Identificação de padrões conformacionais em estruturas experimentais e projeto de meta-heurísticas para a predição da estrutura tridimensional de proteínas.

Mariana Oliveira[‡] and Márcio Dorn[†]

[†]Instituto de Informática, Universidade Federal do Rio Grande do Sul, Brasil.
[‡]Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Brasil.

Introdução

Proteínas são longas sequências formadas por aminoácidos, os quais adotam uma estrutura tridimensional (3D) única quando em meio fisiológico [1]. Predizer a estrutura 3D de uma proteína a partir de sua sequencia de aminoácidos ainda é um problema em aberto para a Bioinformática Estrutural [2]. Sua determinação por métodos experimentais, como as técnicas de cristalografia por raio-X e RMN, é de alto custo e podem levar muito tempo. Assim, o número de proteínas com estrutura 3D conhecida é muito inferior ao número de sequencias conhecidas, sendo, então, fundamental o desenvolvimento e a utilização de métodos computacionais para a predição de estruturas 3D de proteínas (3D-PSP).

Atualmente são encontrados diversos algoritmos e métodos propostos a fim de resolver os problemas 3D-PSP, os quais podem ser classificados em métodos de de primeiros princípios (*ab initio* e predição de novo) e métodos baseados em conhecimentos (alinhavamento e modelagem comparativa) [3,4]. Os últimos experimentos do CASP (*Critical Assesment of Strutucture Prediction* mostraram melhores resultados para os métodos baseados em conhecimentos, os quais utilizam dados experimentais [5]. Frequentemente, são utilizadas meta-heurísticas no processo de busca de possíveis soluções conformacionais.

Meta-heurística é uma classe de algoritmos utilizada para solucionar problemas de otimização, buscando encontrar soluções satisfatórias para problemas em que a solução exata muitas vezes é inviável [6].

Proteínas podem ser representadas computacionalmente por dois ângulos de torção (phi ϕ e psi ψ) por aminoácido [1]. Regiões de estruturas secundárias flexíveis, como alças e voltas, podem ter diferentes combinações de valores de ângulos, ou seja, não possuem padrões regulares. Desta maneira, encontrar valores exatos torna-se custoso devido as inúmeras possibilidades. A utilização de padrões conformacionais observados experimentalmente pode diminuir o espaço de busca para essas regiões flexíveis.

Métodos e Materiais

Banco de Dados: a fim de garantir melhores resultados na predição das estruturas 3D, as estruturas do *Protein Data Bank* [7] foram selecionadas seguindo critérios de qualidade. Todas as estruturas foram determinadas experimentalmente através da técnica de cristalografia com uma resolução $\leq$ 2.5 $\mathring{A}$. Foram removidas as estruturas com R-observado 0.2 e selecionadas apenas aquelas com identidade de pelo menos 30 %. Para os átomos da cadeia principal foram selecionados apenas resíduos com B-factor $\leq$ 30 $\mathring{A}$ e ocupância igual a 1. A estrutura secundária de cada resíduo foi atribuida através do programa STRIDE [8], sendo considerado apenas 6 dos possíveis estados: H (α-helix), G ($\mathbf{3_10}$-helix), E (β-sheet), B (β-bridge), T (turn), C (coil). Para cada estrutura foi gerado um arquivo contendo sua estrutura secundária e informação conformacional (ângulos de torção).

Foi feita a fragmentação de estrutura secundária a fim de encontrar padrões mais específicos, utilizando como critério a existência de estruturas flexíveis (alças e voltas) unindo estruturas rígidas (folhas e hélices). Dentre todos os padrões presentes, foi selecionado um total de 14 padrões, os quais apresentaram maiores ocorrências (**GCG, HCCCCCH, HTTTH, HCCCCH, HTTTE, ECCCH, HCCCH, ECCCE, HCCH, HCCE, ETTTTE, HCCCE, ETTTE, ETTE**).

Meta-Heurística: foi utilizado a meta-heurística baseada em população *Differential Evolution*[9], a qual parte de um conjunto de soluções iniciais e, a cada iteração, melhora a fim de encontrar o melhor grupo de possíveis soluções. Para evitar o alto custo computacional, foi implementado um algoritmo adaptativo *Self-Adapting Differential Evolution*[10], o qual utiliza diferentes parâmetros e estratégias ao longo do processo de busca/evolução, sendo escolhidos aqueles de resultados promissores .

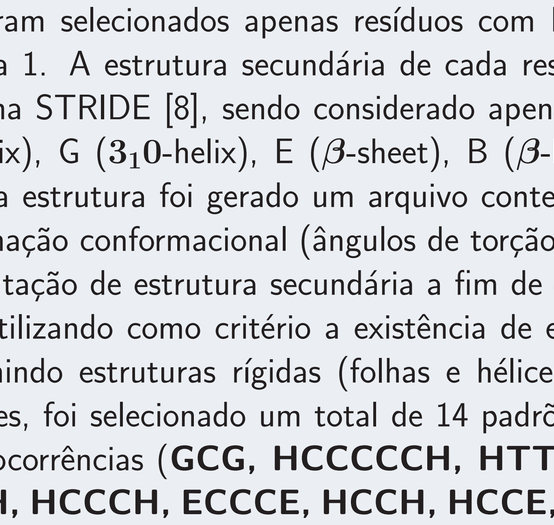
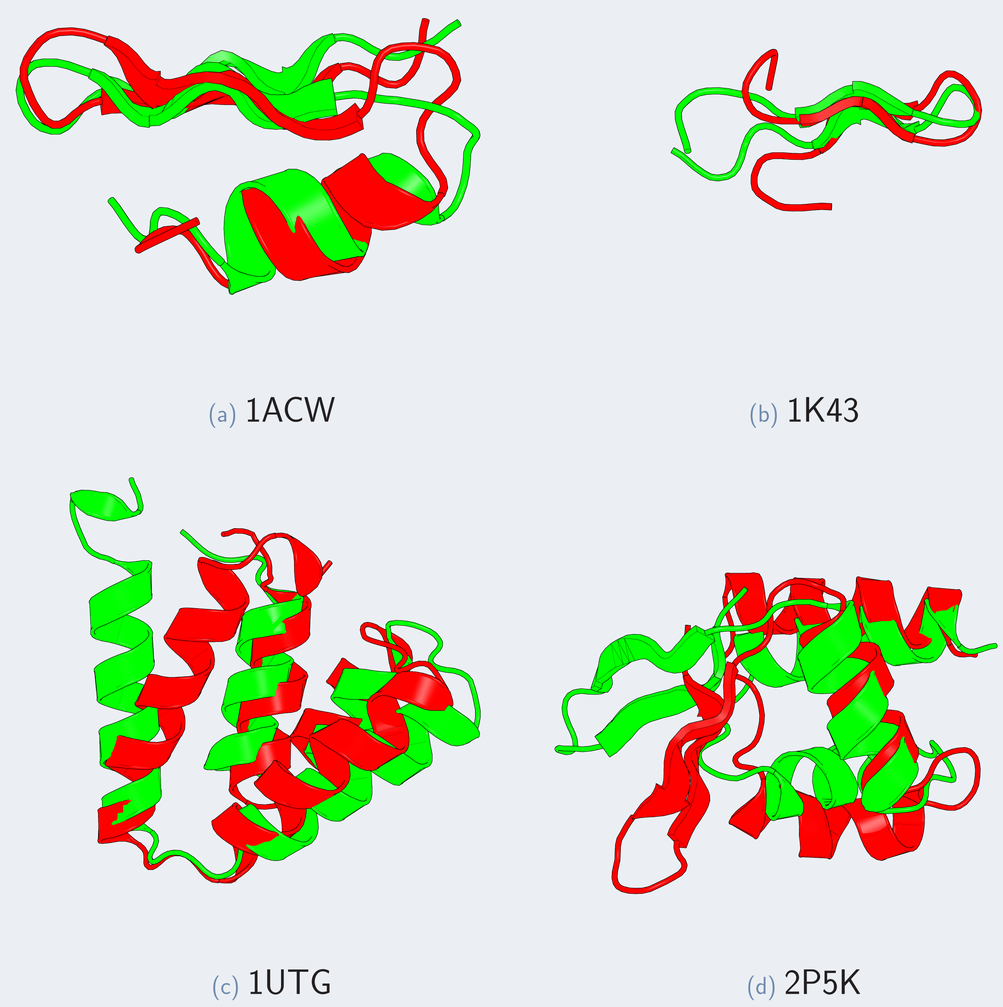


Figura 1: Fluxograma do SADE

A população inicial é formada por um conjunto de vetores contendo a informação conformacional de cada estrutura utilizada. Para as regiões de voltas foram utilizadas as informações do banco de dados gerado, enquanto que para as demais foram utilizadas as informações da APL-1 [11] (Figura 1).

Foi realizado um total de 10 corridas para cada proteína (1ACW, 1K43, 1UTG, 2P5K) utilizando o algoritmo SADE [10] com a biblioteca de voltas e não utilizando a biblioteca de voltas. Para analisar os dados, calculou-se o RMSD e o GDT de cada estrutura gerada em relação a estrutura nativa provinda do PDB.

Resultados e Discussão



(a) 1ACW (b) 1K43 (c) 1UTG (d) 2P5K

Figura 2: Sobreposição da estrutura gerada (vermelho) com estrutura experimental (verde) provinda do PDB.

Os valores de RMSD e GDT das estruturas geradas em relação as estruturas nativas quando utilizando a biblioteca de padrões de voltas (colunas 2 e 4) e quando utilizando apenas a APL-1 para montagem dos indivíduos estão resumidos na Tabela 1. A utilização do banco de dados monstrou melhores valores para ambos os parâmetros para 3 das 4 estruturas testadas. Para a estrutura 1K43 os valores de GDT foram melhores quando utilizando apenas a APL-1, entretanto essa diferença foi pequena.

PDB ID	RMSD±std	RMSD±std	GDT±std	GDT±std
1ACW	3.37 ± 0.70	4.23 ± 2.01	64.91 ± 4.72	62.41 ± 7.11
1K43	1.18 ± 0.18	1.19 ± 0.24	71.61 ± 2.45*	74.64 ± 2.86*
1UTG	6.67 ± 1.62	7.81 ± 1.71	46.14 ± 7.45	45.93 ± 8.38
2P5K	8.24 ± 2.23	9.69 ± 1.67	36.39 ± 4.46	34.40 ± 2.53

Tabela 1: **Valores médios e desvio padrão de RMSD e GDT.** Colunas 2 e 4: valores utilizando o banco de padrões de voltas. Colunas 3 e 5: valores utilizando apenas APL-1.

Conclusion

Neste trabalho, foi proposta uma meta-heurística que utiliza o banco de dados com informações estruturais para regiões de voltas aplicada ao problema de predição de estruturas tridimensionais. Os resultados para as estruturas proteícas testadas monstraram-se favoráveis ao uso do banco de dados de padrões de voltas gerado. Para trabalhos futuros, serão realizados mais testes com as estruturas já utilizadas, além disso, serão testadas outras estruturas já trabalhadas pelo grupo.

References

1 Lehninger, A., Nelson, D., Cox, M., *Principles of Biochemistry.*, W.H. Freeman, New York, USA. 4 ed., 2005.
2 Zhang, Q., Veretnik, S., Bourne, P.E., *Overview of Structural Bioinformatics*, Springer, Heidelberg. 2005.
3 Floudas, C., Fung, H., McAllister, S., Moennigmann, M., Rajgaria, R., *Advances in protein structure prediction and de novo protein design: A review.*, Chem. Eng. Sci. **61** (2006), 966-988.
4 Dorn, M., Barbachan e Silva, M., Buriol, L.S., Lamb, L.C., *Three-dimensional protein structure prediction: Methods and computational strategies.*, Comput. Biol. Chem. **53** (2014b), Part B, 251 – 276.
5 Kryshtafovych, A., Fidelis, K., Moult, J., *CASP10 results compared to those of previous CASP experiments.*, Proteins: Struct., Funct., Bioinf. **82** (2014a), 164–174.
6 Siarry, P., Boussaid, I., Lepagnot, J. *A survey on optimization metaheuristics.*, Information Sciences. **237** (2013), 82-117.
7 **www.rcsb.org** H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. *The Protein Data Bank*, Nucleic Acids Research. **28** (2000), 235-242.
8 Heinig, M., Frishman, D. *STRIDE: a Web server for secondary structure assignment from known atomic coordinates of proteins.*, Nucleic Acids Research. **32** (2004), W500-2.
9 R. Storn and K. Price. *Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces.*, Technical Report TR-95-012, International Computer Science Institute, Berkeley, CA, Mar. 1995.
10 A. K. Qin, V. L. Huang, P. N. Suganthan, *Differential Evolution Algorithm With Strategy Adaptation for Global Numerical Optimization.*, IEEE Transactions on Evolutionary Computation. **13** (2009), no. 2, 398-417.
11 M. Dorn, B. Borguesan, M. Inostroza-Ponta, M. Barbachan e Silva, B.Grisci, *APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction.*, Computational Biology and Chemistry. **59** Part A. (2005), 142–157.

Agradecimentos:

FAPERGS CNPq pro-pesq Pró-Reitoria de Pesquisa CAPES