10th Brazilian Congress on Computational Intelligence (CBIC'2011), November 8 to 11, 2011, Fortaleza, Ceará Brazil

ALGORITMOS GENÉTICOS COM FUNÇÃO DE AVALIAÇÃO DINÂMICA PARA O PROBLEMA DE PREDIÇÃO DE ESTRUTURAS DE PROTEÍNAS

Luis Henrique U. Ishivatari[1], Lariza L. de Oliveira[1], Fernando L. B. da Silva[2], Renato Tinós[1]

[1]Depto. de Computação e Matemática, FFCLRP, Universidade de São Paulo (USP)
Av. Bandeirantes 3900 – 14040-901 – Ribeirão Preto – SP – Brasil

[2]Depto. de Física e Química, FCFRP, Universidade de São Paulo (USP)
Av. Bandeirantes 3900 – 14040-901 – Ribeirão Preto – SP – Brasil

{luishenrique.ln,larizalaura}@usp.br, rtinos@ffclrp.usp.br, fernando@fcfrp.usp.br

Abstract – In this paper, Genetic Algorithms (GAs) are used for the protein structure prediction problem in a dynamic framework. The objective is to investigate if the changes on the fitness function during the evolution process of the GA are beneficial for the optimization process. Changing the fitness function during the evolutionary process can reduce the population's premature convergence problem, common on the search space for the protein structure prediction problem. A C-alfa minimalist model is used to generate the dynamic fitness function for the GA. Tests with static and dynamic fitness functions are presented, and the results indicate an improvement on the RMSD when dynamic fitness functions are used.

Keywords – Genetic Algorithms, Protein Structure Prediction, Dynamic Fitness Function, Protein

1 Introdução

Proteínas, que são formadas por uma sequência de aminoácidos unidos por ligações covalentes, são moléculas constituintes de estruturas das células e responsáveis por diversas atividades biológicas [1]. A determinação da estrutura tridimensional das proteínas pode ser realizada experimentalmente, através de métodos de cristalografia de Raios-X e Ressonância Nuclear Magnética [2]. No entanto, esses métodos possuem algumas limitações, como alto custo e condições especiais para aplicação. De acordo com a hipótese de Afinsen [3], a Predição de Estruturas de Proteínas (PEP) pode, a princípio, ser realizada conhecendo-se apenas sua estrutura primária, ou seja, a seqüência de aminoácidos que forma sua cadeia. Esta determinação através de sistemas computacionais pode ser vista como um problema de otimização, no qual dada uma sequência de aminoácidos, deve-se determinar qual é a estrutura tridimensional da proteína, dentre as muitas estruturas possíveis, que minimiza uma função da energia potencial (função de fitness ou função de avaliação).

A predição de estrutura de proteínas utilizando técnicas de aprendizado de máquina é um tema bastante abordado na literatura [4], [5], [6]. Algoritmos Genéticos (AGs), devido a suas características, como uso de populações de soluções e de operadores estocásticos, são interessantes para problemas complexos como a PEP. De fato, tem havido um crescente interesse por parte dos pesquisadores em aplicar tais algoritmos na determinação da estrutura tridimensional de proteínas. Entretanto, apesar de alguns resultados promissores, AGs apresentam dificuldades neste problema. Os motivos principais desta dificuldade são dois: primeiro, a existência de um número extremamente grande de soluções possíveis e de ótimos locais; segundo, a escolha de uma função de avaliação pertinente à tarefa de otimização da estrutura da proteína. Desta forma, diversas técnicas foram empregadas para tentar promover melhorias nos AGs utilizados no problema de PEP, como por exemplo: manutenção da diversidade da população [7]; inserção de conhecimento no AG [8]; uso de otimização multi-objetivo [9]; soluções híbridas de hill-climbing e AG [10] e remoção de gêmeos na população do AG [11].

Este trabalho tem como principal objetivo investigar o uso de funções de avaliação que mudam durante o processo de otimização realizado por um AG no problema de predição de estruturas tridimensionais de proteínas para um número finito de gerações. Espera-se que tal procedimento possa facilitar o processo de otimização realizado pelo AG ao permitir que a população escape de ótimos locais inerentes ao espaço de busca do problema de PEP. Na Seção 2 os detalhes da metodologia são descritos. Na Seção 3 os experimentos realizados são apresentados e, na Seção 4, os resultados dos experimentos são discutidos. Por fim, na Seção 5 as conclusões e comentários gerais do trabalho são relatados.

2 Metodologia

Neste trabalho, o modelo minimalista do tipo C-alfa proposto em [12] é utilizado para avaliar os indivíduos do AG, que representam possíveis conformações para a proteína. Propõe-se aqui investigar se a mudança periódica da função de avaliação é útil para o problema de PEP usando AGs. A seguir, é feita uma breve introdução aos AGs, bem como o detalhamento da sua implementação. Na Seção 2.2, o modelo minimalista do tipo C-alfa utilizado para calcular a função de fitness do AG é

explicado. Por fim, na Seção 2.3, algumas características de problemas de otimização dinâmica (DOP) usando AGs são discutidas.

2.1 Algoritmos Genéticos

No AG padrão, uma população de indivíduos (ou cromossomos) representando soluções do problema é sujeito a operadores inspirados em mecanismos encontrados na evolução biológica. Em cada geração do processo evolutivo, a população de soluções candidatas passa por dois tipos de operadores: de reprodução e de seleção. Geralmente, para o primeiro tipo, são empregados os operadores de mutação e recombinação (crossover), que são aplicados aos indivíduos com certa probabilidade pré-definida. No crossover, partes dos cromossomos de dois indivíduos selecionados da população são permutadas. Já na mutação, elementos do cromossomo são alterados através de regras estocásticas. Os operadores de seleção estão relacionados com a escolha dos indivíduos para serem reproduzidos ou comporem a população na próxima geração. Como exemplo de tais operadores, pode-se citar o elitismo, que copia o melhor indivíduo da população para a população seguinte, e o método da roleta, que seleciona indivíduos com probabilidade proporcional ao seu fitness relativo.

Neste trabalho, para a composição da população inicial e para dois dos tipos de mutação empregados, é utilizada uma base de dados com os ângulos diedrais φ e ψ, que são respectivamente, os ângulos dos planos formados pelos átomos C'-N-Cα-C' e N-Cα-C'-N. Esses ângulos são obtidos de observações experiementais [13], sendo a base de dados ordenada e dividida por aminoácido e por estrutura secundária. Este procedimento, ajuda o algoritmo a guiar o processo de otimização para regiões promissoras do espaço de busca. A representação dos indivíduos (Figura 1) é do tipo real, sendo que o cromossomo de cada indivíduo tem tamanho igual a duas vezes o número de resíduos da proteína, já que cada alelo do cromossomo é composto por um dos ângulos diedrais φ e ψ associados com cada resíduo. Além disso, para facilitar o processo de utilização dos dados experimentais presentes na base de dados, o índice desta base eventualmente utilizado para gerar os ângulos diedrais φ e ψ (na população inicial e em dois tipos de mutação) é armazenado [8].

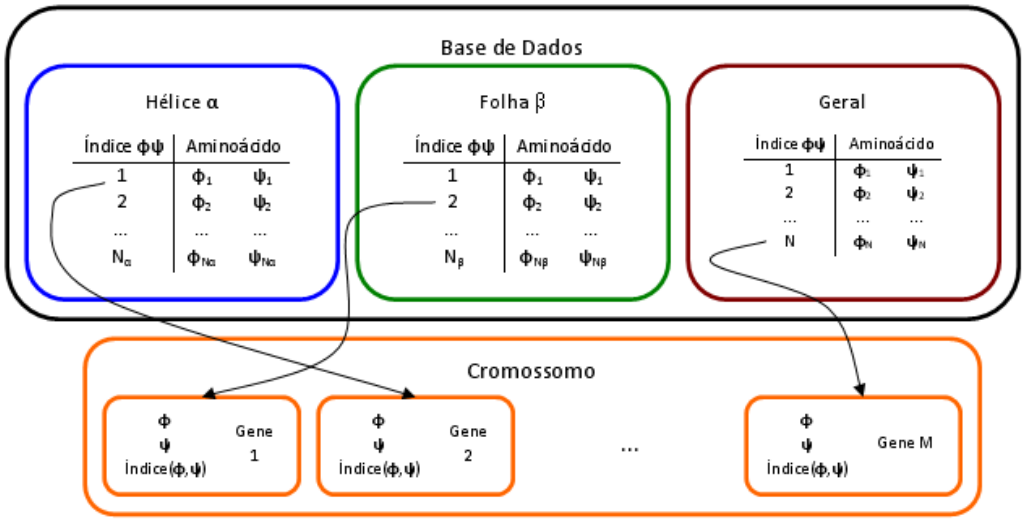


Figura 1. Representação da codificação dos indivíduos do AG [Oliveira et al., 2010].

Dois operadores de seleção são aqui empregados: elitismo, para preservar o melhor indivíduo obtido em cada geração; e torneio, no qual o melhor de três indivíduos aleatoriamente escolhidos é selecionado. Os operadores de reprodução utilizados neste trabalho são crossover de dois pontos e mutação. A mutação possui três variantes, todas com a mesma chance de ocorrer:

- Mutação com auxílio da base de dados. Neste tipo de mutação, utiliza-se o último índice para a base de dados conhecido, que é então alterado aleatoriamente num intervalo de [-5, 5] posições. Como a base está ordenada, espera-se que neste tipo de mutação ocorram pequenas mudanças na estrutura da proteína.
- Mutação com auxílio de toda a base de dados. Por outro lado, nesta mutação é sorteado um novo índice entre [1, *N*], onde *N* é o tamanho da base de dados referente ao resíduo analisado. O intuito desta mutação é tentar promover mudanças drásticas nos ângulos dos aminoácidos.
- Mutação real. Ao ser escolhido o alelo para sofrer mutação, sorteia-se em qual dos dois ângulos ocorrerá a mutação real no intervalo de [-5º, 5º] com distribuição uniforme, mantendo o último valor do índice para a base de dados. O

Excluído: ocorra

objetivo desta mutação é testar outros ângulos possíveis que não estão na base de dados. Apesar do intervalo da mudança ser pequeno, há uma possibilidade de esta mutação resultar na perda de estruturas secundárias já existentes.

2.2 Função de Avaliação

AGs precisam também de uma função de avaliação, que é utilizada para guiar o processo de evolução. A função de avaliação do AG é dependente do problema estudado. Neste trabalho, é utilizado um modelo minimalista do tipo C-alfa, baseado na referência [12]. Modelos minimalistas (*coarse-grained*) assumem que um aminoácido pode ser representado apenas por um ou alguns átomos (*beads*). Outras informações dos resíduos são geralmente incorporadas ao simplificar-se o campo de força. Por outro lado, há modelos *full-atom*, os quais consideram todos os átomos dos aminoácidos ao calcularem a energia potencial. Estes modelos são computacionalmente mais custosos e possuem campos de força bem conhecidos, como o CHARMM 35 [14].

O campo de força utilizado aqui possui quatro termos, correspondentes a energia do ângulo de ligação, energia do ângulo de torção, que têm papel importante na formação das estruturas secundárias; energia de Van de Waals e energia de ligação de hidrogênio, os quais são importantes para tornar as estruturas secundárias mais próximas e rígidas. A equação a seguir apresenta o campo de força descrito em [12]:

$$E = E_{Lig} + E_{Diedral} + E_{VdW} + E_{HB} \quad (1)$$

sendo:

$$E_{Lig} = \sum_{angulos} \frac{1}{2} k_\theta (\theta - \theta_0)^2 \, , E_{VdW} = \sum_{i,\, j \geq i+3} 4\varepsilon_H S_1 \left[(\frac{\sigma}{r_{ij}})^{12} - S_2 (\frac{\sigma}{r_{ij}})^6 \right],$$

$$E_{Diedral} = \sum_{Diedrais} \left[\begin{array}{c} A[1+\cos(\phi+\phi_0)] + B[1-\cos(\phi+\phi_0)] + \\ + C[1+3\cos(\phi-\phi_0)] + D[1+\cos(\phi+\phi_0+\frac{\pi}{4}) \end{array} \right], E_{HB} = \sum_{ligaçõesHidrogênio} U_{HB}$$

nas quais θ é o ângulo de ligação definido por três C-α consecutivos, Φ é o ângulo diedral definido por quatro *beads* de C-α consecutivos e r_{ij} é a distância entre os *beads* *i* e *j*. ε_H é a magnitude do potencial de interação de Van der Waals, θ_0 é o ângulo de ligaçãoda estrutura nativa, *A, B, C, D* e Φ_0 são parâmetros que dependem do tipo diedral, k_θ é a constante de força do potencial harmônico, σ é a soma do raio de van der Waals e S_1 e S_2 são parâmetros que se referem a atração e repulsão dos *beads*. A parametrização e o detalhamento de E_{HB} são descritos pelos autores de [12].

2.3 Problemas de Otimização Dinâmica

Os autores de [15] argumentam que utilizar objetivos que variam ao longo do tempo pode acarretar em uma aceleração do processo de evolução, visto que a mudança de objetivos muda também o espaço de busca, o que pode alterar a posição dos ótimos locais. Eles afirmam também que esta aceleração é diretamente proporcional à complexidade do problema. Esta é uma afirmação importante para o problema de predição de estruturas de proteínas, uma vez que se sabe que o espaço de busca no problema de PEP possui muitos ótimos locais [16], [17].

Existem diversas maneiras de se mudar o espaço de busca. Em [15], são propostos diversos modelos, mas conclui-se que, em geral, apenas dois conseguem produzir aceleração: o *Modular Varying Goal* (MVG), no qual a mudança do espaço de busca é predefinida, e o *Random Varying Goal* (RVG), no qual a mudança é aleatória. A metodologia empregada aqui é inspirada nesses dois modelos.

Aqui, as mudanças no espaço de busca são provocadas pela alteração da função de avaliação, utilizando-se para este fim, pesos em dois dos termos da Eq. (1). Estes pesos, que podem variar com o tempo, são aplicados nos termos E_{VdW} e E_{HB} pelo motivo explicado a seguir. Os autores de [18] comentam que existem duas fases distintas no processo de dobramento da proteína. Primeiro há a "fase explosiva", na qual são formadas as estruturas secundárias. Em seguida ocorre o "colapso hidrofóbico", que provoca a redução no raio da proteína. De acordo com esta afirmação, a proposta é aumentar o enfoque no colapso hidrofóbico, uma vez que modelo já assume informações sobre as estruturas secundárias e de acordo com trabalhos anteriores [8] a formação de estruturas do tipo hélice-α conseguem ser reproduzidas utilizando-se uma função de avaliação estacionária. Desta forma, e de acordo com estudos anteriores feitos pelos autores acerca do efeito de cada termo da Eq. (1) sobre o processo de otimização [19], os pesos para os termos de energia de Van der Waals e Ligação de Hidrogênio são

modificados. É proposto, então, transformar a Equação (1) em uma função dinâmica através da equação abaixo, sendo os pesos (w_{VdW} e w_{HB}) números reais positivos, que podem mudar a cada ciclo de τ gerações.

$$E = E_{Lig} + E_{Diedral} + w_{VdW}(t)E_{VdW} + w_{HB}(t)E_{HB} \quad (2)$$

Os padrões de mudança para os pesos da Eq. (2) são descritos na Seção 3. É interessante notar que a análise de tais pesos pode ajudar na elucidação de quais contribuições são mais significativas em diferentes fases do dobramento protéico.

3. Experimentos

Nos experimentos apresentados neste artigo, foi utilizado um AG padrão para testar se a função de fitness dinâmica no problema de PEP é benéfica. A proteína utilizada para os testes foi a Proteína G (PDB ID: 2GB1), uma vez que esta proteína já fora utilizada em trabalhos anteriores [8], [12], permitindo maior controle do experimento.

Alguns testes iniciais foram realizados para ajustar os parâmetros do AG, como o número de gerações e os valores dos pesos utilizados. Para os cinco experimentos, os parâmetros da taxa de mutação e de crossover foram mantidos fixos em 15% e 60%, respectivamente. Estes valores foram determinados de acordo com experimentos anteriores com a função de fitness fixa [8]. De acordo com os testes, verificou-se que a partir de 50 gerações por resíduo (aproximadamente 2500 gerações no total, visto que a proteína 2GB1 possui 56 resíduos), o fitness em geral apresenta pouca variação. Desta forma, escolheu-se τ – a duração de um ciclo – igual a 2500 gerações, ou seja, o programa muda a função de fitness a cada 2500 gerações. Ao todo foram considerados 23 ciclos de 2500 gerações em cada execução do AG, totalizando 57500 gerações para cada execução. Cada experimento foi executado 10 vezes com diferentes sementes aleatórias.

Nos experimentos em que a função de fitness é dinâmica, foram escolhidos para os pesos da Eq. (2) valores entre 28 e 112 (metade e dobro do número de resíduos, respectivamente). Os detalhes dos experimentos realizados são sumarizados na Tabela 1.

Tabela 1. Experimentos realizados.

Experimento	τ – tempo do ciclo (gerações)	Pesos nos 3 primeiros ciclos
Sem pesos (SP)	Não se aplica	w_{VdW}=w_{PH}=1 para todos os ciclos
Est-VdW	Não se aplica	w_{VdW}= 28 e w_{PH}=1 para todos os ciclos
MVG	2500	w_{VdW}=w_{PH}=1 (ciclo 1); w_{VdW}=28 e w_{PH}=1 (ciclo 2); w_{VdW}=1 e w_{PH}=56 (ciclo 3)
RVG 1	2500	w_{PH}=1 e aleatório entre 1 e 56 para w_{VdW}
RVG 2	2500	Aleatório entre 1 e 56 para w_{VdW} e w_{PH}

No experimento sem pesos (SP), não há ciclos, ou seja, a Eq. (1) é utilizada ao longo das 57500 gerações. Já no MVG, a alteração dos pesos segue a sequência da Tabela 1 e ao terminar os três ciclos de pesos, a sequência é executada novamente, até atingir o número total de gerações. No experimento RVG 1, o peso do termo de Van der Waals é alterado para um novo peso aleatório a cada troca de ciclo. O RVG 2 é similar ao RVG 1, exceto que o peso escolhido pode ser ou w_{VdW} ou w_{PH}, com 50% de chance cada. Os três experimentos dinâmicos, MVG, RVG 1 e RVG 2 são inspirados no trabalho [15]. Por fim, o experimento Est-VdW é similar ao experimento realizado em [19], onde a função de fitness é estática e um peso definido previamente é atribuído a componente de Van der Waals por todas as gerações da execução.

Excluído: 37500

4 Resultados

Os resultados são divididos em três partes. Na primeira parte são apresentados os resultados das energias obtidas, e na segunda parte os resultados relacionados ao *root mean squared deviation* (RMSD) das estruturas. Por fim, a terceira parte apresenta os resultados das estruturas 3D das soluções obtidas.

4.1 Energia

Como o fitness é dinâmico para três dos cinco experimentos, não é possível fazer uma comparação direta entre os valores de fitness dos experimentos. Dessa forma, optou-se por mostrar aqui a evolução dos valores de energia correspondentes à Eq. (1) durante a amostragem. As energias médias do melhor indivíduo de cada uma das dez execuções são apresentadas na Figura 2.

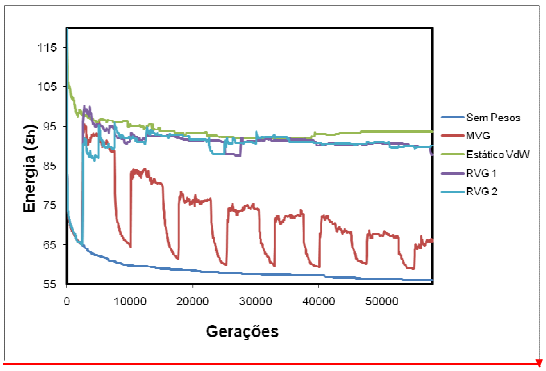
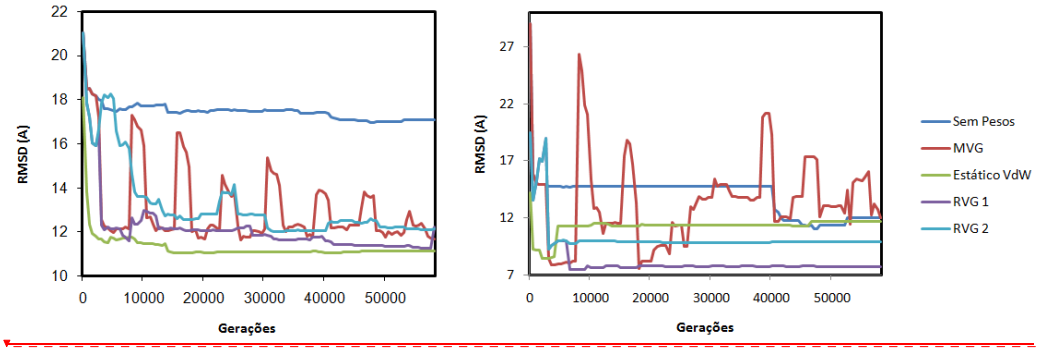


Figura 2. Energia (Eq. 1), em unidades reduzidas (ε_h), ao longo das gerações para a média das execuções dos melhores indivíduos dos cinco experimentos.

Pela Figura 2 é possível observar que o experimento SP (Sem Pesos) possui um decaimento rápido inicial e, apesar de lento nas últimas gerações, é contínuo como esperado, pois não há mudança no espaço de busca. Similarmente, o experimento Est-VdW também possui um decaimento rápido inicial e lento no final da execução, sendo que o Est-VdW possui valores de energia próximas aos dos experimentos RVGs, o que evidencia que a contribuição de Van der Waals não deve ser tão acentuada durante todo o processo do dobramento protéico. Já para o MVG, percebe-se um comportamento que é comum em problemas de otimização dinâmica: a energia diminui e quando há a troca no espaço de busca ele cresce bruscamente, dado que os indivíduos – os mesmos do ciclo anterior – podem não ser boas soluções no novo espaço. Os indivíduos se ajustam ao novo espaço e começam a reduzir o fitness novamente e assim sucessivamente. Ao longo dos ciclos, verifica-se que os vales e os picos em geral também ficam menores. Nos dois experimentos RVG, observa-se um comportamento inicial esperado de experimentos no qual as mudanças são aleatórias. Entretanto verifica-se que à medida que as gerações passam os valores ficam estagnados em torno de 95 ε_h o que pode indicar que a mudança nos valores aleatórios dos pesos pode não ter sido brusca o suficiente. É importante salientar também que o SP não muda o espaço de busca, o que a princípio daria certa vantagem sobre os outros experimentos, entretanto verifica-se que os vales do MVG chegam bem próximos aos valores do SP.

4.2 RMSD

Por ser um problema de predição de estruturas de proteínas, é importante que outras propriedades estruturais possam ser analisadas. Como neste trabalho usa-se uma proteína alvo conhecida (ou seja, a estrutura nativa já foi experimentalmente determinada), pode-se usar o RMSD entre a conformação obtida pelo AG e a estrutura nativa. Em outras palavras, o RMSD diz a proximidade de uma estrutura (geralmente em Å) em relação à outra. A Figura 3a apresenta um gráfico do RMSD para o melhor indivíduo (valor médio sobre as 10 execuções de cada experimento) ao longo das gerações, enquanto que a Figura 3b apresenta os resultados para a execução que obteve menor RMSD para cada experimento.



(a) (b)

Figura 3. (a) Média (em relação a 10 execuções) do RMSD (em Å) dos melhores indivíduos de cada experimento. (b) Valor de RMSD (em Å) para o melhor indivíduo ao longo das gerações onde se observou o menor RMSD.

Pela Figura 3a observa-se o mesmo comportamento discutido na Seção 4.1. Os saltos bruscos no RMSD coincidem com os saltos na energia, enquanto que o SP possui um decaimento lento e contínuo. O experimento Est-VdW possui um decaimento rápido inicial e permanece estável nas gerações finais, o que é típico de problemas em que a população fica presa em um ótimo do espaço de busca.

Pode-se observar que os valores obtidos pelo experimento SP ficam estagnados em aproximadamente 17 Å após 3000 gerações. Já os vales dos outros três experimentos dinâmicos conseguem chegar a valores próximos a 12 Å. Até mesmo os experimentos aleatórios conseguiram chegar a valores menores que o SP. Os valores obtidos para o Est-VdW a princípio parecem ser melhores que todos os outros experimentos, incluindo os experimentos com função de fitness dinâmica. Entretanto, como se trata de uma média das execuções ao longo das gerações, esta interpretação induz ao engano. Tal afirmação pode ser constatada na Figura 4, que apresenta os resultados para a execução que obteve o menor RMSD para cada experimento.

Excluído: para

Verifica-se pela Figura 3b que os três experimentos com fitness dinâmico conseguiram obter valores de RMSD menores que os experimentos estáticos e o experimento RVG 1 conseguiu manter o RMSD menor que o Est-VdW após a geração 18000.

Foi realizado o teste estatístico não-paramétrico *Wilcoxon Signed-Rank*, a fim de comparar os resultados dos experimentos MVG, RVG 1 e RVG 2 (dinâmicos) em relação aos resultados dos experimentos SP e Est-VdW (estáticos). São selecionados os menores RMSDs obtidos em cada execução de cada um dos experimentos. Com esses dados, o teste pareado bi-caudal foi feito utilizando a ferramenta estatística R [20]. O resultado do teste indica se dois conjuntos amostrais são estatisticamente diferentes, de forma que para se afirmar que um conjunto de RMSDs seja estatisticamente menor, a mediana dos conjuntos também deve ser comparada. A Tabela 2 apresenta os resultados obtidos e a comparação estatística (p-valor). Verifica-se que os três experimentos dinâmicos possuem um RMSD estatisticamente menor que o experimento SP (com nível de confiança 99%), mas não é possível afirmar que os experimentos dinâmicos são melhores ou piores que o experimento Est-VdW, apesar de verificar-se que o experimento MVG possui média e mediana menores que as obtidas pelo Est-VdW.

Mesmo que os experimentos dinâmicos não tenham obtido valores de RMSD estatisticamente menores que o Est-VdW, verificou-se a partir de um teste de diversidade da população – distância euclidiana média entre os pares de indivíduos da população – que em média, a distância euclidiana nos experimentos dinâmicos foi de 594, enquanto que no experimento Est-VdW a média foi de 437. Isso significa que com uma diversidade maior na população, os experimentos dinâmicos têm maiores chances de encontrar soluções melhores.

Formatado: Recuo: Primeira linha: 1,25 cm

Tabela 2. Resultados do RMSDs nos experimentos e comparação estatística (p-valor) nas duas últimas colunas. Estes resultados referem-se aos melhores valores obtidos em cada uma das 10 execuções de cada experimento.

	Média (Desvio Padrão)	Mediana	Mínimo	Máximo	SP x Outros	Est-VdW x Dinâmicos
SP	15,390 (±3,104)	15,873	11,004	20,102	Não se aplica	Não se aplica
Est-VdW	10,661 (±1,699)	10,317	8,452	13,563	0,004	Não se aplica
MVG	9,755 (±1,279)	10,081	7,754	11,271	0,002	0,160
RVG 1	10,515 (±1,678)	10,456	7,480	13,616	0,006	0,770
RVG 2	11,153 (±1,213)	11,500	9,276	12,781	0,004	0,492

4.3 Estruturas Tridimensionais

Além do RMSD, na PEP também é possível analisar as estruturas 3D geradas. Apesar de não ser exatamente uma métrica, já que a análise visual é subjetiva, é possível verificar a forma da estrutura em um determinado instante da execução. Isso pode ser importante para AGs com fitness dinâmico, pois é possível verificar o que ocorre ao trocar o espaço de busca. As imagens das Figuras 4 e 5 foram geradas utilizando a ferramenta Jmol [21] e representam as estruturas com menor RMSD obtidas para cada experimento.

Excluído:

Mesmo que os experimentos com pesos tenham diminuído consideravelmente o RMSD, nota-se que as estruturas – principalmente as hélices-α – estão um pouco mais curvadas do que a Fig. 4a (sem pesos). Isso ocorre devido ao fato de as

contribuições de Van der Waals e Ponte de Hidrogênio serem maiores, o que resulta em uma aproximação entre os átomos. A Figura 5 apresenta as estruturas obtidas no fim de cada ciclo de mudança em uma execução do experimento MVG.

Da Figura 5, nota-se que durante a evolução no experimento MVG o algoritmo desenovela a estrutura ao trocar o espaço de busca. Isto pode parecer, a princípio, mais custoso, porém ela pode ser importante no PEP, visto que no processo de desenovelamento as ligações são desfeitas e novas ligações mais promissoras possam ser formadas. Como dito anteriormente, sabe-se que o problema de PEP possui muitos ótimos locais, de modo que este desenovelamento pode ser eventualmente benéfico para retirar o AG dos ótimos locais.

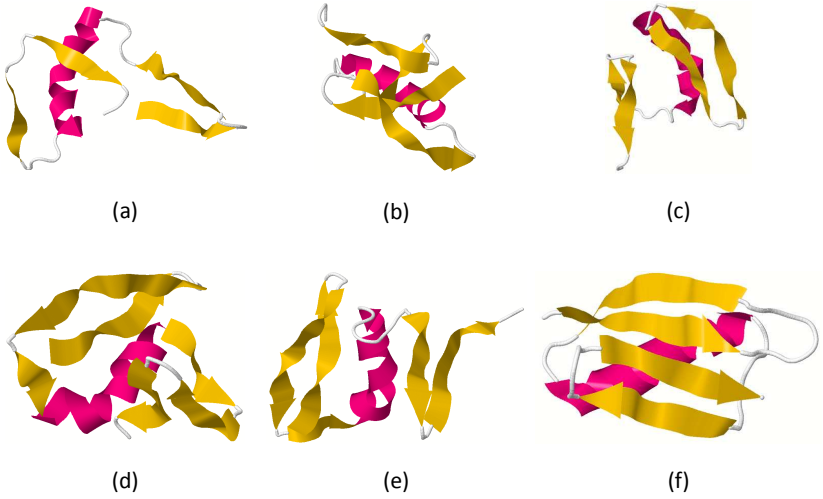


Figura 4. Estruturas 3D com menor RMSD, enfatizando as estruturas secundárias. (a) Experimento SP, 11,699 Å na geração 7000. (b) Est-VdW, 8,452 Å na geração 2000.(c) MVG, 7,574 Å na geração 18000. (d) RVG 1, 7.266 Å na geração 30500. (e) RVG 2, 6,401 Å na geração 8500. (f) Estrutura nativa da proteína 2GB1.

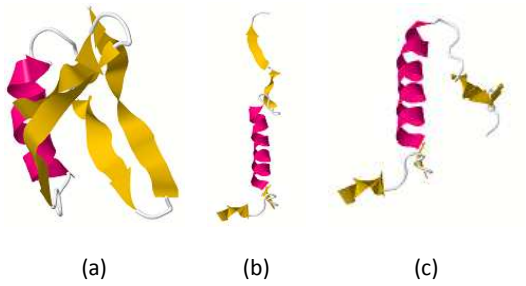


Figura 5. Estruturas 3D de uma execução do experimento MVG, evidenciando as estruturas secundárias. (a) 8.21 Å na geração 7500. (b) 26.368 Å na geração 8000. (c) 21.9 Å na geração 9000.

5 Conclusões

De uma maneira geral, pode-se dizer que os resultados utilizando AGs com função de avaliação dinâmica para o problema de predição de estruturas de proteínas foram positivos nos experimentos realizados neste trabalho. Foi possível obter valores de energia próximos aos obtidos pelo AG com função de avaliação estática, ainda que, como discutido anteriormente, a comparação não seja justa para os algoritmos com fitness dinâmico visto estarmos comparando parametrizações distintas do campo de força. Além disso, os valores de RMSD mostraram-se melhores que os obtidos pelos AGs estáticos – ainda que os resultados comparando com o AG estático com peso de Van der Waals não sejam estatitisticamente significativos – juntamente com um aumento na diversidade da população, corroboram com a hipótese proposta pelos autores de [Kashtan et al., 2007].

Entretanto deve-se ficar atento aos possíveis artefatos que o aumento ou diminuição da contribuição de um dos termos pode trazer, como as estruturas mais curvadas, discutido na seção anterior. Dentre os trabalhos atuais e futuros, pretende-se testar o fitness dinâmico para outras proteínas e também testar variantes do AG desenvolvidos para problemas dinâmicos, como o AG com Hipermutação e AG com Imigrantes Aleatórios, que ajudam a manter a diversidade da população ao longo das gerações.

6 Agradecimentos

A FAPESP pelo auxílio financeiro. Projeto 2009/12931-3.

7 Referências Bibliográficas

[1] Lehninger, A, Nelson, D. L. & Cox, M. M. (1998) "Principles of Biochemistry with an Extended Discussion of Oxigen – Binding Proteins". New York: Worth Publishers Inc., 2 Ed.

[2] Baxevanis, A. & Oullette, B. (2001) "Bioinformatics – A Practical Guide to the Analysis of Genes and Proteins", Lawrence Erlbaum Associates Publishers.

[3] Anfinsen, C. B. (1973) "Principles that Govern the Folding of Protein Chains", Science, v. 181, p. 223-230.

[4] Mansour, N., Kanj, F. & Khachfe, H. (2010) "Evolutionary algorithm for protein structure prediction", In: Sixth International Conference on Natural Computation (ICNC), v. 8, p. 2347-2351.

[5] Becerra, D., Sandoval, A., Restrepo-Montoya, D. & Nino, L. F. (2010) "A parallel multi-objective ab initio approach for protein structure prediction", In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), p. 137-141.

[6] Mathkour, H. & Ahmad, M. (2010) "An Integrated Approach for Protein Structure Prediction Using Artificial Neural Network", In: Second International Conference on Computer Engineering and Applications (ICCEA), v. 2, p. 484-488

[7] Tragante do Ó, V. & Tinós, R. (2009). "Controle da Diversidade da População em Algoritmos Genéticos Aplicados na Predição de Estruturas de Proteínas". Scientia – Unisinos, vol. 20, n. 2, p. 83-93.

[8] Oliveira, L. L., Ishivatari, L. H. U., Silva, F. L. B. & Tinós, R. (2010) "Genetic Algorithms with a Coarse-Grained Model for Protein Structure Prediction", In: 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology – X-Meeting, X-Meeting Abstract Book, p. 39

[9] Lima, T. W. (2006) "Algoritmos Evolutivos para Predição de Estruturas de Proteínas", Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

[10] Su, S. C., Lin, C. J., Ting, C. K. (2010) "An efficient hybrid of hill-climbing and genetic algorithm for 2D triangular protein structure prediction", In: International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), p. 51-56

[11] Hoque, M. T., Chetty, M., Lewis, A., Sattar, A. (2011) "Twin Removal in Genetic Algorithms for Protein Structure Prediction Using Low-Resolution Model", In: Transactions on Computational Biology and Bioinformatics, v. 8, p. 234-245

[12] Yap, E-H., Fawzi, N. L., Head-Gordon, T. (2008) "A Coarse-Grained α-Carbon Protein Model with Anisotropic Hydrogen-Bonding", Proteins: Structure, Function and Bioinformatics, v. 70, p. 626-638

[13] Gopalakrishnan, K., Sheik, S. S., Ranjani, C. V., Udayakumar, A., Sekar, K. (2007) "Conformational Angles Database (CADB 3.0)", Protein and Peptide Letters, v. 14, p. 665-668

[14] Brooks, B. R., Brooks III, C. L., Mackerel, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., Karplus, M. (2009) "CHARMM: The Biomolecular simulation Program", Journal of Computational Chemistry, v. 30, p. 1545-1615

[15] Kashtan, N., Noor, E. & Alon, U. (2007) "Varying Environments can Speed Up Evolution", Proceedings of the National Academy of Sciences, v. 104, p. 13711-13716

[16] Liang, F & Wong, W. H. (2001) "Evolutionary Monte Carlo for Protein Folding Simulations", Journal of Chemical Physics, v. 115, p. 3374-3380

[17] Chandru, V, Dattasharma, A. D. & Kumar, A. (2003) "The Algorithmics of Folding Proteins on Lattices", Discrete Applied Mathematics, v. 127, p. 145-161

[18] Voet, D. & Voet, J. G. (2003) "Biochemistry", New York: J Wiley

[19] Ishivatari, L. H. U. (2009) "Comparação de Diferentes Contribuições em Campos de Força Minimalistas para o Problema de Determinação de Estruturas de Proteínas Através de Algoritmos Genéticos". Monografia de Conclusão de Curso em Informática Biomédica, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo

[20] Ihaka, R., Gentleman, R. (1996) "R: A Language for Data Analysis and Graphics", Journal of Computational and Graphical Statistics, v. 5, p. 299-314

[21] Herráez, A. (2006) "Biomolecules in the computer: Jmol to the rescue", Biochemistry and Molecular Biology Education, v. 34, p. 255-261