UNIVERSIDADE DE SÃO PAULO
FFCLRP - DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

Maísa de Carvalho Silva

Algoritmos Evolutivos Multiobjetivos Aplicados na Otimização de Códigos Genéticos Expandidos

Ribeirão Preto SP

2020

MAÍSA DE CARVALHO SILVA

Algoritmos Evolutivos Multiobjetivos Aplicados na Otimização de Códigos Genéticos Expandidos

Versão Revisada

Monografia de Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) da Universidade de São Paulo (USP), como parte das exigências para a obtenção do título de Mestre em Ciências.
Área de Concentração: Computação Aplicada

Orientador: Renato Tinós

2020

Maísa de Carvalho Silva
Algoritmos Evolutivos Multiobjetivos Aplicados na Otimização de Códigos Genéticos Expandidos.
Ribeirão Preto–SP, 2020.
55p. : il.; 30 cm.

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras
de Ribeirão Preto da USP, como parte das exigências para
a obtenção do título de Mestre em Ciências,
Área: Computação Aplicada.

Orientador: Renato Tinós

1. Algoritmo Genético. 2. Código Genético. 3. Código Genético Expandido

MAÍSA DE CARVALHO SILVA

Algoritmos Evolutivos Multiobjetivos Aplicados na Otimização de Códigos Genéticos Expandidos

Dissertação apresentada à Faculdade de
Filosofia, Ciências e Letras de Ribeirão Preto
(FFCLRP) Da Universidade de São Paulo (USP),
como parte das exigências para a obtenção
do título de Mestre em Ciências.
Área: Computação Aplicada

Aprovado em:

Banca Examinadora:

_____

Professor Orientador
Renato Tinós

_____

Professor Convidado

_____

Professor Convidado

Ribeirão Preto – SP

2020

A sabedoria não é um produto do pensamento.
A sabedoria é um profundo conhecimento que vem do simples ato de dar total atenção a alguém ou a alguma coisa.
A atenção é a inteligência primordial, a própria consciência.
Ela dissolve as barreiras criadas pelo pensamento, levando-nos a reconhecer que nada existe em si e por si.
A inteligência une a pessoa que percebe ao objeto percebido, num campo unificado de percepção.
É a atenção que cura a separação.

(Tolle, O Poder do Silêncio)

AGRADECIMENTOS

Meus mais sinceros agradecimentos a todos aqueles que me apoiaram em todas as etapas desse projeto. Que vivenciaram todos os momentos e fases, permanecendo firmes ao meu lado (para que eu permanecesse também).

Ao meu orientador, Prof. Dr. Renato Tinós, pela parceria imensa, pelos inúmeros conselhos, pela paciência infindável e por ser um orientador de verdade. A melhor representação de calma e respeito! Sou muito grata por poder fazer parte do seu time de pesquisa e ter tido a oportunidade de dar continuidade ao projeto do seu lado, mesmo depois dos perrengues da graduação. Minha admiração por você cresce a cada dia!

À minha co-orientadora de consideração, Dra. Lariza Laura de Oliveira, pela contribuição incrível com a sua pesquisa e seus conhecimentos, estando sempre disponível para tirar dúvidas e compartilhar experiências, enquanto tomamos um café fresquinho na casinha. Sua luz é incrível, Lari! Obrigada por tudo.

Um muito obrigada especial ao meu amigo do coração, Rafael Elias, que segurou as pontas inúmeras vezes e permaneceu fiel ao meu lado. Atendeu aos meus chamados nas diversas horas do dia. De longe, quem mais lidou com minhas instabilidades e surtos internos. E quem mais vibrou comigo também com cada etapa vencida. Sua presença, muitas vezes representado por um “respira, calma, você vai conseguir”, foi fundamental para eu chegar até aqui. Você é único, Cata!

Aos meus familiares, José Luiz, Murilo e Marília, que sempre me deram forças e apoio nas decisões tomadas e confiaram em mim em mais uma etapa.

À minha estrelinha, Maria Aparecida, que nunca deixou de me acompanhar e está sempre junto de mim, em todos os momentos.

À minha amiga de laboratório, Raquel Candido, que sempre alegrava minhas quartas de reunião com boas notícias e vídeos fofinhos. Dividíamos o ombro nos momentos de lamentações também. É muito bom te ter por perto, Raquel!

À equipe Kidopi, que sempre me apoiou no mestrado, me ajudava com possíveis dúvidas e sempre permitiram “uma escapadinha no lab”, até mesmo em horário comercial. Trabalhar com vocês é incrível! Eu amo fazer parte desse time!

Aos meus amigos do PPGCA, da IBM, de Itápolis, que sempre me apoiaram e me incentivaram. A vida é muito melhor com vocês do meu lado!

Aos funcionários da secretaria, Lúcia, Jalmei e Karina, que sempre me recepcionaram com um “bom dia” alegre e perlongaram longas e deliciosas conversas no fim de cada reunião. Obrigada por sempre me ajudarem e facilitarem minha vida!

RESUMO

Recentemente, tem havido grande interesse na criação de organismos geneticamente modificados que utilizam aminoácidos não-naturais, i.e., aminoácidos diferentes dos 20 aminoácidos codificados no código genético padrão. Aminoácidos não-naturais têm sido incorporados em organismos geneticamente modificados visando o desenvolvimento de novos remédios, combustíveis e substâncias químicas. Ao incorporar novos aminoácidos, é necessário mudar o código genético padrão. Os códigos genéticos expandidos têm sido criados sem que a robustez do código seja considerada. O objetivo principal deste trabalho de mestrado é a utilização de algoritmos genéticos (AGs) para a otimização de códigos genéticos expandidos. O AG deve indicar quais códons do código genético devem ser usados para codificar um novo aminoácido não natural. Para tal fim, investigamos aqui três abordagens multiobjetivos diferentes: ponderada, lexicográfica e por Pareto. Busca-se otimizar o código expandido afim de apresentar uma robustez, em relação à polaridade e volume molecular, similar à do código genético padrão, substituindo um número pequeno de aminoácidos. Os experimentos indicam que as abordagens multiobjetivo permitem a obtenção de uma lista de códigos expandidos otimizados. Tais códigos são mais ou menos otimizados de acordo com os diferentes objetivos, permitindo ao especialista a escolha de uma solução otimizada de acordo com as necessidades.

Palavras chave: algoritmo genético. código genético padrão. código genético expandido.

ABSTRACT

Recently, there has been great interest in the creation of genetically modified organisms that use unnatural amino acids, i.e., amino acids other than the 20 amino acids encoded in the standard genetic code. Unnatural amino acids have been incorporated into genetically modified organisms to develop new drugs, fuels and chemicals. When incorporating new amino acids, it is necessary to change the standard genetic code. Expanded genetic codes have been created without considering the robustness of the code. The main objective of this master's work is the use of genetic algorithms (AGs) for the optimization of expanded genetic codes. The AG should indicate which codons in the genetic code should be used to encode a new unnatural amino acid. To this end, we investigate here three different multiobjective approaches: weighted, lexicographic and by Pareto. The aim is to optimize the expanded code in order to present a robustness, in relation to the polarity and molecular volume, similar to that of the standard genetic code, replacing a small number of amino acids. The experiments indicate that multiobjective approaches allow to obtain a list of expanded codes optimized. Such codes are more or less optimized according to the different objectives, allowing the specialist to choose an optimized solution according to the needs.

Keywords: genetic algorithm. standard genetic code. expanded genetic code.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AG - Algoritmo genético

AGP – Algoritmo genético ponderado

AGL – Algoritmo genético lexicográfico

AGMO – Algoritmo genético multiobjetivo (no caso, abordagem por Pareto)

CGP – Código genético padrão

DNA - Ácido desoxirribonucleico

Escherichia coli - *E. coli*

RNA - Ácido ribonucleico

Sumário

1. Introdução

Proteínas são macromoléculas vitais em organismos vivos, desempenhando diferentes funções, tais como, catálise, transporte, armazenamento, motilidade, defesa e regulação [LEHNINGER *et al*, 2005]. Elas são compostas por aminoácidos unidos por ligações covalentes formando séries com diferentes tamanhos e constituições. Alterações na sequência de aminoácidos ocasionam, geralmente, mudanças na estrutura tridimensional da proteína e consequentemente na sua função.

Cada aminoácido é codificado no DNA (*ácido desoxirribonucleico*) por meio de uma sequência de três nucleotídeos, chamada *códon*. Sessenta e um códons especificam aminoácidos e três códons indicam o fim do sequenciamento da proteína, durante a sua síntese, também conhecida como *tradução*. Como geralmente são utilizados 20 tipos de aminoácidos nas proteínas e como existem 4^3=64 combinações possíveis dos quatro nucleotídeos em um códon, vários aminoácidos são codificados por mais de um códon. A associação dos diferentes códons com os diferentes aminoácidos é ditada pelo *código genético*. A maioria dos seres vivos compartilham o mesmo código genético, sendo observadas algumas exceções [VOGEL, 1998]. Este código genético é conhecido como *código genético padrão* (Tabela 1).

Tabela 1: Código genético padrão.

Primeira base		Segunda base U	Segunda base C	Segunda base A	Segunda base G		Terceira base
	U	Fenilalanina (PHE)	Serina (SER)	Tirosina (TYR)	Cisteína (CYS)	U	
		Fenilalanina (PHE)	Serina (SER)	Tirosina (TYR)	Cisteína (CYS)	C	
		Leucina (LEU)	Serina (SER)	Códon de Parada	Códon de Parada	A	
		Leucina (LEU)	Serina (SER)	Códon de Parada	Triptofano (TRP)	G	
	C	Leucina (LEU)	Prolina (PRO)	Histidina (HIS)	Arginina (ARG)	U	
		Leucina (LEU)	Prolina (PRO)	Histidina (HIS)	Arginina (ARG)	C	
		Leucina (LEU)	Prolina (PRO)	Glutamina (GLN)	Arginina (ARG)	A	
		Leucina (LEU)	Prolina (PRO)	Glutamina (GLN)	Arginina (ARG)	G	
	A	Isoleucina (ILE)	Treonina (THR)	Asparagina (ASN)	Serina (SER)	U	
		Isoleucina (ILE)	Treonina (THR)	Asparagina (ASN)	Serina (SER)	C	
		Isoleucina (ILE)	Treonina (THR)	Lisina (LYS)	Arginina (ARG)	A	
		Metionina (MET)	Treonina (THR)	Lisina (LYS)	Arginina (ARG)	G	
	G	Valina (VAL)	Alanina (ALA)	Ácido Aspático (ASP)	Glicina (GLY)	U	
		Valina (VAL)	Alanina (ALA)	Ácido Aspático (ASP)	Glicina (GLY)	C	
		Valina (VAL)	Alanina (ALA)	Ácido Glutâmico (GLU)	Glicina (GLY)	A	
		Valina (VAL)	Alanina (ALA)	Ácido Glutâmico (GLU)	Glicina (GLY)	G	

[Adaptado de OLIVEIRA, 2015]

Uma pergunta que tem intrigado os cientistas há várias décadas é o porquê de um dado aminoácido ser codificado por um determinado códon. Se a associação entre um determinado códon e um aminoácido (ou código de parada) fosse fruto do acaso, então qualquer código genético, entre os cerca de $1{,}4 \times 10^{70}$ códigos possíveis, poderia ter sido selecionado [YOCKEY, 2005]. Alguns pesquisadores têm sugerido que o código genético padrão evoluiu para sua forma presente de tal maneira a torná-lo mais robusto frente a mutações [VOGEL, 1998]. De fato, quando examinamos a organização do código padrão, podemos observar que diversos aminoácidos são codificados por códons similares (Tabela 1). Ou seja, pequenas alterações na sequência de nucleotídeos podem gerar nenhuma alteração na respectiva proteína codificada. Além disso, muitas vezes, alterações nos códons causam pouca alteração nas propriedades físico-químicas dos aminoácidos codificados.

A organização do código genético padrão torna o processo de tradução da informação do DNA para as proteínas robusto, evitando e prevenindo falhas, uma vez que códons similares tendem a codificarem aminoácidos com propriedades semelhantes. Quando comparado com outros códigos gerados aleatoriamente, Freeland e Hurst (1998) observaram que o código genético padrão é mais robusto que um número muito grande de códigos hipotéticos gerados aleatoriamente[1](mais precisamente 1 em 1 milhão). Para calcular a robustez, leva-se em conta uma determinada propriedade físico-química, e.g., a polaridade

[1] O número de códigos hipotéticos piores que o código genético varia conforme a definição da função de robustez, mas em geral este número fica acima de 99,9% [FREELAND & HURST, 1998].

do aminoácido. Nos trabalhos subsequentes de Freeland e colaboradores, apenas uma medida de robustez é levada em consideração.

Em [SANTOS; MONTEAGUDO, 2011], algoritmos genéticos foram ferramenta essencial para encontrar códigos genéticos robustos, já que o intuito principal do trabalho foi estudar a adaptabilidade do código genético padrão canônico. Como em [FREELAND; HURST, 1998], Santos e Monteagudo (2011) consideraram apenas uma medida de robustez como a propriedade a ser considerada.

Em [OLIVEIRA, 2015], propõe-se utilizar mais de uma medida simultaneamente para comparar diferentes códigos genéticos. Para isso, utiliza-se *algoritmos genéticos* multi-objetivo para otimizar *códigos genéticos hipotéticos*. Ou seja, ao invés de se comparar os códigos utilizando uma única medida de robustez baseada em determinada propriedade físico-química, compara-se os códigos utilizando concomitantemente duas ou mais medidas. Em [OLIVEIRA; TINÓS, 2014], além de uma medida de robustez, considera-se também a entropia do código genético. Já em [OLIVEIRA et al., 2015] e [OLIVEIRA et al., 2017], utilizam-se duas ou três medidas de robustez baseadas em diferentes propriedades dos aminoácidos. Tal metodologia resulta em códigos hipotéticos mais similares ao código genético padrão.

As investigações envolvendo a organização do código genético padrão são importantes do ponto de vista científico, pois podem fornecer pistas relevantes ao estudo da evolução molecular. Entretanto, pesquisas envolvendo outros tipos de códigos genéticos são importantes também do ponto de vista tecnológico. Recentemente, tem havido um grande interesse em criar organismos geneticamente modificados que utilizam aminoácidos não-naturais, i.e., aminoácidos diferentes dos 20 aminoácidos codificados no código genético padrão. Estes aminoácidos podem ser interessantes por diversos motivos. Por exemplo, eles podem conter átomos pesados que facilitem alguns estudos cristalográficos envolvendo raio X. Novos aminoácidos têm sido incorporados em organismos geneticamente modificados para produzir remédios, combustíveis e substâncias químicas de grande interesse econômico [ROVNER *et al*., 2015].

Ao adicionar novos aminoácidos aos organismos geneticamente modificados, é necessário modificar o código genético padrão. Códigos genéticos expandidos geralmente são criados por meio da mudança da codificação realizada por códons pouco usados, ou pela

criação de códigos genéticos com códons com quatro nucleotídeos ao invés de três [ANDERSON *et al*., 2004]. Recentemente, propôs-se a utilização de nucleotídeos sintéticos para a criação de novos códons [ZHANG *et al*., 2017].

1.1 Objetivos

O objetivo principal deste trabalho é a investigação da utilização de algoritmos genéticos multiobjetivos para a otimização de códigos genéticos expandidos. De acordo com o conhecimento dos autores, técnicas de otimização, tais como algoritmos genéticos, não foram ainda utilizadas para a otimização de códigos genéticos expandidos. Destaca-se que códigos genéticos expandidos são de grande interesse de indústrias, como das áreas farmacêutica e química. Portanto, o desenvolvimento de códigos genéticos expandidos otimizados tem forte relevância do ponto de vista tecnológico. A otimização dos códigos expandidos por algoritmos genéticos visa principalmente o desenvolvimento de códigos mais robustos.

1.2 Organização do Trabalho

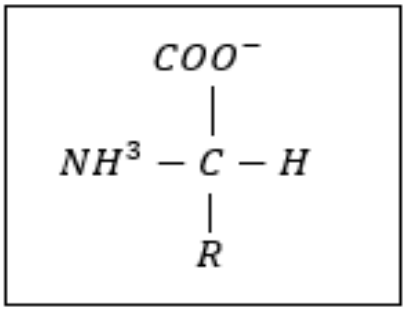
O trabalho está organizado em capítulos, devidamente referenciados no Sumário. O Código Genético está explicado no capítulo 2. Uma breve explicação sobre Algoritmos Genéticos é encontrada no Capítulo 3. A Metodologia proposta é apresentada no Capítulo 4. Experimentos estão no Capítulo 5, seguido da Conclusão no Capítulo 6.

2. Códigos Genéticos

2.1 Código Genético Padrão

As proteínas são as macromoléculas biológicas mais importantes e estruturalmente complexas dos seres vivos. Elas são constituídas por unidades menores, os aminoácidos. Estes são substâncias orgânicas unidos covalentemente uns aos outros por ligações peptídicas e que estão diretamente relacionados com a forma, função, localização celular e evolução de cada proteína, já que cada um possui uma particularidade especial e propriedades químicas distintas. Apesar de existir 20 tipos diferentes de aminoácidos codificados pelo DNA humano, todos possuem a mesma estrutura básica: um carbono central (carbono α) ligado a um grupo amina, a um grupo carboxila ácido, a um átomo de hidrogênio e a uma cadeia lateral, o radical R, que é variável em cada aminoácido (Figura 1) [LEHNINGER *et al*, 2005].

Figura 1: Estrutura comum do aminoácido, representado pelo grupo amina, carboxila, hidrogênio e cadeia lateral (variável).



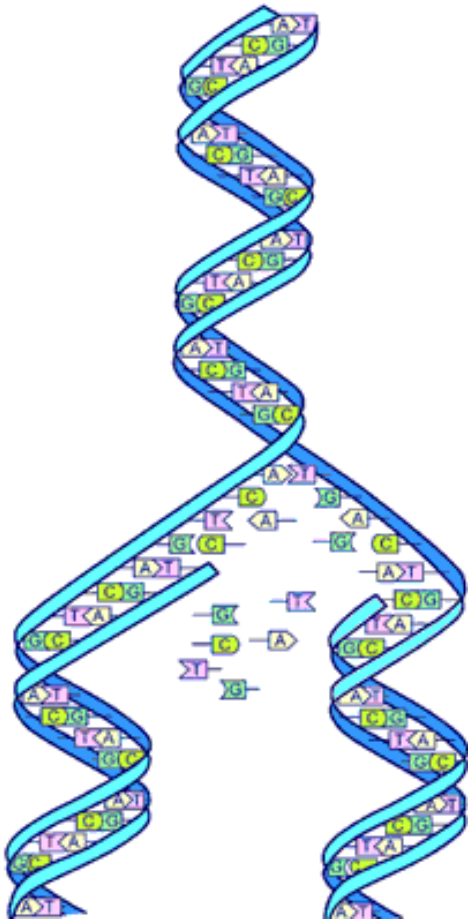
(Adaptado de [LEHNINGER *et al*, 2005].)

O aminoácido é obtido através da decodificação dos códons, constituídos por uma trinca de bases nitrogenadas, contidas dentro do DNA. O DNA é um polímero orgânico linear, longo e fino, que contém toda a informação genética hereditária. Do ponto de vista do armazenamento e processamento de informações, o DNA pode ser visto como uma sequência linear precisa formada por subunidades monoméricas, que são elementos de um alfabeto de quatro bases nitrogenadas[2]: adenina (A), guanina (G), citosina (C) e a timina (T)

[2] As bases nitrogenadas são a parte que distingue os diferentes nucleotídeos.

[LEHNINGER *et al*, 2005]. As duas fitas poliméricas se enrolam uma na outra e formam a dupla hélice, na qual cada subunidade monomérica pareia com a sua complementar na fita oposta (Figura 2).

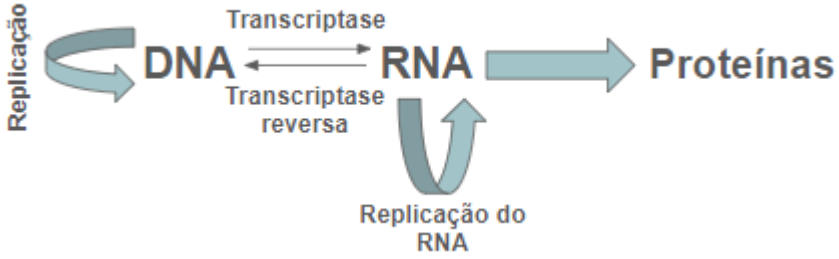
Figura 2: Estrutura do DNA. Na figura, é mostrada a dupla hélice do DNA e como está estruturada no momento da replicação e seus complementos.



(**Fonte**: https://pt.wikipedia.org/wiki/Replica%C3%A7%C3%A3o_do_DNA#/media/File:Dna-)

A informação sobre quais aminoácidos devem compor uma proteína é dada pela decodificação de uma "mensagem" que, no caso, é um fragmento de DNA, que é transcrita em RNA (ácido ribonucleico), processado em RNA mensageiro (mRNA) e levado até o local em que ocorre a síntese de proteínas (os ribossomos, localizados no citoplasma da célula) (Figura 3).

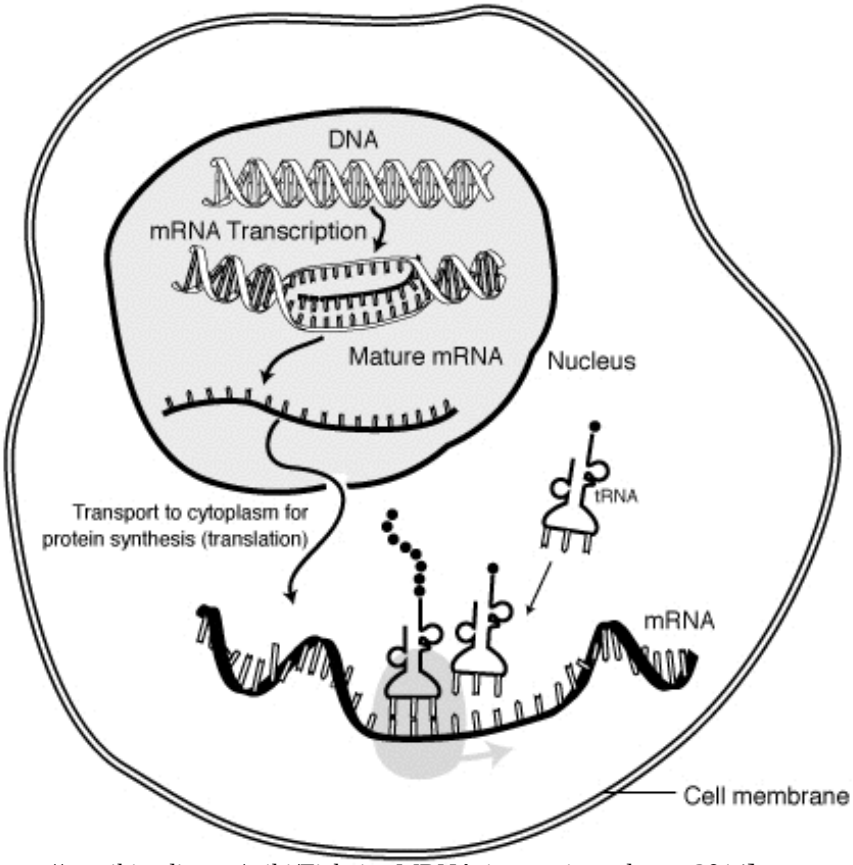
Figura 3: Dogma central da Biologia.



(Adaptado de [ALBERTS, 2002])

A tradução é iniciada por meio do códon de início (AUG – metionina) sendo, a partir desse ponto, sempre traduzido de três em três bases nitrogenadas em um aminoácido. Encontrando algum códon de parada (que podem ser UAA, UAG ou UGA), a tradução é interrompida e todos os aminoácidos que foram traduzidos e unidos por ligações peptídicas no RNA transportador são os constituintes da proteína (Figura 4).

Figura 4: Processo de síntese de proteínas, desde o processo inicial de mRNA no núcleo da célula até a leitura do mesmo pelas moléculas de RNA transportador (tRNA) no citoplasma, que associam os códons com os aminoácidos, formando as proteínas.



[Adaptado de https://pt.wikipedia.org/wiki/Ficheiro:MRNA-interaction_gl.png, 2014]

O RNA transportador (tRNA) é o elemento que garante a tradução de um determinado códon em um determinado aminoácido. O código genético é governado, portanto, pelo conjunto de RNA transportadores presente na célula.

O tRNA, além de manter todos os aminoácidos traduzidos ligados a ele, possui o anticódon, que é uma sequência de nucleotídeos complementar à do códon e funciona como um "mecanismo de segurança" para evitar pareamentos errôneos. Eles se reconhecem pelo pareamento das bases. As aminoacil-tRNA sintetase são as enzimas responsáveis por ligar os aminoácidos ao RNA transportador e por direcionar os tRNA corretos para a decodificação da mensagem, já que elas são capazes de reconhecer os tRNA ao se conectar em duas de suas extremidades (eles possuem o formato de L). A verificação de tRNA é terminada nos ribossomos, onde diferentes mecanismos são realizados para evitar o pareamento errado de códon-anticódon [WATSON *et al*, 2015].

Ou seja, há diferentes verificações pelas quais o DNA passa para que ocorram os mínimos erros possíveis durante a tradução da mensagem em proteína, destacando o quão importante é manter a codificação correta entre códons e aminoácidos. Ainda assim, alguns erros acabam passando, podendo trazer grandes consequências para a proteína codificada e para o organismo.

Como comentado no Capítulo 1, o código genético é robusto a erros no DNA e no processo que gera uma proteína a partir das informações contidas em trechos do DNA. Pode-se observar na Tabela 1 que os códons que codificam um aminoácido são, em geral, semelhantes. Na maioria dos casos, apenas uma letra (quase sempre a última) é alterada. Logo, a chance de erros de transcrição, tradução ou de mutações (que ocorrem frequentemente) alterarem o aminoácido é bem menor, já que caso a mutação ocorra nas últimas letras (que é o tipo mais comum), o aminoácido permanecerá o mesmo, não trazendo mudanças tão prejudiciais à proteína. Também é comum mutações causarem mudanças entre aminoácidos que tem propriedades físico-químicas semelhantes, a fim de causar menos danos na estrutura da proteína [LAJOIE *et al.* 2016].

2.2 Código Genético Modificado

Diversos aminoácidos não-naturais têm sido incorporados em organismos geneticamente modificados, como variedades de *Escherichia coli* (*E. coli*), fungos e células mamárias [LIU & SCHULTZ, 2010]. Xiao e Schultz (2016) citam que mais de 200

aminoácidos não-naturais foram geneticamente codificados até 2015; tais aminoácidos apresentam, muitas vezes, propriedades biológicas, químicas e físicas diferentes das dos aminoácidos naturais. Novos aminoácidos conferem novas funções às proteínas, tais como: i) reação com diferentes compostos; ii) produção de proteínas fluorescentes; iii) facilitação de determinados estudos de cristalografia de raio X.

Ao incorporar um novo aminoácido, o código genético padrão deve ser modificado. A maneira mais comum de se fazer isso é criando, no código genético padrão, novas associações para os códons que raramente são utilizados [ROVNER et al., 2015]. Por exemplo, o códon de parada *UAG* é bastante raro na *E. Coli*. Assim, uma prática comum é desenvolver moléculas de RNA transportador associadas ao *UAG* para que codifiquem o novo aminoácido. Outra alteração possível é expandir o código genético para trabalhar com 4 pares de base ao invés de 3; assim, o número de possibilidades de codificação aumenta, podendo-se agora incluir novos aminoácidos.

Zhang *et al.* (2017) propuseram a utilização de nucleotídeos sintéticos para a criação de novos códons. Assim, os códons já utilizados pelo código genético padrão não precisam ser modificados. Ao incorporar novos nucleotídeos, o alfabeto do DNA cresce, permitindo a utilização de diversas novas combinações das bases nitrogenadas nos códons. Além disso, aumenta-se o isolamento do meio, assegurando-se que estes organismos modificados não recombinem com organismos biológicos naturais.

Vale ressaltar que, de acordo com nosso conhecimento, os códigos expandidos não são otimizados quanto à robustez. Dada uma propriedade físico-química, a robustez de um código genético representado pelo vetor **x** é calculada utilizando-se o erro médio quadrático [HAIG & HURST, 1991] dado por:

$$\boldsymbol{M_s(x)} = \frac{\sum_i \quad \sum_{j \in N(i)} \quad w(i,j)\left(X(i,x) - X(j,x)\right)^2}{T} \qquad (1)$$

sendo $X(i,\mathbf{x})$ a propriedade do aminoácido codificado pelo i-ésimo códon do código **x**, $w(i,j)$ a ponderação correspondente a troca dos códons na posição i e j (algumas posições são mais

suscetíveis a erros que outras), $N(i)$ o subconjunto de códons obtidos por meio de mudanças simples no i-ésimo códon e T o número total de mudanças simples entre códons.

A Eq. (1) é dada pelo erro médio quadrático de todas as alterações possíveis na propriedade dos aminoácidos. A somatória é ponderada por um termo que leva em consideração a posição do nucleotídeo [OLIVEIRA & TINÓS, 2014]. A robustez do código em relação à propriedade X pode ser entendida como o inverso do erro médio quadrático dado na Eq. (1).

Quando a polaridade do aminoácido é considerada como propriedade $X(i,\mathbf{x})$ na Eq. (1), verifica-se que o código genético padrão é mais robusto que a esmagadora maioria dos códigos genéticos hipotéticos. Levando-se em conta a ponderação pela posição da base dentro do códon, a literatura indica que o código genético padrão é mais robusto que 99,9% dos códigos hipotéticos gerados aleatoriamente [FREELAND & HURST, 1998]. Entretanto, quando aminoácidos não-naturais são inseridos, a robustez do código genético é modificada. A proposta aqui é utilizar algoritmos genéticos (AGs) para otimizar os códigos genéticos expandidos.

3. Algoritmos Genéticos

Os AGs são métodos adaptativos inspirados nos processos genéticos de organismos biológicos e na teoria da evolução por seleção natural. AGs são frequentemente utilizados para resolver problemas de busca e otimização [MITCHELL, 1996].

3.1 Algoritmo Genético Padrão

No AG padrão, um conjunto de indivíduos (ou cromossomos) representando soluções do problema é sujeito a operadores de seleção e transformação inspirados em mecanismos encontrados na evolução por seleção natural e na genética. A solução $\mathbf{x}_i$ (também chamada de indivíduo ou cromossomo), para i=1,...,N, sendo N o tamanho da população, é avaliada através de uma função de avaliação, ou *fitness*, $f(\mathbf{x}_i)$.

A função de avaliação é relacionada ao problema de otimização. Aqui é onde cada solução candidata (indivíduo) receberá uma nota. O objetivo do AG é encontrar o melhor fitness, ou seja, encontrar a melhor solução candidata presente no espaço de soluções.

A criação da função de fitness deve ser feita com todo cuidado: quanto mais informações a respeito do problema tiver, melhor. Regras, restrições, exceções, tudo deve estar dentro dessa função. Quanto mais completa, melhor ela representará a qualidade das soluções no problema real e, então, melhor será o resultado do processo de otimização.

O fitness é utilizado para selecionar os indivíduos a serem reproduzidos. Os operadores de seleção funcionam de maneira similar ao que ocorre na seleção natural: diante de todos os indivíduos de uma população, visa selecionar os mais aptos, aqueles que possuem características mais fortes (no caso, os melhores fitness), mas não deixa de considerar os menos aptos, com características menos favoráveis. Assim, ele seleciona com mais frequência os indivíduos com melhor fitness e, ora ou outra, seleciona fitness não tão bons, de maneira a, por exemplo, preservar a diversidade da população.

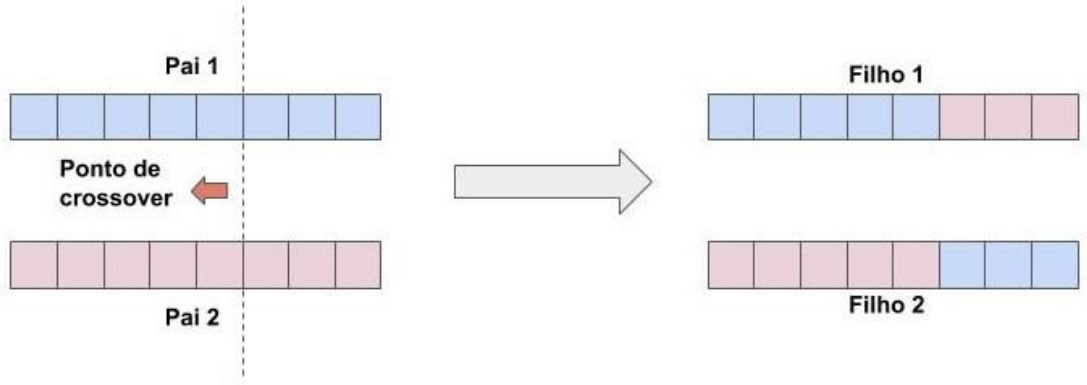
No final, depois de gerados os melhores filhos, os pais são excluídos e substituídos por essa nova melhor população.

Os métodos mais utilizados para selecionar indivíduos são: o método da roleta, no qual a probabilidade de um indivíduo ser selecionado é proporcional ao seu fitness relativo (i.e.,

ao fitness normalizado pela soma de fitness dos indivíduos da população atual); o método de torneio, onde um grupo de indivíduos é selecionado aleatoriamente e aquele que tiver o melhor valor de fitness é selecionado; e o método de elitismo, onde o indivíduo com o melhor fitness é obrigatoriamente selecionado.

Após a seleção dos indivíduos, estes são transformados. Os operadores de transformação mais utilizados são o *crossover* e a mutação. No primeiro, dois indivíduos da população corrente escolhidos por meio do operador de seleção têm algumas das variáveis de decisão trocadas (Figura 5). A probabilidade de se aplicar *crossover* é definida por uma taxa p_c , chamada de taxa de *crossover.*

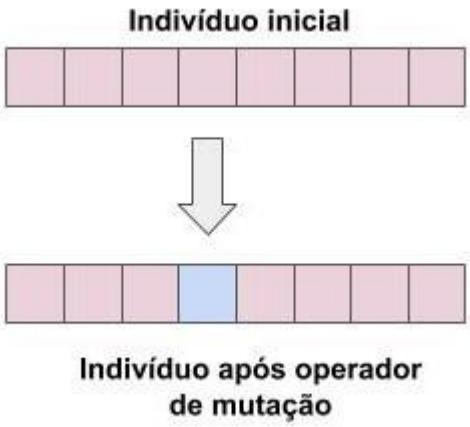
Figura 5: Crossover de um ponto. Exemplo de como o operador genético crossover ocorre na geração de indivíduos



(Adaptado de [LINDEN, 2008])

Na mutação, indivíduos têm alguns de seus elementos alterados por meio de uma regra pré-definida. Por exemplo, quando ocorre mutação no *i*-ésimo elemento do cromossomo para o caso binário, este elemento é negado (Figura 6). O número de genes alterados por mutação é controlado por uma taxa p_m , chamada de taxa de mutação.

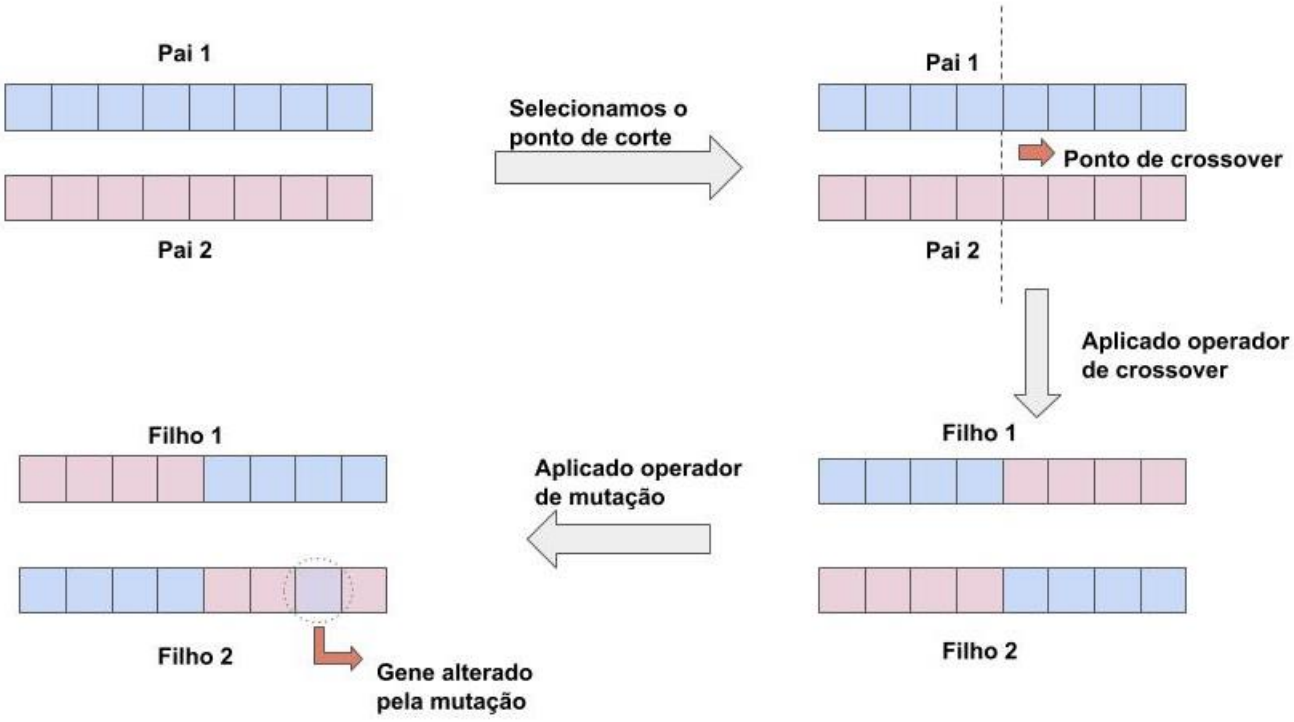
Figura 6: Operador de Mutação.



(Adaptado de [LINDEN, 2008])

Um exemplo da aplicação conjunta dos dois operadores de transformação é mostrado na Figura 7.

Figura 7: Operadores de Crossover e Mutação. Em (a), estão selecionados os pais; em (b), é definido um único ponto de corte, onde irá ocorrer o crossover em (c); em (d), o crossover ocorreu e, em azul, o gene do filho 2 foi mutado.



(Adaptado de [LINDEN, 2008])

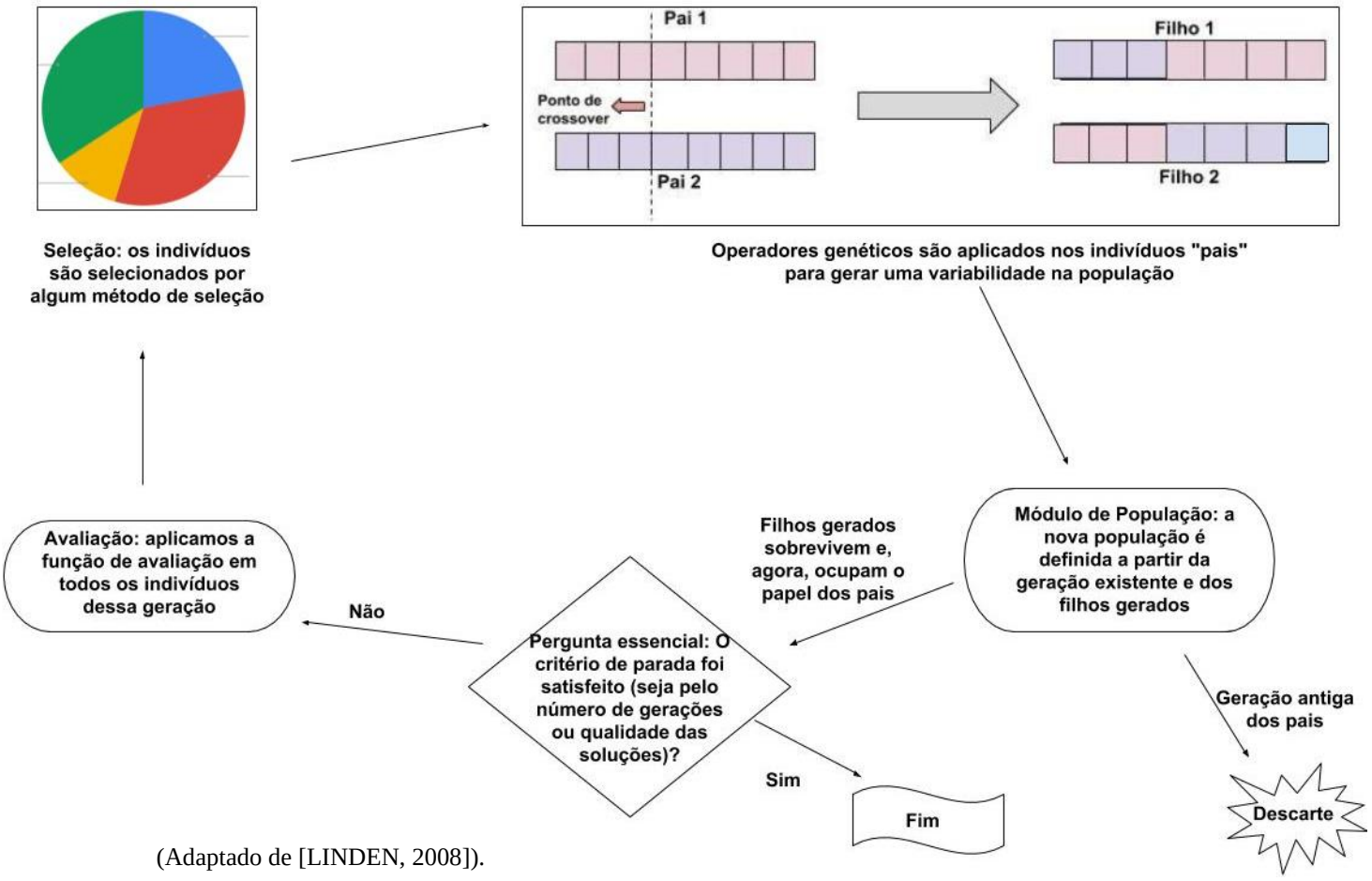
O pseudocódigo simplificado do AG padrão é apresentado na Figura 8, assim como o funcionamento resumido do mesmo é apresentado na Figura 9.

Figura 8: Pseudocódigo do AG padrão.

```
Algoritmo: AG padrão

início
        inicialize a população
        avalie a população inicial
        repita
                se critério de convergência for satisfeito
                        interrompa
                fim se
                selecione indivíduos para a nova população
                aplique mutação e cruzamento nos indivíduos selecionados
                avalie os indivíduos da nova população
        fim repita
fim
Algoritmo: AG padrão
```

Figura 9: Visão do laço principal do Algoritmo Genético. Primeiro selecionamos quais serão os pais do processo, aplicamos os operadores genéticos, crossover e mutação, para variar os genes e, então, encontramos a primeira geração. A partir dela, o algoritmo vai gerando novas gerações, sempre descartando os pais anteriores. A nova geração é avaliada pela função de avaliação (fitness).



(Adaptado de [LINDEN, 2008]).

3.2 Algoritmo Genético Multiobjetivo

Em muitos problemas reais, deve-se otimizar simultaneamente mais de um objetivo. Muitas vezes, esses objetivos são conflitantes. Na indústria e confecção de produtos, por exemplo, visamos ter uma boa qualidade do produto com um baixo custo de produção. Problemas com múltiplos objetivos são chamados de problemas multiobjetivo. Para lidar com tais problemas, são necessários algoritmos multiobjetivo, sendo que sua aplicação se estende para muitas áreas, como telecomunicação e bioinformática [EL-GHAZALLI, 2009].

Os problemas multiobjetivos são muito comuns no dia a dia. Com um grande volume de dados, onde temos muitas características importantes, definir o "melhor resultado" pode ser muito complexo. Para isso, o uso de AGs multiobjetivo é interessante: como o objetivo é sempre otimizar mais de um objetivo, o uso de populações de soluções permite lidar com o problema de forma mais natural.

Nesse trabalho, iremos falar sobre três abordagens evolutivas multiobjetivas: abordagem ponderada, abordagem lexicográfica e abordagem por Pareto.

3.2.1. Abordagem Ponderada

Nessa abordagem, cada objetivo é calculado separadamente e, ao fazer o cálculo final, atribuímos um peso para cada um dos objetivos, de acordo com a importância de cada um deles. Por fim, somamos cada um desses objetivos ponderados, de forma que chegamos em um único resultado. Na equação (2) genérica abaixo, exemplificamos o cálculo:

$$f(\mathbf{x}) = w_1 f_1(\mathbf{x}) + w_2 f_2(\mathbf{x}) + \ldots + w_{n-1} f_{n-1}(\mathbf{x}) + w_n f_n(\mathbf{x}) \quad (2)$$

onde $f_n(x)$ representa a função que calcula o objetivo n e w_n representa o peso atribuído para essa função. Logo, se $f_1(\mathbf{x})$ é mais importante que $f_2(\mathbf{x})$, o valor de w_1 será maior que de w_2, e assim respectivamente. Repare que nesta abordagem, transformamos o problema multiobjetivo em um problema com um único objetivo, podendo-se obter diferentes soluções alterando-se os diferentes pesos.

3.2.2. Abordagem Lexicográfica

Na abordagem lexicográfica, cada objetivo é calculado separadamente e cada objetivo utilizado no cálculo de fitness possui uma ordem de importância. Quando duas

soluções são comparadas, as avaliações de cada objetivo para cada solução são comparadas uma a uma, seguindo uma ordem pré-definida. Para chegar em uma melhor solução, é comparado o primeiro objetivo definido na ordem de importância. Se o valor absoluto da diferença de fitness das soluções para este objetivo for maior que o desvio padrão encontrado na população, a solução com melhor fitness (para o objetivo analisado) é escolhida e os outros objetivos não são analisados. Porém, se o valor absoluto da diferença não for maior que o desvio, considera-se o próximo objetivo e assim por diante. Se não houver um valor objetivo melhor, a solução (no caso, o indivíduo) que apresentar o melhor resultado para o objetivo com maior prioridade é escolhida [FREITAS, 2004].

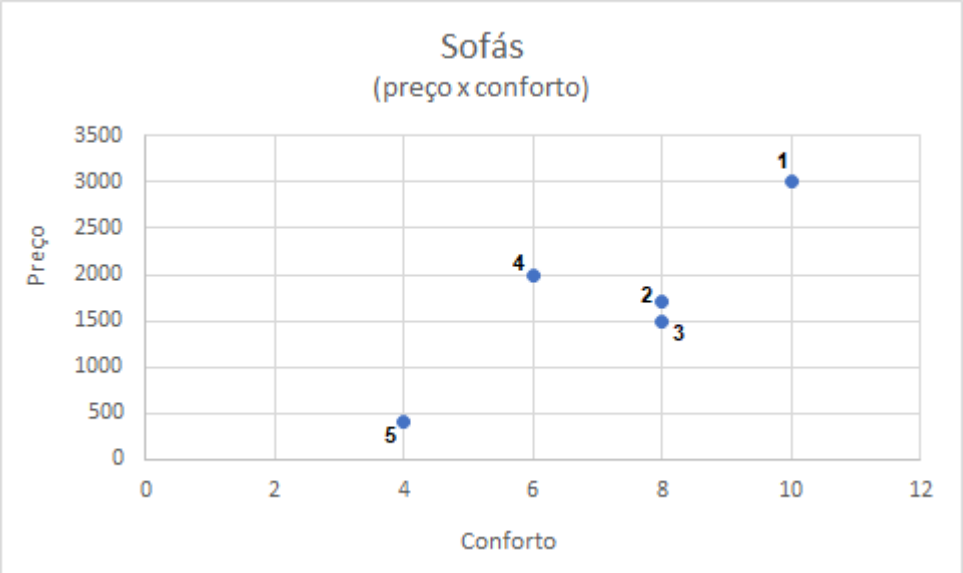
3.2.3 Abordagem por Pareto

Nos problemas multiobjetivos é difícil entendermos qual a "melhor solução", dado um conjunto de soluções. Considere que temos duas soluções, *X* e *Y*, onde *X* apresenta uma melhor avaliação para o objetivo *i* e uma pior avaliação para o objetivo *j*, em comparação com a solução *Y*. Nesse caso, não é possível escolhermos entre essas opções a melhor, se é claro, não houver outros critérios para avaliação das soluções. Porém, se a solução *X* tiver o objetivo *i* melhor avaliado que a solução *Y* e a avaliação do objetivo *j* das duas for igual, podemos dizer que a solução *X* é melhor que a solução *Y*. Neste caso, dizemos que a solução *X* domina a solução *Y* [OLIVEIRA, 2015].

O conjunto de soluções não dominadas de um problema é chamado de **Conjunto de Pareto**. Assim, um algoritmo multiobjetivo utilizando o conceito do conjunto de Pareto visa encontrar o conjunto de soluções não dominadas de um problema multiobjetivo. Por trabalhar simultaneamente com uma população de soluções, AGs são interessantes para se encontrar soluções não dominadas.

Para ilustrar melhor ao conceito de Conjunto de Pareto, olhemos a Figura 10. O ponto 1 possui um preço maior e um conforto alto. Já o ponto 5, possui um preço mais baixo e um conforto menor também. Não é possível dizer qual ponto é melhor nesse caso. Porém, ao olharmos os pontos 2 e 3, que possuem o mesmo conforto, podemos dizer que o ponto 3 é melhor que 2, já que ele possui um preço mais baixo. Dessa forma, 2 não pertence ao Conjunto de Pareto. Já o ponto 4, ao compararmos com os pontos 2 e 3, vemos que também não pertence ao Conjunto de Pareto, pois além de ser mais caro, possui um menor conforto.

Figura 10: Exemplo de problema multiobjetivo. No gráfico, não conseguimos definir qual é a melhor solução, se não tiver um parâmetro a mais para considerarmos. O único que conseguimos definir é entre os pontos 2 e 3, que apresenta os mesmos valores de conforto. Logo, o ponto 3 é melhor que o 2, por apresentar um preço menor. Dessa forma, concluímos que 3 domina 2. Já em comparação com o ponto 3 e 4, vemos que 3 também domina 4, por possuir maior conforte e menor preço. Logo, nosso Conjunto Pareto é formado pelos pontos 1, 3 e 5.



(Adaptado de [OLIVEIRA, 2015]).

Existem vários algoritmos genéticos que utilizam a abordagem por Pareto. Um dos mais utilizados é o NSGA-II (representado na Figura 12). Ele é muito utilizado em problemas multiobjetivo que não tenham muitos objetivos na codificação; geralmente o NSGA-II é utilizado em problemas com 2 ou 3 objetivos.

Figura 11: Pseudocódigo para o algoritmo rápido de ordenação por dominância no algoritmo

```
Início
  Para cada p ∈ P
    S_p = ∅
    n_p = 0
    Para cada q ∈ P
      Se p domina q, então
        S_p = S_p ∪ q
      Senão se q domina p, então
        n_p = n_p + 1
      Fim se
    Fim para
    Se n_p = 0, então
      p_rank = 1
      F_1 = F_1 ∪ p
    Fim se
  Fim para
  i = 1
  Enquanto F_i ≠ ⊘
    Q = ⊘
    Para cada p ∈ F_i
      Para cada q ∈ S_p
        n_p = n_p − 1
        Se n_p = 0, então
          q_rank = i + 1
          Q = Q ∪ q
        Fim se
      Fim para
    Fim para
    i = i + 1
    F_i = Q
  Fim enquanto
Fim
```

([OLIVEIRA, 2015])

O NSGA-II funciona, primeiramente, criando uma população *P* de tamanho *n*, ordenando os indivíduos por dominância e separando-os em diferentes fronteiras. Ele possui um procedimento rápido para realizar essa ordenação por dominância, possuindo um contador de dominância n_p. O contador n_p conta quantos indivíduos dominam uma solução e o S_p é o conjunto de soluções que essa solução domina. Com base nisso, ele define em qual fronteira cada solução *p* irá ficar. As primeiras soluções localizadas são as da primeira fronteira, que possuem $n_p=0$ (ou seja, não são dominadas por nenhuma outra solução).

Depois, dentro de cada solução dessa fronteira, olha-se cada solução contida no conjunto S_p e decrementa 1 do contador n_p. Se, ao decrementar, n_p chegar a 0, essa solução é adicionada a uma lista, fazendo parte da segunda fronteira de dominância. O algoritmo segue até que encontre as fronteiras de todas as soluções obtidas.

Depois de ordenar a primeira população, ele aplica os operadores de reprodução e seleção em cima dessa população, criando uma nova população *Q*, também de tamanho *n*. A população *P* e *Q* é unida, formando uma população *R* com tamanho *2n*. Com essa nova população *R*, é aplicado o operador de torneio, levando em consideração a distância de multidão explicada abaixo para selecionar apenas os melhores indivíduos e, então, a população é ordenada novamente, seguindo o mesmo procedimento citado anteriormente.

Os indivíduos localizados na primeira fronteira possuem valores de fitness melhor que os localizados na segunda fronteira, que são melhores que o da terceira fronteira e assim sucessivamente. Porém, pode ser que haja um grande conjunto de soluções nessa primeira fronteira. A solução "desempate" para esse caso é a chamada *distância de multidão*.

A distância de multidão ordena a população de acordo com o seu objetivo. Depois de ordenado, para cada objetivo, a distância das soluções intermediárias é calculada pela diferença absoluta normalizada dos valores da função das duas soluções vizinhas. Nas soluções da borda, é atribuída um valor de distância infinita. Depois de definir um valor de distância para cada solução, se torna viável comparar duas soluções, de acordo com a sua proximidade com as demais.

Por fim, depois que todas as fronteiras tiveram suas soluções ordenadas com base na distância de multidão, os indivíduos com maiores distâncias são adicionados para formar a nova população P_{t+1}, de tamanho *n*.

O conjunto solução final é composto pelos indivíduos da população final que possuem contador $n_p = 0$, ou seja, não são dominados por nenhuma outra solução e estão localizadas na camada mais inferior. Todas essas soluções ordenadas formam o conjunto de soluções não-dominadas encontrado pelo algoritmo ou o conjunto aproximado do Pareto-ótimo. O objetivo do NSGA-II é encontrar esse conjunto final de soluções dominadas o mais próximo possível do conjunto de Pareto global.

4. Metodologia

Uma maneira de inserir novos aminoácidos, que geralmente é utilizada na prática, é substituir códons que raramente são utilizados [ROVNER *et al*., 2015]. Para tanto, é necessário conhecer a distribuição do uso dos diferentes códons no organismo em questão. Entretanto, como dito anteriormente, não se leva em consideração a robustez do novo código expandido para a inserção de aminoácidos novos nos trabalhos apresentados até aqui na literatura.

Neste trabalho, AGs são utilizados para a otimização de códigos expandidos. Indivíduos do AG representam códigos expandidos hipotéticos que incorporam um novo aminoácido não-natural. Na metodologia aqui proposta, inspirada pela pesquisa apresentada em [OLIVEIRA, 2015] para a investigação de adaptabilidade do código genético padrão, utilizamos três objetivos diferentes para lidar com o problema de otimização: polaridade, frequência e volume molecular. A frequência não era levada em conta em [OLIVEIRA, 2015], pois o objetivo não era obter códigos expandidos nos quais novos aminoácidos são inseridos no código genético. Aqui, investigamos a inserção de apenas um aminoácido no código padrão, apesar de os métodos apresentados poderem ser estendidos para casos com mais de um novo aminoácido. Vale destacar que serão considerados aminoácidos novos hipotéticos. Três abordagens multiobjetivo (ver Capítulo 3) são utilizadas: a abordagem ponderada, a abordagem lexicográfica e a abordagem por Pareto. As três são descritas nas Seções 4.2, 4.3 e 4.4, respectivamente. Antes, aspectos do método baseado em AGs e configurações comuns nas 3 abordagens são apresentados na Seção 4.1.

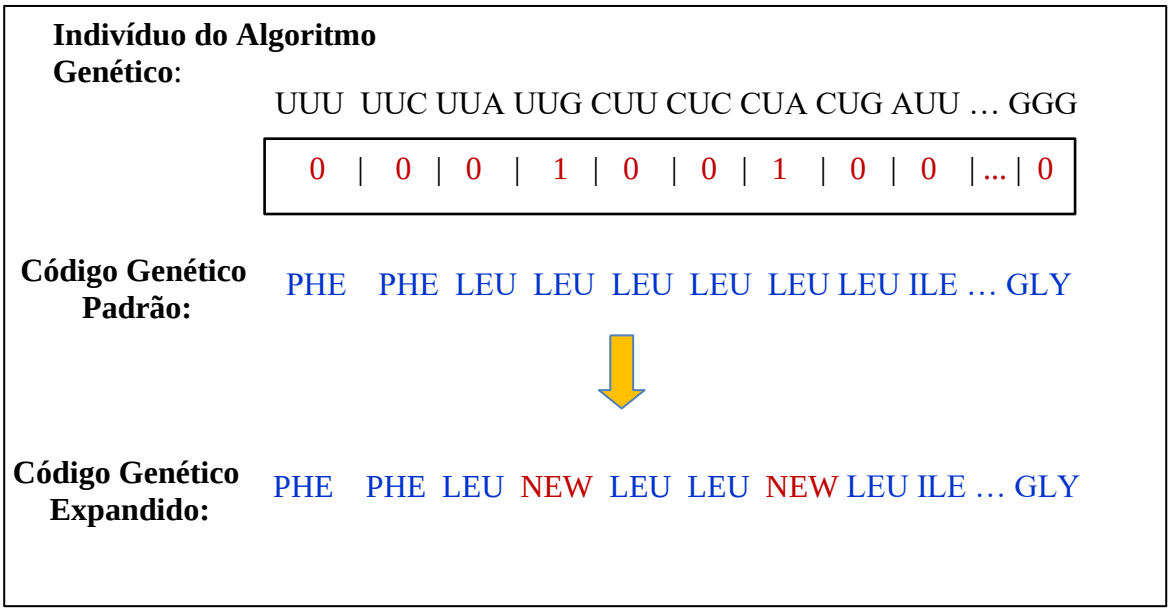
4.1 Aspectos do AG

4.1.1 Aspectos do AG: Codificação

Para realizar os experimentos propostos nesse trabalho de Mestrado, desenvolvemos algoritmos genéticos multiobjetivos em C++.

Para lidarmos com os problemas dos códigos expandidos, utilizamos um vetor binário para representá-los no AG. A Figura 12 mostra um exemplo da decodificação do indivíduo do AG no código expandido. O vetor possui 61 posições, onde cada posição representa um códon específico do código genético. O vetor que contém o código genético padrão possui o valor zero em todas as suas posições, o que significa que não há alteração em nenhum códon com o seu aminoácido. Um elemento igual a 1 no vetor significa que o códon correspondente no código genético padrão será relacionado agora ao novo aminoácido. Os operadores de mutação e reprodução também irão utilizar codificação binária [MITCHELL, 1996]. Pode-se observar que os três códons de parada (*UAA*, *UAG* e *UGA*) são desconsiderados, o que resulta em um vetor binário com 61 posições. Quando os operadores de reprodução removem um dos aminoácidos (naturais ou incorporados), penaliza-se o indivíduo, por exemplo adicionando-se um valor de 10.000 ao seu *fitness*, com o intuito desse indivíduo não ser selecionado.

Figura 12: Ilustração de como a solução é codificada no vetor binário. NEW representa o novo aminoácido introduzido.



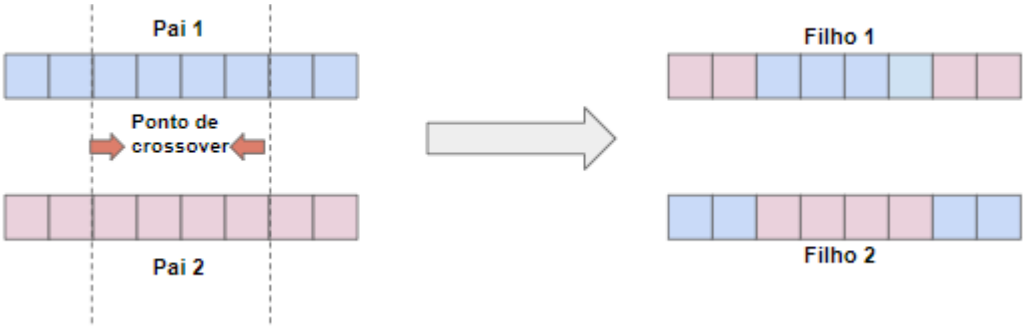
4.1.2 Aspectos do AG: Operadores de Reprodução e Seleção

Nas três abordagens, optamos por utilizar os operadores de seleção: torneio e elitismo. No elitismo, o melhor indivíduo da população é sempre escolhido, levando-o para

as próximas gerações sem alterações. No torneio, o melhor de um conjunto de indivíduos escolhido aleatoriamente é selecionado.

Já para os operadores de reprodução, utilizamos o crossover e mutação. A mutação binária é aplicada com uma taxa de mutação pré-definida. Foi utilizado o crossover de dois pontos, no qual ocorre uma recombinação genética, onde dois indivíduos são selecionados aleatoriamente e têm seu material genético recombinado. Neste tipo de crossover, 2 pontos são aleatoriamente escolhidos (Figura 13). Estes pontos definem as partes que cada filho herda dos pais. Por exemplo, o filho 1 herda os elementos até o ponto 1 e depois do ponto 2 do pai 1, e o restante do pai 2, enquanto que o filho 2 herda o complemento.

Figura 13: Exemplo de crossover de 2 pontos.



4.1.3 Aspectos do AG: Objetivos

Com base em [OLIVEIRA & TINOS, 2015], utilizamos aqui dois objetivos baseados em robustez. O terceiro objetivo considerado aqui é a frequência dos aminoácidos a serem substituídos. Para considerar qual melhor lugar para inserir um novo aminoácido, o objetivo relacionado à frequência prioriza quais são os códons menos utilizados e, consequentemente, onde a substituição traria menor impacto no código genético padrão. Para calcular este objetivo, são utilizadas as frequências de cada códon da *E. coli*, conforme apresentado na tabela 2 abaixo.

Além da frequência, usamos outros dois objetivos relacionados à robustez: em relação à polaridade e em relação ao volume molecular. Estes objetivos foram também utilizados em [OLIVEIRA, 2015], mas considerando apenas códigos genéticos sem novos aminoácidos. Os valores de polaridade e volume molecular para os aminoácidos naturais estão apresentados Tabela 3.

Como utilizamos um novo aminoácido hipotético, consideramos valores também hipotéticos de polaridade (7.4) e volume molecular (85.0).

Tabela 2: Tabela de frequências de uso de códons na *E. coli*.

	U			C			A			G			
U	UUU	Phe	1.9	UCU	Ser	1.1	UAU	Tyr	1.6	UGU	Cys	0.4	U
	UUC	Phe	1.8	UCC	Ser	1.0	UAC	Tyr	1.4	UGC	Cys	0.6	C
	UUA	Leu	1.0	UCA	Ser	0.7	UAA	Stop	0.2	UGA	Stop	0.1	A
	UUG	Leu	1.1	UCG	Ser	0.8	UAG	Stop	0.03	UGG	Trp	1.4	G
C	CUU	Leu	1.0	CCU	Pro	0.7	CAU	His	1.2	CGU	Arg	2.4	U
	CUC	Leu	0.9	CCC	Pro	0.4	CAC	His	1.1	CGC	Arg	2.2	C
	CUA	Leu	0.3	CCA	Pro	0.8	CAA	Gln	1.3	CGA	Arg	0.3	A
	CUG	Leu	5.2	CCG	Pro	2.4	CAG	Gln	2.9	CGG	Arg	0.5	G
A	AUU	Ile	2.7	ACU	Thr	1.2	AAU	Asn	1.6	AGU	Ser	0.7	U
	AUC	Ile	2.7	ACC	Thr	2.4	AAC	Asn	2.6	AGC	Ser	1.5	C
	AUA	Ile	0.4	ACA	Thr	0.1	AAA	Lys	3.8	AGA	Arg	0.2	A
	AUG	Met	2.6	ACG	Thr	1.3	AAG	Lys	1.2	AGG	Arg	0.2	G
G	GUU	Val	2.0	GCU	Ala	1.8	GAU	Asp	3.3	GGU	Gly	2.8	U
	GUC	Val	1.4	GCC	Ala	2.3	GAC	Asp	2.3	GGC	Gly	3.0	C
	GUA	Val	1.2	GCA	Ala	0.1	GAA	Glu	4.4	GGA	Gly	0.7	A
	GUG	Val	2.4	GCG	Ala	3.2	GAG	Glu	1.9	GGG	Gly	0.9	G

[MALOY *et al.*, 1996]

Tabela 3: Valores da Polaridade e Volume Molecular de cada aminoácido.

Aminoácido	Polaridade	Volume molecular
Ala	7	31
Arg	9.1	124
Asp	13	54
Asn	10	56
Cys	4.8	55
Glu	12.5	83
Gln	8.6	85
Gly	7.9	3
His	8.4	96
Ile	4.9	111
Leu	4.9	111
Lys	10.1	119
Met	5.3	105
Phe	5	132
Pro	6.6	32,5
Ser	7.5	32
Thr	6.6	61
Trp	5.2	170
Tyr	5.4	136
Val	5.6	84

(Adaptado de OLIVEIRA, 2014)

Estes três objetivos são utilizados, de acordo com cada abordagem multiobjetiva, para avaliar cada indivíduo da população. Como estamos lidando com um problema multiobjetivo, temos uma função destinada a cada objetivo, descritas abaixo:

$f_1(\mathbf{x})$: dado pela soma da frequência de uso dos códons que codificam os novos aminoácidos no código genético dado por **x** (indivíduo do AG). Com esses dados, adicionamos à função de *fitness* a frequência de cada códon usado pelo novo aminoácido. Para isso, a tabela de frequências de uso dos códons do código genético padrão para o organismo *E. coli* é utilizada (Tabela 2). A penalização é feita de modo a evitar códigos com muitas substituições. Muitas substituições acarretam maior

custo econômico, assim como podem levar a efeitos indesejados do ponto de vista biológico.

$f_2(\mathbf{x})$: dado pelo erro médio quadrático calculado na Eq. (1), considerando-se uma determinada propriedade dos aminoácidos. Aqui, será utilizada a polaridade (Tabela 3). Ressalta-se que, como um novo aminoácido é incorporado ao organismo, a polaridade deste aminoácido é também utilizada na Eq. (1). Ou seja, a Eq. (1) é modificada para incorporar esse novo aminoácido.

$f_3(\mathbf{x})$: dado também pelo erro médio quadrático calculado na Eq. (1), considerando-se uma determinada propriedade dos aminoácidos. Aqui, diferente de $f_2(\mathbf{x})$, será utilizada o volume molecular dos aminoácidos (Tabela 3). Ressalta-se que, como um novo aminoácido é incorporado ao organismo, o volume molecular deste aminoácido é também utilizado na Eq. (1). Ou seja, a Eq. (1) é modificada para incorporar esse novo aminoácido.

A função de avaliação difere para cada abordagem, sendo detalhada a avaliação nas Seções 5.2, 5.3 e 5.4.

4.2 Abordagem Ponderada

Como a abordagem ponderada consiste em atribuir um peso para os objetivos, de acordo com a sua ordem de importância, na metodologia aqui proposta, a função de *fitness* utiliza as funções $f_1(\mathbf{x})$, $f_2(\mathbf{x})$ e $f_3(\mathbf{x})$ para compor a equação 2 abaixo, pela qual o indivíduo é avaliado:

$$f(\mathbf{x}) = w_1 f_1(\mathbf{x}) + w_2 f_2(\mathbf{x}) + w_3 f_3(\mathbf{x}) \quad (3)$$

sendo w_1, w_2 e w_3 os pesos (valores reais positivos normalizados) que ponderam as contribuições dos três objetivos. O problema tratado é de minimização.

4.3 Abordagem Lexicográfica

Nessa metodologia, os objetivos $f_1(\mathbf{x})$, $f_2(\mathbf{x})$ e $f_3(\mathbf{x})$ são ordenados por prioridade, de acordo com a sua importância no resultado.

A função de fitness nessa abordagem calcula o desvio padrão de cada geração para fazer a comparação e os operadores de seleção são codificados para encontrar o melhor indivíduo considerando a prioridade dos objetivos.

Dessa forma, a abordagem lexicográfica compara o mesmo objetivo de dois indivíduos. Suponha, por exemplo, que a robustez em relação à polaridade é o objetivo com maior prioridade. Se o indivíduo 2 apresentar uma polaridade (por exemplo, 4.5) melhor que o indivíduo 1 (que possui polaridade 7) e essa diferença entre as polaridades (2.5) for maior que o desvio padrão da polaridade (suponha que, nesse caso, seja 1.8) da população atual, então temos o indivíduo 2 como a melhor solução.

4.4 Abordagem por Pareto

Tanto na abordagem ponderada quanto na lexicográfica, utilizamos múltiplos objetivos, mas no momento da seleção de indivíduos, o conceito de dominância não é utilizado. Assim, dependendo dos critérios (pesos no caso ponderado e prioridade no caso lexicográfico), apenas uma solução é escolhida como a melhor de uma execução do algoritmo genético. Apenas a abordagem por Pareto trabalha, de fato, com um conjunto de soluções não dominadas. Ao fazer o cálculo, comparamos os indivíduos considerando os vários objetivos e, só farão parte do Conjunto de Pareto aqueles indivíduos não dominados. Para a realização do cálculo multiobjetivo, utilizamos o NSGA-II (Seção 3.2.3). Vamos aqui considerar diferentes combinações dos três objetivos, $f_1(\mathbf{x})$, $f_2(\mathbf{x})$ e $f_3(\mathbf{x})$

5. Experimentos

5.1 Descrição dos experimentos

Nos experimentos descritos a seguir, um novo aminoácido deve ser incorporado ao código genético padrão. Em todos os algoritmos genéticos (nas três abordagens) considera-se uma população de 100 indivíduos. Cada AG é executado 10 vezes, com 1000 gerações em cada execução nas abordagens ponderada e lexicográfica e com 50 gerações[3] na abordagem por Pareto.

Os operadores de seleção por torneio (no qual o melhor de três indivíduos escolhidos aleatoriamente é selecionado) e elitismo são utilizados, assim como *crossover* de dois pontos seguido por mutação binária, com as taxas de 0,6 e 0,01 respectivamente. A maior parte dos parâmetros são iguais àqueles utilizados em [OLIVEIRA, 2015]; testes preliminares também foram feitos para o ajuste de alguns parâmetros. A população inicial é aleatória, assegurando-se que todos os aminoácidos naturais e o novo aminoácido a ser incorporado sejam representados nos indivíduos. Experimentos foram realizados considerando a inserção de um aminoácido hipotético, que teve suas propriedades de polaridade e volume molecular obtidas por meio da média de todos os valores de polaridade e volume molecular dos aminoácidos naturais.

5.1.1 Experimentos: Abordagem Ponderada

Na abordagem ponderada, experimentos foram realizados com o objetivo de testar o impacto dos pesos na Eq. (2), onde quatro algoritmos são considerados. Os pesos são relativos à: f_1 (**x**), que corresponde ao valor de frequência de cada códon; f_2 (**x**) correspondente à robustez segundo a polaridade e f_3 (**x**) correspondente à robustez segundo o volume molecular.

[3] A diferença do número de gerações é decorrente da capacidade do computador utilizada.

No AGP1, $w_1=0$, $w_2=2/3$ e $w_3=1/3$, o que significa que a frequência do código não é levada em consideração e o cálculo da robustez utilizando polaridade tem um maior peso em relação ao mesmo cálculo utilizando o volume.

No AGP2, $w_1=w_2=w_3=1/3$, fazendo com que todos os objetivos tenham o mesmo impacto na avaliação da solução. No AGP3, $w_1=1/3$, $w_2=1/2$ e $w_3=1/6$, sendo que o cálculo de robustez segundo a polaridade tem um maior peso, seguido da frequência do código e da robustez segundo o volume. Finalmente, no AGP4, $w_1=1/2$, $w_2=1/3$ e $w_3=1/6$, o que faz com que a frequência do código tenha um impacto maior na avaliação da solução. Como os valores de fitness dependem dos pesos, aqui mostramos os valores da avaliação para cada objetivo para o melhor indivíduo entre todas as execuções.

5.1.2 Experimentos: Abordagem Lexicográfica

Foram feitas três versões do AG, onde definimos diferentes prioridades. No AGL1, a ordem de prioridade definida foi polaridade, frequência e volume. No AGL2, priorizamos a frequência, seguida da polaridade e do volume, respectivamente. Já no AGL3, consideramos apenas a polaridade e o volume, seguindo essa ordem de importância. Como os valores de fitness dependem da ordem de prioridade, aqui mostramos os valores da avaliação para cada objetivo para o melhor indivíduo entre todas as execuções.

5.1.3 Experimentos: Abordagem por Pareto

Como a abordagem por Pareto é de fato multiobjetivo, não temos uma única solução selecionada: a abordagem traz um conjunto solução com os melhores indivíduos considerando os objetivos utilizados. Após as 10 execuções do NSGA-II, aplica-se novamente o conceito de dominância para determinar o Conjunto de Pareto, ou seja, o Conjunto de Pareto mostrado nos resultados corresponde aos indivíduos não dominados obtidos pela aplicação do conceito de dominância nos conjuntos dos indivíduos não dominados obtidos em cada execução.

Nesse trabalho realizamos quatro experimentos, variando os objetivos considerados pelo NSGA-II: o primeiro utilizando polaridade e frequência, o segundo utilizando polaridade e volume molecular, o terceiro utilizando frequência e volume molecular e, por último, utilizando as três medidas juntas: polaridade, volume molecular e frequência.

5.2 Resultados

5.2.1 Resultados: Abordagem Ponderada

Os resultados sumarizados para as 10 execuções são apresentados na Tabela 4 e no Gráfico 1. O Gráfico 1 mostra que o fitness, que é a soma ponderada da avaliação dos objetivos, é otimizado pelo algoritmo ao longo das gerações.

Gráfico 1: Comportamento do valor do fitness normalizado em cada geração da Execução 1 do AGP 3. Podemos observar o valor de fitness estabilizando na geração 169.

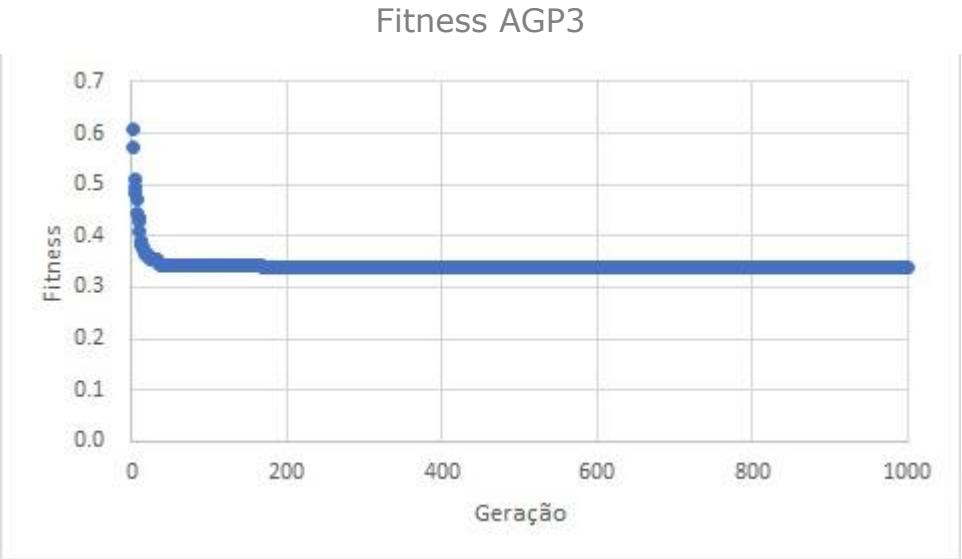


Tabela 4: Avaliação dos melhores indivíduos para os diferentes valores de pesos na Eq. (2): w_1=0, w_2=2/3 e w_3=1/3 (AGP1), w_1=w_2=w_3=1/3 (AGP2), w_1=1/3, w_2=1/2 e w_3=1/6 (AGP3), w_1=1/2, w_2=1/3 e w_3=1/6 (AGP4) . Sendo que f_1 ***(x)*** se refere ao valor da frequência de cada códon, f_2 ***(x)*** à robustez segundo a polaridade e, por último, f_3 ***(x)*** considerando a robustez segundo o volume molecular. Na primeira, segunda e terceira linha exibimos os valores finais das propriedades do novo código. Na quarta linha temos o número de códons que passaram a codificar o novo aminoácido, ou seja, onde houve substituição do aminoácido anterior pelo novo. Finalmente, a última linha mostra o tempo gasto para executar as 1000 gerações das 10 execuções do AGP.

	AGP1	AGP2	AGP3	AGP4
	w_1=0, w_2=2/3 e *w_3=1/3*	*w_1=w_2=w_3=1/3*	*w_1=1/3, w_2=1/2* e *w_3=1/6*	*w_1=1/2, w_2=1/3* e *w_3=1/6*
Polaridade	**2.28494**	2.53071	2.53801	3.37011
Volume	**591.4**	684.4	966.8	1104.7
Frequência	-	43.0	35.8	**24.3**
Substituição	41	38	35	**27**
Tempo (em horas:minutos:segundos)	00:19:00	00:18:05	00:18:22	00:24:00

A Tabela 4 mostra que o novo aminoácido incorporado substituiu a maioria dos aminoácidos naturais quando o AGP1 foi utilizado, afetando drasticamente as frequências de códigos associados a aminoácidos naturais. Já nos experimentos considerando a frequência do novo aminoácido, como no AGP3 e AGP4, um menor número de substituições foi observado, com menos alterações na estrutura do código genético padrão.

Quando a frequência não é considerada no cálculo final (AGP1) ou as três propriedades (volume molecular, frequência e polaridade) são consideradas com a mesma importância, muitas substituições de aminoácidos ocorreram, fazendo com que muitas vezes apenas um códon seja associado com cada um dos aminoácidos naturais. Nesse caso, foi possível, inclusive, obter um código com polaridade melhor que a do próprio código padrão (2.28 e 591.4 do código obtido pelo AGP1 contra 2.63 e 1766.7 do código genético padrão). Porém, realizar muitas substituições é custoso. Também, muitas mudanças não são interessantes devido ao fato de descaracterizarem o código genético natural, e eventualmente modificarem as proteínas codificadas pelos genes do organismo, podendo tornar o código biologicamente inviável. Em experimentos cujos resultados não são mostrados aqui, um menor número de substituições foi observado quando o valor de polaridade do aminoácido hipotético é alto [SILVA; OLIVEIRA; TINOS, 2018].

Quando a frequência dos códons é levada em consideração (AGPs 2, 3 e 4), um número menor de substituições ocorre. Estas substituições ocorrem, em geral, nos códons menos frequentes, o que é bastante interessante do ponto de vista biológico. A Tabela 5 mostra o código genético obtido pelo AGP4, onde nota-se uma grande quantidade de substituições, em geral em aminoácidos menos frequentes; entretanto, vale notar que as substituições visam também aumentar a robustez. Um grande número de códons para um único aminoácido implica em aumentar a robustez pois um eventual erro de tradução em um códon para o novo aminoácido poderia não modificar a tradução, já que muitos códons codificam o mesmo aminoácido. Para o AGP2, a robustez foi menos otimizada do que para o AGP3 e AGP4, já que se utiliza pesos diferentes para esses algoritmos, enquanto o AGP2 atribui o mesmo peso para as três propriedades estudadas. Entretanto, para os melhores códigos obtidos, o AGP3 ocasionou 35 substituições, seguida por 27 substituições do AGP4, ao passo que o AGP2 resultou em 38. Um número menor de substituições poderia ser alcançado aumentando-se ainda mais o peso w_1, i.e, o peso relativo à frequência. Observa-se que, agindo desta forma,

os outros objetivos ficariam piores. Achar os pesos ideais na abordagem ponderada é em geral uma tarefa difícil, pois deve-se buscar um compromisso entre os diferentes objetivos.

Tabela 5: Código genético obtido pelo melhor indivíduo de AGP4 da Abordagem Ponderada. A frequência dos aminoácidos originais do código genético padrão é mostrada na tabela. "New" indica o novo aminoácido.

	U			C			A			G			
U	UUU	Phe	1.9	UCU	New	1.1	UAU	Tyr	1.6	UGU	New	0.4	U
	UUC	Phe	1.8	UCC	New	1.0	UAC	Tyr	1.4	UGC	Cys	0.6	C
	UUA	Leu	1.0	UCA	New	0.7	UAA	Stop	0.2	UGA	Stop	0.1	A
	UUG	Leu	1.1	UCG	New	0.8	UAG	Stop	0.0	UGG	Trp	1.4	G
C	CUU	New	1.0	CCU	Pro	0.7	CAU	His	1.2	CGU	Arg	2.4	U
	CUC	Leu	0.9	CCC	New	0.4	CAC	His	1.1	CGC	Arg	2.2	C
	CUA	New	0.3	CCA	New	0.8	CAA	New	1.3	CGA	New	0.3	A
	CUG	Leu	5.2	CCG	New	2.4	CAG	Gln	2.9	CGG	New	0.5	G
A	AUU	Ile	2.7	ACU	Thr	1.2	AAU	New	1.6	AGU	New	0.7	U
	AUC	Ile	2.7	ACC	Thr	2.4	AAC	Asn	2.6	AGC	Ser	1.5	C
	AUA	New	0.4	ACA	New	0.1	AAA	New	3.8	AGA	New	0.2	A
	AUG	Met	2.6	ACG	Thr	1.3	AAG	Lys	1.2	AGG	New	0.2	G
G	GUU	Val	2.0	GCU	New	1.8	GAU	Asp	3.3	GGU	Gly	2.8	U
	GUC	Val	1.4	GCC	Ala	2.3	GAC	New	2.3	GGC	Gly	3.0	C
	GUA	New	1.2	GCA	New	0.1	GAA	New	4.4	GGA	New	0.7	A
	GUG	Val	2.4	GCG	Ala	3.2	GAG	Glu	1.9	GGG	New	0.9	G

5.2.2 Resultados: Abordagem Lexicográfica

Os resultados do melhor indivíduo para as 10 execuções e 1000 gerações estão mostrados na Tabela 6.

Tabela 6: Avaliação dos melhores indivíduos para os diferentes graus de importância para os AGLs na abordagem lexicográfica. A primeira, segunda e terceira linhas mostram os valores de polaridade, frequência e volume molecular do melhor indivíduo. Na quarta linha, o número de códons que passaram a codificar o novo aminoácido, ou seja, onde houve substituição do aminoácido anterior pelo novo. Finalmente, a última linha mostra o tempo de execução do AGL, considerando as 10 execuções.

	AGL1	AGL2	AGL3
	Polaridade > Frequência > Volume	Frequência > Polaridade > Volume	Polaridade > Volume
Polaridade	3.66949	4.44583	**2.67162**
Volume	1512.8	2188.4	**879.6**
Frequência	24.3	**12.6**	-
Substituições	24	**15**	35
Tempo (em horas:minutos:segundos)	00:14:30	00:10:36	00:09:40

Pela Tabela 6, podemos observar que o AGL2 (Tabela 7) resultou em um menor número de substituições, contando com 15 códons modificados, contra 24 substituições no AGL1 e 35 no AGL3. Como na abordagem ponderada, os algoritmos que não utilizam a frequência trazem um número de substituições bem mais elevado, o que não é interessante para o código genético (conforme ressaltamos na Seção 6.2.1).

Por outro lado, observamos que os valores de polaridade e volume molecular do AGL2 são maiores do que o do CGP, mesmo considerando 1000 gerações. Nesse AGL, obtivemos 4.44 de polaridade e 2188.42 do volume molecular, contra os 2.63 e 1766.7 do CGP. Se a análise fosse feita em cima dos valores do CGP, temos o AGL3 sendo o algoritmo com valores bem próximos para o caso da polaridade e até melhores para o volume molecular. Porém, o elevado número de substituições não torna esse código apropriado para inserção de novos aminoácidos.

Diferentemente da abordagem ponderada, não é possível buscar melhores compromissos entre os objetivos. Na abordagem lexicográfica, apenas a prioridade dos objetivos é levada em conta na otimização.

Tabela 7: Código genético após otimização de AGL2 da Abordagem Lexicográfica

	U			C			A			G			
U	UUU	Phe	1.9	UCU	New	1.1	UAU	Tyr	1.6	UGU	Cys	0.4	U
	UUC	Phe	1.8	UCC	Ser	1.0	UAC	Tyr	1.4	UGC	Cys	0.6	C
	UUA	Leu	1.0	UCA	New	0.7	UAA	Stop	0.2	UGA	Stop	0.1	A
	UUG	Leu	1.1	UCG	Ser	0.8	UAG	Stop	0.03	UGG	Trp	1.4	G
C	CUU	New	1.0	CCU	Pro	0.7	CAU	His	1.2	CGU	New	2.4	U
	CUC	Leu	0.9	CCC	New	0.4	CAC	His	1.1	CGC	Arg	2.2	C
	CUA	New	0.3	CCA	Pro	0.8	CAA	Gln	1.3	CGA	New	0.3	A
	CUG	Leu	5.2	CCG	Pro	2.4	CAG	Gln	2.9	CGG	Arg	0.5	G
A	AUU	Ile	2.7	ACU	Thr	1.2	AAU	Asn	1.6	AGU	Ser	0.7	U
	AUC	Ile	2.7	ACC	Thr	2.4	AAC	Asn	2.6	AGC	Ser	1.5	C
	AUA	New	0.4	ACA	New	0.1	AAA	New	3.8	AGA	New	0.2	A
	AUG	Met	2.6	ACG	Thr	1.3	AAG	Lys	1.2	AGG	New	0.2	G
G	GUU	Val	2.0	GCU	Ala	1.8	GAU	Asp	3.3	GGU	Gly	2.8	U
	GUC	Val	1.4	GCC	Ala	2.3	GAC	New	2.3	GGC	Gly	3.0	C
	GUA	Val	1.2	GCA	New	0.1	GAA	Glu	4.4	GGA	Gly	0.7	A
	GUG	Val	2.4	GCG	Ala	3.2	GAG	New	1.9	GGG	Gly	0.9	G

5.2.3 Resultados: Abordagem por Pareto

Foram considerados 4 AGs na abordagem por Pareto, sendo o primeiro com 3 objetivos e os restantes com 2 objetivos. São eles: AGMO1 (com os 3 objetivos), AGMO2 (com polaridade e volume molecular), AGMO3 (com polaridade e frequência), AGMO4 (com volume molecular e frequência). Os resultados referentes aos conjuntos de Pareto obtidos pelo NSGA-II são apresentados nos Gráficos 2, 3, 4 e 5. Nesta abordagem, o AGMO retorna um conjunto de códigos genéticos que podem ser analisados pelo especialista. A Tabela 8 mostra os valores médios das avaliações dos objetivos para os indivíduos da fronteira de Pareto para cada AGMO. A Tabela 9 mostra os resultados para nove indivíduos obtidos pelo AGMO1. Estes indivíduos são os três que apresentam melhores valores para cada um dos objetivos. Três códigos correspondentes aos melhores indivíduos de cada objetivo são apresentados nas Tabelas 10, 11 e 12.

Gráfico 2: Fronteira de Pareto do AGMO1.

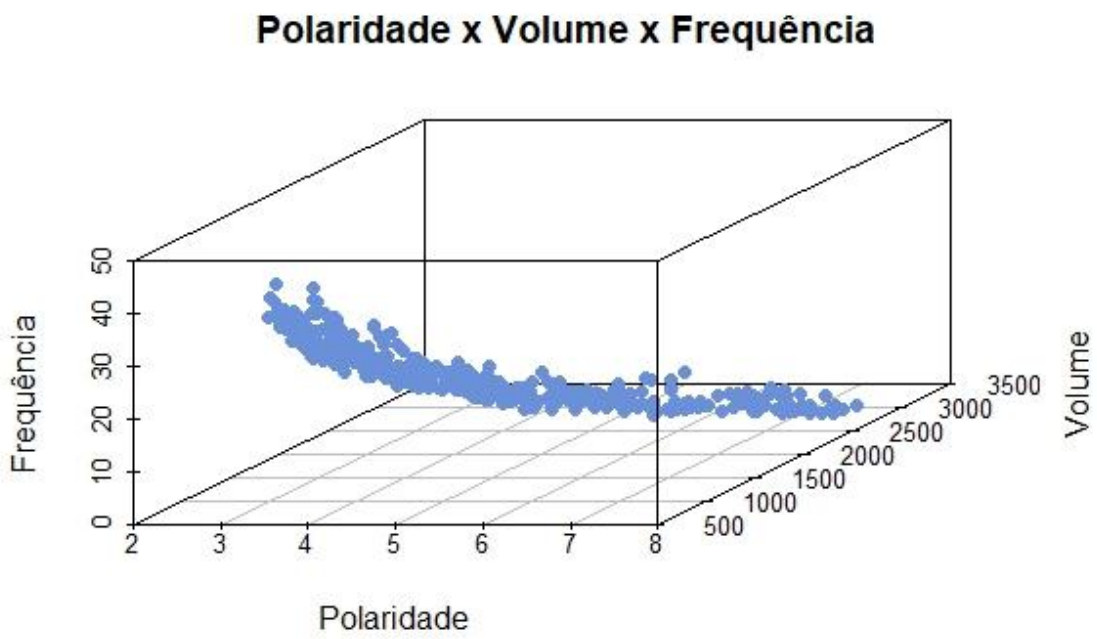


Gráfico 3: Fronteira de Pareto do AGMO2. Nesse caso, foi otimizada a Polaridade e o Volume.

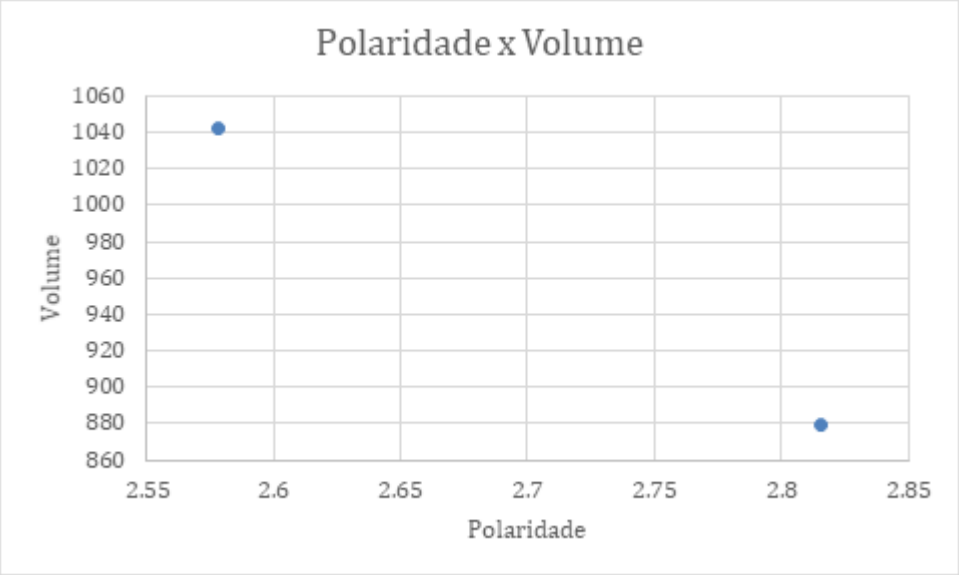


Gráfico 4: Fronteira de Pareto do AGMO3.

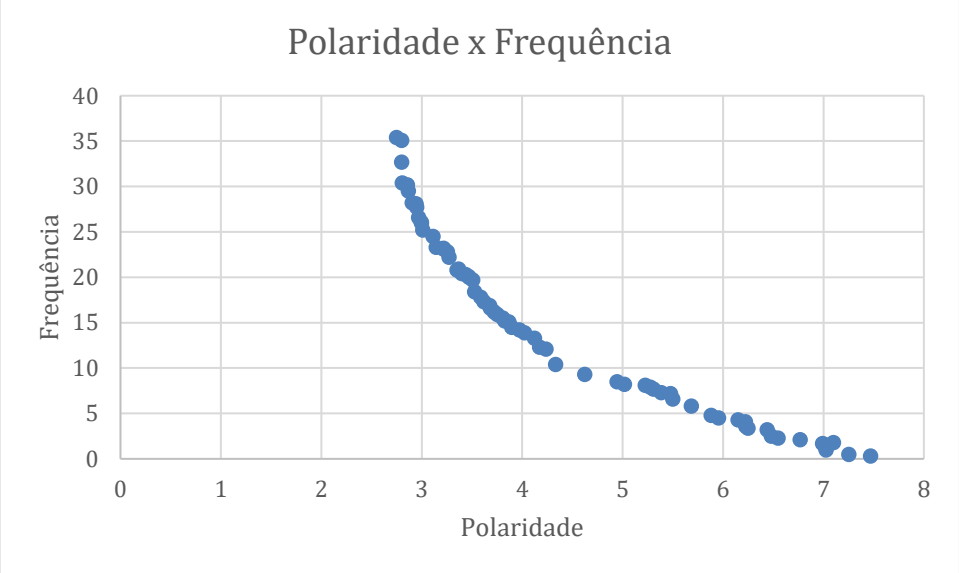


Gráfico 5: Fronteira de Pareto do AGMO4.

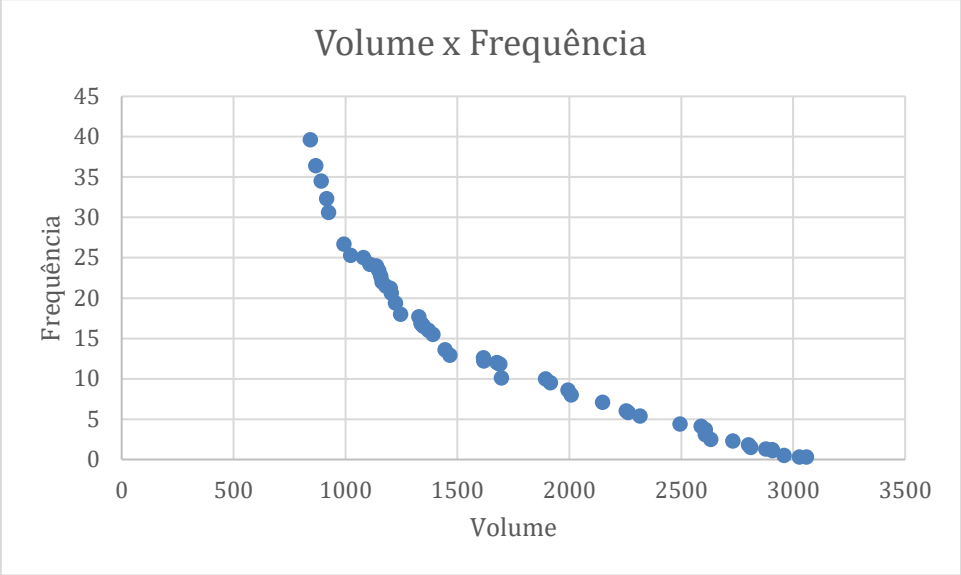


Tabela 8: Resultados médios para os indivíduos localizados na Fronteira de Pareto.

	Polaridade x Volume x Frequência	Polaridade x Volume	Polaridade x Frequência	Volume x Frequência
Média de Polaridade	4.32754	**2.69672**	4,45690	-
Média de Volume Molecular	1762.67	**960.85**	-	1762.21
Média de Frequência	17.62	-	14.82	**14.20**
Média de Substituições	19	33	**16**	**16**
Tempo (em horas:minutos:segundos)	00:32:45	00:24:00	00:26:00	00:25:33

Tabela 9: Amostra de 3 melhores indivíduos para a avaliação cada um dos 3 objetivos para o AGMO1: polaridade, volume molecular e frequência

	Indivíduos com a melhor Polaridade			Indivíduos com o melhor Volume Molecular			Indivíduos com a melhor Frequência		
Polaridade	**2.75074**	2.80132	2.80143	3.15965	3.34447	3.29299	7.47071	7.25174	7.44056
Volume Molecular	1283.59	1630.63	1640.02	**933.60**	953.98	962.47	3028.10	2987.09	2913.61
Frequência	35.4	35.1	32.7	35.9	31.8	30.8	**0.3**	0.5	0.6
Substituições	28	31	30	31	31	30	2	2	3

Tabela 10: Código genético do indivíduo com melhor polaridade otimizado pelo AGMO1 da Abordagem por Pareto.

	U			C			A			G			
U	UUU	Phe	1.9	UCU	New	1.1	UAU	Tyr	1.6	UGU	New	0.4	U
	UUC	Phe	1.8	UCC	New	1.0	UAC	Tyr	1.4	UGC	Cys	0.6	C
	UUA	Leu	1.0	UCA	Ser	0.7	UAA	Stop	0.2	UGA	Stop	0.1	A
	UUG	New	1.1	UCG	New	0.8	UAG	Stop	0.03	UGG	Trp	1.4	G
C	CUU	New	1.0	CCU	Pro	0.7	CAU	His	1.2	CGU	Arg	2.4	U
	CUC	Leu	0.9	CCC	New	0.4	CAC	New	1.1	CGC	Arg	2.2	C
	CUA	Leu	0.3	CCA	New	0.8	CAA	Gln	1.3	CGA	New	0.3	A
	CUG	Leu	5.2	CCG	New	2.4	CAG	Gln	2.9	CGG	New	0.5	G
A	AUU	Ile	2.7	ACU	Thr	1.2	AAU	Asn	1.6	AGU	New	0.7	U
	AUC	Ile	2.7	ACC	Thr	2.4	AAC	New	2.6	AGC	Ser	1.5	C
	AUA	New	0.4	ACA	New	0.1	AAA	New	3.8	AGA	New	0.2	A
	AUG	Met	2.6	ACG	New	1.3	AAG	Lys	1.2	AGG	New	0.2	G
G	GUU	New	2.0	GCU	Ala	1.8	GAU	Asp	3.3	GGU	Gly	2.8	U
	GUC	Val	1.4	GCC	Ala	2.3	GAC	New	2.3	GGC	New	3.0	C
	GUA	Val	1.2	GCA	New	0.1	GAA	New	4.4	GGA	New	0.7	A
	GUG	New	2.4	GCG	Ala	3.2	GAG	Glu	1.9	GGG	New	0.9	G

Tabela 11: Código genético do indivíduo com melhor volume molecular otimizado pelo AGMO1 da Abordagem por Pareto.

	U			C			A			G			
U	UUU	New	1.9	UCU	New	1.1	UAU	Tyr	1.6	UGU	New	0.4	U
	UUC	Phe	1.8	UCC	New	1.0	UAC	Tyr	1.4	UGC	Cys	0.6	C
	UUA	Leu	1.0	UCA	Ser	0.7	UAA	Stop	0.2	UGA	Stop	0.1	A
	UUG	New	1.1	UCG	New	0.8	UAG	Stop	0.03	UGG	Trp	1.4	G
C	CUU	New	1.0	CCU	Pro	0.7	CAU	His	1.2	CGU	Arg	2.4	U
	CUC	Leu	0.9	CCC	New	0.4	CAC	His	1.1	CGC	Arg	2.2	C
	CUA	New	0.3	CCA	New	0.8	CAA	New	1.3	CGA	New	0.3	A
	CUG	Leu	5.2	CCG	New	2.4	CAG	Gln	2.9	CGG	New	0.5	G
A	AUU	Ile	2.7	ACU	Thr	1.2	AAU	Asn	1.6	AGU	New	0.7	U
	AUC	Ile	2.7	ACC	Thr	2.4	AAC	New	2.6	AGC	Ser	1.5	C
	AUA	New	0.4	ACA	New	0.1	AAA	Lys	3.8	AGA	New	0.2	A
	AUG	Met	2.6	ACG	Thr	1.3	AAG	New	1.2	AGG	New	0.2	G
G	GUU	New	2.0	GCU	New	1.8	GAU	Asp	3.3	GGU	New	2.8	U
	GUC	Val	1.4	GCC	New	2.3	GAC	New	2.3	GGC	Gly	3.0	C
	GUA	Val	1.2	GCA	New	0.1	GAA	Glu	4.4	GGA	New	0.7	A
	GUG	New	2.4	GCG	Ala	3.2	GAG	New	1.9	GGG	New	0.9	G

Tabela 12: Código genético do indivíduo com melhor frequência otimizado pelo AGMO1 da Abordagem por Pareto.

	U			C			A			G			
U	UUU	Phe	1.9	UCU	Ser	1.1	UAU	Tyr	1.6	UGU	Cys	0.4	U
	UUC	Phe	1.8	UCC	Ser	1.0	UAC	Tyr	1.4	UGC	Cys	0.6	C
	UUA	Leu	1.0	UCA	Ser	0.7	UAA	Stop	0.2	UGA	Stop	0.1	A
	UUG	Leu	1.1	UCG	Ser	0.8	UAG	Stop	0.03	UGG	Trp	1.4	G
C	CUU	Leu	1.0	CCU	Pro	0.7	CAU	His	1.2	CGU	Arg	2.4	U
	CUC	Leu	0.9	CCC	Pro	0.4	CAC	His	1.1	CGC	Arg	2.2	C
	CUA	Leu	0.3	CCA	Pro	0.8	CAA	Gln	1.3	CGA	Arg	0.3	A
	CUG	Leu	5.2	CCG	Pro	2.4	CAG	Gln	2.9	CGG	Arg	0.5	G
A	AUU	Ile	2.7	ACU	Thr	1.2	AAU	Asn	1.6	AGU	Ser	0.7	U
	AUC	Ile	2.7	ACC	Thr	2.4	AAC	Asn	2.6	AGC	Ser	1.5	C
	AUA	Ile	0.4	ACA	New	0.1	AAA	Lys	3.8	AGA	Arg	0.2	A
	AUG	Met	2.6	ACG	Thr	1.3	AAG	Lys	1.2	AGG	New	0.2	G
G	GUU	Val	2.0	GCU	Ala	1.8	GAU	Asp	3.3	GGU	Gly	2.8	U
	GUC	Val	1.4	GCC	Ala	2.3	GAC	Asp	2.3	GGC	Gly	3.0	C
	GUA	Val	1.2	GCA	Ala	0.1	GAA	Glu	4.4	GGA	Gly	0.7	A
	GUG	Val	2.4	GCG	Ala	3.2	GAG	Glu	1.9	GGG	Gly	0.9	G

Na abordagem por Pareto, não conseguimos definir um melhor indivíduo: por exemplo, para o AGMO1, selecionamos 544 indivíduos, com valores de polaridade, volume molecular e frequência igualmente bons. Na tabela 8 podemos observar que, a média dos valores dos 4 experimentos variam bastante e se assemelham aos experimentos realizados nas outras abordagens. Na tabela 9, ao fazer a comparação com os melhores indivíduos de cada objetivo do AGMO1, essa semelhança fica ainda mais clara. Quando temos um valor de polaridade e volume molecular mais baixo, temos uma alta quantidade de substituições no código e, quando reduzimos muito as substituições, os valores de polaridade e volume molecular se elevam.

5.3 Comparação das Abordagens

Os resultados obtidos mostram que a abordagem multiobjetivo é interessante por resultar em uma lista de códigos otimizados. Esta lista pode então ser analisada pelo especialista. As três técnicas utilizam estratégias diferentes para se obter a lista de códigos otimizados.

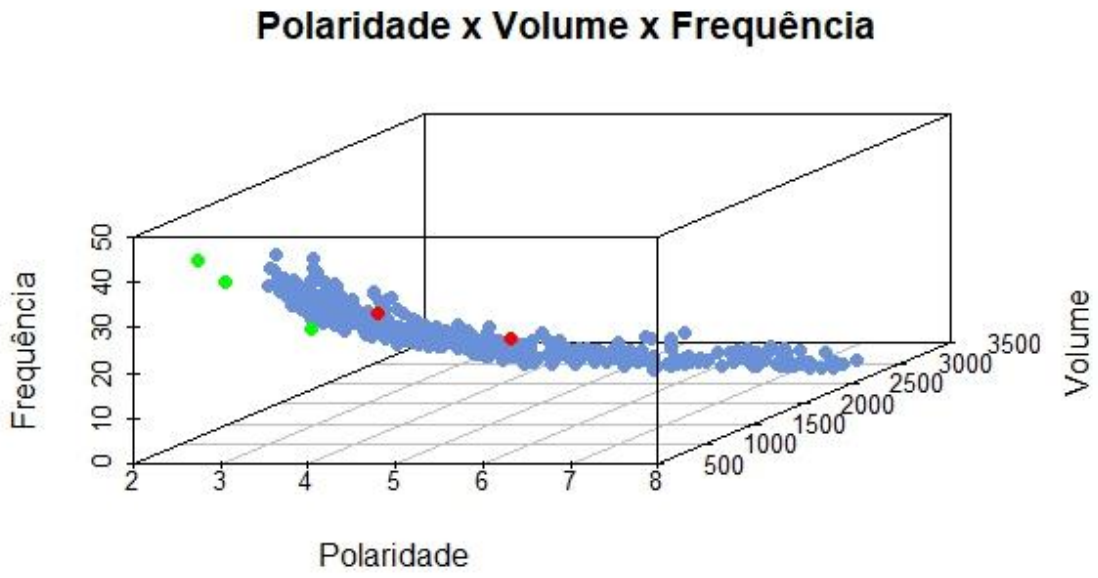
Podemos observar que, na abordagem ponderada, o melhor indivíduo quanto ao número de substituições foi obtido no AGP4, que é o algoritmo que utiliza o peso maior na frequência, seguido de polaridade e volume molecular, respectivamente. Na abordagem lexicográfica, também encontramos um resultado semelhante no AGL2, cuja prioridade

principal é a frequência. As semelhanças entre as duas abordagens se estendem para os demais AGs, onde o número de substituições é semelhante (41 contra 35, quando a frequência não é utilizada; 35 e 24, quando a polaridade tem mais importância; e 27 e 15 quando a frequência é o objetivo mais relevante). Os valores de polaridade e volume molecular são bem baixos quando não considerada a frequência dos códons no cálculo, sendo próximos ou até superior ao do código genético padrão (na abordagem Ponderada, alcançamos 2.28 de polaridade e 591.4 de volume molecular; na Lexicográfica, 2.67 e 879.6, respectivamente. Os valores do CGP são 2.63 para a polaridade e 1766.7 para o volume molecular). Apesar da semelhança entre as duas abordagens, a abordagem ponderada permite um maior grau de liberdade ao possibilitar combinar os diferentes pesos para cada objetivo. Na abordagem lexicográfica, apenas a prioridade dos objetivos é definida, sem que possamos ponderar o impacto de cada um na otimização. Entretanto, vale ressaltar que, atribuir pesos que irão resultar em códigos interessantes não é uma tarefa simples.

Quando comparamos as outras abordagens com a abordagem por Pareto, podemos observar que o padrão se mantém: quando temos apenas volume molecular e polaridade, os valores médios são baixos e muito próximos do CGP (2.69 para a polaridade e 960 para o volume molecular) e um valor muito alto de substituições (33 em média). Quando a frequência é considerada, os valores aumentam um pouco, reduzindo bastante o número de substituições (nos AG 3 e 4, a média foi de 16, mas códigos com apenas 1 substituição foram obtidos). Quando os três objetivos são calculados juntos no AGMO1, podemos observar uma ampla variedade de soluções: os valores médios foram com polaridade de 4.32, volume molecular de 1762.67, frequência de 17.62 e substituições em torno de 19. Os resultados médios se assemelham ao AGMO2 da abordagem lexicográfica. Entretanto, vale ressaltar, que foi possível encontrar códigos com valores bastante baixos para a avaliação dos objetivos.

A abordagem por Pareto mostrou-se ser interessante por não necessitar definir os pesos ou a prioridade para cada objetivo. Além disso, conforme mostrado no Gráfico 6, os melhores resultados dos AGPs e AGLs fazem parte, junto com os indivíduos de AGPO, da Fronteira de Pareto.

Gráfico 6: Comparativo dos resultados entre AGP (em verde), AGL (em vermelho) e AGMO (em azul).



6. Conclusão

Esse trabalho simulou a incorporação da codificação de um novo aminoácido no código genético padrão. As frequências dos aminoácidos utilizadas no primeiro objetivo foram para o organismo *E. coli*. Otimizar a frequência significa buscar códons que são menos utilizados para a substituição do novo aminoácido. Além da frequência, levou-se em consideração a robustez do código calculada considerando-se a polaridade e volume molecular dos aminoácidos. A fim de auxiliar na busca, propôs-se o uso de AGs multiobjetivo para otimizar os códigos genéticos modificados. Os AGs utilizam, portanto, avaliações individuais para os três objetivos: robustez utilizando a propriedade polaridade, robustez utilizando a propriedade volume molecular e a frequência de uso dos códons. Ressalta-se ainda que, de acordo com o conhecimento dos autores, este é o primeiro trabalho que utiliza AGs para a otimização do código genético expandido.

Três abordagens foram utilizadas para realizar os experimentos. A primeira foi a abordagem ponderada, onde foram considerados quatro combinações diferentes dos pesos para cada objetivo: o primeiro deles considerando apenas a otimização da polaridade e volume molecular do código, o segundo considerando a polaridade, volume molecular e a frequência do uso de códons com pesos iguais, o terceiro atribuindo um peso maior a polaridade e o quarto atribuindo um peso maior para a frequência. A segunda abordagem multiobjetivo foi a lexicográfica, na qual foram consideradas três ordens de importância: uma considerando a polaridade e o volume molecular, a segunda considerando a polaridade, volume molecular e a frequência e uma terceira utilizando a frequência, polaridade e volume molecular. Por último, foi considerada a abordagem por Pareto, considerando quatro combinações dos objetivos: polaridade e frequência, polaridade e volume molecular, volume molecular e frequência e os três objetivos juntos.

Os resultados obtidos mostraram que, quando consideramos apenas a polaridade e volume molecular no cálculo de fitness, códigos bastante robustos são obtidos pelo AG; em alguns casos, mais robustos que o próprio código padrão. Entretanto, muitos aminoácidos são substituídos, o que pode descaracterizar o código genético, além de alterar drasticamente a frequência de códons associados a aminoácidos, o que provavelmente resulta em códigos inviáveis do ponto de vista biológico.

Contudo, quando a frequência dos códons é utilizada, menos substituições ocorrem. Observa-se que é possível obter mais ou menos substituições, otimizando também a robustez e preservando a estrutura geral do código genético. Podemos perceber que, a utilização das três abordagens multiobjetivo permite obter uma lista de códigos expandidos otimizados, fornecendo ao especialista liberdade para a escolha do mais interessante de acordo com a aplicação. Neste intuito, a abordagem que se mostrou mais interessante foi a por Pareto, que possibilitou obter uma lista razoavelmente grande de códigos, sem que pesos ou prioridades precisem ser definidas. Como desvantagem, a lista pode ser ampla demais, necessitando outros critérios para a escolha do melhor código. Esse resultado pode ser conveniente para a criação de novos organismos geneticamente modificados e para a produção de proteínas de interesse, uma vez que códigos genéticos mais similares ao padrão foram produzidos com sucesso. Entretanto, vale ressaltar que, este foi um trabalho teórico, considerando aminoácidos hipotéticos e sem levar em conta outras restrições que podem ocorrer do ponto de vista experimental e biológico.

Um possível trabalho futuro, do ponto de vista biológico, é investigar a introdução dos novos aminoácidos por meio da criação de nucleotídeos sintéticos [ZHANG et al., 2017]. Neste caso, o código genético padrão não é modificado; ele apenas é expandido para acomodar os novos códons relacionados aos novos nucleotídeos sintéticos. Por exemplo, supondo que um nucleotídeo sintético *Y* seja criado, além dos códons naturais, teríamos a possibilidade de associar aos novos aminoácidos os novos códons que possuem *Y* em sua constituição, i.e., *AAY*, *ACY*,..., *AYA*,...*YGG*. Usualmente, não se associa todos os novos códons aos aminoácidos. Neste caso, a otimização via AGs mostra-se uma abordagem bastante promissora. Um possível trabalho futuro, do ponto de vista tecnológico, seria utilizar outros algoritmos no cálculo do Conjunto de Pareto, como o SPEA-II (*Strength Pareto Evolutionary Algorithm* 2), que assim como o NSGA-II é um método de otimização baseado em evolução natural para problemas multiobjetivos. Finalmente, um possível trabalho futuro é investigar novos objetivos e novas comparações de algoritmos evolutivos [ZIZTLER; THIELE, 1999] que possam ser interessantes do ponto de vista experimental, técnico e biológico.

Referências Bibliográficas

ALBERTS, B. *et al. (2002).* **Molecular biology of the cell.** *5*th ed., Artmed.

ANDERSON, J. C. *et al*. (2004). An expanded genetic code with a functional quadruplet códon. **PNAS**, 101(20): 7566-7571, 2004.

EL-GHAZALI, T. (2009). **Metaheuristics:** from design to implementation. John Wiley and Sons Inc., Chichester.

FREELAND, S. J. & HURST, L. D. (1998). The genetic code is one in a million. **Journal of Molecular Evolution**, 47(3): 238–248.

FREITAS A. A. (2004). A critical review of multi-objective optimisation in data mining: a position paper. **ACM SIGKDD Explorations,** 6: 77-86

HAIG, D. & HURST, L. D. (1991). A quantitative measure of error minimization in the genetic code. **Journal of Molecular Evolution**, 33: 412–417.

LAJOIE, M. J.; SÖLL, D.; CHURCH, G. M. (2016). Overcoming challenges in engineering the genetic code. **Journal of molecular biology**, *428*(5), 1004-1021.

LEHNINGER, A. L.; NELSON, D. L. & COX, M. M. (2005). **Lehninger Principles Of Biochemistry.** 4th ed., Freeman.

LINDEN, R. (2008). **Algoritmos Genéticos:** Uma importante ferramenta da inteligência computacional. 2ª ed, Brasport.

LIU, C. C.; SCHULTZ, P. G. (2010). Adding new chemistries to the genetic code. **Annual Review of Biochemistry**, *79*: 413-444.

MALOY, S. R.; STEWART, V.J.; TAYLOR, R.K. (1996). **Genetic analysis of pathogenic bacteria**: a laboratory manual. Plainville, USA: Cold Spring Harbor Laboratory Press.

MITCHELL, M. (1996). **An introduction to genetic algorithms**, MIT Press.

OLIVEIRA, L. L. (2015). **Algoritmos Evolutivos Aplicados na Investigação da Adaptabilidade do Código Genético.** Tese de Doutorado, Pós-Graduação em Bioinformática, Universidade de São Paulo.

OLIVEIRA, L. L.; TINÓS, R. (2014). Entropy-based evaluation function in a multiobjective approach for the investigation of genetic code robustness. **Memetic Computing**, 6: 157-170.

OLIVEIRA, L. L.; OLIVEIRA, P. S. L.; TINÓS, R. (2015). A multiobjective approach to the genetic code adaptability problem. **BMC Bioinformatics**, 16(52).

OLIVEIRA, L. L.; FREITAS, A. A.; TINÓS, R. (2017). Multi-objective genetic algorithms in the study of the genetic code's adaptability. **Information Sciences**, 425: 48-61

ROVNER, A. J. *et al*. (2015). Recoded organisms engineered to depend on synthetic amino acids. **Nature,** 518: 89:93.

SANTOS, J.; MONTEAGUDO, Á. Simulated evolution applied to study the genetic code optimality using a model of codon reassignments. **BMC bioinformatics**, BioMed Central, v. 12, n. 1, p. 56, 2011.

SILVA, M. C.; DE OLIVEIRA, L. L.; TINÓS, R. Optimization of Expanded Genetic Codes via Genetic Algorithms. *In*: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 15. , 2018, São Paulo. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2018 . p. 473-484. DOI: https://doi.org/10.5753/eniac.2018.4440.

VOGEL, G. (1998). Tracking the history of the genetic code. **Science**, 281: 329-331.

XIAO, H.; SCHULTZ, P. G. (2016). At the interface of chemical and biological synthesis: an expanded genetic code. **Cold Spring Harbor Perspectives in Biology**, *8*(9): a023945.

YOCKEY, H. P. (2005). **Information Theory, Evolution, and the Origin of Life**, Cambridge University Press, NY.

ZHANG, Y. *et al*. (2017). A semi-synthetic organism that stores and retrieves increased genetic information. ***Nature,*** 551(7682): 644.

ZITZLER, E.; THIELE, L. (1999). Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach. **IEEE Transactions on Evolutionary Computation**, v. 3, n. 4, p. 257-271.

WATSON, J. *et al*. (2015). **Biologia Molecular do Gene,** 7ª ed., Artmed.