***SIMTar: Uma ferramenta para predição de SNPs
interferindo em sítios alvos de microRNAs***

Amanda Rusiska Piovezani

DISSERTAÇÃO APRESENTADA
AO
PROGRAMA INTERUNIDADES
DE
PÓS-GRADUAÇÃO
EM
BIOINFORMÁTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Área de Concentração: **Bioinformática**

Orientadora: **Prof^a. Dr^a. Arianne Machado Lima**

Durante a elaboração deste trabalho, a autora recebeu auxílio financeiro do CNPq

– São Paulo, SP – Novembro de 2013 –

***SIMTar: Uma ferramenta para predição de SNPs
interferindo em sítios alvos de microRNAs***

Este exemplar corresponde à redação final
da dissertação devidamente corrigida e
defendida por Amanda Rusiska Piovezani e
aprovada pela comissão julgadora.

São Paulo, Novembro de 2013

Comissão Julgadora:

- Prof^ª. Dr^ª. Arianne Machado Lima - EACH-USP (orientadora)
- Prof. Dr. Eduardo Moraes Rego Reis - IQ-USP
- Prof. Dr. Luciano Antonio Digiampietri - EACH-USP

RESUMO

PIOVEZANI, AMANDA RUSISKA. **SIMTar: Uma ferramenta para predição de SNPs interferindo em sítios alvos de microRNAs.** 71 p. Dissertação (Mestrado em Bioinformática) - Interunidades em Bioinformática, Universidade de São Paulo, São Paulo, 2013.

Polimorfismos de um nucleotídeo (SNPs) podem alterar, além de códons de leitura da tradução de genes, posições genômicas importantes do processo de regulação gênica, como sítios de iniciação de transcrição, de *splicing* e, mais especificamente, sítios alvos de microRNAs (miRNAs). Em particular, a identificação de SNPs alterando sítios alvos de miRNAs é um problema em aberto, embora venha ganhando um importante destaque nos últimos anos decorrente dos avanços descobertos sobre a capacidade dos miRNAs como elementos reguladores do genoma, associados inclusive a muitas doenças como o câncer e vários transtornos psiquiátricos. Os recursos computacionais atualmente disponíveis para esta finalidade (alguns bancos de dados e uma ferramenta) estão restritos à análise de SNPs na região 3'UTR (*UnTranslated Regions*) de RNAs mensageiros, onde miRNAs geralmente se ligam para reprimir sua tradução. No entanto, essa é uma simplificação do problema, dado que já se conhece a regulação por miRNAs ativando ou reprimindo a transcrição gênica quando ligados à sua região promotora, aumentando a efetividade da regulação negativa da tradução quando ligados à região codificante do gene ou ainda miRNAs se ligando a RNAs não codificantes. Esses recursos se limitam também à identificação de SNPs na região *seed* dos sítios de miRNAs, e portanto só identificam criação ou ruptura de sítios. Porém, SNPs localizados fora dessa região não só podem colaborar na criação e ruptura de sítios como também interferir na estabilidade de ligação de miRNAs e, portanto, na efetividade da regulação. Além disso, considerando toda a extensão do sítio, não somente a *seed*, é possível ocorrer mais de um SNP e, sendo assim, a combinação desses SNPs pode ter uma influência ainda maior na ligação com o miRNA. Os recursos atuais também não informam quais alelos dos SNPs, muito menos quais combinações deles, estão causando qual efeito. Por fim, tais recursos estão restritos a *Homo sapiens* e *Mus musculus*. Assim, este trabalho apresenta a ferramenta computacional SIMTar (*SNPs Interfering in MicroRNA Targets*), desenvolvida para identificar SNPs que alteram sítios alvos de miRNAs e que preenche as lacunas mencionadas. Além disso, é descrita uma aplicação de SIMTar na análise de 114 SNPs associados à esquizofrenia, na qual todos foram preditos interferindo em sítios alvos de miRNAs.

Palavras-chave: microRNAs, predição de alvos de microRNAs, SNPs, SNPs em sítios alvos de microRNAs, ligação de microRNAs em DNA e RNAs não codificantes.

Capítulo 1

Introdução

MICRORNAs (miRNAs) são pequenas moléculas de RNAs não codificadores, com tamanho aproximado de 22 nucleotídeos (nts), que têm revelado uma importante participação no processo de regulação de vários organismos, atuando em nível pré e pós-transcricional (Büssing *et al.* [2008]; Chen *et al.* [2004]; Frankel *et al.* [2008]; Si *et al.* [2006]). Em animais este processo de regulação é realizado através de um pareamento imperfeito com o sítio alvo. Os miRNAs são mais conhecidos pelo papel que desempenham na regulação negativa da tradução, efetuada quando estão pareados à porção 3'UTR de sequências de mRNAs. No entanto, o estudo de miRNAs mostrou recentemente várias novas maneiras de regulação pelas quais os miRNAs podem regular, além da tradução, também a transcrição dos genes. Estes novos estudos identificaram que a regulação pode ser positiva ou negativa, não só quando estão ligados à região 3'UTR mas também quando se ligam inclusive no DNA e sequências de outros RNAs não codificadores (Huang *et al.* [2012]; Jackson *et al.* [2010]; Lee *et al.* [2005]; Lin *et al.* [2011]; Shen *et al.* [2002]; Toscano-Garibay & Aquino-Jarquin [2012]; Vasudevan *et al.* [2007]; Zisoulis *et al.* [2012]). Muitos deles estando associados a doenças complexas como o câncer, diabetes e transtornos psiquiátricos.

Dado que o pareamento em animais ocorre geralmente de forma imperfeita e a região do pareamento abrange por volta de 22 nts, pequenas alterações como é o caso de polimorfismos de um nucleotídeo (*Single Nucleotide Polymorphisms* - SNPs), podem criar um novo sítio, romper um sítio

pré-existente ou mesmo interferir na estabilidade do pareamento dada a sua variação alélica. Isso ocorre porque a variação alélica do SNP pode implicar no aumento ou redução da complementaridade da sequência do miRNA com a sequência alvo e consequentemente na efetividade da regulação pelo miRNA.

SNPs com menor frequência alélica (*Minor Allele Frequency* - MAF) $\geq 1\%$ correspondem a 98% da variabilidade humana [McVean *et al.*, 2012]. Embora estudos de associação (*Genome-Wide Association Study* – GWAS) [Hindorff, 2009] tenham revelado centenas destes SNPs associados a muitas doenças (Jia *et al.* [2010]; Thomas *et al.* [2009]; Zhang *et al.* [2007]), muitos ainda não possuem nenhuma associação. Aproximadamente 90% destes SNPs estão localizados em regiões não codificantes do genoma [Hindorff, 2009], tal como a região promotora dos genes [Savinkova *et al.*, 2013], onde podem alterar sítios de elementos reguladores dessa região e consequentemente alterar muitas vias metabólicas relacionadas à transcrição destes genes. Podem também ocorrer em sítios de excisão de íntrons (*splicing*) e alterar o uso de éxons, implicando na síntese de diferentes isoformas de um mesmo gene (transcritos). Outra possibilidade é que tais SNPs podem criar, romper ou interferir na estabilidade de sítios de ligação de miRNAs [Lu & Clark, 2012].

O estudo computacional de identificação de SNPs que estejam interferindo em sítios alvos de miRNAs é uma área recente de pesquisa e constitui um problema em aberto. Entre as lacunas existentes nos estudos realizados até o momento está o fato desses estudos não considerarem que miRNAs podem se ligar a regiões distintas das 3'UTRs, em RNAs não codificantes ou até mesmo em DNA. Além disso, não relatam o efeito de cada variação alélica, nem tampouco analisam a variação de combinação alélica de mais de um SNP localizado no mesmo sítio alvo, informações importantes, uma vez que a frequência alélica de um SNP pode estar associada a uma doença. Outra limitação é que tais estudos analisam SNPs localizados somente em uma pequena região dos possíveis sítios alvos, conhecida como *seed*. Finalmente, os recursos computacionais utilizados nesses estudos não estão disponíveis como ferramentas de uso local e portanto não podem, por exemplo, serem aplicados a espécies distintas das analisadas no estudo.

Diante disso, este trabalho objetivou o desenvolvimento de uma ferramenta computacional *open-source* capaz de superar as lacunas existentes aqui mencionadas. A ferramenta desenvolvida chama-se

SIMTar (*SNPs Interfering in MicroRNAs Targets*). Para exemplificar o uso da ferramenta, SIMTar foi aplicado a SNPs associados a esquizofrenia.

Este documento corresponde a versão simplificada da dissertação e contém o título, o resumo, a introdução, a conclusão e a bibliografia do trabalho.

Capítulo 2

Conclusões

2.1 Conclusões

Em estudos de regulação por miRNAs, têm crescido as investigações quanto à presença de polimorfismos de um nucleotídeo (SNPs) ocorrendo em sítios alvos de miRNAs. Isso porque a sua variação alélica pode interferir no pareamento com o sítio alvo. Além disso, se tanto o SNP quanto o miRNA estiver associado a algum traço fenotípico, a predição de interferência entre eles aumenta ainda mais a conexão de regulação.

O fato é que a investigação envolvendo estes dois elementos deve incluir as várias formas de regulação por miRNAs e, portanto, todas as possíveis regiões em que eles puderem se ligar. No entanto, os recursos computacionais existente para predizer essa relação apenas a fazem sobre a primeira via de regulação conhecida, que é sobre a região 3'UTR de mRNAs. Com isso todas as outras regiões não estão sendo analisadas. Além disso, não está sendo considerada a possibilidade de mais de um SNP estar localizado em um mesmo sítio alvo e também estar interferindo no pareamento.

Dessa forma, a principal questão levantada nesse trabalho foi a metodologia utilizada na identificação de SNPs que podem alterar sítios alvos de miRNAs em função de sua variação alélica. Para isso foram discutidos os fatores que podem estar relacionados a esta ligação. Em linhas gerais foram considerados e discutidos: a escolha da ferramenta de predição de alvos de miRNAs, que pode res-

tringir a análise para alvos conservados evolutivamente ou não, o delineamento para análise de cada alelo do SNP e também de SNPs vizinhos inseridos em um mesmo sítio alvo, o contexto genômico da região alvo, a posição do SNP em relação ao sítio alvo, que considera o SNP dentro ou fora da seed, a análise de sequência de DNA (genômica) ou RNA (apenas éxons) a existência e disponibilidade para uso local de uma ferramenta de predição de tais SNPs e a espécie analisada.

Com o levantamento realizado ficou claro que as tecnologias disponíveis no estudo de SNPs e miRNAs resolvem apenas uma parte do problema. Neste sentido, este trabalho desenvolveu a ferramenta computacional **SIMTar: SNPs Interfering on MicroRNA Target sites**, a fim de superar as principais lacunas existentes.

A ferramenta analisa SNPs localizados tanto em regiões codificadoras como não codificadoras de proteínas. Os resultados são gerados para cada alelo ou combinação de alelos do SNP, caso existam mais um localizados em um mesmo sítio alvo. Para análise de SNPs localizados em DNA, os resultados mostram se naquela região existem sequências transcritas de RNAs codificadores e não codificadores de proteínas. Por fim, SIMTar é uma ferramenta de código aberto e que pode ser aplicada a outras espécies além da humana. No presente trabalho, SIMTar foi aplicado a dados de esquizofrenia como estudo de caso, mas esta pode ser utilizada para qualquer lista de SNPs.

No estudo de caso utilizando SNPs associados à esquizofrenia, 95% dos SNPs foram preditos interferindo em sítios alvos de miRNAs. Além disso, nestes resultados foram identificados 8 de 18 miRNAs também associados à esquizofrenia. Com a análise dos resultados foi possível identificar quais os alelos dos SNPs estão afetando os sítios alvos e os nomes dos miRNAs envolvidos. Quando todos os dados dos SNPs e miRNAs preditos são plotados em uma rede de interação, é possível identificar a conexão entre mais de um gene, quando ambos já foram observados em estudos de associação como é o caso da esquizofrenia, seria interessante investigar se estes miRNAs não estariam agindo sobre os sítios alvos preditos e regulando a região, seja em nível transcricional ou traducional. Esta regulação pode estar envolvendo processos biológicos e vias metabólicas importantes para o indivíduo e que pode estar danificada no estado de doença.

Neste estudo, a lista de SNPs foi obtida a partir de 38 genes candidatos em esquizofrenia, partindo assim de uma lista consideravelmente pequena, dado que os SNPs correspondem à forma mais

frequente de variação do genoma humano. Por outro lado, se for considerado o conjunto de dados gerado por exemplo por estudos de associação, onde centenas de SNPs são obtidos para cada doença e a maioria ainda não possui papel biológico definido, SIMTar pode ser aplicado para indicar um subconjunto mais promissor de SNPs para alguma validação biológica. Por exemplo, é possível selecionar facilmente os SNPs que interferem em um grande número de sítios. Uma das saídas do SIMTar é justamente a lista de SNPs ordenada decrescentemente pelo número de sítios alvos interferidos pelo SNP.

Um artigo descrevendo o SIMTar e sua aplicação em esquizofrenia está em fase final de escrita. Além disso está sendo solicitado o registro do SIMTar como software aberto.

Referências Bibliográficas

- BÜSSING, I., SLACK, F. J. & GROSS, H. (2008). let-7 microRNAs in development, stem cells and cancer. *Trends in Molecular Medicine* **14**(9), 400–409.
- CHEN, C.-Z., LI, L., LODISH, H. F. & ET AL (2004). MicroRNAs Modulate Hematopoietic Lineage Differentiation. *Science* **303**(5654), 83–86.
- FRANKEL, L. B., CHRISTOFFERSEN, N. R., JACOBSEN, A. & ET AL (2008). Programmed Cell Death 4 (PDCD4) Is an Important Functional Target of the Microrna miR-21 in Breast Cancer Cells. *Journal of Biological Chemistry* **283**(2), 1026–1033.
- HINDORFF, L. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. In: *Proceedings of the National Academy of Sciences*, vol. 106.
- HUANG, V., PLACE, R. F., PORTNOY, V. & ET AL (2012). Upregulation of cyclin B1 by miRNA and its implications in cancer. *Nucleic Acids Research* **40**, 1695–1707.
- JACKSON, R. J., HELLEN, C. U. & PESTOVA, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biology* **11**, 113–127.
- JIA, P., SUN, J., GUO, A. Y. & ET AL (2010). SZGR: a comprehensive schizophrenia gene resource. *Molecular psychiatry* **15**(5), 453–462.
- LEE, S. J., JEON, H. S., JANG, J. S. & ET AL (2005). DNMT3B polymorphisms and risk of primary lung cancer. *Carcinogenesis* **26**, 403–409.

- LIN, C.-C., LIU, L.-Z., ADDISON, J. B. & ET AL (2011). A KLF4–miRNA-206 Autoregulatory Feedback Loop Can Promote or Inhibit Protein Translation Depending upon Cell Context. *Molecular Cell Biology* **31**, 2513–2527.
- LU, J. & CLARK, A. G. (2012). Impact of microRNA regulation on variation in human gene expression. *Genome Research* **22**(7), 1243–1254.
- MCVEAN, G. A., DONNELLY, P., LUNTER, G. & ET AL (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65.
- SAVINKOVA, L., DRACHKOVA, I., ARSHINOVA, T. & ET AL (2013). An Experimental Verification of the Predicted Effects of Promoter TATA-Box Polymorphisms Associated with Human Diseases on Interactions between the TATA Boxes and TATA-Binding protein. *PLoS ONE* **8**(2), e54626.
- SHEN, H., WANG, L., SPITZ, M. & ET AL (2002). A novel polymorphism in human cytosine DNA-methyltransferase-3B promoter is associated with an increased risk of lung cancer. *Cancer Research* **62**, 4992–4995.
- SI, M. L., ZHU, S., WU, H. & ET AL (2006). miR-21-mediated tumor growth. *Oncogene* **26**(19), 2799–2803.
- THOMAS, G., JACOBS, K. B., KRAFT, P. & ET AL (2009). A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p112 and 14q241 *rad51l1*. *Nature Genetics* **41**, 579–84.
- TOSCANO-GARIBAY, J. D. & AQUINO-JARQUIN, G. (2012). Regulation exerted by miRNAs in the promoter and UTR sequences: MDR1/P-gp expression as a particular case. *DNA Cell Biology* **31**, 1358–64.
- VASUDEVAN, S., TONG, Y. & STEITZ, J. A. (2007). Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318**(5858), 1931–1934.
- ZHANG, H., LIU, L., WANG, X. & ET AL (2007). Guideline for data analysis of genomewide association studies. *Cancer genomics & proteomics* **4**(1), 27–34.

ZISOULIS, D. G., KAI, Z. S. & CHANG, R. K. (2012). Autoregulation of microRNA biogenesis by let-7 and Argonaute. *Nature* **486**, 541–544.