Revista Brasileira de Computação Aplicada, Abril, 2020

DOI: 10.5335/rbca.v12i1.10287
Vol. 12, Nº 1, pp. 122–133
Homepage: seer.upf.br/index.php/rbca/index

ISSN 2176-6649

ARTIGO ORIGINAL

Um estudo sobre funções de energia com modelo HP-2D no algoritmo de colônia de formigas com método de *backtracking*

A study on energy functions with the 2D/HP model in the ant colony optimization with backtracking method

Christiane R. S. Brasil[1] and Julia M. Dias[1]

[1]Universidade Federal de Uberlândia

*christiane@ufu.br; juliamanfrindias@gmail.com

Recebido: 27/11/2019. **Revisado:** 12/02/2020. **Aceito:** 23/03/2020.

Resumo

Este trabalho aplica um algoritmo de otimização computacional: o Algoritmo de Colônia de Formigas (ACO, do inglês *Ant Colony Optimization*), com o método de *backtracking* para correção de soluções infactíveis para o problema de predição de estruturas de proteínas, considerado um problema de alta complexidade. Este problema é um grande desafio e uma importante questão nesta área de pesquisa, uma vez que a partir da estrutura conhecida de uma proteína há a possibilidade do conhecimento de suas funcionalidades serem exploradas, colaborando efetivamente para o avanço no desenvolvimento de novos fármacos. Neste sentido, o objetivo principal deste trabalho foi analisar o desempenho do algoritmo ACO com método de *backtracking* para o problema de PSP usando duas funções de energia diferentes no modelo de representação HP-2D em uma abordagem *ab initio*, isto é, sem nenhum conhecimento prévio. Utilizou-se a energia de Lau e Dill, e a energia simplificada, ambas encontradas na literatura, a fim de realizar uma comparação entre elas do ponto de vista computacional e bioquímico. Os experimentos mostraram bons resultados do ACO, principalmente com a energia simplificada.

Palavras-Chave: Função de Energia; Modelo HP-2D; Otimização de Colônia de Formigas; Predição de Estrutura de Proteínas.

Abstract

This work applies a computational optimization algorithm: the Ant Colony Optimization (ACO) with backtracking method for correction of infeasible solutions to the Protein Structures Prediction problem (PSP), considered a problem of high complexity. This problem is a great challenge and an important issue in this research area, since from the known structure of a protein there is the possibility of its functionalities being explored, effectively collaborating for advances in the development of new medicines. In this sense, the main objective of this work was to analyze the performance of the ACO with backtracking method for PSP using two different energies in the 2D/HP representation model applying an ab initio approach, that is, without any prior knowledge. The Lau and Dill energy and the simplified energy, both found in the literature, were used in order to make a computational and biochemical comparison between them. The experiments showed good results of the ACO, mainly with the simplified energy.

Keywords: Ant Colony Optimization; Energy Function; Protein Structures Prediction; 2D/HP Model.

1 Introdução

Existem diversos problemas no mundo que gastam um tempo inviável para encontrar a sua melhor resposta, ou ainda nem obtê-la, uma vez que o espaço de busca para soluções é tão vasto que pode custar anos, ou mesmo séculos, para retornar a solução correta. Isto ocorre porque o tempo para resolver tais problemas usando algum algoritmo determinístico cresce rapidamente à medida que a instância aumenta. Neste contexto, um desses grandes desafios da atualidade é o problema de predição de estruturas de proteína (PSP, do inglês *Protein Structures Prediction Problem*), que visa prever estruturas proteicas ainda desconhecidas na natureza. Esse problema pode ser descrito computacionalmente como NP-completo (Crescenzi et al., 1998).

A estrutura de uma proteína está diretamente relacionada à função que esta exerce nos organismos vivos (Cox and Doudna, 2012). Por conseguinte, a partir do conhecimento das funcionalidades de uma proteína é possível avançar no entendimento de muitas doenças. Deste modo, o estudo e a compreensão de doenças afetam positivamente na elaboração de novos fármacos que possam atuar de maneira efetiva no tratamento dessas doenças.

Neste sentido, o PSP é um problema de grande importância dentro da área de Biologia Molecular, bem como para a humanidade em questões não resolvidas da saúde. Contudo, métodos convencionais usados (cristalografia e ressonância nuclear magnética) para descobrir a estrutura das proteínas são ainda ineficientes em termo de tempo e custo, assim como algoritmos determinísticos também são não recomendados pela alta complexidade do problema. Deste modo, pesquisadores de diversas áreas (físicos, bioquímicos, biomédicos, cientistas da computação, entre outros) vêm unindo esforços para otimizar a busca por estruturas de proteínas. A saber, a estrutura de uma proteína é obtida quando sua conformação se encontra no estado mais estável, ou seja, com a menor energia possível. Portanto, encontrar a energia mínima de uma estrutura de proteína é um forte indicativo que a estrutura mais estável foi obtida. Desta maneira, existe uma área de pesquisa que trabalha com algoritmos de otimização computacional (Yang, 2008) aplicados ao problema de PSP que busca por soluções com energia mínima para uma determinada proteína, ou a aproximação de uma energia mínima. Este valor da energia pode ser calculado de diversos modos, uma vez que existem muitas energias envolvidas no processo de formação de uma estrutura proteica. Essa abordagem baseada somente na sequência proteica e nas energias envolvidas é denominada *ab initio*, ou seja, sem nenhum conhecimento prévio, almeja-se encontrar a estrutura de uma proteína considerando apenas estes critérios (Brasil, 2012). Essa área tem crescido significativamente ao longo dos anos, apresentando bons resultados, tais como os apresentados por Shmygelska et al. (2002), Shmygelska and Hoos (2003, 2005), Song et al. (2006), Hu et al. (2008), Thalheim et al. (2008), Huang et al. (2010), Brasil (2012), Gabriel (2012), Zhang et al. (2013), Maher et al. (2014), Dias and Brasil (2017), Duc et al. (2018), Cavalcanti and Brasil (2019), entre tantos outros.

O objetivo deste trabalho foi desenvolver uma análise comparativa entre duas energias no problema de predição do ponto de vista computacional e bioquímico: (i) a energia de Lau and Dill (1989), largamente conhecida, e (ii) a energia simplificada (Zhang et al., 2013), ainda não muito aplicada. Para tal, foi desenvolvido o algoritmo bioinspirado chamado Algoritmo de Otimização de Colônia de Formigas (ACO, do inglês *Ant Colony Optimization*) (Dorigo and Gambardella, 1997). Foi escolhido este algoritmo por ser muito utilizado na área de PSP com sucesso, e por ainda não ter sido testado com a energia simplificada. A proposta foi analisar se com essa função de energia os resultados seriam ainda melhores. Neste trabalho, foi adotado um procedimento com mecanismo de *backtracking* para corrigir soluções infactíveis, embora acrescente algum tempo computacional, para garantir que somente soluções possíveis fossem consideradas. Este procedimento colabora para uma convergência mais eficiente para a melhor solução, conforme mostrado em outros trabalhos na literatura, como Hu et al. (2008) e Dias and Brasil (2017). O Algoritmo de Otimização de Colônia de Formigas foi aplicado para o conjunto de proteínas usado no trabalho de Zhang et al. (2013). Este trabalho foi selecionado como referência, pois ele aplica a função de energia simplificada, criada pelo autor, além da energia de Lau e Dill. No entanto, diferente da proposta do presente trabalho, Zhang tratou o problema com o Algoritmo de Vagalumes (Zhang et al., 2013). O modelo de representação usado neste trabalho foi HP-2D.

2 Problema de predição de estruturas de proteínas

Todo ser vivo é constituído de uma ou mais proteínas (Brasil, 2012, Cox and Doudna, 2012). Uma proteína pode ser definida como uma cadeia polipeptídica, que é composta por um conjunto de aminoácidos (ou monômeros). Cada aminoácido contém um grupo amino **–NH**, um grupo carboxila **–COOH** e um radical **R**, também chamado de cadeia lateral (Fig. 1). É a cadeia lateral R que torna um aminoácido diferente do outro.

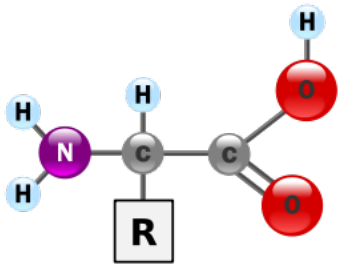

Figura 1: Representação de um aminoácido (Brasil, 2012).

Na natureza existem vinte e dois tipos de aminoá-

cidos, mas apenas vinte são representados no código genético universal. Dois aminoácidos são conectados por meio de uma ligação peptídica, que é realizada pela união do grupo amino de um aminoácido com o grupo carboxila de outro aminoácido, com a formação de um dipeptídeo e a perda de uma molécula de água (Fig. 2).

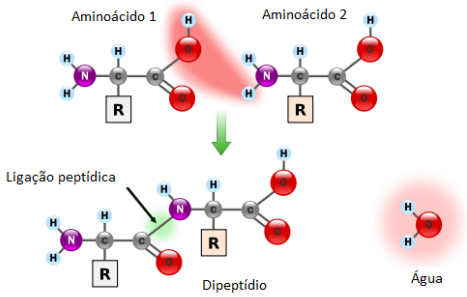

Figura 2: Ligação peptídica entre dois aminoácidos (Brasil, 2012).

A diferença entre as proteínas está na ordem em que os aminoácidos estão ligados entre si, ou seja, para que duas proteínas sejam consideradas iguais elas devem ter a mesma sequência de aminoácidos. Cada sequência corresponde a uma organização molecular, estabelecendo a maneira em que os aminoácidos interagem entre si e/ou com o meio. Existem níveis de organização molecular, que são denominadas estruturas (Brasil, 2012).

O problema de predição de estruturas de proteínas (PSP) visa prever o formato 3D em que as proteínas se encontram na natureza, uma vez que há inúmeras proteínas em que as estruturas tridimensionais ainda são desconhecidas. Essas estruturas são importantes devido sua relação com as funcionalidades em organismos vivos (Brasil, 2012, Cox and Doudna, 2012). Portanto, determinar a estrutura de uma dada proteína pode auxiliar na compreensão de diversas utilidades das mesmas ou até mesmo disfunções, que geram doenças, muitas vezes até incuráveis. Deste modo, estas descobertas podem contribuir fortemente para o surgimento de novos fármacos. Sob o ponto de vista matemático, o PSP é um problema combinatório, e sob o ponto de vista computacional, pode ser classificado como problema NP-Difícil, no qual o espaço de busca cresce exponencialmente em relação ao tamanho da proteína (Hart and Newman, 2001).

Neste contexto, a estrutura tridimensional de uma proteína, que representa a conformação 3D em que esta é encontrada na natureza, é definida pelo dobramento (enovelamento) da sequência de aminoácidos que a compõem, que ocorre devido a vários critérios intra e intermoleculares, tais como as diversas energias (campos de força) envolvidas neste processo.

Existem diferentes métodos para predizer a estrutura de uma proteína. Os métodos convencionais são a cristalografia de raio-x e a ressonância nuclear magnética. No entanto, esses métodos apresentam alto custo computacional e financeiro. Neste sentido, algoritmos de otimização computacional surgiram como uma alternativa a estes métodos ainda restritos. As abordagens computacionais para o problema de PSP são *ab initio*, semi *ab initio*, *threading* e homologia (Brasil, 2012).

A abordagem *ab initio* gera possíveis soluções a partir da sequência dos aminoácidos e das energias envolvidas no processo de dobramento, ou seja, sem utilizar nenhum conhecimento prévio. Para tal, essa abordagem trabalha com algoritmos de otimização computacional que lidam com duas questões importantes: (i) a especificação da função de minimização e (ii) e a escolha do algoritmo de busca. As funções de minimização são baseadas em leis físicas presentes na estabilização do sistema em questão (campos de força), sendo este um dos maiores desafios do problema, uma vez que se conhece a existência de diversas energias que atuam na molécula (interna e externamente).

A abordagem semi *ab initio* acessa banco de dados de estruturas de proteínas a fim de realizar uma busca, com o objetivo de gerar novas estruturas a partir da combinação de outras menores ou segmentos de estruturas. A abordagem *threading* é usada quando uma proteína não tem sequência com alta similaridade com outra encontrada em uma base de dados, mas apresenta estruturas tridimensionais muito semelhantes. Isto é feito com alinhamento das sequências para que se detecte alguma similaridade, mesmo não muito alta, e a partir disto são procuradas estruturas parecidas. Na modelagem por homologia este alinhamento deseja alta similaridade na sequência de aminoácidos, buscando regiões de máxima similaridade.

Neste trabalho o algoritmo utiliza a abordagem *ab initio* e trabalha com duas possíveis funções de energia (a energia de Lau e Dill e a energia simplificada). O modelo de representação da proteína aplicado neste problema é descrito a seguir.

3 Modelo HP

Os modelos *ab initio* de representação da proteína são: lattice, *off*-lattice e *full atom* (Brasil, 2012). O modelo lattice é um reticulado (malha) quadrado ou cúbico. Este modelo tem uma representação bastante simples, em que cada vértice da malha posiciona um aminoácido da sequência, não sendo necessário mostrar a estrutura interna do aminoácido. Essas simplificações reduzem significativamente os custos computacionais, ainda preservando características relevantes do processo de enovelamento. Resumidamente, o modelo *off*-lattice é mais próximo de uma estrutura real que o lattice, pois permite trabalhar com alguns ângulos entre os átomos que constituem os aminoácidos, enquanto o modelo *full-atom* trabalha com mais possibilidades de ângulos internos da molécula, aumentando o realismo na representação. Neste trabalho o modelo de representação escolhido foi o lattice, pela simplicidade e preservação de critérios importantes para predição.

O modelo lattice hidrofóbico-polar (HP), criado por Lau and Dill (1989), é uma representação simplificada da estrutura de uma proteína. Este modelo é caracteri-

zado pelo uso de malhas, que podem ter duas ou três dimensões. Neste trabalho, o modelo HP-2D (do inglês, 2D/HP) foi utilizado, que trata de duas dimensões. Este modelo considera a hidrofobicidade dos aminoácidos para representação computacional. Como fora mencionado anteriormente, uma proteína é composta por uma sequência de aminoácidos. Estes podem ser: (i) hidrofóbicos, ou apolares (H), e são aqueles que não iteragem com o solvente (água), tendendo a se concentrar no interior da molécula; (ii) ou aminoácidos hidrofílicos, ou polares (P), que interagem facilmente com o solvente. Neste sentido, as interações entre aminoácidos hidrofóbicos de uma proteína são fundamentais no processo de dobramento da mesma, uma vez que influenciam no modo em que os aminoácidos H "fogem"da água, voltando-se para o interior da molécula, e os aminoácidos P tendem a estar na superfície molecular assumindo maior contato com a água. Desta maneira, o efeito hidrofóbico contribui significativamente na conformação 3D da proteína. Na Tabela 1 são apresentados: o conjunto de vinte aminoácidos, a abreviatura e a classificação de cada um em relação à hidrofobicidade.

Tabela 1: Conjunto de vinte aminoácidos e sua classificação em relação à hidrofobicidade (adaptado de Lau and Dill (1989)).

Aminoácido	Abreviatura	Classificação
Ácido Aspártico	Asp	Polar, hidrofílico (P)
Ácido Glutâmico	Glu	Polar, hidrofílico (P)
Alanina	Ala	Hidrofóbico (H)
Arginina	Arg	Polar, hidrofílico (P)
Asparagina	Asn	Polar, hidrofílico (P)
Cisteína	Cys	Polar, hidrofílico (P)
Felalanina	Phe	Hidrofóbico (H)
Glicina	Gly	Polar, hidrofílico (P)
Glutamina	Gln	Polar, hidrofílico (P)
Histidina	His	Polar, hidrofílico (P)
Isoleucina	Ile	Hidrofóbico (H)
Leucina	Leu	Hidrofóbico (H)
Lisina	Lys	Polar, hidrofílico (P)
Metionina	Met	Hidrofóbico (H)
Prolina	Pro	Hidrofóbico (H)
Serina	Ser	Polar, hidrofílico (P)
Tirosina	Tyr	Polar, hidrofílico (P)
Treonina	Thr	Polar, hidrofílico (P)
Triptofano	Trp	Hidrofóbico (H)
Valina	Val	Hidrofóbico (H)

A partir da Tabela 1 é possível converter qualquer sequência de aminoácidos de uma proteína para a representação no modelo HP. Deste modo, a entrada do algoritmo pode ser a sequência de aminoácidos, que será internamente convertida para a nomenclatura H e P.

Nas próximas subseções serão descritas as funções de energia relacionadas ao modelo HP-2D utilizadas neste trabalho.

4 Energia de Lau e Dill no modelo HP

Uma função de energia muito comum a ser utilizada no modelo HP é a energia de Lau e Dill. Essa energia é baseada na quantidade de aminoácidos hidrofóbicos vizinhos não consecutivos (Lau and Dill, 1989). A Fig. 3 mostra uma sequência de aminoácidos de uma certa proteína, configurada no modelo HP em uma malha 2D. As linhas pontilhadas indicam os aminoácidos hidrofóbicos vizinhos não consecutivos na malha. Os aminoácidos consecutivos não são considerados no cálculo da energia. A energia de Lau e Dill é o inverso do número de ligações hidrofóbicas não consecutivas. Por exemplo, na Fig. 3 há três ligações hidrofóbicas não consecutivas, logo, o valor desta energia é -3.

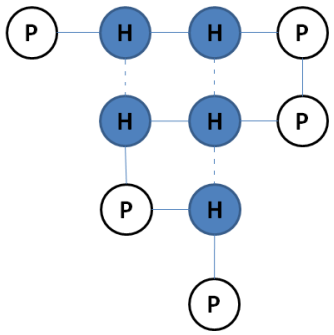

Figura 3: Proteína representada usando o modelo HP-2D.

Essa energia de conformação é dada pela Eq. (1):

$$E = \beta_{ij} \sum \sigma(r_i r_j) \tag{1}$$

Em que β assume o valor 1 se os aminoácidos são do tipo H, e 0, caso contrário; a função σ assume o valor -1 se os aminoácidos r_i e r_j são vizinhos não conectados, e 0, caso não sejam vizinhos ou sejam vizinhos conectados. Portanto, a energia de uma conformação usando este modelo é obtida considerando as interações H-H de aminoácidos não conectados (isto é, não consecutivos). Deste modo, pelo cálculo apresentado pode-se perceber que essa energia é inversamente proporcional à quantidade de interações H-H, lembrando que quanto menor a energia, mais estável tende ser a

estrutura, que é justamente o que se busca. A energia de Lau e Dill aplicada no modelo HP para o PSP é largamente utilizada em diversos trabalhos na literatura, tais como Huang et al. (2010), Gabriel (2012), Citrolo and Mauri (2013), Maher et al. (2014), Dias and Brasil (2017), Cavalcanti and Brasil (2019), sendo aplicada também neste trabalho.

5 Energia simplificada do modelo HP

A energia simplificada é calculada pela distância entre todos os aminoácidos hidrofóbicos não-consecutivos representados no modelo HP-2D. Essa energia é apresentada originalmente no artigo de Zhang et al. (2013). Para este cálculo, deve-se considerar os aminoácidos em coordenadas cartesianas, nos quais cada aminoácido hidrofóbico é representado por um ponto (x, y). Para cada ponto cartesiano é calculada a distância euclidiana entre todos os outros pontos. Os resultados são armazenados em uma matriz de distâncias, e a energia simplificada é calculada pela soma dos valores da matriz triangular superior (ou inferior, uma vez que ambos terão os mesmos valores). Considerando um exemplo de alguns pontos onde os aminoácidos hidrofóbicos estão alocados nas seguintes coordenadas: p1 (2,1), p2 (3,1), p3 (0,4) e p4 (3,2), tem-se a seguinte matriz de distância (Tabela 2):

Tabela 2: Exemplo de matriz de distância.

	p1	p2	p3	p4
p1	0	1	3.6	1.41
p2	1	0	4.24	1
p3	3.6	4.24	0	3.6
p4	1.41	1	3.6	0

Portanto, o cálculo da energia simplificada para este exemplo será: Es = 2*1 + 2*3.60 + 1.41 + 4.24 = 14.85. Essa energia da conformação é dada pela Eq. (2):

$$Es = \sum d(r_i r_j) \tag{2}$$

Em que d é a distância euclidiana dos aminoácidos r_i e r_j, enfatizando que a soma total considera somente os valores da matriz triangular superior (ou inferior).

A seguir, é explicado o algoritmo de otimização desenvolvido para aplicar ao problema de PSP, usando modelo HP-2D.

6 Algoritmo de Otimização de Colônia de Formigas

O Algoritmo de Otimização de Colônia de Formigas (ACO, do inglês *Ant Colony Optimization*) foi criado por Dorigo and Gambardella (1997). A proposta do algoritmo é imitar o comportamento das formigas na natureza. Em seu habitat natural, as formigas inicialmente buscam por alimentos seguindo caminhos aleatórios, os quais, com o tempo, deixam de ser aleatórios devido à comunicação estabelecida entre as formigas por meio de uma substância química denominada feromônio.

Deste modo, o ACO gera soluções probabilísticas a partir de um processo computacional que mimetiza as formigas na busca pelo seu alimento com a ajuda do feromônio. O algoritmo inicial proposto tem algumas variações, como o ACS (*Ant Colony System*) e o MMAS (*Min-Max Ant System*). A principal diferença entre essas variações está na fórmula de atualização do feromônio. No ACS, ocorre a atualização do feromônio local e globalmente. Isto significa que cada formiga, após concluir o processo de busca pela solução, depositará feromônio, sendo uma atualização local. Após todas as formigas gerarem suas soluções, somente a melhor formiga encontrada na população até o momento atualizará o feromônio, sendo esta uma atualização global. O MMAS foi desenvolvido por Stützle and Hoos (1998), e também tem a atualização do feromônio por uma única formiga, que pode ser a melhor formiga da iteração ou a melhor formiga encontrada no momento. No MMAS o valor do feromônio não ultrapassa os limites de valor mínimo e máximo pré-estabelecidos, a fim de evitar uma convergência prematura. Essa abordagem foi aplicada neste trabalho.

O cálculo da probabilidade usado para decidir qual caminho seguir é dado pela Eq. (3):

$$P_{ij} = \frac{(\tau_{ij})^{\alpha}(\eta_{ij})^{\beta}}{\sum_{x=1}^{N}(\tau_{ix})^{\alpha}(\eta_{ix})^{\beta}} \tag{3}$$

Em que i e j significam, respectivamente, o ponto em que a formiga está localizada e o ponto que a formiga pode ir, τ_{ij} é o valor do feromônio no local pretendido, η_{ij} é o valor da informação heurística, os símbolos α e β representam, respectivamente, quão importantes o feromônio e a informação heurística serão para o cálculo, o valor x pertence aos números naturais, e N é o tamanho do conjunto de possíveis caminhos que a formiga pode escolher.

A atualização do feromônio é dada pela Eq. (4):

$$\tau_{ij} = \rho(\tau_{ij}) + \rho\Delta Best \tag{4}$$

Em que τ_{ij} é o valor do feromônio no local pretendido; o símbolo ρ é o parâmetro que representa a taxa de evaporação do feromônio no local e seu valor pode variar de 0 a 1; $\Delta Best = Esol/Eopt$ é a energia obtida a partir do caminho que a formiga tem escolhido, e opt é a melhor energia conhecida da proteína. Se essa informação for desconhecida, a menor energia encontrada até o momento da execução é considerada.

O algoritmo começa com uma quantidade inicial de feromônio em toda população de formigas. As formigas estão representadas pela variável k, e variam até m. Todas as formigas constroem uma provável solução por meio da Eq. (3), são avaliadas pelo cálculo da energia (utilizando a Eq. (1) ou a Eq. (2) neste trabalho específico), e para cada formiga a quantidade de feromônio é

atualizada com a Eq. (4). Ao final de um certo número de iterações, é retornada a melhor solução encontrada. O pseudocódigo do ACO desenvolvido neste trabalho é descrito no Algoritmo 1.

Algoritmo 1: Algoritmo de Otimização de Colônia de Formigas.

```
1: Inicializar feromônio de toda população de formigas
2: enquanto número de iterações não atingido faça
3: para k = 1 até m faça
4: Construir possível solução, usando Eq. (3)
5: Avaliá-la com a Eq. (1) ou a Eq. (2)
6: Atualizar feromônio, utilizando Eq. (4)
7: fim para
8: Guardar melhor solução até o momento
9: fim enquanto
10: Retornar melhor solução encontrada
```

A seguir, é descrito o processo de movimentação da formiga, com relação ao modelo HP-2D, e o processo de *backtracking*, executado sempre que necessário.

6.1 Movimentação da Formiga

Considerando uma formiga localizada na malha 2D no modelo HP, a Fig. 4 mostra a formação de uma possível configuração de representação da proteína cuja sequência é HPHHPH. Note que o aminoácido H, destacado em vermelho, indica o ponto de partida dos movimentos da formiga. O ponto de partida, nesta implementação, está no centro na malha. As setas em cinza indicam quais foram os locais por onde a formiga passou. Ao passar por um local (sítio), a formiga "deixa"um aminoácido, e posteriormente, seleciona qual o próximo local (a seta azul indica as possibilidades) para o aminoácido seguinte da cadeia.

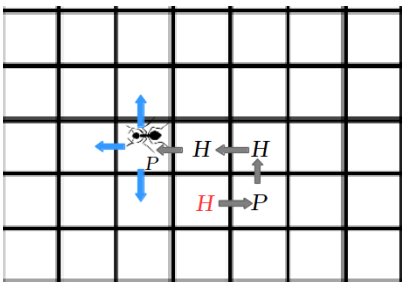

Figura 4: Deslocamento da formiga.

Deste modo, para percorrer a malha a formiga pode escolher entre três movimentos possíveis: Frente (F), Direita (D) ou Esquerda (E). São analisados os cálculos de probabilidade de ir para uma destas opções, baseados na Eq. (1). Ao escolher um desses movimentos, a orientação na malha é alterada. Por exemplo, se a formiga

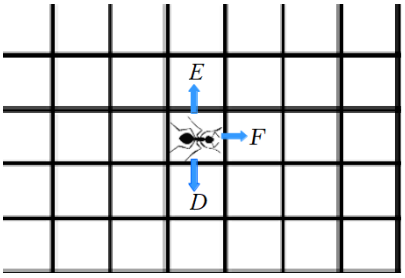

Figura 5: Formiga com orientação na horizontal.

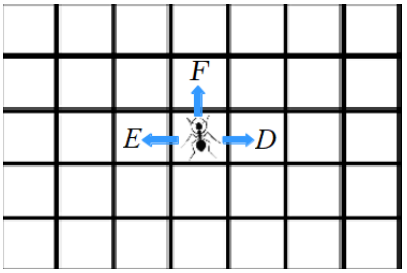

Figura 6: Formiga com orientação na vertical.

que estava no ponto de origem no sentido horizontal (Fig. 5) escolher o movimento de ir para esquerda, logo após realizar esse movimento a formiga muda de orientação, estando no sentido vertical (Fig. 6).

6.2 Procedimento de *Backtracking*

Neste trabalho, foi aplicado o procedimento de *backtracking* sempre que a formiga encontrasse uma posição que em haveria colisão.

Na malha 2D não é permitido que dois aminoácidos ocupem o mesmo lugar ao mesmo tempo. Quando isso ocorre, há uma colisão. Para prevenir essa colisão, a formiga verifica se há a possibilidade de ir para, pelo menos, um dos três sítios possíveis, de acordo com os movimentos (F, D ou E), sabendo que a formiga não anda para trás normalmente. Caso não seja possível (ou seja, os três sítios estarem ocupados), a formiga inicia o processo de *backtracking*. Este processo consiste em marcar a posição atual da formiga como uma posição inválida, de modo que a formiga não acesse novamente este sítio. Em seguida, a formiga retrocede para posição anterior à sua posição atual. O *backtracking* pode fazer voltar a posições anteriores, quantas vezes forem necessárias, até encontrar uma posição que impossibilite a ocorrência de colisão.

Neste trabalho, o ACO foi elaborado em duas versões, com a finalidade de comparar a energia de Lau e Dill com a simplificada, uma vez que este algoritmo vem apresentando bons resultados para o PSP com a energia de Lau e Dill em diversos trabalhos, tais como: (Shmygelska et al., 2002, Shmygelska and Hoos, 2003, 2005, Song et al., 2006, Thalheim et al., 2008, Gabriel, 2012, Dias and Brasil, 2017, Cavalcanti and Brasil, 2019). A ideia foi verificar se com função de energia simplificada

os resultados seriam ainda melhores, analisando os valores de energia, o tempo computacional e o resultado visual das estruturas obtidas.

7 Trabalhos correlatos

Shmygelska et al. (2002) foram os primeiros pesquisadores a aplicarem o ACO para o problema de PSP no modelo HP-2D. Em 2003, os mesmos pesquisadores desenvolveram uma melhoria no algoritmo, focando na busca local para o próximo movimento (Shmygelska and Hoos, 2003), e em 2005, trabalharam com essa abordagem agregando ao modelo HP-3D (Shmygelska and Hoos, 2005). Chu and Zomaya (2005) elaboraram o ACO com o paradigma de programação distribuída, aplicando ao modelo HP-2D e 3D. Hu et al. (2008) propuseram uma nova versão do ACO para o PSP com modelo HP-2D, chamado de *Flexible Ant Colony Algorithm* (FAC), que utiliza um método de *backtracking* para reparar conformações inválidas geradas em tempo de execução. Citrolo and Mauri (2013) apresentaram um algoritmo que hibridizava ACO e Monte Carlo para o PSP no modelo HP . Liu and Yang (2014) desenvolveram uma nova abordagem do ACO para o PSP com modelo HP usando o método de busca por *pull moves* como técnica de busca local e o método de cópia parcial para gerenciar conformações inviáveis: a nova abordagem foi chamada de *Heuristic Ant Colony Optimization* (HACO). Thilagavathi and Amudha (2015) elaboraram uma versão do ACO baseada em classificação, chamada de *Rank Based Ant Colony Optimization*. Llanes-Castro et al. (2016) propuseram o ACO com paralelismo, aplicada ao PSP no modelo HP, usando *Compute Unified Device Architecture* (CUDA). Recentemente, Wang (2018) propôs uma abordagem híbrida do ACO com *Artificial Fish Swarm Algorithm* para o PSP com modelo HP-2D. Como pode se observar pela literatura, o ACO é um algoritmo bioinspirado largamente aplicado para o problema de PSP usando o modelo HP, seja usando-o de modo puro ou híbrido com outros algoritmos de otimização.

8 Experimentos

Neste trabalho, foram implementadas duas versões do ACO para o problema de PSP com o modelo HP-2D: uma com a energia de Lau e Dill, e outra com a energia simplificada. Em ambas as versões, foram utilizados os seguintes parâmetros: *NumF* para representar o número de formigas, *NumI* para indicar o número de iterações,α para expressar o importância do feromônio, β para indicar a importância da heurística da informação, e ρ para atualização do feromônio. Os resultados foram obtidos com *NumF* = 100, *NumI* = 500, α = 1, β = 1 e ρ = 0.5 (Zhang et al., 2013, Dias and Brasil, 2017). O conjunto de proteínas foi advindo do artigo de Zhang et al. (2013), que foi aplicado originalmente no Algoritmo de Vagalumes para PSP com o modelo HP-2D. A Tabela 3 mostra as proteínas usadas, com suas respectivas sequências de aminoácidos hidrofóbicos (H) e polares (P), e o tamanho de cada sequência.

Tabela 3: Conjunto de proteínas do trabalho de Zhang et al. (2013)

ID	Tam	Sequência
1	18	PHHHHPPPPPPHHHHHHH
2	26	PPHHHHHHHPPHHHHHHHHHHHH HHHPP
3	30	PPPHHHHHHHPPPPHHHHHHPPP PHHHHHPP
4	39	PPHHHHPPPPPHHHHHHHHHH HHHHHHPPPPPPHHHHPP
5	42	PPHHHHHHHHHHHHHPPPPHH HHHHHHHHHHHPPPPPPPHHH
6	49	HHHHHHPPPPPPPPPPHHHHH HHHHHHHHHHPPPPHHHPPPPP PPPHH
7	53	PHHHHPPPPPPPHHPPPPPPP PPHHHHHHHHHHPPPPPHHH HHHHHPPHHH
8	66	HPPPPHHHHHHHHHHPPPPPHH HHHHHHHHHHHHHHHHHHHH HHHHHHHHHHPPPPPPHHHH HHPHP
9	72	HHHPPPPPPPPPHHHHHHHHHP PPPHHHHHHHPPPPHHHHHH HPPPPHHHHHHHHHHHHHH HHHHHHPPP
10	79	PHHHHHHHHHHHHPPPPPPPP PPPHHHHHHHHHHHHHPPPPP PPPPPHHHHHHHHHHHHHHH HHHPPPPPPHHHH
11	83	PPHHHHPPPPPPPPHHHHPPPH HHHPPPPHHHHPPPPHHHHH HHHHHHHHHHHHHHHPPPP PPHHPPPPPPPPHHHHHHH
12	89	PHHHHHHHHPPPPPPPPHHHH HHHHHHHHHHHPPPPHHHHH HPPPPPHHHHHHHHHPPPPH HHHHPPPPPPHHHHHHHHH HHP
13	91	PHHHHHHHHHHHHHHHHPPPH HPPPPPPHHHHHHHHHPPPHH HHHHHHHHHHHHHHHPPHHH HHHHHHHHPPPPHHHHHHHH HHHHH
14	100	PPPHHHHHHHHPPPHHHHHH PPPHHHHHHHHHHHHHHPP PHHHHHHHPHHHHHHHPPHH HHHHPPPHHHHHHHHHPPPP PPHHHHHHHHPPPPP

8.1 Resultados

As Tabelas 1 e 2 apresentam os resultados dos experimentos com as proteínas mostradas na Tabela 3. Foram realizadas 20 execuções de cada algoritmo para cada proteína, onde 10 execuções foram para versão do ACO com energia de Lau e Dill, e 10, usando a energia simplificada. Para nossos experimentos, foi usado o sistema operacional Linux com processador i7 e 4G de memória, usando a linguagem de programação C.

Na Tabela 1, *E* é a melhor energia de Lau e Dill encontrada no artigo de referência em 10 execuções (Zhang

Tabela 1: Energias obtidas no trabalho de Zhang et al. (2013) e pelas duas versões do ACO

ID	E	E_{ACO}	Es	Es_{ACO}
1	-5	-6	100.97	100.97
2	-13	-13	457.23	455.99
3	-11	-11	305.16	300.96
4	-20	-17	716.77	716.77
5	-20	-21	1174.79	1150.56
6	-19	-19	885.69	878.12
7	-21	-21	1082.36	1074.93
8	-39	-39	4587.09	4514.47
9	-37	-37	4168.48	4165.35
10	-38	-38	4611.02	4569.59
11	-35	-35	4319.72	4512.71
12	-44	-44	6981.92	7080.43
13	-58	-58	11721.87	11491.49
14	-53	-55	11175.25	11553.40

et al., 2013) usando o Algoritmo de Vagalumes (a média e o desvio padrão não foram mostrados, pois no artigo de referência ficaram omitidos esses dados); e E_{ACO} é a melhor energia de Lau e Dill encontrada pelo ACO em 10 execuções para cada proteína. Embora o trabalho de Zhang enfatize a energia simplificada, a energia de Lau e Dill é facilmente obtida por meio das estruturas geradas por ele. Note que esta versão do ACO obteve, em 3 casos, valores com energia mais baixa que o trabalho de Zhang; enquanto que o artigo de Zhang et al. (2013) teve apenas 1 proteína com valor mais baixo que o ACO, sendo que para o restante das proteínas mostrou-se o mesmo resultado em ambos os algoritmos. Ainda na Tabela 1, Es é a energia simplificada obtida por Zhang (a média e o desvio padrão calculado das 10 execuções não foi mostrado no artigo), e Es_{ACO} é a energia simplificada resultante com o ACO em 10 execuções para cada proteína. Pode-se verificar que esta versão do ACO encontrou melhores valores (menores) que o algoritmo de Zhang em 9 das 14 proteínas, mostrando que a aplicação desta energia no ACO superou o resultado com Algoritmo de Vagalumes, em relação ao valor da energia.

A partir dos experimentos, pode-se perceber na Tabela 2 que a média do tempo de execução (**Ts**), em segundos, do ACO usando a energia simplificada é maior que aquela com a energia de Lau e Dill (**T**) em quase todos os casos testados. Isto pode ser explicado pela necessidade de gerar a matriz de distância para cada solução. O trabalho de Zhang não publicou o tempo de execução em seus experimentos, por isso não foi possível realizar uma comparação neste sentido.

8.2 Estruturas

A seguir, são apresentadas as melhores estruturas geradas pelas duas versões do ACO (com energia simplificada e energia de Lau e Dill), comparando-as com as melhores soluções obtidas por Zhang. Pelas estruturas, pode-se confirmar a influência da hidrofobicidade do dobramento com as três abordagens, explicitando que os aminoácidos hidrofóbicos (de cor **rosa**) se encontram

Tabela 2: Média do tempo de execução, em segundos, do ACO usando energia de Lau e Dill (T) e a energia simplificada (Ts)

ID	T	Ts
1	1.0458 s	1.1144 s
2	1.6757 s	2.2581 s
3	2.1002 s	2.5510 s
4	3.1908 s	3.8144 s
5	3.5965 s	4.3123 s
6	4.6765 s	4.5909 s
7	5.4108 s	5.2403 s
8	7.9610 s	10.3216 s
9	9.2852 s	10.6251 s
10	10.9371 s	12.0157 s
11	12.0600 s	12.2694 s
12	13.7183 s	15.5919 s
13	14.1207 s	19.3843 s
14	16.8713 s	20.5518 s

no interior da estrutura em todas as soluções, enquanto os hidrofílicos (de cor **verde**) estão na superfície, tendo um contato maior com o solvente.

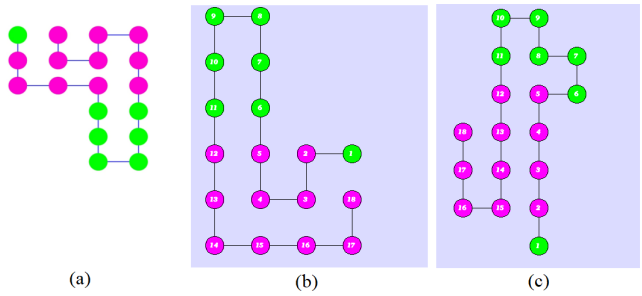

Figura 7: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 1.

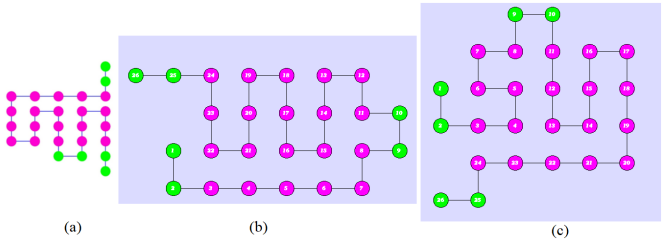

Figura 8: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 2.

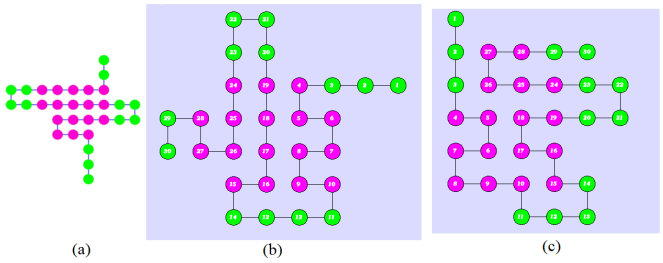

Figura 9: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 3.

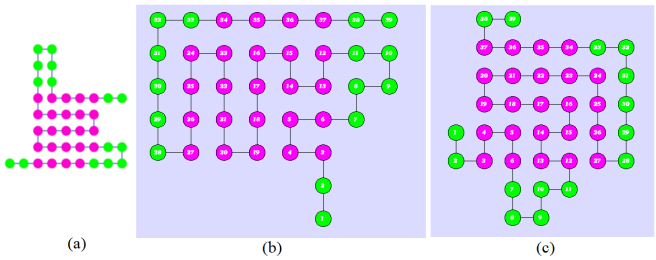

Figura 10: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 4.

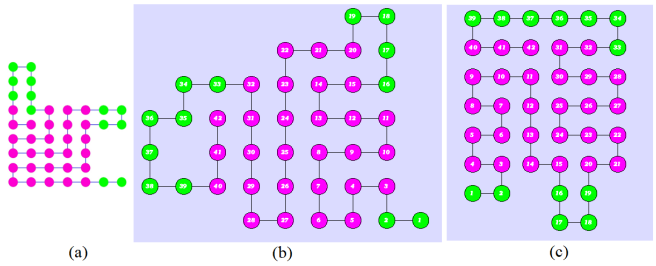

Figura 11: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 5.

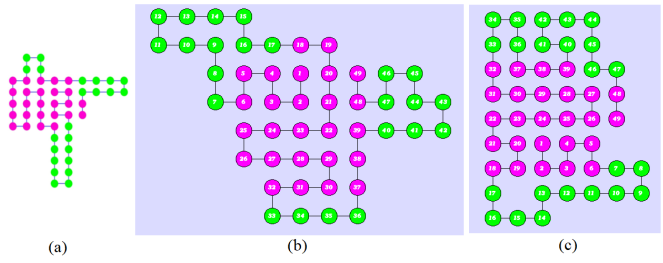

Figura 12: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 6.

9 Conclusões

Neste trabalho, o objetivo principal foi analisar o desempenho do algoritmo ACO com método de *back-tracking* para o problema de PSP em uma abordagem *ab initio*, usando duas funções de energia diferentes no modelo de representação HP-2D.

Para tal, foram aplicadas a energia de Lau e Dill e a energia simplificada, ambas no ACO, a fim de comparar o desempenho tanto do ponto de vista computacional quanto bioquímico. Para implementação e experimentos, foi usado um processador i7 com 4G de memória, utilizando a linguagem de programação C.

De acordo com os experimentos realizados e os resultados obtidos para o problema de PSP com o modelo HP-2D, pode-se concluir que, em relação aos valores de energia de Lau e Dill, a versão do ACO apresentou energia mínima compatível à literatura (Zhang et al., 2013) para o conjunto de proteínas utilizado, sendo que

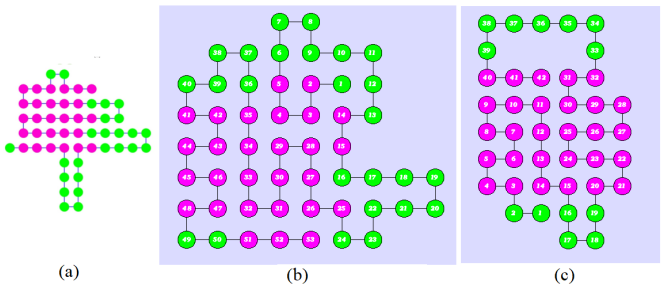

Figura 13: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 7.

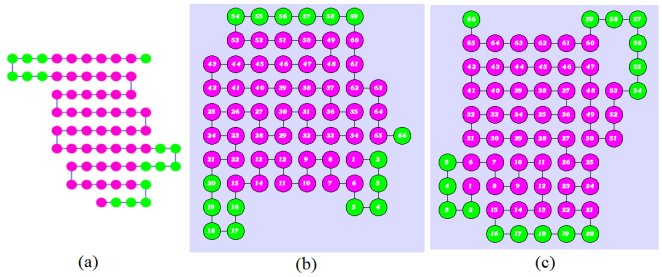

Figura 14: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 8.

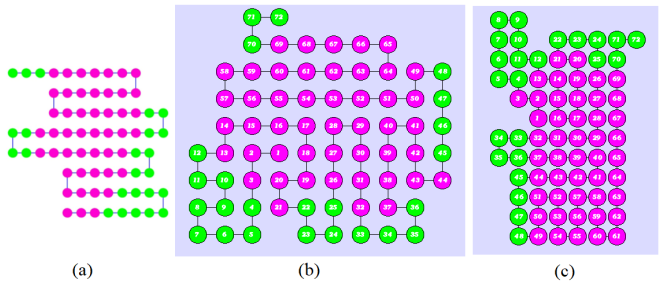

Figura 15: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 9.

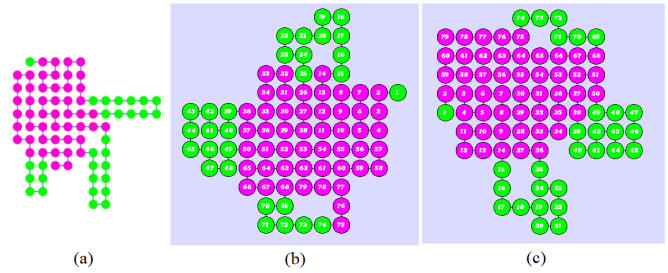

Figura 16: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 10.

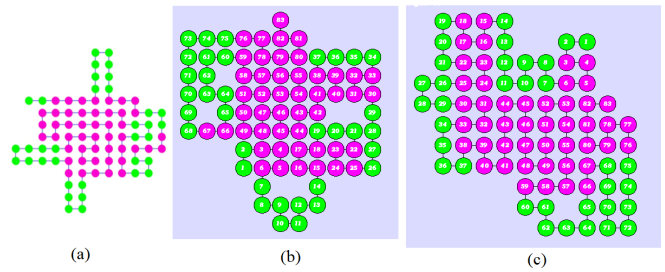

Figura 17: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 11.

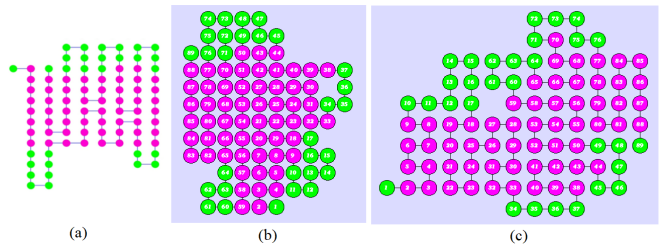

Figura 18: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 12.

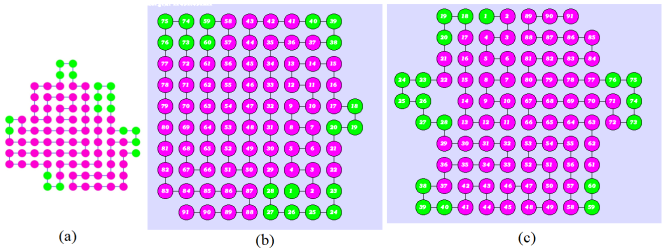

Figura 19: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 13.

em 3 casos o ACO deste trabalho superou o resultado dos valores de energia do trabalho de Zhang, o qual utilizou

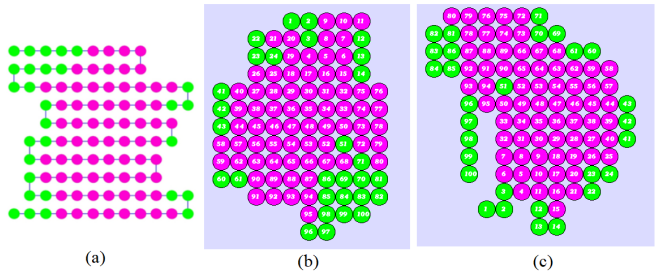

Figura 20: Melhores estruturas geradas por Zhang (a), pelo ACO com energia simplificada (b) e pelo ACO com energia de Lau e Dill (c) para Proteína 14.

outro algoritmo de otimização. Comparando a energia simplificada do ACO com o artigo de Zhang, também pode-se verificar resultados melhores na maioria dos casos (9 de 14 proteínas). Logo, o ACO apresentou melhores resultados em relação aos valores de energia com a versão da energia simplificada.

Comparando o tempo computacional das duas versões do ACO, pode-se verificar que com a energia de Lau e Dill o algoritmo teve melhor desempenho do que com a energia simplificada; mas com isto não se pode excluir o uso da mesma, uma vez que a diferença de tempo não é tão drástica, isto é, a aplicação da energia simplificada ainda apresenta resultados satisfatórios em relação ao custo computacional. Ainda discutindo sobre o tempo de execução, em ambas as versões do ACO foi utilizado o método de *backtracking* para corrigir possíveis colisões, inserindo um custo que é justificado pela garantia de se considerar apenas soluções factíveis.

Quanto às melhores estruturas obtidas, foi possível observar nas duas versões do ACO a forte influência da hidrofobicidade na formação das conformações, o que também ocorre no trabalho de Zhang, uma vez que os aminoácidos hidrofóbicos claramente ficaram concentrados no interior da molécula, enquanto que os polares se localizaram na parte externa da estrutura, permitindo um maior contato com o solvente, atingindo o efeito esperado.

Portanto, com os resultados mencionados deste trabalho confirma-se que o ACO com método de *backtracking* é uma abordagem adequada para o problema de PSP, inclusive melhorando em alguns casos com o uso da energia simplificada (que até o momento não tinha sido aplicada neste algoritmo), do ponto de vista bioquímico (valores de energia), principalmente. Considerando que para este problema tão desafiador todo refinamento de resultados é uma melhoria significativa e importante, sobretudo quando se trata de uma abordagem *ab initio*, pode-se concluir que o ACO com energia simplificada impulsiona, de fato, para resultados promissores. Em trabalhos futuros, almeja-se elaborar uma função de energia ponderada, que combine características de ambas as funções usando modelo lattice HP-2D.

10 Agradecimentos

Nossos agradecimentos à Universidade Federal de Uberlândia pelo constante estímulo à pesquisa de docentes e alunos.

Referências

Brasil, C. R. S. (2012). *Algoritmo evolutivo de muitos objetivos para predição ab initio de estrutura de proteínas*, Doutorado em ciência da computação, Instituto de Ciências Matemáticas e de Computação. Disponível em http://www.teses.usp.br/teses/disponiveis/55/55134/tde-20072012-163056/pt-br.php.

Cavalcanti, D. and Brasil, C. R. S. (2019). Algoritmos de inteligência de enxame com busca por pull move aplicados ao problema de predição de estrutura de proteínas, *XII Encontro Academico de Modelagem Computacional (EAMC2019)*, pp. 111–120. Disponível em http://www.eamc.lncc.br/PastEditions/EAMC2019/XIIEAMC_Soares2019.pdf.

Chu, D. and Zomaya, A. (2005). Parallel ant colony optimization for 3d protein structure prediction using the hp lattice model, *19th IEEE International Parallel and Distributed Processing Symposium*. Disponível em http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.125.8883&rep=rep1&type=pdf.

Citrolo, A. and Mauri, G. (2013). A hybrid monte carlo ant colony optimization approach for protein structure prediction in the hp model, *Italian Workshop on Artificial Life and Evolutionary Computation*. Disponível em https://www.researchgate.net/publication/257201307_A_Hybrid_Monte_Carlo_Ant_Colony_Optimization_Approach_for_Protein_Structure_Prediction_in_the_HP_Model.

Cox, M. and Doudna, J. (2012). *Biologia molecular: Princípios e técnicas*.

Crescenzi, P., Goldman, D., Papadimitriou, C. and Piccolboni, A. (1998). On the complexity of protein folding, *Journal of Computational Biology* **50**(3): 423–466. https://www.ncbi.nlm.nih.gov/pubmed/9773342.

Dias, J. M. and Brasil, C. R. S. (2017). Comparando algoritmos de otimização computacional aplicados ao problema de predição de estruturas proteicas com modelo hp-2d, *Revista Brasileira de Computação Aplicada* **9**(3): 87–99. "http://seer.upf.br/index.php/rbca/article/view/7005.

Dorigo, M. and Gambardella, L. M. (1997). Ant colony system: A cooperative learning approach to the traveling salesman problem, *IEEE Transactions on Evolutionary Computation*. Disponível em http://people.idsia.ch/~luca/acs-ec97.pdf.

Duc, D. D., Anh, V. T. N., Dinh, P. T. and Linh-Trung, N. (2018). An efficient ant colony optimization algorithm for protein structure prediction, *12th International Symposium on Medical Information and Communication Technology (ISMICT)*. Disponível em https://www.researchgate.net/publication/329652458_An_Efficient_Ant_Colony_Optimization_Algorithm_for_Protein_Structure_Prediction.

Gabriel, P. H. (2012). *Algoritmos evolutivos e modelos simplificados de proteínas para predição de estruturas terciárias*, Mestrado em ciência da computação, Instituto de Ciências Matemáticas e de Computação. Disponível em http://www.teses.usp.br/teses/disponiveis/55/55134/tde-14052010-143653/pt-br.php.

Hart, W. and Newman, A. (2001). *The Computational Complexity of Protein Structure Prediction in Simple Lattice Models*, CRC Press. Disponível em http://dimacs.rutgers.edu/~alantha/papers2/alantha-bill-bc.pdf.

Hu, X.-M., Zhang, J. and Li, Y. (2008). *Flexible protein folding by ant colony optimization*, Vol. 151, Springer-Verlag Berlin Heidelberg. https://link.springer.com/chapter/10.1007/978-3-540-70778-3_13.

Huang, C., Yang, X. and He, Z. (2010). Protein folding simulations of 2d hp model by the genetic algorithm based on optimal secondary structures, *Computational Biology and Chemistry* **34**(3): 137–142. https://www.ncbi.nlm.nih.gov/pubmed/20627698.

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. https://pubs.acs.org/doi/abs/10.1021/ma00200a030.

Liu, Z. and Yang, Z. (2014). Heuristic ant colony optimization algorithm for predicting the structures of 2d hp model proteins, *IEEE. Biomedical Engineering and Informatics (BMEI)*, pp. 719–723. Disponível em https://ieeexplore.ieee.org/document/7002867.

Llanes-Castro, A., Velez, C., Sánchez, A. M., Pérez-Sánchez, H. and Cecilia, J. (2016). Parallel ant colony optimization for the hp protein folding problem, *International Conference on Bioinformatics and Biomedical Engineering*, p. 615–626. Disponível em https://link.springer.com/chapter/10.1007/978-3-319-31744-1_54.

Maher, B., Albrecht, A., Loomes, M., Yang, X.-S. and Steinhöfel, K. (2014). A firefly-inspired method for protein structure prediction in lattice models, *Biomolecules* **1**(4): 56–57. https://www.ncbi.nlm.nih.gov/pubmed/24970205.

Shmygelska, A., Hernández, R. A. and Hoos, H. H. (2002). An ant colony optimization algorithm for the 2d hp protein folding problem, *Proceedings of the Third International Workshop on Ant Algorithms*, p. 40–53. Disponível em https://link.springer.com/chapter/10.1007%2F3-540-45724-0_4.

Shmygelska, A. and Hoos, H. H. (2003). An improved ant colony optimisation algorithm for the 2d hp protein folding problem, *Conference of the Canadian Society for Computational Studies of Intelligence Canadian AI 2003: Advances in Artificial Intelligence*,

pp. 400–417. Disponível em https://link.springer.com/chapter/10.1007/3-540-44886-1_30.

Shmygelska, A. and Hoos, H. H. (2005). An ant colony optimisation algorithm for the 2d and 3d hydrophobic polar protein folding problem, *BMC Bioinformatics*, Vol. 6. Disponível em https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-30.

Song, J., Cheng, J. and Zheng, T. (2006). Protein 3d hp model folding simulation based on aco, *Sixth International Conference on Intelligent Systems Design and Applications*, Vol. 6. Disponível em https://ieeexplore.ieee.org/document/4021474.

Stützle, T. and Hoos, H. H. (1998). Improvements on the ant-system: Introducing the max-min ant system, *International Conference in Norwich*, p. 245–249. Disponível em https://link.springer.com/chapter/10.1007/978-3-7091-6492-1_54.

Thalheim, T., Merkle, D. and Middendorf, M. (2008). Protein folding in the hp-model solved with a hybrid population based aco algorithm, *IAENG International Journal of Computer Science*, Vol. 3. Disponível em http://www.iaeng.org/IJCS/issues_v35/issue_3/IJCS_35_3_06.pdf.

Thilagavathi, N. and Amudha, T. (2015). *Rank based ant algorithm for 2D-HP protein folding*, Vol. 3, Springer. Disponível em https://link.springer.com/chapter/10.1007/978-81-322-2202-6_40.

Wang, S. (2018). *Improved swarm intelligence algorithm for protein folding prediction*, Springer. Disponível em https://link.springer.com/article/10.1007%2Fs10586-018-2257-1.

Yang, X.-S. (2008). Luniver Press. Disponível em https://dl.acm.org/citation.cfm?id=1628847.

Zhang, Y., Wu, L. and Wang, S. (2013). Solving two-dimensional hp model by firefly algorithm and simplified energy function, *Mathematical Problems in Engineering* **2013**. https://doi.org/10.1155/2013/398141.