

Análises iniciais para otimização da técnica GARN

Jhonatan Silva¹, Daniel Cordeiro¹, Johanne Cohen²

¹Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)

²LRI, Université Paris-Sud, CNRS, Université Paris-Saclay, France

{jhonatansilva, daniel.cordeiro}@usp.br, johanne.cohen@lri.fr

Abstract. *GARN is a technique to predict RNA 3D structures, and uses game theory concepts in its algorithm. The study authors suggest an optimisation of the technique, performing new analyses. The purpose of this article is to show the initial steps developed for this optimisation. Analyses were performed on 3D structures obtained in the laboratory, and applied to the game settings. This article presents some preliminary results resulting from these changes.*

Resumo. *GARN é uma técnica desenvolvida para prever estruturas 3D de RNAs, e utiliza conceitos de teoria dos jogos em seu algoritmo. Os autores do estudo sugerem uma otimização da técnica, realizando novas análises. O objetivo desse artigo é mostrar os passos iniciais desenvolvidos para essa otimização. Foram realizadas análises sobre estruturas 3D obtidas em laboratório, e aplicadas nas configurações do jogo. Estão presentes nesse artigo alguns resultados preliminares decorrentes dessas alterações.*

1. Introdução

Ácidos Ribonucleicos (RNA) são moléculas presentes em seres vivos, responsáveis por funções como a síntese de proteínas através da tradução do código genético do indivíduo, catalisação de reações bioquímicas e desempenho de papéis reguladores no indivíduo. Estruturas 3D de RNAs são importantes, por exemplo, na produção de certas drogas e podem proporcionar a descoberta de mutações causadoras de doenças [Dawson et al. 2016].

Para obter tais estruturas com a menor probabilidade de erros, são usadas técnicas experimentais em laboratório que, no entanto, são demoradas e caras [Shapiro et al. 2007]. Técnicas computacionais começaram a ser desenvolvidas na tentativa de prever tais estruturas, porém há limitações, tais como tempo de execução e/ou tamanho da molécula a ser modelada.

Um estudo recente [Boudard et al. 2017] demonstrou suprir essas limitações utilizando conceitos de teoria dos jogos, desenvolvendo uma técnica denominada GARN (*Game Algorithms for RNA 3D sampling*). Teoria dos jogos é uma área da matemática cujo objetivo é analisar interações entre dois ou mais elementos dentro de um conjunto, onde as ações de um elemento devem ser tomadas sempre verificando as possíveis ações dos outros [Figueiredo 1994].

A técnica, no entanto, deixou uma lacuna relacionada à predição de moléculas pequenas, que divergem bastante quando comparadas a estruturas homologadas decorrentes de técnicas experimentais. Uma possível solução citada pelos autores é alterar as configurações do jogo para obter estruturas mais refinadas, isto é, visualmente mais

próximas de estruturas obtidas por técnicas experimentais. Esse artigo tem por objetivo mostrar os passos iniciais realizados para otimizar as configurações do jogo.

2. Materiais e Métodos

Na técnica GARN, agrupamentos de unidades que compõem RNAs chamadas nucleotídeos são representadas por jogadores, e o jogo consiste em dispôr todos os nucleotídeos no espaço 3D, dado um conjunto de 12 estratégias possíveis, que são intervalos de ângulos de 0 a 360 graus. A técnica usa o conceito de Equilíbrio de Nash, onde, após várias jogadas, não é possível obter maior ganho (ou menor perda). Para atingir o Equilíbrio de Nash, a técnica usa o algoritmo EXP3 [Auer et al. 2002] e aplica diferentes pontuações para cada estratégia escolhida, baseando-se no tipo de jogador. As fórmulas usadas para aplicar as pontuações são:

- **Lennard Jones** [Lennard-Jones 1931]: $-A_{i,j} \cdot \left(\left(\frac{B_{i,j}}{d} \right)^{12} - 2 \cdot \left(\frac{B_{i,j}}{d} \right)^6 \right)$, onde $A_{i,j}$ and $B_{i,j}$ são constantes definidas, dependentes do tipo de jogador. Esse tipo de pontuação tem apenas um modo, que é a melhor distância entre dois jogadores.
- **Lennard Jones Modificado**: Segue com a mesma fórmula anterior, porém qualquer resultado negativo é transformado em zero.
- **Distribuição normal**: $A_{i,j} \cdot \frac{1}{\sigma_{i,j} \sqrt{2\pi}} e^{-\frac{(d-\mu_{i,j})^2}{2\sigma_{i,j}^2}}$, onde $-A_{i,j}$, $\sigma_{i,j}$ e $\mu_{i,j}$ são constantes definidas, dependentes do tipo de jogador. Esse tipo de pontuação possui vários modos, que é um conjunto com as melhores distâncias entre dois jogadores.

Para otimizar a técnica GARN, inicialmente foram realizadas análises sobre distância e ângulo dos átomos em várias moléculas obtidas em laboratório. As estruturas utilizadas estão presentes nas bases de dados RCSB PDB [Berman et al. 2000] e RNA FRABASE [Popenda et al. 2010]. Os resultados dessas análises foram usados para atualizar as constantes definidas nos algoritmos referentes à distância entre dois jogadores, e as constantes para os 3 tipos de pontuação citados anteriormente. Para gerar as estruturas com a técnica, foi usada uma máquina com processador 2,2 GHz Intel Core i7 6-Core e memória de 16GB.

Para analisar o quanto a estrutura gerada pela técnica diverge de uma estrutura obtida em laboratório, é utilizado o RMSD (*Root Mean Square Deviation*) [Kufareva and Abagyan 2011], definido por $\sqrt{\frac{1}{p} \sum_{i=1}^p \|m_i - n_i\|^2}$, onde m é a estrutura obtida em laboratório e n é a estrutura resultante da técnica, p é a quantidade de jogadores agrupados nessa estrutura, e m_i e n_i referem-se a posição do jogador i nas estruturas m e n .

3. Resultados Preliminares

As análises realizadas até o momento visam otimizar um dos padrões estruturais mais encontrados nas estruturas de RNA, que são as hélices [Leontis and Westhof 2001]. Essas hélices, por sua vez, são formadas por vários elementos chamados pares de bases. Na Figura 1 pode ser visto um resultado usado, calculando a distância entre pares de bases dentro da mesma hélice e pares de bases de outras hélices. Ao lado esquerdo, na mesma figura, há um grafo 3D representado os baricentros de uma estrutura de RNA, onde as sequências em branco são os pares de bases das hélices, e os elementos em azul

são elementos não-pareados. Essa análise busca atualizar as constantes definidas para as fórmulas de pontuação. Similar a esta análise, na Figura 2 também é calculada a distância entre os pares de bases, mas apenas entre os elementos dentro da mesma hélice.

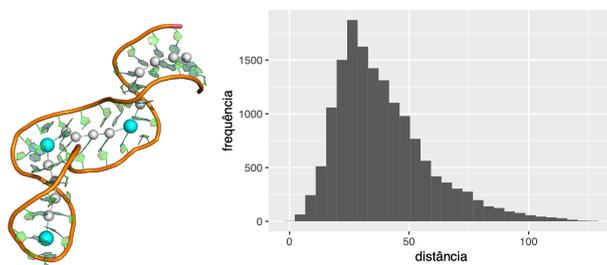


Figura 1. Distância entre pares de bases dentro da mesma hélice e os pares de bases de outras hélices, na mesma estrutura. Em branco, no grafo à esquerda, há um exemplo de pares de bases.

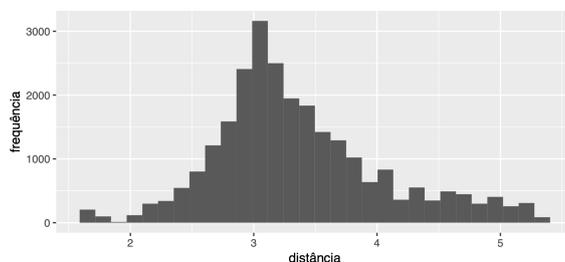


Figura 2. Distância entre pares de bases, apenas dentro da mesma hélice

Com o resultado dessas análises sobre as moléculas, as constantes definidas na técnica GARN para as 3 fórmulas de pontuação foram atualizadas, conforme os valores da Figura 1. Adicionalmente, o algoritmo da técnica foi atualizado para que sejam adicionadas mais vértices ao grafo 3D gerado pela técnica e, conseqüentemente, a distância entre os jogadores das hélices também foi atualizado.

Após essas atualizações, para cada uma das moléculas listadas na Tabela 1, foram geradas 5 estruturas 3D usando os diferentes tipos de pontuação. Nessa tabela, é possível verificar o RMSD máximo e mínimo, de acordo com a molécula e o tipo de pontuação, tal como o tempo de execução.

4. Conclusão

Nas moléculas usadas para esse teste, pode-se verificar que as melhores pontuações foram obtidas usando a fórmula de Lennard-Jones, original e modificada. No entanto, com uma quantidade maior de moléculas, o resultado pode ser diferente.

Em trabalhos futuros, é interessante que haja uma quantidade maior de moléculas testadas, para que seja possível buscar padrões nos resultados, considerando características específicas das moléculas e, assim, atualizar o algoritmo com o melhor tipo de pontuação para cada tipo de molécula. Posteriormente, é necessário comparar os resultados com outras técnicas na literatura.

Este resultado abre caminho para o desenvolvimento de um algoritmo paralelo e distribuído mais eficiente e que resulte em predições mais precisas.

Tabela 1. Comparação dos resultados obtidos usando Lennard-Jones (LJ), Lennard-Jones Modificado (LJM) e Distribuição Normal (DN)

Molécula	Nucleotídeos	Ponto	RMSD			Tempo de execução		
			LJ	LJM	DN	LJ	LJM	DN
17RA	21	max	3.9	3.8	3.61	1s	1s	1s
		min	2.26	2.2	2.91			
3SUH	101	max	17.34	16.21	18.64	13s	13s	10s
		min	13.31	10.66	11.99			
3D0U	161	max	26.45	29.37	27.3	1m 43s	1m 49s	1m 23s
		min	20.12	21.02	20.55			
4C4Q	233	max	27.15	31.65	33.73	6m 14s	6m 45s	5m 9s
		min	18.6	23.74	23.04			
2A2E	300	max	30.49	34.17	26.5	5m 30s	5m 53s	4m 31s
		min	20.52	22.62	23.1			

Referências

- Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.
- Boudard, M., Barth, D., Bernauer, J., Denise, A., and Cohen, J. (2017). GARN2: coarse-grained prediction of 3D structure of large RNA molecules by regret minimization. *Bioinformatics*, 33(16):2479–2486.
- Dawson, W. K., Maciejczyk, M., Jankowska, E. J., and Bujnicki, J. M. (2016). Coarse-grained modeling of RNA 3D structure. *Methods*, 103:138–156.
- Figueiredo, R. S. (1994). Teoria dos jogos: conceitos, formalização matemática e aplicação à distribuição de custo conjunto. *Gestão & Produção*, 1(3):273–289.
- Kufareva, I. and Abagyan, R. (2011). Methods of protein structure comparison. In *HOMOLOGY Modeling*, pages 231–257. Springer.
- Lennard-Jones, J. E. (1931). Cohesion. *Proceedings of the Physical Society*, 43(5):461–482.
- Leontis, N. B. and Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499–512.
- Popenda, M., Szachniuk, M., Blazewicz, M., Wasik, S., Burke, E. K., Blazewicz, J., and Adamiak, R. W. (2010). RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC bioinformatics*, 11(1):231.
- Shapiro, B. A., Yingling, Y. G., Kasprzak, W., and Bindewald, E. (2007). Bridging the gap in RNA structure prediction. *Current Opinion in Structural Biology*, 17(2):157–165.