Avaliação de Desempenho em Bioinformática: Estudo de caso de sistemas computacionais para a investigação de microRNAs

Rosana R. Aguiar[1], Leandro A. Ambrosio[1], Gonzalo Sepúlveda-Hermosilla[2], Vinicius Maracaja-Coutinho[2,3,4], Alexandre Rossi Paschoal[1,*]

* paschoal@utfpr.edu.br

[1]Federal University of Technology - Paraná, Brazil.

[2] Centro de Genómica y Bioinformática, Universidad Mayor, Santiago, Chile.

[3] Beagle Bioinformatics, Santiago, Chile.

[4] Instituto Vandique, João Pessoa, Brazil.

Abstract. *This paper describe in details the computational evaluation of the miRQuest (http://mirquest.integrativebioinformatics.me/) system which is a webserver and middleware architecture integrating different standardized softwares for microRNA (miRNA) investigation.*

Resumo. *Este artigo apresenta em detalhes a avaliação do sistema web miRQuest (http://mirquest.integrativebioinformatics.me/) que foi implementado e fez a integração dos principais preditores de microRNAs (miRNAs) em uma arquitetura de middleware, aplicado em um estudo com de caso em bioinformática.*

1. Introdução

A bioinformática é uma área interdisciplinar a fim de auxiliar as pesquisas para analisar e interpretar a informação de valor contida na massa de dados biológicos que vem sendo gerados [ARAÚJO et al., 2005; PADILHA et al., 2008]. Na bioinformática tem-se o estudo dos RNAs não-codificadores, do inglês *non-coding RNAs* (ncRNAs), que são sequências de ácidos ribonucleicos transcritas, e que diferentemente dos mais conhecidos RNAs mensageiros (mRNA), não são traduzidos em proteínas, mas ainda assim possuem funções importantíssimas para a regulação e manutenção dos processos biológicos celulares [MACHADO-LIMA et al., 2008; OLIVEIRA et al., 2011].

O miRNA ou microRNA é uma classe de ncRNA que desperta maior interesse de pesquisa pela comunidade científica. miRNAs são pequenos RNAs, contendo cerca de 22 nucleotídeos (nt), responsáveis pelo controle pós-transcricionais dos níveis de RNA mensageiro nas células via o pareamento estrutural complementar entre as duas moléculas de RNAs: miRNA-mRNA [ZHANG et al., 2007]. Este interesse se deve pelo papel de controle e regulação das redes gênicas dentro das células, estando assim envolvidas em diversos processos biológicos e doenças como câncer ou em planta, por exemplo [BARROS-CARVALHO et al. 2014; SEVERINO et al., 2013].Outro exemplo é que de acordo com o banco de dados NRDR (*The Non-coding RNA Databases Resource*) [PASCHOAL *et al.*, 2012], 51% (70 de 137, versão 2.0) dos bancos indexados que são referentes as moléculas de RNAs são específicos apenas para a classe de miRNAs.

Atualmente, se sabe que existem diversos programas para identificação de miRNA, que foram desenvolvidos utilizando as mais diferentes linguagens e métodos computacionais como: o agrupamento (*cluster*) ou aprendizagem de máquina [MACHADO-LIMA et al., 2008]. Entretanto, entender o funcionamento de cada um deles e aplicá-los torna-se uma tarefa complexa. Este artigo apresenta a avaliação de desempenho do middleware miRQuest um Web Server que integra quatro dos principais preditores para investigação de miRNAs.

2. ncRNAs e miRNAs

A partir dos resultados da análise de projetos de transcriptomas, se descobriu que grande parte do que era transcrito nos genomas (p.ex. camundongo e humano) eram regiões não traduzidas, ou seja, que não geravam proteínas como um produto final. Evidências apontam que o aumento na complexidade dos organismos vivos, observados ao longo do processo evolutivo, parece estar diretamente associado à quantidade e variedade de ncRNAs, indicando que quanto mais próximo dos eucariotos superiores, maior a quantidade de ncRNAs [TAFT et al., 2007]. Em 1993, Lee e colaboradores (1993) caracterizaram e descreveram uma classe específica de ncRNA definida como miRNA ou miRNA, responsáveis por regular negativamente os níveis de expressão mRNAs nas células. O banco de dados miRBase [KOZOMARA e GRIFFITHS-JONES, 2014] sendo o estado da arte em informações referente a miRNAs, é um repositório que descreve essas moléculas como grandes responsáveis por diversos processos biológicos e doenças, resultando na classe de ncRNA mais estudada [PASCHOAL et al., 2012].

3. Trabalhos relacionados

Em 2008, Machado-Lima e colaboradores (2008) descreveram uma extensa lista de abordagens e técnicas computacionais para pesquisa de ncRNAs e que possuem uma relação direta com os miRNA, visto que o entendimento a respeito da modelagem computacional associada à biologia de RNAs é a base para os algoritmos de identificação deste tipo molécula. Já em 2012, Allmer e Yousef (2012) comentaram sobre algumas ferramentas e abordagens para predição de miRNAs, discutindo também os aspectos e padrões utilizados para predição de miRNAs (p. ex. estrutura, sequência, tamanho). Em linha similar, Gomes e colaboradores (2013) discorrem em uma revisão sobre ferramentas computacionais para identificação miRNAs quinze abordagens e técnicas de diferentes programas de predição. Apesar do grande número de revisões bibliográficas publicadas, verificou-se a falta e necessidade de um sistema integrado de modo a facilitar o uso de preditores para identificação de miRNAs, em especial para profissionais com poucos conhecimentos de ambientes computacionais não amigáveis, como o caso dos biólogos.

4. Materiais e metodologias

4.1. Busca dos preditores de miRNA

Foi realizada nos mecanismos de busca PubMed, ACM, IEEE e Google Scholar uma investigação por programas de identificação de miRNA. Os resultados foram analisados e os seguintes preditores escolhidos: Triplet-SVM [XUE et al., 2005]; MiPred [JIANG et al., 2007]; HHMMiR [KADRI et al., 2009] e NovoMIR [TEUNE e STEGE, 2010].

4.2. Conjunto de dados e experimentos para avaliação de desempenho

Foram usados 1872 sequencias miRNAs de *Homo sapiens* e 298 de *Arabidopsis thaliana* do miRBase versão 19. Na avaliação de desempenho, primeiramente foi

analisado o tempo de execução. Dados com sequências de distintos tamanhos foram usados. Para humano usou-se 10, 50, 100, 500 e 1.000 sequências e planta de 10, 30, 50, 100 e 200 sequências. O segundo experimento foi de performance. As sequências de ambos genomas foram usadas como conjunto de dados positivo (os dados de treinamento foram retirados). Já o conjunto de dados negativo foi extraído de um estudo publicado por Janssen e colaboradores (2008), que contém sequências de proteínas e um grupo de sequências randômicas aleatórias que sabidamente não correspondem miRNAs. Para medir o desempenho usou-se as seguintes métricas: sensibilidade; sensibilidade; precisão; acurácia e F1-Score (Powers, 2011).

4.3. Desenvolvimento do miRQuest

O miRQuest foi desenvolvido como aplicação web na linguagem Java com o emprego do conceito em camadas utilizando o Tomcat versão 8 como Web Server. O XML é usado como formato padrão para comunicação entre as etapas de processamento do middleware e Shell Scripts, em linguagem PERL, foram desenvolvidos para execução de etapas específicas como RNAfold [BRAMEIER e WIUF, 2007] e RNAshapes [STEFFEN et al., 2006]. RNAfold e RNAshapes são programas de predição de estrutura secundária, uma etapa usada por alguns preditores para identificação de características referentes à estrutura dos miRNAs identificados.

O miRQuest foi também implementado usando *Contexts and Dependency Injection* (CDI) para integração das camadas, Apache Shiro Framework para administração de seção e *Commons Email API* para envio de email. Cada ferramenta de predição foi instalada conforme as regras descritas em suas documentações. E ainda, para a execução das ferramentas foram aplicados os parâmetros e padrões conforme recomendado no manual de uso de cada ferramenta.

5. Resultados e discussão da avaliação do miRQuest

As tabelas 1 e 2 apresentam os resultados para humano e a tabela 3 para planta. Analisando-se os dados destas tabelas observou-se a diferença no tempo de execução quando comparamos a execução *stand-alone* com a execução via web.

Tabela 1. Resultado do teste de tempo dos programas Triplet-SVM **(A)** e HHMMiR **(B)** com 10, 50, 100, 500 e 1.000 sequências em humano. Em **azul** o tempo executando comandos *stand-alone*; **vermelho** via web com miRQuest.

a	10		50		100		500		1000	
b	1,1 KB		5,4 KB		10,7 KB		54,6 KB		107,5 KB	
	A	B	A	B	A	B	A	B	A	B
1	0m0.051s	0m0.053s	0m0.183s	0m0.193s	0m0.343s	0m0.360s	0m1.849s	0m1.962s	0m3.689s	0m3.692s
2	0m0.049s	0m0.092s	0m0.086s	0m0.107s	0m0.132s	0m0.139s	0m0.510s	0m0.094s	0m0.939s	0m0.205s
3	0m0.011s	0m0.092s	0m0.015s	0m0.144s	0m0.016s	0m0.422s	0m0.036s	0m0.908s	0m0.046s	0m1.754s
	0m0.111s	**0m0.237s**	**0m0.284s**	**0m0.444s**	**0m0.491s**	**0m0.921s**	**0m2.395s**	**0m2.964s**	**0m4.674s**	**0m5.651s**
	0m3.00s	**0m2.00s**	**0m4.00s**	**0m3.00s**	**0m4.00s**	**0m4.00s**	**0m5.00s**	**0m5.00s**	**0m10.00s**	**0m9.00s**

Fonte: o autor.

Notas. [a] e [b] significam respectivamente tamanho da sequência e tamanho do arquivo. E os números 1, 2 e 3 correspondem ao conjunto de comandos que cada ferramenta possui para o processamento das sequencias. Na execução *stand-alone* cada comando precisa ser processado separadamente, o resultado em azul corresponde a soma desses tempos.

Tabela 2. Resultado do teste de tempo do preditor MiPred.

[a]	10	50	100	500	1000
[b]	1,1 KB	5,4 KB	10,7 KB	54,6 KB	107,5 KB
Local	1m44.183s	11m17.601s	25m48.489s	148m30.318s	278m11.054s
Web	2m33.00s	15m5.00s	27m47s	189m25.00s	329m04.00s

Fonte: o autor.
Nota. [a e b] tem o mesmo significado da tabela 1.

O sistema miRQuest expressou um tempo maior de execução devido a integração dos Preditores - que é um custo normal necessário. E, nos testes *stand-alone*, não foi avaliado o tempo que o usuário leva para escrever os comandos, foi considerado apenas a execução real dos comandos. Outro aspecto perceptível é que o MiPred (Tabela 2) é a ferramenta que mais demanda tempo de execução. HHMMiR (Tabela 1) é a ferramenta mais rápida para execução, em geral, quando comparada ao Triplet- SVM e ao NovoMIR (Tabela 1 e 3, respectivamente).

Tabela 3. Teste de tempo utilizando organismo de planta na ferramenta NovoMIR com 10, 30, 50, 100 e 200 sequências.

[a]	10	30	50	100	200
[b]	2,2 KB	5,5 KB	8,8 KB	19,3 KB	42,6 KB
Local	0m2.203s	0m4.328s	0m7.338s	0m.18.539s	0m.44.789s
Web	0m.3.00s	0m.7.00s	0m10.00s	0m24.00s	0m56.00s

Fonte: o autor.
Nota. [a e b] tem o mesmo significado da tabela 1.

Na análise de performance - comparação entre as ferramentas (*benchmarking*) - os resultados demonstram que o NovoMIR é o melhor preditor para plantas e o MiPred para humano, conforme dados da Tabela 4. A surpresa foram os resultados obtidos para o preditor HHMMiR, uma vez que essa ferramenta foi desenvolvida e treinada, principalmente, para predição de miRNAs em humanos.

Tabela 4. Resultado do *benchmarking* dos preditores. Testes feitos em humano e planta utilizando o sistema miRQuest. Em verde está destacado as melhores pontuações.

	Triplet-SVM	MiPred	HHMMiR	NovoMIR	
A.thaliana	64,09%	34,90%	9,06%	77,27%	Sensibilidade
	87,47%	96,48%	91,91%	99,50%	Acurácia
	16,99%	51,49%	6,32%	90,27%	Precisão
	26,86%	41,60%	7,45%	83,27%	F1-Score
H. sapiens	59,42%	92,81%	0,66%	40,28%	Sensibilidade
	83,20%	97,72%	79,95%	88,56%	Acurácia
	52,42%	94,18%	2,44%	98,56%	Precisão
	55,70%	93,49%	1,04%	57,19%	F1-Score
Ambos	98,78%	99,86%	95,00%	98,78%	Especificidade

Fonte: o autor.

6. Conclusão

Os resultados apresentaram um tempo maior para execução via web, mas o miRQuest, mostrou-se viável por não apresentar grande discrepância de tempo com relação a execução *stand-alone*. No que se refere ao *benchmarking*, realizado pelo miRQuest, demonstrou-se os bons resultados para o NOVOMIR e o MiPred. Tem-se interesse futuro de adicionar outros preditores ao miRQuest e testar o seu desempenho em arquitetura *Cloud Computing*.

7. Agradecimentos

ARP agradece pelo suporte do projeto ao CNPq - Edital MCTI/CNPQ/Universal 14/2014 - Faixa A - até R$ 30.000,00 Processo: 454505/2014-0. Esta pesquisa é parte do resultado de mestrado da aluna RRA do Programa de Pós-Graduação em Informática (Profissional) UTFPR Cornélio Procópio, PR, Brasil.

Referências

Allmer J1, Yousef M. Computational methods for ab initio detection of microRNAs. Front Genet. 2012 Oct 10; 3:209. doi: 10.3389/fgene.2012.00209. eCollection 2012.

Araújo DAM, Maracaja-Coutinho V, Padilha IQM, Rego TG. (2005) Genômica e Bioinformática: importância e perspectivas para o Nordeste. Ciência e Cotidiano. 1:5-9.

Barros-Carvalho GA, Paschoal AR, Marcelino-Guimarães FC, Hungria M. (2014) Prediction of potential novel microRNAs in soybean when in symbiosis. Genet Mol Res. 13(4):8519-29. doi: 10.4238/2014.October.20.28.

Brameier M, Wiuf C (2007). Ab initio identification of human microRNAs based on structure motifs. BMC Bioinformatics, 8:478.

Gomes CP1, Cho JH, Hood L, Franco OL, Pereira RW, Wang K. (2013) A Review of Computational Tools in microRNA Discovery. Front Genet. 15; 4:81. doi: 10.3389/fgene.2013.00081.

Janssen S. et. al. (2008). Shape based indexing for faster search of RNA family databases. BMC Bioinformatics, 9:131.

Jiang P. et. al. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids, 35:339-344.

Kadri S. et al. (2009) HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. BMC Bioinformatics. 10 Suppl 1:S35.

Kozomara A. and Griffiths-Jones S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. NAR 42: D68-D73

Lee RC, Feinbaum RL, Ambros V; Feinbaum; Ambros (1993). "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14". Cell 75 (5): 843–54. Doi: 10.1016/0092-8674(93)90529-Y. PMID 8252621.

Machado-Lima, A.; del Portillo, H.A. ; DURHAM, A.M. (2008) Computational methods in noncoding RNA research. Journal of Mathematical Biology, v. 56, p. 15-49.

Oliveira KC, Carvalho MLP, Maracaja-Coutinho V, Kitajima JP, Verjovski-Almeida S (2011). Non-coding RNAs in schistosomes: an unexplored world. Anais da Academia Brasileira de Ciências, 83(2): 673-694.

Padilha IQM, Durbano JPM, Martins AB, Almeida RS, Maracaja-Coutinho V, Araújo DAM. (2008) A bioinformática como instrumento de inserção digital e difusão da biotecnologia. Revista Eletrônica Extensão Cidadã, 5.

Paschoal AR, Maracaja-Coutinho V, Setubal JC, Simões ZL, Verjovski-Almeida S, Durham AM. (2012) Non-coding transcription characterization and annotation: A guide and web resource for non-coding RNA databases. RNA BIOL, 9:274-282.

Powers, D.M.W. (2011) Evaluation: FromPrecision, Recall and F-Measureto ROC, Informedness, Markedness e Correlation. Journal of Machine Learning Technologies, 2 (1) 37-63.

Severino P, Oliveira LS, Torres N, Andreghetto FM, Klingbeil Mde F, Moyses R, Wünsch-Filho V, Nunes FD, Mathor MB, Paschoal AR, Durham AM. (2013) High-throughput sequencing of small RNA transcriptomes reveals critical biological features targeted by microRNAs in cell models used for squamous cell cancer research. BMC Genomics. 14:735. doi: 10.1186/1471-2164-14-735.

Steffen P. et al. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics. 22(4):500-3.

Taft RJ, Pheasant M, Mattick JS. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays. 29(3):288-99.

Teune J.-H. and Steger, G. (2010) NOVOMIR: De Novo Prediction of MicroRNA-Coding Regions in a Single Plant-Genome. Journal of Nucleic Acids.

Xue C. et al. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics, 6:310.

Zhang, B. et al. (2007) MicroRNAs and their regulatory roles in animals and plants. J. Cell. Physiol., 210:279–289.