Comportamento da energia no Algoritmo de Colônia de Formiga para predição de estruturas de proteínas

Christiane Regina S. Brasil[1], Júlia M. Dias[1]

[1]Faculdade de Computação – Universidade Federal de Uberlândia (UFU)
Uberlândia – MG - Brasil

Resumo: *Os algoritmos de otimização computacional são frequentemente usados na resolução de problemas complexos. Neste trabalho o algoritmo de otimização estudado foi o Algoritmo de Colônia de Formiga (ACO) com backtracking. Este algoritmo foi aplicado ao problema de predição de estruturas de proteínas, considerado um problema de alta complexidade. Deste modo, o objetivo deste trabalho foi utilizar o ACO para encontrar uma solução para o problema e analisar o comportamento da energia durante o processo de predição, comparando estes resultados com a literatura.*

Abstract: *Computational optimization algorithms are often used to solve complex problems. In this work the optimization algorithm studied was the Ant Colony Optimization (ACO) with backtracking, which is inspired by the natural process that the ants do in the search for their food. This algorithm was applied to the problem of Protein Struct Problem, considered a problem of high complexity. Therefore, the objective of this work was to use the ACO to find a solution for the problem and analyze the energy behavior during the prediction process, comparing these results with the literature.*

1. Introdução

Há, no mundo real, diversos problemas que gastam um tempo inviável para encontrar sua melhor resposta, cujo espaço de busca por soluções é imenso e podem levar anos, ou até mesmo séculos, para retornar a solução correta. Estes são problemas combinatórios, e são classificados computacionalmente como problemas NP, Crescenzi (1998). Nesse contexto, NP significa "tempo polinomial não determinístico", isto é, o tempo necessário para resolver tal problema usando qualquer algoritmo conhecido cresce rapidamente à medida que o tamanho do problema aumenta. Um destes problemas complexos é o Problema de Predição de Estruturas de Proteínas (PSP), que representa um grande desafio na atualidade e visa gerar estruturas de proteínas ainda não conhecidas na natureza. A estrutura de uma proteína está diretamente relacionada à função que ela exerce, onde tais funções são fundamentais para manter a vida de qualquer ser vivo na Terra, Cox (2012). Além disso, a partir do conhecimento das funções proteicas pode-se investir no avanço da medicina.

Deste modo, o PSP é um problema de extrema importância biológica, especialmente no âmbito da saúde mundial. No entanto, os métodos convencionais utilizados (cristalografia e ressonância nuclear magnética) para descobrir a estrutura das proteínas ainda são poucos eficientes em termos de tempo e custo. Neste sentido, aplicar algoritmos de otimização a este problema representa uma alternativa que vem

apresentando bons resultados, tais como Shmygelska (2002), Hu (2008), Gabriel (2010), Brasil (2012), Dias (2017).

O objetivo deste trabalho foi realizar uma análise sobre o comportamento da energia encontrada durante o processo de predição com o modelo HP-2D usando o Algoritmo de Colônia de Formiga, um importante método de otimização computacional, usando *backtracking* para correção de soluções infactíveis. A maioria dos trabalhos na literatura não utiliza correção de soluções, uma vez que torna o processo mais lento. Neste trabalho foi adotado este procedimento, pois embora aumente, de fato, o tempo computacional, tem-se a garantia de se estar considerando apenas soluções factíveis, colaborando para uma convergência mais eficiente para a melhor solução. O método foi aplicado sobre algumas proteínas extraídas do trabalho de Zhang (2013), comparando os resultados com esse artigo de referência.

2. Algoritmo de Colônia de Formiga (ACO)

O Algoritmo de Colônia de Formiga (do inglês, *Ant Colony Optimization*), foi desenvolvido por Dorigo e Gambardella (1997), baseado no comportamento das formigas. Na natureza, as formigas fazem, inicialmente, caminhos aleatórios para encontrar o alimento. Este caminho perde a aleatoriedade após um tempo. Isto ocorre devido a comunicação entre as formigas por meio de uma substância química chamada feromônio.

Com base nesta análise, surgiu o ACO, que mimetiza o feromônio e as formigas e os utilizam para computar suas soluções probabilísticas na busca por uma solução viável. Do algoritmo inicial criado por Dorigo, surgiram algumas variações. As mais comuns são o ACS (*Ant Colony System*) e o MMAS (*Min-Max Ant System*). Na maioria das vezes a diferença entre as variações consiste na fórmula de atualização do feromônio. O MMAS foi desenvolvido por Stützle e Hoos (1998) e consiste em manter valor mínimos e máximos para o feromônio. Para este trabalho foi utilizada a abordagem MMAS.

O cálculo de probabilidade executado para a formiga decidir qual caminho seguir é dado pela fórmula a seguir:

$$P_{ij} = (\tau_{ij})^{\alpha} (\eta_{ij})^{\beta} / \sum_{e \in N} (\tau_{ie})^{\alpha} (\eta_{ie})^{\beta}$$

onde i e j significam, respectivamente o ponto em que a formiga está 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 o quão importante será o feromônio e a informação heurística, respectivamente, para o cálculo; e N é o conjunto de prováveis caminhos que a formiga tem para escolher.

A atualização do feromônio é dada pela fórmula a seguir:

$$\tau_{ij} = \rho (\tau_{ij}) + \rho \Delta^{melhor}$$

onde τ_{ij} é o valor do feromônio no local pretendido; o símbolo ρ é o parâmetro que representa a taxa de evaporação do feromônio naquele local e seu valor varia no intervalo de 0 a 1; Δ_{melhor} = E sol / E opt . E sol é a energia obtida a partir do caminho que a formiga percorreu. E opt é a melhor energia conhecida da proteína. Se esta

informação for desconhecida, usa-se a melhor energia encontrada até o momento da execução. O delta (Δ) pode sofrer algumas modificações dependendo da abordagem do ACO.

Neste trabalho foi desenvolvido o método ACO por ser de fácil entendimento e apresentar bons resultados para o PSP, Shmygelska (2002), Hu (2008), Dias (2017), com *backtracking*, que pode fazer a formiga voltar quantas vezes forem necessárias, até encontrar uma posição em que não dê possibilidades de ocorrer colisões.

3. Problema de Predição de Estruturas de Proteínas

O objetivo do PSP (do inglês, *Protein Structure Problem*) é encontrar uma estrutura tridimensional de uma proteína, uma vez que esta conformação está relacionada com a funcionalidade da proteína, Brasil (2012). Portanto, encontrar a estrutura de uma dada proteína pode representar um avanço para a saúde mundial, pois descobrindo suas funções pode-se investir em medicamentos ou vacinas, por exemplo, para doenças graves ou até incuráveis. No entanto, o PSP é classificado como problema NP, no qual seu espaço de busca cresce exponencialmente em relação ao tamanho da proteína, Dorigo and Gambardella (1997).

Neste contexto, a estrutura tridimensional apresenta o formato em que a proteína está quando encontrada na natureza. Essa conformação 3D da proteína é definida pelo enovelamento dos aminoácidos que a compõem, que pode ocorrer devido a vários critérios intra e itermoleculares. Dentre tantos, pode-se destacar a hidrofobicidade, descrito a seguir.

3.1 Modelo de representação HP

O modelo Hidrofóbico-Polar (HP), criado por Lau e Dill (1989), é um modelo de representação da estrutura de uma proteína, e se caracteriza por usar malhas quadráticas reticuladas em duas dimensões (2D) ou em três dimensões (3D). Neste trabalho foi usado o modelo HP-2D.

Este modelo considera a hidrofobicidade e a polaridade dos aminoácidos para a representação computacional. A hidrofobicidade, no contexto biológico, define as moléculas que não se dissolvem nem interagem com a água. Aminoácidos hidrofóbicos, que não interagem com a água, são classificados como (H); enquanto que os aminoácidos hidrofílicos (P) possuem interação com a água. As interações entre aminoácidos hidrofóbicos são importantes para a formação da estrutura proteica, pois induzem com que as cadeias de aminoácidos se dobrem, uma vez que aqueles hidrofóbicos tendem a “fugir” da água.

Deste modo, a energia tradicional de uma conformação obtida por este modelo provém das interações H-H de aminoácidos não conectados. Essa energia é inversamente proporcional à quantidade de interações H-H, uma vez que se busca a menor energia possível. Neste trabalho utilizou-se o modelo HP para o PSP, por ser um modelo simples e efetivo, Shmygelska (2002), Hu (2008), Gabriel (2010), Dias (2017).

4. Resultados

Neste trabalho, o ACO utiliza os seguintes parâmetros: α para expressar a importância do feromônio, β para indicar a importância da informação heurística, ρ para a atualização do feromônio. Os resultados foram obtidos com α = 1, β = 1 e ρ = 0.5. Foram analisadas cinco proteínas advindas do artigo Zhang (2013), que aplica o método bioinspirado *Firefly Algorithm* para o PSP com modelo HP-2D.

A Tabela 1 apresenta os resultados das execuções destas proteínas e os parâmetros utilizados. Foram feitas 10 execuções do algoritmo para cada proteína. A energia calculada pelo ACO é a função tradicional de energia.

Tabela 1: Resultados das execuções do ACO com as cinco proteínas.

Proteína	Tam	NumF	NumI	E_{ACO}	E_{ref}	$Freq_{ACO}$	Tempo
1	18	100	500	-6	-5	10/10	1045,8 ms
2	26	100	500	-13	-13	10/10	1675,7 ms
3	30	100	500	-11	-11	10/10	2100,2 ms
4	42	100	500	-21	-20	01/10	3596,5 ms
5	53	100	500	-21	-20	02/10	5410,8 ms

Onde *Tam* é a quantidade de aminoácidos de uma proteína; *NumF* é a quantidade de formigas usadas em cada execução do ACO; *NumI* é o número de iterações; E_{ACO} é a melhor energia tradicional encontrada entre as dez execuções do ACO; E_{ref} é a melhor energia tradicional encontrada no artigo de referência (Zhang, 2013); $Freq_{ACO}$ é a frequência encontrada da melhor energia em dez execuções do ACO, por exemplo, 2/10 significa que a melhor energia foi encontrada em 2 execuções das 10, *Tempo* é o tempo gasto, em média, nas dez execuções. Note que se manteve o número de formigas e iterações nas diferentes proteínas para fins de comparação com o trabalho de Zhang (2013).

Embora o trabalho de Zhang não tenha focado na energia tradicional como o ACO, é possível calcular a mesma por meio das estruturas obtidas. Na Tabela 6 pode-se verificar que o ACO encontrou a mesmas energias de Zhang para a maioria dos casos, inclusive, encontrou até melhores para alguns deles.

4.1. Resultados da Proteína 1

A Figura 1 representa as estruturas da proteína com 18 aminoácidos. A estrutura da Figura 1(a) foi gerada pelo ACO e a estrutura (b) foi encontrada pelo algoritmo *Firefly* implementado por Zhang (2013). Nas figuras, os aminoácidos H são representados pelos círculos mais escuros. Pode-se perceber certa semelhança no dobramento das mesmas. Observa-se na Figura 2 que para a Proteína 1 a energia tradicional é rapidamente encontrada pelo ACO. Na maioria das execuções (em 8 das 10 execuções) do ACO para esta proteína, o gráfico apresentou este comportamento da energia.

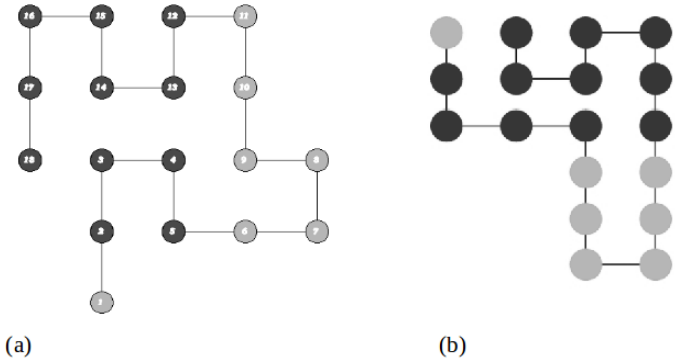


Figura 1. Estruturas encontradas para a Proteína 1.

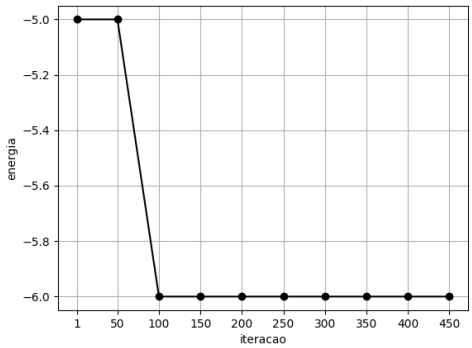


Figura 2. Comportamento da energia durante as iterações do ACO para Proteína 1.

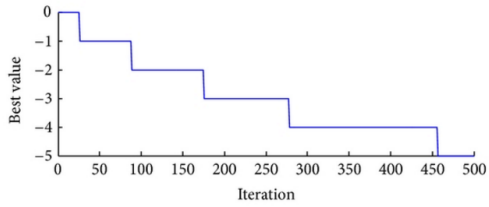


Figura 3. Gráfico de execução gerado por Zhang (2013).

A Figura 3 mostra o resultado de uma execução do algoritmo *Firefly* implementado por Zhang (2013) com a energia tradicional. Nota-se na Figura 2 que o ACO encontrou a energia ótima facilmente nas iterações iniciais, enquanto a energia do algoritmo de Zhang (2013) foi sendo minimizada gradativa e mais lentamente. Vale ressaltar que o ACO, inclusive, consegue obter uma energia mínima melhor que a gerada em Zhang (2013) neste caso.

4.2. Resultados da Proteína 2

A Figura 4 mostra uma comparação entre as estruturas encontradas pelo ACO (Figura 4(a)) e por Zhang (2013) (Figura 4(b)) com a proteína de 26 aminoácidos.

As estruturas apresentadas na Figura 4 mostram-se bem semelhantes entre si. Outro detalhe é que ambas apresentam o mesmo número de ligações entre os aminoácidos H não consecutivos, ou seja, possuem a mesma energia tradicional, de -13.

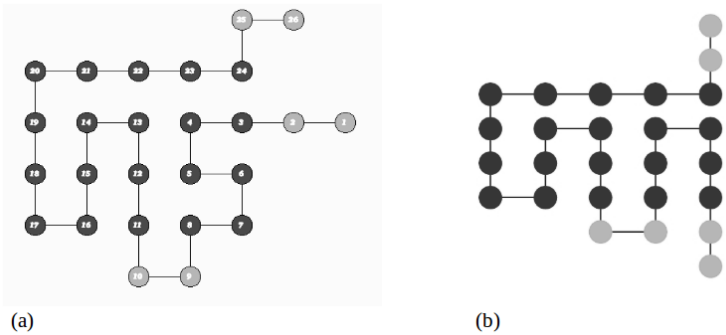


Figura 4. Estruturas encontradas para a Proteína 2.

A Figura 5 mostra o comportamento da energia no ACO em algumas das execuções realizadas. Note que para esta proteína o ACO não precisou de muitas iterações para encontrar a menor energia de conformação. O autor do artigo, Zhang (2013), não disponibilizou o gráfico de execução desta proteína, nem das restantes.

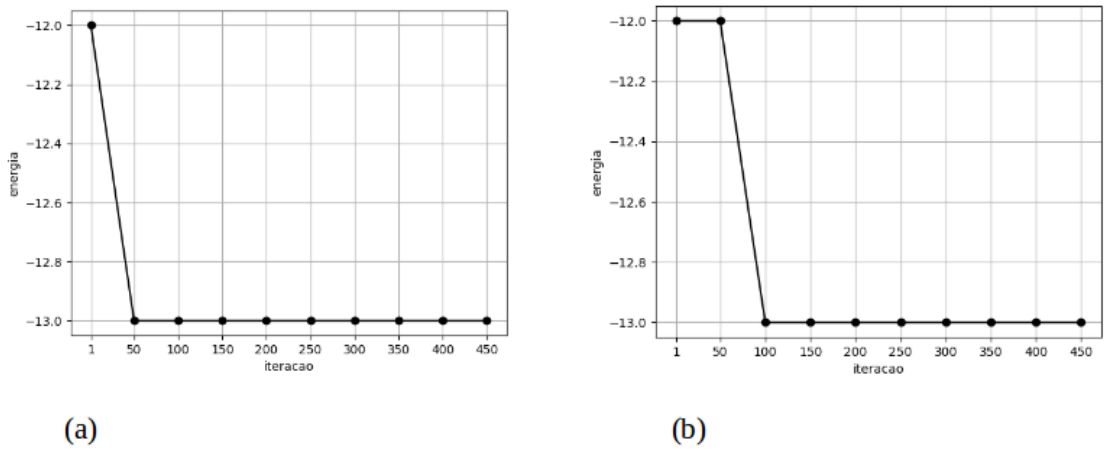


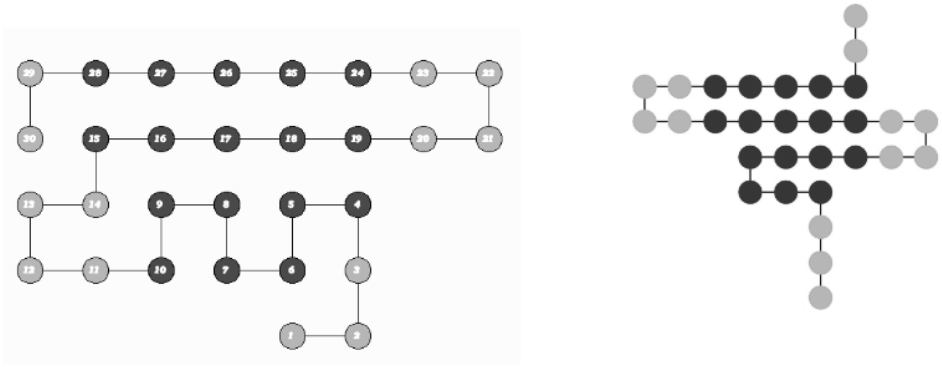
Figura 5. Comportamento da energia durante as iterações do ACO para Proteína 2.

4.3. Resultados da Proteína 3

Na Figura 6, as estruturas da proteína de 30 aminoácidos encontradas pelo ACO (Figura 6(a)) e por Zhang (Figura 6(b)), são mostradas.

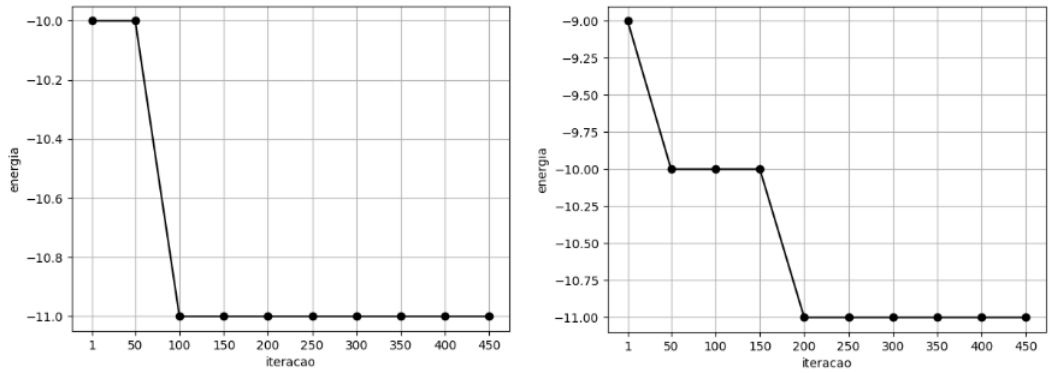
Embora as melhores estruturas obtidas pelos dois métodos não sejam as mesmas, nota-se que ambas apresentam uma folha-beta, descrita em Huang (2010), caracterizada por duas sequências paralelas de aminoácidos dispostos em linha reta. Ambas também têm a mesma energia mínima de -11.

A Figura 7 apresenta o gráfico de algumas execuções do ACO. Os dois gráficos são os resultados mais comuns da execução do ACO para a proteína de 30 aminoácidos. Para esta proteína o artigo de Zhang (2013) não apresenta o gráfico de execução do algoritmo.



(a) (b)

Figura 6. Estruturas encontradas para a Proteína 3.



(a) (b)

Figura 7. Comportamento da energia durante as iterações do ACO para Proteína 3.

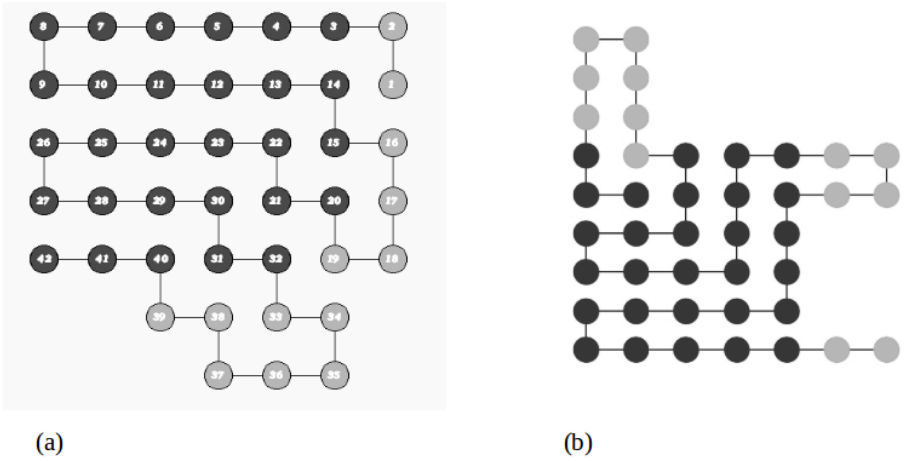
4.4. Resultados da Proteína 4

A Figura 8 apresenta as estruturas de conformação para a proteína de 42 aminoácidos.

Há algumas semelhanças entre as estruturas apresentadas na Figura 9 no que se refere a formação de folhas-beta. No entanto, o número de ligações entre os aminoácidos H não consecutivos, e consequentemente a energia tradicional, na Figura 8(a) obtida pelo ACO é -21, o que difere da Figura 8(b) gerada por Zhang, onde a energia é -20.

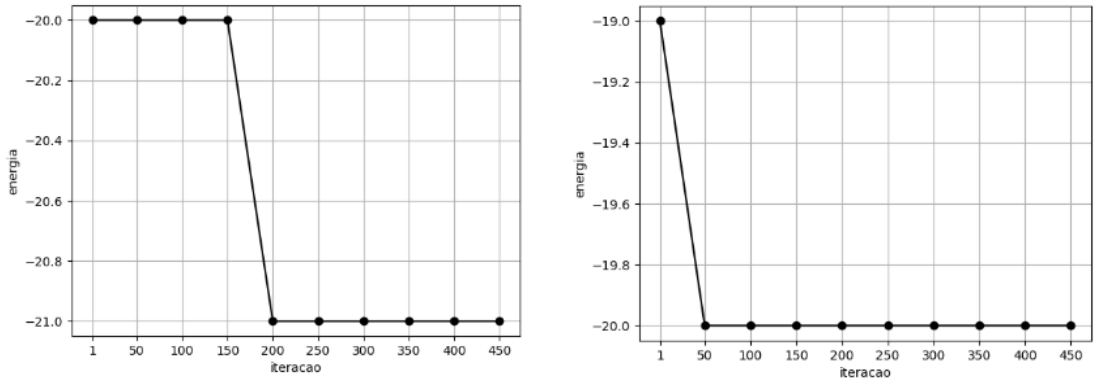
A Figura 9 ilustra dois exemplos de execução: uma onde a energia -21 foi encontrada (Figura 9(a)) e outra onde não foi (Figura 9(b)). Na Figura 9(a) a energia -21 demora um pouco a ser encontrada, mas a partir da iteração 200 foi alcançada, enquanto

na Figura 9(b) aparece a energia -20 sendo encontrada rapidamente, porém permanece com este valor até o fim da execução.



(a) (b)

Figura 8. Estruturas encontradas para a Proteína 4.



(a) (b)

Figura 9. Comportamento da energia durante as iterações do ACO para Proteína 4.

4.5. Resultados da Proteína 5

A Figura 10 se refere às estruturas encontradas pelo ACO (Figura 10(a)) e por Zhang (Figura 10(b)) para a proteína com 53 aminoácidos.

Na Figura 10 pode-se verificar que as estruturas têm algumas semelhanças entre si, mas se diferem na energia, onde a energia tradicional resultante de Zhang nesta estrutura é – 20, enquanto o ACO encontrou a energia de – 21, portanto, uma energia menor. Ainda que tenha sido em 2 execuções de 10, ainda assim, foi melhor que Zhang que não encontrou em nenhuma vez esse valor, do modo que aconteceu na proteína de 42 aminoácidos.

A Figura 11 mostra duas execuções: uma onde a energia -21 foi encontrada (Figura 11(a)) e outra onde não foi (Figura 11(b)). Pode-se perceber, neste caso, que a

menor energia não foi encontrada nas iterações iniciais em nenhum dos dois exemplos, indicando que aumentar o número de iterações para proteínas maiores poderia ser mais viável para se encontrar energias melhores.

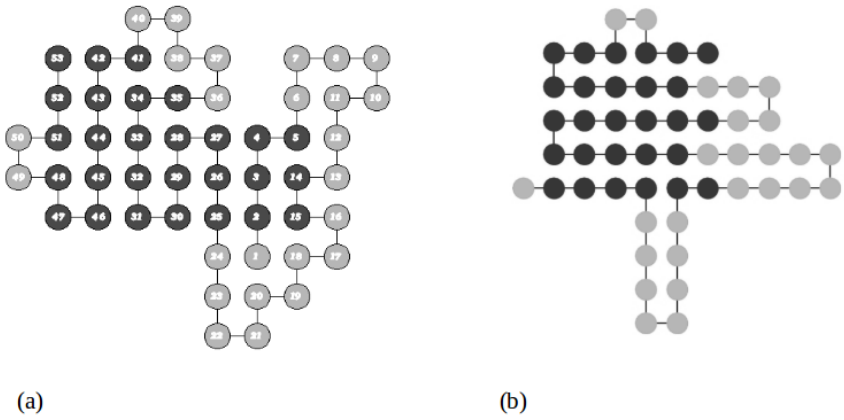


Figura 10. Estruturas encontradas para a Proteína 5.

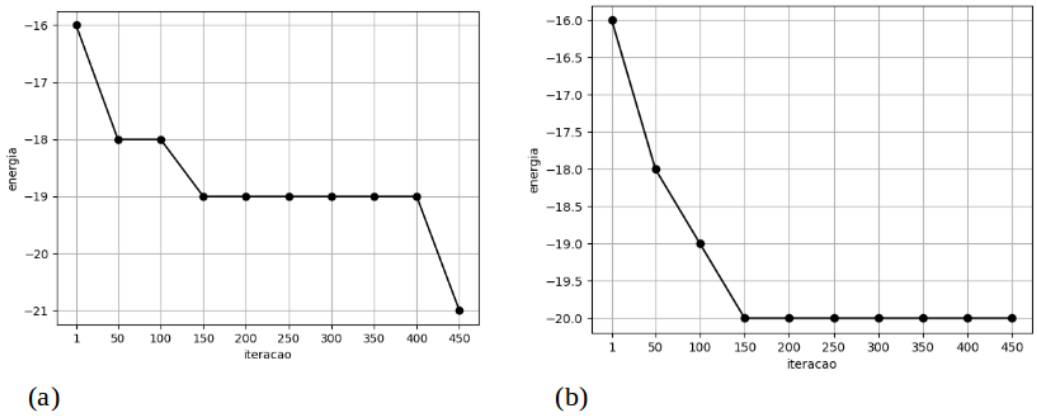


Figura 11. Comportamento da energia durante as iterações do ACO para Proteína 5.

5. Considerações Finais

De acordo com os resultados obtidos, pode-se concluir que o ACO implementado neste trabalho com *backtracking* mostrou-se bastante eficiente para o problema em questão, tanto para o valor da energia quanto para a estrutura gerada, mas principalmente em relação à energia, confirmando conclusões obtidas em Dias (2017) para outro conjunto de testes. Como observado, a energia converge rapidamente na maioria dos casos testados, indicando que seria necessária uma quantidade menor de iterações no ACO que a quantidade proposta por Zhang (2013) no algoritmo *Firefly*, e que somente para proteínas maiores seria importante um maior número de iterações, uma vez que nestes casos houve uma minimização mais gradual da energia. Vale destacar que o ACO, inclusive, consegue obter uma energia mínima melhor que a gerada em Zhang (2013) para as proteínas 1, 4 e 5.

Referências

Brasil, C. R. S. (2012), *Algoritmo evolutivo de muitos objetivos para predição ab initio de estrutura de proteínas.* Tese (Doutorado) — USP, São Carlos.

Crescenzi, P., Goldman, D., Papadimitriou, C.H., Piccolboni, A., Yannakakis, M. (1998) On the Complexity of Protein Folding. *Journal of Computational Biology. v.* 50, p. 423–466.

Cox, M. and Doudna, J. A. (2012), Biologia Molecular: Princípios e Técnicas, Artmed Editora.

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,* v. 9, n. 3, ISSN 2176-6649.

Dorigo, M. and Gambardella, L. M. (1997), Ant colony system: A cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, Press, Piscataway, NJ, USA, v. 1, n. 1, p. 53–66, ISSN 1089-778X.

Gabriel, P. H. R.(2010), *Algoritmos evolutivos e modelos simplificados de proteínas para predição de estruturas terciárias.* Dissertação (Mestrado) — USP, São Carlos.

Hu, X.; Zhang, J.; Xiao, J. and Li, Y. (2008) Protein Folding in Hydrophobic-Polar Lattice Model: A Flexible Ant-Colony Optimization Approach. *Protein and Peptide Letters*, v. 15, n. 5, p. 469 - 477.

Huang, Y., 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*, v. 34, n. 3, p. 137–142, jun. 2010. ISSN 1476- 9271.

Lau, K. F. and Dill, K. A. (1989), A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, v. 22, n. 10, p. 3986–3997.

Shmygelska, A.; Hernández, R. A. and Hoss, H. (2002) An ant colony optimization algorithm for the 2D HP protein folding problem. In: *Proceedings of the Third International Workshop on Ant Algorithms.* [S.l.]: Springer Verlag, p. 40–53.

Stützle, T. and Hoos, H. (1998) Improvements on the ant-system: Introducing the max-min. Artificial Neural Nets and Genetic Algorithms: Proceedings of the ant system. *In: International Conference in Norwich*. p. 245–249. ISBN 978-3-7091-6492-1.

Zhang, Y.; Wu, L.; and Wang, S. (2013) Solving Two-Dimensional HP Model by Firefly Algorithm and Simplified Energy Function, *Mathematical Problems in Engineering*, vol. 2013, Article ID 398141, 9 pages, doi:10.1155/2013/398141