UNIVERSIDADE FEDERAL DE UBERLÂNDIA

CAROLINA ALVES DA SILVA

UMA ANÁLISE DE PARÂMETROS DO ALGORITMO EVOLUTIVO VOLTADO PARA O PROBLEMA DE PREDIÇÃO DE PROTEÍNAS

Uberlândia, Brasil

2023

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

CAROLINA ALVES DA SILVA

UMA ANÁLISE DE PARÂMETROS DO ALGORITMO EVOLUTIVO VOLTADO PARA O PROBLEMA DE PREDIÇÃO DE PROTEÍNAS

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.
Orientadora: Profa. Dra. Christiane Regina Soares Brasil

Universidade Federal de Uberlândia – UFU

Faculdade de Ciência da Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2023

Resumo

Existem problemas computacionais complexos que buscam uma melhor solução dentro de todas as combinações das possíveis soluções, classificados como problemas não polinomiais (NP). Podemos citar o Problema de Predição de Estruturas de Proteínas (PSP), que é também considerado um NP, que busca encontrar estruturas tridimensionais de proteínas ainda desconhecidas. Este trabalho busca realizar uma análise de resultados de um método computacional, mais especificamente o Algoritmo Evolutivo (AE), que trata o problema de PSP com o modelo de energia 2D HP utilizando a energia simplificada. Foi dado enfoque na análise dos parâmetros: tamanho da população, número de gerações e o tipo de seleção (torneio). Desta forma, foi avaliado o impacto de cada um para aprimorar e evoluir os resultados. Este estudo é de suma relevância para auxiliar na busca de melhores estruturas de proteínas e suas funções, e assim encontrar novos métodos que investigam soluções/fármacos para doenças graves ou até incuráveis até o momento, além de confirmar a importância do ajuste dos melhores valores para cada parâmetro.

Palavras-chave: Otimização, Algoritmo Evolutivo, Predição de Estrutura de Proteína, Modelo HP, Função de Energia, Seleção de Torneio, Tamanho da População, Número de Gerações.

Lista de ilustrações

Lista de abreviaturas e siglas

AE	Algoritmo Evolutivo
AG	Algoritmo Genético
NP	*Non-Deterministic Polynomial time*
PSP	*Protein Structure Problem*

Sumário

1 Introdução

Existem vários problemas computacionais em que algoritmos determinísticos são ineficientes na busca da melhor solução, neste contexto, podemos destacar problemas combinatórios, que necessitam da busca de uma melhor solução dentro de todas as combinações das possíveis soluções. Alguns dos seus clássicos exemplos computacionais são: Caixeiro viajante, Problema da mochila e Árvore de extensão de custo mínimo. [4]

Um problema é chamado NP (*Non-Deterministic Polynomial time*) se a sua solução pode ser encontrada em tempo polinomial por uma máquina não determinística. Esta máquina consegue encontrar ou predizer o resultado correto (ou uma aproximação dos mesmo), produzindo resultados diversos com os mesmos dados de entrada, diferente de uma máquina determinística, que sempre produz o mesmo resultado para os mesmos dados de entrada.

Neste sentido, também temos o Problema de Predição de Estruturas de Proteínas (do inglês, *Protein Structure Problem* - PSP), que é um dos principais desafios da biologia computacional atualmente [2]. Neste problema há um grande conjunto de possibilidades de conformação tridimensional de uma proteína, sendo poucas consideradas estáveis, ou seja, as que possuem energia mínima.

No problema de PSP são buscadas estruturas tridimensionais de proteínas ainda desconhecidas, pois a partir do conhecimento de suas estruturas pode-se obter as suas funções. Deste modo, é possível utilizar estas estruturas no estudo de novas soluções/fármacos para doenças graves ou até incuráveis até o momento. Existem métodos tradicionais em laboratórios que determinam a sequência de aminoácidos que compõem uma proteína, sendo estes considerados simples, como a cristalografia de raio X e a ressonância nuclear magnética (RNM) [2]. Como esses métodos são caros e lentos, além de apresentarem limitações com relação ao tamanho das proteínas, existem métodos computacionais de otimização que lidam com o problema de PSP a partir do resultado da sequência de aminoácidos. Por isso, são necessários estes métodos computacionais de otimização que tratam o problema de PSP, como pode mostrar a literatura. [2, 5, 8, 10, 14, 16, 17, 21, 26–28, 30, 31, 35]

O método estudado neste projeto é o Algoritmo Evolutivo, que é inspirado em processos que ocorrem na natureza e em teorias evolucionistas de Darwin. O objetivo principal deste trabalho foi estudar o Algoritmo Evolutivo, uma vez que é um método clássico de otimização computacional, usando o problema de PSP como uma aplicação do mundo real.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo geral deste trabalho foi realizar um estudo com análise de resultados de um algoritmo de otimização computacional, mais especificamente o Algoritmo Evolutivo, para o problema de predição de estruturas de proteínas com o modelo de representação Hidrofóbico-Polar (HP), utilizando como função objetivo a energia simplificada, a fim de comparar os resultados obtidos com um artigo de referência. No modelo HP, os aminoácidos são classificados em H (hidrofóbico ou não-polar) ou P (hidrofílico ou polar), buscando alcançar simplicidade na representação.

1.1.2 Objetivo específico

O objetivo específico deste trabalho foi realizar experimentos visando analisar diferentes parâmetros do AE implementado. Deste modo, foi dado enfoque em experimentos alterando o tamanho da população, o número de gerações e o tipo de seleção (torneio): avaliando o impacto de cada um e aprimorando os resultados obtidos em cada experimento.

1.2 Justificativa

É importante investigar os métodos de otimização voltados para o problema de predição, a fim de analisar as possibilidades de encontrar soluções adequadas, em um tempo viável, pois os métodos tradicionais atuais disponíveis não são rápidos e são caros para predizer estruturas de proteínas. A solução deste problema é importante para a busca pelo conhecimento profundo sobre as funcionalidades das proteínas e suas estruturas tridimensionais, que são fundamentais na compreensão da origem de diversas doenças, e consequentemente na elaboração de soluções de diversas doenças para as mesmas [1, 6, 9, 15, 19, 32, 34]. Além disso, com este trabalho, podemos confirmar a importância da etapa de ajustes dos parâmetros no AE.

1.3 Considerações Finais

Este trabalho está organizado da seguinte maneira:

- Capítulo 1: apresenta o objetivo principal, o objetivo específico e a justificativa do trabalho, contextualizando-o no problema de predição de estrutura de proteínas.

- Capítulo 2: mostra os principais conceitos abordados no problema de predição de estrutura de proteínas, assim como expõe fundamentos importantes do Algoritmo

Evolutivo e os trabalhos relacionados a este algoritmo referente ao estudo da importância de seus parâmetros.

- Capítulo 3: mostra como foram obtidos os resultados dos experimentos com o Algoritmo Evolutivo implementado, bem como apresenta a análise destes resultados baseada em um comparativo com um artigo de referência.

- Capítulo 4: aponta as conclusões finais alcançadas a partir deste Trabalho de Conclusão de Curso.

2 Referencial Teórico

Neste capítulo serão mostrados conceitos importantes utilizados neste Trabalho de Conclusão de Curso, configurando o referencial teórico para o desenvolvimento do mesmo: será explicado o problema de predição de estruturas de proteínas com suas principais abordagens; o algoritmo de otimização computacional usado para trabalhar com este problema será descrito, neste caso, o Algoritmo Evolutivo; e, por fim, alguns trabalhos científicos serão apresentados por mostrarem a importância de ajustar os parâmetros do algoritmo evolutivo para se obter melhores resultados.

2.1 Problema de Predição de Estruturas de Proteínas

Pesquisadores de várias áreas estão crescentemente buscando conhecimentos para a descoberta de novos fármacos e, deste modo, conseguirem a amenização, e até mesmo a cura de várias doenças. Isso tem ocorrido principalmente nas áreas de Biologia Molecular, Bioquímica e Farmácia, além da busca pelo conhecimento profundo sobre as funcionalidades das proteínas e suas estruturas tridimensionais, que são fundamentais na compreensão da origem de diversas doenças [1, 6, 9, 15, 19, 32, 34]. Existem dois métodos experimentais utilizados na determinação da estrutura terciária de uma proteína: a Cristalografia, e a Ressonância Nuclear Magnética. Estes métodos são caros e lentos, por isso, é necessário um método computacional que seja rápido e confiável para predizer estruturas de proteínas a partir de sequências proteicas [2]. Desta maneira, a predição de estruturas de proteínas por simulação computacional pode impulsionar o desenvolvimento de pesquisas em áreas que requerem a determinação de estruturas de novas proteínas ou de complexos moleculares envolvendo proteínas.

2.1.1 Estrutura de Proteínas

As proteínas são essencialmente importantes para os organismos vivos, pois elas compõem mais de metade do peso seco de uma célula [23].

Elas são formadas por um conjunto de aminoácidos ligados entre si através de ligações peptídicas, ou seja, as proteínas são compostas por moléculas de carbono (C), hidrogênio (H), oxigênio (O) e nitrogênio e (N). Existem vinte tipos de aminoácidos na natureza.

A estrutura da molécula de uma proteína, que é definida através da forma em que os seus aminoácidos interagem entre si e/ou com o meio, pode ser classificada como: primária, secundária, terciária e quaternária.

A estrutura primária, que é a principal e a mais simples, é basicamente a sequência de aminoácidos que diferencia uma proteína da outra.

A estrutura secundária é a conformação tridimensional, no qual os aminoácidos estão dispostos interagindo entre si. Os principais tipos de estruturas secundárias são Hélices, Folhas e Voltas.

Nas Hélices formam-se ligações de hidrogênio entre os aminoácidos, e são como uma corda enrolada em torno de um tubo imaginário. Nas folhas as ligações de hidrogênio entre um aminoácido e outro gera uma estrutura achatada e rígida. Nela, são envolvidos entre 5 a 10 resíduos e podem ser paralelas ou antiparalelas. Já as voltas são as estruturas responsáveis pela inversão da direção da cadeia polipeptídica, e normalmente envolvem de 3 a 4 resíduos [2].

A diferença entre a estrutura secundária e a terciária é que a secundária é determinada pela interação estrutural de curta distância de aminoácidos e a terciária pela longa distância, como por exemplo interações hidrofóbicas.

A estrutura terciária é equivalente à sua disposição tridimensional, ou seja, é a maneira com que as estruturas secundárias estão arranjadas no espaço 3D. É a partir dessa conformação que as atividades biológicas são atribuídas às proteínas. Proteínas constituídas por várias cadeias polipeptídicas são chamadas por subunidades.

Quando estas subunidades se juntam, interagindo entre si, originando a sua estrutura quaternária.

2.1.2 Modelos de representação de energia

Uma vez que que os métodos tradicionais, como a cristalografia de raio X e a ressonância nuclear magnética (RNM), que determinam a estrutura tridimensional de uma proteína são caros e lentos, algoritmos de otimização computacional foram desenvolvidos como uma tentativa de reverter esse problema. Os métodos computacionais para o problema de PSP são *ab initio*, semi *ab initio*, *threading* e homologia. [2]

Neste trabalho o algoritmo estudado utiliza a modelagem *ab initio*, que tem como base algoritmos de otimização que aborda primeiramente a especificação da funcão de minimização e depois a escolha do algoritmo de busca. A grande diferença nessa modelagem é que ela não utiliza de nenhum conhecimento prévio, ou seja, utiliza apenas a sequência de aminoácidos como entrada no algoritmo. Ela possui três modelos de representação de energia, que são: *lattice*, *off-lattice* e *full-atom*.

Os modelos *off-lattice* e *full-atom* possuem um alto custo computacional pois se aproximam de uma estrutura mais realista.

O modelo *lattice* é o mais simples de todos e foi proposto por Shakhnovich [25].

Ele é considerado simples pois em cada vértice da malha um aminoácido é posicionado, não sendo necessário representar a estrutura interna de cada resíduo e mesmo assim ainda consegue preservar interações polares entre os indivíduos da malha. Além disso ele possui um baixo custo computacional. Esse foi o modelo de representação de energia utilizado no algoritmo estudado.

2.1.2.1 Modelo *lattice* hidrofóbico-polar (HP)

O modelo HP foi criado por *Lau and Dill* [18] em 1989, possui uma representação simplificada de uma proteína e se baseia na hidrofobicidade dos aminoácidos para realizar a representação computacional.

Ele utiliza malhas que podem ter duas ou três dimensões. O algoritmo estudado utiliza o modelo HP-2D, que possui duas dimensões. Nele, uma proteína é composta por: P, aminoácidos polares (hidrofílicos) que interagem facilmente com o solvente que é a água; H indica resíduos hidrofóbicos, ou seja, que não interagem com solvente. Os aminoácidos hidrofóbicos tendem a se mover para o núcleo da estrutura enquanto os hidrofílicos tendem a se mover para a superfície.

2.1.3 Funções de energia

2.1.3.1 Energia de Lau e Dill

Essa energia é muito utilizada no modelo HP. Ela se baseia na quantidade de aminoácidos hidrofóbicos vizinhos não consecutivos presentes na malha. Ela é uma energia clássica, e muito utilizada em diversos trabalhos encontrados na literatura. Essa energia de conformação é obtida considerando as interações H-H de aminoácidos não conectados e é dada pela seguinte equação:

$$\mathbf{E} = \beta ij \Sigma \sigma (r\dot{i} r\dot{j})$$

β: Assume o valor de 1 se os aminoácidos forem do tipo H, e 0 caso contrário;

σ: Essa função assume o valor -1 se os aminoácidos ri e rj forem não conectados, e 0 se eles não forem vizinhos ou forem vizinhos conectados.

As linhas pontilhadas da Figura 1 indicam os aminoácidos hidrofóbicos vizinhos não consecutivos na malha. Nela, o valor de energia é -2.

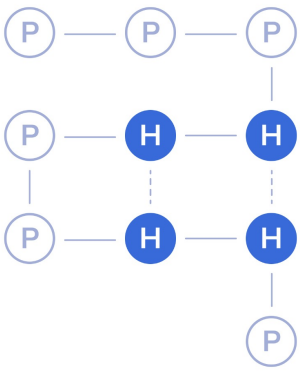

Figura 1 – Sequência de aminoácidos (P-P-P-H-H-P-P-H-H-P)

Fonte: Autora própria (2023)

2.1.3.2 Energia simplificada

Essa função de energia foi apresentada originalmente em 2013, no artigo de Zhang [35], é utilizada no modelo HP e calcula a distância entre todos os aminoácidos hidrofóbicos não consecutivos presentes na malha. No entanto, não é fácil identificar os dois aminoácidos hidrofóbicos que sejam consecutivos ou não consecutivos, então é utilizado um cálculo para encontrar a energia através da matriz de distância.

Para este cálculo é necessário considerar os aminoácidos em coordenadas cartesianas, sendo assim, um aminoácido hidrofóbico é representado por um ponto (x, y). Para cada ponto cartesiano se calcula a distância euclidiana entre todos os outros pontos hidrofóbicos. Os resultados devem estar armazenados em uma matriz de distância para depois serem calculadas a soma dos valores da matriz triangular, superior ou inferior, para encontrar a energia simplificada. A energia de conformação é dada pela seguinte equação:

$$\mathbf{E}s = \Sigma d(r i r j)$$

d: Valores da matriz triangular, superior ou inferior, da distância euclidiana dos aminoácidos ri e rj.

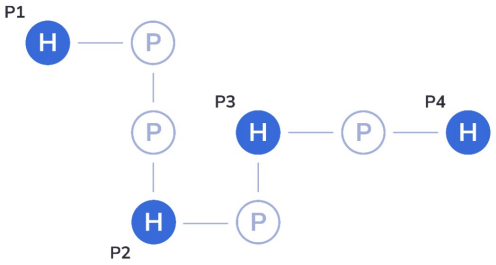

Figura 2 – Sequência de aminoácidos (H-P-P-H-P-H-P-H)

Fonte: Autora própria (2023)

Considerando a Figura 2, os pontos onde os aminoácidos hidrofóbicos estão alocados nas seguintes coordenadas: p1(0,2), p2(2,0), p3(3,1) e p4(5,1). A matriz de distância obtida (Figura 3) foi:

	p1	**p2**	**p3**	**p4**
p1	0	2.82	3.16	5.09
p2	2.82	0	1.14	3.16
p3	3.16	1.14	0	2
p4	5.09	3.16	2	0

Figura 3 – Matriz de distância

Fonte: Autora própria (2023)

Com esses resultados, o cálculo da energia simplificada será:

Es = 2*3.16 + 5.09 + 2.82 + 1.14 + 2 = 17.37.

Após o cálculo da matriz de distância, o valor da energia é 17.37.

Essa energia de Zhang é mais recente que a energia de Lau e Dill, portanto, não existem muitos trabalhos que mostrem o seu desempenho no Algoritmo Evolutivo. Por essa razão, essa energia foi aplicada no Algoritmo Evolutivo 2D-HP para o problema de proteína a fim de analisar/comprovar sua eficiência em encontrar boas soluções.

Na Seção 2.2 serão apresentados os principais fundamentos do Algoritmo Evolutivo encontrados na literatura.

2.2 Algoritmos Evolutivos

A Computação Evolutiva (CE) é um ramo de pesquisa que tem sido cada vez mais utilizado, pois com ela é possível criar algoritmos que encontrem soluções de problemas que ainda não foram encontrados por outras técnicas computacionais e foi por meio dela

que se originou os Algoritmos Evolutivos (AE), que são métodos considerados simples. Estes métodos utilizam a Teoria da Evolução e Genética e podem ser escritos em poucas linhas de código. Os AEs eram em seu início utilizados como ferramentas de modelagem e simulação computacional, mas conforme o acesso aos computadores foi aumentando, e essas técnicas de otimização começaram a ser mais utilizadas. Neles, o número de iterações necessárias para obter uma solução na maioria dos casos, é menor do que outras técnicas de busca e otimização, pois eles trabalham com conjuntos de soluções, ao contrário das outras. [13]

Três abordagens de AEs foram criadas: a programação evolutiva (PE), as estratégias evolutivas (EE) e os algoritmos genéticos (AGs). Em todas elas, quando é dada uma população de indivíduos, pressões do ambiente, ou seja, processos que favorecem as melhores soluções encontradas, permitem que somente os mais aptos sejam selecionados. Dada uma função a ser otimizada, seja maximizada ou minimizada, primeiramente escolhe-se um conjunto de soluções que sao criadas aleatoriamente. Todo esse conjunto é avaliado de acordo com uma função que mede a qualidade das soluções candidatas e então é atribuído um valor que calcula adequação de cada solução. Esse valor de adequação, ou aptidão, é chamado *fitness* . Depois que as melhores soluções são selecionadas com base no *fitness* , temos a criação de uma nova população através da aplicação de operadores de recombinação e/ou mutação. [13]

A recombinação imita o processo de reprodução sexuada. Esse operador é aplicado em duas ou mais soluções candidatas, que são chamadas pais, e resulta em duas ou mais novas soluções, chamadas descendentes ou filhos. A mutação é aplicada em uma candidata a fim de gerar outra. Ao finalizar os processos de seleção, recombinação e mutação, e escolhida a candidata que possua a melhor solução qualificada (que tenha o melhor *fitness*), comparando as novas candidatas com as candidatas da geração anterior, para assumir um lugar na nova população. Caso ela não seja encontrada, esse processo irá se repetir até que um número máximo de iterações seja obtido ou até ter encontrado uma solução ótima. [13]

2.2.1 Algoritmos Genéticos

Os Algoritmos Genéticos (AGs) foram fundamentados na década de 1970 por John Henry Holland [13] e seus alunos e eles foram inspirados na Teoria da Evolução de Darwin. Para Holland, a evolução natural é um processo simples, poderoso e robusto. Com os AGs é possível produzir soluções de qualidade, além de poder ser utilizado para achar soluções computacionais em problemas de otimização.

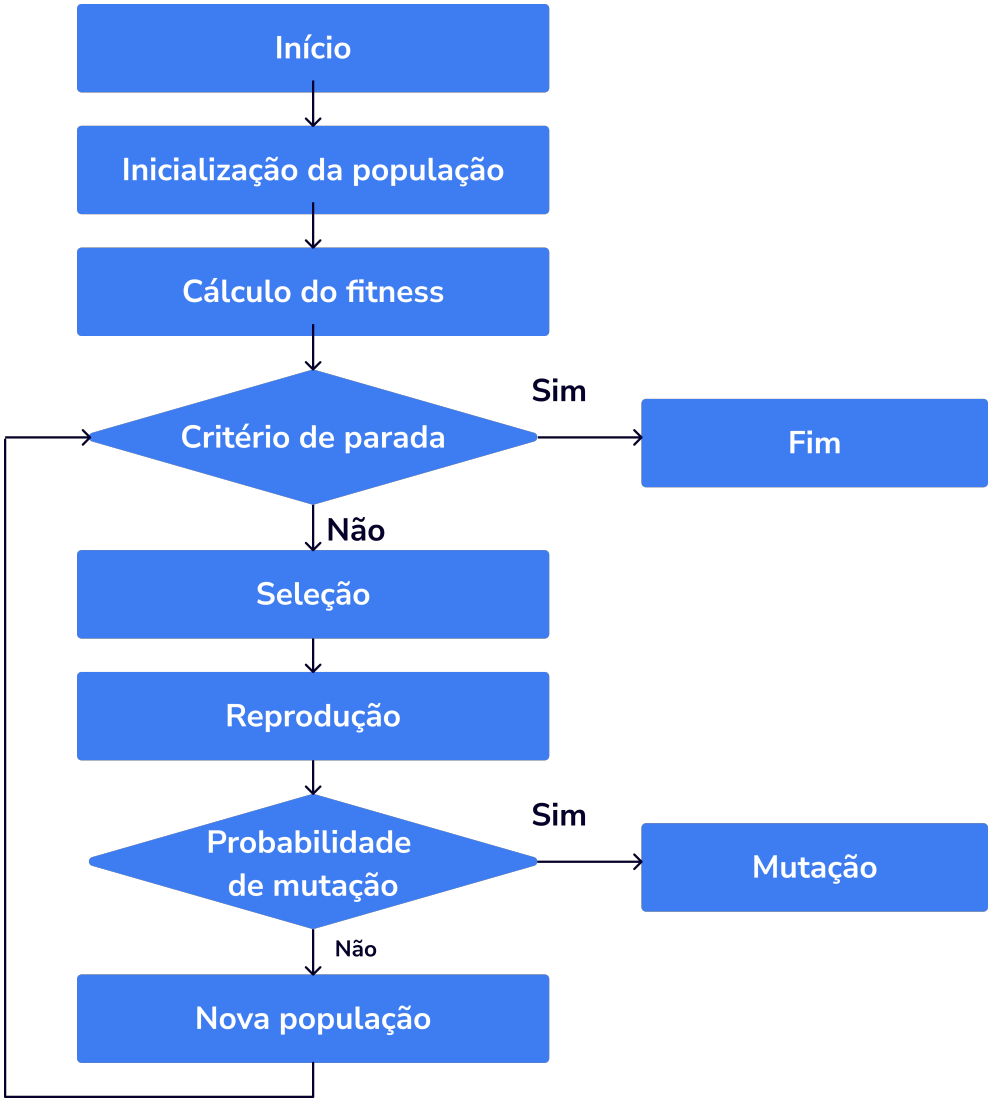

Figura 4 – Diagrama de Fluxo de um Algoritmo Genético

Fonte: Autora própria (2023)

A seguir, mostram-se os principais passos de um AG (Figura 4):

- Primeiramente, escolhe-se uma população inicial;
- Avalia-se cada indivíduo produzindo uma medida de aptidão, ou *fitness*;
- Em seguida, escolhem-se os indivíduos com melhor *fitness* através do operador de seleção;
- Novas soluções potenciais são formadas, quando há recombinação e mutação em alguns indivíduos;
- São selecionados indivíduos sobreviventes para a próxima geração;
- Este processo é repetido até que tenha uma solução encontrada ou um número máximo de iterações.

O desempenho de AGs ainda pode ser melhorado utilizando a *seleção elitista* [12, 22], que é manter sempre o melhor indivíduo da geração atual na geração seguinte, ou mesmo uma porcentagem dos melhores indivíduos. O AG também pode simplesmente forçar a escolha do melhor indivíduo encontrado em todas as gerações do algoritmo.

Existe a *seleção por classificação* [12, 22], que utiliza determina a probabilidade de seleção através das posições dos indivíduos quando ordenados de acordo com o *fitness*.

Há também a *seleção por torneio*: em um subconjunto da população com k indivíduos é sorteado e os melhores indivíduos desse grupo são selecionados para decidir qual irá reproduzir. Nela, não é necessário um conhecimento global da população. [29, 33]

2.2.1.1 Seleção

O processo de seleção se baseia no princípio da seleção natural, ou seja, apenas os indivíduos mais aptos de uma espécie conseguem sobreviver. Eles são aqueles que herdaram as características mais adequadas para determinadas condições naturais. [7] Existem vários métodos utilizados para selecionar esses indivíduos considerados aptos, sendo alguns deles:

Seleção por classificação: Todos os indivíduos de uma população são ordenados pelo seu valor de aptidão, assim sua probabilidade de escolha é atribuída através da posição que cada um ocupa.

Seleção elitista: Nesta forma de seleção existe uma taxa de preservação dos melhores indivíduos, que são copiados para a próxima geração.

Seleção por torneio: Um número n de indivíduos de uma população é escolhido aleatoriamente. Depois, o indivíduo que possuir o melhor valor de aptidão entre o grupo é selecionado para gerar uma nova população.

A Figura 5 demonstra uma seleção por torneio de 2. Sendo que, 4 indivíduos foram escolhidos de forma aleatória, onde 2 são ganhadores (que possuíam o melhor*fitness*) gerar o próximo indivíduo.

Figura 5 – Representação da seleção por torneio

Fonte: Autora própria (2023)

Seleção por roleta: Os indivíduos de uma população são escolhidos através do sorteio de uma roleta, ou seja, cada indivíduo é representado na roleta pelo seu valor de aptidão. sendo assim, os indivíduos com um valor alto de aptidão irão possuir uma porção maior da roleta (Figura 6). A roleta deve ser girada um número de vezes igual ao tamanho da população, e serão escolhidos os da próxima geração, aqueles sorteados na roleta.

Indivíduo S_i	Aptidão $f(S_i)$	Apitão Relativa
S_1 10110	2.22	0.14
S_2 11000	7.26	0.47
S_3 11110	1.04	0.07
S_4 01001	3.36	0.21
S_5 00110	1.70	0.11

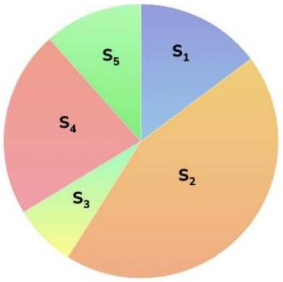

Figura 6 – Representação da Seleção por roleta.

Fonte: Autora própria (2023)

2.2.1.2 Cruzamento

O cruzamento é um processo que possibilita a mistura de genes de indivíduos, ou seja, a junção de material genético de um indivíduo de gerações anteriores em novos indivíduos, e isso ocorre através da reprodução sexuada.

Existem vários tipos de operadores de cruzamento, a seguir serão descritos alguns deles.

Cruzamento de um ponto: Nesse tipo de cruzamento é escolhido um ponto P da cadeia do cromossomo aleatoriamente e se copia uma parte dos genes de cada pai, sendo a posição P do *pai1* e a posição $P + 1$ do *pai2*, gerando assim um novo filho. Neste método é comum os pais gerarem dois filhos. A Figura 7 demonstra um exemplo deste método de cruzamento.

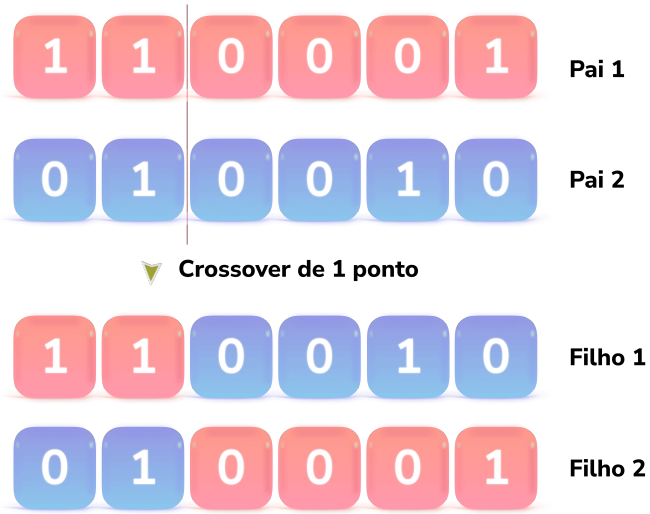

Figura 7 – Representação do cruzamento de um ponto

Fonte: Autora própria (2023)

Cruzamento de dois pontos: Diferente do cruzamento de um ponto, nesse tipo de cruzamento são escolhidos dois pontos *P1* e *P2* da cadeia do cromossomo aleatoriamente e são copiados uma parte dos genes de cada pai, com base nos dois pontos de corte selecionados gerando assim, dois novos filhos. A Figura 8 demonstra um exemplo deste método de cruzamento.

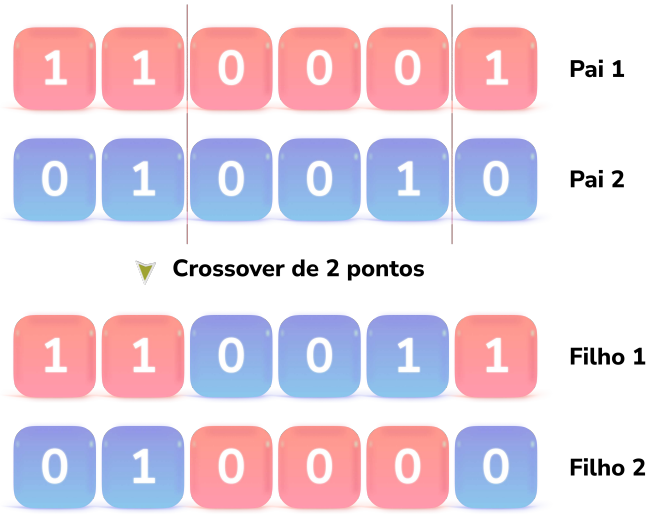

Figura 8 – Representação do cruzamento de dois pontos

Fonte: Autora própria (2023)

Cruzamento uniforme: Nesse tipo de cruzamento, representado na Figura 9, os filhos são gerados através de uma máscara binária gerada aleatoriamente. Na criação do cromossomo, todas as posições dessa máscara são percorridas e quando o valor de uma posição for 0, o gene do *pai1* é copiado, caso o valor seja 1, o gene do *pai2* é copiado.

Figura 9 – Representação do cruzamento uniforme

Fonte: Autora própria (2023)

2.2.1.3 Mutação

A variabilidade genética ocorre através da mutação. Nela, há alterações na estrutura de um gene de um indivíduo sorteado aleatoriamente com uma determinada probabilidade de mutação. Sendo assim, vários indivíduos de uma nova população podem ter um de seus genes alterado aleatoriamente.

Mutação genética: Alterações que afetam um único gene, que originam novas versões dos genes. Essa condição pode produzir novas características nos portadores da mutação. Os seus tipos são: substituição, inserção e deleção. [7]

Mutação genética por substituição: Nesse tipo de mutação genética, um gene é sorteado aleatoriamente, e então ele é alterado por outro gene (Figura 10). [7]

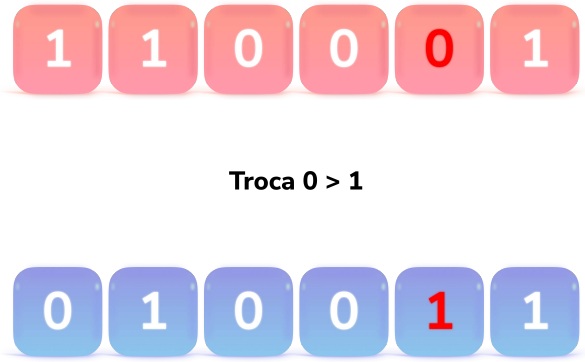

Figura 10 – Representação da mutação por substituição

Fonte: Autora própria (2023)

Na próxima seção serão mostrados trabalhos científicos que confirmam a relevância de se ajustar os parâmetros do Algoritmo Evolutivo, a fim de se obter melhores resultados.

2.3 Trabalhos relacionados

Nesta seção serão apresentados alguns trabalhos que visam analisar diferentes parâmetros de um AE e a importância de realizar os ajuste deste parâmetros, dando enfoque na seleção por torneio. Existem outros trabalhos que também possuem esse objetivo, porém, estes artigos apresentados a seguir possuem semelhança com o presente projeto, ainda que tratem de outros problemas.

1. *Population Size Influence on the Genetic and Ant Algorithms Performance in Case of Cultivation Process Modeling.*

Neste artigo, foi feita uma investigação da influência do tamanho da população no desempenho do Algoritmo Genético (AG) e Otimização de Colônias de Formigas (ACO) para um problema de identificação de parâmetros. Sua conclusão foi que, apesar de encontrar bons valores de parâmetros não seja uma tarefa simples e que requer experiência humana, bem como tempo, uma boa seleção dos parâmetros do algoritmo melhora o tempo de computação e também a precisão da solução. [24]

2. *Comparative Review of Selection Techniques in Genetic Algorithm*

Neste artigo são comparadas diversas técnicas de seleção, utilizando um AG com o objetivo de fazer um estudo sobre a importância da identificação da técnica de seleção apropriada. É enfatizado o quanto o processo de seleção desempenha um papel importante na resolução da convergência prematura que se pode ocorrer, devido à falta de diversidade na população. Portanto, a seleção da população em cada geração é muito importante. Na sua conclusão observou-se que a seleção por torneio tem melhor desempenho do que outras técnicas em termos de taxa de convergência e complexidade de tempo. [29]

3. *Comparative Study of Different Selection Techniques in Genetic Algorithm*

Neste artigo, seis tipos de estratégias de seleção foram descritas, utilizando um AG com o objetivo de encontrar a solução ótima, comparando o desempenho e o número de gerações de cada uma das seleções. Algumas das técnicas utilizadas foram: seleção por roleta, seleção por classificação, seleção por torneio e seleção por elitismo. Sua conclusão foi que, apesar de todos os métodos de seleção possuírem o mesmo objetivo, eles se diferem na maneira como os indivíduos mais aptos serão escolhidos, logo, alguns métodos são mais robustos do que outros. Por isso, é importante a escolha de um método que seja satisfatório e viável para determinado AG. [33]

4. *A Many-Objective Evolutionary Algorithm Using A One-by-One Selection Strategy*

Como a maioria dos algoritmos evolutivos multiobjetivos existentes possuem dificuldades em resolver problemas de otimização, neste artigo foi proposto um novo algoritmo evolutivo usando uma estratégia de seleção um por um, numa tentativa de solucionar esse problema. Nesse tipo de seleção, quando um indivíduo é selecionado, os outros são excluídos através de uma técnica de nicho, garantindo a diversidade da população, na qual a similaridade entre os indivíduos é avaliada por meio de um indicador de distribuição. O algoritmo proposto foi comparado empiricamente com oito algoritmos evolutivos. Os resultados comparativos demonstram que o desempenho geral do algoritmo proposto é superior aos algoritmos comparados nos problemas de otimização estudados neste artigo. [20]

5. *On Proportions of Fit Individuals in Population of Mutation-Based Evolutionary Algorithm with Tournament Selection*

Neste artigo, foi utilizado um AE não elitista, com o método de seleção por torneio com o objetivo de buscar uma convergência mais segura das soluções obtidas. A sua conclusão foi que, com os resultados obtidos, nota-se que aumentar o tamanho do torneio melhora o desempenho do AE. [11]

2.4 Considerações finais

Neste capítulo foram apresentados conceitos importantes no contexto do problema de predição de estrutura de proteínas, bem como foram descritos os principais fundamentos do algoritmo de otimização computacional utilizado neste trabalho: o Algoritmo Evolutivo. Finalizando o capítulo, mostrou-se alguns artigos científicos que confirmam a relevância do estudo dos parâmetros do Algoritmo Evolutivo, e o quanto ajustes nestes parâmetros podem influenciar em bons resultados, enfatizando na análise da seleção por torneio.

3 Metodologias e Resultados

Neste capítulo serão apresentados os resultados obtidos do Algoritmo Evolutivo implementado para o problema de predição de proteínas. A função objetivo aplicada neste AE foi a energia simplificada e as soluções obtidas foram comparadas com o artigo de referência [3], que usou a mesma função de energia com modelo HP-2D, com a diferença de aplicá-la no Algoritmo de Colônia de Formigas.

Cada indivíduo da população do AE desenvolvido foi representando por uma sequência de aminoácidos, classificados como hidrofóbicos (H) ou polares (P), que são posicionados em coordenadas específicas em uma malha 2D. Logo, cada proteína tem um conjunto de indivíduos que representam as diferentes possibilidades de posicionamento dos aminoácidos. Por exemplo, um indivíduo pode ser a sequência: P-P-P-H-H-P-P-H-H-P, onde cada aminoácido tem uma posição diferente em um plano cartesiano, e deste modo, cada proteína pode ter inúmeras soluções distintas, dependendo das posições de cada um dos aminoácidos da proteína.

Neste trabalho, foram feitas 10 execuções do AE para cada configuração utilizando a energia simplificada, tendo como referência os resultados de um estudo sobre as funções de energia com modelo HP-2D [3], onde foram variados os números de indivíduos e gerações, sendo N o tamanho da proteína; ES a melhor energia encontrada no artigo de referência; E a melhor energia encontrada; P a pior energia encontrada; M a média aritmética das melhores energias encontradas; D o desvio padrão das melhores energias encontradas; T a média aritmética do tempo (em segundos) de execução do algoritmo das melhores energias encontradas.

Os parâmetros utilizados nesse AE que não foram alterados, ou seja, que se mantiveram fixos em todos os experimentos são: cruzamento, taxa de cruzamento, mutação, taxa de mutação e o elitismo. Os tipos de cruzamento utilizados foram: 1 ponto e de 2 pontos; a taxa de cruzamento utilizada foi: 50%, 70% e 100% de forma incremental durante as gerações; o tipo de mutação utilizada foi: de 1 ponto; a taxa de mutação utilizada foi: 5% e 30% de forma incremental durante as gerações. Foi utilizado elitismo, com uma taxa de 50%.

Para implementação e experimentos, foi usado um processador i5 com 4GB de memória, utilizando a linguagem de programação C.

Primeiramente, os testes foram feitos com a seleção de torneio de 2. As configurações utilizadas foram número de indivíduos = 100, e o de gerações = 500.

N	ES	E	P	M	D	T
18	100.97	100.97	106.07	101.61	1.84	19.1
26	455.99	457.23	502.33	475.15	15.72	22.5
30	300.96	316.47	364.48	331.65	16.79	25.2
39	716.77	745.96	790.54	767.38	16.56	18.8
42	1150.56	1195.56	1526.06	1290.46	69.22	19.1
49	878.12	961.56	1149.83	1003.41	51.21	19.8

Figura 11 – Resultados 1 (torneio = 2; nº de indivíduos. = 100; nº de gerações = 500)

Fonte: Autora própria (2023)

A proteína de tamanho 18 foi a única em que o resultado alcançado foi igual ao do artigo de referência e isso é um indício que é promissor alterar a forma de seleção. Em todas as outras proteínas o menor resultado foi maior do que a do artigo de referência.

Nos próximos experimentos, a forma de seleção foi alterada para torneio de 3, com o objetivo de aumentar a probabilidade dos pais selecionados aleatoriamente possuírem um bom valor de aptidão, e as configurações utilizadas também foram número de indivíduos = 100, e o de gerações = 500.

N	ES	E	P	M	D	T
18	100.97	100.97	100.97	100.97	0	17.9
26	455.99	457.23	495.97	469.82	11.16	19.5
30	300.96	306.75	347.16	326.76	11.87	18.7
39	716.77	735.04	797.02	760.75	19.49	18.0
42	1150.56	1204.45	1320.30	1266.91	49.51	19.2
49	878.12	902.23	1130.28	981.69	73.37	20.5

Figura 12 – Resultados 2 (torneio = 3; nº de indivíduos = 100; nº de gerações = 500)

Fonte: Autora própria (2023)

Na Figura 12, apenas a proteína de tamanho 18 alcançou o resultado esperado novamente, mas desta vez em todas as execuções. Na proteína de tamanho 26, o menor resultado foi o mesmo que o da tabela de resultados 1, mas o seu desvio foi menor, o que indica que seus resultados foram mais consistentes. Entretanto, na proteína de tamanho 30, o seu menor resultado e desvio foram melhores que o da tabela de resultados 1. Na proteína de tamanho 39, o menor resultado foi melhor que o da tabela de resultados 1, mas o seu desvio foi pior. Na proteína de tamanho 42 o seu menor resultado foi pior do que o da tabela de resultados 1, mas o seu desvio foi melhor. Enquanto na proteína de

tamanho 49, o menor resultado foi melhor do que o da tabela de resultados 1, mas o seu desvio foi pior.

Apenas a proteína de tamanho 18 alcançou o resultado esperado, isso é um indicativo que é promissor aumentar o número de torneio novamente. A forma de seleção dos próximos testes foi alterada para torneio de 4.

N	ES	E	P	M	D	T
18	100.97	100.97	100.97	100.97	0	18.2
26	455.99	454.82	483.54	469.49	10.46	20.3
30	300.96	306.75	347.58	323.70	11.09	21.4
39	716.77	730.20	780.10	749.43	16.93	21.2
42	1150.56	1180.63	1329.32	1248.14	47.55	24.2
49	878.12	915.81	1089.46	987.37	50.39	23.4

Figura 13 – Resultados 3 (torneio = 4; nº de indivíduos = 100; nº de gerações = 500)

Fonte: Autora própria (2023)

Na Figura 13, a proteína de tamanho 18 o resultado esperado foi alcançado em todas as execuções. Na proteína de tamanho 26 o menor resultado foi melhor que a do artigo de referência e o seu desvio foi menor do que o da Figura 12, indicando que os resultados foram mais consistentes. Entretanto, na proteína de tamanho 30, o menor resultado foi o mesmo do da tabela anterior, mas o seu desvio foi menor. Nas proteínas de tamanho 39 e 42, o menor resultado e o desvio foram melhores do que o da tabela anterior. Na proteína de tamanho 49, o menor resultado foi maior do que o da tabela anterior, mas o seu desvio foi menor. Como os resultados foram promissores utilizando torneio de 4, foi dada continuidade nos testes mantendo a forma de seleção e o número de gerações e aumentando somente o número de indivíduos.

A configuração utilizada nos resultados abaixo foi: número de indivíduos = 200 e o de gerações = 500.

N	ES	E	P	M	D	T
18	100.97	100.97	100.97	100.97	0	18.2
26	455.99	457.23	470.21	461.23	5.26	20.9
30	300.96	300.96	359.61	325.90	19.53	21.8
39	716.77	719.92	805.35	768.40	26.71	21.2
42	1150.56	1214.14	1357.99	1267.88	45.35	25.3
49	878.12	885.69	1032.63	952.69	49.09	22.4

Figura 14 – Resultados 4 (torneio = 4; nº de indivíduos = 200; nº de gerações = 500)

Fonte: Autora própria (2023)

Na Figura 14, a proteína de tamanho 18 alcançou novamente o resultado esperado em todas as execuções. Na proteína de tamanho 26 o menor resultado foi maior do que o da Figura 13, mas o seu desvio foi menor, o que significa que os resultados foram mais consistentes. A proteína de tamanho 30 também alcançou o resultado esperado, apesar do seu desvio ter sido maior do que o da tabela anterior. Na proteína de tamanho 39, o menor resultado foi melhor do que o da tabela anterior, apesar do seu desvio ter sido maior, indicando resultados inconsistentes. Entretanto, na proteína de tamanho 42, o menor resultado foi pior do que o da tabela anterior, mas o seu desvio foi melhor. Na proteína de tamanho 49, ambos menor resultado e desvio foram melhores do que os da tabela anterior.

Foi dada continuidade nas execuções utilizando as configurações: número de indivíduos = 300 e o de gerações = 500, com o objetivo de aumentar a variabilidade genética da população.

N	ES	E	P	M	D	T
18	100.97	100.97	100.97	100.97	0	18.5
26	455.99	454.82	469.43	460.24	5.83	20.9
30	300.96	305.16	328.96	310.89	8.62	21.8
39	716.77	716.77	782.10	752.80	22.19	21.2
42	1150.56	1167.35	1272.12	1207.08	33.28	25.3
49	878.12	919.14	1004.43	958.41	23.05	23.2

Figura 15 – Resultados 5 (torneio = 4; nº de indivíduos = 300; nº de gerações = 500)

Fonte: Autora própria (2023)

Na Figura 15, novamente em todas as execuções a proteína de tamanho 18 alcançou o resultado esperado. Na proteína de tamanho 26 apesar do desvio ter sido maior que o da

tabela de resultados 4, o menor resultado foi melhor do que a do artigo de referência. A de tamanho 39 alcançou o resultado o esperado e o seu desvio também foi menor que o da tabela anterior. As proteínas de tamanho 30, 42, 49 não alcançaram o resultado esperado mas os seus desvios foram menores que o da tabela anterior, que significa que o resultado foi mais homogêneo.

No geral, os resultados foram melhores que os anteriores, o que é um indicador para dar continuidade nos testes aumentando somente o número de indivíduos.

A próxima configuração utilizada foi: número de indivíduos = 400 e o de gerações = 500.

N	ES	E	P	M	D	T
18	100.97	100.97	100.97	100.97	0	35.5
26	455.99	457.23	470.21	461.33	4.93	46.3
30	300.96	300.96	318.11	306.45	4.82	46.2
39	716.77	731.04	765.86	749.76	11.72	50.1
42	1150.56	1186.62	1338.48	1228.17	43.05	46.2
49	878.12	885.69	992.01	937.21	31.86	52.6

Figura 16 – Resultados 6 (torneio = 4; nº de indivíduos = 400; nº de gerações = 500)

Fonte: Autora própria (2023)

Na Figura 16, em todas as execuções a proteína de tamanho 18 alcançou o resultado esperado. Na proteína de tamanho 30, o resultado esperado foi alcançado, e o seu desvio foi menor do que o da tabela de resultados 5. Nas proteína de tamanho 26 e 39, apesar do seu menor resultado ter sido maior do que que o da tabela anterior, o seu desvio foi menor, indicando consistência nos resultados. Na proteína de tamanho 42, ambos menor resultado e desvio foram maiores do que os da tabela anterior. Na proteína de tamanho 49, apesar do menor resultado ter sido melhor do que o da tabela anterior, o seu desvio foi maior.

Os melhores resultados obtidos foram com as configurações de: número de indivíduos = 300 e o de gerações = 500. Por isso, nos próximos testes o número de indivíduos foi mantido, alterando apenas o número de gerações, já que o número de gerações anterior não foi suficiente pra alcançar o resultado esperado.

Na Figura 17, as configurações utilizadas foram: número de indivíduos = 300 e o de gerações = 600.

N	ES	E	P	M	D	T
18	100.97	100.97	100.97	100.97	0	38.0
26	455.99	457.23	487.19	463.69	12.53	38.2
30	300.96	300.96	335.97	317.22	12.86	39.8
39	716.77	722.27	759.61	739.91	14.68	43.6
42	1150.56	1183.17	1291.49	1221.19	38.06	41.2
49	878.12	900.29	1008.98	947.29	34.12	48.9

Figura 17 – Resultados 7 (torneio = 4; nº de indivíduos = 300; nº de gerações = 600)

Fonte: Autora própria (2023)

Na Figura 17, outra vez em todas as execuções a proteína de tamanho 18 alcançou o resultado esperado. Na proteína de tamanho 26, o menor resultado foi o mesmo que o da Figura 16, mas o seu desvio foi maior, o que significa que os seus resultados não convergiram da maneira esperada. Na proteína de tamanho 30, o resultado esperado foi alcançado, mas o seu desvio também foi maior do que o da Figura 16. Na proteína de tamanho 39, o menor resultado foi melhor do que o da tabela de seu anterior, mas o seu desvio também foi maior. Na proteína de tamanho 49, ambos menor resultado e desvio foram maiores do que os da tabela de anterior.

Continuamos realizando os experimentos aumentando de 100 em 100 o número de gerações até 1000, e não obtivemos uma melhora nos resultados, as melhores energias ficaram maiores do que todas as anteriores.

Os melhores resultados obtidos foram com as configurações de: número de indivíduos = 300 e o de gerações = 500 utilizando torneio de 4, logo, as próximas execuções utilizaram essas configurações, alterando a forma de seleção para torneio de 5. Com isso, podemos avaliar que quanto maior o número de gerações, há uma chance de perdermos indivíduos com alta aptidão a cada nova geração. Nas próximas execuções as configurações utilizadas foram: número de indivíduos = 300 e o de gerações = 500, com a forma de seleção de torneio de 5.

N	ES	E	P	M	D	T
18	100.97	100.97	100.97	100.97	0	40.0
26	455.99	455.82	470.21	460.32	5.25	40.9
30	300.96	305.16	333.97	314.58	9.40	45.4
39	716.77	722.27	801.15	755.09	25.44	44.8
42	1150.56	1177.07	1273.11	1225.30	39.96	52.5
49	878.12	918.65	1063.81	988.48	36.70	48.3

Figura 18 – Resultados 8 (torneio = 5; nº de indivíduos = 300; nº de gerações = 500)

Fonte: Autora própria (2023)

Na Figura 18, novamente em todas as execuções a proteína de tamanho 18 alcançou o resultado esperado. Na proteína de tamanho 26, o resultado foi melhor do que o do artigo de referência. Na proteína de tamanho 30, o menor resultado foi maior que o da tabela anterior, mas o seu desvio foi menor. Na proteína de tamanho 39, ambos menor resultado e desvio foram maiores do que os da tabela anterior. Enquanto que, na proteína de tamanho 42, ambos menor resultado e desvio foram menores do que os da tabela anterior, que significa que os resultados estão mais consistentes. Na proteína de tamanho 49, o menor resultado foi maior do que o da tabela anterior, mas o seu desvio foi menor.

Como os resultados foram promissores utilizando torneio de 5, foi dada continuidade nos testes mantendo a forma de seleção e o número de gerações, aumentando somente o número de indivíduos. Abaixo, as configurações utilizadas foram: número de indivíduos = 300 e o de gerações = 600.

N	ES	E	P	M	D	T
18	100.97	100.97	100.97	100.97	0	39.1
26	455.99	455.99	470.21	463.18	6.20	39.2
30	300.96	300.96	350.39	318.98	13.02	39.4
39	716.77	716.77	794.12	765.93	37.07	41.2
42	1150.56	1157.15	1315.68	1221.49	38.81	42.6
49	878.12	889.54	1015.65	935.97	39.13	44.9

Figura 19 – Resultados 9 (torneio = 5; nº de indivíduos = 300; nº de gerações = 600)

Fonte: Autora própria (2023)

Na Figura 19, outra vez em todas as execuções a proteína de tamanho 18 alcançou o resultado esperado. Na proteínas de tamanho 26, 30 e 39, os resultado foram alcançados, mas os seu desvios foram maiores do que os da Figura 18. Na proteína de tamanho 42,

o menor resultado foi bem próximo ao do artigo de referência, e também menor do que o da tabela anterior, e o seu desvio também foi menor. Entretanto, proteína de tamanho 49, o menor resultado também foi menor do que o da tabela anterior, mas o seu desvio foi maior.

Os resultados foram um dos melhores obtidos até o momento, por isso aumentando somente o número de indivíduos. Abaixo, as configurações utilizadas foram: número de indivíduos= 300 e o de gerações = 700.

N	ES	E	P	M	D	T
18	100.97	100.97	100.97	100.97	0	40.0
26	455.99	454.82	470.21	460.32	5.02	40.9
30	300.96	305.16	333.97	314.58	8.40	45.4
39	716.77	722.27	801.15	755.09	25.39	44.8
42	1150.56	1177.07	1273.11	1225.30	32.40	52.5
49	878.12	918.65	1063.81	988.48	52.75	48.3

Figura 20 – Resultados 10 (torneio = 5; nº de indivíduos = 300; nº de gerações = 700)

Fonte: Autora própria (2023)

Na Figura 20, em todas as execuções a proteína de tamanho 18 alcançou o resultado esperado. Na proteína de tamanho 26 o menor resultado foi melhor do que a do artigo de referência, e o seu desvio também foi menor que o da Figura 19. Nas proteína de tamanho 30, 39, 42, e 49, os menores resultados foram maiores que o da tabela anterior, mas os desvio foi menores.

Como os resultados não foram promissores, a forma de seleção dos próximos testes foram alteradas para torneio de 6 utilizando as configurações número de indivíduos = 300 e o de gerações = 600, pois com elas foram obtidas energias mais próximos do artigo de referência.

N	ES	E	P	M	D	T
18	100.97	100.97	100.97	100.97	0	39.3
26	455.99	455.99	474.52	461.38	6.62	39.4
30	300.96	300.96	315.02	318.98	8.61	40.2
39	716.77	716.77	855.46	765.40	13.56	41.2
42	1150.56	1027.08	1252.75	1191.28	39.08	40.1
49	878.12	879.67	968.77	928.3	59.52	44.4

Figura 21 – Resultados 11 (torneio = 5; nº de indivíduos = 300; nº de gerações = 600)

Fonte: Autora própria (2023)

Na Figura 21, novamente em todas as execuções a proteína de tamanho 18 alcançou o resultado esperado. Nas proteínas de tamanho 26, 30 e 39 os resultado foram alcançados, mas os seu desvios foram maiores do que os da Figura 20. Na proteína de tamanho 42, o resultado alcançado foi menor do que o do artigo de referência, apesar do desvio ter sido maior do que o da tabela anterior. Embora que, na proteína de tamanho 49, o menor resultado não foi alcançado, eles foi menor do que os da tabela anterior, e o seus desvio também.

Como os resultados foram muito promissores e os melhores obtidos até aqui, podemos notar que houve uma melhora significativa a cada vez que a forma de seleção foi alterada. Com isso, a forma de seleção dos próximos testes foram alteradas para torneio de 7 utilizando as configurações número de indivíduos = 300 e o de gerações = 600.

N	ES	E	P	M	D	T
18	100.97	100.97	100.97	100.97	0	39.1
26	455.99	455.99	474.52	461.38	5.54	39.5
30	300.96	300.96	315.02	318.98	9.14	41.4
39	716.77	716.77	855.46	765.40	15.85	43.2
42	1150.56	1150.50	1252.75	1191.28	33.83	40.2
49	878.12	878.12	980.76	946.88	53.20	42.4

Figura 22 – Resultados 12 (torneio = 7; nº de indivíduos = 300; nº de gerações = 600)

Fonte: Autora própria (2023)

Novamente, em todas as execuções a proteína de tamanho 18 alcançou o resultado esperado. Nas proteínas de tamanho 30 e 39 os resultado foram alcançados, mas os seu desvios foram maiores do que os da Figura 21. Nas proteína de tamanho 26, 42 e 49, os

resultados também foram alcançados, apesar dos desvios terem sido maiores do que os da tabela anterior.

3.1 Considerações finais

Neste capítulo foram apresentados como foram obtidos os resultados dos experimentos realizados com AE no modelo 2D-HP aplicando como função objetivo a energia simplificada.

Pode-se observar que aumentar o tamanho da população no AE significa aumentar a variabilidade genética da população, pois quanto maior ela for, maior a chance de ser diversificada. Portanto, com uma população grande há uma maior probabilidade de se gerar um indivíduo forte, isto é, que tenha chance de estar mais bem adaptado ao meio e com uma boa carga genética. Apesar disso, populações muito grandes necessitam de um esforço computacional maior, por isso este limitante deve ser considerado na definição do tamanho populacional.

Outro ponto importante é o número de gerações, que representa um dos critérios de parada de um AE, pois é ele quem define o número máximo de ciclos de evolução. Um número muito pequeno de gerações pode causar um desempenho ruim, pois o ciclo evolutivo acaba sendo muito curto, podendo levar a uma convergência prematura e não chegando à soluções esperadas. Por outro lado, ciclos grandes demandam um tempo maior de processamento e também podem causar a perda de diversidade, ou seja, não ocorrer uma melhora significativa na solução com o aumento das gerações.

Na seleção por torneio de N, temos que, N indivíduos devem ser escolhidos de forma aleatória, e os dois que possuírem os melhores valores de aptidão serão os pais selecionados para gerar uma nova população. A partir dos experimentos realizados, pode-se notar que um torneio pequeno diminui a probabilidade de encontrar bons indivíduos, sendo assim, aumentar o torneio significou aumentar as chances de encontrar pais que possuam bons genes e, consequentemente, melhorar a próxima geração. Foi importante ter um cuidado ao escolher um N não tão próximo do tamanho da população, pois praticamente toda a população seria escolhida neste caso.

4 Conclusões

Neste Trabalho de Conclusão de Curso, o objetivo geral foi desenvolver um AE voltado para o problema de predição de estrutura de proteínas, aplicando a energia simplificada como função objetivo, a fim de analisar quão bem este algoritmo se comportaria com esta energia, uma vez que não foram encontrados trabalhos que relatassem este experimento. O o artigo de referência [3] utilizado foi aplicado com o Algoritmo de Colônia de Formigas e mostrou bons resultados. O objetivo específico foi realizar experimentos alterando alguns parâmetros do AE neste problema em questão: o tamanho da população, o número de gerações e a forma de seleção (torneio), para melhor avaliar o desempenho do algoritmo implementado.

Sob o ponto de vista do objetivo principal do trabalho, pode-se concluir que o Algoritmo Evolutivo obteve resultados promissores utilizando a energia simplificada para o problema de predição de proteínas, comparando com o artigo de referência que aplicou a mesma energia em um algoritmo diferente, alcançando o mesmo valor e até, em alguns casos, melhores valores que o artigo de referência com o Algoritmo de Colônia de Formigas.

Focando o objetivo específico deste trabalho, tem-se a confirmação da fundamentação teórica apresentada na literatura, [11, 20, 29, 33] a partir dos experimentos. Durante o processo de aprimoramento dos parâmetros, percebeu-se que foi necessário aumentar a população e o número de gerações, mas apenas isso não foi o suficiente para alcançar os melhores resultados. A fim de refinar os resultados obtidos, foi importante aumentar o N da seleção por torneio, e deste modo, obteve-se as melhores soluções, ratificando o que afirma a literatura [11, 29, 33] sobre a importância da seleção por torneio e seus ajustes no Algoritmo Evolutivo.

Referências

1. P. Bradley, K. M. S. Misura, and D. Baker; *Toward high-resolution de novo structure prediction for small proteins*, (Science, 309(5742): 1868-1871, 2005). Citado 2 vezes nas páginas 7 e 9.

2. C. R. S. Brasil; *Algoritmo evolutivo de muitos objetivos para predição ab initio de estrutura de proteínas*, (Tese de Doutorado, Universidade de São Paulo, 2012). Citado 3 vezes nas páginas 6, 9 e 10.

3. C. R. S. Brasil and J. M. Dias; *Um estudo sobre funções de energia com modelo HP-2D no algoritmo de colônia de formigas com método de backtracking*, (Artigo, 2019). Citado 2 vezes nas páginas 22 e 32.

4. Britannica, The Editors of Encyclopaedia.; *NP-complete problem*, (Encyclopedia Britannica , 2021). Citado na página 6.

5. Cavalcanti, D. and Brasil, C. R. S.; *Algoritmos de inteligência de enxame com busca por pull move aplicados ao problema de predição de estrutura de proteínas*, (XII Encontro Acadêmico de Modelagem Computacional (EAMC2019), pp. 111–12, 2019). Citado na página 6.

6. V. Cutello, G. Narzisi, and G. Nicosia; *A multiobjective evolutionary approach to the protein structure prediction problem*, (J. R. Soc. Interface, 83:1-13, 2005). Citado 2 vezes nas páginas 7 e 9.

7. C. Darwin, *Evolução de vertebrados. Mutação genética.* Citado 2 vezes nas páginas 16 e 19.

8. Dias, J. M. and Brasil, C. R. S.; *Comparando algoritmos de otimização computacional aplicados ao problema de predição de estruturas proteicas com modelo hp-2d*, (Revista Brasileira de Computação Aplicada 9(3): 87–9, 2017). Citado na página 6.

9. F. DiMaio, T.C. Terwilliger, R, J. Read, and et al; *Improved molecular replacement by density – and energy-guided protein structure optimization*, (Nature, 473(7348):540-543-2011). Citado 2 vezes nas páginas 7 e 9.

10. Duc, D. D., Anh, V. T. N., Dinh, P. T. and Linh-Trung, N.; *An eficient ant colony optimization algorithm for protein structure prediction*, (12th International Symposium on Medical Information and Communication Technology (ISMICT), 2018). Citado na página 6.

11. A. V. Eremeev, *On Proportions of Fit Individuals in Population of Mutation-Based Evolutionary Algorithm with Tournament Selection*, (Evolutionary Computation, vol. 26, no. 2, pp. 269-297, June 2018) Citado 2 vezes nas páginas 21 e 32.

12.D. B. Fogel; *An introduction to simulated evolution. IEEE Transactions on Neural Networks*, (v.5, n.1., 1994). Citado na página 16.

13.P. H. R. Gabriel, A. C. B. Delbem; *Fundamentos de Algoritmos Evolutivos*, (Universidade de São Paulo, 2008). Citado na página 14.

14.P. H. R. Gabriel; *Algoritmos evolutivos e modelos simplicados de proteínas para predição de estruturas terciárias*, (Dissertação de Mestrado em ciência da computação, Instituto de Ciências Matemáticas e de Computação, 2012). Citado na página 6.

15.G. Helles; *A comparative study of the reported performance of ab initio protein structure prediction algorithms*, (Journal of the Royal Society Interface the Royal Society, 5(21):387-396, 2008). Citado 2 vezes nas páginas 7 e 9.

16.Hu, X.-M., Zhang, J. and Li, Y; *Flexible protein folding by ant colony optimization*, (Vol. 151, Springer-Verlag Berlin Heidelberg, 2008). Citado na página 6.

17.Huang, C., Yang, X. and He, Z; *Protein folding simulations of 2d hp model by the genetic algorithm based on optimal secondary structures*, (Computational Biology and Chemistry, 34(3): 137–142, 2010). Citado na página 6.

18.Lau, K. F. and Dill, K. A. (1989). *A lattice statistical mechanics model of the conformational and sequence spaces of proteins*, Macromolecules 22(10): 3986–3997. Citado na página 11.

19.A. Leaver-Fay, M. Tyka, S. M. Lewis, and et al. *ROSETTA3: an object-oriented software suíte for the simulation and design of macromolecules. Methods in enzymology*, (487:545-574,574, 2011). Citado 2 vezes nas páginas 7 e 9.

20.Y. Liu, D. Gong, J. Sun and Y. Jin, *A Many-Objective Evolutionary Algorithm Using A One-by-One Selection Strategy*, in IEEE Transactions on Cybernetics, vol. 47, no. 9, pp. 2689-2702, Sept. 2017. Citado 2 vezes nas páginas 21 e 32.

21.Maher, B., Albrecht, A., Loomes, M., Yang, X.-S. and Steinhöfel, K., *A firefly-inspired method for protein structure prediction in lattice models*, Biomolecules 1(4):56–57. 2014. Citado na página 6.

22.Z. Michalewicz; *Genetic algorithms + data structures = Evolution programs*, (3 ed. Berlin: Springer, 1996). Citado na página 16.

23.D. L. Nelson and M; *Cox. Lehninger Principles of Biochemistry*, (2004). Citado na página 9.

24.Roeva, O., Fidanova, S., Paprzycki, M. (2015). *Population Size Influence on the Genetic and Ant Algorithms Performance in Case of Cultivation Process Modeling.* In: Fidanova, S. (eds) Recent Advances in Computational Optimization. Studies in Computational Intelligence, vol 580. Springer. Citado na página 20.

25.E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus; *Protein folding bottlenecks: A lattice Monte Carlo simulation. Physical Review Letters*, (7(12):1665-1668, Sept. 1991). Citado na página 10.

26.Shmygelska, A., Hernández, R. A. and Hoos, H. H. *An ant colony optimization algorithm forthe 2d hp protein folding problem*, (Proceedings of the Third International Workshop on Ant Algorithms, p. 40–53, 2002). Citado na página 6.

27.Shmygelska, A. and Hoos, H. H. *An improved ant colony optimization algorithm for the 2dhp protein folding problem*, (Conference of the Canadian Society for Computational Studies of Intelligence Canadian AI 2003: Advances in Articial Intelligence, pp. 400–417, 2003). Citado na página 6.

28.Shmygelska, A. and Hoos, H. H. *An ant colony optimisation algorithm for the 2d and 3d hydrophobic polar protein folding problem*, (BMC Bioinformatics, Vol. 6, 2005). Citado na página 6.

29.A. Shukla, H. M. Pandey, D Mehrotra; *Comparative review of selection techniques in genetic algorithm*, (Article, 2017). Citado 3 vezes nas páginas 16, 20 e 32.

30.Song, J., Cheng, J. and Zheng, T. *Protein 3d hp model folding simulation based on aco*, (Sixth International Conference on Intelligent Systems Design and Applications, Vol. 6, 2017). Citado na página 6.

31.Thalheim, T., Merkle, D. and Middendorf, M; *Protein folding in the hp-model solved with a hybrid population based aco algorithm*, (AENG International Journal of Computer Science, Vol. 3, 2008). Citado na página 6.

32.D. Xu, J. Zhang, A. Roy, and Y. Zhang; *Automated protein structure modeling in casp9 by i-tasser pipeline combined with quark-based ab initio folding nd fg-md-based structure refinement*, (Proteins: Structure, Function, and Bioinformatics, 79(S10):147-160, 2011). Citado 2 vezes nas páginas 7 e 9.

33.S. L. Yadav; *Comparative Study of Different Selection Techniques in Genetic Algorithm*, (Article, 2017). Citado 3 vezes nas páginas 16, 20 e 32.

34.Y. Zhang; *Template-based modeling and free modeling by i-tasser in casp7*, (Proteins, 69:108-117, 2007). Citado 2 vezes nas páginas 7 e 9.

35.Zhang, Y., Wu, L. and Wang, S.; *Solving two-dimensional hp model by firefly algorithm and simplied energy function*, (Mathematical Problems in Engineering, 2013). Citado 2 vezes nas páginas 6 e 12.