Optimization of Expanded Genetic Codes via Genetic Algorithms

Maísa de Carvalho Silva, Lariza Laura de Oliveira, Renato Tinós

Departamento de Computação e Matemática – Universidade de São Paulo (USP) – Ribeirão Preto – SP – Brasil

maisa.carvalho.silva@usp.br, larizalaura@gmail.com, rtinos@ffclrp.usp.br

Abstract. *In the last decades, researchers have proposed the use of genetically modified organisms that utilize 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 engineered organisms for the development of new drugs, fuels and chemicals. When new amino acids are incorporated, it is necessary to modify the standard genetic code. Expanded genetic codes have been created without considering the robustness of the code. The objective of this work is the use of genetic algorithms (GAs) for the optimization of expanded genetic codes. The GA indicates which codons of the standard genetic code should be used to encode a new unnatural amino acid. The fitness function has two terms; one for robustness of the new code and another that takes into account the frequency of use of amino acids. Experiments show that, by controlling the weighting between the two terms, it is possible to obtain more or less amino acid substitutions at the same time that the robustness is minimized.*

1. Introdução

Proteínas são macromoléculas vitais para os organismos vivos, desempenhando diferentes funções biológicas. Proteínas são essenciais em tarefas relacionadas à catálise, ao transporte, ao armazenamento, à motilidade, à defesa e à regulação [LEHNINGER *et al*., 2005]. Proteínas são compostas por aminoácidos unidos por ligações covalentes, formando séries com diferentes tamanhos e constituições. Alterações na sequência de aminoácidos de uma proteína ocasionam geralmente mudanças em sua estrutura tridimensional e também em sua função.

Cada aminoácido é codificado no DNA (*ácido desoxirribonucléico*) através de uma sequência de três nucleotídeos chamada códon. Sessenta e um códons especificam aminoácidos e três códons indicam o término da sequência 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 compartilha o mesmo código genético, sendo observadas raras exceções [VOGEL, 1998]. Além disso, as exceções observadas

são mínimas, fazendo com que os códigos difiram muito pouco do código genético mais frequentemente encontrado na maioria dos seres vivos. Este código genético é conhecido como código genético padrão, sendo sua organização apresentada na Figura 1.

Primeira base		Segunda base				Terceira base
		U	C	A	G	
	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

FIGURA 1. Código genético padrão [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 fosse fruto do acaso, então qualquer código genético, entre os cerca de 1,4 x 10^{70} códigos possíveis, poderia ter sido selecionado [YOCKEY, 2005]. Alguns pesquisadores têm sugerido que o código genético padrão evoluiu para sua organização atual de tal forma a torná-lo mais robusto frente a mutações [VOGEL 1998]. De fato, quando examinamos a organização do código padrão, podemos observar que aminoácidos são codificados por códons similares (Figura 1). Ou seja, muitas das 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 extremamente robusto, evitando e prevenindo falhas, uma vez que códons similares tendem a codificarem aminoácidos com propriedades semelhantes. Freeland e Hurst (1998) observaram que o código padrão é mais robusto que um número muito grande de códigos hipotéticos gerados aleatoriamente, mais precisamente que o código padrão é 1 em 1 milhão. Contudo, tanto nos trabalhos subsequentes de Freeland, quanto nos de outros autores, somente medidas de robustez do código são levadas em consideração para comparar os diferentes códigos.

Em [OLIVEIRA, 2015], propõe-se utilizar mais de uma medida simultaneamente para comparar diferentes códigos genéticos. Utilizam-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 uma 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], utiliza-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 facilitam alguns estudos cristalográficos envolvendo raio X. Aminoácidos não-naturais têm ainda 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]. Atualmente, de acordo com o conhecimento dos autores, as modificações nos códigos genéticos ocorrem de forma manual, sem que a robustez do novo código seja avaliada.

O objetivo principal deste trabalho é a investigação do uso de algoritmos genéticos 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 visa o desenvolvimento de códigos mais robustos.

Códigos genéticos expandidos são discutidos na Seção 2 deste trabalho. A metodologia proposta é apresentada na Seção 3. Os resultados experimentais e as conclusões são apresentados respectivamente nas seções 4 e 5.

2. Códigos Genéticos Expandidos

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 & 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 (*ácido ribonucléico*) transportador associadas ao *UAG* para que codifiquem o novo aminoácido.

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:

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

sendo $X(i,\mathbf{x})$ a propriedade do aminoácido codificado pelo i-ésimo códon do código **x**, $w(i,j)$ a ponderação correspondente à posição do nucleotídeo no códon (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 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. Metodologia

Uma maneira de inserir novos aminoácidos é substituir códons que raramente são utilizados [ROVNER *et al*., 2015]. Para tanto, é necessário conhecer a distribuição do uso dos diferentes códons em um determinado organismo geneticamente modificado. Entretanto, como dito anteriormente, não se leva em consideração a robustez do novo código no desenvolvimento deste 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, a função de *fitness* é composta por dois termos:

$f_1(\mathbf{x})$: 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 (Figura 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, é utilizada a polaridade dos aminoácidos. Ressalta-se que, como novos aminoácidos são incorporados ao organismo, a polaridade destes aminoácidos é também utilizada na Eq. (1). Ou seja, a Eq. (1) é modificada para incorporar os novos aminoácidos.

Ao utilizar dois objetivos, o problema de otimização torna-se multi-objetivo. Existem três abordagens principais para tratar problemas multi-objetivos utilizando-se AGs [FREITAS, 2004]: abordagem ponderada, lexicográfica ou por Pareto. Neste artigo, propomos utilizar a primeira delas. Outras deverão ser exploradas no futuro. Na abordagem ponderada utilizada aqui, o indivíduo do AG (código genético expandido) é avaliado por:

$$f(\mathbf{x}) = w_1 f_1(\mathbf{x}) + w_2 f_2(\mathbf{x}) \qquad (2)$$

sendo w_1 e w_2 os pesos (valores reais positivos ou iguais a zero) que ponderam as contribuições dos dois objetivos. O problema tratado é de minimização.

	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

FIGURA 2. Tabela de frequências de uso de códons na *E. coli* [MALOY *et al*., 1996].

Os códigos expandidos são representados no AG por meio de um vetor binário. A Figura 3 mostra um exemplo da decodificação do indivíduo do AG no código expandido. Um elemento igual a 1 associado a um códon no vetor significa que o códon correspondente no código genético padrão será relacionado agora ao novo aminoácido. A representação é binária, sendo que operadores de mutação e reprodução para codificação binária são utilizados [MITCHELL, 1996]. 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 adicionando-se um valor de 10.000 ao seu *fitness*.

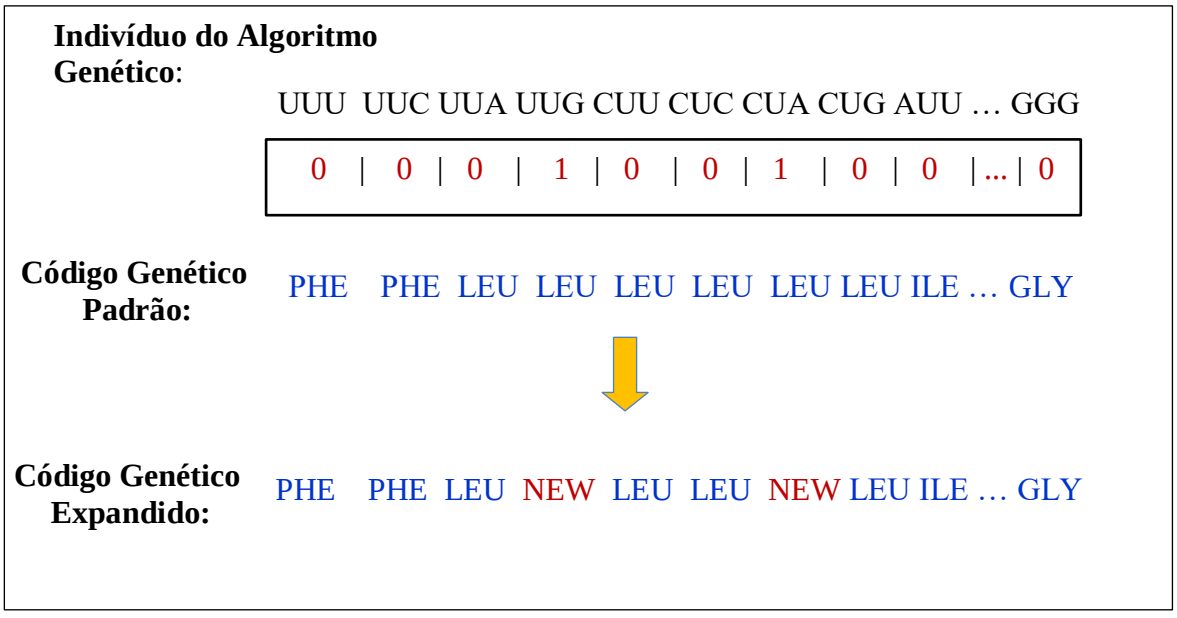

FIGURA 3. Representação utilizada no primeiro método. *NEW* representa o novo aminoácido introduzido.

4. Resultados

Nos experimentos descritos a seguir, um novo aminoácido deve ser incorporado ao código genético padrão. O AG padrão utilizado na otimização do código foi empregado considerando-se uma população de 100 indivíduos. O AG é executado 10 vezes, com 50 gerações em cada execução. Os operadores de seleção por torneio e elitismo são utilizados, assim como *crossover* de dois pontos seguido por mutação binária, com as taxas de 0,6 e 0,01 respectivamente. Testes preliminares foram realizados para se determinar os melhores parâmetros do AG. 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-se três valores hipotéticos de polaridade (baixa, média e alta) para o novo aminoácido. Com o objetivo de testar o impacto dos pesos na Eq. (2), três algoritmos são considerados. No AG 1, $w_1=0$ e $w_2=1$, o que significa que apenas a robustez do código é levada em consideração. No AG 2, $w_1=w_2=1$, fazendo com que frequência de uso dos códons e robustez tenham o mesmo impacto na avaliação da solução. Finalmente, no AG 3, $w_1=1$ *e* $w_2=5$, o que faz com que a robustez tenha um impacto muito maior na avaliação da solução.

Os resultados para as 10 execuções são apresentados na Tabela 1. Os melhores códigos obtidos para cada algoritmo quando a polaridade média é considerada são apresentados nas Figuras 4, 5 e 6.

TABELA 1. Avaliação dos indivíduos para os diferentes valores de pesos na Eq. (2): w_1=0 e w_2=1 (AG 1), w_1= w_2=1 (AG 2), w_1=1 e w_2=5 (AG 3) . Experimentos foram realizados considerando-se três valores para a polaridade do aminoácido a ser incorporado: 4,9 (baixa), 7,4 (média) e 13,0 (alta). A primeira linha mostra o melhor resultado das 10 execuções. A segunda e a terceira linhas mostram respectivamente a média e o desvio padrão dos melhores indivíduos obtidos nas 10 execuções. As duas linhas seguintes mostram os valores de cada termo na Eq. (2) para o melhor indivíduo obtido pelo AG. Finalmente, a última linha mostra o número de substituições (de um aminoácido natural para o novo aminoácido) contidas no melhor indivíduo.

	AG1			AG2			AG3		
	7.4	13.0	4.9	7.4	13.0	4.9	7.4	13.0	4.9
Melhor Indivíduo	2,27122	7,21778	3,52858	7,62896	8,28851	7,77836	31,4253	38,9917	34,2392
Média Indivíduos	2,642739	7,220144	4,009634	8,3407	9,37154	8,5273	32,0695	42,9644	35,208
Desvio Padrão	0,854021	0,016883	0,813134	3,16098	3,58773	3,1855	5,76968	2,3	6,16912
Robustez	2,27122	7,21778	3,52858	7,52896	8,18851	7,57836	4,68505	7,33835	5,76785
Frequência	-	-	-	0,1	0,1	0,2	8	2,3	5,4
nº Substituições	41	2	37	1	1	1	11	1	3

	U		C		A		G		
U	UUU	Phe	UCU	New	UAU	New	UGU	Cys	U
	UUC	New	UCC	New	UAC	Tyr	UGC	New	C
	UUA	New	UCA	New	UAA	Stop	UGA	Stop	A
	UUG	New	UCG	New	UAG	Stop	UGG	Trp	G
C	CUU	New	CCU	Pro	CAU	New	CGU	New	U
	CUC	Leu	CCC	New	CAC	His	CGC	New	C
	CUA	New	CCA	New	CAA	Gln	CGA	New	A
	CUG	New	CCG	New	CAG	New	CGG	New	G
A	AUU	Ile	ACU	New	AAU	New	AGU	New	U
	AUC	New	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	New	ACA	New	AAA	New	AGA	Arg	A
	AUG	Met	ACG	New	AAG	Lys	AGG	New	G
G	GUU	New	GCU	New	GAU	Asp	GGU	Gly	U
	GUC	New	GCC	Ala	GAC	New	GGC	New	C
	GUA	Val	GCA	New	GAA	Glu	GGA	New	A
	GUG	New	GCG	New	GAG	New	GGG	New	G

FIGURA 4. Melhor solução (código expandido) obtida pelo AG 1 (w_1=0 e w_2=1) para o novo aminoácido com polaridade média (7,4).

	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	New	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

FIGURA 5. Melhor solução (código expandido) obtida pelo AG 2 (w_1=1 e w_2=1) para o novo aminoácido com polaridade média (7,4).

	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	New	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	New	CCA	Pro	CAA	Gln	CGA	New	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	New	ACA	New	AAA	New	AGA	New	A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	New	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	New	GGC	Gly	C
	GUA	Val	GCA	New	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	New	GGG	Gly	G

FIGURA 6. Melhor solução (código expandido) obtida pelo AG 3 (w_1=1 e w_2=5) para o novo aminoácido com polaridade média (7,4).

De acordo com a Tabela 1, os melhores indivíduos obtidos pelos experimentos com o AG 1, AG 2 e AG 3 foram obtidos para a polaridade de 7,4, sendo que o menor valor de robustez foi obtido para o AG 1, que considera apenas a robustez. Entretanto, a Figura 4 mostra que o novo aminoácido incorporado substituiu a maioria dos aminoácidos naturais quando o AG 1 foi utilizado, afetando drasticamente as frequências de códigos associados a aminoácidos naturais. Já nos experimentos considerando o AG 2 e AG 3, um menor número de substituições foi observado, com destaque para o indivíduo da Figura 5 que alterou muito pouco a estrutura do código genético padrão.

Quando apenas a robustez é considerada (AG 1), muitas substituições de aminoácidos ocorreram, fazendo com que muitas vezes apenas um códon seja associado com cada um dos aminoácidos naturais (Figura 1). Isso ocorreu principalmente para os valores de polaridade 7,4 e 4,9. Nesse caso, foi possível, inclusive, obter um código com robustez melhor que a do próprio código padrão (2,27122 do código obtido pelo AG 1 contra 2,62835 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. Um menor número de substituições foi observado quando o valor de polaridade do aminoácido foi igual a 13. Isso se deve ao fato de que dentro dos 20 aminoácidos essenciais, o valor de polaridade 13 é o maior valor observado. Dessa forma, a fim de minimizar a função da robustez, o algoritmo tenderá a aumentar a frequência de aminoácidos mais próximos ao valor médio da polaridade, que seria em torno de 7. Este fato explica o porquê de o melhor resultado para robustez ter sido alcançado para o aminoácido com polaridade média (4,9).

Quando a frequência dos códons é levada em consideração (AGs 2 e 3), 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. Para o AG 2, a

robustez foi menos otimizada do que para o AG 3, já que utiliza-se w_1=1 e w_2=5. Entretanto, para os melhores códigos obtidos, considerando-se um valor de polaridade média (7,4) para o novo aminoácido, o AG 3 ocasionou 11 substituições, ao passo que o AG 2 resultou em apenas uma. Vale ressaltar que para o AG 2, o códon modificado (*GCA*) é um dos com menor frequência na *E. coli* (Figura 2). Além disso, observando os outros valores de polaridade considerados (4,9 e 13), o AG 2 também obteve soluções com poucas substituições e em códons de baixa frequência, o que é importante, uma vez que pode ser interessante incorporar aminoácidos com valor de polaridade diferentes da média.

5. Conclusões

Esse trabalho simulou a incorporação de aminoácidos modificados no código genético padrão. O organismo escolhido foi a *E. coli*. Para tanto, levou-se em consideração a robustez do código calculada considerando-se a polaridade dos aminoácidos. Além disso, utilizou-se a frequência de ocorrência dos códons no organismo de estudo. A fim de auxiliar na busca, propôs-se o uso de AGs para otimizar os códigos genéticos modificados. O AG utiliza uma função de *fitness* composta pelos dois termos acima mencionados: robustez e 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 experimentos foram realizados: o primeiro deles considerando apenas a otimização da robustez do código, o segundo considerando a robustez e a frequência do uso de códons com pesos iguais e o terceiro atribuindo um peso maior à robustez. Os resultados obtidos mostraram que, quando apenas a robustez é otimizada, 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, controlando a ponderação entre os dois termos, é possível obter mais ou menos substituições, otimizando também a robustez e preservando a estrutura geral do código genético. 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.

Vale ressaltar que a abordagem ponderada foi utilizada para o problema multi-objetivo. Assim, uma continuação natural para a abordagem proposta é utilizar outras abordagens multi-objetivo, como a abordagem por Pareto e a abordagem lexicográfica. Outro possível trabalho futuro 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.

6. Agradecimentos

Os autores agradecem à FAPESP (Proc. 2011/00561-7 e 2015/06462-1) e ao CNPq pelo apoio financeiro.

Referências Bibliográficas

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

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.

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

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",* 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.

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.