Algoritmo genético aplicado à predição da estrutura de proteínas utilizando o modelo 3D-HP *Side Chain*

César Manuel Vargas Benítez, Heitor S. Lopes*

[1]Laboratório de Bioinformática
Programa de Pós-Graduação em Engenharia Elétrica
e Informática Industrial – CPGEI
Universidade Tecnológica Federal do Paraná
Av. 7 de setembro, 3165 – 80230-901 Curitiba (PR), Brasil

cbenitez@cpgei.ct.utfpr.edu.br, hslopes@pesquisador.cnpq.br

Abstract. *This work presents a parallel genetic algorithm (PGA) for the protein folding problem, using the 3DHP-SC model. This model has been sparsely studied in the literature due to its complexity. A new fitness function was proposed, based on the free-energy and compacity of the folding. Since there is no benchmark available to date, a set of 5 sequences was used, based on a simpler model. The PGA obtained biologically coherent results, suggesting its adequacy for the problem. Future work will include new knowledge-based genetic operators and the expansion of the benchmark.*

Keywords: Genetic Algorithm, Bioinformatics, Protein Folding, 3DHP-SC.

Resumo. *Este trabalho apresenta um algoritmo genético paralelo (AGP) para o problema de dobramento de proteínas, utilizando o modelo 3DHP-SC. Este modelo tem sido pouco abordado devido ao elevado grau de complexidade envolvido. Foi proposta uma função de fitness baseada na energia livre e na compacidade do dobramento. Devido a não existir, até então, benchmarks para teste deste modelo, foi proposto um conjunto de 5 sequências baseado em outro modelo mais simplificado. O AGP obteve dobramentos biologicamente coerentes, sugerindo a adequabilidade da metodologia proposta. Trabalhos futuros incluirão a proposição de operadores genéticos baseados em conhecimento, bem como a expansão do benchmark.*

Palavras-chave: Algoritmos genéticos, Bioinformática, Dobramento de proteínas, 3DHP-SC.

1. Introdução

Proteínas são polímeros compostos por uma cadeia de aminoácidos (também chamados de resíduos)que são ligados linearmente através de ligações peptídicas. Os aminoácidos são caracterizados pela existência de um átomo de carbono central (Cα) ao qual estão ligados um átomo de hidrogênio, um grupo amina (NH_2), um grupo carboxílico (COOH) e uma cadeia lateral (também chamada de radical R) que define

*Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, protocolos 550977/2007-4, 481207/2007-0 e 309262/2007-0.

a função do aminoácido. Dois aminoácidos formam uma ligação peptídica quando o grupo carboxílico de um deles reage com o grupo amina do outro. As proteínas, à medida que formadas no ribossomo vão se dobrando sobre si, originando uma conformação tridimensional única, também conhecida como conformação nativa. Este processo é conhecido como dobramento de proteínas. A função biológica de uma proteína depende da sua estrutura tridimensional, que, por sua vez, depende da estrutura primária, isto é, da sequência de aminoácidos que a compõe. Sabe-se que proteínas mal-formadas (devido a um dobramento errôneo) podem originar diversas enfermidades, como por exemplo, mal de Alzheimer, alguns tipos de câncer, fibrose Cística, etc. Devido à importância do dobramento de proteínas para a medicina e a bioquimica, pesquisadores têm se concentrado no estudo deste processo e, consequentemente, gerado uma quantidade considerável de informações disponíveis para a comunidade científica. Portanto, adquirir conhecimento sobre a estrutura tridimensional de proteínas e, consequentemente, sobre a sua função é muito importante para o desenvolvimento de novas drogas com funcionalidade específica.

Graças aos inúmeros projetos de sequenciamento genômico no mundo, uma grande quantidade de proteínas tem sido descoberta. No entanto, apenas uma pequena porção delas possui estrutura tridimensional conhecida. Isto se deve à dificuldade envolvida no dobramento de proteínas tanto do ponto de vista bioquímico quanto computacional. A ciência da computação desempenha um papel importante nisto, desenvolvendo modelos computacionais e soluções para o problema de dobramento de proteínas (PDP). A simulação de modelos computacionais que levam em consideração todos os átomos de uma proteína são inviáveis computacionalmente. Consequentemente, diversos modelos que abstraem a estrutura real da proteína foram propostos. Tais modelos, embora irreais, utilizam algumas propriedades bioquímicas dos aminoácidos, e podem apresentar algumas características interessantes e úteis para se observar o comportamento de proteínas sintéticas.

O modelo mais simplificado para o estudo do dobramento de proteínas é conhecido como modelo Hidrofóbico-Polar (*Hydrophobic-Polar*–HP), nas versões bi (2D-HP) e tridimensional (3D-HP) [Dill 1985]. Contudo, a abordagem computacional para este modelo leva a um problema NP-difícil [Berger and Leighton 1998]. Este fato enfatiza a necessidade de se utilizar métodos heurísticos para lidar com o problema. Neste cenário, métodos de computação evolucionária tem se mostrados muito eficientes e, dentre estes, os algoritmos genéticos (AGs) têm se destacado [Lopes 2008].

O objetivo deste trabalho é aplicar um AG paralelo para o PDP utilizando o modelo 3D-HP *Side Chain*, um modelo que aumenta o realismo da simulação, mas torna mais difícil o problema.

2. O modelo 3D-HP *Side Chain*

O modelo HP, já citado, representa a abstração mais simples para o PDP. Este modelo divide os 20 aminoácidos proteinogênicos em duas classes, de acordo com a sua solubilidade em meio aquoso: hidrofílicos (ou polares – P) e hidrofóbicos – H. Quando uma proteína é dobrada em sua conformação nativa, a maioria dos aminoácidos hidrofóbicos tendem a ficar na região interna da proteína (dentro do novelo

formado), interagindo entre si, protegidos do contato com a água pelos aminoácidos polares. Acredita-se que a conformação nativa, por ser a estrutura mais estável da proteína, esteja no estado de energia livre mínima. O modelo HP considera que a interação entre os aminoácidos hidrofóbicos representa a contribuição mais significativa para a energia livre da proteína. Nos modelos HP, o dobramento de uma proteína é representado em uma treliça, usualmente quadrada (para o 2DHP) ou cúbica (para o 3DHP). Ambos os modelos (2DHP e 3DHP) tem sido frequentemente explorados na literatura [Lopes 2008].

O próximo passo para simular características mais realistas das proteínas é incluir uma elemento representando a cadeia lateral (*Side Chain* – SC) dos aminoácidos [Li et al. 2002]. Para isto, a proteína é representada por um *backbone* (comum para todos os aminoácidos) e uma cadeia lateral, hidrofóbica (H) ou polar (P). Esta representação define o modelo 3DHP-SC. Ao contrário dos modelos 2DHP e 3DHP, o 3DHP-SC tem sido muito pouco explorado na literatura recente. Para este modelo, a energia livre de uma conformação leva em consideração a posição espacial da cadeia lateral, e pode ser descrita pela equação 1 [Li et al. 2002]:

$$H = \epsilon_{bb} \sum_{i=1,j>i+1}^{N} \delta_{r_{ij}^{bb}} + \epsilon_{bs} \sum_{i=1,j\neq i}^{N} \delta_{r_{ij}^{bs}} + \epsilon_{ss} \sum_{i=1,j>i}^{N} \delta_{r_{ij}^{ss}} \quad (1)$$

Onde ϵ_{bb}, ϵ_{bs} e ϵ_{ss} representam a ponderação na energia de cada tipo de interação possível: *backbone*/*backbone* (BB-BB), *backbone*/*side-chain* (BB-SC) e *side-chain*/*side-chain* (SC-SC); e r_{ij}^{bb}, r_{ij}^{bs} e r_{ij}^{ss} são as distâncias (no espaço tridimensional) entre o i-ésimo e o j-ésimo resíduos das interações BB-BB, BB-SC e SC-SC, respectivamente (para efeitos de simplificação, neste trabalho foi utilizada distância unitária entre os resíduos).

Acredita-se que as interações hidrofóbicas correspondem à força principal que dirige o processo de dobramento das proteínas. Durante este processo, a energia livre da proteína tende a diminuir. Sabe-se que a energia livre de uma conformação tridimensional é inversamente proporcional ao número de contatos hidrofóbicos (HnC). Portanto, o procedimento algorítmico para o dobramento que maximiza o HnC encontrará, reciprocamente, a conformação com o estado de menor energia livre possível.

3. Metodologia

Os AGs foram criados na década de 60 e, desde então, têm sido utilizados para a solução de muitos problemas de otimização e busca em engenharia e na computação [Michalewicz 1996]

Dentre as várias abordagens computacionais utilizadas para o PDP, certamente a mais utilizada é o AG, devido à sua simplicidade e eficiência em encontrar boas soluções em um espaço de busca complexo e fortemente restrito [Lopes 2008].

AGs possuem a capacidade de encontrar boas soluções em um tempo de processamento razoável. Porém, este tempo aumenta consideravelmente para problemas complexos, como é o caso do PDP. Para tais situações, AGs paralelos tem

sido utilizados com mais eficácia do que as versões sequenciais. Neste trabalho, foi desenvolvido um AG paralelo do tipo *master-slave* síncrono [Cantú-Paz 2000]. Nesta abordagem a carga de processamento é dividida entre vários processadores (*slaves*), sobre a coordenação de um processador central *master*. Esta abordagem é particularmente interessante para problemas onde a computação da função de *fitness* é muito custosa, como é o caso deste trabalho (ver seção 3.3). O *master* é responsável por inicializar a população, executar o procedimento de seleção e os operadores genéticos (*crossover* e mutação) e por distribuir os indivíduos aos *slaves*. Os *slaves* são responsáveis por processar a função de *fitness* de cada indivíduo recebido.

3.1. Codificação

AGs evoluem uma população de indivíduos, onde cada um representa uma solução candidata ao problema. Cada indivíduo é codificado sob a forma de um ou mais cromossomos. A codificação utiliza o alfabeto binário ou outro que seja adequado à natureza das variáveis sendo manipuladas. A forma de codificação e o tamanho do cromossomo definem o tamanho do espaço de busca e influenciam fortemente a dificuldade em resolver o problema, devido ao estabelecimento da epistasia entre os genes (variáveis codificadas) do cromossomo.

Para o PDP há várias abordagens para representar um dobramento em um cromossomo [Lopes 2008]: matriz de distâncias, coordenadas cartesianas ou coordenadas internas. Com base no estudo de [Krasnogor et al. 1999] para o modelo 2DHP, foram utilizadas coordenadas internas relativas neste trabalho. Neste sistema de coordenadas, uma dada conformação da proteína é representada como sendo um conjunto de movimentos sobre uma treliça cúbica. Assim, a posição de cada aminoácido na cadeia é relativa ao seu predecessor. No modelo 3DHP-SC, os aminoácidos da proteína são representados por um *backbone* (B) e uma cadeia lateral, hidrofóbica (H) ou polar (P). No espaço tridimensional há cinco movimentos relativos possíveis para o *backbone* (**E**squerda, **F**rente, **D**ireita, **B**aixo, **C**ima) e outros cinco para a cadeia lateral, relativos ao *backbone* (**e**squerda, **f**rente, **d**ireita, **b**aixo, **c**ima). Portanto, a combinação dos possíveis movimentos do *backbone* e da cadeia lateral leva a 25 possibilidades, representadas pelo conjunto: {Ee, Ef, Ed, Eb, Ec, Fe, Ff, Fd, Fb, Dc, De, Df, Dd, Db, Dc, Be, Bf, Bd, Bb, Bc, Ce, Cf, Cd, Cb, Cc}. Cada elemento desde conjunto é representado como um único símbolo, conforme a tabela 1, sendo este o alfabeto utilizado para codificar o cromossomo do AG. Considerando o dobramento de uma proteína com n aminoácidos, um cromossomo com $n-1$ genes representará o conjunto de movimentos do *backbone* e da cadeia lateral na treliça.

Para representar a posição dos aminoácidos na treliça cúbica, a coordenada Cartesiana do *backbone* e da cadeia lateral será representada por x_i (linha), y_i (coluna), z_i (profundidade), e obtida a partir do movimento relativo do aminoácido atual e da posição do aminoácido predecessor. Portanto, um procedimento recursivo é necessário, iniciando desde o primeiro *backbone*, situado na origem do sistema de coordenadas (posição (0, 0, 0)) e com cadeia lateral situada na posição (0, -1, 0).

3.2. População inicial

O uso de coordenadas relativas internas para o PDP cria um problema na inicialização do AG quando a população inicial é gerada. Desde que a geração seja realizada

Tabela 1. Esquema de codificação das coordenadas relativas internas para o PDP.

Movimentos		*Backbone*				
		E	**F**	**D**	**B**	**C**
Cadeia Lateral	**e**	0	5	A	F	K
	f	1	6	B	G	L
	d	2	7	C	H	M
	b	3	8	D	I	N
	c	4	9	E	J	O

de maneira aleatória, o número de colisões entre elementos (*backbone* e cadeias laterais) tende a ser grande [Lopes 2008]. Consequentemente, na geração da população inicial não se pode garantir indivíduos válidos. Isto conduz o AG a um gasto de tempo de processamento e geração de conformações inválidas antes que bons resultados possam ser obtidos. Para contornar esta condição, este trabalho propõe um método especializado para a geração da população inicial. A população é dividida em duas partes geradas aleatoriamente, sendo que uma delas é composta de indivíduos livres de colisão. A taxa de indivíduos livres de colisão pode ser configurada pelo usuário através de um parâmetro (ver seção 3.4). A geração dos indivíduos sem colisão é realizada utilizando uma estratégia de *backtracking*, da seguinte maneira:

O *backbone* do primeiro aminoácido é situado na origem, com a sua cadeia lateral na posição (0, -1, 0). O movimento do próximo aminoácido é selecionado aleatoriamente. Se o movimento conduz a uma colisão com o *backbone* ou com a cadeia lateral de outro aminoácido já posicionado na treliça, o *backtracking* é realizado e outro movimento é selecionado aleatoriamente. Neste caso, o movimento selecionado pertence a um conjunto de movimentos possíveis que não conduzam a uma colisão.

O método proposto para a geração da população inicial consome um tempo de processamento significativo. Contudo, garante a qualidade dos indivíduos da primeira geração, permitindo a evolução do AG para boas soluções.

3.3. Função objetivo

A cada geração os indivíduos são avaliados de acordo com a sua capacidade em apresentar uma solução ao PDP. Um cromossomo contendo um *string* de movimentos relativos é decodificado em um vetor de coordenadas Cartesianas. Estas, por sua vez, são utilizadas para cálculo de uma função objetivo que fornece o valor de *fitness* do indivíduo.

A função objetivo proposta neste trabalho é composta por termos que levam em consideração não somente a energia livre da conformação, mas também penaliza o número de colisões. Esta função, apresentada na Equação (2), também incorpora termos que mensuram a compacidade dos aminoácidos hidrofóbicos e polares.

$$fitness = Energia \cdot RaioG_H \cdot RaioG_P \tag{2}$$

Nesta equação, *Energia* corresponde ao termo que leva em consideração o

número de contatos hidrofóbicos, interações hidrofílicas, e interações com o *backbone*, o número de colisões (considerado como penalidades) e o peso destas penalidades; $RaioG_H$ representa o raio de giração dos resíduos hidrofóbicos; $RaioG_P$ representa o raio de giração dos resíduos hidrofílicos. Estes termos são detalhados a seguir. Esta função foi originalmente proposta por [Scapin and Lopes 2008] para o modelo 2DHP e adaptada neste trabalho para o modelo 3DHP-SC.

A energia livre de uma conformação proteica com cadeias laterais é dada na Equação (1), onde todas as interações possíveis, entre resíduos e *backbone* são levadas em consideração. Com esta equação também é possível atribuir ponderações (todos os ϵ_i) para as interações entre resíduos hidrofóbicos ou polares.

A conformação em avaliação pode apresentar colisões, seja entre cadeias laterais de aminoácidos diferentes ou entre cadeias laterais e o *backbone*. Desta maneira, a conformação é considerada como fisicamente inválida mas, no entanto, pode conter algum material genético promissor no cromossomo. Neste caso ela é penalizada no seu valor de *fitness* através do termo $Energia$. Esta penalização é composta pelo número de posições na treliça ocupadas por mais de um elemento (NC – número de colisões), multiplicado por um peso de penalização (PP), conforme apresentado na Equação (3).

$$Energia = H - (NC \cdot PP) \tag{3}$$

Uma questão importante a ser levada em consideração ao tentar predizer a estrutura tridimensional de uma proteína (em modelos de treliça) está relacionada com a sua (hiper)superfície de energia (ou *energy landscape*), ou seja, como a energia livre está distribuída ao se considerar todas as possíveis conformações.

O modelo 2DHP original utiliza apenas o número de interações hidrofóbicas para avaliar indivíduos [Dill 1985, Lopes 2008]. Esta abordagem gera grandes regiões planas na superfície de energia [Krasnogor et al. 1999] e muitos mínimos locais, tornando ineficiente métodos de busca local.

Para aumentar a eficiência da busca na superfície de energia foi proposto por [Scapin and Lopes 2008] o uso do conceito físico de raio de giração, como parte da função objetivo. O raio de giração indica quão compacto se encontra um conjunto de pontos (neste caso, aminoácidos em uma treliça). Conjuntos mais compactos possuem menor raio de giração. Esta medida de compacidade incorporada em dois termos da função objetivo (equação 2) avalia separadamente resíduos hidrofóbicos e polares através da equação 4.

$$RG_{aa} = \sqrt{\frac{\sum_{i=1}^{N_{aa}} (x_i - \overline{X}) + (y_i - \overline{Y}) + (z_i - \overline{Z})}{N_{aa}}} \tag{4}$$

Onde x_i, y_i e z_i são as coordenadas do i-ésimo resíduo do tipo "aa" da proteína, sendo "aa" ou hidrofóbico (H) ou polar (P); $\overline{X}$, $\overline{Y}$ e $\overline{Z}$ são as médias de todos os x_i, y_i e z_i e N_{aa} é o número de resíduos do tipo "aa" da proteína. Desta maneira, o raio de giração hidrofóbico ($RaioG_H$) deve ser minimizado para se obter um núcleo

hidrofóbico com grande número de contatos, conforme a equação 5, onde $maxRG_H$ é o valor do raio de giração quando toda a cadeia está esticada. De maneira oposta, o raio de giração polar ($RaioG_P$) deve ser maximizado de modo a levar os resíduos polares para as extremidades da conformação, como mostrado na equação 6.

$$RaioG_H = maxRG_H - RG_H \tag{5}$$

$$RaioG_P = \begin{cases} 1 & \text{se } (RG_H - RG_P \geq 0) \\ \frac{1}{1-(RG_P - RG_H)} & \text{senão} \end{cases} \tag{6}$$

3.4. Parâmetros de controle do AG

Não há um procedimento específico para se realizar o ajuste dos parâmetros de um AG. Optou-se por realizar uma série de experimentos com diversas combinações de valores de parâmetros e rodar várias vezes cada experimento com sementes aleatórias diferentes. Outros procedimentos de auto-ajuste de parâmetros foram propostos por [Maruo et al. 2005], mas este trabalho não contempla este assunto. O melhor conjunto de parâmetros encontrado, analisando-se os resultados médios, foi escolhido para os experimentos posteriores. Foram utilizados os seguintes parâmetros: número de gerações (3000), tamanho da população (472), probabilidade de *crossover* de dois pontos (80%), probabilidade de mutação multibit (3%), escalonamento linear ($C = 1.40$), método de seleção (torneio estocástico), taxa de indivíduos sem colisão (20% da população).

4. Experimentos e Resultados

Os experimentos realizados neste trabalho foram executados em um *cluster* de 30 computadores *core 2-quad* rodando Linux e utilizando o pacote MPICH2[1], para a implementação da interface de troca de mensagens entre processos.

Nos experimentos, cinco sequências sintéticas de aminoácidos foram utilizadas como *benchmark* (Tabela 2). Estas sequências tem sido largamente utilizadas por outros pesquisadores para o modelo 3DHP, desde que foram propostas por [Yue and Dill 1993]. Entretanto, esta é a primeira vez em que são utilizadas para o modelo 3DHP-SC. Como o modelo computacional empregado neste trabalho tem sido muito pouco explorado na literatura, não há *benchmarks* conhecidos para o modelo 3DHP-SC. Apenas para fins de comparação, o valor máximo conhecido para o número de contatos hidrofóbicos para tais sequências é apresentado na coluna (E_{3DHP}) da tabela, obtidos por [Yue and Dill 1993]. Como mostrado na seção 2, o modelo 3DHP-SC tem mais graus de liberdade do que o 3DHP tradicional. Assim, para uma mesma sequência, é de se esperar que o número de contatos hidrofóbicos para o modelo 3DHP-SC seja maior do que para o modelo 3DHP.

Para cada sequência de *benchmark*, o AG foi executado 100 vezes com sementes aleatórias diferentes. Os resultados são apresentados na Tabela 3. Nesta tabela, a primeira coluna identifica a sequência, as três seguintes identificam o número da geração em que o melhor indivíduo foi encontrado e o número máximo e médio de

[1]Disponível em: http://www.mcs.anl.gov/research/projects/mpich2/

Tabela 2. Benchmarks utilizados para teste do modelo 3DHP.

Ordem	Número de aminoácidos	Sequência HP	E_{3DHP}
1	27	$HP^4H^4P(PH)^3H(HP)^2PH^2P^2H$	16
2	27	$HP^3H^4(PH)^2HP^3HPH(HP)^2P^2HP$	15
3	27	$HPH^2(PPHH)^2H(HPPP)^2H^3P^2H$	16
4	31	$(HHP)^3H(HHHHHPP)^2H^7$	28
5	36	$PH(PPH)^{11}P$	14

gerações para encontrar o melhor indivíduo. A coluna T_p se refere ao tempo de processamento médio. As três últimas colunas mostram, respectivamente, o número máximo de contatos hidrofóbicos encontrado, o valor médio para todas as rodadas e o desvio padrão correspondente.

Tabela 3. Resultados para os benchmarks.

Ordem	Geração			T_p(s)	$E_{3DHP-SC}$		
	melhor	max	média		max	média	σ
1	2603	2991	1513,05	242,58	19	14,69	1,99
2	769	2863	1598,55	236,94	16	12,58	1,69
3	1305	2982	1717,65	239,38	21	14,91	2,05
4	1161	2954	1884,15	310,5	34	27,79	2,69
5	2875	2996	2182,05	488,64	14	11,95	2,08

Para cada sequência testada, o melhor dobramento encontrado é representado graficamente nas Figuras 1(a), 1(b), 1(c), 1(d) e 1(e). Os resíduos hidrofóbicos e polares são representados, respectivamente, por esferas vermelhas e azuis. O *backbone* e as conexões entre aminoácidos são mostrados em cinza.

5. Conclusões e trabalhos futuros

O PDP ainda é um problema em aberto do ponto de vista computacional, posto que sua solução recai em um problema NP-difícil. Nem mesmo utilizando os modelos mais simplificados é possível encontrar a conformação nativa de proteínas reais (com várias dezenas ou centenas de aminoácidos). A alternativa tem sido a utilização de métodos heurísticos, notadamente aqueles oferecidos pela computação evolucionária.

O modelo 3DHP tem sido bastante explorado na literatura recente. Entretanto, o 3DHP-SC, embora tenha maior expressividade do que o 3DHP, tem sido muito pouco abordado devido à maior complexidade envolvida. Como consequência, também não há *bechmarks* para este modelo e neste trabalho foram adaptados *bechmarks* para o 3DHP. Portanto, os resultados obtidos neste trabalho representam uma contribuição importante em relação a este assunto.

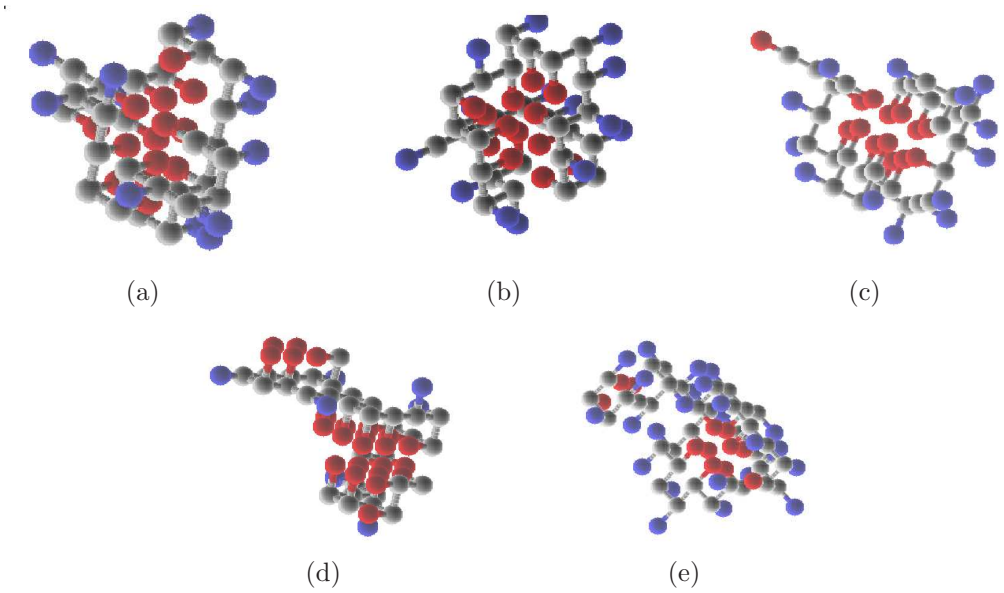

Figura 1. Melhor dobramento encontrado para as sequências 1 (a), 2 (b), 3 (c), 4 (d) e 5 (e).

Este trabalho mostra que os AGs são uma ferramenta eficiente para lidar com o PDP utilizando o modelo 3DHP-SC. Embora os resultados obtidos não possam ser considerados como ótimos, eles são coerentes com o modelo, pois o número de contatos hidrofóbicos encontrados no 3DHP-SC é sempre maior (ou igual) que no 3DHP.

Nas Figuras 1(a) – 1(e) pode ser observada a formação de um núcleo hidrofóbico, parcialmente protegido por aminoácidos polares. Isto é particularmente observável nas sequências menores. Devido à natureza do problema de otimização, a formação deste núcleo já era esperada. Isto sugere que a função de *fitness* proposta é capaz de levar a conformações que mimetizam propriedades bioquímicas de proteínas reais durante o processo de dobramento.

Foi observado que, à medida que o tamanho da sequência aumenta, o AG diminui a sua eficácia. Este comportamento é conhecido quando se aplica métodos metaheurísticos (com o AG) para problemas de complexidade NP. Isto pode ser observado sutilmente nas figuras 1(c), 1(d) e 1(e), onde o núcleo hidrofóbico ou está bipartido ou não está completamente protejido por resíduos polares. Este fato sugere a necessidade de operadores genéticos especiais que incorporem conhecimento biológico, bem como a hibridização do AG com técnicas de busca local para aumentar a sua efetividade.

A utilização de computação paralela para o PDP foi promissora, pois permitiu a obtenção de resultados satisfatórios em tempos de processamento razoáveis, estando de acordo com o que foi preconizado por [Lopes 2008]. Trabalhos futuros considerarão, também, o uso de abordagens baseadas em *hardware* reconfigurável, tal como [Armstrong Junior et al. 2007], para acelarar o processamento.

De maneira geral, os resultados são bons e promissores para suportar a con-

tinuidade deste trabalho. Acredita-se que este trabalho seja uma contribuição interessante para esta área de pesquisa, pois este modelo simula as características das proteínas de uma maneira mais realista do que o modelo 3DHP. Em trabalhos futuros serão feitos mais experimentos com estes e outros *benchmarks*, bem como o teste de outras funções de *fitness* mais complexas e operadores especializados baseados em conhecimento biológico.

Referências

Armstrong Junior, N., Lopes, H., and Lima, C. (2007). Reconfigurable computing for accelerating protein folding simulations. *Lecture Notes in Computer Science*, 4419:314–325.

Berger, B. and Leighton, F. (1998). Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *Journal of Computational Biology*, 5(1):27–40.

Cantú-Paz, E. (2000). *Efficient and Accurate Parallel Genetic Algorithms*. Springer.

Dill, K. (1985). Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6):1501–1509.

Krasnogor, N., Hart, W., Smith, J., and Pelta, D. (1999). "Protein structure prediction with evolutionary algorithms". In *International Genetic and Evolutionary Computation Conference (GECCO)*, volume 4, pages 1596–1601.

Li, M. S., Klimov, D. K., and Thirumalai, D. (2002). Folding in lattice models with side chains. *Computer Physics Communications*, 147(1-2):625–628.

Lopes, H. (2008). "Evolutionary algorithms for the protein folding problem: a review and current trends.". In Smolinski, T., Milanova, M., and Hassanien, A.-E., editors, *Applications of Computational Intelligence in Bioinformatics and Biomedicine: Current Trends and Open Problems*, volume I, pages 297–315, Heidelberg. Springer-Verlag.

Maruo, M., Lopes, H., and Delgado, M. (2005). Self-adapting evolutionary parameters: encoding aspects for combinatorial optimization problems. *Lecture Notes in Computer Science*, 3448:154–165.

Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, Berlin, 3^{rd} edition.

Scapin, M. and Lopes, H. (2008). "A hybrid genetic algorithm for the protein folding problem using the 2D-HP lattice model.". In Yang, A., Shan, Y., and Thu, L., editors, *Success in Evolutionary Computation*, number 92 in Studies in Computational Intelligence, pages 205–224, Heidelberg. Springer-Verlag.

Yue, K. and Dill, K. (1993). Sequence-structure relationships in proteins and copolymers. *Physical Review E.*, 48(3):2267–2278.