Algorítmo Genético Distribuído Baseado em Conhecimento para Predição de Estruturas de Proteínas

Jonas Bohrer†, Márcio Dorn†
jsbohrer@inf.ufrgs.br, mdorn@inf.ufrgs.br

†Instituto de Informática, Universidade Federal do Rio Grande do Sul, Brasil.

Abstract

O estudo da predição de sua estrutura tridimensional é um dos problemas mais desafiadores em Bioinformática Estrutural. Nos últimos anos, diversas estratégias computacionais foram propostas como soluções para este problema. Conforme revelado pelos experimentos do último CASP, os melhores resultados são obtidos por métodos baseados em conhecimento. Apesar dos avanços no desenvolvimento de métodos computacionais, sistemas e algorítmos para solucionar esse problema complexo, mais pesquisas precisam ser feitas. Neste trabalho, apresentamos uma estratégia computacional para obter informações estruturais de estruturas de proteínas determinadas experimentalmente e um Algorítmo Genético Distribuído baseado em conhecimento para predizer a estrutura tridimensional de proteínas. O método proposto foi testado com oito sequências de resíduos de aminoácidos. Os resultados mostram que as estruturas 3-D preditas são topológicamente comparáveis à seus correspondentes experimentais, assim corroborando a efetividade de nossa proposta.

Introdução

Proteínas têm papéis essenciais em diversos processos fisiológicos em organismos vivos. Uma molécula de proteína é composta por uma cadeia linear ordenada de resíduos de aminoácido que realizam uma variedade de funções ao assumir uma estrutura tridimensional (3-D) particular [7]. A predição da estrutura 3-D de uma proteína é experimentalmente cara e custosa. Nos últimos anos, diversas estratégias computacionais foram propostas como soluções ao problema de Predição de Estruturas de Proteínas (PSP). Métodos computacionais para o problema de predição (PSP) podem ser estudados em quatro classes [4, 3]: (a) métodos de primeiros principios sem informações de database; (b) métodos de primeiros principios com informação de database; (c) métodos de fold recognition; e (d) métodos de modelamento comparativo. Como revelado pelos últimos experimentos CASP, os melhores resultados foram atingidos por métodos baseados em conhecimento (grupo b, c e d). Esses métodos são dependentes de dados experimentais e apresentam dois grandes desafios: (1) definir uma estratégia computacional para identificar e recuperar informações estruturais do *Protein Data Bank* (PDB) [1]; e (2) definir uma estratégia de pesquisa baseada em conhecimento para encontrar estruturas de proteína similares às naturais. Para lidar com estes desafios, desenvolvemos uma estrategia computacional para extração de preferencias conformacionais de proteínas modelo, chamada *Angle Probability List* - APL e um Algorítmo Genético Distribúido baseado em conhecimento para predição da estrutura 3-D de proteínas.

Implementação

Algorítmos genéticos (GA) [5] são algorítmos de pesquisa adaptáveis baseados em ideias evolucionárias. GAs são modelados através do uso de uma população de indivíduos que passam por seleções na presença de operadores que induzem variações, como a mutação e a recombinação (*crossover*). Uma função de avaliação é utilizada para avaliar indivíduos, e o sucesso reprodutivo varia com a adequação do resultado. O sucesso do GA depende principalmente de uma exploração balanceada do espaço de soluções. Quando esse balanço é disproporcional, problemas de convergência prematura podem ocorrer, e o GA vai perder eficiência. Em predição de estruturas de proteínas, a rugosidade da superfície de energia da proteína representa um desafio significante para a técnicas de otimização como os GAs. Uma aproximação para lidar com esse problema considera o uso de Algorítmos Genéticos Distribuídos [3]. A ideia básica de um GA distribuído é manter, em paralelo, populações independentes. Nós propomos um Algorítmo Genético Distribuído baseado na preferência conformacional de resíduos de aminoácido em proteínas com estruturas experimentalmente determinadas. Cada população incorpora a APL derivada de dados experimentais para gerar as populações iniciais e novos indivíduos, aumentando a diversidade do modelo. A Figura 1 mostra a organização geral do método proposto e cada parte do Algorítmo Genético Distribuído é apresentada nos seguintes subtópicos.

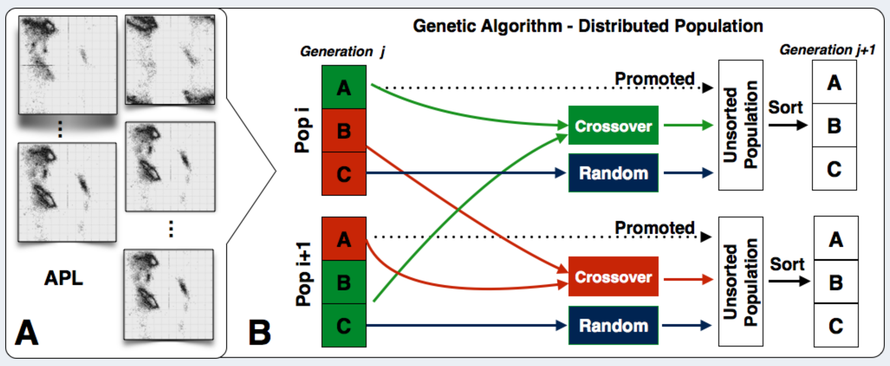

Figura 1: Algorítmo Genético Distribuído Baseado em Conhecimento. (A) APL representando as preferências conformacionais de resíduos de aminoácidos em proteínas. (B) Representação esquemática do Algorítmo Genético distribuído.

Populações: o modelo utilizado é composto de indivíduos que representam possíveis aproximações à estrutura 3-D real. Cada indivíduo armazena um conjunto de ângulos de cadeia principal (ϕ, ψ) e ângulos de cadeia lateral (χ). Essa população é classificada por adequação, e os 10% melhores indivíduos são classificados como Classe A, os 50% seguintes como classe B e os 40% restantes como classe C.

Função de Classificação: uma função de energia que considera todos os atomos foi utilizada para classificar a adequação de cada indivíduo. Empregamos a função de energia do Rosetta implementada por PyRosetta [2], a qual é uma implementação baseada em Python da função Rosetta. Esta função de energia incorpora mais de 20 termos de pêso de energia, estando estes em constante aperfeiçoamento [2].

Crossover: gera uma nova solução, o filho, usando angulos (ϕ, ψ, χ's) aleatórios de dois pais diferentes. Este procedimento é usado para criar novas soluções a cada geração, mas os pais são escolhidos de formas diferentes em duas situações: **(1)** gerações normais: pais são selecionados da classe A (pai 1) e classes B ou C (pai 2). **(2)** gerações de troca: pais são selecionados de um Buffer de Trocas (EB) compartilhado entre dois processos (pai 1), e das classes B ou C (pai 2). A Figura 1 mostra como o crossover funciona na segunda situação. Em ambas as situações, damos maior probabilidade dos ângulos vindos do pai 1 (cerca de 50%-70%), que tem maior adequação, e o restante do pai 2.

Procedimento de troca: dependendo do tempo gasto por cada geração, computamos um número de transações esperadas entre dois processos. No procedimento de troca, cada processo copia seus indivíduos mais adequados (classe A) para o *Buffer* de Troca. Estas soluções são utilizadas como pais no próximos *crossover* em outros processos, e a prole gerada é inserida em sua próxima população. Posteriormente, os indivíduos no *Buffer* são descartados, assim evitando soluções repetidas em soluções diferentes e melhorando a diversidade da população.

Computação da população seguinte: indivíduos da classe A são automaticamente promovidos para a próxima geração. Todas as soluções resultantes do **Crossover** são inseridas na próxima geração. Finalmente, os indivíduos da classe C são descartados, e novos indivíduos são gerados com ângulos escolhidos aleatóriamente da APL. Assim que a população está completa, as soluções são classificadas por seus valores de adequação. No final de cada geração, as melhores soluções estão sempre no topo da população.

Experimentos Computacionais

O método proposto foi usado para predizer a estrutura tridimensional de oito sequências de proteínas do PDB: 1L2Y (Fig. 2a), 1WQC (Fig. 2b), 2F4K (2c), 2MR9 (Fig. 2d), 2MTW (Fig. 2e), 3P7K (Fig. 2f), 1K43 (Fig. 2g) e 1ACW (Fig. 2h). Estes casos de estudo foram selecionados de forma a testar o método proposto com diferentes padrões de enovelamento [8]. Primeiramente, analisamos a contribuição do uso da APL. Para seis casos de estudo (Tabela 1, coluna 1) executamos o algorítmo de predição 16 vezes por 12 horas para cada sequência de proteína com e sem a APL. Os testes foram executados em ambiente Linux de uma SuperWorkstation com Intel Xeon CPU E5-2650 2.00GHz com 20MB de Cache e 32GB de memória RAM. Tabela 1 mostra a média e o desvio padrão para a melhor solução encontrada (menor energia) para cada sequência e o correspondente RMSD. Como pode ser observado, o algorítmo usando a APL atinge soluções com melhor RMSD e energia. Assim, selecionar ângulos da APL guia o algorítmo à soluções mais corretas. De forma a avaliar nosso método, executamos o Algorítmo Genético Distribuído proposto 16 vezes por 12 horas para outras duas proteínas (2F4K, 1ACW) usando a APL (Tabela 2).

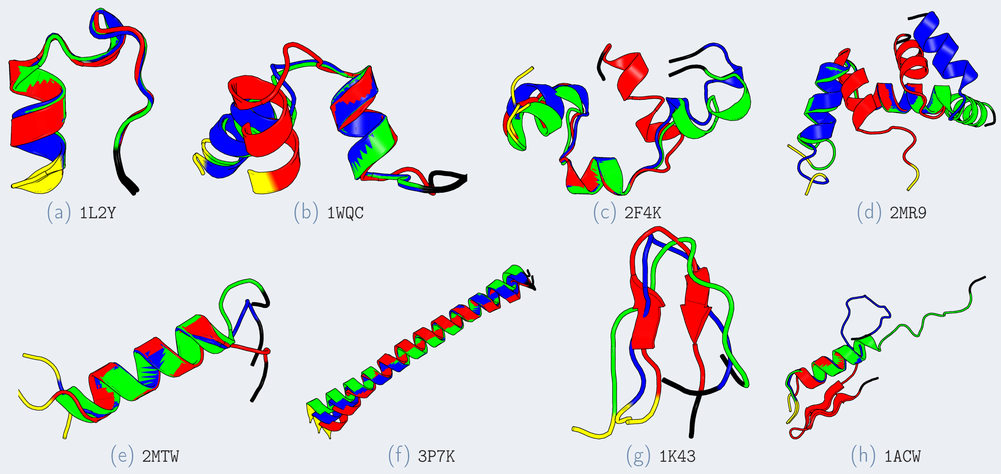

Figura 2: Representação das estruturas 3-D experimental (vermelha), menor RMSD (azul) e menor energia (verde). A C_α das estruturas experimentais e preditas estão alinhadas. Cadeias laterais não são mostradas por razões de claridade. Representação gráfica preparada utilizando PyMOL.

Proteína ID	Energia Média com APL	Rmsd Médio com APL	Energia Média sem APL	Rmsd Médio sem APL
1L2Y	-13.49 (±2.08)	4.26 (±1.83)	193.03 (±34.74)	5.61 (±1.08)
1WQC	-14.16 (±2.82)	5.44 (±1.05)	122.47 (±32.80)	7.70 (±1.05)
2MR9	-43.59 (±2.94)	10.36 (±1.67)	225.36 (±35.40)	14.55 (±2.23)
2MTW	-19.57 (±1.33)	2.16 (±0.44)	36.83 (±1.28)	5.61 (±0.78)
3P7K	-59.84 (±1.22)	2.01 (±0.36)	205.23 (±9.10)	11.11 (±2.05)
1K43	-4.91 (±1.19)	3.00 (±0.51)	27.18 (±1.78)	4.61 (±1.14)

Tabela 1: Valores de média de energia e Rmsd encontrados pelo algorítmo quando utilizando a APL (colunas 2-3) ou não (colunas 4-5). Valores de Rmsd são expressados em Å. Valores de energia são representados em $\boldsymbol{Kcal/mol^{-1}}$.

PDB ID	Menor Energia		Menor Rmsd		Energia Média(Std.)	Rmsd Média(Std.)	Média Gerações
	E	Rmsd	Rmsd	E			
1L2Y	-19.96	0.56	0.56	-19.96	-13.49 (±2.08)	4.26 (±1.83)	36265
1WQC	-20.51	3.34	2.98	-14.77	-14.16 (±2.82)	5.44 (±1.05)	25330
2F4K	-28.00	5.03	4.63	-25.25	-21.89 (±2.67)	7.76 (±1.72)	19147
2MR9	-48.47	9.25	6.78	-45.31	-43.59 (±2.94)	10.36 (±1.67)	13402
2MTW	-21.73	2.53	1.45	-20.47	-19.57 (±1.33)	2.16 (±0.44)	27684
3P7K	-61.30	1.79	1.22	-60.86	-59.84 (±1.22)	2.01 (±0.36)	14939
1K43	-7.09	3.36	1.53	-2.62	-4.91 (±1.19)	3.00 (±0.51)	35032
1ACW	-10.77	11.10	7.52	-5.04	-7.59 (±1.86)	11.01 (±1.45)	19219

Tabela 2: Resultado da simulação do algorítmo. Colunas 2 e 3 mostram, respectivamente, a menor energia potencial (*Kcal/mol*) e a energia média das 16 execuções do algorítmo proposto. Colunas 6 e 7 mostram, respectivamente, o RMSD e valores de Energia para a estrutura com menor energia e estruturas com o menor RMSD encontrado na execução de nosso método.

Conclusão

Como corroborado pelos resultados do CASP através dos últimos anos [6], há uma crescente necessidade de novas estratégias para extraír, representar e manipular dados estruturais de estruturas de proteínas 3-D experimentalmente determinadas, assim é necessário o desenvolvimento de estratégias computacionais que façam uso dessas informações de forma a predizer a corresponde estrutura 3-D de proteínas.

Neste trabalho, propomos um nova estratégia de pesquisa baseada em conhecimento para o problema de predição de estruturas de proteínas (*Protein Structure Prediction*). A estratégia de pesquisa é baseada em um Algorítmo Genético Distribuído com populações estruturadas. Como demonstrado pelos experimentos, o método desenvolvido pode produzir predições precisas, onde as estruturas de proteínas 3-D são comparáveis às duas correspondentes experimentais. Quando comparado com outros métodos de predição de primeiros principios que usam informações de database, nossa aproximação apresenta vantagens em termos de tempo exigido para produzir estruturas de proteínas 3-D similares às experimentais.

Referências

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bath, H. Weissig, I.N. Shindyalov, and P.E. Bourne.
The protein data bank.
Nucleic Acids Res., 28(1):235–242, 2000.

[2] S. Chaudhury, S. Lyskov, and J. J. Gray.
Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta.
Bioinformatics, 26(5):689–691, 2010.

[3] M. Dorn, M. Barbachan e Silva, L. S. Buriol, and L. C. Lamb.
Three-dimensional protein structure prediction: Methods and computational strategies.
Comp. Biol. and Chem., 53, Part B:251 – 276, 2014.

[4] C.A. Floudas, H.K. Fung, S.R. McAllister, M. Moennigmann, and R. Rajgaria.
Advances in protein structure prediction and de novo protein design: A review.
Chem. Eng. Sci., 61(3):966–988, 2006.

[5] D.E. Goldberg.
Kluwer Academic Publishers, Boston, 1 edition, 1989.

[6] A. Kryshtafovych, K. Fidelis, and J. Moult.
Casp10 results compared to those of previous casp experiments.
Proteins: Structure, Function, and Bioinformatics, 82:164–174, 2014.

[7] A.L. Lehninger, D.L. Nelson, and M.M. Cox.
Principles of Biochemistry.
W.H. Freeman, New York, USA, 4 edition, 2005.

[8] A. Liljas, L. Liljas, J. Pskur, P. Lindblom, G. amd Nissen, and M. Kjeldgaard.
World Scientific Printers, Singapore, 2001.

Agradecimentos: - Bruno Borguesan