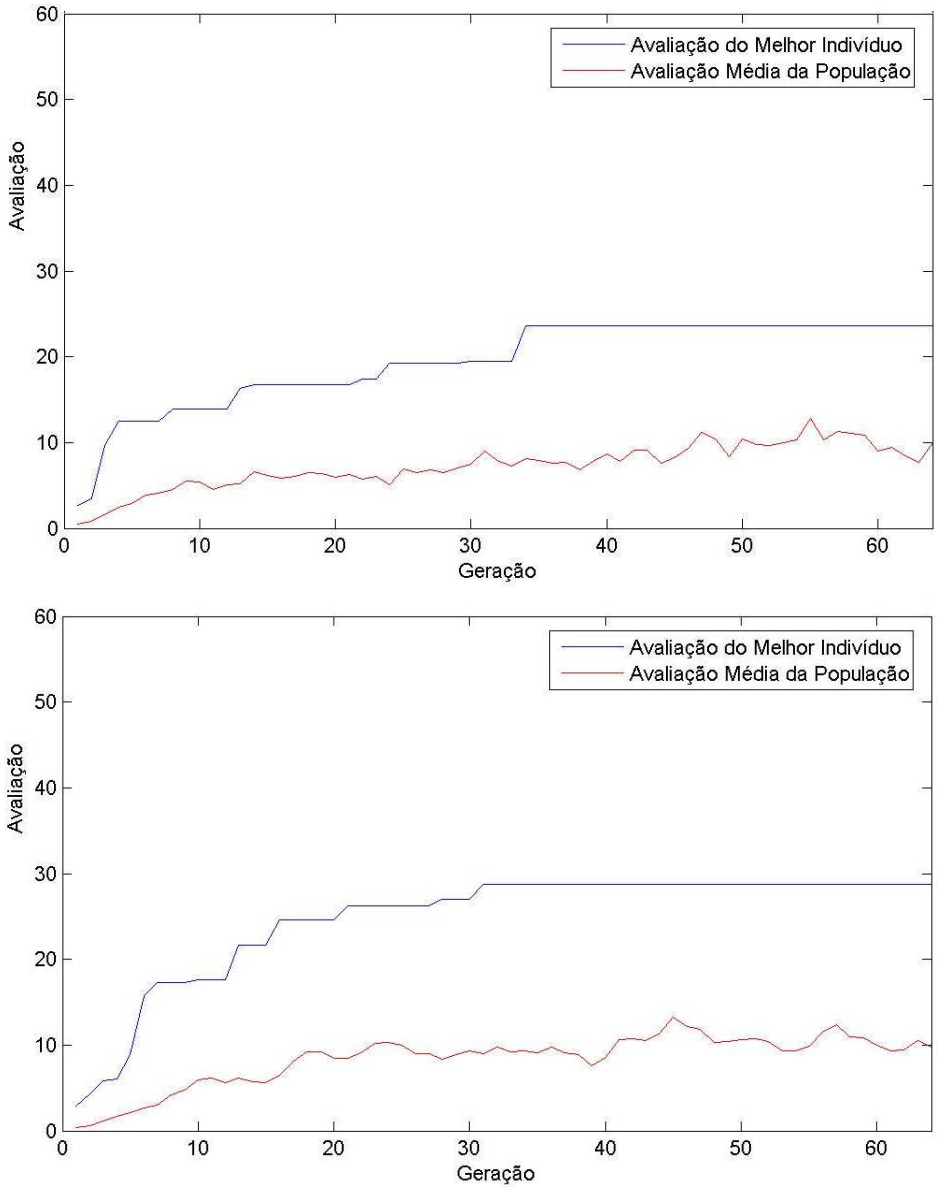


Figura 6.4: Comparação entre os testes 3 e 6, com probabilidade de recombinação $p_r = 0,6$, para a estratégia 3. No primeiro gráfico, correspondente ao teste 3 (com população de 64 indivíduos), podemos ver que a avaliação do melhor indivíduo e a média da avaliação é inferior a do segundo gráfico, correspondente ao teste 6 (com população de 128 indivíduos).

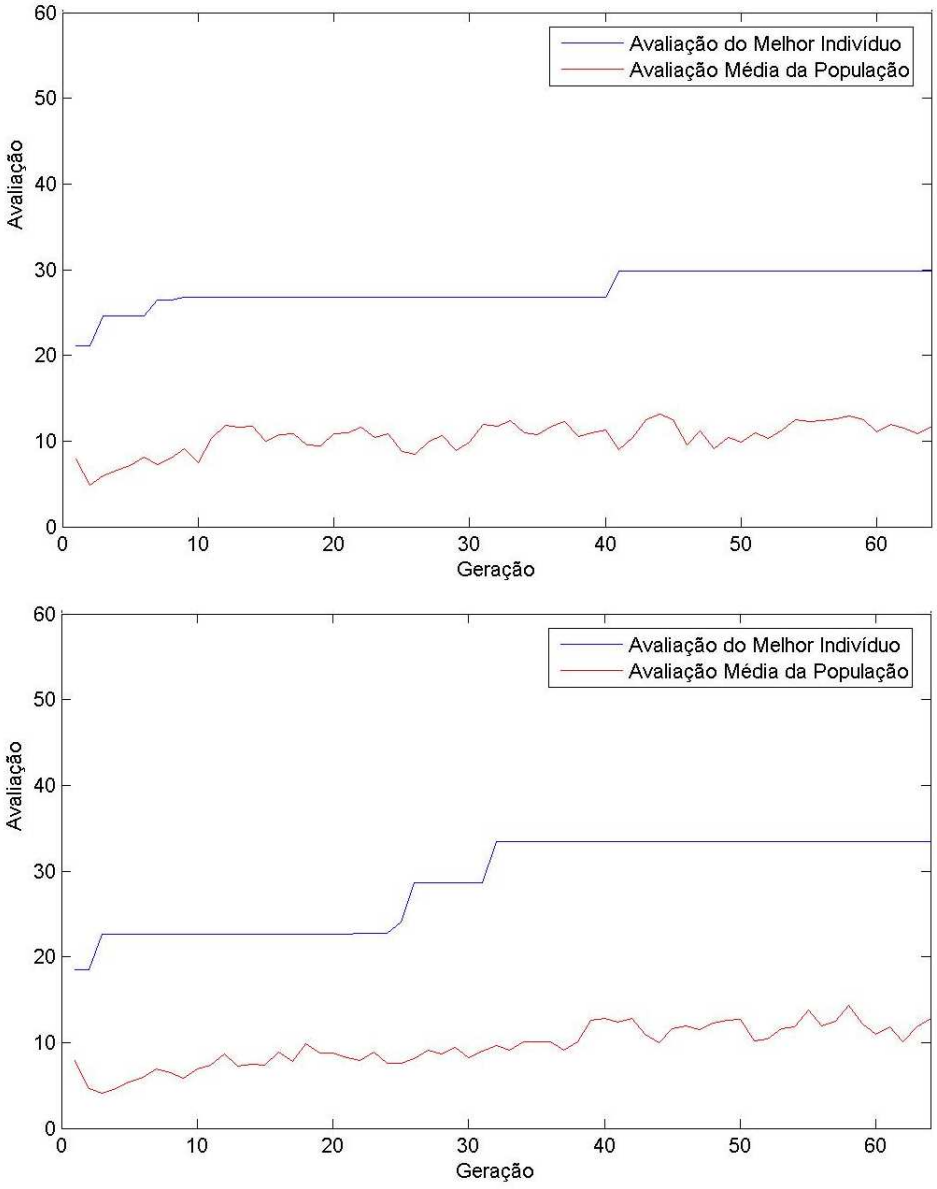


Figura 6.5: Comparação entre os testes 3 e 6, com probabilidade de recombinação $p_r = 0,6$, para a estratégia 4. No primeiro gráfico, correspondente ao teste 3 (com população de 64 indivíduos), podemos ver que a avaliação do melhor indivíduo e a média da avaliação é inferior a do segundo gráfico, correspondente ao teste 6 (com população de 128 indivíduos).

6.1.3 Influência da Variação da Probabilidade de Recombinação

Devido ao tipo de espaço de busca do problema tratado nesse trabalho os testes nos quais a probabilidade de recombinação era mais baixa aliada à uma população de tamanho maior, obtiveram resultados melhores do que os com a probabilidade mais alta. Quando a recombinação ocorre com mais freqüência, características desejáveis de alguns indivíduos da população são perdidas, mesmo com o uso do elitismo, tornando a busca menos eficiente. Por outro lado, para os testes com uma população maior os testes com probabilidade de recombinação mais alta foram mais bem sucedidos. Nesse caso, a probabilidade de recombinação possibilita que a população se torne mais variada, fazendo uma busca melhor.

Na figura 6.6 ilustramos tal efeito para a estratégia 4 utilizando probabilidade de recombinação $p_r = 0,8$, $p_r = 0,7$ e $p_r = 0,6$.

6.1.4 Influência da Estratégia Utilizada

Após todos os testes terem sido realizados, observou-se que a segunda estratégia foi a que teve o pior desempenho nos testes. Isso se deveu ao fato de que boa parte dos fatores e posições que eram codificados no indivíduo não eram válidos, tornando assim a busca ineficaz. Em oposição à segunda estratégia, a quarta estratégia foi a melhor, por resolver o problema do excesso de fatores e posições inválidas geradas. Na figura 6.7 ilustramos a diferença entre os gráficos para as quatro estratégias.

Na tabela 6.2 mostramos os melhores indivíduos gerados pela estratégia 4, para os seis testes realizados. É importante notar que alguns fatores foram representados como *vazio* pois o fator codificado no indivíduo não era válido. Isso pode acontecer devido à recombinação ou à mutação, que podem modificar o valor que antes era

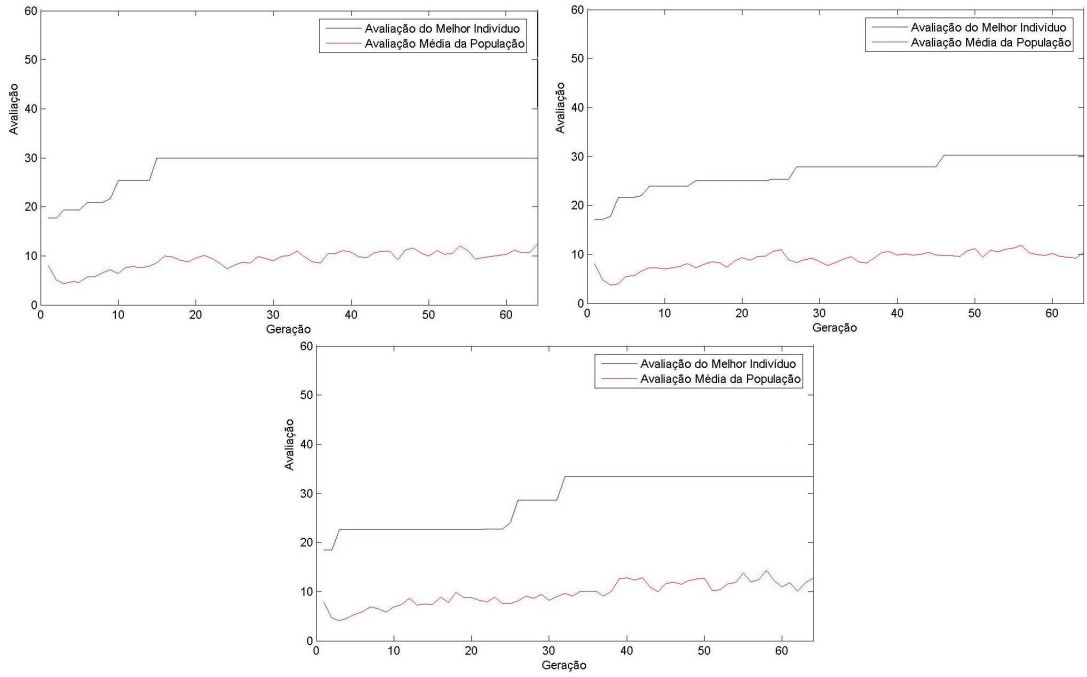


Figura 6.6: Comparação dos gráficos para os testes 4, 5 e 6 da estratégia 4, respectivamente. O teste que utiliza uma probabilidade de recombinação menor é ligeiramente superior aos anteriores.

válido para um inválido.

Tabela 6.2: Melhores indivíduos gerados pela abordagem 4

Abordagem 4		
Teste	Melhor indivíduo	Melhor indivíduo (posição em relação ao promotor)
1	[CaiF 274] [ArgR 141] [CaiF 217] [CRP 97]	[CaiF -84] [ArgR -217] [CaiF -141] [CRP -261]
2	[CaiF 274] [IHF 16] [CRP 378] [CaiF 217]	[CaiF -84] [IHF -342] [CRP 20] [CaiF -141]
3	[CaiF 217] [MarA 183] [NarL 332] [CaiF 274]	[CaiF -141] [MarA -175] [NarL -26] [CaiF -84]
4	[vazio 320] [CaiF 217] [CaiF 217] [CaiF 274]	[vazio -38] [CaiF -141] [CaiF -141] [CaiF -84]
5	[CaiF 274] [vazio 367] [CaiF 217] [CaiF 217]	[CaiF -84] [vazio 9] [CaiF -141] [CaiF -141]
6	[vazio 185] [CaiF 274] [CaiF 217] [CaiF 274]	[vazio -173] [CaiF -84] [CaiF -141] [CaiF -84]

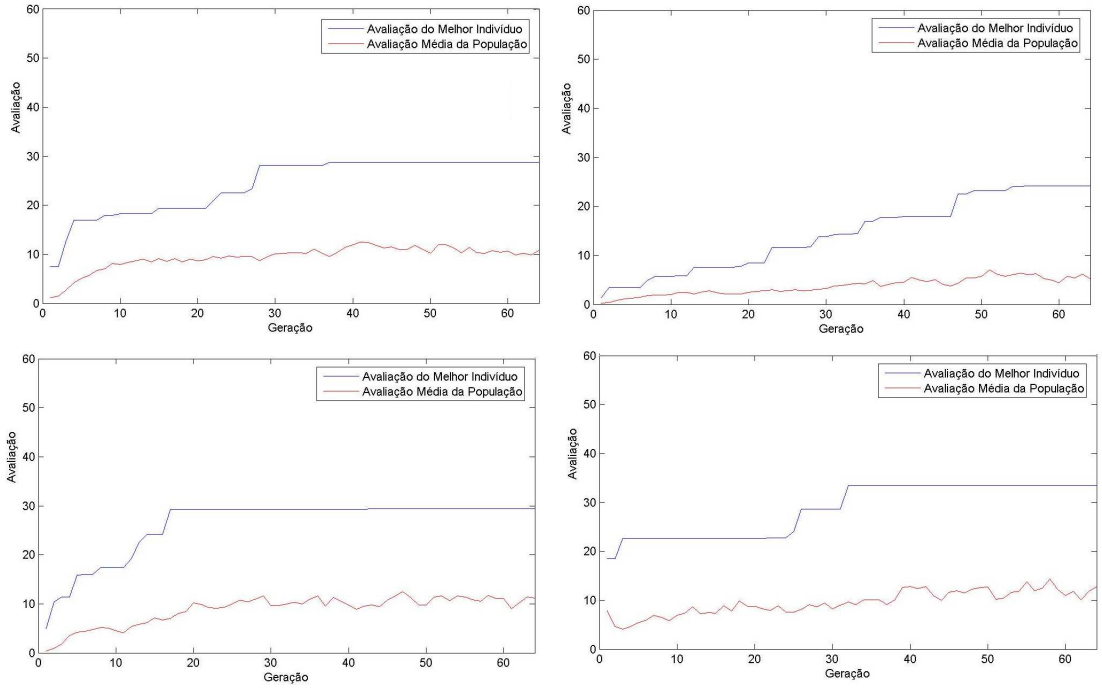


Figura 6.7: Comparação dos gráficos para as estratégias 1, 2, 3 e 4, no sentido horário. Pelos valores da função de avaliação podemos ver que a segunda estratégia foi a que obteve o pior desempenho, enquanto a quarta foi a melhor.

6.2 Análise Preliminar das Características dos Indivíduos Gerados

Segundo (COLLADO-VIDES, 1993a) é necessária a existência de um fator de transcrição no sítio proximal e não deve ocorrer sobreposição entre os fatores de transcrição. Nessa seção analisamos preliminarmente os resultados de acordo com as características levantadas por Collado-Vides, verificando se existem indivíduos que têm sobreposição entre fatores ou que não possuem pelo menos um fator no sítio proximal que no caso da seqüência utilizada vai da posição 298 até a posição 378.

A existência de fatores no sítio proximal não é garantida, mas ocorre em alguns indivíduos da população. Na tabela 6.3, mostramos a porcentagem de indivíduos

que atendem os requisitos descritos acima para os resultados obtidos pela abordagem quatro com elitismo.

Para um determinado fator estar no sítio proximal é necessário que o nucleotídeo central de sua seqüência esteja dentro do intervalo mencionado acima. Por exemplo, o indivíduo *CaiF* 274 *Fur* 23 *CaiF* 217 *GadX* 343, não possui sobreposição e possui um fator no sítio proximal, o *GadX* (que possui tamanho 16, ver tabela 2.1, assim seu nucleotídeo central está na posição 350) na posição 343. Na figura 6.8 podemos ver a correspondência entre as posições como geradas pelo PATSER (levando em consideração que o início da seqüência é no nucleotídeo mais à esquerda) e as posições levando em consideração o início da seqüência sendo na posição +1.

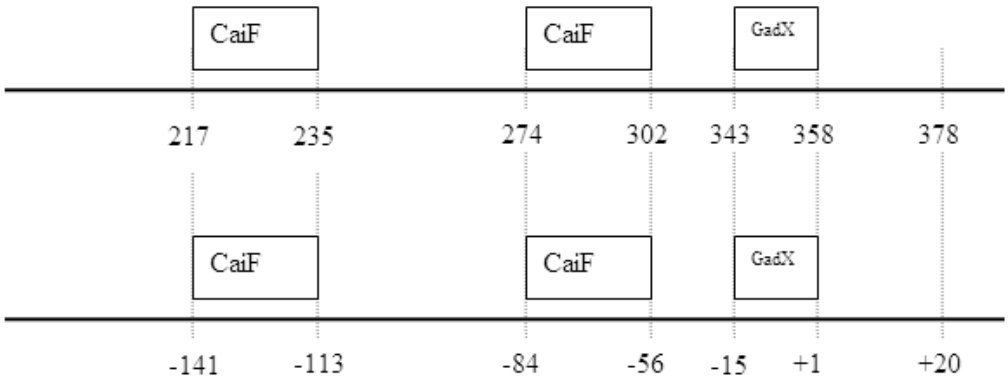


Figura 6.8: Representação gráfica parcial do indivíduo apresentado em 6.2

Tabela 6.3: Estatísticas dos indivíduos gerados pela abordagem 4 (com elitismo)

Abordagem 4 com elitismo				
Teste	Número de indivíduos na população final	Indivíduos com sítio proximal	Indivíduos sem sobreposição	Indivíduos com sítio proximal e sem sobreposição
1	64	9,37%	70,31%	6,25%
2	64	3,12%	92,18%	3,12%
3	64	21,87%	100,00%	21,87%
4	128	41,40%	94,53%	39,06%
5	128	17,96%	91,40%	16,40%
6	128	22,65%	87,50%	18,75%

6.3 Análise Comparativa com os Bancos de Dados

De posse das posições previstas pelo RegulonDB e pelo TractorDB (exibidas nas tabelas 6.4 e 6.5), realizamos análises comparativas com os resultados obtidos pelo AG. As posições do RegulonDB foram previstas utilizando quatro métodos diferentes, GEA (análise de expressão gênica), HIBSCS (inferência humana baseada em similaridade a seqüências CONSENSUS), BPP (ligação de proteínas purificadas) e AIBSCS (inferência automatizada baseada em similaridade com seqüências CONSENSUS), divididos pela força da evidência. As evidências consideradas fortes são aquelas provenientes de dados experimentais, que dão alto nível de certeza da existência do fator. Todas as outras são consideradas fracas.

6.3.1 Análise Quantitativa

A primeira análise realizada, chamada de quantitativa, tinha por objetivo verificar se os fatores de transcrição apontados pelos bancos de dados eram os mais freqüentes na população do AG.

Tabela 6.4: Posições previstas pelo RegulonDB para os fatores CRP, CaiF e FNR.

Fator de Transcrição	Posição	Tipos de Evidência	Força da Evidência
CRP	-137	GEA e HIBSCS	Fracas
CRP	-80	GEA e HIBSCS	Fracas
CaiF	-144	BPP e GEA	Forte e Fraca
CaiF	-84	BPP e GEA	Forte e Fraca
FNR	-204	AIBSCS E GEA	Fracas

Tabela 6.5: Posições previstas pelo TractorDB para os fatores CRP, CaiF e FNR

Fator de Transcrição	Posição	Tipo de Evidência
CRP	-280	Modelo Estatístico
CRP	-279	Modelo Estatístico
CRP	-178	Expressões Regulares
CRP	-160	Modelo Estatístico
CRP	-121	Expressões Regulares
CRP	-103	Modelo Estatístico
FNR	-278	Modelo Estatístico
FNR	-229	Modelo Estatístico
CaiF	-192	Modelo Estatístico

Para calcular quantas vezes cada fator de transcrição aparecia na população do AG foram percorridos todos os 60 arquivos da estratégia 4 (correspondentes a 10 rodadas para cada um dos seis testes propostos na tabela 6.1) gerados para o caso em que o algoritmo utilizava elitismo. Nos casos em que um fator aparecia mais de uma vez no mesmo indivíduo do AG, todas as ocorrências eram consideradas, pois é natural a existência de um ou mais fatores iguais na mesma seqüência regulatória.

Pode-se ver na tabela 6.6 que o AG consegue apontar como fatores de transcrição mais freqüentes na população os mesmos apontados pelo TractorDB e RegulonDB, estando entre os seis primeiros fatores mais freqüentes na população. A tabela para os demais fatores de transcrição utilizados pode ser encontrada no Apêndice B.

O fator de transcrição representado como *vazio* denota que alguns dos indivíduos encontrados pelo AG não possuem o número máximo de fatores possíveis para uma UIG (4, no caso da estratégia analisada).

Tabela 6.6: Resultados da análise quantitativa para quarta abordagem I

Fator de transcrição	No. de FT encontrados	% na população
CaiF	**7517**	**32,29%**
Vazio	4944	21,24%
CRP	**1018**	**4,37%**
IHF	809	3,48%
NarL	520	2,23%
FNR	**513**	**2,20%**

6.3.2 Análise qualitativa

Após a análise quantitativa, nosso objetivo era verificar se o AG acertava não somente os fatores de transcrição apontados pelos bancos de dados, mas também se acertava a posição correta. De modo semelhante à análise anterior, também foram percorridos todos os 60 arquivos para a estratégia 4 com elitismo, totalizando 23280 posições possíveis de serem preditas.

Além da contagem de predições do fator para as 450 posições da região regulatória realizamos uma análise posterior para até 8 pares de base à esquerda e à direita da posição exata indicada pelos bancos de dados (ver tabelas 6.4 e 6.5), pois segundo (GUÍA et al., 2005) dentro do DNA bacteriano a posição do fator de transcrição pode variar dentro desse intervalo de 16 pares de base.

Nos gráficos 6.9, 6.10 e 6.11 mostramos os resultados obtidos para a varredura completa para todas as 450 posições da região regulatória. Para os três fatores de transcrição preditos pelos bancos, as melhores respostas do algoritmo genético dis-

tam no máximo de 5 pares de base (ver tabelas 6.7, 6.8 e 6.9) e, para os fatores CaiF e FNR, houveram resultados que acertavam a posição exata predita. O fator CaiF foi o que obteve a predição mais precisa, tendo a posição -84 sido prevista 5976 vezes pelo AG.

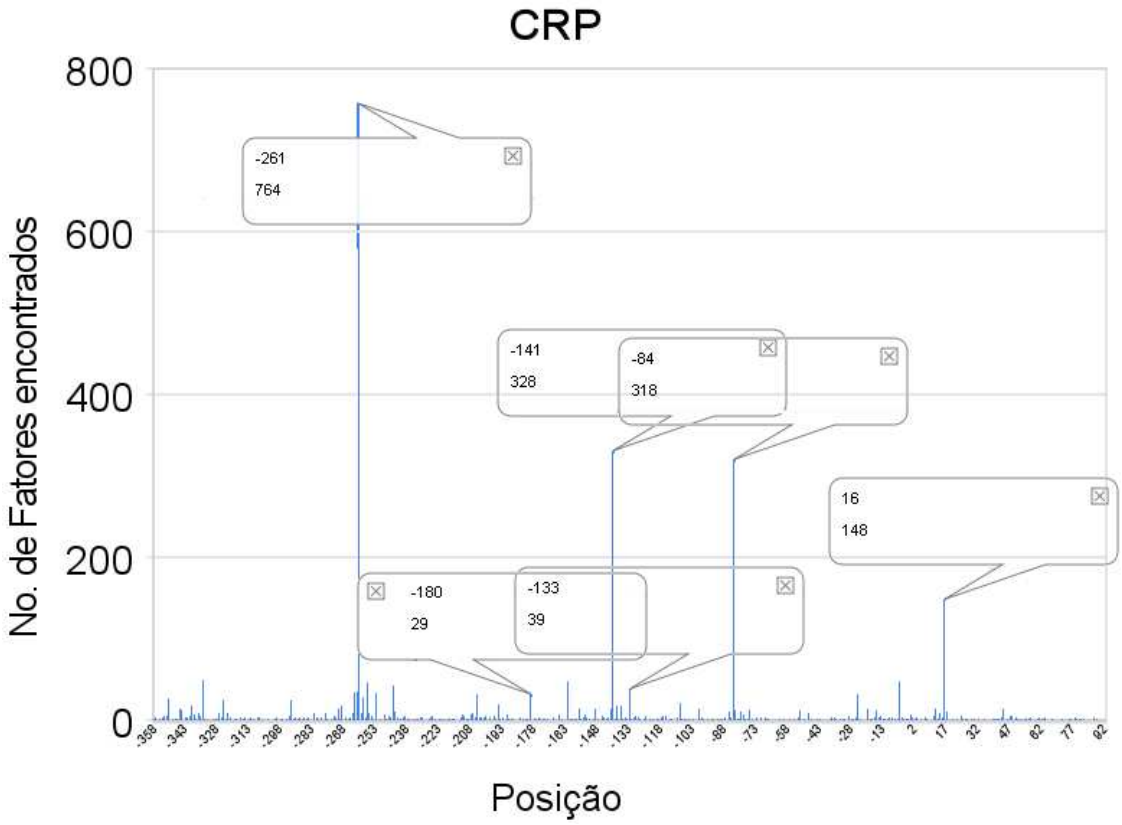


Figura 6.9: Resultados da análise qualitativa para o fator de transcrição CRP. Os balões no gráfico apontam a posição e o número de vezes que ela foi predita. Além das posições mostradas na tabela 6.7, destacamos as posições -261 e 16 por terem apresentado bons resultados

Para o fator de transcrição CRP ainda destacamos (figura 6.9) as posições -261 que foi prevista 764 vezes e a posição 16, prevista 148 vezes. O fato da primeira posição ter sido a mais prevista para esse FT pode indicar que esse seria um possível sítio de CRP.

Tabela 6.7: Melhores resultados da análise qualitativa para o fator de transcrição CRP

Posição	Quantidade encontrada	Posição no BD	Distância	BD
-180	29	-178	2	TractorDB
-162	48	-160	2	TractorDB
-133	37	-137	4	RegulonDB
-141	328	-144	3	RegulonDB
-84	315	-80	4	RegulonDB

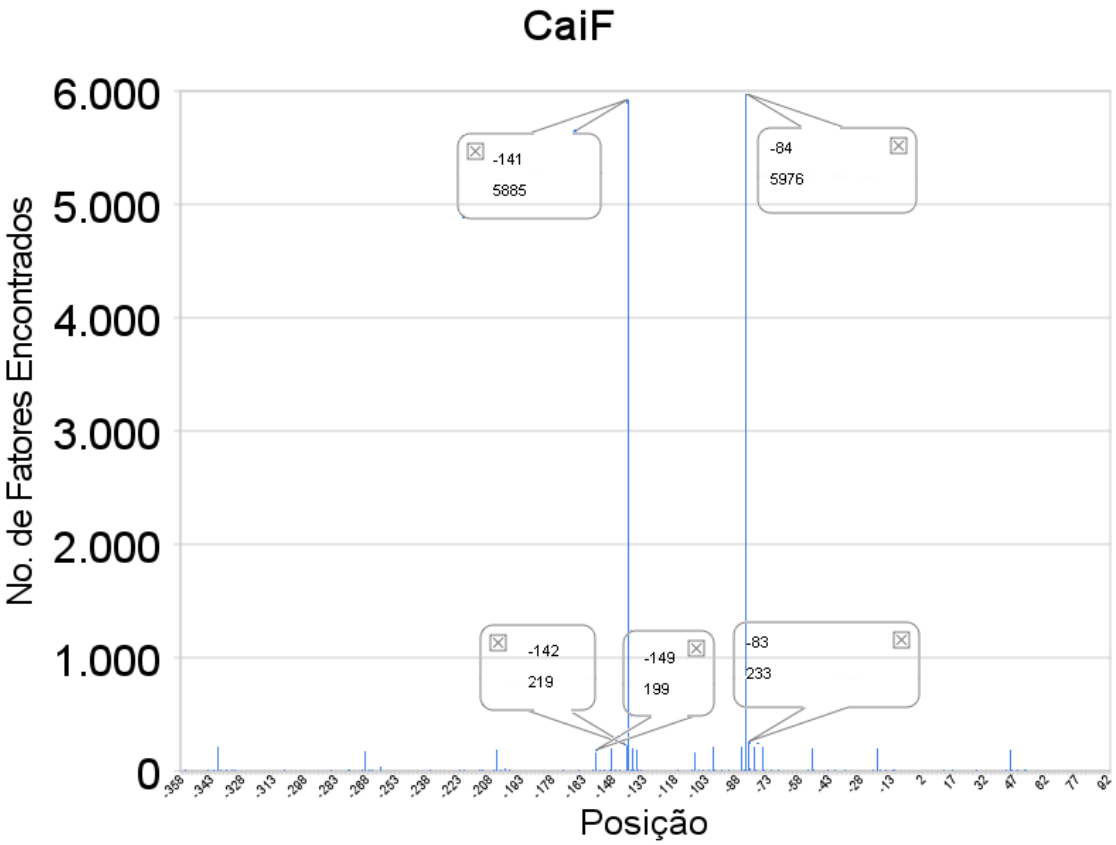


Figura 6.10: Resultados da análise qualitativa para o fator de transcrição CaiF

Para o fator de transcrição *CaiF* nenhuma posição além das já destacadas na figura 6.10 pode ser apontada com um possível sítio de ligação. Contudo, os re-

Tabela 6.8: Melhores resultados da análise qualitativa para o fator de transcrição CaiF.

Posição	Qtde. encontrada	Posição no BD	Distância	Banco de Dados
-142	219	-144	2	RegulonDB
-141	5885	-144	3	RegulonDB
-149	197	-144	5	RegulonDB
-83	233	-84	1	RegulonDB
-80	213	-84	4	RegulonDB
-84	5976	-84	0	RegulonDB

sultados para esse FT foram os mais promissores, sendo a primeira posição mais predita coincidente com a indicada pelo RegulonDB (predição realizada utilizando o método BPP, considerado como uma forte evidência da existência do sítio) e a segunda distando apenas três pares de base. Tais resultados indicam que a metodologia proposta por esse trabalho está obtendo resultados conhecidos ou próximo aos conhecidos.

Tabela 6.9: Melhores resultados da análise qualitativa para o fator de transcrição FNR

Posição	Quantidade encontrada	Posição no BD	Distância	Banco de Dados
-278	20	-278	0	TractorDB
-274	13	-278	4	TractorDB
-201	19	-204	3	RegulonDB
-199	81	-204	5	RegulonDB
-205	57	-204	1	RegulonDB

Para o fator de transcrição FNR ainda destacamos (figura 6.11) a posição -141 que foi prevista 216 vezes e a posição -84, prevista 196 vezes. Tal qual ocorreu para o fator de transcrição CRP, o fato da primeira posição ter sido a mais prevista para esse FT pode indicar que esse seria um possível sítio de FNR.

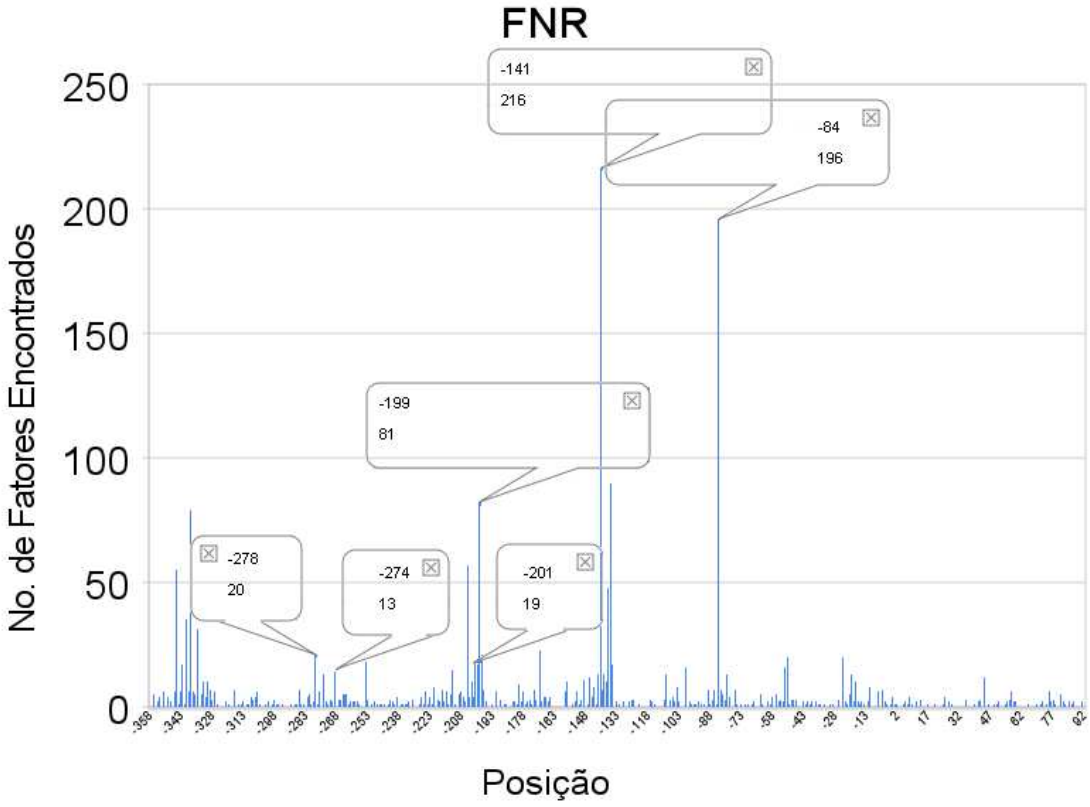


Figura 6.11: Resultados da análise qualitativa para o fator de transcrição FNR.Os balões no gráfico apontam a posição e o número de vezes que ela foi predita.

6.4 Considerações

Nesse capítulo apresentamos os resultados obtidos pela metodologia apresentada em 5 e os analisamos sob alguns pontos de vista.

Na subseção 6.1.1 mostramos que o uso do elitismo para o problema tratado nesse trabalho é essencial para a obtenção de bons resultados. Nas subseções 6.1.2 e 6.1.3 mostramos a influência do tamanho da população e a taxa de recombinação utilizada no algoritmo genético, bem como a interação entre esses dois parâmetros

do AG. Como esperado, os testes que utilizavam uma população maior foram mais bem sucedidos em relação aos de população menor. Finalizando a primeira sessão de análises foi feita, na subseção 6.1.4 uma comparação entre as quatro estratégias propostas, na qual a quarta (subseção 5.2.4) foi a mais bem sucedida.

Na seção 6.2 fizemos uma breve análise dos resultados do ponto de vista das características do indivíduos e na seção 6.3 comparamos os resultados do presente trabalho com os dados encontrados nos bancos de dados RegulonDB (seção 4.4) e TRACTOR_DB (seção 4.3). Tal comparação forneceu indícios de que os resultados obtidos por esse trabalho estão corretos e podem indicar bons candidatos a UIG.

No capítulo 7 apresentaremos as conclusões desse trabalho e proporemos caminhos a seguir em trabalhos futuros.

7 CONCLUSÕES E TRABALHOS FUTUROS

A proposta do presente trabalho era obter combinações de fatores de transcrição que fossem bons candidatos a unidades de informação genética (UIG). Tínhamos como hipótese de que tal objetivo seria possível de ser alcançado utilizando algoritmos genéticos com uma codificação e operadores adequados.

Para a geração de tais candidatos à UIG utilizamos os dados fornecidos pelo programa PATSER (HERTZ; STORMO, 1999), que indicam qual a probabilidade de um determinado fator de transcrição ligar-se a uma posição do DNA, como parte da função de avaliação do algoritmo genético. Para evitar direcionar a busca o algoritmo genético não possuía nenhum tipo de restrição (para obrigar a existência de indivíduo no sítio proximal ou evitar a sobreposição de fatores), sendo utilizados somente os operadores simples e uma função de avaliação adequada ao problema.

Apresentamos, então, os bancos de dados RegulonDB (SALGADO et al., 2006) e TRACTOR_DB (GONZÁLEZ et al., 2005) e seus métodos de inferência de posicionamento de fatores de transcrição no DNA e utilizamos seus dados para avaliar a qualidade dos indivíduos gerados pelo AG.

Uma dificuldade encontrada durante a proposição das estratégias para resolver o problema surgiu de uma particularidade do mesmo. O número de posições do DNA que apresentavam pontuação positiva atribuida pelo PATSER era muito baixo, em contraposição às 450 escolhas possíveis e isso tornava o espaço de busca muito complexo. Somente na quarta estratégia proposta é que tal característica do problema foi contornada, utilizando-se um mapeamento dinâmico para as posições cuja pontuação associada era positiva.

É importante ressaltar que a metodologia proposta por esse trabalho pode ser aplicada a outros organismos procariotos. Escolhemos a *Escherichia coli K12* como estudo de caso por este ser um organismo amplamente conhecido e estudado, facilitando a obtenção de dados para teste.

Devido aos resultados apresentados no capítulo 6 acreditamos que o objetivo do trabalho foi cumprido. A metodologia aqui proposta foi capaz de encontrar com um nível razoável de confiabilidade os resultados já conhecidos, quando comparados aos bancos de dados supracitados.

7.1 Sugestões para Trabalhos Futuros

Ao longo do trabalho aqui apresentado algumas idéias surgiram, porém optou-se por restringir o escopo do trabalho, resultando em algumas sugestões para trabalhos futuros. São elas:

7.1.1 Proposição de Novos Operadores de Recombinação e Mutação

Um dos problemas notados durante a análise dos resultados desse trabalho foi o poder disruptivo dos operadores de recombinação e mutação. Essa questão está pre-

sente em diversos trabalhos que utilizam algoritmos genéticos(DE JONG; SPEARS, 1990), contudo, dado o tipo de espaço de busca do problema da regulação gênica é importante refinar esses operadores, a fim de reduzir seus efeitos disruptivos.

7.1.2 Criação de uma Nova Função de Avaliação

No presente trabalho optamos por não fazer restrições ao algoritmo genético porém, é necessária a proposição de uma função de avaliação que leve em consideração a sobreposição entre fatores de transcrição. Uma possibilidade seria penalizar o indivíduo que tivesse sobreposição, de acordo com o número de fatores de transcrição com os quais isso ocorre. Uma solução proposta por (LINDEN, 2006) é transformar a função de avaliação $f(x)$ em

$$f'(x_{inad}) = f(x_{inad}) - Q(x_{inad})$$

onde $Q(x_{inad})$ é a penalidade associada ao descumprimento das condições. Uma outra possível solução é tentar reparar o indivíduo que não obedece à restrição, subtraindo do valor de sua avaliação o custo do reparo.

7.1.3 Elitismo na Presença de Fator de Transcrição no Sítio Proximal

Para respeitar a obrigatoriedade da presença de um fator de transcrição no sítio proximal uma sugestão para trabalhos futuros é gerar uma certa porcentagem dos indivíduos da população inicial que contenham um FT. Porém, também é necessário preservar esses indivíduos e para tal pode-se utilizar um operador de elitismo.

7.1.4 Extensão da Metodologia Proposta para Casos mais Complexos

De posse das demais seqüências regulatórias de 450 pares de base de outros operon da *E.coli K12* seria possível estender o presente trabalho para os demais casos e realizar um estudo para toda a rede de regulação da *E.coli*.

REFERÊNCIAS

ALBERTS, B.; BRAY, D.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; P.WALTERS. **Fundamentos da Biologia Celular**. 1.ed. [S.l.]: ArtMed Editora, 1999.

BIOLOGY. [S.l.]: Wikipedia, 2007.

COLLADO-VIDES, J. The Search of a Grammatical Theory of Gene Regulation is Formally Justified by Showing the Inadequacy of Context-free Grammars. **CABIOS**, Oxford, v.7, n.3, p.321 – 326, 1991.

COLLADO-VIDES, J. A linguistic representation of the regulation of transcription initiation. I. An ordered array of complex symbols with distinctive features. **BioSystems**, Irlanda, p.87 – 104, 1993.

COLLADO-VIDES, J. A linguistic representation of the regulation of transcription initiation. II. Distinctive features of sigma 70 promoters and their regulatory binding sites. **BioSystems**, Irlanda, n.29, p.105 – 128, 1993.

COLLOMB, D. **RiboNucleic Acids**. [S.l.: s.n.], 2007.

DE JONG, K. A.; SPEARS, W. M. An Analysis of the Interacting Roles of Population Size and Crossover in Genetic Algorithms. In: FIRST WORKSHOP PARALLEL PROBLEM SOLVING FROM NATURE, 1990, Berlim. **Proceedings...** Springer-Verlag, 1990. p.38 – 47.

FRIEDL, J. E. F. **Mastering Regular Expressions**. 2.ed. Sebastopol, CA: O'Reilly, 1997.

GALAS, D. J.; SCHMITZ, A. DNAse footprinting: a simple method for the detection of protein-dna binding specificity. **Nucleic Acids Res.**, Oxford, v.5, n.9, p.3157 – 3170, 1978.

GAMA-CASTRO, S.; JIMÉNEZ-JACINTO, V.; PERALTA-GIL, M.; SANTOS-ZAVALETA, A.; PEÑALOZA-SPINOLA, M. I.; CONTRERAS-MOREIRA, B.; SEGURA-SALAZAR, J.; MUÑIZ-RASCADO, L.; MARTÍNEZ-FLORES, I.; SALGADO, H.; BONAVIDES-MARTÍNEZ, C.; ABREU-GOODGER, C.; RODRÍGUEZ-PENAGOS, C.; MIRANDA-RÍOS, J.; MORETT, E.; MERINO, E.; HUERTA, A. M.; TREVIÑO-QUINTANILLA, L.; COLLADO-VIDES, J. RegulonDB (version 6.0): gene regulation model of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation.**Nucleic Acids Research**, Oxford, v.36, n.1, p.120–124, 2008.

GARNER, M. M.; REVZIN, A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the escherichia coli lactose operon regulatory system. **Nucleic Acid Research**, [S.l.], v.9, n.13, p.56–65, 1981.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization and learning**. Massachusetts: Addison-Wesley, 1989.

GONZÁLEZ, A. D.; ESPINOSA, V.; VASCONCELOS, A. T.; PÉREZ-RUEDA, E.; COLLADO-VIDES, J. TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes. **Nucleic Acids Research**, [S.l.], v.33, p.98–102, 2005.

GUÍA, M. H.; PÉREZ, A. G.; ANGARICA, V. E.; VASCONCELOS, A. T.; COLLADO-VIDES, J. Complementing computationally predicted regulatory sites

in Tractor_DB using a pattern matching approach. **In Silico Biology**, [S.l.], v.5, n.2, p.209–219, 2005.

HERTZ, G. Z.; HARTZELL, G. W.; STORMO, G. D. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. **CABIOS**, [S.l.], v.6, n.2, p.81–92, 1990.

HERTZ, G. Z.; STORMO, G. D. Identifying DNA and protein patterns with statiscally significant alignments of multiple sequences. **BIOINFORMATICS**, Oxford, v.15, n.7, p.563–577, 1999.

JACOB, F.; MONOD, J. Genetic regulatory mechanisms in the synthesis of proteins. **Journal of Molecular Biology**, [S.l.], v.3, p.318–356, 1961.

LAWRENCE, C. E.; ALTSCHUL, S. F.; BOGUSKI, M.; LIU, J.; NEUWALD, A.; WOOTTON, J. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. **Science**, [S.l.], v.262, n.5131, p.208–214, 1993.

LINDEN, R. **Algoritmos Genéticos**. 1.ed. Rio de Janeiro: Brasport, 2006.

LODISH, H.; BERK, A.; ZIPURSKY, L.; MATSUDAIRA, P.; BALTIMORE, D.; DARNELL, J. **Molecular Cell Biology**. 4.ed. [S.l.]: W. H. Freedman, 2000.

MAKING the Modern World. [S.l.]: The Science Museum, 2004.

MICHALEWICZ, Z. **Genetic Algorithms + Data Structures = Evolution Programs**. 3.ed. Berlim: Springer, 1996.

MULLIS, K. The Unusual Origin of the Polymerase Chain Reaction. **Scientific American**, [S.l.], v.262, n.4, p.56–65, 1990.

MUNAKATA, T. **Fundamentals of the New Artificial Intelligence** 3.ed. New York: Springer-Verlag, 1998.

NEUWALD, A. F.; LIU, J. S.; LAWRENCE, C. E. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. **Protein Science**, Cambridge, v.4, n.8, p.1618–1632, 1995.

PASSARGE, E. **Color Atlas of Genetics**. 1.ed. New York: Thieme, 1995.

ROSENBLUETH, D. A.; THIEFFRY, D.; HUERTA, A. M.; SALGADO, H.; COLLADO-VIDES, J. Syntactic recognition of regulatory region in Escherichia coli. **CABIOS**, Oxford, v.12, n.5, p.415 – 422, 1996.

RUSSEL, S.; NORVIG, P. **Inteligência Artificial**. 2.ed. Rio de Janeiro: Elsevier, 2004. tradução da segunda edição.

SALGADO, H.; GAMA-CASTRO, S.; PERALTA-GIL, M.; DÍAZ-PEREDO, E.; SÁNCHEZ-SOLANO, F.; SANTOS-ZAVALETA, A.; MARTÍNEZ-FLORES, I.; JIMÉNEZ-JACINTO, V.; BONAVIDES-MARTÍNEZ, C.; SEGURA-SALAZAR, J.; MARTíNEZ-ANTONIO, A.; COLLADO-VIDES, J. RegulonDB (version 5.0): escherichia coli k-12 transcriptional regulatory network, operon organization, and growth conditions. **Nucleic Acids Research**, Oxford, v.34, n.1, p.394–397, 2006.

SILHAVY, T. J.; BECKWITH, J. R. Uses of lac fusions for the study of biological problems. **Microbiological Reviews**, [S.l.], v.49, n.4, p.398–418, 1985.

SOARES, J. L. **Biologia**. 4.ed. São Paulo: Scipione, 1994.

SOARES, J. L. **Biologia**. 5.ed. São Paulo: Scipione, 1995.

TOH, H.; HORIMOTO, K. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. **BIOINFORMATICS**, Oxford, v.18, n.2, p.287–297, 2002.

WANDERLEY, M. F. B. **Aplicação de Algoritmos Genéticos no Problema da Regulação Gênica**. 2007. Monografia de Projeto Final de Curso — Instituto de Matemática, UFRJ, Rio de Janeiro, RJ, Brasil.

WANDERLEY, M. F. B.; SILVA, J. C. P. da; BORGES, C. C. H.; VASCONCELOS, A. T. R. Application of Genetic Algorithms to the Genetic Regulation Problem. In: ADVANCES IN BIOINFORMATICS AND COMPUTATIONAL BIOLOGY: THIRD BRAZILIAN SYMPOSIUM ON BIOINFORMATICS, 2008.**Proceedings...** Springer, 2008. p.140–151.

WHITLEY, D. **A Genetic Algorithm Tutorial**. [S.l.]: Colorado State University, 1993.

ZHANG, F.; KUO, M. D.; BRUNKHORS, A. E. coli Promoter Prediction using Feed-Forward Neural Networks. In: PROCEEDINGS OF THE 28TH IEEE EMBS ANNUAL INTERNATIONAL CONFERENCE, 2006, New York City. **Proceedings...** IEEE Service Center, 2006. p.2025–2027.

ApêndiceA GRÁFICOS COMPARATIVOS

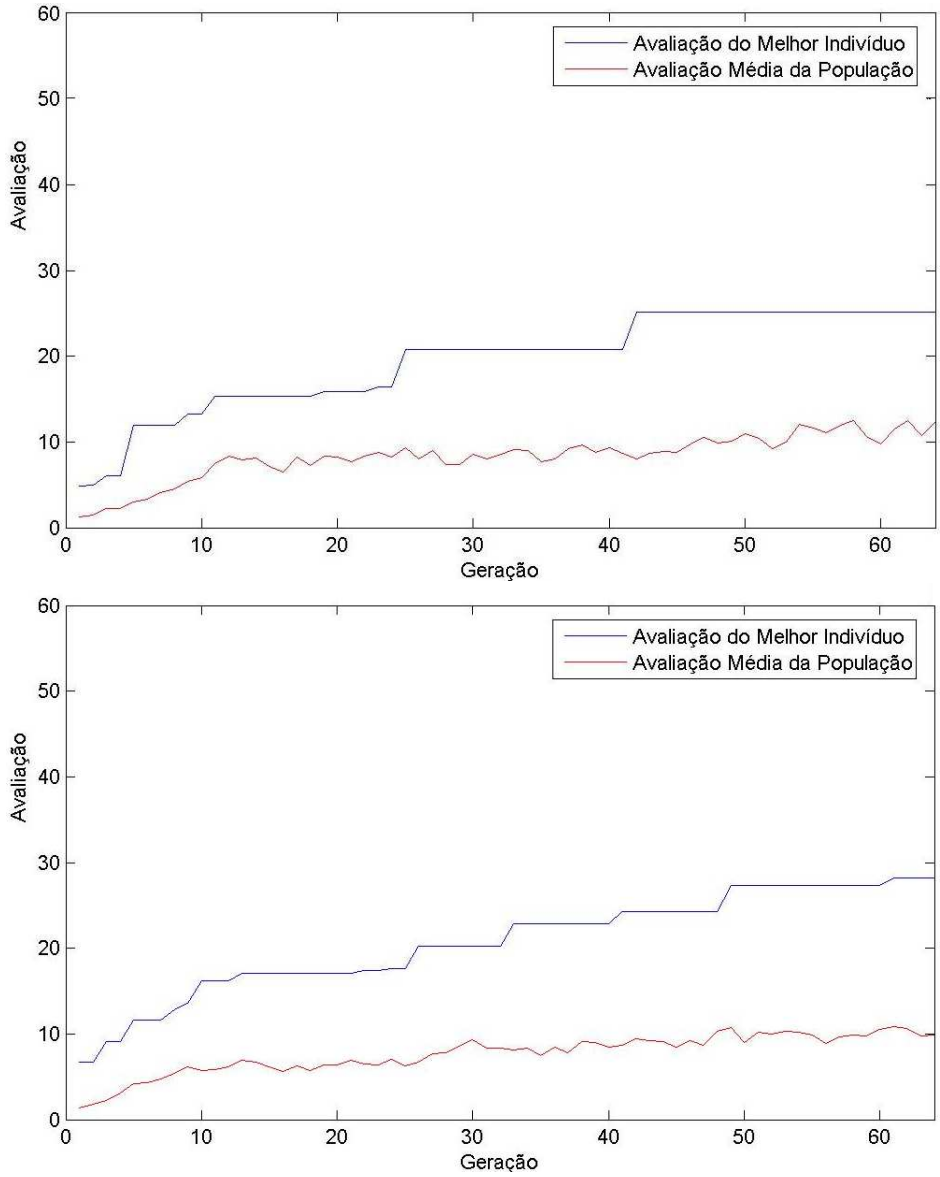


Figura A.1: Comparação entre os testes 2 e 5 para a estratégia 1. No primeiro gráfico, correspondente ao teste 2 (com população de 64 indivíduos), podemos ver que a avaliação do melhor indivíduo e a média da avaliação é inferior a do segundo gráfico, correspondente ao teste 5 (com população de 128 indivíduos).

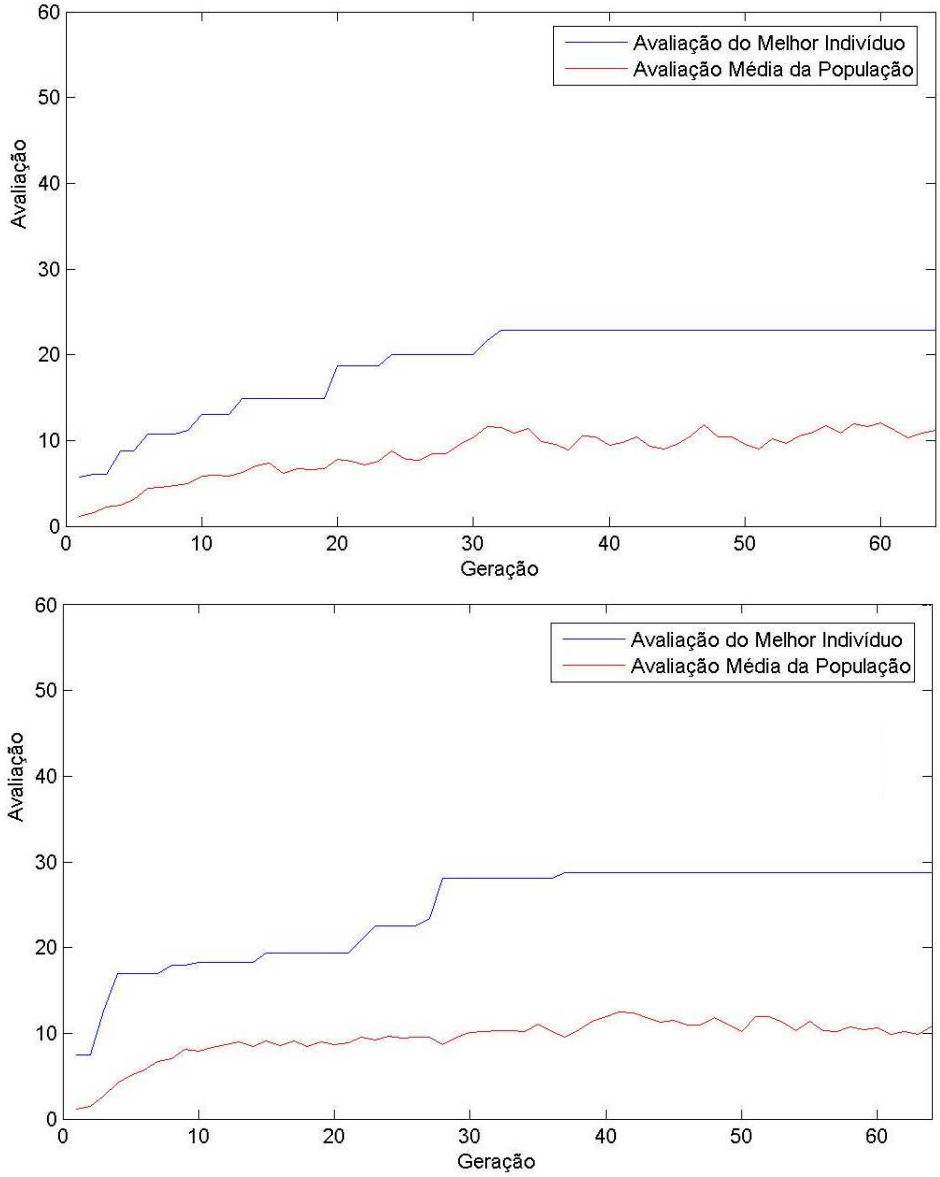


Figura A.2: Comparação entre os testes 3 e 6 para a estratégia 1. No primeiro gráfico, correspondente ao teste 3 (com população de 64 indivíduos), podemos ver que a avaliação do melhor indivíduo e a média da avaliação é inferior a do segundo gráfico, correspondente ao teste 6 (com população de 128 indivíduos).

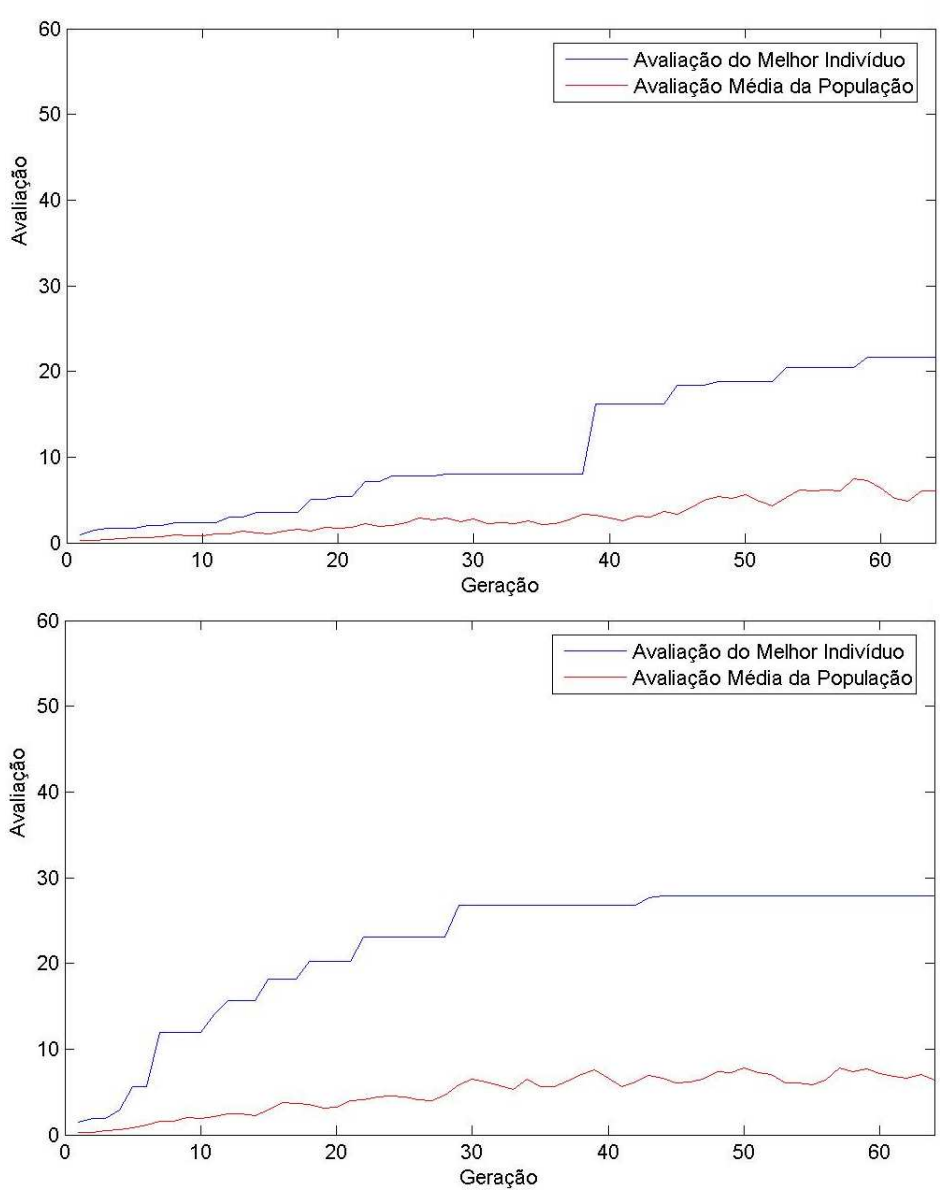


Figura A.3: Comparação entre os testes 1 e 4 para a estratégia 2. No primeiro gráfico, correspondente ao teste 1 (com população de 64 indivíduos), podemos ver que a avaliação do melhor indivíduo e a média da avaliação é inferior a do segundo gráfico, correspondente ao teste 4 (com população de 128 indivíduos).

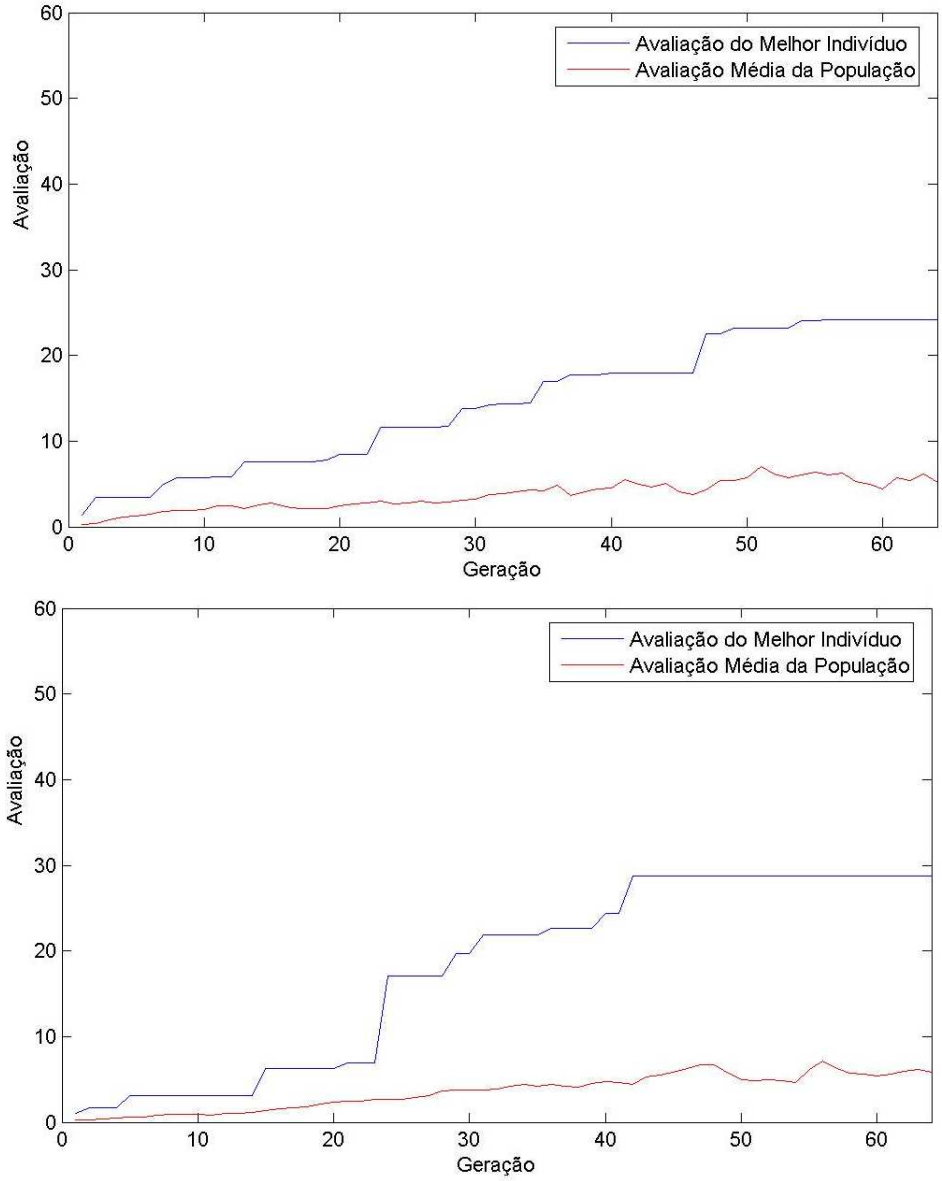


Figura A.4: Comparação entre os testes 3 e 6 para a estratégia 2. No primeiro gráfico, correspondente ao teste 3 (com população de 64 indivíduos), podemos ver que a avaliação do melhor indivíduo e a média da avaliação é inferior a do segundo gráfico, correspondente ao teste 6 (com população de 128 indivíduos).

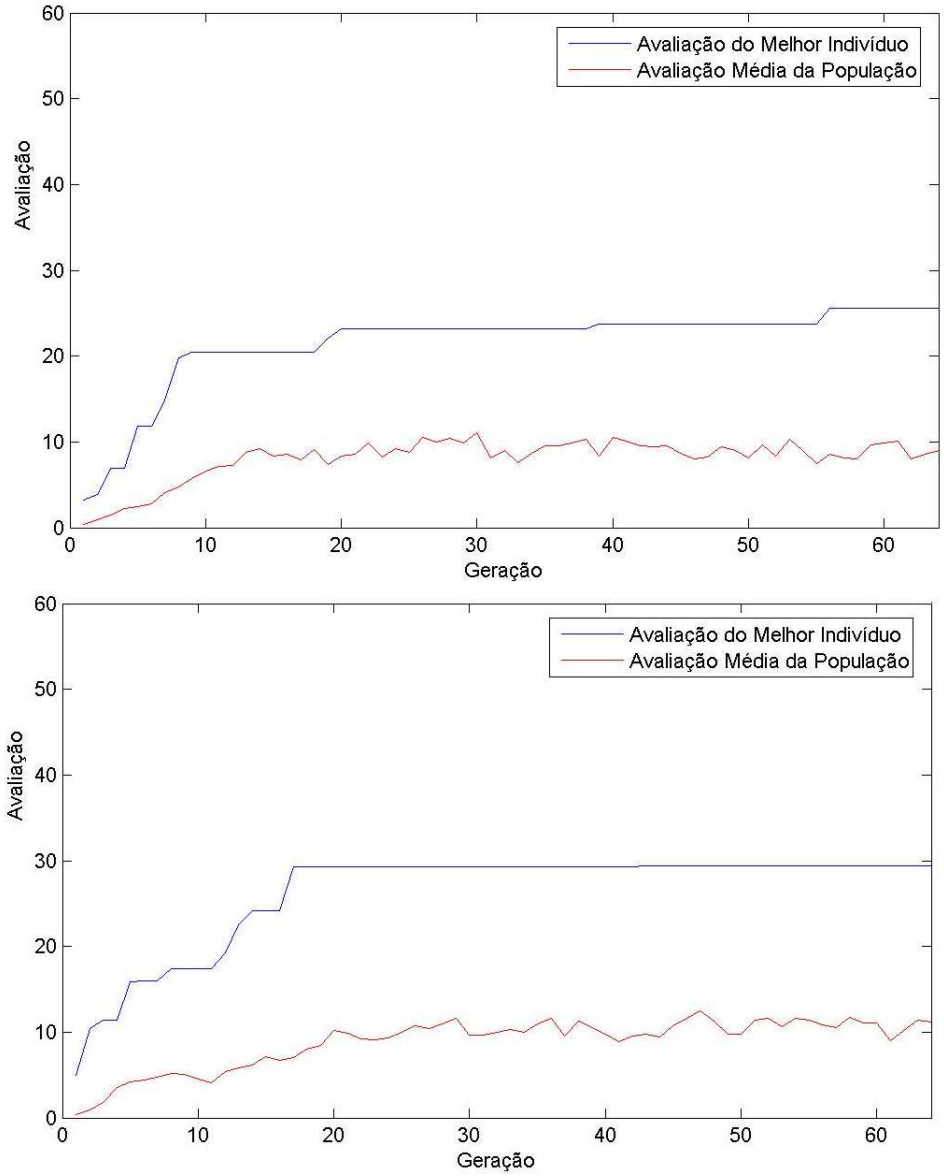


Figura A.5: Comparação entre os testes 1 e 4 para a estratégia 3. No primeiro gráfico, correspondente ao teste 1 (com população de 64 indivíduos), podemos ver que a avaliação do melhor indivíduo e a média da avaliação é inferior a do segundo gráfico, correspondente ao teste 4 (com população de 128 indivíduos).

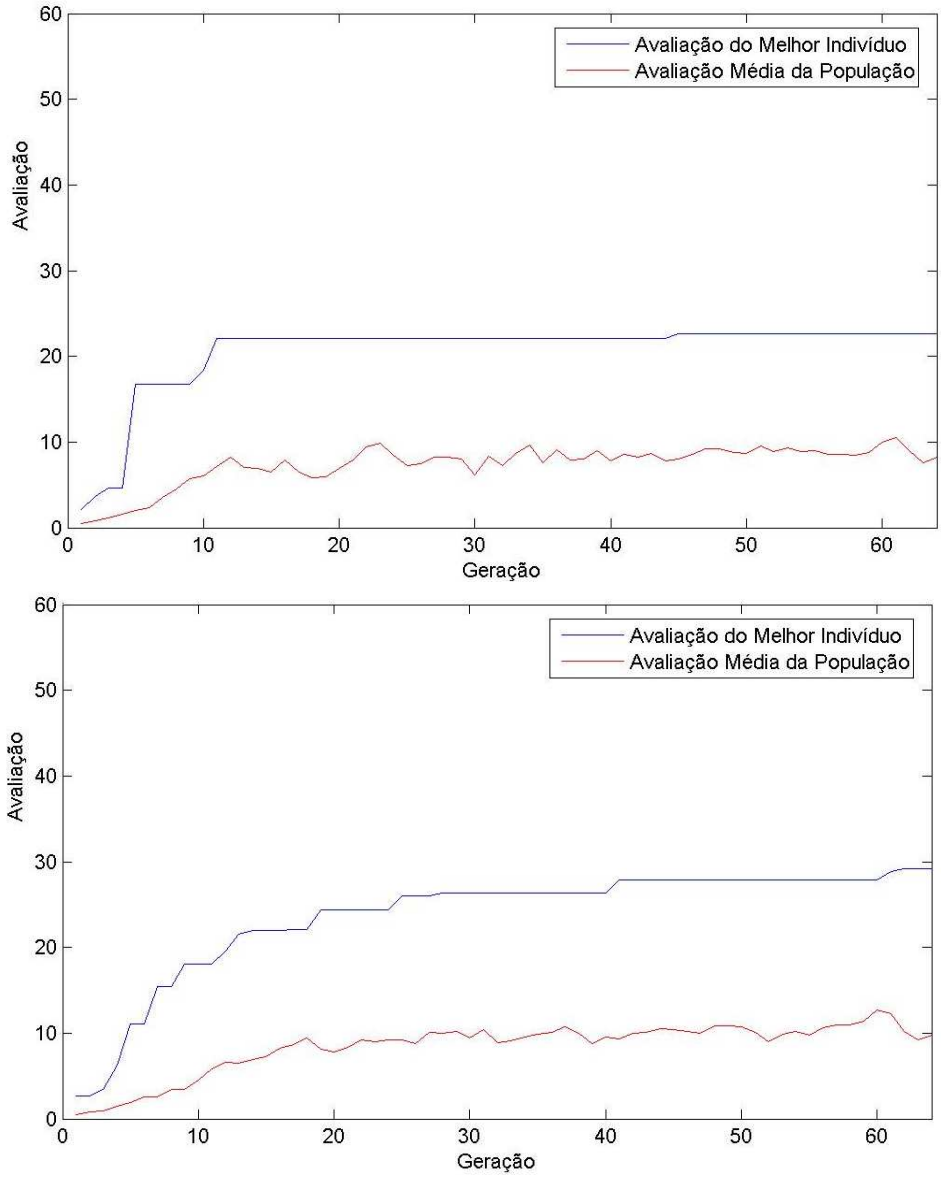


Figura A.6: Comparação entre os testes 2 e 5 para a estratégia 3. No primeiro gráfico, correspondente ao teste 2 (com população de 64 indivíduos), podemos ver que a avaliação do melhor indivíduo e a média da avaliação é inferior a do segundo gráfico, correspondente ao teste 5 (com população de 128 indivíduos).

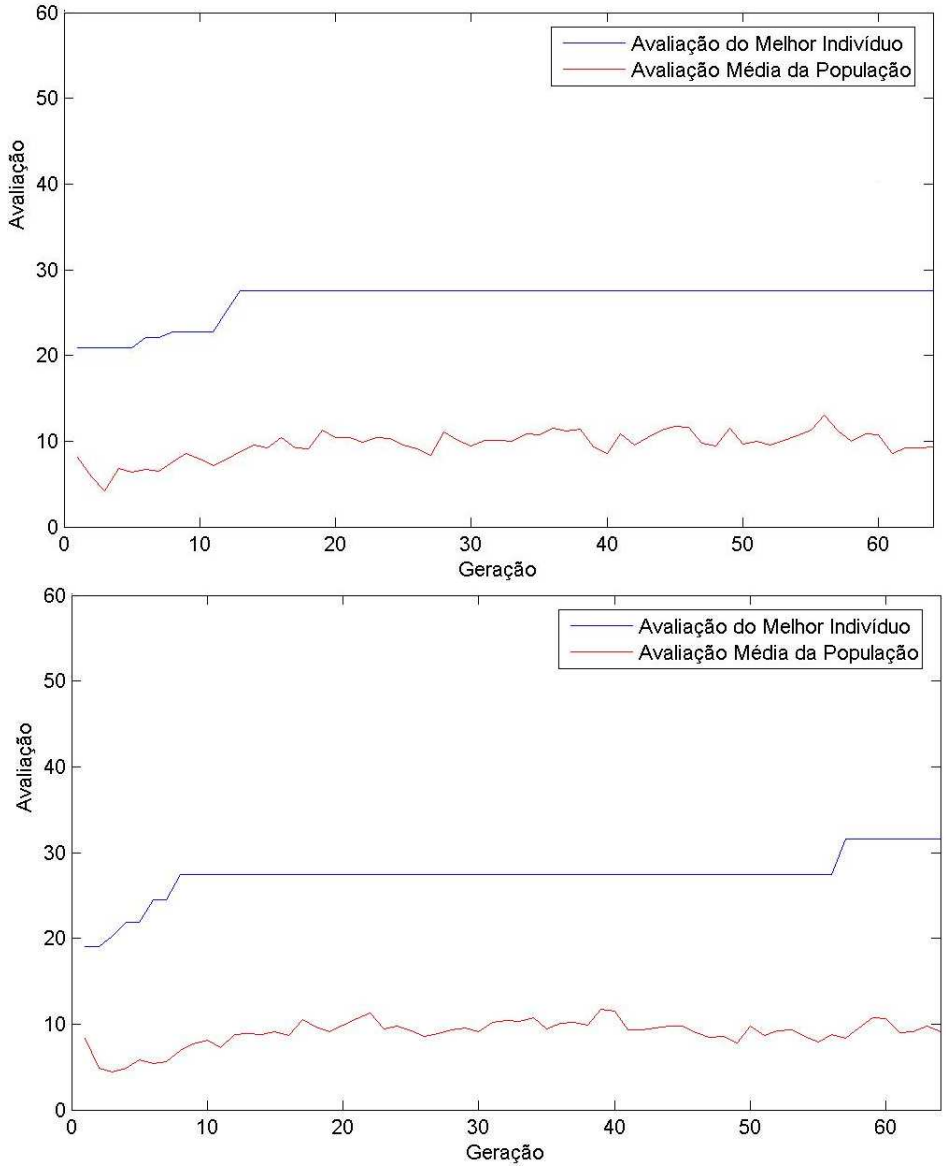


Figura A.7: Comparação entre os testes 1 e 4 para a estratégia 4. No primeiro gráfico, correspondente ao teste 1 (com população de 64 indivíduos), podemos ver que a avaliação do melhor indivíduo e a média da avaliação é inferior a do segundo gráfico, correspondente ao teste 4 (com população de 128 indivíduos).

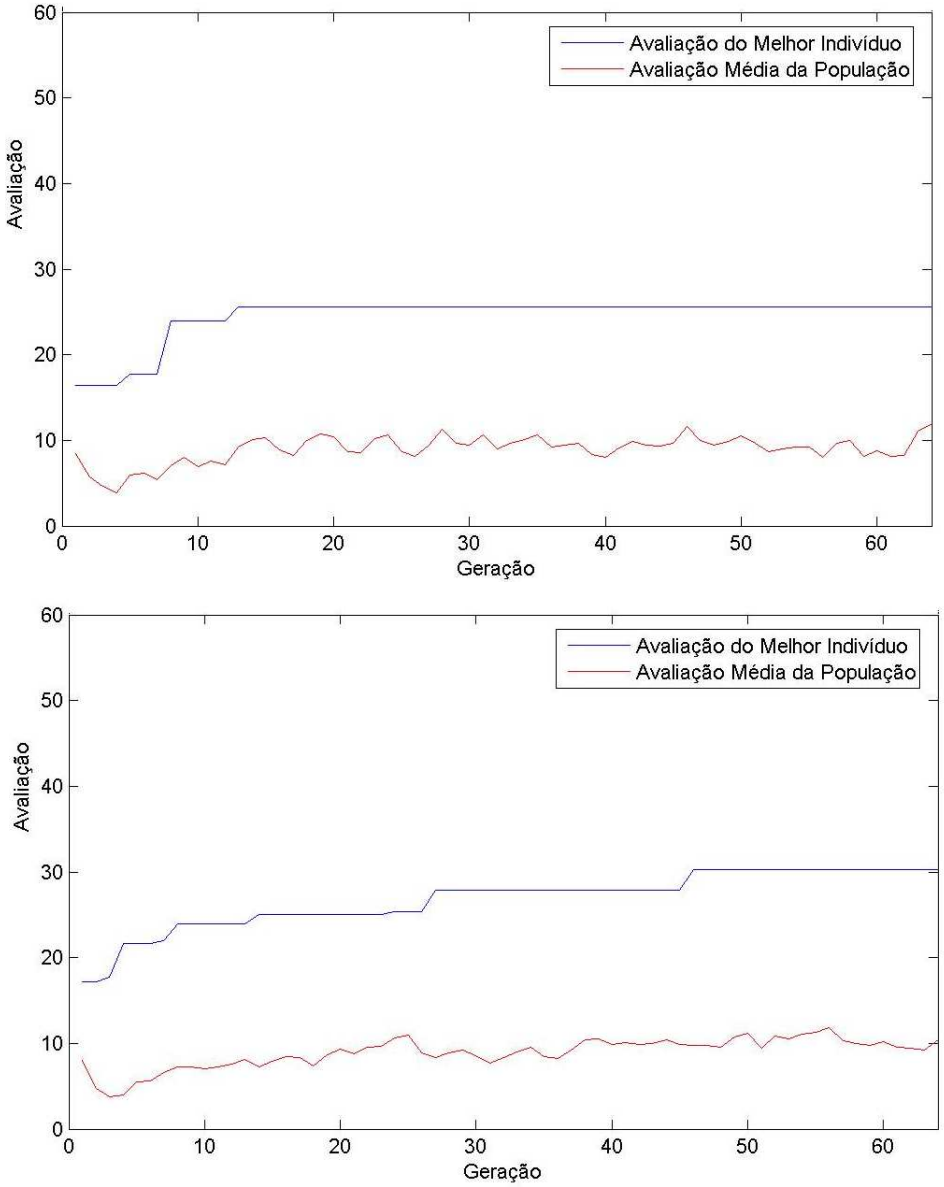


Figura A.8: Comparação entre os testes 2 e 5 para a estratégia 4. No primeiro gráfico, correspondente ao teste 2 (com população de 64 indivíduos), podemos ver que a avaliação do melhor indivíduo e a média da avaliação é inferior a do segundo gráfico, correspondente ao teste 5 (com população de 128 indivíduos).

ApêndiceB RESULTADOS COMPLEMENTARES

Tabela B.1: Resultados da análise qualitativa para quarta abordagem II

Fator de transcrição	No. de FT encontrados	% na população
OmpR	504	2,16%
Fur	488	2,10%
MalT	464	1,99%
CpxR	439	1,89%
GadX	423	1,82%
PhoB	385	1,65%
Lrp	382	1,64%
CytR	370	1,59%
FIS	341	1,46%
ArgR	317	1,36%
MelR	290	1,25%
ArcA	275	1,18%
CysB	256	1,10%
Rob	233	1,00%

Tabela B.2: Resultados da análise qualitativa para quarta abordagem III

Fator de transcrição	No. de FT encontrados	% na população
MarA	219	0,94%
AraC	213	0,91%
GlpR	211	0,91%
DnaA	209	0,90%
MetJ	205	0,88%
FadR	192	0,82%
FruR	181	0,78%
OxyR	178	0,76%
LexA	173	0,74%
PurR	169	0,73%
XylR	154	0,66%
NtrC	150	0,64%
ModE	141	0,61%
SoxS	140	0,60%
TyrR	140	0,60%
TorR	117	0,50%