Incorporação de evidências biológicas para a identificação de SNPs interferindo em sítios alvos de miRNAs

PAULA PRIETO OLIVEIRA

Tese apresentada ao Programa de Interunidades em Bioinformática da Universidade de São Paulo para a obtenção do título de Doutora em Ciências

Programa: **Bioinformática**
Orientadora: **Profa. Dra. Ariane Machado Lima**
Co-orientadora: **Profa. Dra. Helena P. Brentani**

São Paulo, SP - dezembro de 2018

Incorporação de evidências biológicas para a identificação de SNPs interferindo em sítios alvos de miRNAs

Este exemplar corresponde à redação final da tese devidamente corrigida e defendida por Paula Prieto Oliveira e aprovada pela Comissão Julgadora.
Área de concentração: Bioinformática.

Orientadora: Profa. Dra. Ariane Machado Lima.
Co-orientadora: Profa. Dra. Helena P. Brentani.

Incorporação de evidências biológicas para a identificação de SNPs interferindo em sítios alvos de miRNAs

Este exemplar corresponde à redação final da tese devidamente corrigida e defendida por Paula Prieto Oliveira e aprovada pela Comissão Julgadora.
Área de concentração: Bioinformática.

Orientadora: Profa. Dra. Ariane Machado Lima.
Co-orientadora: Profa. Dra. Helena Brentani.

Comissão Julgadora:

- Profª. Drª. Ariane Machado Lima - EACH-USP (orientadora)
- Profª. Drª. Fátima L. S. Nunes Marques - EACH-USP
- Profª. Drª. Mariana Maschietto - Hospital Infantil Boldrini
- Prof. Dr. Israel Tojal da silva - Fundação Antônio Prudente, Hospital AC Camargo
- Prof. Dr. Ricardo Giordano - IQ-USP

A meus pais

Que sempre me apoiaram nos momentos difíceis e são a razão da minha vida.

A meu irmão

Que é um grande amigo e companheiro.

A minha tia Sônia

Que sempre esteve disposta a me ajudar com muito carinho e me deu força para superar os desafios.

Agradecimentos

À Profa Dra Ariane Machado Lima por ser minha orientadora, pela oportunidade de realizar o doutorado, por todos os ensinamentos e pela paciência.

À Profa Dra Helena Brentani por ser minha co-orientadora.

Ao meu amigo Denis Jacob Machado, pela amizade, pelos ensinamentos e pelo apoio.

Ao meu amigo Lucas, aluno de iniciação científica, por me ajudar diante das minhas dificuldades em programação.

Ao CEPID-CeMEAI - Centro de Ciências Matemáticas Aplicadas a Indústria, pela concessão de uso de um equipamento computacional servidor que permitiu as análises aqui realizadas (Processo FAPESP 2013/07375-0).

Resumo

SIMTar (*SNPs Interfering in MicroRNA Targets)* é uma ferramenta computacional que realiza a predição de interferência de SNPs em sítios alvos de miRNAs. Dada uma lista de SNPs, SIMTar analisa uma janela ao redor de cada SNP e verifica se tal SNP cria ou rompe um sítio de miRNA utilizando como base programas de predição de sítios de miRNAs nos vários alelos desta janela. Se o resultado da predição para um dado miRNA é diferente para diferentes alelos, então tal SNP está potencialmente interferindo em tal sítio. O presente trabalho teve como objetivo melhorar o processo de identificação de interferência de SNPs em sítios alvos de miRNAs, usando um novo programa de predição de sítios de miRNAs, assim como a incorporação de resultados biológicos experimentais. A hipótese deste trabalho é que com estas modificações poderia-se diminuir a expectativa de falsos positivos sem diminuir a sensibilidade. Os resultados obtidos validam a hipótese e ainda mostram que o SIMTar possui vantagens quando comparado com outras ferramentas ou bancos de dados de propósito similar.

Palavras-chave: SNPs, sítios de microRNAs, interferência de SNPs.

Abstract

SIMTar (SNPs Interfering in MicroRNA Targets) is a computational tool that performs the prediction of SNPs interference in miRNA target sites. Given a list of SNPs, SIMTar analyzes a window around each SNP and checks whether such SNP creates or breaks a miRNA site based on miRNA target prediction programs in the various alleles of this window. If the prediction result for a given miRNA is different for different alleles, then such SNP is potentially interfering in such site. The present study had the aim to improve the SNPs interference identification in miRNA sites, using a new prediction program of miRNAs targets, as well as the incorporation of experimental biological results. The hypothesis of this study is that with these modifications the expectation of false positives could be reduced without diminishing the sensitivity. The results obtained validate the hypothesis and also show that the SIMTar has advantages when compared to other tools or databases with similar purpose.

Keywords: SNPs, microRNAs sites, SNPs interference.

SUMÁRIO

CAPÍTULO 1- INTRODUÇÃO

Embora "o genoma humano" seja disponibilizado basicamente como uma sequência para cada cromossomo, essas sequências não são as mesmas para todos os seres humanos, sendo este um dos fatores pelos quais não somos todos iguais. Estas variações nas sequências genômicas podem ter vários diferentes impactos, um deles sendo a criação ou o rompimento de sítios de ligação de microRNAs, moléculas estas envolvidas principalmente na regulação pós-transcricional da expressão gênica.

O contexto deste projeto é a identificação de variações de uma única base que interferem em tais sítios de microRNAs. Uma ferramenta computacional com tal finalidade havia sido criada em nosso grupo de pesquisa. Resumidamente, esta ferramenta baseia-se em executar programas de predição de sítios de miRNAs nas sequências variantes, identificando uma interferência se, para algum miRNA, o sítio era predito em uma variante da sequência e não era predito na outra variante. No entanto, estimava-se que tal ferramenta relatava muitos falsos positivos, uma vez que a maioria das variações testadas eram sugeridas como interferindo em sítios de miRNAs. Com o aumento de dados na literatura advindos de técnicas experimentais e ou métodos de análise para melhor estudo de miRNAs tornou-se possível a integração destes resultados à nossa ferramentas de predição. A hipótese deste projeto é, portanto, que a incorporação das informações advindas de tais experimentos e a utilização de programas mais adequados de predição de sítios alvos de microRNAs poderiam conferir maior confiabilidade aos resultados. Para tanto foi utilizada uma amostra como padrão ouro, ou seja, com sítios que interferem em miRNAs validados experimentalmente.

No restante deste capítulo são introduzidos os conceitos básicos importantes para o entendimento deste projeto.

1.1- Organização de genes eucariotos

Uma espécie eucariota é aquela que possui células nucleadas, em cujo núcleo se localiza pelo menos grande parte de seu DNA.

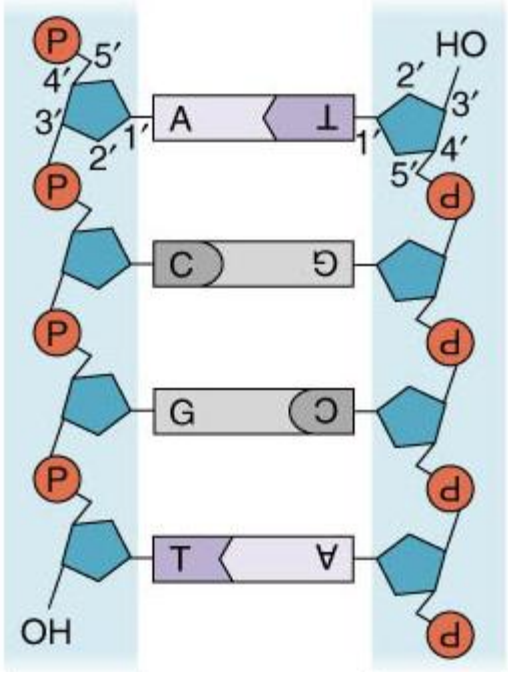
Uma molécula de ácido desoxirribonucleico (DNA) é formada por duas longas cadeias polipeptídicas (fitas de DNA) compostas por quatro tipos de nucleotídeos, ligados covalentemente entre si. Os nucleotídeos do DNA são constituídos por açúcares com cinco carbonos (desoxirriboses), aos quais se ligam um ou mais grupos fosfatos e uma base nucleotídica, que pode ser: adenina (A), timina (T), guanina (G) e citosina (C). As ligações de hidrogênio entre estas bases mantêm as duas fitas unidas. As formas e a estrutura química das bases permitem que as

interações entre elas sejam formadas eficientemente apenas entre as bases A e T, e C e G (Johnson et al, 2010).

A estrutura do DNA está ilustrada na figura 1. Os nucleotídeos estão ligados covalentemente por ligações fosfodiéster entre o grupo 3’-hidroxila (-OH) de um açúcar e o 5’-hidroxila-fostato (P) de próximo. Dessa forma, cada fita tem uma polaridade e as duas extremidades são quimicamente diferentes. A extremidade 5’ do DNA é, por convenção, a do grupo fosfato, enquanto a 3’ é a da hidroxila. Os pares de bases somente se encaixam se as duas fitas estiverem dispostas de forma antiparalela (polaridade de uma fita em direção oposta à da outra) (Johnson et al, 2010).

Assim, conhecida a sequência de uma fita de DNA, chamada fita positiva (denotada “+”), a sequência da fita oposta, negativa (“-”), é obtida pelas operações de reversão e complementação. A reversão reposiciona os nucleotídeos em ordem inversa (ou seja, “de trás para a frente”) enquanto a complementação troca um nucleotídeo pelo seu complementar (A por T e C por G). Por isso se diz que uma fita é o reverso-complementar da outra. Por convenção, a fita depositada nos bancos de dados genômicos oficiais é considerada a fita “+”.

Figura 1- Estrutura do DNA. A, C, G e T representam as bases nucleotídicas, 1', 2', 3', 4' e 5' identificam os cinco carbonos, OH o término 3’-hidroxila e P o fosfato. A fita da esquerda está descrita de cima para baixo em seu sentido 5' → 3', enquanto a fita da direita segue em sentido contrário.



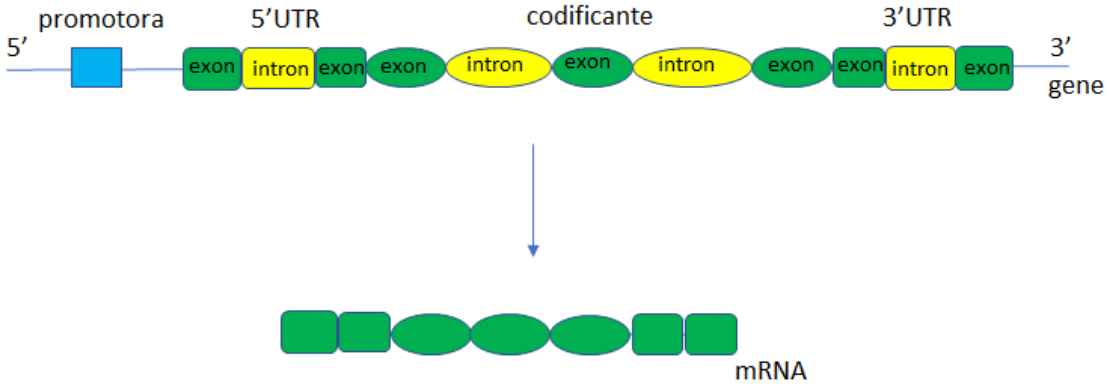
Fonte: http://bioblogkarolla.blogspot.com/2010/10/estrutura-do-dna.html

O genoma humano é composto por dois cromossomos sexuais (um cromossomo X e um cromossomo Y em homens ou um par de cromossomos X em mulheres) e 22 pares de cromossomos não sexuais (autossomos) numerados de 1 a 22 (Johnson et al, 2010).

Os genes são subsequências específicas dos cromossomos que contêm trechos que serão transcritos em RNA (figura 2). A região promotora se situa no extremo 5' (anterior) do gene e é responsável pelo início da transcrição, por meio da ligação a fatores gerais de transcrição e a RNA polimerase. O RNA transcrito pode conter trechos (chamados íntrons) que são eliminados por um processo chamado de *splicing*, sendo os trechos que permanecem no RNA final chamados éxons. Um mesmo gene pode dar origem a diferentes transcritos em um processo chamado *splicing alternativo*, no qual diferentes éxons são utilizados para a formação do RNA final (Johnson et al, 2010).

Um RNA pode ser codificante, se ele pode ser traduzido em uma proteína, ou não codificante, caso contrário. Um RNA codificante é formado por três regiões consecutivas: 5'UTR (*UnTranslated Region*), CDS (*CoDing Sequence*) e 3'UTR, das quais apenas a região CDS é traduzida em proteína (Brown, 2018).

Figura 2- Regiões do gene codificante e mRNA resultante. O retângulo azul representa a região promotora, as regiões verdes representam os éxons enquanto as regiões amarelas representam os íntrons. Além disso, no mRNA já processado, os retângulos com cantos arredondados representam as regiões não traduzidas (UTRs) e as elipses representam a região codificante (CDS). Embora esta seja uma única molécula, a imagem mostra a concatenação dos éxons transcritos do DNA, tanto das regiões UTRs quanto da CDS.



Fonte: Paula Prieto Oliveira, 2018

1.2- Polimorfismo de nucleotídeo único (SNP)

Considerando que uma base nucleotídica possa ser A (adenina), C (citosina), G (guanina) ou T (timina), uma variação de um nucleotídeo entre os indivíduos é dita *bialélica* se são observados na população 2 alelos (ou variantes, por exemplo A ou C, variação esta denotada por A/C), *trialélica* se observados 3 alelos e *tetra alélica* se observados os 4 alelos. Dentre os alelos observados na população um deles é o de menor frequência (MAF - *Minor Allele Frequence*).. Os SNPs podem ser classificados como comuns, de baixa frequência ou raros dependendo de seus valores de MAF na população em estudo. Os comuns apresentam MAF > 0.05, os de baixa frequência MAF entre 0.01 e 0.05 e os raros MAF abaixo de 0.01. Assim SNPs contribuem para variabilidade fenotípica normal, no entanto em algumas situações podem aumentar o risco de doença (Panoutsopoulou et al, 2013).

Os SNPs podem alterar a função ou regulação da expressão de proteínas de várias formas. A mais evidente é a variação não sinônima que, ocorrendo em uma região codificante do gene, promove a troca de um aminoácido por outro na proteína traduzida, e consequentemente altera a sequência da proteína formada. Alguns outros SNPs podem causar alterações nos sítios de *splicing*, resultando em variantes proteicas que se distinguem pelos éxons apresentados. Outros ocorrem em regiões promotoras e podem afetar a regulação e a expressão gênica (Kwok, 2003; Piovezani, 2013).

De forma semelhante, SNPs podem estar presentes em sítios alvos de microRNAs de tal forma a romper um sítio, ou alterar a estabilidade do pareamento, enfraquecendo ou fortalecendo a ligação. É possível ainda que um SNP crie um novo sítio de microRNA (Chen et al., 2008; Piovezani, 2013).

1.3- MicroRNAs

MicroRNAs (miRNAs) são RNAs não codificantes de fita única e aproximadamente 22 nucleotídeos de comprimento, envolvidos em processos de regulação da expressão gênica (Krol et al., 2010). A figura 3 resume a biogênese dos miRNAs. Os miRNAs são transcritos a partir de genes independentes ou processados a partir de genes hospedeiros codificantes ou não (Chu e Rana, 2007; Krol et al., 2010). Estes genes são transcritos geralmente pela RNA polimerase II, embora alguns casos possam ser transcritos pela RNA polimerase III, em um longo microRNA primário (pri-miRNA), que é clivado pela Drosha RNase III endonuclease no núcleo, resultando em um microRNA precursor (pré-miRNA) (Bartel, 2004). Este apresenta cerca de 70

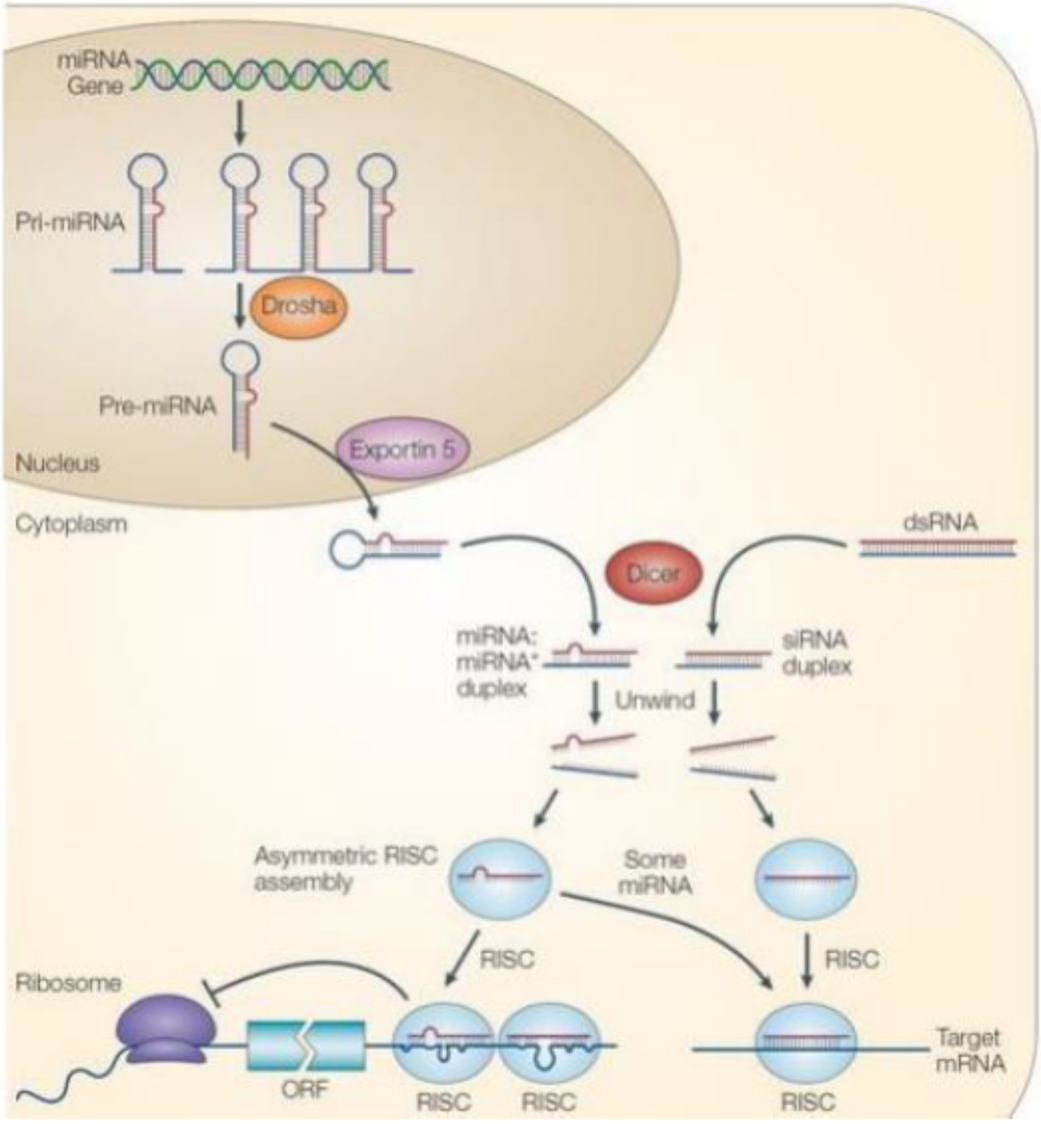
pares de bases, um trecho de fita dupla e uma alça de fita simples, formando um *hairpin*. O pré-miRNA é então transportado para o citoplasma, onde é clivado pela Dicer, gerando um duplex (fita dupla) de dois miRNAs maduros de 22 nucleotídeos cada. O complexo de silenciamento induzido por RNA (RISC), composto por várias proteínas dentre elas as argonautas (proteínas AGO), é responsável por romper o duplex e guiar um dos miRNAs maduros para o seu alvo (Kim et al., 2009; Krol et al., 2010; Thomas et al., 2010). Há quatro tipos de proteínas argonautas em humanos: AGO1, AGO2, AGO3 e AGO4. Todas elas se ligam a miRNAs com complementaridade parcial ao mRNA alvo promovendo seu silenciamento (Meister, 2013).

A regulação gênica exercida pelo miRNA pode ser pré ou pós-transcricional. A pré-transcricional ocorre por meio da ligação do miRNA à região promotora do DNA, ativando ou silenciando a transcrição (Toscano-Garibay e Aquino-Jarquin, 2012). No entanto, o mecanismo de ação mais conhecido é a regulação pós-transcricional em genes codificadores de proteínas por meio do pareamento com o RNA mensageiro (mRNA) dos mesmos, geralmente na porção 3'UTR do mRNA (Chen et al., 2008). Essa regulação pode ocorrer de três formas: degradação do mRNA, repressão traducional ou ativação da tradução (Wu e Belasco, 2008; Iwasaki e Tomari, 2009). Quando a complementaridade entre o miRNA e o mRNA é completa ou quase completa, há a clivagem direta e a degradação do mRNA. Mas quando a complementaridade é parcial, a tradução é inibida com posterior degradação do mRNA por meio da deadenilação (remoção da cauda poli-A) (Zhang et al., 2007; Wu e Belasco, 2008; França et al., 2010). Essa complementaridade parcial permite que o miRNA se ligue a vários sítios de sequências distintas, tornando mais difícil a predição computacional de alvos (Chu e Rana, 2007). Essa complementaridade parcial e imperfeita é bastante comum em animais, enquanto que em plantas o pareamento é perfeito ou quase-perfeito (Bartel, 2004; Chaudhuri e Chatterjee, 2007).

Há cinco tipos de sítio de ligação, um sendo o sítio canônico e quatro tipos de sítios não canônicos. O sítio canônico (figura 4), considerado o mais comum, possui de seis a sete pareamentos de bases complementares entre o miRNA e o RNA alvo em uma região chamada *seed,* que envolve os nucleotídeos 2-8 contando a partir do extremo 5’ do miRNA, (Bartel, 2009; Witkos et al., 2011; Moore et al., 2015). Os sítios não canônicos são o 3’ compensatório, sítio com pareamento central, sítio com pareamento central sem envolvimento de seed e pivô (figura 5). O sítio 3’ compensatório apresenta interação ruim na região *seed*, com presença de pelo menos uma base não pareada no meio desta região, mas compensado por uma boa complementaridade concentrada nos nucleotídeos 13-16 (acima de quatro bases) do miRNA (Bartel, 2009). O sítio de pareamento central possui 11-12 emparelhamentos contíguos de Watson-Crick na região central (nucleotídeos 4-15) (Shin et al., 2010). O pareamento central sem envolvimento de *seed* apresenta interações confinadas ao meio e ao extremo 3’ do microRNA (Helwak et al., 2013). Já o sítio pivô

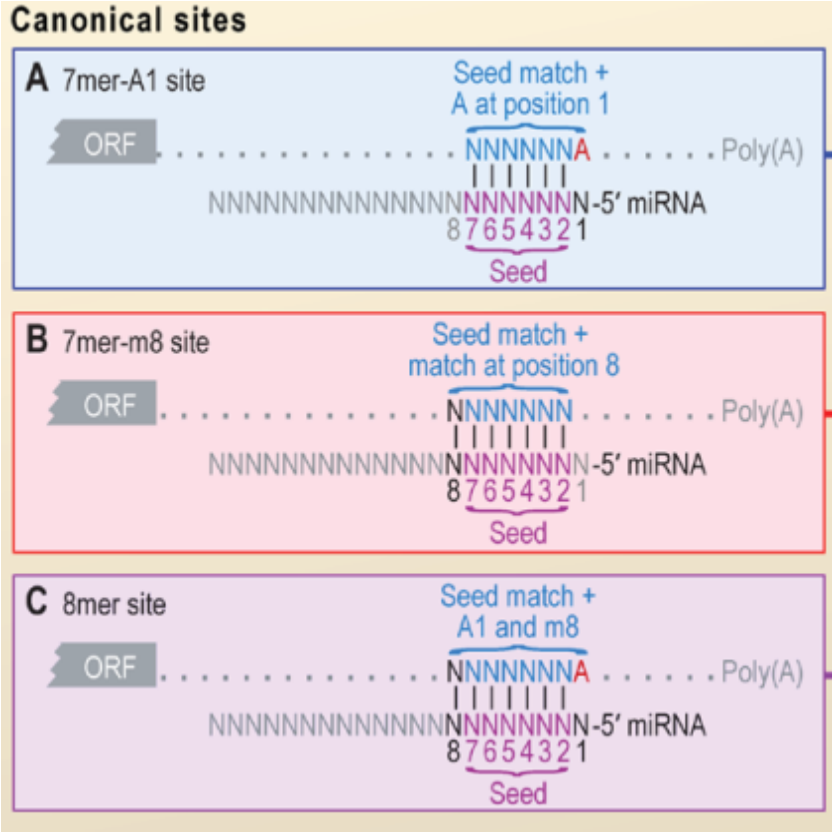
possui um pivô no nucleotídeo 6, com ligações das bases 2-6 e arqueamento nas posições 5-6 (Chi et al., 2012).

Figura 3- Biossíntese e mecanismo de ação dos miRNAs (vide texto).



Fonte: (He e Hannon, 2004)

Figura 4- Sítios canônicos.

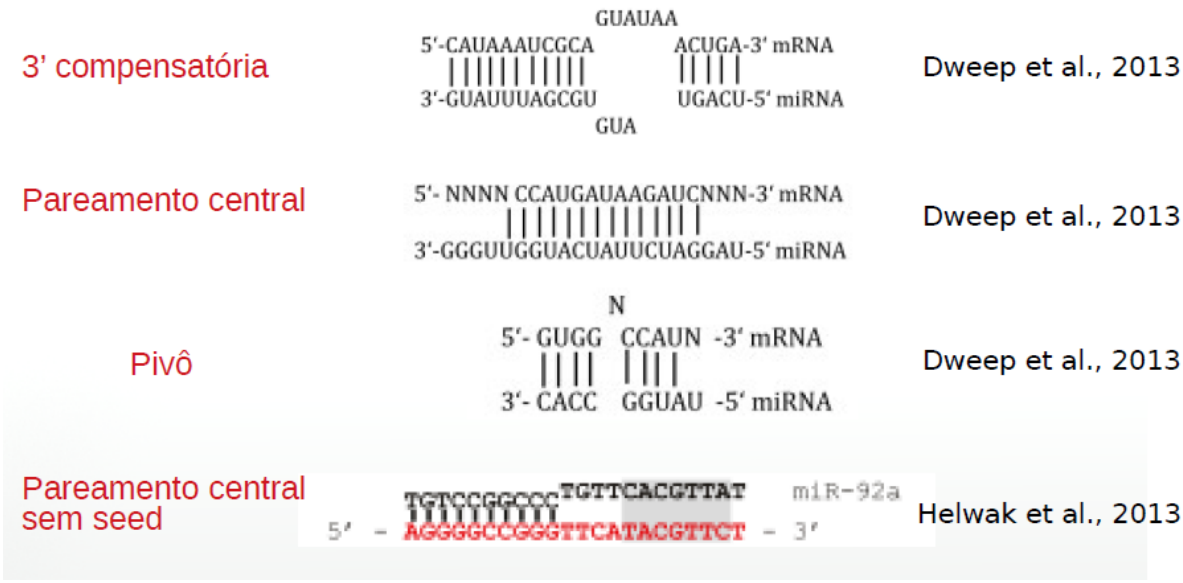


Fonte: Bartel et al, 2009.

Os miRNAs estão envolvidos em vários processos biológicos, como por exemplo: neurodesenvolvimento, formação de sinapses neuronais, proliferação celular, morte celular, defesa contra infecção viral, diferenciação e metabolismo (Huang et al., 2010; Thomas et al., 2010). Além disso, a expressão aberrante de vários miRNAs possuem um importante papel na carcinogênese e no desenvolvimento de metástases (Li et al., 2010; Huang et al., 2010). Sendo assim, a identificação dos genes alvos de microRNAs pode auxiliar no entendimento de seus mecanismos fisiopatológicos e muitas ferramentas computacionais foram desenvolvidas com esse objetivo (Huang et al., 2010).

Acreditava-se que os miRNAs se ligavam somente à região 3'UTR do mRNA pelo extremo 5' (região *seed*) (Lewis et al., 2005). Entretanto, estudos mostram que eles podem se ligar também às regiões 5'UTR e CDS do mRNA e promotora do DNA (Ørom et al, 2008; Chi et al., 2009; Hafner et al., 2010; Toscano-Garibay e Aquino-Jarquin, 2012).

Figura 5 - Os quatro tipos de sítios não canônicos. Em cada linha tem-se à esquerda o tipo de sítio, no centro um exemplo do padrão de pareamento entre o miRNA e o mRNA alvo, e à direita a citação bibliográfica da fonte da imagem central.



Fonte: Paula Prieto Oliveira, 2018

1.4- Predição computacional de alvos de microRNAs

Há atualmente várias ferramentas computacionais destinadas à predição de alvos de miRNAs em animais (Lewis et al., 2005; Grimson et al., 2007; Friedman et al., 2009; Enright et al., 2003; John et al., 2004; Kertesz et al., 2007; Rehmsmeier et al., 2004; Sturm et al., 2010; Radfar et al., 2013; Ahmadi et al. 2013, Xu et al., 2014; Li et al., 2014; Kim et al., 2006; Wang et al., 2013; Park e Kim, 2013; Menor et al., 2014; Liu et al., 2010; Ragan et al., 2009; Ogul et al., 2011; Reczko et al., 2012; Bandyopadhyay et al., 2015; Thadani e Tammi, 2006; Burgler e Macdonald, 2005; Incarnato et al., 2013; Mitra e Bandyopadhyay, 2011; Moxon et al., 2008; van Dongen et al., 2008; Saetrom et al., 2005; Bandyopadhyay e Mitra, 2009; Oulas et al., 2012; Rusinov et al., 2005; Miranda et al., 2006; Gumienny e Zavolan, 2015). Cada uma delas se baseia em características envolvidas na ligação entre o miRNA e seu alvo, dentre elas: padrão de pareamento de bases, estabilidade termodinâmica do miRNA-mRNA, conservação evolutiva, acessibilidade do sítio alvo e quantidade de sítios alvos (Min e Yoon, 2010; Piovezani, 2013).

Com relação ao padrão de pareamento de bases, as ferramentas levam em consideração um ou alguns dos cinco subtipos de sítios descritos acima. Muitas ferramentas, por

exemplo, se restringem a predizer alvos canônicos, como TargetScan (Lewis et al., 2005; Grimson et al., 2007; Friedman et al., 2009), HomoTarget (Ahmadi et al., 2013) e TargetS (Xu et al., 2014).

A estabilidade termodinâmica do miRNA-alvo é analisada por meio do cálculo da energia livre (ΔG) da ligação putativa. Quanto menor a energia livre do duplex miRNA/RNA alvo maior a energia necessária para rompê-lo e, portanto, mais estável é a interação (Huang et al., 2010).

Para a identificação de conservação evolutiva, *a*nálises comparativas de sequências ortólogas em várias espécies são realizadas para avaliar se os sítios alvos são conservados evolutivamente. O objetivo dessa estratégia é reduzir o número de falsos positivos; entretanto, deixa de prever resultados espécie-específicos, aumentando a quantidade de falsos negativos (Min e Yoon, 2010; Ritchie e Rasko, 2014).

A característica de acessibilidade do sítio alvo representa a necessidade da estrutura secundária local do RNA permitir que o miRNA se ligue no sítio específico, ou seja, a região do sítio não deve estar envolvida em pareamentos que impossibilite esta interação (Grimson et al., 2007).

A relevância da informação da quantidade de alvos vem do fato de que vários miRNAs são coexpressos e regulam uma mesma molécula de forma coordenada, e um determinado miRNA pode ter vários sítios em um mesmo RNA. Por essa razão, algumas ferramentas avaliam a presença e o número de múltiplos alvos ao fazer a predição, já que sítios isolados poderiam ser indícios de falsos positivos (Min e Yoon, 2010).

Os algoritmos não necessariamente utilizam todas as características colocadas acima, mas apenas algumas dependendo do algoritmo utilizado. Uma deficiência na predição é a falta de convergência entre os resultados produzidos por eles, pois cada um utiliza diferentes critérios para selecionar os alvos. Além disso, os miRNAs apresentam várias formas de interação, visto que podem se ligar à região 3'UTR, 5'UTR, CDS do mRNA ou promotora do DNA e existem sítios canônicos e não canônicos, com efeitos diferentes na regulação gênica. Sendo assim, as ferramentas treinadas em uma forma (por exemplo, ligação em 3' UTR de mRNAs), podem não detectar as outras (como ligação no DNA em regiões promotoras) (Ritchie e Rasko, 2014).

Alguns estudos foram conduzidos com o intuito de comparar esses algoritmos de predição.

Alexiou et al (2009) analisaram a sensibilidade e a precisão das seguintes ferramentas de predição: DIANA-microT 3.0, EIMMo, miRanda, miRBase, Pictar, PITA, RNA22 e TargetScan 5.0. Os autores utilizaram dados das mudanças de expressão proteica mediadas por miRNA do estudo de Selbach e al (2008), disponíveis em http://psilac.mdc-berlin.de, para classificar um par miRNA-alvo como positivo ou negativo e então estimar a precisão e sensibilidade das ferramentas.

Da Silva (2013) realizou um estudo comparativo entre os algoritmos Miranda, PITA, RNAhybrid e TargetScan avaliando, dentre outras métricas, a sensibilidade e a precisão. Foram gerados dois tipos de controles positivos e negativos: um para o TargetScan e outro para as demais ferramentas. A amostra positiva foi composta de pares de miRNA-mRNA validados experimentalmente com posição exata conhecida, obtidos do banco miRecords (Xiao et al., 2009). A amostra negativa foi obtida utilizando-se o mesmo conjunto de miRNAs e mRNAs da amostra positiva, só que com as sequências de mRNAs contendo os nucleotídeos embaralhados de forma a manter o mesmo tamanho e composição de bases dos mRNAs originais. Para o TargetScan, contudo, foi necessário utilizar os alinhamentos múltiplos dos mRNAs alvos, sendo que na amostra negativa foram utilizados os embaralhamentos das colunas desses alinhamentos.

A Tabela 1 compara as taxas de precisão e sensibilidade dos dois estudos. A precisão indica quantos por cento dos sítios preditos como positivos são realmente positivos (isto é, pertencem à amostra positiva). Logo, quanto maior a precisão, menor a taxa de falsos positivos. Já a sensibilidade indica quantos por cento dos sítios positivos foram preditos como positivos. Logo, quanto maior a sensibilidade menor a taxa de falsos negativos.

Tabela 1- Resultados dos estudos de Alexiou et al (2009) e da Silva (2013)

Ferramentas	**Alexiou et al., 2009**		**da Silva, 2013**	
	precisão	sensibilidade	precisão	sensibilidade
Miranda	29%	**20%**	6,71%	**69,13%**
PITA	26%	6%	1,39%	4,92%
RNAHybrid	-	-	32,26%	10,93%
TargetScan	**51%**	12%	**46,70%**	43,62%
Pictar	49%	10%	-	-
DIANA-microT	48%	12%	-	-
RNA 22	24%	6%	-	-

Fonte: Alexiou et al, 2009; da Silva, 2013.

Embora utilizando estratégias distintas de estimação de precisão e sensibilidade das ferramentas, ambos os estudos concordam que a ferramenta TargetScan é a que apresenta a maior precisão e a ferramenta Miranda é a que apresenta a maior sensibilidade, o que é coerente com o fato do TargetScan exigir conservação evolutiva dos alvos para diminuir a taxa de

falsos positivos, enquanto o Miranda não. Mas o mais surpreendente é o quanto as taxas de precisão e sensibilidade, mas em especial a precisão, ainda são muito baixas entre as ferramentas de predição.

Fan e Kurgan (2015) analisaram sensibilidade, especificidade, precisão, MCC (*Matheus Correlation Coeficient*), área sob a curva ROC (AUC), SNR+ = Verdadeiro Positivo (VP)/ Falso Positivo (FP), SNR- = Verdadeiro Negativo (VN)/ Falso Negativo (FN), e PNR = VP +FP/ VP + FN de sete ferramentas (TargetScan, miRanda, PicTar, DIANA-microT, miRmap, EIMMo e MirTarget2), conforme mostra a tabela 2. Considerando as predições dos miRNA-mRNA duplexes, TargetScan e DIANA-microT apresentaram os maiores valores de AUC. Além disso, DIANA-microT obteve o mais alto MCC. TargetScan apresentou a maior sensibilidade, enquanto o PicTar obteve a maior especificidade. DIANA-microT apresentou o mais alto SNR+, enquanto o TargetScan obteve o maior SNR- e o maior PNR. A análise de significância estatística mostrou que a diferença entre os valores de AUC de TargetScan, DIANA-microT e miRmap não é estatisticamente significante, mas essas três ferramentas são significativamente melhores que as outras quatro. Para os pares miRNA-gene, o TargetScan apresentou o maior valor de AUC, enquanto o EIMMo obteve o segundo maior AUC e o melhor MCC. O miRmap obteve a maior sensibilidade e o TargetScan o maior balanço entre sensibilidade e especificidade. O MirTarget2 apresentou a melhor especificidade, precisão e SNR+; ele prediz apenas alguns alvos funcionais mas com uma alta taxa de sucesso. O TargetScan obteve o maior PNR. Análise de significância dos valores de AUC mostrou que o TargetScan foi significativamente melhor que os demais. AUCs de EIMMo e miRmap não são significativamente diferentes e são significativamente melhores que os outros métodos.

Tabela 2- Resultados do estudo de Fan e Kurgan (2015).

Prediction type	Predictor	AUC	MCC	Sen.	Spe.	Prec.	SNR+	SNR-	PNR
At the duplex level	TargetScan	**0.674**	0.200	**0.823**	0.389	0.855	1.346	**2.194**	**0.962**
	DIANA-microT	0.673	**0.273**	0.627	0.722	**0.908**	**2.256**	1.934	0.690
	miRmap	0.658	0.158	0.741	0.444	0.854	1.333	1.713	0.867
	miRanda	0.560	0.081	0.437	0.667	0.852	1.310	1.184	0.513
	EIMMo	0.552	0.116	0.696	0.444	0.846	1.253	1.463	0.823
	PicTar	0.538	0.069	0.272	**0.806**	0.860	1.400	1.107	0.316
	MirTarget2	0.519	0.055	0.285	0.778	0.849	1.282	1.088	0.335
At the gene level	TargetScan	**0.748**	0.386	0.733	0.652	0.733	2.108	2.446	**1.000**
	EIMMo	0.725	**0.391**	0.707	0.687	0.746	2.257	2.342	0.947
	miRmap	0.714	0.353	**0.800**	0.539	0.694	1.736	**2.696**	1.153
	DIANA-microT	0.637	0.225	0.520	0.704	0.696	1.759	1.467	0.747
	miRanda	0.636	0.239	0.467	0.765	0.722	1.988	1.435	0.647
	MirTarget2	0.627	0.298	0.327	**0.922**	**0.845**	**4.174**	1.369	0.387
	PicTar	0.588	0.196	0.340	0.835	0.729	2.058	1.265	0.467

Fonte: Fan e Kurgan, 2015.

1.5 - Predição computacional de interferência de SNPs em sítios alvos de miRNAs (trabalhos correlatos)

Existem já na literatura trabalhos com o intuito de identificar SNPs interferindo em sítios alvos de miRNAs. Alguns são aplicações web nas quais o usuário pode analisar SNPs de interesse, outros são apenas interfaces web para acessar dados já processados e armazenados nos bancos. Como detalhado a seguir, a maioria é dedicada a SNPs humanos. Mesmo os que possuem a possibilidade de análise em outras espécies, as análises se restringem às espécies armazenadas nos bancos, não sendo possível o usuário utilizar seus próprios dados de uma espécie diferente. Além disso, todos esses programas ou bancos tratam apenas SNPs presentes na região 3' UTR de mRNAs, sendo que praticamente metade deles se restringe a SNPs localizados na região *seed* de sítios canônicos. Essas informações são sumarizadas na Tabela 3.

Foram encontradas na literatura cinco aplicações web - miRNASNP (Gong et al., 2012; Gong et al., 2015), MicroSNiPer (Barenboim et al., 2010), mrSNP (Deveci et al., 2014), mirsnpscore (Thomas et al., 2011) e SNPinfo (Xu e Taylor, 2009) - e quatro bancos de dados de SNPs que interferem em sítios alvos de miRNAs - MirSNP (Liu et al., 2012), PolymiRTS (Bao et al., 2007; Ziebarth et al., 2012; Bhattacharya et al., 2014), miRdSNP (Bruno et al., 2012) e Patrocles (Hiard et al., 2010).

MiRNASNP (Gong et al., 2015) é um banco de dados de SNPs preditos que interferem na ligação miRNA-alvo, presentes no pré-miRNA, miRNA maduro ou no alvo do microRNA. Com o objetivo de melhorar o entendimento relacionado a microRNAs e SNPs, os autores acrescentaram as seguintes características: dados de expressão de miRNA e mRNA, oriundos do The Cancer Genome Atlas (TCGA), bem como a correlação entre eles; SNPs em alvos experimentalmente validados, sendo estes retirados do Tarbase, starBase, miRecords, miRTarBase e miR2disease; e informações sobre frequência do alelo menor (MAF) e GWAS (genome-wide association studies) relacionados aos SNPs. Além disso, foram adicionados três algoritmos online para predição de: SNPs na região 3'UTR do mRNA que alteram a interação miRNA-mRNA, SNPs na região seed do microRNA que interferem no pareamento miRNA-mRNA, e SNPs que causam impacto na estrutura do pré-miRNA. Para detectar SNPs que interferem na interação miRNA-mRNA, as ferramentas TargetScan e miRanda são utilizadas. Os resultados disponibilizados são restritos à 3'UTR de mRNAs. Se múltiplos SNPs estão presentes em um determinado microRNA ou 3'UTR, apenas um pode ser testado por vez.

Tabela 3- Ferramentas/recursos de predição do efeito de SNPs em sítios alvos de miRNAs.

Nome	Tipo de tecnologia	Região estudada	Posição do SNP	Estratégia	Espécie	Referência
MicroSNiPer	Aplicação web	3'UTR	seed	FASTA para parear *seed* do mir com a região que contém o SNP	humana, camundongo	Barenboim et al., 2010
mrSNP	Aplicação web	3'UTR	*seed* ou qualquer quando a *seed* é considerada fraca	pareamento com *seeds* fortes ou RNAhybrid para *seeds* fracas	humana, mosca, verme, galinha, peixe, rato, camundongo	Deveci et al., 2014
mirsnpscore	Aplicação web	3'UTR	qualquer	SVM duas etapas	humana	Thomas et al., 2011
miRNASNP	Aplicação web e banco de dados	3'UTR	seed, pré-miRNA	TargetScan e miRanda	humana, chimpanzé, peixe-zebra, camundongo, rato, cachorro, vaca, galinha	Gong et al., 2015
MirSNP	banco de dados	3'UTR	seed	miRanda	humana	Liu et al., 2012
PolymiRTS	banco de dados	3'UTR	*seed*	TargetScan	humana, camundongo	Bao et al., 2007; Ziebarth et al., 2012; Bhattacha

						rya et al., 2014
Patrocles	banco de dados	região promotora, sequência codificante e 3'UTR	qualquer, pré-miRNA	pesquisa, em bancos públicos, de polimorfismos em sítios alvos, pré-miRNAs e maquinaria de silenciamento	homem, camundongo, chimpanzé, rato, cachorro, vaca e galinha	Hiard et al., 2010
miRdSNP	banco de dados	3'UTR	qualquer	miRanda	humana	Bruno et al., 2012
SNPinfo	aplicação web	3'UTR	qualquer	miRanda	humana	Xu e Taylor, 2009

Fonte: Paula Prieto Oliveira, 2018.

MicroSNiPer (Barenboim et al., 2010) é uma ferramenta web que foca na análise de impacto de SNPs humanos presentes na 3' UTR, especificamente na suposta região *seed* alvo de algum miRNA. Utiliza o programa FASTA para identificar o pareamento entre o miRNA com a região onde se localiza o SNP, e exige um mínimo de seis pareamentos consecutivos na região seed. Este emparelhamento é usado como critério de predição e apenas os casos em que há alteração são mostrados. É possível predizer o impacto de até seis SNPs em conjunto presentes em uma mesma 3'UTR, por meio da construção de variantes desta sequência, com todas as possíveis combinações de alelos, que servem como entrada para o FASTA. Não é possível testar todos os pares miRNA-alvo simultaneamente. Foi enviado um email para os autores perguntando sobre essa possibilidade mas não foi obtida resposta.

MrSNP (Deveci et al., 2014) possui como entrada os SNPs e procura nas sequências 3'UTRs onde estes SNPs estão localizados. Se o SNP não é localizado em uma 3'UTR, nenhum cálculo é realizado. Se o SNP é encontrado na 3'UTR, a ferramenta retorna 79 bases da sequência que contém o SNP no centro. Esta sequência é duplicada de acordo com o número de alelos, que são colocados na posição correta. Em seguida, é checado se alguma subsequência forma pelo menos seis pareamentos consecutivos de Watson-Crick[1], começando pela posição dois da região *seed* do miRNA. É permitido um único par de *wobble*[2] para as sequências contendo de 7 a 9 pareamentos consecutivos. Se os critérios de pareamento não forem satisfeitos, é concluído que o mRNA em questão não é alvo do miRNA e nenhuma investigação adicional é realizada. Mas se o mRNA apresenta pelo menos sete pares de Watson-Crick na região *seed* do miRNA, é considerado um alvo. Para interações mais fracas, aquelas contendo 6 pareamentos consecutivos ou aquelas contendo 7 a 9 pareamentos mas com um pareamento *wobble*, é calculada a energia de ligação por meio do programa RNAhybrid de predição de alvos de miRNAs. Nessas situações, são considerados alvos aquelas interações que apresentarem energia de ligação acima de 74% da máxima. Para um determinado par SNP-miRNA, se um sítio é predito em um alelo mas não em outro, considera-se que o SNP interferiu no sítio daquele miRNA. Entretanto, o mrSNP não está mais disponível. Foi enviado um email para os autores perguntando como acessá-lo, mas houve falha na entrega da mensagem.

Mirsnpscore (Thomas et al., 2011) foca em SNPs associados a doenças. Além de predizer o efeito de SNPs em sítios alvos de miRNA, utiliza desequilíbrio de ligação para mapear esses SNPs a dados de doença de GWAS. Para analisar todos os SNPs presentes em uma determinada 3'UTR ao mesmo tempo, é construída uma sequência para cada combinação de alelos. Em seguida, os autores utilizam uma ferramenta de predição de alvo de miRNA baseada em Máquina de Vetor de Suporte (SVM) de dois passos (Saito e Sætrom, 2010), atribuindo um escore a cada sequência de alelos. Estes scores são comparados entre si para predizer SNPs candidatos a interferir na regulação gênica. Se uma mesma 3'UTR apresenta mais de um sítio candidato, é atribuído um score total para o gene alvo. O mirsnpscore apresenta um banco de dados com os SNPs preditos. A pesquisa no entanto é limitada a apenas um gene de cada vez e no máximo seis SNPs; mas há a alternativa de buscar por uma interação específica, incluindo o SNP id e o nome do miRNA.

MirSNP (Liu et al., 2012) é um banco de dados de SNPs tanto em sítios alvos preditos de miRNAs como também em genes de pré-miRNA. A predição dos alvos é realizada pelo miRanda, considerando apenas as 3'UTRs e exigindo pelo menos sete nucleotídeos pareados na

[1] Os pareamentos Watson-Crick são os pareamentos canônicos A-T e C-G.
[2] Um pareamento *wobble* é um pareamento entre as bases G-U.

região *seed* do microRNA. Também leva em consideração a conservação da região *seed*, o escore da ferramenta mirSVR dado ao sítio predito e informação sobre o MAF (*minor alele frequence*, considerando que maiores valores de MAF indicam maior relevância do SNP). Os efeitos dos SNPs são classificados em quatro grupos: criação do sítio, rompimento do sítio, aumento da interação ou redução da interação. É possível fazer uma busca individual para cada SNP ou colocar uma lista de SNPs para pesquisar todos de uma só vez.

PolymiRTS (Bao et al., 2007; Ziebarth et al., 2012; Bhattacharya et al., 2014) é um banco de dados que identifica o impacto de SNPs e *indels* na região *seed* do miRNA e no sítio alvo deste. Integra dados de alvos validados por experimentos de CLASH, além de avaliar a variação do escore de contexto, dado pelo TargetScan, causada pelo polimorfismo na interação miRNA-mRNA. Também inclui vias de sinalização biológicas para identificar o impacto dos polimorfismos no processos biológicos. Foram baixadas todas as vias de sinalização KEGG e a lista de genes de cada via é comparada com genes de sítio alvo polimórfico presentes no banco de dados PolymiRTS. Os autores também incluem dados de GWAS e eQTL e classificam o efeito dos SNP em quatro classes: rompe um alvo conservado; rompe um sítio não conservado; cria um novo alvo; “outros” para casos em que o alelo ancestral não pode ser determinado. É possível fazer uma busca individual para cada SNP ou colocar uma lista de SNPs para pesquisar todos de uma só vez.

Patrocles (Hiard et al., 2010) é um banco de dados resultante da pesquisa, em bancos públicos, de SNPs e outros polimorfismos em sítios alvos, pré-miRNAs e maquinaria de silenciamento, com o objetivo de auxiliar na identificação de polimorfismos que interferem na regulação mediada pelo miRNA. Inclui sete espécies de vertebrados (homem, camundongo, chimpanzé, rato, cachorro, vaca e galinha), além de informações de coexpressão miRNA-alvo e dados de eQTL. No entanto o banco parece ter sido descontinuado, não tendo sido encontrado nenhum síte na internet para ele.

MiRdSNP (Bruno et al., 2012) é um banco de dados que analisa a relação (distância) entre dSNPs (SNPs associados a doenças) nas 3’UTRs e sítios de miRNAs. Os alvos são obtidos por meio do TargetScan 5.1 e PicTar, mas também há dados experimentalmente validados, advindos do TarBase, miRTarBase, miRecords e miR2disease. Os autores mencionam terem realizado uma análise para a predição de novos sítios de miRNAs criados por dSNPs, utilizando o miRanda. No entanto os resultados estão apenas descritos no artigo (não estão presentes no banco de dados).

SNPinfo (Xu e Taylor, 2009) é uma ferramenta responsável pela predição de SNPs que promovem alteração da função biológica, como: estrutura e função da proteína, regulação transcricional, *splicing* do RNA e ligação ao miRNA. Para a identificação da inferência de SNPs

que interferem na interação do alvo com o miRNA, são extraídos 20 bases de cada lado do SNP, presente na região 3'UTR do gene, e procuram-se possíveis sítios de ligação para as sequências de cada alelo, utilizando o miRanda com parâmetros padrões. Foram excluídos os SNPs presentes em alvos não conservados em sequências homólogas de rato ou camundongo. É possível fazer uma busca individual para cada SNP ou colocar uma lista de SNPs para pesquisar todos de uma só vez.

Resumidamente, percebe-se que ainda se carece de uma ferramenta que seja disponível, de fácil uso (comentários sobre essa característica serão feitos no capítulo 4, na seção que descreve os testes dessas ferramentas), que seja capaz de identificar SNPs interferindo em alvos dentro ou fora da 3'UTR, sítios não necessariamente canônicos, cujos SNPs estejam presentes em qualquer região do alvo (não apenas na *seed*) e que ainda permita ser executada em qualquer espécie para a qual o usuário possua um conjunto de dados mínimo.

1.6- SIMTar: versão inicial de uma ferramenta computacional para predição de interferência de SNPs em sítios alvos de miRNAs

A ferramenta computacional SIMTar foi desenvolvida como fruto do projeto de mestrado de uma aluna de nosso grupo de pesquisa (Piovezani, 2013). SIMTar significa *SNPs Interfering in MicroRNATargets* e é uma ferramenta para a predição de SNPs interferindo em sítios alvos de miRNAs, concebida com o intuito de sanar algumas das deficiências observadas nas ferramentas citadas anteriormente com o mesmo propósito (Seção 1.5). Diferente de outros estudos com objetivos similares (Barenboim et al., 2010; Ziebarth et al., 2012; Liu et al., 2012; Gong et al., 2012; Bruno et al., 2012), SIMTar pode ser aplicado para qualquer espécie e é capaz de analisar SNPs em regiões promotoras, intrônicas, CDS e 5'UTR, não se limitando apenas à região 3'UTR. Também não se limita a SNPs na região *seed* de sítios canônicos. Além de identificar qual SNP (ou combinação de SNPs) está interferindo no sítio alvo do miRNA, indica qual a variação ou combinação alélica de SNPs vizinhos que estão interferindo no sítio alvo, e mostra o efeito predito para cada alelo: criação de um novo sítio, rompimento de um pré-existente ou alteração da estabilidade do pareamento (Piovezani, 2013).

Para a predição de sítios alvos de microRNAs ao redor dos SNPs, a versão inicial do SIMTar utilizava ferramentas já existentes de predição de alvos de miRNAs: TargetScan versão 5 (Lewis et al., 2005) e/ou Miranda (Enright et al., 2003). TargetScan permite a identificação de sítios evolutivamente conservados, enquanto Miranda permite a identificação também de sítios espécie-específicos.

A entrada do SIMTar é uma lista de SNPs fornecida pelo usuário, sendo os SNPs identificados pelo identificador no banco dbSNP[3] (denotado como rsX, rs significando *reference sequence* e sendo X um número inteiro). O SIMTar obtém, para cada SNP, sua localização genômica (nome do cromossomo e posição) e então, para cada transcrito que ocorre naquela localização[4], obtém uma sequência de 51 nucleotídeos centrada no SNP[5], como mostra a figura 6.

Para cada uma dessas sequências de 51 bases são geradas *x* sequências similares, sendo *x* = número de alelos do SNP, e cada sequência similar sendo diferente apenas na posição do SNP, em cada uma delas assumindo um diferente alelo. Tais sequências eram então analisadas pelos programas miRanda[6] e/ou TargetScan[7], com o objetivo de identificar se cada uma das sequências variantes é um sítio de miRNA.

Após a execução desses programas, o SIMTar identifica quais os SNPs que estão interferindo nos sítios verificando se, para um dado miRNA, um sítio é predito em um ou mais alelos mas não predito nos demais.

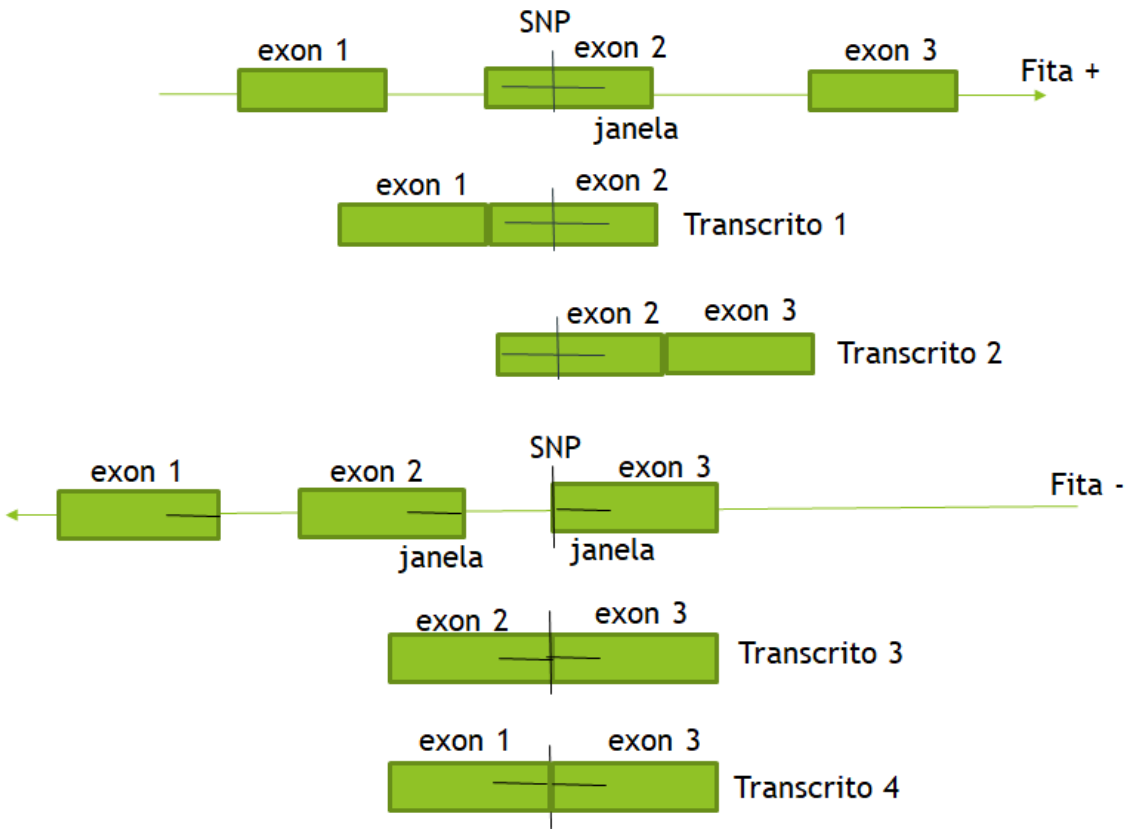
Em testes preliminares, a maioria dos SNPs testados (> 95%) foram identificados pela versão inicial do SIMTar como interferindo em sítios de miRNAs. Tais resultados motivaram o presente projeto para propor melhorias no processo de análise de forma que uma nova versão do SIMTar fornecesse resultados mais confiáveis, respaldados por outros resultados de experimentos biológicos relacionados com o problema.

1.7- Experimentos biológicos envolvendo miRNAs ou SNPs

Há atualmente vários experimentos biológicos envolvendo miRNAs ou SNPs que poderiam, se combinados com os resultados de predição do SIMTar, conferir maior confiabilidade aos resultados. Neste trabalho são considerados experimentos que validem a ligação de um miRNA a um alvo (como CLASH), CLIP, eQTL e GWAS.

[3] https://www.ncbi.nlm.nih.gov/projects/SNP/
[4] Na versão de análise de RNAs, considerada neste trabalho, apenas são considerados os transcritos nos quais o SNP localiza-se em um éxon, considerando-se tanto a fita positiva como a negativa.
[5] O tamanho desta sequência é baseado no tamanho esperado de um miRNA de um lado ou de outro do SNP.
[6] Dos resultados do programa miRanda eram selecionados os resultados com score e energia livre acima de um certo limiar dado pelo usuário (valores padrões de 90 e -17, respectivamente, definidos pelos autores da ferramenta).
[7] Como o TargetScan utiliza dados de conservação filogenética das sequências alvo, não basta apresentar a este programa a sequência variante mas o alinhamento múltiplo inter-espécies desta sequência.

Figura 6 - Cenários possíveis de localização das janelas de 51 nucleotídeos, centradas em um determinado SNP, em regiões exônicas de transcritos.



Fonte: Paula Prieto Oliveira, 2018.

1.7.1 - Cross-linking ligation and sequencing of hybrids (CLASH)

CLASH é uma técnica utilizada para descobrir sítios de ligação AGO-mRNA e interações RNA-RNA que se formam dentro do complexo AGO-mRNA (Helwak e Tollervey, 2014). Primeiramente, células vivas são preparadas para expressar PTH (protein A–tobacco etch virus (TEV) cleavage site–6×His)-AGO, e são irradiadas com UV para formar ligações covalentes entre RNA e proteína. Em seguida, as células são lisadas, PTH-AGO é purificada, RNases são adicionadas para cortar os duplexes RNA-RNA e as fitas de RNA são ligadas para formar moléculas quiméricas miRNA-mRNA. O complexo AGO-RNA é isolado e utilizado para produzir cDNA, que é sequenciado (Illumina) (Helwak e Tollervey, 2014; Felekkis e Voskarides, 2015).

1.7.2 - Cross linking immunoprecipitation (CLIP)

Cross linking immunoprecipitation (CLIP) é uma técnica de biologia molecular que utiliza luz ultravioleta (UV) e imunoprecipitação para identificar ligações entre RNAs e RBPs (RNA

binding proteins) (Wang et al., 2015). Primeiramente, o material é irradiado com luz UV para a formação de ligações covalentes entre as proteínas e os RNAs (*cross linking*). Em seguida, a amostra é lisada e colocam-se nucleases para cortar os RNAs associados às proteínas em tamanhos de 50-150 nucleotídeos (Darnell, 2012). O lisado é limpo de ribossomos e então é feita a imunoprecipitação: o material é incubado com o anticorpo contra a proteína de interesse para precipitar o antígeno (Lenz, 2004; Darnell, 2012). Após a ligação do anticorpo ao antígeno (proteína + RNA), as proteínas não ligadas ao anticorpo são lavadas. O RNA associado à proteína é desfosforilado nos extremos 3' e 5' com fosfatase, e um RNA vinculador é ligado ao extremo 3' do mesmo por meio da RNA ligase, enquanto o extremo 5' é fosforilado pelo tampão PNK na presença de ATP (Darnell, 2012). Os complexos RNAs-proteínas são isolados dos RNAs livres por meio de gel SDS (dodecil sulfato de sódio) e membrana de transferência de nitrocelulose (Ascano et al., 2012; Darnell, 2012). A digestão com proteinase K é realizada para remover a proteína do complexo RNA-proteína. Após a ligação do RNA vinculador ao extremo 5' do RNA, este é transformado em cDNA por meio da transcrição reversa. O cDNA é amplificado através do PCR, e então sequenciado (Darnell, 2012).

A argonauta (AGO) é uma proteína que participa do complexo RISC e é guiada pelo microRNA para se ligar ao seu sítio em um mRNA alvo, promovendo inibição ou até mesmo ativação deste (Wu e Belasco, 2008; Iwasaki e Tomari, 2009). Dessa forma, se os dados de CLIP apontam um sítio de ligação do mRNA à AGO, temos uma evidência de que esse RNA é um sítio de ligação de um microRNA.

1.7.3 - Expression quantitative trait loci (eQTL)

Estudos têm revelado a associação entre polimorfismos genômicos e expressão diferencial (KUDARAVALLI et al., 2009; CHEUNG e SPIELMAN, 2009). Assim, enquanto o termo QTL (*Quantitative Trait Loci)* é utilizado para indicar um loco genômico polimórfico associado a uma característica quantitativa, o termo eQTL (*expression QTL*) tem sido utilizado para indicar um loco genômico polimórfico associado à expressão diferencial de um ou mais transcritos, uma vez que a expressão de um transcrito é também considerado uma característica quantitativa (HANSEN et al., 2008).

Um eQTL pode estar próximo ou não do(s) transcrito(s) com expressão diferencial. Quando se localiza no éxon do transcrito, a variabilidade genômica desse locus pode ser apenas um marcador dessa expressão diferencial ou ser uma das causas da mesma, por exemplo alterando, criando ou rompendo um sítio de miRNA.

1.7.4 - Estudos de associação

Os GWAS (*genome-wide association studies*) são estudos de associação realizados sobre todo o genoma. Estudos de associação genética são analisam variantes genéticas comuns, geralmente SNPs, para testar a associação entre estas variantes e um fenótipo de interesse, muitas vezes com o objetivo de explicar a herdabilidade do mesmo (Bush e Moore, 2012). Cada SNP é avaliado, na maioria das vezes por regressão logística ou linear, para determinar se um alelo é significativamente associado com o fenótipo (Power et al, 2017).

O fato de um SNP estar associado a um fenótipo não implica necessariamente que ele seja a causa do mesmo, e tampouco que ele interfira no sítio alvo de um miRNA. Mas indica que ele tem relevância para aquele fenótipo e pode auxiliar o pesquisador a priorizar SNPs positivos (cuja interferência sobre um sítio de miRNA foi predita) para validação biológica por exemplo.

1.8 - Objetivos e hipóteses

O objetivo geral deste projeto foi aprimorar a predição computacional de interferência de SNPs em sítios alvos de miRNAs, com o intuito de diminuir a expectativa[8] de falsos positivos, por meio da integração de informações biológicas oriundas de distintos aspectos envolvidos no problema assim como de uma reanálise das ferramentas utilizadas para predição de sítios alvos de miRNAs. Tal aprimoramento utilizará como ponto de partida a atual versão do programa SIMTar. Lembrando-se que falsos positivos neste contexto são SNPs relatados pelo SIMTar como interferindo em sítios de miRNAs mas que, na verdade, não se localizam em sítios de miRNAs e nem criam novos sítios ou, caso se localizem em sítios, não rompam os mesmos.

Para alcançar esse objetivo geral, são propostos os seguintes objetivos específicos:

1- Inclusão de uma nova ferramenta de predição de alvos de miRNAs ou reanálise dos programas utilizados para predição de sítios de microRNAs na versão inicial do SIMTar;
2- Incluir informação de sítios alvos experimentalmente validados de miRNAs;
3- Incluir resultados de experimentos de CLIP nas predições de sítios alvos de miRNAs;
4- Incluir informação de SNPs já associados a eQTLs;
5- Incluir informação de SNPs já associados a fenótipos de interesse;
6- Comparar o SIMTar com outras ferramentas de predição de interferência de SNPs em sítios alvos de microRNAs.

[8] Utiliza-se aqui o termo “expectativa de falsos positivos” porque não tem acesso a uma amostra de SNPs sabidamente negativa. Ou seja, não há como dizer, com certeza, que um dado SNP não interfere em um sítio alvo de nenhum miRNA.

A hipótese deste projeto é que a incorporação de informações de diferentes tipos de experimentos relacionados ao problema pode diminuir a expectativa de falsos positivos sem diminuir significativamente a sensibilidade.

1.9 - Resumo dos métodos utilizados

Esta seção visa a fornecer um panorama geral dos métodos utilizados neste trabalho, que são descritos de forma mais detalhada no capítulo 3.

Foram realizadas: a) uma revisão bibliográfica exploratória para fazer um levantamento dos trabalhos correlatos, b) várias revisões bibliográficas exploratórias para identificar artigos e bancos que contivessem resultados de experimentos relacionados ao problema, c) uma revisão bibliográfica sistemática acerca dos programas de predição de sítios de miRNAs e d) uma outra revisão bibliográfica sistemática para identificar SNPs com validação experimental de interferência em sítios de miRNAs, SNPs estes que compuseram uma amostra positiva de teste para avaliação da nova versão do SIMTar e comparação com outras ferramentas.

Os artigos e bancos de dados contendo resultados de experimentos biológicos (mencionados nos objetivos específicos descritos na seção 1.7) foram analisados e processados para obtenção das informações relevantes.

Os programas de predição de sítios de miRNAs utilizados na versão inicial do SIMTar foram a princípio abandonados e o TargetScan versão 7 foi incorporado.

Para integrar todas as informações acerca de SNPs, genes, transcritos, resultados de experimentos biológicos e resultados de predições do TargetScan 7, foi modelado e implementado um banco de dados relacional.

A nova versão do SIMTar, assim como algumas das ferramentas correlatas, foram testados com uma amostra positiva de SNPs (com validação experimental de sua interferência) e com 100 amostras de SNPs aleatórios (cada amostra aleatória contendo o mesmo número de SNPs da amostra positiva). O racional é que os resultados sobre as amostras aleatórias fornecem uma estimativa do número de resultados positivos esperados simplesmente por chance, dando uma ideia, portanto, da significância do número de resultados positivos observados na amostra positiva.

1.10 - Organização deste documento

O restante deste documento está organizado da seguinte forma. O capítulo 2 traz a revisão bibliográfica sistemática acerca dos programas de predição de sítios alvos de miRNAs, como cumprimento do objetivo específico 1. O capítulo 3 detalha os materiais e métodos utilizados neste trabalho. O capítulo 4 apresenta e discute os resultados obtidos. Finalmente, o capítulo 5 traz as conclusões e trabalhos futuros. O Apêndice A traz o protocolo da revisão sistemática apresentada no capítulo 2 enquanto o Apêndice B descreve o processo e o resultado da pesquisa dos SNPs com interferência experimentalmente validada em sítios de miRNAs.

CAPÍTULO 2 - REVISÃO BIBLIOGRÁFICA SISTEMÁTICA: FERRAMENTAS DE PREDIÇÃO DE SÍTIOS ALVOS DE MICRORNAS

A revisão sistemática acerca das ferramentas de predição de alvos de miRNA foi realizada com o objetivo de rever os programas utilizados pelo SIMTar para predição de sítios de miRNAs nos vários alelos de forma a contribuir para a diminuição da taxa de falsos positivos do SIMTar.

A revisão sistemática é um levantamento da literatura de evidências sobre um determinado assunto (Biolchini et al., 2005), e possui um protocolo bem definido para condução da revisão, planejando-se previamente as palavras-chaves a serem utilizadas na buscas, as bases de dados e os critérios de inclusão e exclusão dos artigos, e documentando-se todos os passos. Logo, uma revisão sistemática pode ser mais facilmente auditada e atualizada (Moher et. al, 2009).

2.1- Métodos

O protocolo completo definido para a realização desta revisão encontra-se no Apêndice A. Resumidamente, a busca pelos artigos foi realizada no portal Pubmed em abril de 2015, utilizando as seguintes palavras-chave: (microRNA*[Title] OR miRNA*[Title]) AND target*[Title] AND (prediction[Title/Abstract] OR identification[Title/Abstract] OR detecting[Title/Abstract] OR detection[Title/Abstract]) AND (tool[Title/Abstract] OR approach[Title/Abstract] OR method[Title/Abstract] OR algorithm[Title/Abstract] OR program[Title/Abstract]). Tal pesquisa retornou 321 resultados, sobre os quais foram aplicados os critérios de inclusão e exclusão definidos no protocolo, primeiro pela leitura dos títulos e depois dos resumos. Dos 321 artigos, 31 ferramentas foram aceitas, sendo três adicionadas por se enquadrarem nos critérios e não terem aparecido na busca. Os estudos desses 34 algoritmos foram lidos na íntegra e um panorama dos mesmos é apresentado abaixo.

2.2- Análise Quantitativa

Nesta seção são apresentados alguns dados sumarizadores com o objetivo de fornecer uma visão geral sobre os algoritmos encontrados.

No total foram revisados 34 artigos relativos a 30 ferramentas distintas. A Figura 7 mostra o número de estudos publicados por ano. Percebe-se que, desde 2003 até 2015 (quando foi realizada essa revisão sistemática), artigos propondo programas de predição de sítios alvos de miRNAs vêm sendo publicados regularmente, de 2 a 5 por ano, tendo seu pico em 2013. Isso mostra o grande interesse da comunidade científica no tema e também que o problema ainda está em aberto.

Como já mencionado anteriormente, muitas são as características utilizadas pelas ferramentas de predição. A Figura 8 mostra quais são estas características e a proporção de algoritmos que requerem cada um delas. Percebe-se que a maioria dos programas (87%) usam algum tipo de alinhamento entre o miRNA e o sítio alvo e/ou informação termodinâmica (73%) relacionada à energia de ligação entre as duas moléculas. As características menos utilizadas foram as baseadas em dados de expressão gênica, números de sítios e outras, cada uma sendo requerida por apenas dois algoritmos (7%). As demais características foram usadas por 7 a 9 ferramentas, a saber conservação filogenética (27%), acessibilidade do alvo (23%), características composicionais (30%), posicionais (30%) e estruturais (27%).

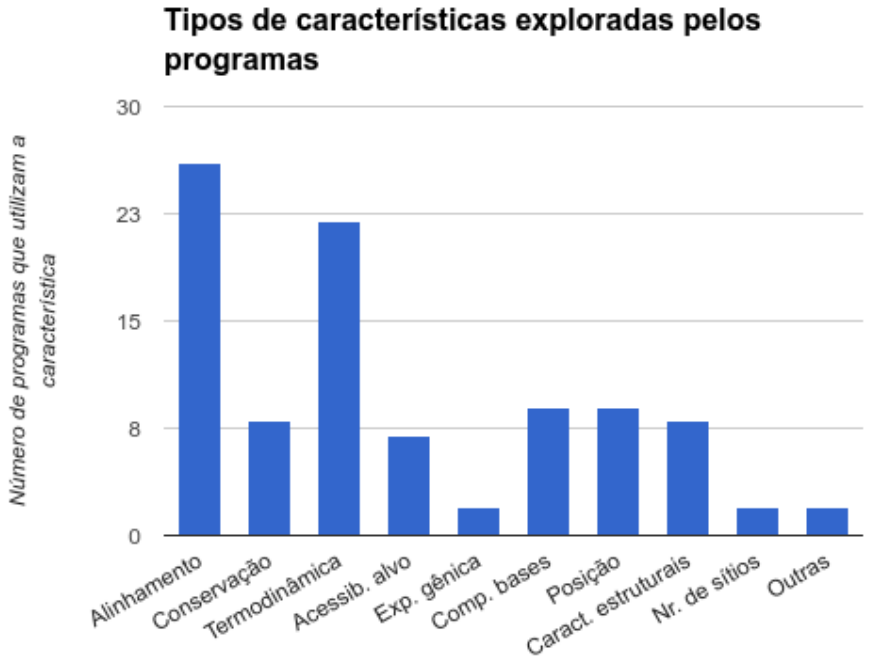
Embora se saiba atualmente que há sítios legítimos de miRNAs diferentes do canônico (revisado na Seção 1.1), mais da metade das ferramentas (63%) exigem a presença de uma *seed* canônica no alvo, e outras utilizam a informação da presença desta região de forma não obrigatória (Tabela 4). Além disso, embora esteja descrito na literatura que microRNAs se ligam não apenas à 3'UTR mas também à 5'UTR, à CDS e até ao DNA, um terço delas (33%) utilizam informações posicionais considerando que o sítio situa-se na 3' UTR. Uma delas (mirMark) se diz adaptável para outras regiões. Alguns algoritmos (Sylamer e Baymir) não requerem localização na 3' UTR, mas assumem efeito repressor do miRNA, que é o oposto do que pode acontecer quando este se liga à região promotora de genes alvos (Tabela 4).

Figura 7- Número de artigos de ferramentas de predição de alvos de miRNAs publicados

por ano.

Fonte: Paula Prieto Oliveira, 2018.

Figura 8- Principais características da interação miRNA-alvo exploradas pelos programas de predição de sítios de microRNAs.



Fonte: Paula Prieto Oliveira, 2018.

2.3- Análise Qualitativa

Nesta seção, as ferramentas revisadas são apresentadas individualmente, em ordem cronológica de publicação, com o intuito de fornecer também um panorama histórico. A Tabela 4 apresenta um resumo de suas principais características.

MiRanda

Miranda (Enright et al., 2003; John et al., 2004) foi o primeiro programa disponibilizado para predição de sítios alvos de miRNAs. Este programa recebe como entrada dois arquivos no formato FASTA: um apresentando sequências de microRNAs, e o outro com sequências nucleotídicas dos potenciais genes alvos. Trata-se de um algoritmo de programação dinâmica, semelhante ao Smith-Waterman (Smith e Waterman, 1981). Entretanto, ao invés de se basear na identidade entre os nucleotídeos, se baseia na complementaridade entre eles. São atribuídos pesos diferentes aos pareamentos para calcular o escore: +5 para G-C ou A-U, +2 para G-U e -3 para os demais pares. O algoritmo utiliza *affine gaps*: -8 para abertura do gap e -2 para a sua extensão. Além disso, os escores das primeiras 11 posições são multiplicados por um fator escalar (2.0) para refletir a assimetria 5'-3' (a região inicial tende a apresentar mais pareamentos perfeitos do que a final). Quatro regras empíricas são também aplicadas, com o início da contagem no extremo 5' do miRNA: proibição de despareamentos entre as posições 2 e 4, menos que 5 *mismatches* entre as posições 3 e 12, ao menos um despareamento entre as posições 9 e L-5 (sendo L o comprimento total do alinhamento), menos que dois despareamentos nas últimas cinco posições. Com esses parâmetros, o algoritmo de programação dinâmica otimiza o escore de complementaridade entre o miRNA e o mRNA, soma todas as posições alinhadas e encontra todos os alinhamentos não-sobrepostos em ordem decrescente de escore. O limiar do escore de alinhamento é 80. Apenas os alvos com pontuação acima deste valor são considerados. No entanto, o ponto de corte pode ser modificado pelo usuário.

Para cada interação miRNA-alvo predita, a energia livre do pareamento é calculada por meio das rotinas de dobramento do "Viena 1.3 RNA secondary structure programming library" (RNAlib) (Wuchty et al, 1999). Essa ferramenta utiliza programação dinâmica e considera a energia livre total como a soma das interações entre os pares de base (Wuchty et al, 1999). O limiar de escore de energia livre é -14. Apenas os alvos com valor abaixo desse limiar são considerados. No entanto, o ponto de corte pode ser modificado pelo usuário.

Tabela 4: Principais características relacionadas à interação miRNA-alvo utilizadas pelas ferramentas de predição de sítios de miRNAs.

Programa	Ano	Exige seed?	Assume 3' UTR?	Alinham ento	Conservação filogenética	Termodinâmica	Acessib. do alvo	Expressão gênica	Caract. composicionais	Caract. posicionais	Caract. estruturais	Nr. de sítios	Outras
miRanda	2003, 2004	canônica	não	X		X			X	X			
TargetScan	2003, 2005, 2007, 2009	canônica	sim	X	X	X			X	X			
RNAHybrid	2004	opcional	não			X							
Moving Targets	2005	canônica	sim	X		X						X	
MicroInspector	2005	canônica	não	X		X							
TargetBoost	2005	não	não	x									
miTarget	2006	não	não	X		X				X			
RNA22	2006	canônica	não	X		X							
MicroTar	2006	canônica	não	X		X							
PITA	2007	canônica	não	X		X	X						
SBM	2008	não	não	X				X					
Sylamer	2008	canônica	não	X									
FASTH	2009	canônica	não	X		X							
TargetMiner	2009	canônica	sim	X					X	X	X		
TargetSpy	2010	não	sim	x		X			X	X	X		
SVMicrO	2010	canônica	sim	X	X	X	X		X	X	X	X	
MuTIMiTar	2011	canônica	não	X					X	X	X		
(Ogul et al., 2011)	2011	não	não	X									
DIANA-microT-ANN	2012	canônica	não	X	X	X	X						
Targetprofiler	2012	não	sim	X	X	X				X			
miRTarPri	2013	não	não										X
HomoTarget	2013	canônica	não	X		X				X	X		
mirTarHunter	2013	não	não	X		X							
Baymir	2013	canônica	não	X	X			X					
MREdictor	2013	não	sim	X		X	X						X
TargetS	2014	canônica	não	X		X	X						
(Li et al., 2014)	2014	canônica	não			X			X		X		
miRMark	2014	canônica	sim	X	X	X	X		X	X	X		
MBSTAR	2015	canônica	sim	X		X			X	X	X		
MIRZA-G	2015	não	sim		X	X	X		X	X			

Fonte: Paula Prieto Oliveira, 2018

TargetScan

Na tentativa de diminuir falsos positivos, TargetScan (Lewis et al., 2003; Lewis et al., 2005; Grimson et al., 2007; Friedman et al., 2009) utiliza análise de conservação filogenética para prever sítios alvos de microRNA conservados entre múltiplas espécies. É composto por três módulos que calculam três valores: probabilidade de conservação evolutiva do alvo (pct), pontuação do contexto do sítio alvo e tipo de *seed*. A pontuação de contexto considera que o sítio está presente na região 3' UTR, e leva em consideração características do posicionamento do sítio dentro do mRNA, por exemplo proximidade do códon de terminação e proporção de bases A e U ao redor do sítio predito. É fortemente baseado na identificação de uma *seed* canônica conservada entre espécies, por isso também é chamado TargetScanS. Os tipos de *seed* identificadas são: 6mer, 7mer-A1, 7mer-m8 e 8mer.. A *seed* 6mer possui somente o pareamento de 6 bases. A região 7mer-A1 apresenta 6 pares de base (pb) e uma adenina na posição 1 do miRNA. A *seed* 7mer-m8 possui um pareamento com o alvo na posição 8, além dos 6 pb. Já a região 8mer apresenta 8 pares de bases, com adenina na posição 1 e pareamento na posição 8.

Em 2015, Agarwal e colaboradores (2015) propõem a mais nova versão do TargetScan (versão 7), baseada em novos resultados experimentais que mostram que, embora miRNAs se liguem em várias regiões dos mRNAs alvos utilizando diferentes tipos de sítios, muitos não causam a repressão da síntese proteica. Com isto, mostram que a grande maioria dos sítios funcionais são sítios canônicos presentes na 3'UTR dos mRNAs alvos. Além disso os autores analisaram várias características envolvidas no pareamento miRNA/alvo (incluindo informações de contexto ao redor do sítio, como por exemplo acessibilidade), na molécula alvo como um todo (ex: tamanho da 3'UTR) e no miRNA em questão (ex: número total de sítios em todas as 3'UTRs anotadas). As 14 características mais relacionadas com a repressão da molécula alvo foram incluídas nesta nova versão do TargetScan.

RNAhybrid

O programa RNAhybrid (Rehmsmeier et al., 2004; Kruger e Rehmsmeier, 2006) utiliza a técnica de programação dinâmica para calcular a energia livre mínima (MFE) de todas as possíveis hibridizações entre miRNAs e alvos, ou seja, considerando todas as possíveis posições iniciais no miRNA e no alvo. Loops convexos (trechos de nucleotídeos não pareados em uma das sequências) e loops internos (trechos de nucleotídeos não pareados em ambas as sequências) são restritos a um comprimento máximo constante de 15 como um valor padrão.

O usuário pode forçar um pareamento perfeito na extremidade 5’ do microRNA, por exemplo, na região *seed*. Ele pode, se desejar, definir qual parte do miRNA tem que formar uma

hélice perfeita, e apenas as estruturas que apresentem essa definição são consideradas na otimização da programação dinâmica.

Devido ao tamanho pequeno dos microRNAs, boas MFEs podem ocorrer frequentemente ao acaso. Quanto maior a sequência alvo putativa, melhores serão tais energias randômicas. Consequentemente, as boas MFEs não são significativas se elas forem resultantes de sequências longas. Por essa razão, as MFEs são normalizadas para eliminar a influência do tamanho da sequência. Seja *e* a energia livre mínima, *m* o comprimento da sequência alvo pesquisada e *n* o comprimento do miRNA, a energia negativa normalizada e_n é definida como:

$$e_n = -\frac{e}{\log(mn)}.$$

O resultado da teoria da probabilidade determina que o valor máximo de variáveis aleatórias independentes seguem uma distribuição de probabilidades conhecida como distribuição de valores extremos (EVD). Dessa forma, pode-se assumir que os valores de MFE seguem o mesmo padrão das EVDs, assumindo que os valores de energia das ligações entre miRNAs e alvos são resultado de um processo de otimização que produz um valor mínimo.

Para cada duplex miRNA-mRNA predito com uma certa MFE, calcula-se a probabilidade de tal MFE ocorrer ao acaso, ou seja, seu p-valor. O número esperado de MFEs ocorridas ao acaso, o E-valor, é o produto do p-valor e do número de sequências alvos do banco de dados. Os valores p-valor e E-valor indicam a significância estatística das MFEs normalizadas observadas. Quando estes valores são pequenos, considera-se improvável que a MFE observada seja resultante de uma complementaridade aleatória entre o microRNA e o alvo.

Os parâmetros da EVD são calculados previamente por meio do programa RNAcalibrate e fornecidos via linha de comando ao RNAhybrid. Caso não sejam fornecidos, os parâmetros são estimados a partir da energia mínima do duplex miRNA-alvo, assumindo dependência linear.

O RNAcalibrate estima os parâmetros de localização e de escala de uma EVD para um determinado microRNA, a partir de um conjunto de sequências alvos aleatórias. Esses conjuntos podem ser fornecidos como entrada ou gerados pelo próprio programa, contendo sequências com o mesmo tamanho e distribuição de um dado conjunto de possíveis alvos.

O programa RNAeffective também é utilizado e define se as sequências ortólogas são estatisticamente independentes, ou seja, utiliza-se do fato de que nem sempre as sequências relacionadas podem ser tratadas como independentes estatisticamente. Esse valor pode ser usado para analisar com mais precisão os parâmetros de energia livre mínima em uma análise entre múltiplas espécies. Com tal informação, é possível restringir a análise das predições realizadas pelo RNAhybrid, concentrando apenas os alvos que são conservados filogeneticamente.

MovingTargets

A ferramenta MovingTargets (Burgler e Macdonald, 2005), específica para alvos em 3'UTRs de *D. melanogaster* e *D. pseudoobscura*, apresenta duas etapas: a criação de um banco de dados dos alvos potenciais e avaliação de todos os pares miRNA/mRNA possíveis em relação a restrições biológicas sugeridas pela análise de interações conhecidas.

A seleção de sequências para o banco de dados foi guiada pela compreensão das ações e características dos miRNAs e seus alvos. O banco de dados foi limitado a sequências 3'UTRs, que foram divididas em segmentos menores que 50 nucleotídeos. Cada segmento foi testado para descobrir se é alvo ou não.

O MovingTargets aplica cinco restrições biológicas para todos os possíveis alinhamentos de cada miRNA com as sequências dos bancos de dados, produzindo um conjunto de alvos preditos. As restrições são: mínimo de três sítios alvos no mRNA, máxima energia livre de hibridização de -15 kcal/mole, mínimo de sete pareamentos consecutivos no extremo 5' do miRNA, e máximo de um par G:U na região 5' do miRNA.

Os autores ranquearam potenciais interações miRNA/target de acordo com a força de hibridização entre eles, medida pela energia livre de ligação. Esta foi predita por meio do sofware DINAMelt Server (Markham e Zuker, 2005).

As sequências alvos potenciais do banco de dados correspondem a 3'UTRs de D. melanogaster e sequências conservadas de D. pseudoobscura. As primeiras foram obtidas do *Drosophila Genome Project* http://www.fruitfly.org. Já as últimas foram retiradas do *Berkeley Genome Pipeline*.

MicroInspector

O MicroInspector (Rusinov et al., 2005), para um dado par miRNA/mRNA, escaneia a sequência de mRNA simultânea e independentemente por duas janelas de seis nucleotídeos. A primeira janela representa os nucleotídeos 1-6 (a partir do extremo 5' do miRNA), e a segunda nucleotídeos 2-7. Elas deslizam na sequência alvo por passos de um nucleotídeo e o programa realiza uma análise de complementaridade.

A complementaridade pré-filtro procura em cada uma das duas janelas por domínios com cinco pareamentos de Watson-Crick ou quatro de Watson-Crick e um G:U. Se nenhuma das duas janelas cumprirem esse requerimento, os dados são ignorados e as janelas se movem um nucleotídeo em direção ao extremo 5' do mRNA. Quando a análise de sequência identifica pelo menos uma das situações descritas acima, o programa inicia uma avaliação detalhada desse sítio, extraindo uma sequência de 32 nucleotídeos de mRNA. Posteriormente, o par miRNA-mRNA é submetido a um algoritmo para cálculo da energia livre mínima de ligação entre as duas moléculas.

Para calcular as propriedades termodinâmicas do duplex miRNA-mRNA predito, foram integradas algumas rotinas do *Vienna RNA secondary structure programming library* (RNAlib) do pacote Vienna RNA versão 1.5 (Hofacker et al, 1994; Hofacker, 2003). Essa análise calculará a energia livre e a estrutura secundária da interação miRNA-mRNA. Hits abaixo do limiar selecionado de energia livre (valor padrão = -20 kcal/mol) são submetidos à análise pós-filtro.

A análise pós-filtro inspeciona a estrutura miRNA-mRNA e elimina qualquer sítio caracterizado por dois nucleotídeos não pareados no lado 5' ou 3' do miRNA. O filtro também exclui estruturas com baixos valores de energia de dobramento que são resultantes da auto-complementaridade em uma das duas fitas de RNA.

TargetBoost

O TargetBoost (Saetrom et al., 2005) combina programação genética e boosting para criar o classificador. A programação genética desenvolve padrões de sequência individuais que possam diferenciar melhor entre alvos e não alvos, a partir dos dados de treinamento. Os padrões de sequência são expressões gerais que descrevem as propriedades comuns dos sítios de ligação ao miRNA. Estas expressões gerais são traduzidas em queries específicas para cada miRNA. Então, o algoritmo boosting atribui pesos para os padrões de sequência baseado no desempenho dos mesmos no conjunto de treinamento. Cada padrão responde sim (1) ou não (-1) caso seja um alvo ou não. O classificador final é a média dos vários classificadores. Essa ferramenta não utiliza complementariedade de sequência, estabilidade termodinâmica ou conservação evolutiva.

O TargetBoost foi comparado ao Nucleus e ao RNAHybrid. Avaliando as curvas ROC de validação cruzada 10-fold, O TargetBoost apresentou melhor desempenho que os demais e maior sensibilidade.

miTarget

O miTarget (Kim et al., 2006) é um classificador fundamentado em Máquina de Vetores de Suporte (SVM) e utiliza função kernel de base radial como medida de similaridade para características estruturais, termodinâmicas e posicionais.

Todas as características são baseadas na estrutura secundária do RNA predita pelo programa RNAfold (Hofacker, 2003). Este necessita de uma sequência única e linear de RNA como entrada. Por essa razão, o extremo 3' do mRNA alvo é conectado ao extremo 5' do miRNA através de uma sequência ligante "LLLLLL". As posições no alinhamento são numeradas a partir da posição mais 5' da região *seed*. Alinhamentos são estendidos até a vigésima posição e as posições restantes são descartadas.

Para as características estruturais e termodinâmicas, os autores dividiram o alinhamento secundário em três partes: 3', 5' e alinhamento total. Cada valor de contagem de matches, mismatches, pareamentos GC, pareamentos AU, pareamentos GU e outros mismatches das três partes são considerados como característica estrutural. Os valores de energia livre das partes 5', 3' e de todo o alinhamento miRNA-mRNA são características termodinâmicas e são calculadas pelo RNAfold.

Em relação às características baseadas em posição, Doench et al (2004) e Brennecke et al (2005) observaram que uma mutação pontual única poderia inibir a função do miRNA dependendo da posição da mesma. Características baseadas em posição corresponderam a mutações pontuais nos dois estudos. Cada posição tinha um dos quatro valores nominais: pareamento G:C, pareamento A:U, pareamento G:U e mismatch. Com o objetivo de tornar esses valores disponíveis para o SVMlight, os autores do miTarget traduziram-nos em valores de um a quatro, respectivamente, e os normalizaram.

RNA22

O RNA22 (Miranda et al., 2006) não se baseia em conservação filogenética. Primeiramente, entradas duplicadas de miRNAs maduros presentes no Release 3.0 (janeiro, 2004) do RFAM foram removidas, e o restante utilizado pelo algoritmo Teiresias para descobrir padrões (motifs de comprimento variado) nas sequências de microRNAs maduros. Esses padrões devem apresentar um tamanho mínimo de quatro nucleotídeos e pelo menos 30% das posições especificadas, além de aparecerem pelo menos duas vezes em 354 miRNAs. Em seguida, procuram-se sítios complementares reversos dos padrões dentro do mRNA de interesse e determinam-se sítios com múltiplos padrões alinhados (chamados de "hot spots").

A partir de dados genômicos atuais, utiliza-se uma cadeia de Markov de segunda-ordem para estimar a significância estatística de cada padrão, descartando aqueles com log de probabilidade >= -38.

Os autores utilizam o termo "ilha alvo" para se referir a uma região de 36 nucleotídeos que contenha pelo menos 30 padrões. Cada microRNA é pareado com cada ilha alvo gerada e um vinculador GCGGGACGC é inserido entre as duas sequências. O Vienna package é então utilizado para predizer a estrutura e a energia livre de Gibbs do duplex.

Três características, definidas pelo usuário, são consideradas: o número mínimo de pareamentos de base entre o microRNA e o alvo (excluindo os pares de base no vinculador), número máximo de bases não pareadas permitidas na região *seed*, e a energia livre máxima permitida. O mRNA é predito como alvo se todas essas condições forem satisfeitas. Não há

restrições em relação ao número de pareamentos G:U que possam estar presentes na região seed do híbrido.

MicroTar

O MicroTar (Thadani e Tammi, 2006) se baseia na identificação de *seeds* canônicas e na diferença da energia livre entre o mRNA ligado ao miRNA e o mRNA não ligado. Primeiramente, o algoritmo calcula a energia livre mínima de cada mRNA. Paralelamente, procura por sítios *seed* de um dado miRNA e, para cada um deles, recalcula a energia livre mínima do mRNA exigindo que a *seed* esteja pareada com sua respectiva região complementar no alvo. A saída é uma lista de duplexes putativos nos quais estes são mais estáveis que o mRNA livre, bem como imagens da estrutura secundária do mRNA ligado e não ligado. Este resultado é submetido a uma análise estatística para determinar a significância de cada pareamento miRNA-mRNA.

O microTar foi comparado ao PicTar e apresentou maior sensibilidade.

PITA

PITA (Kertesz et al., 2007) significa "*Probability of Interaction by Target Accessibility*" (Probabilidade de interação por acessibilidade de alvo). A ferramenta percorre cada mRNA alvo, procurando sítios potencialmente favoráveis à ligação de miRNAs. Em seguida, filtra os sítios que apresentam pareamento quase perfeito com a *seed* do microRNA. Nenhum não pareamento ou loops são permitidos, e apenas um par de bases G-U é aceito no 7- ou 8-mers.

Com o objetivo de medir a força de repressão do miRNA sobre o alvo, o software calcula a pontuação de energia da interação mRNA-miRNA: ΔΔG= ΔGduplex – ΔGopen. O ΔGduplex é a energia livre adquirida da ligação entre o microRNA e o mRNA, ou seja, a energia livre da estrutura duplex miRNA-mRNA. Já o ΔGopen corresponde à energia livre necessária para desfazer essa ligação.

Na maioria dos casos experimentais analisados pelos autores da ferramenta, as sequências adjacentes ao alvo mostram fortes interações de pareamento com o microRNA, devido ao RISC. Por essa razão, o cálculo do ΔGopen foi modificado para incluir os custos da quebra de ligação entre as bases adjacentes ao alvo e o microRNA. Sendo assim, são considerados 70 nucleotídeos a mais de cada lado do alvo no cálculo do ΔGopen. O valor de 70 nucleotídeos foi escolhido para reduzir a complexidade das computações, e pelo fato de que há uma menor probabilidade de interações de pareamento de bases da estrutura secundária entre nucleotídeos que são separados por mais de 70 nucleotídeos.

SBM

Stacking Binding Matrix (SBM) (Moxon et al., 2008) incorpora todos os alvos conhecidos de um dado miRNA em uma busca por alvos adicionais. Este algoritmo captura aspectos dos mecanismos de ligação específicos extraindo informações particulares de um conjunto de alvos validados. Isso torna o método genérico e ele pode ser aplicado para qualquer organismo sem nenhum conhecimento prévio de mecanismos específicos de reconhecimento de alvo.

O SBM é computado de um alinhamento múltiplo de sequências que consiste do reverso complementar do miRNA em questão, juntamente com alvos conhecidos, utilizando um pacote de alinhamento ClustalW (Chenna et al, 2003). A matriz resultante é então usada para escanear e pontuar um conjunto de sítios de ligação potenciais. Sequências apresentando score acima de um limiar definido pelo usuário (valor padrão 1) são consideradas como alvos.

Sylamer

Sylamer (van Dongen et al., 2008) é um método de detecção de alvos de miRNA a partir de dados de expressão. A entrada é uma lista ranqueada de genes super ou sub-regulados de acordo com experimentos de miRNA.

Os autores aplicaram Sylamer a palavras complementares à *seed* dos microRNAs nas regiões 3'UTRs dos genes. Então, se o enriquecimento das palavras *seed* nas 3'UTRs correlaciona com o ranking de genes de acordo com a mudança deles durante o experimento de microRNA, parte das alterações de expressão podem ser atribuídas a efeitos diretos.

Sylamer rapidamente acessa a super- e sub-representação de palavras de nucleotídeos de tamanho específico em lista de genes ranqueados e testa esta utilizando vários pontos de corte. Para cada ponto de corte, é avaliado se uma determinada palavra é mais ou menos abundante no topo da lista que o esperado quando comparado ao resto da mesma. Isso é feito até o ponto de corte incluir o conjunto inteiro de sequências a ser analisado. A significância é calculada assumindo uma distribuição hipergeométrica.

O Sylamer é utilizado primeiro para estabelecer se qualquer miRNA tem um efeito significativo e para escolher um limiar apropriado. Se um claro pico de enriquecimento é encontrado perto do início da lista de genes ranqueados, resultados para hexâmeros, heptâmeros e octâmeros deveriam ser comparados e a forma das curvas e a localização dos picos deveriam coincidir aproximadamente. O pico mais próximo do início do ranking pode ser escolhido como um limiar conservador. Acima desse limiar, uma lista de genes cujas sequências contém pareamento de palavra apropriada com um miRNA específico é considerada como um conjunto de alvos candidatos, suportada por dados de expressão.

As 3'UTRs foram obtidas por meio do mapeamento do conjunto de sondas IDs para identificadores transcritos RefSeq, utilizando arquivos de anotação provenientes do Affymetrix. Quando o conjunto de sondas não mapeou diretamente as sequências RefSeq, ou não tem uma 3'UTR, os autores pegaram as 3'UTRs do maior transcrito anotado Ensembl.

FASTH

FASTH (Ragan et al., 2009) pode ser usado tanto para DNA como para RNA, e utiliza uma estratégia de busca semelhante à do FASTA. Em um primeiro estágio, sítios putativos (alvos) dos miRNAs (queries), são selecionados e ranqueados utilizando energia livre de ligação como critério. Após dois passos de pré-processamento, o complemento da sequência do miRNA é localizado no banco de dados de mRNAs, e a energia livre mínima é calculada por meio do programa hybrid-min no pacote UNAfold, considerando: energia livre do par de base e do par de base empilhado em hélices perfeitas, contribuições desfavoráveis de energia do interior dos loops e protuberâncias, contribuições favoráveis de bases de cadeias simples ou *mismatches* adjacentes a pares de bases, e energia de iniciação.

O FASTH compara palavras do reverso complementar de cada query com palavras do banco de dados, utilizando o alfabeto de duas letras R,Y, que fornece pares WC, GU e AC. Pares WC e GU são aceitos, enquanto os AC são descartados.

Dada uma query, FASTH computa todas as palavras no reverso complementar. No alfabeto de duas letras R,Y, todo pareamento exato no banco de dados dessas palavras é rapidamente tabulado. Se w é o tamanho da palavra, então um pareamento perfeito significa que há um índice j no banco de dados e um índice i na query, tal que j, j+1,..., j+w-1 no banco de dados são potencialmente complementares às bases i, i-1,..., i-w+1 na query, respectivamente. É dito que o match ocorre na diagonal i+j. O FASTH varre essas diagonais em que pelo menos c pareamentos ocorrem, sendo c um valor de corte definido pelo usuário. Todas as outras posições no banco de dados são ignoradas.

Para cada diagonal d que é varrida, o FASTH considera a hibridização de diagonais vizinhas d-b, d-b+1, …, d, d+1, …, d+f em que b é o "parâmetro de busca backward" e f é o "parâmetro de busca forward". Um procedimento ad hoc é utilizado para determinar a hibridização inicial e a energia livre correspondente. Os pares de base encontrados nas diagonais adjacentes são adicionados à hibridização se eles não entram em conflito com pares de base existentes. A adição desses pares de base extra introduz pequenas protuberâncias ou loops interiores. O escore da hibridização resultante é a soma das energias livres das hélices resultantes. A energia dos pares de base e do empilhamento dos pares de base são consideradas, mas loops e mismatches são ignorados.

Em um segundo estágio da ferramenta, ocorre a identificação de sítios de ligação do miRNA por meio da filtragem da lista inicial de alvos, aplicando critérios baseados em informação biológica sobre alvos validados:

- pareamento WC perfeito na *seed*, ou apenas um par GU
- emparelhamento WC na região 3' do miRNA
- *score* de energia livre do duplex complementar reverso, com pareamento WC perfeito entre o mRNA e todo o comprimento do miRNA. Um limiar de energia livre de 40% em relação ao melhor score observado é aplicado em todos os casos.

Para cada microRNA, alvos potenciais são então ranqueados pelo valor de energia livre.

TargetMiner

TargetMiner (Bandyopadhyay e Mitra, 2009) é um classificador SVM treinado com função base radial kernel (RBF). Um conjunto de 90 características relevantes contexto-específicas da interação miRNA-alvo são extraídas. Os autores dividiram o miRNA em regiões *seed* (posições 1-8) e *out-seed* (parte restante). O sítio de pareamento com a *seed* é categorizado em 6mer/ 7mer-A1/ 7mer-m8/ 8mer. As características computadas são:

- Frequência de nucleotídeos únicos (A, C, U, G) e dinucleotídeos na regiões *seed* e *out-seed* flanqueadora (30 nucleotídeos *upstream* e 30 nucleotídeos *downstream*) do alvo;
- Frequência dos seis tipos de pares de base (A:U, U:A, C:G, G:C, G:U e U:G);
- Número de sítios de pareamento com a *seed* "efetivos", ou seja, que não residem a menos de 15 nucleotídeos do códon de parada nem próximo do meio da 3'UTR;
- Composição de bases da região flanqueadora da *seed*: é atribuído um valor de 1 se o conteúdo de AU for maior que 60%, caso contrário o valor é 0;
- Para cada interação miRNA-alvo efetiva, é verificado se um pareamento de Watson-Crick está presente nas posições 13-16 do microRNA. Se sim, o valor é um, caso contrário o valor é 0.
- Frequência de dois pareamentos de base consecutivos na região *seed*, como A:U-G:C (frequência do emparelhamento G:C imediatamente após A:U). São 32 características de pareamento consecutivo ao invés de 36, pois apenas um par G:U é permitido.

Os pares miRNA-mRNA positivos foram retirados do miRecords. Um conjunto de 59 exemplos negativos de homem, rato e Drosophila foram extraídos da última versão do TarBase. Foram utilizados pares miRNA-mRNA humanos preditos por uma ou mais das seguintes ferramentas: miRanda, TargetScanS, PicTar e DIANA-microT. As sequências de miRNA maduros humanos foram coletadas do miRBase. As 3'UTRs humanas foram extraídas de University of California, Santa Cruz (UCSC) Genome Bioinformatics site (http://genome.ucsc.edu).

Para detectar exemplos negativos potenciais, os autores selecionaram um conjunto de todos os alvos de miRNA preditos por uma ou mais ferramentas computacionais descritas acima. Com o objetivo de identificar tais não-alvos potenciais, foram utilizados dados de expressão de um miRNA e de seus alvos e medida a especificidade tecidual de ambos. Um par miRNA-mRNA foi considerado tecido-específico caso o perfil de expressão fosse consideravelmente diferente em no máximo dois tecidos comparado aos outros. Se um par miRNA-mRNA era significativamente super-expresso no mesmo tipo de tecido específico, esse par foi escolhido como exemplo potencial negativo.

Os não alvos potenciais foram adicionalmente testados contra outro conjunto de dados de expressão gênica, utilizado por Johnson et al (2003). Para este conjunto de dados, os autores ranquearam os níveis de expressão de cada gene não alvo potencial extraído do conjunto de dados anterior nos tecidos de interesse. O ranking de tal gene é medido em um tipo de tecido específico em que o miRNA e esse gene estão super-expressos. Se o rank de tal gene em um tipo de tecido específico é maior que o rank da média, significa que o gene está super-expresso, caso contrário está sub-expresso. Os autores reduziram a lista de genes não alvo potenciais eliminando aqueles que não estavam super-expressos em um tecido específico.

Para a seleção de não alvos com alta confiança, foi fixado um limiar de escore da energia de interação miRNA-mRNA, $\Delta\Delta G > 0$ kcal/mol. O limiar do escore de conservação definido foi <0.5. Para a validação dos não alvos potenciais, foi utilizado o conjunto de dados pSILAC.

O TargetMiner apresentou maiores Mathew's correlation coeficient (MCC), average classwise accuracy (ACA) e balanço entre sensibilidade e especificidade em relação a outras ferramentas (DIANA-microT, miRanda, TargetScan, PicTar, MicroInspector, MirTarget2, NBmiRTar, PITA, RNA22, RNAhybrid).

TargetSpy

O TargetSpy (Sturm et al., 2010) utiliza aprendizagem de máquina, por meio da técnica Multiboost com *decision stumps* como base de aprendizado, para predizer os alvos de microRNAs. O Multiboost é um método que combina vários classificadores, chamados de classificadores fracos, para gerar um único classificador, chamado de classificador forte (Webb, 2000; Chaves, 2012). O classificador fraco utilizado nessa ferramenta é o *decision stumps*, que consiste de uma árvore de decisão de um nível, com um nó raiz e dois nós folhas. O *decision stumps* realiza a predição baseado em apenas uma característica e é também chamado de 1-rule (Holte, 1993; Chaves, 2012). A técnica utilizada para a seleção de características do classificador é a CFS (*Correlation-based Feature Selection*), juntamente com o algoritmo *best-first search*. Foi obtido um conjunto de sete características: compacidade (valor médio entre o número de

pareamentos/tamanho do microRNA e o número de pareamentos/ tamanho do sítio alvo); razão do conteúdo G+C entre microRNA e alvo; comprimento do trecho mais longo de emparelhamentos consecutivos em qualquer lugar do duplex; assimetria de ligação (razão entre a quantidade de pares de base na 3' versus a região 5' do microRNA, considerando oito nucleotídeos de cada lado); conteúdo G+C do alvo; número de pareamentos da seed 8-mer do microRNA; e a posição do sítio de ligação na 3'UTR. Essa ferramenta não utiliza pareamento com a *seed* nem conservação evolutiva, e é capaz de prever sítios espécie-específicos e 3' compensatórios.

Essa ferramenta apresenta dois arquivos multifasta como entrada: um contendo sequências 3'UTRs e outro com sequências de microRNAs maduras. Os microRNAs são colocados contra as 3'UTRs para gerar zonas candidatas (zonas de elevada atração entre o microRNA e o mRNA alvo), da seguinte forma:

1- para um par miRNA-mRNA, todas as possíveis estruturas duplex preditas pelo RNAduplex são ordenadas de acordo com a posição da âncora delas. A âncora de cada híbrido é o primeiro nucleotídeo do alvo que pareia com a extremidade 5' do microRNA. Os duplexes que apresentam a mesma posição de âncora são agrupados;

2- Os duplexes com menor energia livre em cada grupo são selecionados. A atração de áreas individuais do mRNA em relação a um determinado miRNA foi medida em termos de valores da energia livre de Gibbs;

3- As áreas do mRNA com forte atração pelo miRNA, chamadas zonas candidatas, são identificadas com base na exigência de que todos os duplexes preditos possuam valor de energia abaixo de um limiar x, e pelo menos um deles, chamado de representante, possua valor de energia abaixo de um limiar y;

4- Para cada zona candidata, o duplex energeticamente mais favorável que apresente pareamento de base dentro dos dois primeiros nucleotídeos no extremo 5' do microRNA é selecionado como representante.

5- É calculado um escore dado pelo classificador para cada representante, funde-se as zonas candidatas sobrepostas e ranqueia-se as predições de acordo com o escore delas.

Com o objetivo de obter as melhores predições, os autores criaram um subconjunto com elevada sensibilidade e alta especificidade. Os sítios alvos que apresentaram taxa de falso-positivo menor que 5% (como avaliado na validação cruzada 10-fold) foram atribuídos ao subconjunto sensível, e aqueles com taxa de falso-positivo de 1% ou menos ao subconjunto específico.

Com o objetivo de comparar o TargetSpy com outros métodos, os algoritmos foram divididos em três classes: I- sem pareamento com a seed, sem conservação; II- requer pareamento com a seed; III- requer pareamento com a seed e conservação.

A avaliação dos dados experimentais de mosca sugerem que o TargetSpy tem um desempenho tão bom quanto os algoritmos correntes de última geração, quando aplicado o critério de pareamento com a seed. Além disso, a predição sem seed é notavelmente melhor que a do RNA22, a outra ferramenta testada que não requer pareamento perfeito com a seed.

SVMicrO

SVMicrO (Liu et al., 2010) apresenta uma estrutura de dois estágios, incluindo site-SVM e UTR-SVM, e se baseia em 21 características site e 18 características UTR para fazer a predição.

Primeiramente, um filtro é aplicado, que utliliza a sequência de miRNA para escanear ao longo da 3'UTR para buscar sítios de ligação possíveis. Além disso, regiões da 3'UTR que obedecem uma das regras de interação com a *seed* são consideradas como alvos potenciais:

- Mais de quatro pareamentos WC contínuos;
- Mais de cinco emparelhamentos contínuos (incluindo pares GU) e mais que dois contínuos WC;
- Mais de seis pareamentos no total e três contínuos WC, mas nenhum gap é permitido;
- Emparelhamentos WC nos nucleotídeos 2 a 4 do miRNA, mais que três WC e quatro no total, mas nenhum gap é permitido;
- Mais que cinco pareamentos e cinco WC, é permitido apenas um gap na sequência de miRNA ou na 3'UTR.

Em seguida, os alvos potenciais identificados pelo filtro são submetidos ao site-SVM, que extrai características de cada sítio e atribui um escore para indicar a confiança de predição do alvo como um sítio verdadeiro. Finalmente, os escores dos alvos, juntamente com outras características UTR, são considerados pelo UTR-SVM pra produzir uma predição final desta região como um sítio de ligação.

O conjunto de dados positivos veio do miRecords, enquanto o conjunto de dados negativos foi baseado em 20 microarranjos do NCBI Gene Expression Omnibus, cada um gerado pela super-expressão de um miRNA diferente. Os autores assumiram que os alvos negativos são menos prováveis de estarem com baixa expressão diante da super-expressão do miRNA.

O sítio de ligação foi dividido em duas subregiões, que são a região *seed* e a 3'. Apenas os 20 primeiros nucleotídeos são considerados na extração de características.

Os grupos de características site são: pareamento perfeito com a *seed* (6mer, 7mer-A1, 7mer-m1, 7mer-m8, 8mer-A1, 8mer-m8); estrutura de ligação de pares: foi utilizado o miRNAbind, que é uma versão modificada do RNAduplex, para gerar a estrutura secundária da ligação do miRNA; estrutura de ligação regional: para cada região, foram considerados os números de pares

WC, pares GU, despareamentos e gaps como características regionais, os números de estruturas protuberantes e nucleotídeos protuberantes em cada região de ligação também foram considerados; conservação; energia de ligação da *seed* e energia de acessibilidade; características da região contexto da *seed*, que corresponde a duas sequências de 10 nucleotídeos em ambos os finais da seed; localização do sítio potencial em relação ao códon de parada e à 3'UTR.

Os grupos de características UTR são: comprimento da 3'UTR, densidade do alvo e escore do sítio de ligação gerado pelo site-SVM.

O algoritmo mínima redundância máxima relevância (mRMR) (Ding e Peng, 2005) é utlilizado para selecionar as características. As 21 características site são:

1. pareamento da seed 6mer
2. escore de conservação da região contexto 3'
3. número de pareamentos na região *seed*
4. pareamento da *seed* 7merA1
5. pareamento da *seed* 7merm8
6. pareamento da *seed* 7merm1
7. pareamento da *seed* 8merA1
8. energia de acessibilidade
9. pareamento da *seed* 8merm1
10. statuts do sexto 2mer
11. número de pareamentos na região total
12. energia de ligação da região *seed*
13. escore de conservação da *seed*
14. status do pareamento da posição 7
15. tipo de nucleotídeo da primeira posição da região contexto 5'
16. energia de ligação da região total
17. escore de conservação da região contexto 5'
18. número de *mismatches* na região *seed*
19. status do pareamento da posição 5
20. status do pareamento da posição 2
21. status do pareamento da posição 12

As 18 características UTR são:

1. escore do sítio máximo
2. total de escore positivo
3. número de sítios positivos

4. número máximo de sítios positivos dentro de 100 nucleotídeos
5. densidade dos sítios positivos
6. número de sítios potenciais com 8merA1
7. número de sítios positivos com 8merA1
8. escore máximo com 8merA1
9. número de sítios potenciais com 8merm1
10. número de sítios positivos com 8merm1
11. escore máximo com 8merm1
12. escore máximo da 7merm1
13. escore máximo com 7merA1
14. escore máximo com 6mer
15. escore máximo sem a *seed* perfeita
16. número de sítios potenciais com 7merA1
17. número de sítios positivos com 7merA1
18. comprimento da UTR

O desempenho do SVMicrO foi comparado ao do TargetScan, miRanda, MirTarget, PicTar e PITA. O SVMicro apresentou maior AUC, com taxa de verdadeiros positivos mais alta, especialmente para taxa de falso positivo baixa, e melhor precisão, especificidade e sensibilidade.

MulTiMitar

MulTiMitar (Mitra e Bandyopadhyay, 2011) utiliza o AMOSA (Arquived Multi-objective Simulating Annealing)[9] integrado ao SVM, o AMOSA-SVM, que extrai um conjunto de características não redundantes para aumentar o poder preditivo do classificador.

A sequência de miRNA foi dividida em *seed* (nucleotídeos de 1 a 8) e *outer-seed* (parte restante). Os tipos de *seed* foram categorizados em 6mer, 7mer-A1, 7mer-m8 e 8mer. O MulTiMitar procura primeiro regiões complementares *seed* 6mer na 3'UTR do mRNA. Um único par GU é considerado, se presente no sítio de ligação à *seed*.

Para a interação miRNA-alvo, um pareamento de Watson-Crick adicional pode estar presente nas posições 13-16 do miRNA, que pode ser estendido dos nuceotídeos 12 a 17. Se isso ocorrer, o valor da característica correspondente é 1, caso contrário é 0. Os autores também avaliaram se o sítio *seed* reside preferencialmente em uma região rica em AU ou não. Para tal, foi considerada a composição de bases da região flanqueadora à *seed upstream* e *downstream* (30

[9] AMOSA é um algoritmo de otimização multi-objetivo baseado em *simulated annealing*, em que várias soluções ótimas igualmente importantes podem existir (Bandyopadhyay et al, 2008).

nucleotídeos cada). Foi atribuído o valor 1 para o conteúdo da AU acima de 60%, caso contrário o valor é 0. Outras características incluídas são:

- frequência de cada nucleotídeo no sítio de pareamento com a *seed*;
- frequência de cada nucleotídeo no sítio de pareamento *outer-seed*;
- frequência de di-nucleotídeos no sítio de pareamento com a *seed*;
- frequência de di-nucleotídeos no sítio de pareamento *outer-seed*;
- características da interação miRNA-mRNA na região *seed* (frequência dos possíveis pares de base);
- frequência dos possíveis pareamentos di-nucleotídeos da interação miRNA-mRNA, como: AU-AU, AU-CG, etc.

Comparado a outros métodos, o MulTiMitar apresentou maior taxa de verdadeiros positivos e menor de falsos positivos que o TargetMiner, TargetSpy, TargetScan, NBMiRTar, PicTar, miRanda, RNAhybrid, Micro Inspector, RNA22 e MirTarget2 . Além disso, apresentou maior coeficiente de correlação de Mathews (MCC) e average class-wise accuracy (ACA) que TargetMiner, PITA, miRanda, TargetScan, TargetSpy, MicroInspector, MirTarget2, Pictar, NBmiRTar, RNAhybrid, DIANA Micro T 3.0 e RNA22. Para avaliar a sensibilidade e especificidade do MulTiMitar em relação ao TargetMiner, a área sob a curva ROC (AUC) foi calculada. O MulTiMitar apresentou maior AUC. Considerando o conjunto de dados pSILAC, o MulTiMitar apresentou um resultado balanceado em termos de precisão e recall, enquanto as outras ferramentas (RNA22, PITA, miRanda, EIMMO, TargetScanS, Pictar e TargetSpy) sofreram de baixa precisão ou baixo recall.

A probabilistic approach to microRNA-target binding

Ogul e colaboradores (2011) definiram um modelo probabilístico que realiza um alinhamento complementar entre o miRNA e o mRNA. Este emparelhamento é transformado em uma nova sequência definida por letras do alfabeto, que representam os diferentes tipos de pareamentos de nucleotídeos, incluindo mismatches e gaps. O alinhamento é realizado por meio do algoritmo de programação dinâmica, com penalidade de -1 para mismatches e gaps. A probabilidade do duplex miRNA-mRNA é analisada utilizando a Cadeia de Markov de Comprimento Variável. Este modelo avalia o conteúdo da sequência baseado na ordem dos arranjos locais, quantificando a probabilidade de ocorrência de um símbolo específico após uma certa sub-sequência, com variação de comprimento menor que o máximo pré-definido. Tal processo permite calcular a probabilidade de toda a sequência multiplicando as probabilidades locais.

Nos testes realizados pelos autores, esse método apresentou melhor desempenho que RNAhybrid e Mirtif, em termos de sensibilidade, acurácia e área abaixo da curva ROC, com especificidade semelhante à o Mirtif e maior que a do RNAhybrid.

DIANA-microT-ANN

DIANA-microT-ANN (Reczko et al., 2012) combina uma rede neural artifical com características dos alvos de microRNA (tipo de ligação, energia termodinâmica mínima, conservação evolutiva e acessibilidade estrutural).

O programa identifica os sítios de ligação putativos por meio de um algoritmo de programação dinâmica, que seleciona o melhor alinhamento entre a região *seed* estendida (nucleotídeos de 1 a 9 a partir do extremo 5' do miRNA) e qualquer janela de nove nucleotídeos na 3'UTR. Um mínimo de quatro pareamentos de Watson e Crick (WC) consecutivos são necessários, iniciando pela posição um ou dois da *seed* estendida do miRNA. Um único emparelhamento de wobble é permitido para os sítios de ligação que apresentam mais de seis WC consecutivos. Uma única protuberância ou despareamento é consentido se houver oito pareamentos WC consecutivos.

Todos os sítios de ligação com menos de seis pareamentos WC consecutivos (4mers, 5mers), bem como aqueles contendo uma imperfeição (protuberância, wobble, mismatch), são filtrados baseado na energia livre, que é medida por meio do RNAhybrid. A energia de ligação entre o miRNA e seu reverso complementar é considerada como energia complementar perfeita. O sítio somente é selecionado se a razão entre a sua energia livre e a energia complementar perfeita for superior a um certo limiar, determinado para cada categoria de ligação: 0,7 para sítio de ligação 7mer com um pareamento wobble; 0,6 para todos os outros com imperfeições; 0,4 para 4mers e 5mers.

Para todos os alvos identificados, são calculadas a conservação e a acessibilidade estrutural. A conservação evolutiva dos sítios de ligação identificados é avaliada por meio do escore de conservação, baseado em 16 espécies. Um filtro inicial retém apenas os alvos que são conservados em todas as posições pareadas da *seed* em pelo menos três espécies. O escore de conservação é definido como a razão do número de espécies em que as posições de ligação da região *seed* estendida são conservadas e o número máximo de espécies apresentando qualquer conservação na 3'UTR inteira.

Para estimar a acessibilidade estrutural, a amostragem estatística para a ocorrência de regiões de fita única, como implementado no programa Sfold, é utilizada.

Todas as características do alvo são apresentadas à rede neural artificial (RNA) utilizando um vetor com sete componentes. O primeiro componente é a acessibilidade estrutural do alvo. Os outros seis correspondem a classes de ligação (4mer, 5mer, 6mer, 7mer, 8mer, 9mer) e

tem como valor o escore de conservação relativo. Esse vetor serve como entrada para um neurônio linear que atribui o MRE escore multiplicando as características com sete pesos. O MRE escore serve como entrada para um neurônio não linear que integra todos os MRE escores em um miTG escore final, por meio de um feedback não linear. O miTG escore reflete o grau de repressão do gene codificante.

O DIANA-microT-ANN foi comparado ao Pictar e TargetScan 5.0 usando três diferentes limiares de escore: estrito, médio e frouxo. No limiar médio, em que todas as ferramentas apresentam precisão semelhante, o DIANA-microT-ANN apresentou maior sensibilidade. Além disso, este obteve uma fração de verdadeiros positivos 14% maior, alvos não preditos pelos outros algoritmos. No limiar estrito, o DIANA-microT-ANN obteve maior precisão e menor sensibilidade, enquanto no limiar frouxo ocorreu o contrário.

Targetprofiler

No Targetprofiler (Oulas et al., 2012), profile HMMs são treinadas para reconhecer certas características biológicas das interações miRNA-mRNA.

Uma janela deslizante de 32 nucleotídeos percorre inteiramente as 3'UTRs humanas se movendo um nucleotídeo por vez. Os alvos potenciais são filtrados de acordo com o escore de conservação[10] >= 2, escore HMM >= 3 e energia livre < -8.0 (predita pelo RNAcofold). O outro filtro utilizado está associado à localização do sítio na 3'UTR: os alvos dentro dos nucleotídeos nos primeiros 0.3% do extremo 5' ou dos últimos 0.2% do extremo 3' das sequências 3'UTRs são eliminados. Então, o modelo de HMM atribui um escore de probabilidade ao sítio de ligação.

MirTarPri

Os autores (Wang et al., 2013) realizaram uma análise sistemática de alvos de miRNA experimentalmente validados usando dados de genômica funcional, e encontraram associações funcionais significativas entre genes que são alvos de um mesmo miRNA. Baseado nesses achados, foi desenvolvido um método de priorização de sítios de ligação de microRNA para ranquear as listas de alvos preditos por ferramentas comumente usadas para este fim.

MicroRNAs humanos e alvos associados foram extraídos de três bancos de dados: TarBase, miR2Disease e miRecord. Foram utilizados os alvos de miRNA preditos pelas ferramentas TargetScan, PicTar, miRanda, PITA e DIANA-microT. Já que o RNAhybrid não fornece os resultados preditos, os autores utilizaram-no para predizer alvos de miRNAs em transcritos humanos. As sequências de transcritos humanos foram extraídas do Ensembl.

[10] Baseado em alinhamentos múltiplos de várias espécies construídos com o software MULTIZ.

Um passo importante nesse método é medir as associações funcionais entre alvos de miRNA, por meio da similaridade funcional baseada em anotações GO e proximidade de rede baseada em redes de interação proteína-proteína (PPI).

Foi utilizado o conteúdo da informação (IC) para definir a similaridade entre dois termos GO, calculada para aqueles que dividiam informações ontológicas comuns, indicadas por um ancestral comum específico. O IC é uma medida de quanto específico é um termo. O valor IC de um termo t é calculado como o log negativo da razão entre o número n de genes mapeados ao termo t, e o total N de genes de todo o genoma humano:

$$IC(t) = -\log (n/N)$$

À medida que o valor de IC aumenta, a função do termo se torna mais específica. O escore da similaridade funcional (FS) entre dois genes alvos, g1 e g2, foi calculado como o IC do ancestral comum mais informativo entre os termos mapeados por g1 e por g2, como se segue:

$$FS(g1,g2) = \max_{t \in T(g1,g2)} IC(t)$$

T(g1,g2) representa o conjunto de todos os termos ancestrais comuns mapeados por g1 e g2. Um escore FS mais alto indica que dois genes apresentam mais informação em comum e são mais semelhantes. O escore da média de similaridade funcional (AFS) entre um alvo candidato g e um grupo de n alvos experimentalmente validados G foi definido da seguinte forma:

$$AFS(g,G) = \sum_{i=1 \text{ a } n} FS(g,gi)/n, \text{ em que gi é um membro do grupo G}$$

O escore de proximidade da rede de dois genes alvos g1 e g2 foi definido como o inverso da distância recíproca mais curta (DIS) entre os nós dos produtos gênicos na rede, da seguinte forma:

$$NC(gi,g2) = 1/DIS(g1,g2)$$

O escore da média de proximidade da rede (ANC) entre um sítio candidato g e um conjunto de alvos experimentalmente validados G foi definido como se segue:

$$ANC(g,G) = \sum_{i=1 \text{ a } n} NC(g,gi)/ n, \text{ em que gi é um membro do grupo G}$$

O mirTarPri funciona da seguinte maneira: os sítios candidatos são priorizados de acordo com sua similaridade e proximidade a alvos experimentalmente validados de um determinado miRNA. Há três etapas para a priorização. Primeiramente, cada sítio potencial é mapeado aos termos GO de ontologias ortogonais (processos biológicos, funções moleculares e componentes

celulares), para calcular seu AFS escore em relação aos alvos experimentalmente validados. A lista de sítios candidatos é então ranqueada de acordo com o AFS escore. Na segunda etapa, cada alvo potencial é mapeado à rede PPI para medir seu ANC escore baseado em sítios já conhecidos. Os possíveis alvos são então ranqueados de acordo com o ANC escore. Na terceira etapa, os ranqueamentos fundamentados nos escores AFS e ANC são combinados em apenas um, utilizando o método estatístico Q. Este ranqueamento final indica a prioridade de cada lista de alvos potenciais.

O mirTarPri foi comparado a outros métodos integrados: myMIR, MAGIA e HOCTAR, e obteve maior número de verdadeiros positivos, com maior precisão.

HomoTarget

HomoTarget (Ahmadi et al., 2013) utliliza uma rede neural artificial (ANN) modificada para prever os sítios de ligação de microRNAs, e um padrão de reconhecimento de redes neurais (PRNN) para classificação dos alvos, associado a análise de componentes principais (PCA). Este modelo se baseia em doze características estruturais, termodinâmicas e posicionais das interações mRNA-miRNA para selecionar os alvos candidatos:

1. Escore total = obtido pela soma dos escores dos pares: WC +5, G-U +1, mismacth -3, gap -1 e extensão do gap -1
2. Escore *seed* = obtido pela soma dos escores dos pares na região *seed*.
3. Número de pares WC no duplex
4. Número de pareamentos de *wobble* no híbrido
5. Número de mismatches no duplex
6. Número de arqueamentos no híbrido
7. Proporção de “A” no duplex
8. Proporção de “C” no híbrido
9. Proporção de “G” no duplex
10. Proporção de “U” no híbrido
11. Proporção de pareamentos A-U no duplex
12. Energia livre mínima = calculada pelo RNA fold para o híbrido miRNA-alvo.

Primeiramente, o miRNA de entrada é alinhado a uma dada sequência alvo (mRNA) utilizando o algoritmo de alinhamento local Smith-Waterman modificado. Ao procurar por todos os emparelhamentos possíveis em cada mRNA-miRNA, os primeiros 10% com maior escore são selecionados. Este escore é computado de acordo com as regras descritas anteriormente. O próximo passo é filtrar os alinhamentos. O filtro mais importante aplicado é baseado na definição padrão de *seed*. A *seed* apresenta comprimento entre seis e oito pares de bases, com início na

posição dois do microRNA. Não são permitidos mismatches ou loops, mas apenas um único pareamento de *wobble* dentro dos hepta ou octâmeros. Em seguida, para cada sítio alvo potencial, as doze características listadas anteriormente são extraídas e normalizadas formando um vetor, que é submetido ao PCA. [11]Os componentes principais do vetor de características são calculados e finalmente adicionados ao classificador.

Foi realizada uma comparação entre esse modelo (PRNN) e outros classificadores (KNN, árvore, SVM, RBFNN, LVQ, GRNN, PNN, CFNN, FFNN). Os melhores resultados foram do PRNN e FFNN, mas o primeiro apresentou menor erro quadrático e portanto foi considerado a melhor ferramenta. O HomoTarget também apresentou melhor especificidade que os demais (EIMMO, MiRanda, MirTarget2, PicTar, Rna22, TargetScan, TargetScanS, MTAR, e MiREE).

MirTarHunter

MirTarHunter (Park e Kim, 2013) é baseado em programação dinâmica, incluindo vários tipos de sítios alvos e características espécie-específicas, e não utiliza pareamento com a *seed* e nem conservação evolutiva.

Sequências de miRNA e mRNA servem como entrada, os alvos são preditos por meio do algoritmo de Smith-Waterman modificado e a energia livre do duplex é calculada. O resultado final é ordenado e ranqueado.

Os autores modificaram o algoritmo de Smith-Waterman e criaram uma versão diferencialmente ponderada, que é otimizada para pesquisar alinhamentos locais máximos entre um miRNA e qualquer posição do mRNA. O algoritmo do mirTarHunter é muito semelhante ao Smith-Waterman. Entretanto, ao invés de o alinhamento ser baseado no nucleotídeo correspondente (A-A ou U-U, por exemplo), é baseado no nucleotídeo complementar (A-U ou G-C), permitindo também pareamentos de wobble (G-U). São atribuídos os seguintes escores: +7 para pares G:C, +5 para A:T, +1 para G:U e -3 para todos os outros pares não complementares. O algoritmo utiliza penalidades para abertura do gap (-8) e extensão deste (-2). Devido à assimetria de ligação existente entre os extremos 5' e 3' do microRNA, os escores de complementaridade da região 5' seed (posições de 1 a 11 contando a partir do extremo 5' do miRNA) e da região 3' (posições de 12 a 15) são multiplicados por um fator escalar de 4 e 2, respectivamente. Os tipos de sítios alvos de miRNA incluídos são: 7mer-A1, 7mer-m8, 8mer, 6mer, 6mer-deslocada, suplementar 3' e compensatório 3', definidos por Bartel, 2009.

Além disso, o mirTarHunter aplica quatro regras para assegurar que os duplexes miRNA-mRNA sigam os padrões determinados experimentalmente: nenhum mismatch entre as

[11]Este é um método matemático que utiliza a transformação ortogonal para converter um conjunto de observações de variáveis possivelmente correlacionadas em um conjunto de valores de variáveis não correlacionadas chamadas de componentes principais.

posições 2 e 4 (contando a partir do extremo 5'); menos que cinco mismatches entre as posições 3 e 12; pelo menos um mismatch entre as posições 9 e L-5 (L é o comprimento do híbrido); e menos que dois mismatches nas últimas cinco posições do duplex. Utilizando esses parâmetros, esta ferramenta otimiza o escore do alinhamento complementar entre miRNA e mRNA, soma todas as posições alinhadas e cria um ranking, em ordem decrescente, de todos os emparelhamentos não sobrepostos que apresentam score abaixo de um determinado limiar (valor padrão de 100).

Posteriormente, a energia livre das possíveis interações detectadas na etapa anterior é calculada por meio da rotina de dobramento de energia livre mínima (MFE fold), uma das rotinas de dobramento do RNAlib. Esse cálculo é baseado no algoritmo de programação dinâmica desenvolvido por Zuker e Stiegler (1981).

As sequências de miRNA e mRNA são unidas por oito bases X. A energia livre dessa estrutura é então calculada pelo MFE fold e colocada contra um valor limiar. Em seguida, os alvos preditos para cada miRNA são classificados em tipos de sítios e ordenados primeiro de acordo com o escore do alinhamento e depois de acordo com a energia livre.

Uma desvantagem desse método é que se múltiplos miRNAs se ligam a um mesmo sítio apenas aquele com maior escore de alinhamento e menor escore de energia livre será reportado.

Em relação ao desempenho, o mirTarHunter apresentou sensibilidade de 100%; 98,12% e 77,4% para os limiares de 100, 120 e 140, respectivamente, sendo bem mais alta que a de outros programas. Em um estudo prévio, a maioria das ferramentas apresentaram sensibilidade em torno de 60-65% (Sethupathy et al., 2006). Semelhante a outros programas, o mirTarHunter apresentou uma baixa especificidade de 42,3% para um limiar de 100. Entretanto, apresentou uma especificidade relativamente alta de 78.8% e 96,2% para os limiares de 120 e 140, respectivamente.

Baymir

Baymir (Radfar et al., 2013) é um modelo de regressão linear Bayesiana esparsa, e calcula o grau em que a regulação decrescente dos mRNAs pode ocorrer devido à atividade dos miRNAs, integrando evidências de sequência e expressão.

Inicialmente, para cada microRNA, essa ferramenta identifica um conjunto de alvos baseados na presença de sítios conservados complementares à região *seed* do miRNA na 3'UTR do mRNA. Em seguida, Baymir extrai vetores de expressão do mRNA (variável medida) associados a alvos selecionados de um conjunto de dados de expressão gênica. Cada vetor de expressão consiste da abundância transcricional do sítio em uma das amostras coletadas de experimentos de microarray. Para encontrar o vetor de atividade de um dado miRNA, é calculada a média dos vetores de expressão normalizados de seus alvos. Em seguida, os vetores de atividades do

microRNA são usados como regressores em um modelo de regressão linear Bayesiana, considerando cada mRNA.

Considerando K miRNAs e M amostras de *microarray*:

$\mathbf{y}^i$ é um vetor de M posições que contém, em cada posição k, o nível de expressão do i-ésimo mRNA na amostra k;

$\mathbf{\Delta y}^i$ é um vetor de M posições obtido da subtração de $\mathbf{y}^i$ pela média de seus valores, ou seja, é a diferença do nível de expressão do i-ésimo mRNA em relação ao valor médio de expressão pelas M amostras;

W é uma matriz de dimensão MxK na qual w_{mk} representa o nível de atividade de do miRNA k na amostra m;

$\mathbf{h}^i$ é um vetor de K posições no qual cada posição k que indica a contribuição do miRNA k em regular negativamente a expressão do i-ésimo mRNA.

Considerando um erro **e**, o modelo de regressão é dado pela equação

$$\mathbf{\Delta y}^i = \mathbf{W}\,\mathbf{h}^i + \mathbf{e}.$$

Ou seja, é atribuído um coeficiente de regressão (h) a cada interação miRNA-mRNA. Este coeficiente indica a força da repressão mediada por miRNAs em um determinado mRNA, considerando todos os microRNAs que se ligam a ele, e é chamado de “Baymirscore”. O “Baymirscore” é calculado por meio da regressão de rede elástica. Trata-se de um método que reduz o número de classificadores e seleciona os mais importantes, considerando a correlação entre eles. Quando um classificador de um determinado grupo é escolhido, o grupo inteiro é selecionado (Zou e Hastie, 2005).

Já que muitos alvos estão sob controle de múltiplos microRNAs, a ferramenta aplica um modelo linear que relaciona o vetor de expressão do alvo (variável medida) a uma combinação ponderada de vetores de atividade de miRNAs (variável regressora).

Também é considerada a variabilidade na expressão gênica de um mRNA para diferenciar alvos funcionais e não-funcionais de um determinado microRNA. O índice de variação gênica de cada mRNA é calculado como a variância dos níveis de expressão gênica em todas as amostras.

Para avaliar o desempenho da predição do Baymir, os autores utilizaram alvos experimentalmente validados do Tarbase (Papadopoulos et al., 2009). O número de mRNAs alvos validados dessa fonte é 491, sendo 279 com sítio alvo conservado. A ferramenta predisse 203 desses alvos conservados (72,8%).

MREdictor

MREdictor (Incarnato et al., 2013) detecta elementos de reconhecimento de proteínas Pumilio na proximidade de sequências *seed* dentro de estruturas pouco acessíveis. As proteínas

Pumilio são necessárias para a atividade do miRNA em sítios termodinamicamente indisponíveis. Este algoritmo permite a identificação de alvos em regiões pouco acessíveis e não é restrito a um pareamento perfeito com a *seed*.

Dado um alvo para testar e um miRNA, MREdictor obtém sequências 3'-UTR do UTRdb e sequências de miRNA do miRBase. A ferramenta extrai os primeiros oito nucleotídeos do extremo 5' do miRNA e computa o reverso complementar. A partir deste, são considerados os seguintes pareamentos com a seed: 8mer; 8mer-A1; 7mer-m8; 7mer-A1; 6mer; 5mer; 4mer; 7 pareamentos contíguos, com um ou dois pares G:U; um nucleotídeo do elemento de reconhecimento de miRNA (MRE) forma uma protuberância e não participa da nucleação do duplex; um nucleotídeo da seed forma uma protuberância e não participa da nucleação do duplex; um nucleotídeo da seed não é complementar ao seu correspondente no MRE.

Para toda interação possível, o algoritmo extrai uma janela de 200 nucleotídeos, que é centrada no pareamento com a *seed*, e a acessibilidade local é avaliada. Regiões excedendo uma energia *ΔGaccess* de -10 kcal/mol são sujeitas a uma PWM para localizar possíveis PRE (elementos de reconhecimento de pumilio) motifs dentro da mesma janela. Qualquer sítio presente dentro de uma região inacessível que falte um PRE motif próximo é descartado. O *ΔGaccess* é calculado como a diferença entre o ΔG de um conjunto de estruturas 3'-UTR e o ΔG de um conjunto de estruturas em que uma restrição é imposta para fazer posições de nucleotídeos entre 15 nucleotídeos *upstream* e três *downstream* da *seed* não pareada. Os sítios que superam esse primeiro filtro são submetidos a uma simulação da formação do duplex e são novamente filtrados de acordo com a energia livre (ΔGduplex).

Comparado a outras ferramentas (Miranda, TargetScan e PITA), MREdictor obteve uma maior fração da verdadeiros positivos e a maior acurácia, apresentando também maior sensibilidade.

TargetS

TargetS (Xu et al., 2014) se baseia no pareamento da região 5' do microRNA e na estabilidade termodinâmica do duplex microRNA-mRNA para predizer os alvos, e não se restringe à região 3'UTR do mRNA. Os alvos podem ser procurados também nas regiões CDSs, 5'UTRs ou promotoras. Essa ferramenta não utiliza conservação evolutiva e é capaz de predizer interações espécie-específicas e genomas não conservados.

Em relação à interação da região 5' do microRNA, são considerados cinco tipos de pareamentos com a seed: 8mer-A1, 7mer-m8, 7mer-A1, 6mer, e pareamento das posições 1 a 8 com presença de um par G:U

A estabilidade termodinâmica é medida por meio da ΔGduplex (energia de ligação entre o miRNA e o mRNA ganha para formar o duplex miRNA-mRNA) e da ΔΔG (energia de acessibilidade, que é a diferença entre a energia livre, ΔGduplex, e a necessária para desparear os nucleotídeos de miRNA-alvo, ΔGopen). A energia de ligação foi calculada pelo RNAhybrid. Para cada par miRNA-mRNA, foi calculada a ΔGduplex utilizando a sequência do miRNA e 58 nucleotídeos flanqueando a região de pareamento da *seed* na sequência do mRNA, incluindo o sítio de pareamento com a *seed*, 30 e 20 nucleotídeos conectados à região 5' e 3' do pareamento com a *seed*, respectivamente. A ΔGopen foi calculada baseada nos 58 nucleotídeos do mRNA. São calculadas todas as ΔGduplex e ΔΔG para todos os pareamentos da *seed* encontrados em cada par miRNA-mRNA. Já que diferentes tipos de *seed* apresentam diferentes eficácias em relação ao alvo (8mer-A1>7mer-m8>7mer-A1>6mer em 3'UTR), são atribuídos pesos distintos para cada *seed* de acordo com o tipo e a localização no mRNA. Os autores estabelecem dois valores de corte, um para ΔGduplex e outro para ΔΔG. Quando ambos apresentam valor abaixo do de corte, o mRNA é considerado um potencial alvo do miRNA.

O TargetS foi comparado ao TargetScanS, Pictar e MicroT-CDS. O TargeS obteve acurácia semelhante ao TargetScanS e MicroT-CDS e menor que a do Pictar. Entretanto, obteve maior número de verdadeiros positivos, sensibilidade e especificidade em relação às demais.

New support vector machine-based method for microRNA target prediction

Li e colaboradores (2014) definiram um método de predição de sítios alvos de microRNAs baseado em um modelo de Máquina de Vetores Suporte (SVM), com uma função kernel de base radial como medida de similaridade, incluindo características estruturais, termodinâmicas e de conservação.

Os sítios alvos utilizados como treinamento da SVM foram obtidos do Tarbase, pSILAC e miRecords.

Para melhorar a acurácia da pesquisa de sítios alvos de microRNAs, tanto a amostra de treinamento quanto os potenciais alvos iniciais a serem analisados pela SVM devem respeitar as seguintes regras:

- A região entre os nucleotídeos 2 e 8 do extremo 5' do microRNA é definida como região *seed*, e o número de pares de bases deve ser pelo menos cinco;
- O número de mismatches deve ser menos que quatro em toda a interação miRNA-mRNA;
- O número de mismatches consecutivos deve ser menor que três em todo o duplex;
- O valor da energia livre do híbrido dever ser menor que -30, calculado pelo RNAHybrid.

Os autores pesquisaram dois sítios alvos potenciais em um mesmo mRNA: um somente na região 3'UTR e outro nas demais áreas.

Foram definidas 133 características, classificadas em estruturais primárias, estruturais secundárias e de conservação da sequência:

- Características estruturais primárias: frequência de nucleotídeos, frequência de dinucleotídeos, frequência de trinucleotídeos e razão GC;
- Características dos dois sítios de interação: a energia livre mais baixa e o P valor;
- Características estruturais secundárias: razão do número de bases pareadas/número total de bases, razão do número de bases não pareadas/total de bases, energia livre de redobramento, número de hélices de vários tamanhos, número de *loops* de vários tamanhos, total de hélices e *loops*, tamanhos de todos os tipos de hélices e *loops*;
- Escore do vetor de características de um sítio de conservação da sequência.

A estrutura secundária do RNA mensageiro foi predita por meio do RNAfold, incluindo aproximadamente 120 bases. O algoritmo FastCompare (Elemento e Tavazoie, 2007) foi utilizado para calcular os escores de conservação, calculando o escore dos sítios de ligação 4mer. A soma de todos os escores de conservação das regiões 4mer foi considerada como o escore total de toda a sequência.

Em relação aos outros métodos, o classificador SVM de dois sítios obteve melhor taxa de verdadeiros positivos que o classificador SVM de um sítio e o Pictar. Também obteve maior taxa de verdadeiros positivos que o miranda e o mirtarget2 com menor taxa de falsos positivos. Utilizando amostras publicadas recentemente, o método dos autores predisseram maior número de pares miRNA-mRNA que as demais ferramentas (TargetScan, PITA, TargetSpy, TargetMiner, Miranda e MirTarget2).

MirMark

MirMark (Menor et al., 2014) é uma ferramenta baseada em aprendizado de máquina que, no processo de aprendizado, considerou mais de 700 características e utilizou Seleção de Característica baseada em Correlação (CFS) para escolher as características mais relevantes e menos redundantes, com uma variedade de métodos estatísticos para classificar os alvos de microRNA. Neste artigo, apenas alvos na região 3'UTR são considerados, mas o método é adaptável para predição de sítios alvos dentro das regiões codificantes (CDSs).

Os dados positivos foram obtidos do miRecords e do miRTarBase. Os dados negativos foram gerados utilizando miRNAs falsos, que são permutações randômicas de uma sequência de miRNA maduro real, sem sobreposição com regiões *seed* de miRNAs conhecidos. Essas sequências randômicas foram geradas por meio do algoritmo Fisher-Yates shuffle (Knuth, 2014).

Os grupos de características site-level são: diferentes valores de energia livre, tipos de pareamento com a seed (8mer, 8merA1, 7mer1, 7mer2, 7merA1, 6mer2, 6mer1, 6mer1GU,

6mer2GU), tipo de pareamento miRNA-mRNA dos 20 primeiros nucleotídeos do miRNA, acessibilidade do alvo, composição do sítio, conservação do alvo, localização do sítio em relação à 3'UTR.

Os grupos de características UTR-level são: resumo das características site-level (valor total, mínimo, máximo e médio das 151 características site-level ; valor total, mínimo, máximo e médio da probabilidade *a posteriori* do classificador site-level baseado em *random forest*, escore de alinhamento do Miranda, posição inicial e final do alvo potencial. Além disso, o comprimento da 3'UTR e o número de sítios candidatos para um par gene-miRNA são considerados. A densidade do alvo potencial é computada como número de sítios/ comprimento da UTR, como feito no SVMicro. Uma outra medida de densidade é computada contando o número máximo de sítios alvos candidatos que estão dentro de 100 nucleotídeos entre si.

Para o classificador site-level, houve a seleção de um subconjunto de 12 características, ranqueadas de acordo com a informação mútua, que são:

- número de arqueamentos no sítio alvo;
- número de nucleotídeos dentro dos arqueamentos no sítio alvo;
- tipo de ligação da posição 1 do miRNA, como pareamento GC, AU ou GU;
- número de mismatches no sítio alvo;
- número de pareamentos AU no sítio alvo;
- tipo de ligação da posição 8 do miRNA;
- tipo de ligação da posição 4 do miRNA;
- escore de conservação da região *seed*;
- tipo de ligação da posição 3 do miRNA;
- tipo de ligação da posição 15 do miRNA;
- escore de acessibilidade da posição 1 da região *seed*;
- número de arqueamentos na região *seed*.

Primeiramente, os sítios alvos candidatos são identificados utilizando o algoritmo de alinhamento do Miranda. O classificador site-level utiliza as características selecionadas para encontrar os mais fortes. Em seguida, o classificador UTR-level integra os sítios alvos candidatos para determinar se um gene é alvo de um determinado miRNA.

Os autores compararam o desempenho do random forest (RF), SVM gaussiana e regressão logística (LR) com ferramentas de predição de alvos de miRNAs publicamente disponíveis: TargetScan, Miranda, SVMicro, RNAhybrid e PITA. Os classificadores site-level RF, SVM gaussiano e LR obtiveram menor taxa de falso positivo que os demais métodos, sendo que o RF apresentou maior AUC. Além disso o mirMark apresentou maior número de verdadeiros positivos que o TargetScan.

Em relação ao classificador UTR-level, um total de 15 características foram selecionadas pelo CFS e ranqueadas de acordo com a informação mútua:

- sore do alinhamento máximo entre miRNA e alvo (Miranda);
- proporção de sítios alvos com pareamento WC entre as posições 2 e 7;
- tipo de ligação da posição 1 do miRNA;
- proporção de sítios alvos com pareamento WC entre as posições 2 e 8;
- proporção de sítios alvos com pareamento WC entre as posições 1 e 7;
- energia livre mínima da região *seed* dos duplexes miRNA-alvo;
- energia livre mínima média da região 3' dos duplexes miRNA-alvo;
- composição dos dímeros UC do sítio alvo candidato;
- tipo de ligação da posição 9 do miRNA;
- tipo de ligação da posição 2 do miRNA;
- número médio de pareamentos GU na região *seed;*
- tipo de ligação da posição 7 do miRNA;
- distância mínima dos sítios alvos ao extremo 5' da 3'UTR;
- tipo de ligação da posição 19 do miRNA;
- tipo de ligação da posição 15 do miRNA.

Os autores compararam o desempenho do classificador UTR-level da RF, SVM gaussiana e LR com TargetScan, Miranda, RNAHybrid, PITA e SVMicro. Os classificadores RF, SVM gaussiano e LR do mirMark apresentaram melhor desempenho que as demais ferramentas, com maior AUC. O classificador SVM foi o que apresentou maior AUC, bem próximo do AUC do RF.

MBSTAR

MBSTAR (Bandyopadhyay et al., 2015) trata cada sítio de ligação potencial ao miRNA como uma instância individual e utiliza a abordagem *multiple instance learning* para encontrar os alvos funcionais. *Multiple instance learning* é uma variação do aprendizado supervisionado em que, ao invés de receber um conjunto de instâncias, o classificador recebe um conjunto de bolsas que são classificadas como positivas ou negativas. A bolsa é considerada negativa se todas as instâncias dentro dela são negativas. Por outro lado, a bolsa é considerada positiva se pelo menos uma instância dentro dela for positiva.

Exemplos positivos biologicamente verificados de pares miRNA-mRNA foram obtidos do miRecords, enquanto exemplos negativos foram coletados de um trabalho prévio dos autores (Bandyopadhyay et al., 2009).

Para cada par miRNA-mRNA, a ferramenta identifica os sítios da 3'UTR complementares à região *seed* do miRNA, considerando quatro categorias de sítios de ligação: 6-mer, 7-mer-A1, 7mer-M8 e 8-mer. Pareamentos de *wobble* são aceitos. Selecionam-se somente os candidatos alvos posicionados não muito próximos ao códon de parada (menor ou igual a 15 nucleotídeos) nem próximo do meio da 3'UTR.

O próximo passo é a extração de características estruturais e de sequência dos potenciais sítios de ligação ao miRNA e regiões vizinhas (mais ou menos 30 nucleotídeos ao redor da possível região alvo). Além disso, calcula-se a energia livre mínima utilizando o programa RNAcofold. A seleção de características não supervisionada Laplacian é utilizada para ranqueá-las de acordo com sua importância e escolher 40 características para treinar o classificador.

O MBSTAR foi comparado ao TargetScan, miRanda, SVmicrO e MirTarget2 e apresentou melhor desempenho que as outras ferramentas, com maior taxa de verdadeiros positivos, maior área abaixo da curva ROC e menor de falsos positivos. Além disso, obteve maior acurácia e maior F-score:

F-score= 2*PPV*Sn/ PPV + Sn, sendo PPV o valor preditivo positivo e Sn a sensibilidade.

MIRZA-G

MIRZA-G (Gumienny e Zavolan, 2015) utiliza as seguintes características para a predição de alvos de miRNA: energia de ligação, composição nucleotídica ao redor síto putativo, acessibilidade estrutural deste, localização dentro das 3'UTRs e conservação evolutiva. Os autores consideraram alvos canônicos e não canônicos.

As sequências de miRNAs foram retiradas do miRBase e as de 3'UTRs do TargetScan.

Primeiramente, os autores consideraram alvos canônicos utilizados pelo TargetScan e as 3'UTRs foram escaneadas em busca de pareamentos com a *seed* do miRNA. Posteriormente, os autores identificaram sítios não canônicos. As 3'UTRs foram escaneadas com o MIRZA, utilizando uma janela de 50 nucleotídeos, deslizando 30 bases por vez. Mas só foram aceitas janelas com um escore de qualidade do alvo de pelo menos 50. Em seguida, os autores calcularam a melhor estrutura miRNA-mRNA e inferiram a região do mRNA que se ligaria à *seed* do miNA. Este sítio âncora foi utilizado para definir o alvo completo do miRNA, composto pelo pareamento da *seed* e 21 nucleotídeos *upstream*. Para cada um desses sítios, foram computadas as seguintes características:

- Escore de qualidade do alvo dado pelo MIRZA: este é um modelo que atribui valores de energia livre de ligação a todos os possíveis híbridos miRNA-mRNA, e quantifica a afinidade dos diferentes fragmentos de mRNA em relação ao RNA-induced silencing complex (RISC) baseado em dados de CLIP, gerando o score de qualidade (Korshid et al, 2013);

- Acessibilidade do alvo: definida como a probabilidade de o sítio alvo apresentar uma conformação de fita única dentro do mRNA. Essa probabilidade foi computada por meio do CONTRAfold, que foi aplicado na região cobrindo o pareamento da região *seed*, e 50 nucleotídeos *upstream* e 50 *downstream*;
- Composição de nucleotídeos na região flanqueadora ao sítio alvo: foi avaliada a proporção de G e de U dentro de 50 nucleotídeos *upstream* e 50 nucleotídeos *downstream* da região de pareamento miRNA-mRNA;
- Conservação evolutiva: as sequências 3'UTR foram alinhadas ao genoma humano (hg19) com GMAP. Os alinhamentos do genoma humano (hg19) com genomas de 41 espécies foi obtido do UCSC. Esses alinhamentos foram utilizados para acessar o grau de conservação evolutiva dos sítios alvos putativos. Para cada alvo putativo, os autores fizeram a seguinte computação. Baseado no alinhamento de 3'UTRs humanas com todas as outras espécies, foi extraída a região que corresponde ao sítio alvo putativo nas 3'UTRs humanas em todas as outras espécies. Os autores cortaram os sítios alvos putativos em todas as espécies em 50 nucleotídeos. Então, foi computado o escore de qualidade do alvo o, e o sítio foi considerado conservado quando escore de qualidade do alvo é de pelo menos 50. Em seguida, baseado nas distâncias evolutivas ao longo da árvore providenciada por UCSC, os autores computaram a fração de distância evolutiva total na árvore filogenética ao longo de qual sítio foi conservado. Essa medida foi chamada de *branch length score*. Todas as manipulações da árvore filogenética foram realizadas pelo DendroPy package.
- Posição do sítio alvo na 3'UTR: foi determinada a distância do limite mais próximo da 3'UTR como o mínimo entre a distância do começo da região complementar à *seed* e o códon de parada e a cauda poliA.

Os autores observaram que os modelos que contemplam conservação evolutiva apresentam melhor desempenho que aqueles que não contemplam. Considerando conservação evolutiva, os alvos preditos pelo modelo que considera apenas alvos canônicos (seed-MIRZA-G-C) sofrem maior regulação em resposta à transfecção do miRNA, seguido pelos alvos preditos por DIANA-microT; TargetScanPCT; o modelo MIRZA-G-C, que também contempla sítios não canônicos; e miRanda-mirSVR. Em relação aos modelos que não consideram conservação evolutiva, o modelo seed-MIRZA-G, que considera apenas sítios canônicos, obteve o melhor desempenho, seguido pelo modelo que inclui sítios não canônicos e TargetScan Context+.

2.4- Discussão

A maioria das ferramentas consideram apenas alvos que interagem com a região *seed* do miRNA e muitas se restringem à região 3'UTR (Tabela 4). Entretanto, a maior parte dos alvos estão presentes na região codificante (Smalheiser e Torvik, 2004; Helwak et al., 2013). Alguns artigos mostram que a predição é semelhante nos alvos dentro da 3'UTR e naqueles dentro da CDS e 5'UTR, e as regras de pareamento são semelhantes (Goshal et al., 2015; Moore et al., 2015) desconsiderando, logicamente, as características explicitamente dependentes da localização dentro da 3' UTR. Além disso, todas as ferramentas, com exceção do TargetS, consideram que a molécula alvo é um mRNA. TargetS permite que o sítio analisado seja DNA (região promotora).

Também tem-se mostrado que muitas interações miRNA-mRNA apresentam pareamentos não-canônicos (Smalheiser e Torvik, 2004; Helwak et al., 2013; Goshal et al., 2015; Moore et al., 2015). Segundo Helwak et al (2013), a maioria das ligações miRNA-mRNA incluem a região *seed*, mas 60% delas possuem sítios não-canônicos, com bases não pareadas e arqueamentos. Este estudo também mostra que 18% das interações envolvem a extremidade 3' do miRNA, com pouca evidência de pareamento no extremo 5', e algumas foram funcionalmente validadas. Apenas 13% dos mRNAs alvos apresentam emparelhamento acima de sete bases começando pela posição um ou dois da extremidade 5' do microRNA (Smalheiser e Torvik, 2004). Além disso, a taxa de falsos positivos dos algoritmos existentes é bastante elevada (Alexiou et al., 2009; da Silva, 2013).

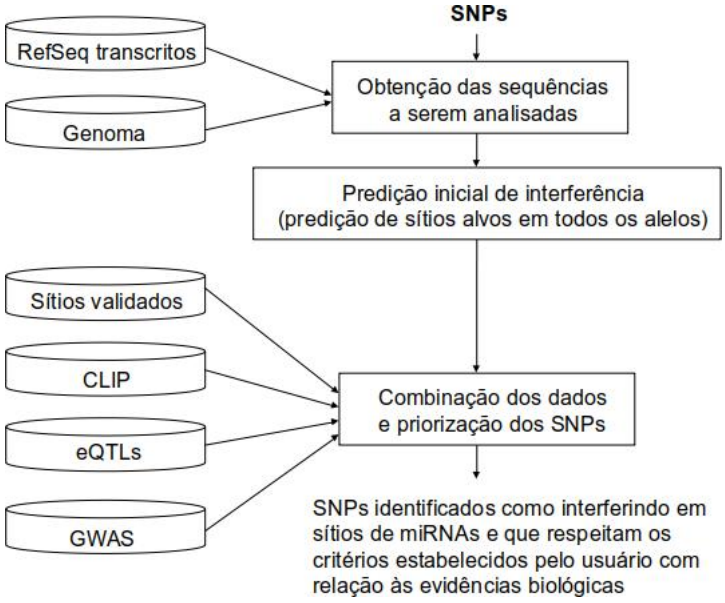
Apesar de todos esses resultados mostrando que miRNAs se ligam em várias regiões do RNA alvo, com diferentes tipos de sítios e não apenas o canônico que pressupõe uma *seed* bem definida, resultados mais recentes mostram que muitas dessas ligações não convencionais não causam a repressão da expressão da molécula alvo (Agarwal et al., 2015). Mais especificamente, Agarwal e colaboradores (2015) concluem que a vasta maioria dos sítios funcionais são canônicos, presentes na região 3'UTR. As características do miRNA, do sítio e da moléculas alvo que mais estão relacionadas com a resposta de repressão fazem parte da nova versão da ferramenta TargetScan (versão 7), proposta neste artigo. Considerando tais resultados, TargetScan 7 foi a ferramenta escolhida para ser utilizada na nova versão do SIMTar, uma vez que ela busca identificar não apenas a ligação de um miRNA em um dado sítio mas sim se tal ligação causará repressão. Tal funcionalidade vai ao encontro do objetivo central do SIMTar, que é identificar se um SNP tem real impacto na regulação por miRNA da molécula na qual o SNP se localiza.

CAPÍTULO 3 - MÉTODOS

Este capítulo descreve os métodos utilizados neste projeto para cumprir cada um dos objetivos específicos a menos do objetivo específico 1 (revisão bibliográfica dos programas de predição de sítios de miRNAs), tendo estes já sido descritos no capítulo 2.

A figura 9 resume a arquitetura do novo SIMTar detalhada ao longo deste capítulo. Dada uma lista de SNPs (cada SNP representado por seu código no dbSNP - “código rs”), o SIMTar obtém a localização de cada SNP tanto no genoma quantos nos transcritos de referência (RefSeqs). Para cada transcrito de referência no qual o SNP se localiza, a sequência exônica ao redor do SNP (figura 6) tem todas as suas variantes alélicas analisadas pelo programa de predição de sítios de miRNA (atualmente TargetScan 7) para cada miRNA humano. Se, para algum miRNA, uma ou mais sequências variantes apresentaram um sítio predito e outras não, tal SNP é um candidato a interferidor de sítio de miRNA. Para corroborar este resultado, o SIMTar então procura 4 tipos de resultados experimentais: 1) se há dados de validação experimental de que o miRNA identificado pelo TargetScan realmente se liga e reprime a molécula alvo; 2) se há dados de CLIP indicando que uma argonauta se liga àquele sítio e que portanto tal sítio deve ser alvo de algum miRNA; 3) se o SNP é um eQTL da molécula em que ele se localiza, indicando que o SNP está associado à variação de expressão desta molécula e 4) se há resultados de GWAS (*Genome wide association studies*) associando o SNP a algum fenótipo especificado pelo usuário.

Figura 9: Arquitetura do novo SIMTar.



Fonte: Paula Prieto Oliveira, 2018.

O objetivo da incorporação destas quatro fontes de informação biológica é priorizar para o usuário os SNPs que possam ser de seu maior interesse. Os dados de GWAS, embora não evidenciem nada especificamente sobre sítios de miRNAs, mostram se um SNP está associado a um determinado fenótipo de interesse. O usuário pode escolher se cada uma dessas quatro fontes de informações é um critério classificatório ou eliminatório. De forma bem genérica, o pseudocódigo abaixo expressa a lógica do SIMTar. Maiores detalhes são descritos ao longo deste capítulo.

ENTRADA: lista de SNPs

PARÂMETROS: se cada uma das quatro fontes de experimentos biológicos será um critério eliminatório ou apenas classificatório. Além disso, para GWAS, o usuário pode selecionar um subconjunto de fenótipos de interesse.

```
candidatos = conjunto vazio
para cada SNP rs
    para cada transcrito refseq em que o SNP cai em um éxon
        obtém janela w de 51 bases centrada no SNP
        para cada miRNA humano mir
            executa programa de predição de sítio deste miRNA em cada variante de w
            se o sítio é predito em alguma(s) variante(s) e não nas demais
                candidatos = candidatos ∪ {(rs, refseq, mir)}
para cada trio (rs, refseq, mir) ∈ candidatos
        se há dados de validação do sítio do miRNA mir no refseq
                (rs, refseq, mir) ← (validado = sim)
para cada par (rs, refseq) dos trios ∈ candidatos
        se há dados de CLIP de argonauta se ligando àquele refseq na região do rs
                (rs, refseq, *) ← (clip = sim)    /* isto é, para todo miRNA */
        se há dados indicando que rs é um eQTL do refseq
                (rs, refseq, *) ← (eqtl = sim) /* isto é, para todo miRNA */
para cada SNP rs dos trios ∈ candidatos
        se há dados indicando que rs está associado em algum GWAS aos fenótipos selecionados
                (rs, *, *) ← (gwas = sim)  /* isto é, para todo transcrito refseq e para todo miRNA */
para cada trio (rs, refseq, mir) ∈ candidatos
        aplica critérios de eliminação e/ou classificação definidos pelo usuário (com base nos atributos "validado", "clip", "eqtl", "gwas")
```

Para facilitar o armazenamento e o processamento de todas as informações mencionadas na figura 9, foi criado um banco de dados relacional cujo modelo conceitual será apresentado por partes neste capítulo, culminando no modelo completo apresentado na seção 4.1. Este particionamento da apresentação do modelo visa a facilitar a compreensão de como os dados foram obtidos e processados com vistas ao modelo planejado. A seção 3.1 apresenta as entidades e relacionamentos fundamentais do SIMTar. As seções 3.2 a 3.6 apresentam os métodos para cumprir os objetivos específicos 2 a 7, que são relacionados com a predição de sítios alvos nos vários alelos e incorporação das informações advindos de experimentos biológicos, e como tais informações foram incorporadas ao modelo conceitual. Por fim a seção 3.7 descreve a comparação do SIMTar com ferramentas correlatas.

Embora a arquitetura possa ser utilizada para qualquer espécie, o banco foi povoado assumindo apenas a espécie humana. Mais especificamente, foi considerada a versão hg38 do genoma humano, e as versões mais recentes dos bancos de dados externos utilizados. O banco do SIMTar foi implementado utilizando o gerenciador Postgresql versão 10.5. O modelo conceitual entidade-relacionamento foi elaborado na ferramenta Creately[12].

3.1- Entidades fundamentais do SIMTar

3.1.1 - Genes, sinônimos e Refseqs

Uma vez que o foco deste trabalho é a identificação de SNPs interferindo em sítios de miRNAs em RNAs alvos, a caracterização destes RNAs alvos e seus respectivos genes é fundamental. A figura 10 mostra a parte do modelo conceitual que envolve as entidades **Genes**, **Synonyms** e **RefSeqs** e os relacionamentos entre elas.

A entidade **Genes** armazena as informações acerca dos genes humanos, obtidas do banco Gene do NCBI[13] (https://www.ncbi.nlm.nih.gov/gene). Esta entidade armazena o código do gene neste banco (GeneID), código este já amplamente utilizado na comunidade de Bioinformática, o tipo de gene (codificante, não codificante ou pseudogene), sua descrição e *gene symbol*. O *gene symbol* é como um nome oficial para o gene. No entanto é comum que um gene seja conhecido por outros nomes. Desta forma, a entidade **Synonyms** armazena todos os nomes utilizados para cada gene, informação esta também obtida do banco Gene do NCBI.

A entidade **RefSeqs** armazena as informações das sequências de referência (não redundantes) dos vários transcritos humanos (incluindo as isoformas de *splicing alternativo*,

[12] https://creately.com/app/
[13] *National Center for Biotechnology Information*.

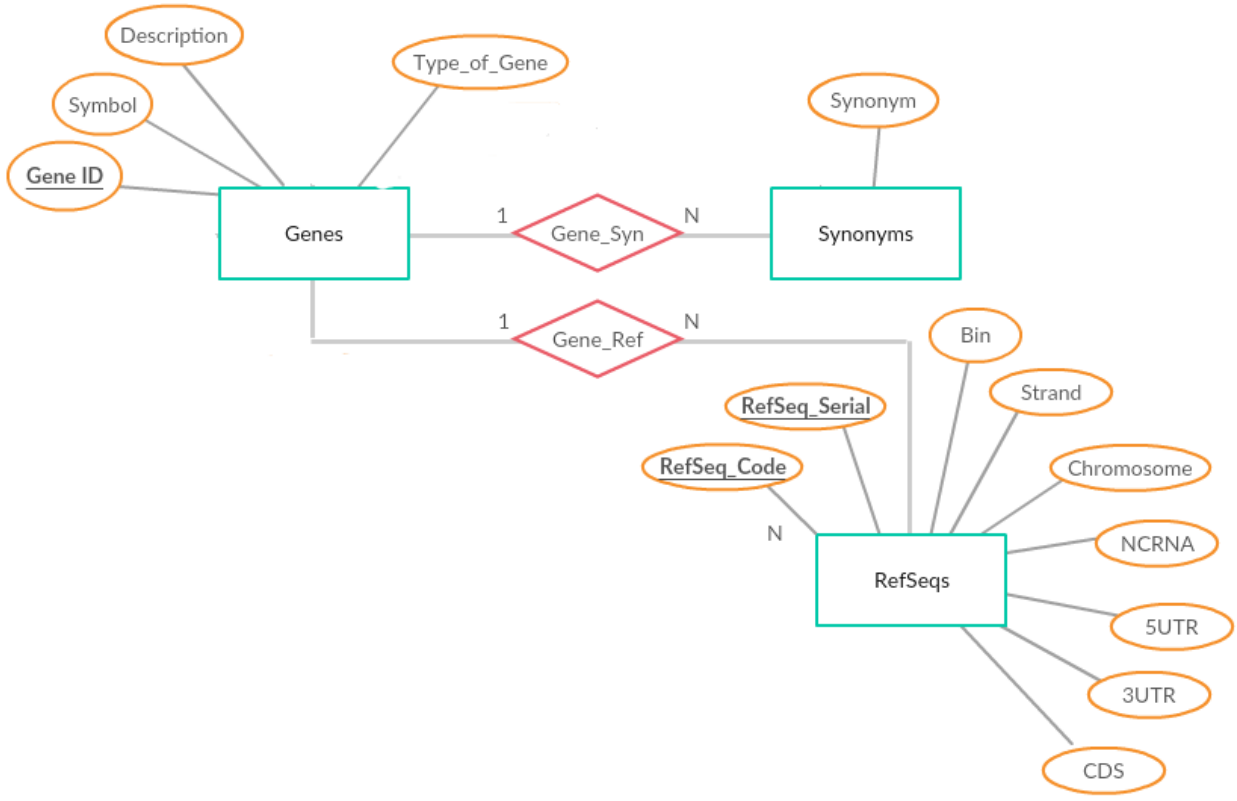
quando há), obtidas do banco UCSC[14] referente ao genoma humano versão hg38 (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/refGene.txt.gz), que traz os dados do banco RefSeqGene[15] (transcritos RefSeq com anotação curada de localização). Este banco traz as coordenadas genômicas de vários RefSeq (incluindo coordenadas de início e fim de todo o transcrito, de cada éxon e da CDS no caso de transcrito codificante), informações estas utilizadas para compor os atributos *chromosome* (cromossomo), *strand* (fita "+" ou "-"), *NCRNA* (no caso de transcritos não codificantes) e *5UTR, CDS e 3UTR* (no caso de transcritos codificantes). Estes últimos quatro atributos são *strings* que descrevem as posições de início e fim de cada éxon da respectiva região. Ex: 3UTR = "*a-b,c-d*" indica que a 3'UTR do respectivo RefSeq é formada por 2 éxons, o primeiro indo da posição *a* até *b* (incluindo *a* e *b*) e o segundo indo da posição *c* até *d*, considerando 1 como sendo a primeira posição do cromossomo, e considerando que $a < b < c < d$, independente da fita de origem do transcrito.

No banco RefSeq, um código único no formato LL_D^+ (L sendo uma letra e D^+ sendo uma sequência de um ou mais dígitos) é atribuído a uma sequência transcrita específica (ex: NM_001205206). Este é um banco considerado não redundante porque não considera possíveis cópias do gene correspondente em localizações distintas; se todos transcrevem exatamente a mesma sequência de RNA maduro (pós processo de *splicing*), todos recebem o mesmo código RefSeq. Logo, a tabela refGene do banco UCSC possui várias entradas com um mesmo código RefSeq, pois embora os transcritos finais tenham a mesma sequência eles se originaram de localizações genômicas diferentes. Por isso, a entidade **RefSeqs** no banco SIMTar utiliza como chave primária o campo *RefSeq_serial*, aqui definido como sendo o código RefSeq acrescido de um número inteiro, serial, para cada localização distinta de um mesmo RefSeq. A diferenciação da localização genômica é fundamental no SIMTar devido ao fato do SNP estar localizado no DNA, ou seja, na origem do transcrito.

As informações dos transcritos presentes no banco RefSeqGene contém, para cada transcrito, qual o gene ao qual o transcrito se refere. Essa informação, para a maior parte dos transcritos, é dada em forma de *gene symbol*. No entanto, para algumas entradas é apresentado um sinônimo. Para padronizar a informação e facilitar a recuperação de qual gene se refere determinado transcrito, no banco proposto do SIMTAR uma entrada na entidade **RefSeqs** se relaciona com um gene (entidade **Genes**) pelo GeneID.

[14] *University of California Santa Cruz Genome Browser* - http://genome.ucsc.edu/
[15] https://www.ncbi.nlm.nih.gov/refseq/rsg/

Figura 10- Parte do modelo conceitual do SIMTar que ilustra as entidades Genes, Synonyms e RefSeqs e os relacionamentos entre elas.



Fonte: Paula Prieto Oliveira, 2018.

3.1.2 - SNPs

As principais informações acerca de SNPs são armazenadas na entidade **SNPs** que, além de sua chave primária numérica sequencial, tem como atributos o identificador no banco dbSNP (código rs), cromossomo, posição no cromossomo, alelos (representados por quatro atributos binários - A, C, G e T - cujo valor 0 representa ausência e 1 representa presença de tal alelo dentre as variações encontradas), MAF e o alelo de menor frequência. Todas essas informações foram obtidas do banco dbSNP (build 151), referentes aos SNPs do genoma humano versão hg38, a partir do link ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/chr_rpts/.

Um SNP afeta as duas fitas de DNA e em ambas pode haver um gene, mais especificamente, a posição deste SNP pode coincidir com regiões exônicas de genes. Mesmo considerando cada fita separadamente, pode haver mais de uma isoforma de um gene. Logo, um

SNP pode estar associado a vários RefSeqs distintos, e para cada um deles é obtida uma janela de 51 bases ao redor do SNP (figura 6). Este comprimento de 51 bases considera que o SNP pode ocorrer em qualquer posição do início ao fim de um sítio, e por isso considera 25 bases de cada lado do SNP.

As informações de cada uma dessas janelas é armazenada no banco do SIMTar em **RNA_SNPCenteredWindows**, que é na verdade um relacionamento N:N entre **SNPs** e **RefSeqs**, como mostra a figura 11. Aqui fica evidente a necessidade de cada entrada da entidade **RefSeqs** ser relativa a uma determinada localização genômica de um dado transcrito de referência, já que este relacionamento é sobre a localização de um SNP (que ocorre no DNA) impactando um RNA dali transcrito. **RNA_SNPCenteredWindows** possui como atributos a sequência desta janela, a posição do SNP central relativa à sequência Refseq e contexto do SNP no Refseq (5' UTR, CDS, 3' UTR ou NCRNA).

Além disso, cada janela de 51 bases de um transcrito, centrada em um SNP, pode conter vários outros SNPs não centrais, o que é representado pelo relacionamento N:N **NonCentralSNPs** entre SNPs e **RNA_SNPCenteredWindows**. Tal relacionamento possui um atributo que define a posição relativa do SNP não central na janela.

A possibilidade de presença de vários SNPs em uma janela tem como consequência a possibilidade dessa janela apresentar várias combinações alélicas distintas de todos esses SNPs. Por isso, **RNA_SNPCenteredWindows** pode ser enxergada como uma entidade que possui um relacionamento N:N com **SNPs** (neste caso SNPs não centrais na janela) e com a entidade **AlleleCombinations**, que descreve cada combinação de alelos dos vários SNPs de uma janela. Por exemplo, se ao redor de um dado SNP em um dado RefSeq é definida uma janela na qual aparecem os SNPs s_1 (A/T), s_2 (A/C) s_3 (C/T), dos quais um é o SNP central e os demais são SNPs não centrais, há 8 possíveis combinações de variação destes SNPs: AAC, AAT, ACC, ACT, TAC, TAT, TCC, TCT. Assim, cada janela, representada na entidade **RNA_SNPCenteredWindows**, possui um relacionamento 1:N (**RNAWindowAlleles**) com a entidade AlleleCombinations.

E é justamente a existência de janelas com combinações alélicas distintas apresentando diferentes resultados de predição de sítio de um dado miRNA que indica a interferência de tais SNPs (ao menos um deles) em sítios de miRNAs. Os resultados das predições dos sítios de miRNAs são armazenados no relacionamento **PredictedSites** entre **RNAWindowAlleles**, **MiRNAs** e **PredictionTools**.

3.2 - Reanálise dos programas de predição de sítios de miRNAs e incorporação no SIMTar

Como já exposto na seção 1.4, há muitos programas computacionais para a predição de sítios alvos de miRNAs, e muitos deles ainda apresentam uma alta taxa de falsos positivos. Se um programa de predição de sítios não tiver boa acurácia, o SIMTar, que parte desses resultados, também será prejudicado. Assim, neste trabalho o foco foi escolher um programa que apresentasse maior precisão.

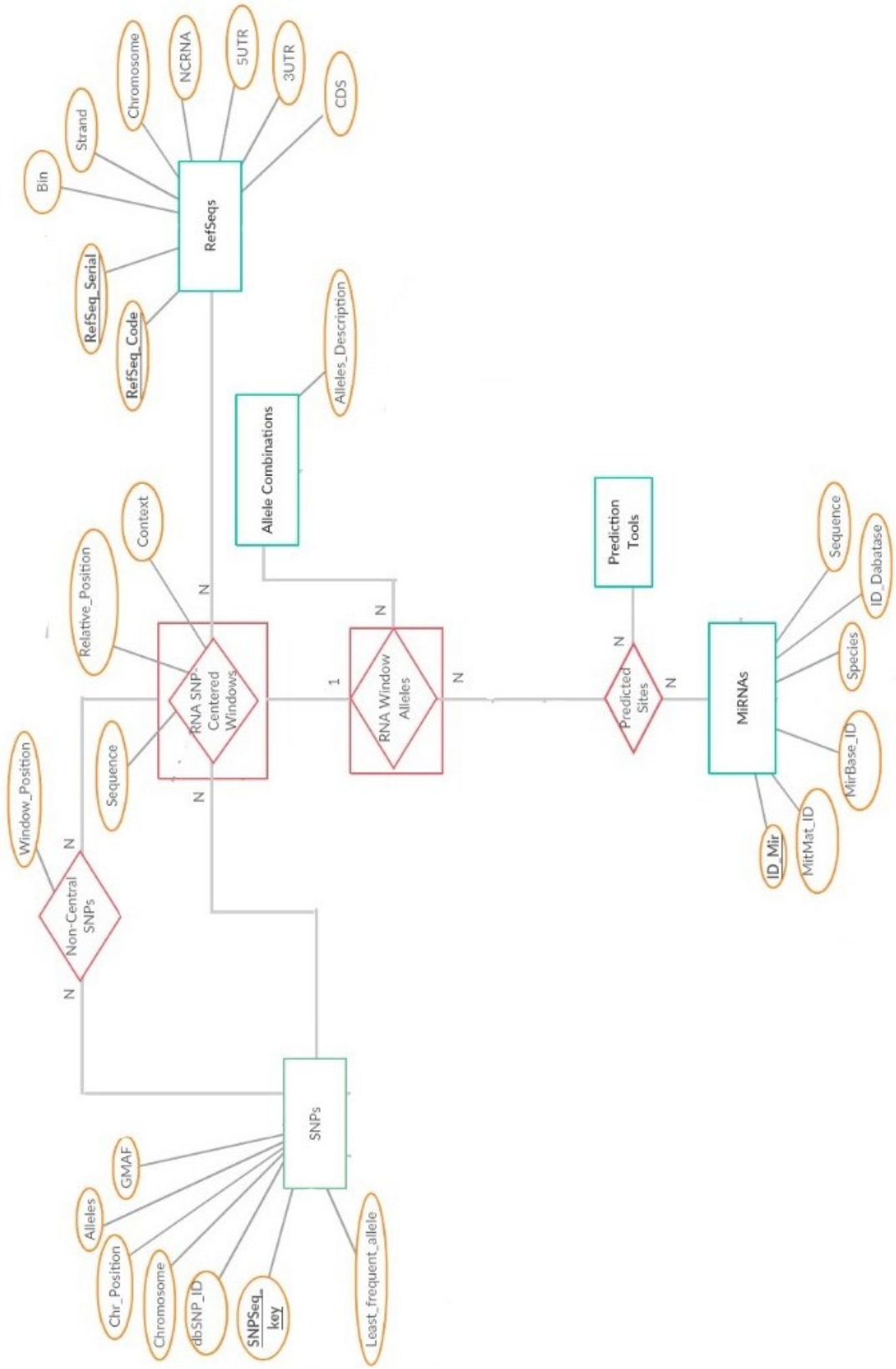
Dentre os programas analisados na revisão sistemática apresentada no capítulo 2, destaca-se a nova versão do TargetScan (versão 7). O programa foi reformulado com base em novos resultados experimentais que mostram que, embora miRNAs se liguem em várias regiões dos mRNAs alvos utilizando diferentes tipos de sítios, muitos não causam a repressão da síntese proteica. Também mostram que a grande maioria dos sítios funcionais são sítios canônicos, presentes na 3'UTR dos mRNAs alvos. A nova versão do programa analisa ao todo 14 características, várias delas não analisadas em versões anteriores, mais relacionadas com a repressão da molécula alvo. Como o objetivo do SIMTar é identificar SNPs que realmente produzem algum impacto na regulação por miRNAs, essa nova versão do TargetScan foi escolhida para integrá-lo.

Como o TargetScan baseia-se em grande parte na conservação filogenética dos sítios alvos, ele recebe como entrada um alinhamento múltiplo dos possíveis alvos assim como os miRNAs a serem investigados. O site do TargetScan 7 disponibiliza tais alinhamentos de UTRs e de CDS envolvendo 84 espécies, assim como o arquivo de miRNAs agrupados por famílias.

O procedimento aqui adotado consiste em, para cada SNP considerado, identificar nesses alinhamentos a localização deste SNP[16] e gerar um novo alinhamento para cada alelo. Em cada alinhamento gerado apenas a sequência relativa à espécie

[16] A identificação da localização do SNP foi realizada da seguinte maneira: 1) para cada alinhamento de UTR disponibilizado no site do TargetScan, gerou-se a sequência fasta (sem os símbolo "-" de gap) apenas da sequência humana; 2) para cada janela de 51 bases ao redor do SNP, identificou-se a qual UTR humana ela pertence (via alinhamento local com as sequências geradas no passo 1, utilizando o programa blast), e a localização de início desta janela, a fim de identificar a posição específica do SNP naquela UTR (digamos posição *i*); 3-) identificou-se, no alinhamento da UTR identificada no passo 2, a coluna do alinhamento em que se localiza o i-ésimo nucleotídeo da sequência humana (que corresponde à localização do SNP).

Figura 11- Parte do modelo conceitual do SIMTar que ilustra as entidades SNPs, RefSeqs e Allele Combinations, e os relacionamentos existentes entre elas.



Fonte: Paula Prieto Oliveira, 2018.

humana é alterada de forma a refletir a variação alélica. O TargetScan é então executado sobre cada alinhamento gerado e, sobre os resultados, é identificado se há diferença de predição entre eles para ao menos um miRNA (isto é, para um alelo foi predito o sítio e para outro não).

Os resultados do TargetScan sobre os vários alelos são armazenados no relacionamento PredictedSites apresentado na figura 11. Embora no momento o TargetScan seja o único programa de predição de sítios de miRNAs utilizado no SIMTar, a existência da entidade PredictionTools permite que outros programas possam ser incorporados no futuro.

3.3- Inclusão de informação de sítios alvos experimentalmente validados de miRNAs

A fim de se ter maior evidência de que um SNP interfere em um sítio de miRNA precisa-se, primeiro, saber se esse SNP localiza-se em uma região na qual, para pelo menos um dos alelos, há realmente um sítio de miRNA. Para isso, foram obtidos e integrados ao SIMTar as informações acerca de sítios validados experimentalmente, restringindo-se apenas aos dados relativos à espécie humana, das seguintes bases de dados: miRecords (Xiao et al., 2009), miRTarBase (Hsu et al, 2011; Hsu et al., 2014; Chou et al, 2016; Chou et al, 2018), miR2Disease (Jiang et al, 2009) e MtiBase (Guo et al, 2015).

O banco de dados do SIMTAR armazena esse tipo de informação por meio do relacionamento **ValidatedSites**, um relacionamento triplo entre as tabelas **MiRNAs**, **Refseqs** e **Databases** (Figura 12). Este relacionamento armazena a informação de que existe, em uma determinada base de dados, o relato de um sítio validado de um determinado miRNA ocorrendo em um transcrito Refseq. Com relação à localização do sítio, algumas das bases de dados encontradas informam exatamente a posição do sítio na sequência do RNA, outras apenas a região do transcrito quando em RNAs codificantes (5' UTR, CDS ou 3'UTR) e outras não possuem nenhuma informação de localização. Com relação à sequência alvo, algumas bases de dados informam exatamente o transcrito no qual o sítio foi validado (informando o código Refseq ou similar) e outras informam apenas o nome do gene ou a coordenada genômica. Desta forma, a fim de armazenar estas diferentes informações, o relacionamento **ValidatedSites** foi implementado possuindo os seguintes atributos:

- 2 atributos para as posições inicial e final do sítio no transcrito (valor -1 se a posição exata for desconhecida);
- 2 atributos para as regiões nas quais se localizam o início e o fim do sítio (valor -1 se a região exata for desconhecida, e códigos 0 a 3 para 5' UTR, CDS, 3' UTR e ncRNA, respectivamente);

- tipo de localização informada (valor 0 para posição exata, 1 para apenas região e 2 para apenas o transcrito ou gene alvo);
- evidência da molécula alvo (valor 0 se o transcrito exato foi informado, 1 se foi informado apenas o nome do gene, 2 se foi informada a localização genômica do sítio alvo).

A ideia em diferenciar essas informações é permitir que o usuário opte por ser mais ou menos conservador na identificação de interferências de SNPs. Ou seja, o usuário poderá exigir (na interface gráfica de acesso ao SIMTar), por exemplo, que haja um sítio validado de miRNA na posição exata do SNP ou então que baste que o sítio se localize na mesma região do SNP.

A seguir são descritos os métodos de obtenção dos valores desses atributos para cada uma das bases de dados pesquisadas.

Base de dados miRecords

Na base de dados miRecords[17] (Xiao et al., 2009) todos os sítios são relativos a transcritos específicos (o código Refseq é informado), e portanto todas as entradas do Relacionamento **ValidatedSites** tiveram o atributo "evidência da molécula alvo" preenchido com o valor 0 (transcrito exato). Parte dos dados são disponibilizados com informação de posicionamento exato, parte com informação apenas de região e parte com informação do transcrito alvo apenas, sendo portanto, para cada um deles, preenchido devidamente o atributo "tipo de localização informada".

Para os dados com localização de posicionamento exato (374 entradas), foi necessário corrigir tais posições, pois foi identificado que as coordenadas do sítio alvo informadas no miRecords, relativas aos transcritos, não coincidiam com a sequência do sítio alvo informada nesta mesma base[18]. Para obter a informação correta, foi desenvolvido um programa Perl que alinha a sequência do sítio alvo informada no miRecords com a sequência do respectivo Refseq. Tal alinhamento é feito fazendo chamadas à ferramenta blastn do pacote blast+[19] (com parâmetro *word_size*=11) e obtendo a posição correta de início do sítio relativa ao transcrito[20]. A posição final do sítio foi calculada como sendo a posição inicial + o tamanho da sequência do sítio -1. Feito isso, foram identificadas as regiões do transcrito nas quais essas posições de início e fim se localizavam. Isso foi feito com base nas informações contidas na tabela de RefSeqGenes

[17] Dados obtidos em 2016, sendo que a última atualização do banco foi em abril de 2013.
[18] Tal discrepância deve ser devida às atualizações que são realizadas sobre as sequências de referência (refseq) ao longo do tempo, gerando diferentes versões.
[19] https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&P
[20] Tomou-se o cuidado de certificar-se de que o alinhamento era o correto, ou seja, com total de cobertura e identidade.

humanos do UCSC Genome Browser[21], como descrito a seguir. Essa tabela (refGene.txt) contém, em coordenadas genômicas, as fronteiras éxons/íntrons e UTRs/CDS. Foi então desenvolvido um programa que processa essa tabela e a traduz em fronteiras em termos de coordenadas do próprio RNA (posições de 1 até o tamanho do RNA).

Cabe mencionar que, das 374 entradas do miRecords com posicionamento exato, 10 foram descartadas porque a sequência alvo possuía muitos N's e uma porque se referia a um refseq que foi removido do banco de Refseqs[22]. Outros ajustes manuais tiverem que ser feitos, por exemplo de alguns casos em que a sequência alvo e a sequência do miRNA estavam em colunas trocadas, e casos em que a sequência alvo apresentava caracteres estranhos. Após uma cuidadosa correção dos dados, 363 entradas do MiRecords, com localização exata, puderam ser incluídas.

No total, dentre os vários tipos de informação de localização, 1725 sítios validados (pares miRNA/refseq) do miRecords foram incluídos no relacionamento **ValidatedSites**.

Bases de dados miRTarBase e miR2Disease

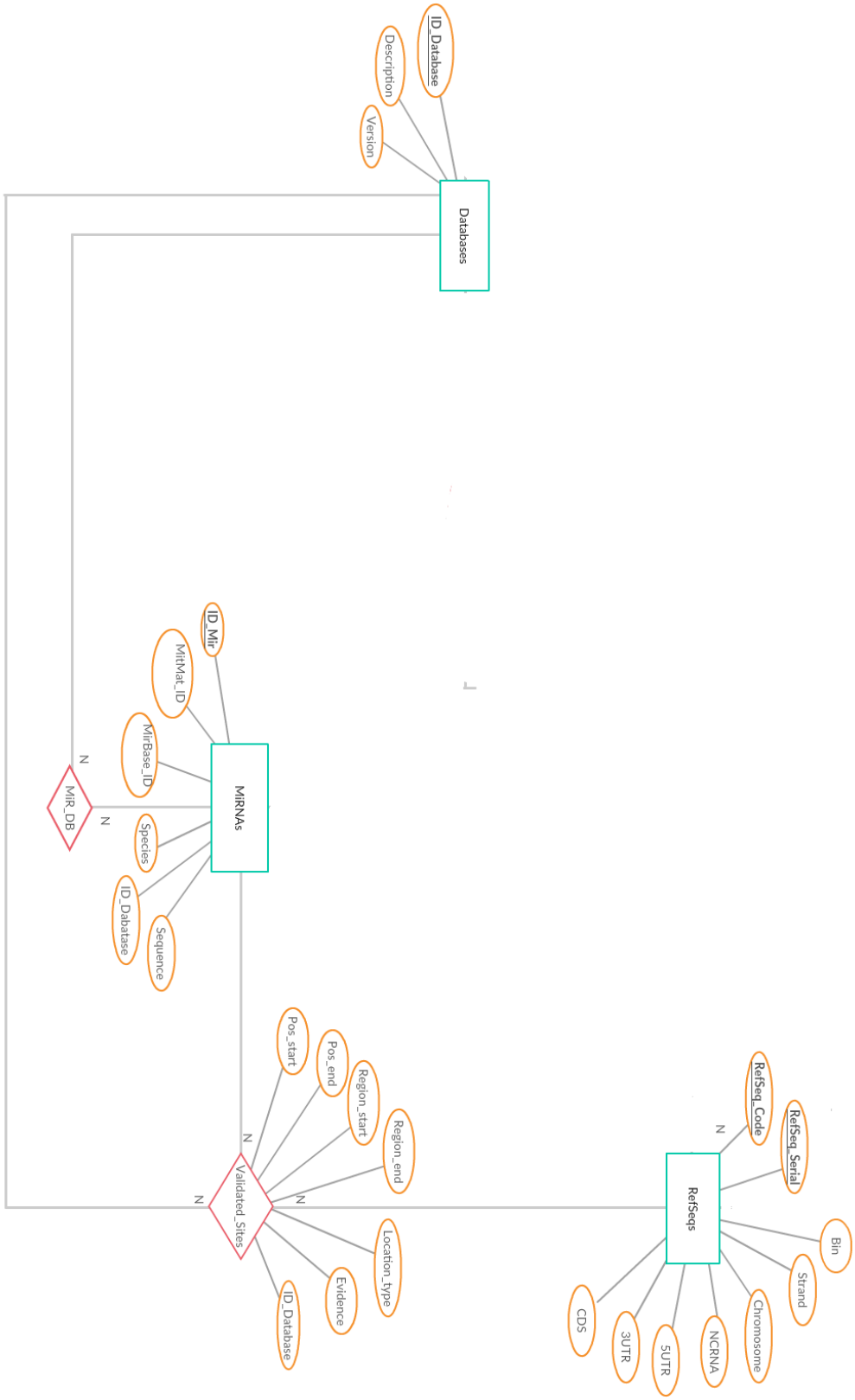
Os dados dos bancos miRTarBase (Hsu et al, 2011; Hsu et al., 2014; Chou et al, 2016, Chou et al, 2018) e miR2Disease (Jiang et al, 2009) possuem apenas informação relacionando a validação de um miRNA em um gene alvo, ou seja, não informa a localização do sítio nem o transcrito específico. Por isso as entradas desta base de dados foram inseridas no relacionamento **ValidatedSites** com os atributos de posição e região iguais a -1 (valor desconhecido), tipo de localização = 2 (apenas gene alvo) e evidência = 1 (gene).

Além disso, para cada par miRNA/gene destes bancos de dados, foram inseridas potencialmente várias entradas no relacionamento **ValidatedSites**, uma para cada refseq do respectivo gene, como descrito a seguir.

No caso do miRTarBase[23], para obter todos os refseqs de um dado gene (miRTarBase informa o Entrez GeneID), foram utilizadas as tabelas wgEncodeGencodeEntrezGeneV24 (que relaciona EntrezGeneID com ID's de transcritos do banco Ensembl) e wgEncodeGencodeRefSeqV24 (que relaciona ID's de transcritos do banco Ensembl com RefSeq), ambas obtidas do site da UCSC (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/).

[21] http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/refGene.txt.gz
[22] NM_138931.
[23] O arquivo fornecido por essa base possui duplicações de entradas, que foram removidas antes de seu processamento.

Figura 12- Parte do modelo conceitual do SIMTar que ilustra as as entidades MiRNAs, Refseqs e Databases, e o relacionamento triplo entre elas (Validated Sites).



Fonte: Paula Prieto Oliveira, 2018.

Já o miR2Disease informa o gene pelo seu nome (*gene symbol*), sendo que para muitas entradas é utilizado o *gene symbol* oficial e para outras um nome alternativo. Desta forma o primeiro passo foi realizar um mapeamento de um Entrez Gene ID com os vários nomes possíveis (*gene symbols* oficiais e sinônimos)[24] a fim de obter o Gene ID correto. Tal mapeamento é armazenado no SIMTar nas tabelas Genes e Synonyms (Figura 10).

Os dados do miRTarBase foram obtidos da versão 7 deste banco[25], que possui 502653 pares mir/gene humanos. O processamento descrito acima gerou 810681 entradas no SIMTar (pares mir/refseq).

O banco miR2Disease possui uma única versão disponível. Os dados, após processados e removidas as redundâncias, deram origem a 1710 pares mir/refseq.

Base de dados MtiBase

O banco de dados MtiBase não está mais disponível, mas foi possível obter 28 sítios alvos validados de miRNAs, em regiões de CDS e 5'UTR, no material suplementar[26] do artigo (Guo et al, 2015).

Tais sítios são descritos pelas suas coordenadas genômicas na versão hg19 do genoma humano. Desta forma:

1. obteve-se a sequência de nucleotídeos referentes a essas coordenadas (utilizando as sequências genômicas[27] da versão hg19);
2. obteve-se o código refseq do transcrito que localiza-se em tal coordenada, utilizando o mapeamento dos transcritos[28] na versão hg19;
3. obteve-se a posição de início e fim da sequência (obtida no passo 1) no referido transcrito (obtido no passo 2, via programa blastn do pacote blast+ (com parâmetro *word_size*=11). Desta forma, foram considerados apenas os refseqs cuja localização do sítio alvo deu-se em regiões exônicas.

Assim, as entradas desta base de dados foram inseridas no relacionamento **ValidatedSites** com os atributos de posição iguais às identificadas no passo 3, regiões iguais às informadas pelas própria base (e conferidas na tabela de mapeamento do refseqs), tipo de localização = 0 (posição exata) e evidência = 2 (posição genômica). No entanto, como o Mtibase baseia-se em validações sobre os RNAs transcritos e não DNA, o relacionamento foi feito com

[24] Informações obtidas de ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz
[25] Última versão do miRTarBase, de setembro de 2017.
[26] Tabela S2 do material suplementar.
[27] Obtidas de http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz.
[28] Obtido de http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/refGene.txt.gz

todas as entradas refseqs (da entidade **RefSeqs**) que contém o respectivo código refseq independente de sua localização genômica (isto é, para todos os números seriais), desde que tenha sido encontrado um match entre a sequência do sítio alvo e a sequência (exônica) do refseq. Tal processo deu origem a 154 entradas na tabela **ValidatedSites**, todas referentes a regiões CDS e 5'UTR.

Dados de CLASH: bases doRiNA e StarBase

Experimentos de CLASH (*Cross-linking, Ligation and Sequencing of Hybrids*) são experimentos que identificam interações RNA–RNA ocorridas em um complexo proteico. Os bancos de dados doRiNA (Anders et al, 2012; Blin et al, 2015) e starBase (Yang et al, 2011; Li et al, 2014) armazenam resultados de CLASH de ligação miRNA/RNA, e portanto fornecem também uma forma de validação dos sítios alvos.

Os dados de CLASH do banco doRiNA apresentam as posições iniciais e finais dos alvos, em coordenadas genômicas versão hg19, para refseqs específicos. As posições exatas de ligação dos miRNAs nos RNAs alvos, assim como as regiões nas quais se localizam, foram obtidas da seguinte forma:

1. primeiro verificou-se se as coordenadas do alvo eram consistentes[29] com o RefSeq informado, sendo o dado descartado caso não fosse;
2. obteve-se a sequência de nucleotídeos relativa às coordenadas genômicas informadas para o alvo (utilizando a versão hg19 do genoma humano);
3. mapeou-se essa sequência na sequência RNA RefSeq informada como alvo (versão atual) via programa blastn do pacote blast+ (com parâmetro *word_size*=11).

Alguns dos RefSeqs mencionados na base não existem mais na versão atual do banco RefSeq, tendo sido substituídos por outros. Estes casos foram manualmente verificados, atualizando-se o código do RNA alvo para o código RefSeq atual.

Dos 36996 dados de CLASH (pares refseq/miRNA), 33763 eram consistentes com relação às coordenadas de sítio alvo e RefSeq, e destes 30035 tiveram a posição alvo corretamente identificada e inseridos no SIMTar.

O StarBase (Yang et al, 2011; Li et al, 2014) não apresenta os refseqs. Para obtê-los, foi elaborado um programa que cruza as informações da tabela clashStarBase2.bed com as da tabela wgEncodeGencodeRefSeqV24 (GencodeID e RefSeq), pois ambas possuem GencodeID. As

[29] A planilha disponibilizada pelo doRiNA apresenta as coordenadas do gene alvo e do sítio alvo. Foram consideradas inconsistentes as entradas nas quais tais coordenadas diferiam na fita (3055 entradas) ou nas bordas dos genes e sítios, isto é, quando as coordenadas eram tal que o início do gene > início do sítio ou fim do gene < fim do sítio (212 entradas). Destes dois conjuntos, 35 entradas eram inconsistentes tanto quanto à fita quanto às bordas dos genes e sítios.

demais informações descritas acima também foram incluídas no programa. Este banco não apresenta informações de posição nem de região.

3.4 - Inclusão de dados de CLIP

Cross linking immunoprecipitation (CLIP) é uma técnica de biologia molecular que utiliza luz ultravioleta (UV) e imunoprecipitação para identificar ligações entre RNAs e RBPs (*RNA binding protein*s)[30] (Wang et al., 2015). A argonauta (AGO) é uma dessas proteínas, que participa do complexo RISC e é guiada pelo microRNA para se ligar ao seu sítio em um mRNA alvo, promovendo inibição ou até mesmo ativação deste (Wu e Belasco, 2008; Iwasaki e Tomari, 2009). Dessa forma, se os dados de CLIP apontam um sítio de ligação do mRNA à AGO, temos uma evidência de que esse mRNA é um sítio de ligação de um microRNA.

Como para aumentar a confiabilidade da predição de que um SNP interfere em um sítio de miRNA, precisa-se primeiro aumentar a confiabilidade de que um de seus alelos seja um alvo de miRNAs, dados de CLIP de RNAs ligados a Argonautas foram incorporados ao SIMTar.

Foi realizada uma busca na literatura e em sites disponíveis na Internet por bancos de CLIP relacionados às quatro proteínas argonautas (AGO1, AGO2, AGO3 e AGO4) que apresentem as coordenadas genômicas da região onde a proteína se ligou, em humanos.

Foram encontrados quatro bancos de CLIP nessas condições: AURA (Dassi et al, 2012; Dassi et al., 2014), CLIPdb (Yang et al., 2015), DoRiNA (Anders et al., 2012; Blin et al., 2015) e starBase (Yang et al., 2011; Li et al., 2014). Esses bancos se baseiam, total ou parcialmente, em um mesmo conjunto de artigos publicados de experimentos de CLIP, cujos dados (*data sets*) podem ser obtidos do banco de dados *Gene Expression Omnibus* (GEO) *Datasets* (Barrett et al., 2013) ou do material suplementar dos artigos de origem. Os artigos referem-se a estudos independentes, cada um deles utilizando diferentes tecidos ou linhagens celulares, compostos por uma ou várias amostras distintas sob diversos tratamentos, e para uma dada argonauta em

[30] Primeiramente, o material é irradiado com luz UV para a formação de ligações covalentes entre as proteínas e os RNAs (cross linking). Em seguida, a amostra é lisada e colocam-se nucleases para cortar os RNAs associados às proteínas em tamanhos de 50-150 nucleotídeos (Darnell, 2012). O lisado é limpo de ribossomos e então é feita a imunoprecipitação: o material é incubado com o anticorpo contra a proteína de interesse para precipitar o antígeno (Lenz, 2004; Darnell, 2012). Após a ligação do anticorpo ao antígeno (proteína + RNA), as proteínas não ligadas ao anticorpo são lavadas. O RNA associado à proteína é desfosforilado nos extremos 3' e 5' com fosfatase, e um RNA vinculador é ligado ao extremo 3' do mesmo por meio da RNA ligase, enquanto o extremo 5' é fosforilado pelo tampão PNK na presença de ATP (Darnell, 2012). Os complexos RNAs-proteínas são isolados dos RNAs livres por meio de gel SDS (dodecil sulfato de sódio) e membrana de transferência de nitrocelulose (Ascano et al., 2012; Darnell, 2012). A digestão com proteinase K é realizada para remover a proteína do complexo RNA-proteína. Após a ligação do RNA vinculador ao extremo 5' do RNA, este é transformado em cDNA por meio da transcrição reversa. O cDNA é amplificado através do PCR, e então sequenciado (Darnell, 2012).

particular. No entanto, a sobreposição desses bancos não é clara. AURA 2.0 (Dassi et al., 2014) é centrado em UTRs, e armazena os sítios nessas regiões exatamente como processados pelos autores dos estudos considerados. DoRiNA armazena dados de experimentos do próprio grupo de pesquisa e também de outros trabalhos publicados, sendo estes últimos armazenados como disponibilizados nos artigos originais. Starbase 2.0 (Li et al., 2014) também manteve os dados originais a menos dos dados brutos obtidos por meio da técnica de PAR-CLIP, que foram re-analisados utilizando as ferramentas FASTX-Toolkit e PARalyzer. CLIPdb armazena os resultados de suas reanálises dos dados brutos de todos os estudos considerados por eles e utiliza duas ferramentas de análise para cada amostra: Piranha e outra específica da tecnologia CLIP empregada no estudo da amostra. Piranha é uma ferramenta de identificação de sítios de CLIP que independe da tecnologia utilizada, porém não fornece a informação da fita genômica da qual o sítio se origina. Já as ferramentas tecnologia-específicas são capazes de identificar sítios fita-específicos: PARalyser para a técnica PAR-CLIP, CIMS para HITS-CLIP, CIMS/CITS para a técnica iCLIP. Também foi encontrado um banco de CLIP chamado CLIPZ, mas este se encontra desligado e não está mais disponível.

A fim de agregar toda essa informação ao SIMTar, os estudos de todos esses quatro bancos foram incorporados ao SIMTar[31] por meio do relacionamento **CLIPs**, entre **Databases** e **RNA_SNPCenteredWindows** (Figura 13). Ou seja, armazena-se a informação de qual janela de RNA (centrada em um SNP) possui um resultado de CLIP de acordo com um destes bancos de dados de CLIP. Este relacionamento tem como atributos as coordenadas (cromossomo, posição inicial, posição final e fita) na qual o CLIP foi identificado (valores de fita podendo ser “+”, “-” ou “.”, este último para resultados fita-inespecíficos) e a argonauta (AGO 1, 2, 3 ou 4).

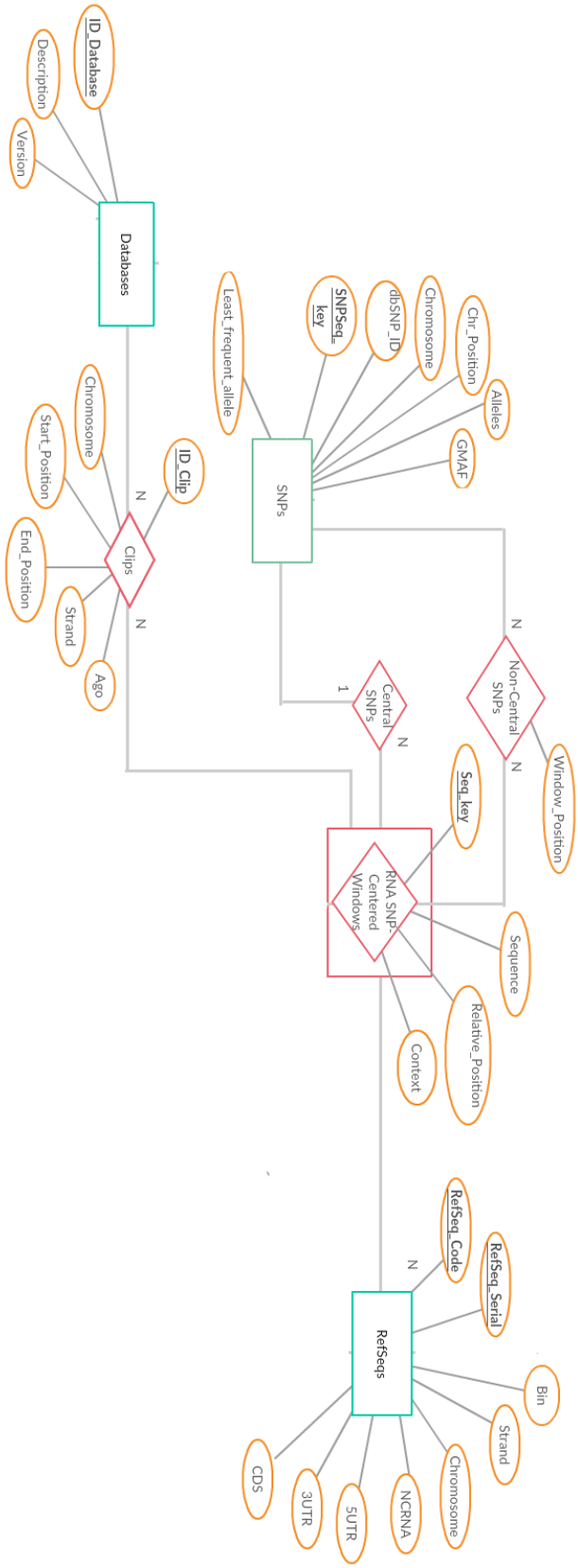
3.5 - Inclusão de dados de eQTL

Como já mencionado, *expression quatitative trait loci* (eQTLs) são variantes da sequência de DNA associadas a variações no nível de expressão de um gene (Albert e Kruglyak, 2015). Predições do SIMTar de interferências de SNPs que sejam eQTLs associados à variação de expressão do gene que possui tal SNP possuem maior evidência de serem verdadeiras.

Foram obtidos dados de eQTL dos bancos GTEX (Carithers e Moore, 2015), bloodEQTL [32], seeQtl (Xia et al., 2012) e bloodeqtlbrowser (Westra et al., 2013). Além disso foram também considerados resultados de estudos publicados (Myers et al., 2007; Stranger et al., 2007; Gibbs et

[31] Apenas os estudos de argonautas em amostras humanas.
[32] https://molgenis58.target.rug.nl/bloodeqtlbrowser/2012-12-21-CisAssociationsProbeLevelFDR0.5.zip

Figura 13- Parte do modelo conceitual do SIMTar que ilustra o relacionamento CLIPs e as entidades envolvidas nele.



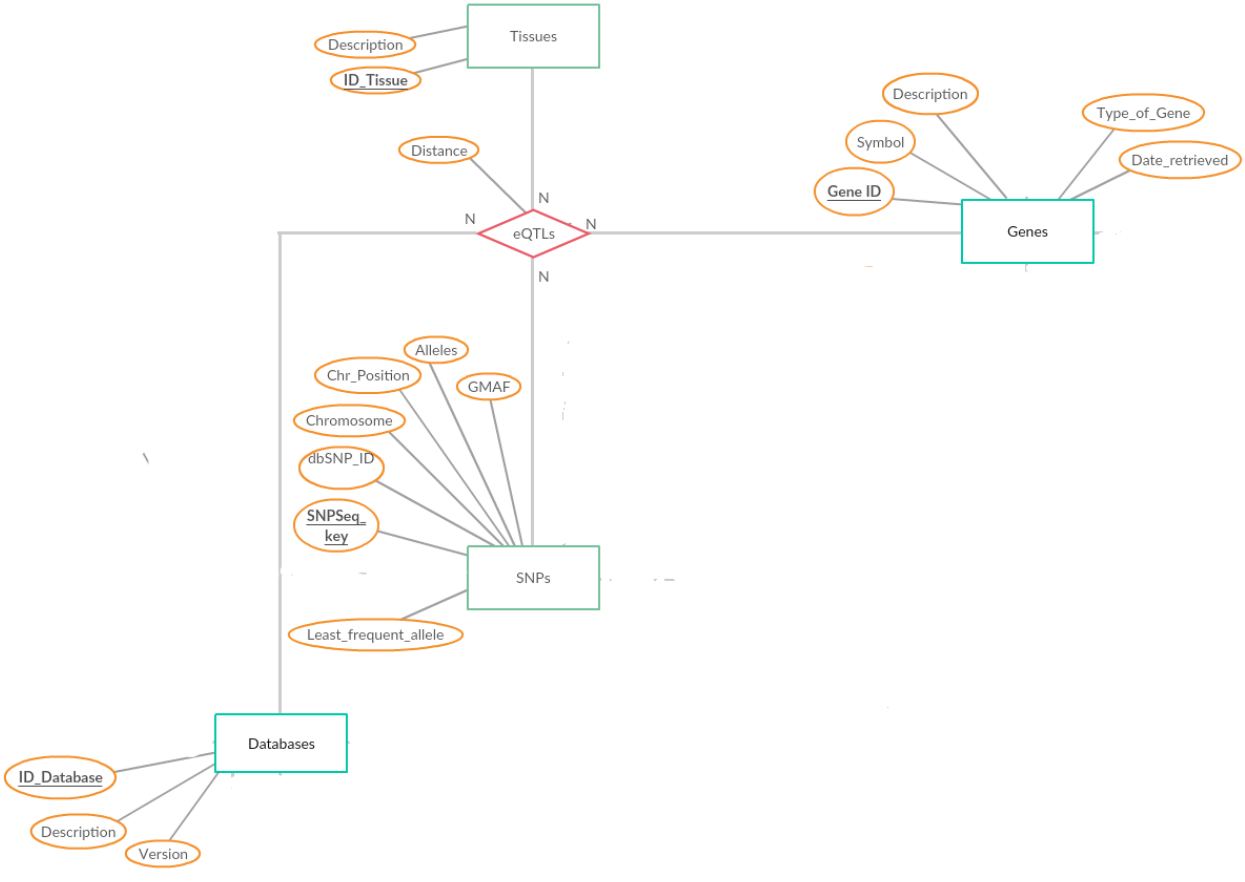
Fonte: Paula Prieto Oliveira, 2018.

al., 2010; Colantuoni et al., 2011; Liang et al., 2013) e dos seguintes websites encontrados: http://csg.sph.umich.edu/junding/eQTL/TableDownload/ e http://www.scandb.org

Foi elaborado um programa, para cada uma dessas fontes de dados, a fim de selecionar apenas os eQTLs que são SNPs e que estão localizados dentro do gene cuja expressão é alterada. Um SNP foi considerado “dentro do gene” quando ele se localiza entre o início mais *upstream* e o fim mais *downstream* dentre suas isoformas.

Além das informações sobre o SNP e gene, foi também armazenada a informação do tecido celular no qual foi realizado o experimento. A informação de tecido pode ser utilizada pelo usuário para priorizar SNPs de interesse. Por exemplo, um pesquisador da área de Psiquiatria pode desejar priorizar a investigação mais aprofundada de SNPs que sejam eQTLs em tecidos cerebrais. A figura 14 mostra a modelagem realizada, na qual **eQTLs** é um relacionamento quádruplo entre um SNP, um gene, um tecido (entidade **Tissues**) e um banco de dados.

Figura 14- Parte do modelo conceitual do SIMTar que ilustra as entidades Genes, Tissues, Databases e SNPs, e o relacionamento eQTLs entre elas.



Fonte: Paula Prieto Oliveira, 2018.

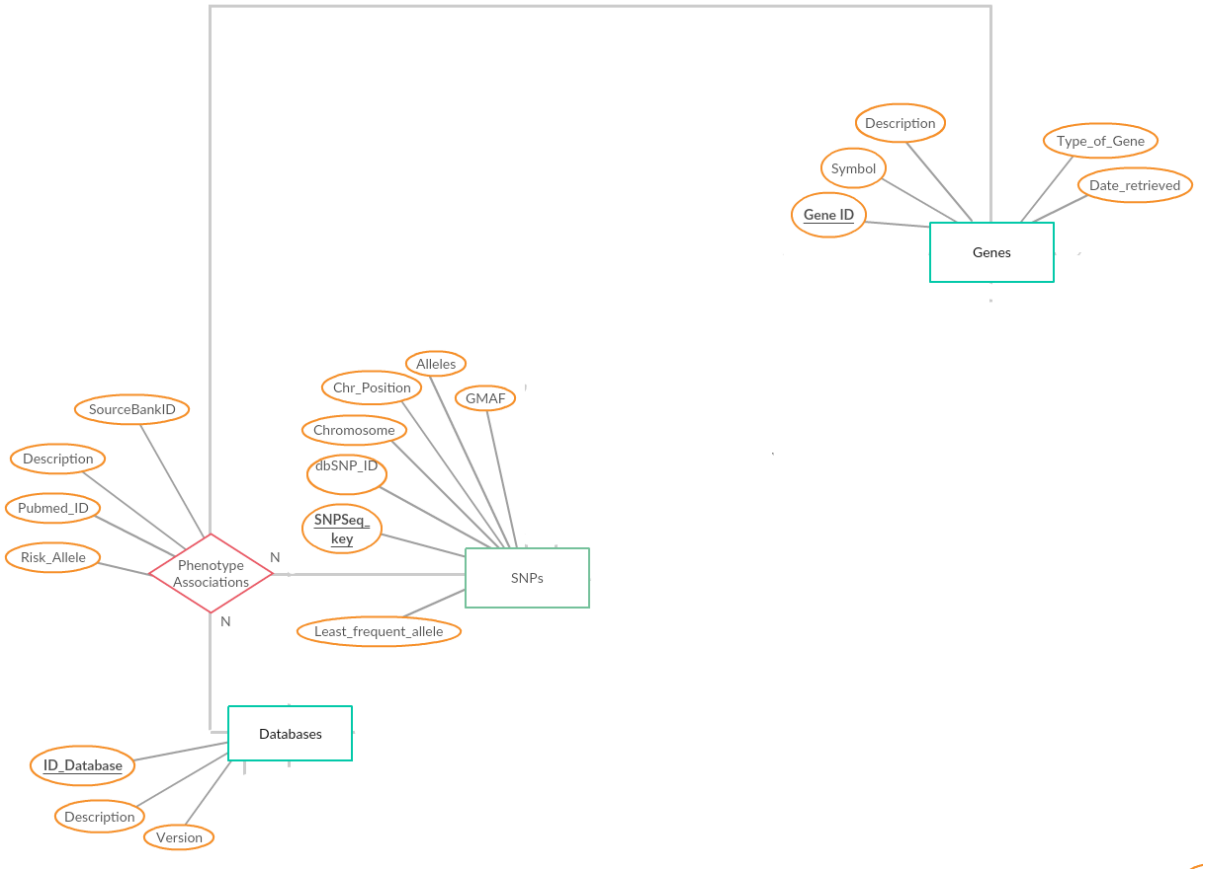
3.6- Incorporação de informação acerca de SNPs já associados a fenótipos de interesse

Se o usuário assim optar, o SIMTar poderá priorizar os SNPs com associação a algum fenótipo de interesse. SNPs envolvidos em estudos de associação foram obtidos das bases de dados GWASdb, GWAS catalog, Clinvar, dbDSM (*database of Deleterious Synonymous Mutation*), GRASP e PolymiRTS. Elaborou-se um programa para cada banco com o objetivo de selecionar apenas SNPs, o alelo risco caso a informação seja fornecida, e o fenótipo. As informações foram integradas ao SIMTar, como modelado na figura 15, por meio do relacionamento triplo, **PhenotypeAssociations**, entre SNPs, Genes e Databases. Diferente do que foi feito na modelagem de eQTL, em que os tecidos foram armazenados na entidade Tissue, aqui os fenótipos são diretamente descritos por um string no atributo “Description”. A decisão deste tipo de modelagem se deu pelo fato de não ter sido possível realizar uma padronização dos fenótipos, já que cada banco possui um nomenclatura própria. Sendo um campo texto, a busca por um fenótipo é realizada por *string matching*.

Em relação ao GWASdb, foram incluídos dois arquivos: gwasdb_20150819_snp_trait e gwasdb_20150819_snp_drug, com seleção das seguintes colunas: SNPID, PMID, GWAS_TRAIT, HPO_TERM, DO_TERM, DOLITE_TERM, MESH_TERM, EFO_TERM e RISK_ALLELE. Os termos de HPO_TERM, DO_TERM, DOLITE_TERM, MESH_TERM e EFO_TERM não são exclusivos do GWASdb, mas já utilizados na literatura: Human Phenotype Ontology (HPO), Disease Ontology (DO), Disease Ontology Lite (DOLite), Medical Subject Headings (MESH) e Experimental Factor Ontology (EFO). Para o gwasdb_20150819_snp_drug, foram selecionadas também mais duas colunas: DRUG_NAME e DRUG_ANNO. A DRUG_ANNO se refere à propriedades biológicas da droga.

Do GWAS catalog, foi utilizado apenas um arquivo, o gwas_catalog_v1.0.1-associations_e91_r2018-02-28.tsv, com inclusão das colunas PUBMEDID, DISEASE/TRAIT, CHR_ID, CHR_POS, STRONGEST SNP, RISK ALLELE MAPPED_TRAIT e MAPPED_TRAIT_URI. A coluna STRONGEST SNP se refere ao SNP mais fortemente associado à doença, a MAPPED_TRAIT indica o termo EFO (Experimental Factor Ontology) para a característica, e a MAPPED_TRAIT_URI é o link para o termo EFO.

Figura 15- Parte do modelo conceitual do SIMTar que ilustra as entidades Genes, Databases e SNPs, e o relacionamento Phenotype Associations entre elas.



Fonte: Paula Prieto Oliveira, 2018.

Clinvar[33] (Landrum et al, 2014; Landrum et al, 2016) é um banco que apresenta vários tipos de variações, não apenas SNPs. Para compor os SIMTar foram selecionados apenas as variações do tipo SNV (*single nucleotide variant*) com algum fenótipo associado especificamente ao SNV em questão[34]. Para isto foram utilizadas as seguintes colunas: cromossomo, posição, ID do Clinvar, RS, CLNDN, GENEINFO e CLINDISDB. A coluna RS se refere ao ID do SNP, a CLNDN é o nome preferido do Clinvar para a doença identificada no CLINDSDB, a GENEINFO indica o símbolo do gene da variante e o gene id, que estão separados por dois pontos. Já a CLINDSDB se refere aos bancos de doenças e seus respectivos números de identificação no banco. As colunas cromossomo e posição foram incluídas para a identificação das SNVs que não apresentam RS.

[33] Site do Clinvar: https://www.ncbi.nlm.nih.gov/clinvar/
[34] O banco ClinVar descreve fenótipos associados tanto a um SNV específico quanto a todo o haplótipo do qual o SNV participa.

Para o DSM, utilizaram-se as colunas SNP, Disease, Chr e Location. Em relação aos casos desprovidos de SNP id, foram selecionados apenas aqueles que apresentam a localização exata do SNP.

Em relação ao GRASP, incluíram-se as colunas SNPid(dbSNP134), PMID, Phenotype e PaperPhenotypeDescription; e para o PolymiRTS as colunas SNP e Disease.

3.7- Comparação do SIMTar com outras ferramentas correlatas

Uma ferramenta de predição de interferência de SNPs em sítios alvos de miRNAs é um classificador que, dado um SNP, classifica-o como positivo quando tal SNP interfere no sítio de algum miRNA e negativo caso contrário (isto é, não interfere no sítio de nenhum miRNA). Logo, medidas de desempenho desse classificador, como acurácia, precisão e revocação poderiam ser calculadas. No entanto, embora existam alguns poucos exemplos positivos (SNPs validados experimentalmente como interferindo no sítio de algum miRNA), não há como conhecer com certeza exemplos negativos. Isso porque não se pode nem precisar que um dado SNP não está localizado em um sítio de nenhum miRNA.

Na carência de exemplos negativos, a análise dos resultados do SIMTAR e sua comparação com as ferramentas correlatas foi realizada seguindo a seguinte estratégia:

a) identificou-se exemplos positivos a fim de calcularmos os números de verdadeiros positivos (VP) e falsos negativos (FN), permitindo o cálculo de sensibilidade para essa amostra positiva, dada por VP/(VP+FN);
b) criou-se várias listas de um certo número de SNPs sorteados aleatoriamente (cada lista com o mesmo número de SNPs da amostra positiva) e, para cada lista, calculou-se a proporção do número de resultados positivos / tamanho da lista. Estas proporções permitem visualizar uma distribuição desses valores que nos dão uma estimativa da significância da significância calculada no item a).

Detalhes destes dois passos são descritos a seguir.

3.7.1 - Obtenção da amostra positiva

Foi realizada uma revisão bibliográfica sistemática com o objetivo de obter SNPs na literatura que apresentam interferência, validada experimentalmente, em sítios alvos de miRNAs. O protocolo detalhado para a condução dessa revisão encontra-se no apêndice B.

Resumidamente, a busca pelos artigos foi realizada no portal Pubmed em maio de 2018, utilizando a seguinte *string* de busca: (miRNA* [Title/Abstract] OR microRNA* [Title/Abstract] OR

miR [Title/Abstract]) AND (target [Title/Abstract] OR "binding site" [Title/Abstract]) AND (SNP [Title] OR variation[Title] OR variant*[Title] OR allele [Title] OR polymorphism*). Foram aplicados os critérios de inclusão e exclusão definidos no protocolo sobre os resultados, primeiro pela leitura dos títulos e depois pela leitura dos resumos.

A busca no pubmed retornou 1156 artigos. Dos artigos incluídos foram identificados 267 pares gene-miRNA que apresentam interferência validada experimentalmente na ligação entre eles devido à presença de um SNP na região alvo do microRNA. Dois destes SNPs apresentam códigos rs (código no banco dbSNP) que não foram encontrados no dbSNP, e por isso foram descartados[35].

Realizado esse levantamento, foram identificados 208 SNPs distintos com validação experimental de estarem interferindo no sítio de ao menos um miRNA. Estes compõem a amostra positiva considerada nos testes.

Foi identificado que estes 208 SNPs estão distribuídos pelas seguintes regiões: nenhum na 5'UTR, 15 em CDS, 184 em 3' UTR e 9 em éxons de RNAs não codificantes (NCRNA)[36].

Esses 208 SNPs foram utilizados como entrada tanto do SIMTar como das ferramentas correlatas, como descrito nas seções 3.7.3 e 3.7.4, e a sensibilidade de cada uma dessas ferramentas foi estimada com base nesta amostra (número de SNPs relatados como interferindo em algum sítio de miRNA / 208).

3.7.2 - Geração das amostras aleatórias e estimação de densidade de positivos aleatórios

Foram obtidos do banco dbSNP[37] todos os SNPs de todos os cromossomos humanos oficiais 1 a 22, X e Y, não sendo consideradas suas versões alternativas, e selecionados aqueles presentes em cada região aqui considerada (5'UTR, CDS, 3'UTR e NCRNA).

A partir dos SNPs destas regiões, foram geradas 100 listas de SNPs aleatórios, sendo cada lista composta por 208 SNPs sorteados aleatoriamente mantendo a mesma proporção de região de origem observada na amostra positiva (nenhum na 5'UTR, 15 em CDS, 184 em 3' UTR e 9 em NCRNA).

---

[35] São eles rsrs5030762 e rs6186923.

[36] A informação de localização de um SNP foi obtida ao inspecionar nos arquivos dos cromossomos do dbSNP (formato ASN.1 flatfile) o campo *fxn-class* (classe funcional). Para valores "*utr-variant-5-prime*" foi atribuída a região 5'UTR, para valores "*synonymous-codon*", "*missense*", "*stop-lost*" e "*stop-gained*" foi atribuída a região CDS, para valores "*utr-variant-3-prime*" foi atribuída a região 3'UTR e para valores "*nc-transcript-variant*" foi atribuída a região de éxon de RNA não codificante.

[37] dbSNP versão 151, que apresenta coordenadas do genoma humano versão hg38.

Para cada uma destas 100 listas foram executados, para a maioria das ferramentas, os mesmos procedimentos de teste executados para a amostra positiva (detalhes na seção 3.7.4) e calculada uma taxa de positividade (número de SNPs relatados como interferindo em algum sítio de miRNA / 208). Os 100 valores de taxa de positividade foram plotados em um histograma e sobre ele estimada uma densidade não paramétrica. Tal densidade foi utilizada para calcular, para cada ferramenta testada, um p-valor para a sensibilidade estimada. Esse p-valor é interpretado como sendo a probabilidade daquele valor de sensibilidade ser observado puramente por chance em um conjunto de 208 SNPs advindos das mesmas regiões.

Ou seja, esta análise é um teste de hipótese, no qual testa-se a hipótese nula de que a taxa de positivos observada na amostra positiva (digamos x) é igual às taxas de positivos observadas em amostras aleatórias. Tal hipótese é rejeitada se o p-valor calculado para este x for menor, por exemplo, que 0.01.

Por exemplo, se uma ferramenta apresentou uma sensibilidade de 80%, o seu p-valor foi calculado como sendo a área debaixo da curva da densidade estimada a partir do histograma da taxa de positividade das 100 listas aleatórias. Neste exemplo fictício, o histograma gerado corresponde ao apresentado na figura 16, e a área debaixo desta curva para valores >= 80% corresponde ao p-valor = 0.0523 da sensibilidade de 80% observada na amostra positiva. Neste caso a hipótese nula não pode ser rejeitada, indicando que embora 80% seja um valor alto de sensibilidade, a significância não é a desejada, ou seja, a ferramenta tem a tendência de produzir resultados falso positivos.

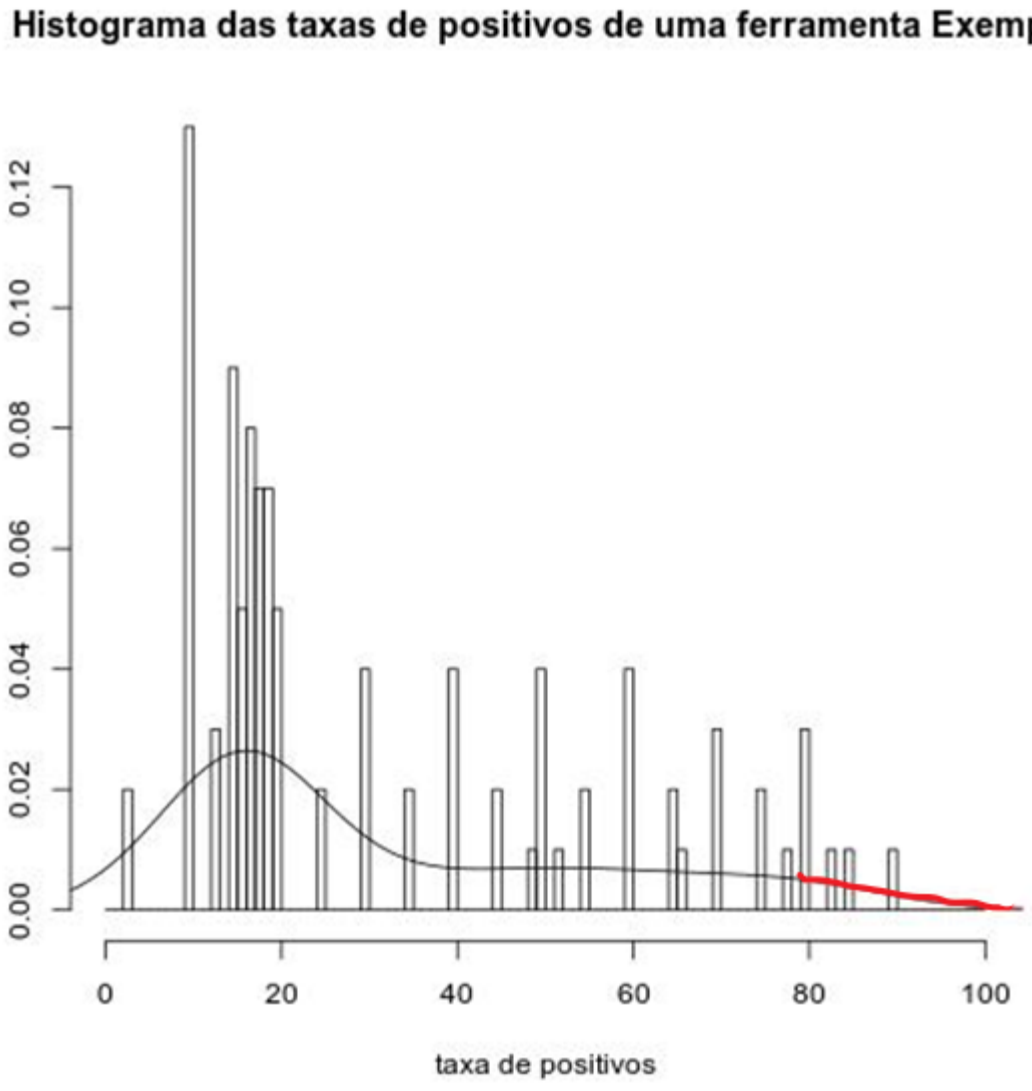
3.7.3 - Procedimento de obtenção de resultados do SIMTar

Cada lista de SNPs considerada (uma lista de SNPs validados e 100 listas de SNPs aleatórios) foi submetida ao SIMTar. Após a etapa de predição de interferência dos SNPs com base nos resultados da predição de sítios alvos em todos os alelos, três critérios de utilização das informações dos experimentos biológicos foram testados para a seleção final dos SNPs considerados como interferidores em sítios de miRNAs:

1. o SNP deve, obrigatoriamente, estar positivamente relacionado com ao menos uma das duas informações biológicas que relaciona um SNP a um miRNA (critério VC: validated OR clip);
2. o SNP deve, obrigatoriamente, estar positivamente relacionado com ao menos uma das quatro informações biológicas (critério VECP: validated OR eqtl OR clip OR phenotype);

3. nenhuma informação de experimentos biológicos é utilizada como critério eliminatório (critério none).

Figura 16- Exemplo de um histograma das taxas de positivos das 100 listas de SNPs aleatórios. A curva é a densidade estimada para os dados, sendo o seu trecho em vermelho correspondente a valores >= 80%.



Fonte: Paula Prieto Oliveira, 2018.

3.7.4 - Procedimento de obtenção de resultados das ferramentas correlatas

Enquanto as ferramentas MirSNP, Polymirts e SNPinfo permitem que o usuário forneça como entrada uma lista de SNPs, as ferramentas MiRNASNP, MicroSNiPer e Mirsnpscore permitem apenas a consulta de um SNP por vez. O SNPinfo, apesar disso de permitir *upload* de um arquivo com vários SNPs, apresenta os resultados na forma de um link para cada SNP, que ao ser seguido mostra as predições de cada alelo. Estes fatos inviabilizaram o teste das ferramentas SNPinfo, MiRNASNP, MicroSNiPer e Mirsnpscore, pois é necessário que cada ferramenta analise 21009 SNPs (208 validados + 20800 SNPs aleatórios).

Assim apenas as ferramentas MirSNP e Polymirts foram comparadas ao SIMTar. Um programa foi implementado para cada uma delas para interpretar os arquivos de saída.

MirSNP

O mirSNP está disponível no link http://bioinfo.bjmu.edu.cn/mirsnp/search/. A busca foi executada na opção “Batch search”. Há casos em que o SNP de entrada não é reconhecido pela ferramenta, tido como SNP inexistente. Para estes SNPs identificamos se havia um código sinônimo no dbSNP, e os testamos em seus lugares.

PolymiRTS

O PolymiRTS está disponível no site http://compbio.uthsc.edu/miRSNP/. A busca foi executada na opção “Batch search”, sem estipular nenhum valor de conservação mínima. Este banco de dados, ao invés de mostrar diretamente as ligações associadas ao número de referência do SNP digitado, ele mostra todos os SNPs que causam interferência e estão presentes na 3’UTR do gene na qual o SNP de entrada se localiza, identificando apenas um transcrito para cada gene. O resultado é uma lista com o pareamento de todos os microRNAs preditos que se ligam aos alelos de referência e aos SNPs.

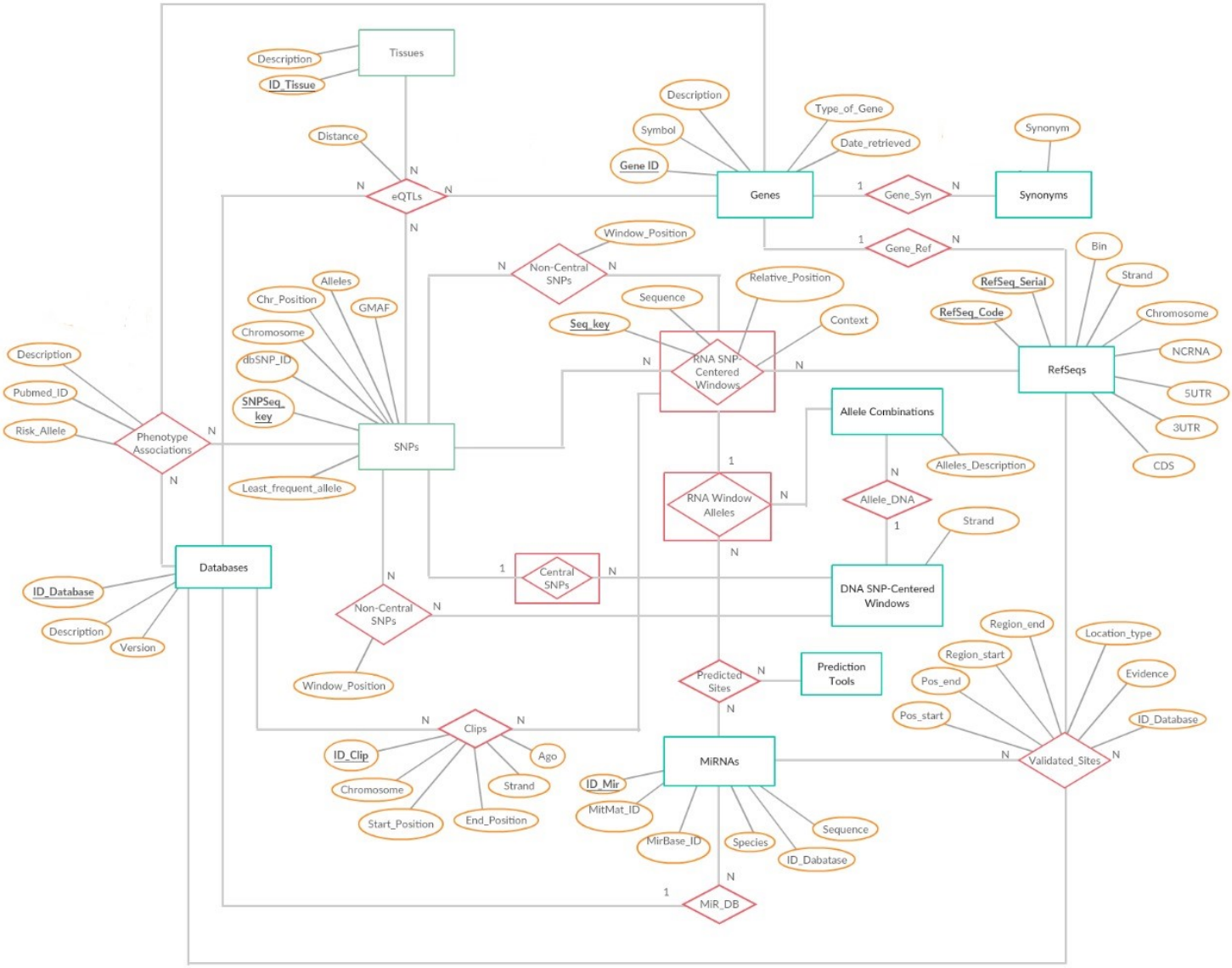
CAPÍTULO 4 - RESULTADOS E DISCUSSÕES

Este capítulo apresenta e discute os resultados do modelo final e dos testes tanto do SIMTar como de trabalhos correlatos.

4.1 - SIMTar: modelo conceitual

O novo SIMTar conta agora com um banco de dados relacional para armazenar todas as informações envolvidas e permitir a análise dos resultados. A figura 17 apresenta o modelo conceitual completo, detalhado no capítulo 3.

Figura 17: modelo conceitual completo do novo SIMTar.



Fonte: Paula Prieto Oliveira, 2018

O centro do modelo conceitual do SIMTar é RNA_SNPCenteredWindows, que modela a ocorrência de um SNP em um dado transcrito Refseq, e analisa sua interferência.

4.2- Comparação do novo SIMTar com outras ferramentas de predição de interferência de SNPs em sítios alvos de microRNAs

Esta seção apresenta os resultados do SIMTar, MirSNP e PolymiRTS com respeito à análise de 101 listas de SNPs, cada uma contendo 208 SNPs distintos: uma lista de SNPs experimentalmente validados e 100 listas de SNPs aleatórios (respeitando as mesmas proporções de região de origem) obtidas e testadas como descrito na seção 3.7.

A Tabela 5 apresenta na coluna “Sensibilidade” a porcentagem de SNPs validados identificados como positivos[38] por cada ferramenta. A coluna “p-valor” apresenta a significância desta sensibilidade calculado com base nos histogramas apresentados nas figuras 18 a 21 (como descrito na seção 3.7.2).

Os resultados corroboram a hipótese inicial deste projeto, de que a incorporação de informações de diferentes tipos de experimentos relacionados ao problema pode diminuir a expectativa de falsos positivos sem diminuir significativamente a sensibilidade. De fato, ao utilizar o critério 3 (apenas as predições baseadas nos sítios identificados pelo TargetScan, sem considerar nenhuma informação de experimentos biológicos) o SIMTar identificou 100% dos SNPs validados mas também relatou como positivos praticamente todos os SNPs aleatórios, resultando em um p-valor de 0.99. Ou seja, a probabilidade de obter 100% de sensibilidade puramente por chance em um conjunto de 208 SNPs advindos das mesmas regiões é de 99%. Ao incorporar os quatro tipos de experimentos biológicos (critério 2 - VECP), a sensibilidade caiu para 83.2%, mas com um p-valor = 0. Além disso, mesmo com essa queda de sensibilidade, 83.2% ainda é maior que as sensibilidade das outras duas ferramentas testadas, MirSNP (76.4%) e PolymiRTS (59.6%).

Ao aplicar o critério 1 (VC - apenas SNPs com alvos validados ou com dados de CLIP), a sensibilidade do SIMTar cai para 30.3%, menor que a sensibilidade das duas outras ferramentas correlatas. Embora o p-valor obtido com o critério 2 seja também baixo, os SNPs identificados sob o critério 1 (VC) tendem a ser mais confiáveis por possuirem validação experimental de serem alvos de miRNAs. Estes tendem a ser priorizados para uma validação experimental de interferência do SNP. Por esse motivo, independente do critério escolhido pelo usuário, os resultados apresentados pelo SIMTar sumarizam todas estas características.

[38] Lembrando que, no contexto deste trabalho, um SNP positivo segundo uma ferramenta é um SNP em que a ferramenta identificou que ele interfere no sítio de algum miRNA, e um SNP validado é um SNP que foi validada sua interferência no sítio de algum miRNA.

Cabe destacar também que o PolymiRTS analisa apenas SNPs em regiões *seed* do alvo. Já o SIMTar tem a vantagem de analisar o SNP independente de sua localização no sítio alvo.

Outra questão refere-se à facilidade de uso da ferramenta para muitos SNPs. Muitas das ferramentas correlatas exigem que a pesquisa seja feita por um SNP apenas, dificultando a análise de um conjunto de SNPs de tamanho mesmo que moderado.

Foi também observado que muitos SNPs não foram encontrados nos bancos das ferramentas correlatas testadas neste trabalho, tendo havido a necessidade de buscar IDs sinônimos no banco dbSNP. O SIMTar armazena dados de todos os SNPs presentes na última versão do dbSNP, incluindo sinônimos.

Tabela 5: resultados do SIMTar e ferramentas correlatas.

Ferramenta	Sensibilidade	p-valor
SIMTar (critério 1 - VC) SIMTar (critério 2 - VECP) SIMTar (critério 3 - none)	30.3% **83.2%** 100%	0 0 0.99
MirSNP	76.4%	0
PolymiRTS	59.6%	0

Figura 18: Histograma da taxa de positivos do SIMTar (critério 1 - VC, que define que o SNP deve, obrigatoriamente, estar positivamente relacionado com um ao menos uma das duas informações biológicas que relaciona um SNP a um miRNA).

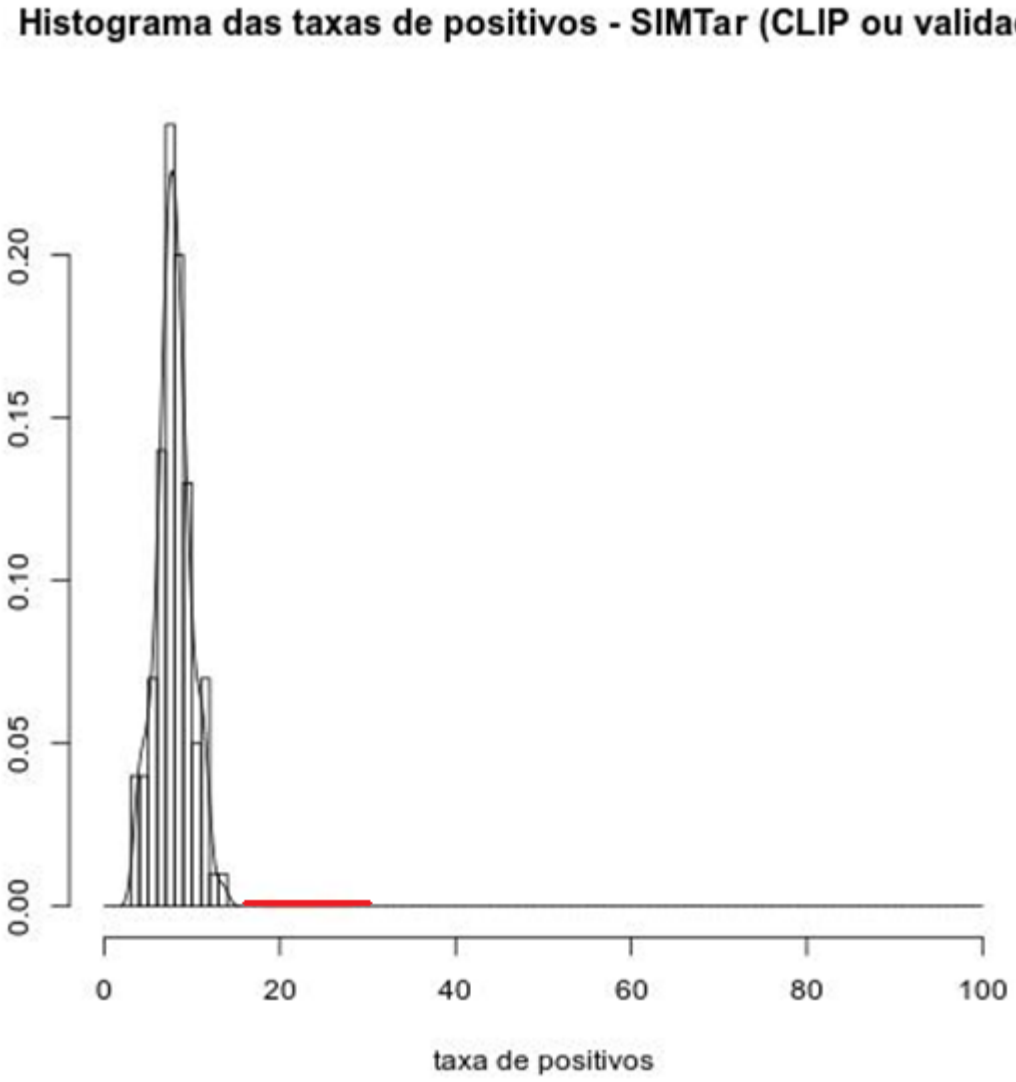


Figura 19: Histograma da taxa de positivos do SIMTar (critério 2 - VECP, que define que o SNP deve, obrigatoriamente, estar positivamente relacionado a alguma informação dos experimentos biológicos).

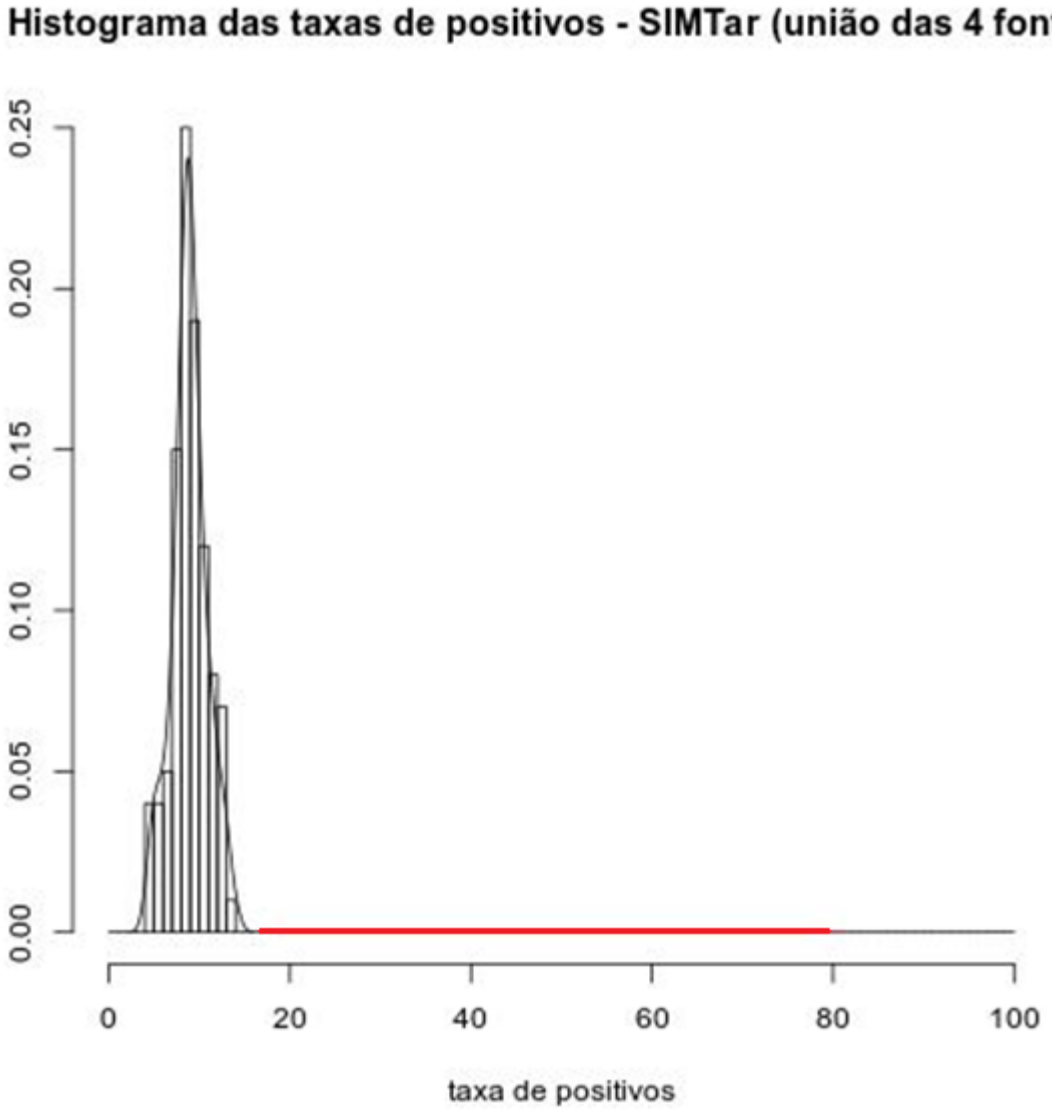


Figura 20: Histograma da taxa de positivos do MirSNP.

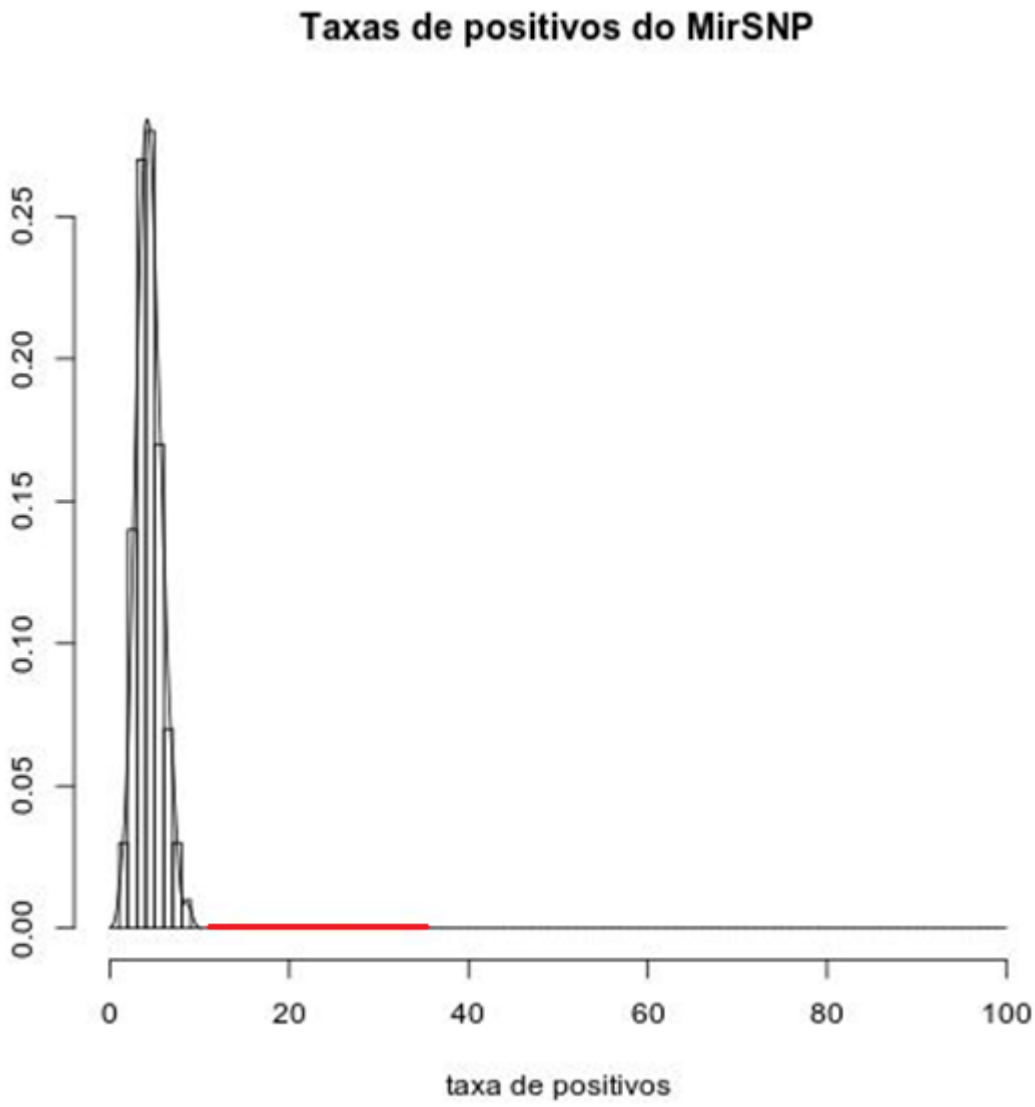
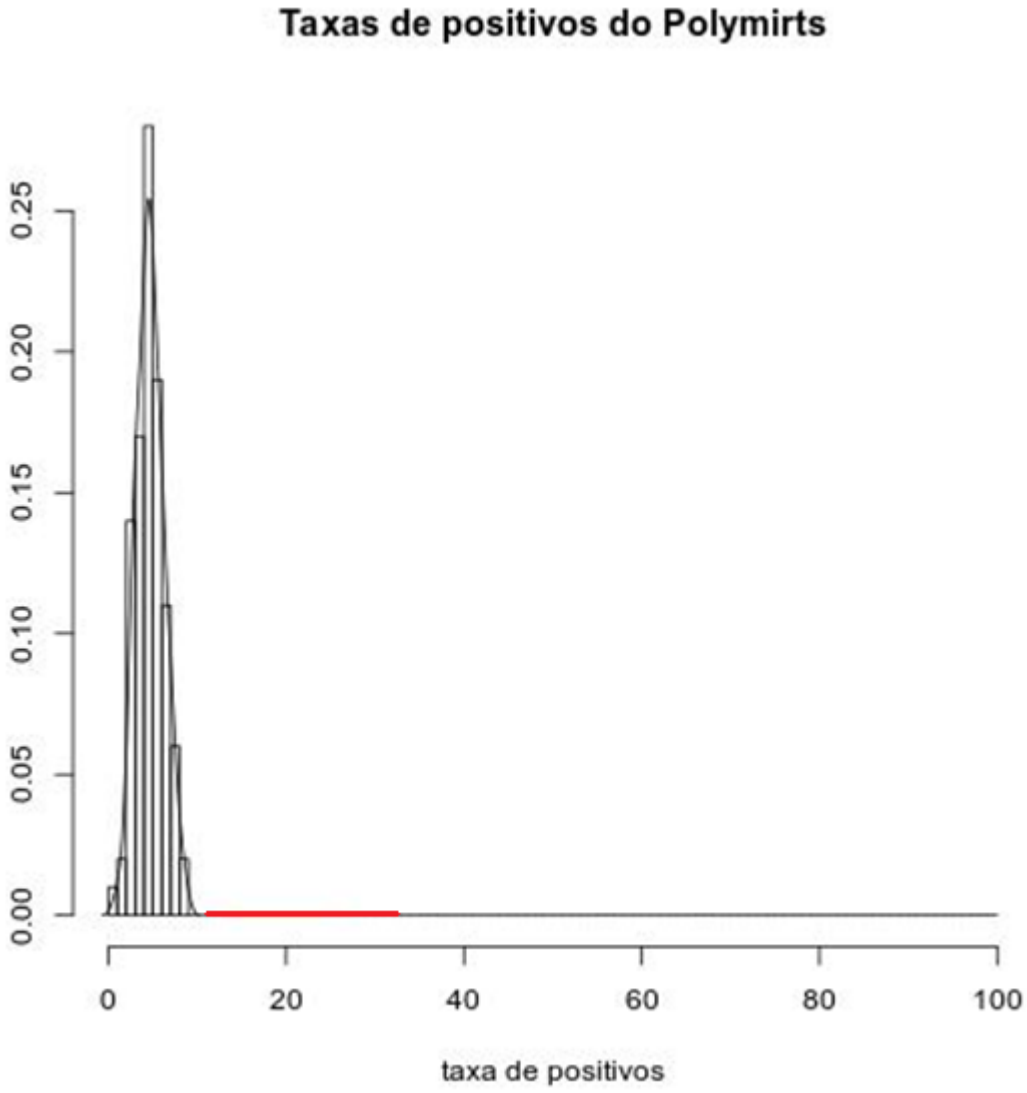


Figura 21: Histograma da taxa de positivos do PolymiRTS.



CAPÍTULO 5 - CONCLUSÃO

Este trabalho apresentou o projeto de remodelagem do SIMTar, uma ferramenta dedicada à identificação de SNPs interferindo em sítios alvos de miRNAs, com o intuito de diminuir a expectativa de falsos positivos em relação à sua versão anterior. Para atingir este objetivo, foram incorporados um novo programa de predição de sítios de miRNAs e resultados de várias fontes distintas de evidências biológicas que corroboram as predições dos sítios alvos de miRNAs nas regiões onde se localizam os SNPs, as associações destes SNPs com a alteração da expressão dos transcritos nos quais se localizam e as associações destes SNPs com fenótipos de interesse.

Para permitir a escolha do novo programa de predição de sítios de miRNAs, foi realizada uma revisão bibliográfica sistemática das ferramentas computacionais existentes com esse propósito, apresentada no Capítulo 2.

A hipótese testada neste trabalho foi de que a incorporação de informações de diferentes tipos de experimentos relacionados ao problema pode diminuir a expectativa de falsos positivos sem diminuir significativamente a sensibilidade. De fato, os resultados mostraram que tal hipótese mostrou-se verdadeira, tendo a incorporação dessas informações diminuído a sensibilidade de 100% (mas com p-valor de 0.99) para 83.2% (mas com p-valor de 0). Tal sensibilidade foi a maior dentre as ferramentas correlatas testadas.

O SIMTar se restringe apenas a analisar SNPs que estão presentes em sítios alvos de microRNAs e não avalia SNPs localizados em miRNAs.

5.1- Principais contribuições

A principal contribuição deste trabalho é a nova versão do SIMTAR capaz de identificar SNPs interferindo em sítios de miRNAs com maior confiabilidade, integrando uma variedade de informações biológicas não contemplada em outras ferramentas correlatas. Espera-se com isso que o SIMTar auxilie a comunidade científica a, por exemplo, priorizar SNPs a serem investigados no estudo de diversas doenças.

O SIMTar também foi planejado para ser de fácil uso pelos usuários. Permite que a entrada seja um arquivo com vários SNPs e disponibiliza todos os resultados positivos em forma de uma tabela. Também permite que o usuário escolha os critérios (eliminatórios ou classificatórios) a serem aplicados sobre os dados (e a operação binária entre eles - “E” ou “OU”, para fazer a intersecção ou a união das informações, respectivamente).

Além disso, este trabalho traz uma revisão bibliográfica sistemática das ferramentas computacionais de predição de sítios de miRNAs. Desta revisão foi escrito um artigo e submetido para a revista “Journal of Computational Biology and Bioinformatics Research (JCBBR)”, a qual mostrou interesse no artigo e solicitou-nos um *major review*.

Também está sendo escrito um artigo com os resultados desse trabalho que será submetido em breve.

5.2- Trabalhos futuros

No momento o banco do SIMTar só é acessado via linha de comando. Em breve será criada uma interface gráfica para acesso remoto na qual o usuário poderá fornecer sua lista de SNPs (*upload* de um arquivo), selecionar os principais parâmetros e receber os resultados tanto para *download* quanto para por meio de uma visualização gráfica dos resultados.

Há uma inconsistência da nomenclatura utilizada na literatura para identificar as doenças, pois *Genome-wide association studies* (GWASs) e bancos de fenótipos diferentes utilizam termos distintos para identificar o mesmo traço. Por essa razão, é necessário integrar as diversas descrições para facilitar o estudo de snps relacionados a uma determinada característica. Uma alternativa é utilizar o software MapIn (http://jjwanglab.org/mapin/, não publicado), descrito no artigo do Gwasdb v2 (Li et al, 2015). Esta ferramenta é capaz de calcular similaridades entre strings, fazendo o mapeamento de várias nomenclaturas a ontologias bem definidas, como Human Phenotype Ontology (HPO), Disease Ontology (DO) e Disease Ontology Lite (DOLite).

A atual versão do SIMTar só lida com SNPs ocorrendo em regiões exônicas. No entanto sabe-se que miRNAs também se ligam tanto em outras regiões do transcrito (íntrons) quanto no DNA (região promotora, por exemplo). Logo, SNPs nessas regiões poderiam também interferir em tais sítios. Para analisar tais SNPs o SIMTar teria que ser expandido para análise em íntrons e em DNA, e o usuário poderia definir onde analisar o efeito dos SNPs (múltipla escolha): RNA exônico, RNA intrônico, DNA.

REFERÊNCIAS BIBLIOGRÁFICAS

Ahmadi H, Ahmadi A, Azimzadeh-Jamalkandi S, Shoorehdeli MA, Salehzadeh-Yazdi A, Bidkhori G, Masoudi_Nejad A. HomoTarget: A new algorithm for prediction of microRNA targets in Homo sapiens. Genomics. 2013 Feb;101(2):94-100.

Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015 Apr;16(4):197-212.

Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG. Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics. 2009 Dec 1;25(23):3049-55.

Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). Hum Mutat. 2011 May;32(5):564-7.

Anders G1, Mackowiak SD, Jens M, Maaskola J, Kuntzagk A, Rajewsky N,et al.. doRiNA: a database of RNA interactions in post-transcriptional regulation. Nucleic Acids Res. 2012 Jan;40(Database issue):D180-6.

Ascano M, Hafner M, Cekan P, Gerstberger S, Tuschl T. Identification of RNA–protein interaction networks using PAR-CLIP. Wiley Interdiscip Rev RNA. 2012 Mar-Apr;3(2):159-77.

Azlan A1, Dzaki N1, Azzam G2. Argonaute: The executor of small RNA function. J Genet Genomics. 2016 Aug 20;43(8):481-94.

Bandyopadhyay S, Ghosh D, Mitra R, Zhao Z. MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets. Sci Rep. 2015 jan 23;5:8004.

Bandyopadhyay S, Mitra R. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. Bioinformatics. 2009 Oct 15;25(20):2625-31.

Bao L, Zhou M, Wu L, Lu L, Goldowitz D, Williams RW, et al.. PolymiRTS Database: linking polymorphisms in microRNA target sites with complex. Nucleic Acids Res. 2007 Jan;35(Database issue):D51-4.

Barenboim, M., Zoltick, B. J., Guo, Y. & Weinberger, D. R. MicroSNiPer: a web tool for predicion of snp e_ects on putative microRNA targets. Human Mutation 2010;31(11):1223–1232.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M et al. NCBI GEO: arquive for funcional genomics data sets-update. Nucl Acids Res. 2013;41:D991-5.

Bartel DP. MicroRNAs: genomics, biogenesis, mechanism and function. Cell. 2004 Jan 23;116(2):281-97.

Bartel DP. MicroRNAS: target recognition and regulatory functions. Cell. 2009 jan 23;136(2):215-33.

Bhattacharya A, Ziebarth JD, Cui Y. PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. Nucleic Acids Res. 2014 Jan;42(Database issue):D86-91.

Bickerton GR1, Higueruelo AP, Blundell TL. Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the PICCOLO database. BMC Bioinformatics. 2011 Jul 29;12:313.

BIOLCHINI, J.; MIAN, P.G.; NATALI, A.C.C.; TRAVASSOS, G.H. Systematic review in software engineering. Technical Report, Systems Engineering and Computer Science Department COPPE / UFRJ, Rio de Janeiro, 2005. Disponível em: < http://disciplinas.stoa.usp.br/pluginfile.php/92788/course/section/27982/Biolchini2005_Systematic_Review_in_Software_Engineering.pdf >.

Blin K, Dieterich C, Wurmus R, Rajewsky N, Landthaler M, Akalin A. DoRiNA 2.0--upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. Nucleic Acids Res. 2015 Jan;43(Database issue):D160-7.

Boutla A, Delidakis C, Tabler M. Developmental defects by antisense-mediated inactivation of micro-RNAs 2 and 13 in Drosophila and the identification of putative target genes. Nucleic Acids Res. 2003 Sep 1;31(17):4973-80.

Brennecke J, Hipfner DR, Stark A, Russell RB, Cohen SM. bantam Encodes a developmentally regulated miRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. Cell. 2003 Apr 4;113(1):25-36.

Brennecke J1, Stark A, Russell RB, Cohen SM. Principles of microRNA-target recognition. PLoS Biol. 2005 Mar;3(3):e85.

Brown TA. Genomes 4. 4. ed. New York: Garland Science; 2018.

Bruno, A. E., Li, L., Kalabus, J. L. & et al. miRdSNP: a database of disease-associated SNPs and microRNA target sites on 3'UTRs of human genes. BMC Genomics 2012;13:14.

Burgler C, Macdonald PM. Prediction and verification of microRNA targets by Moving Targets, a highly adaptable prediction method. BMC Genomics. 2005 jun 8;6:88.

Bush WS, Moore JH. Chapter 11: Genome-wide association studies.PLoS Comput Biol. 2012;8(12):e1002822.

Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. Biopreserv Biobank. 2015 Oct;13(5):307-8.

Carthew RW, Sontheimer EJ. Origins and mechanisms of miRNAs and siRNAs. Cell. 2009;136(4):642–55.

Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, et al.. MINT, the molecular interaction database: 2009 update. Nucleic Acids Res. 2010 Jan;38(Database issue):D532-9.

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, et al.. MINT: the Molecular INTeraction database. Nucleic Acids Res. 2007 Jan;35(Database issue):D572-4.

Chaves BB. Estudo do algoritmo adaboost de aprendizagem de máquina aplicado a sensores e sistemas embarcados [dissertação]. São Paulo: Universidade de São Paulo; 2012.

Chaudhuri K, Chatterjee R. MicroRNA Detection and Target Prediction: Integration of Computational and Experimental Approaches. DNA Cell Biol. 2007 May;26(5):321-37. https://doi.org/10.1089/dna.2006.0549

Chen K, Song F, Calin GA, Wei Q, Hao X, Zhang W. Polymorphisms in microRNA targets: a gold mine for molecular epidemiology. Carcinogenesis 2008;29(7):1306-11.

Chenna R, Sugawara H, Koike T,Lopez R, Gibson TJ, Higgins DG, et al.. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003, 31(13): 3497-3500.

Cheung VG, Spielman RS. Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat Rev Genet. 2009 Sep;10(9):595-604.

Chi SW, Zang JB, Mele A, Darnell RB: Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature 2009, 460(7254):479-486.

Chi SW, Hannon GJ, Darnell RB. An alternative mode of microRNA target recognition. Nat Struct Mol Biol. 2012 feb 12;19(3):321-7.

Cho S, Jang I, Jun Y, Yoon S, Ko M, Kwon Y et al. MiRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting. Nucleic Acids Res. 2013 Jan;41(Database issue):D252-7.

Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, Lee WH et al.. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res. 2016 Jan 4;44(D1):D239-47.

Chu CY, Rana TM. Small RNAs: regulators and guardians of the genome. J Cell Physiol. 2007;213:412-19.

Colantuoni C, Lipska BK, Ye T, Hyde TM, Tao R, Leek JT, et al.. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature. 2011 Oct 26;478(7370):519-23.

Coronnello C, Benos PV. ComiR: Combinatorial microRNA target prediction tool. Nucleic Acids Res. 2013 jul;41(Web Server issue):W159-64.

da Silva IF. Análise comparativa de ferramentas computacionais de predição de alvos de microRNAs (monografia). São Paulo: Universidade de São Paulo; 2013.

Darnell R. CLIP (cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein. Cold Spring Harb Protoc. 2012 Nov 1;2012(11):1146-60.

Dassi E, Malossini A, Re A, Mazza T, Tebaldi T, Caputi L et al.. AURA: Atlas of UTR Regulatory Activity. Bioinformatics. 2012 Jan 1;28(1):142-4.

Dassi E, Re A, Leo S, Tebaldi T, Pasini L, Peroni D. AURA 2: Empowering discovery of post-transcriptional networks. Translation (Austin). 2014 Jan 29;2(1):e27738.

Deveci M, Catalyürek UV, Toland AE.mrSNP: software to detect SNP effects on microRNA binding. BMC Bioinformatics. 2014 Mar 15;15:73.

Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol. 2005 Apr;3(2):185-205.

Doench JG, Sharp PA. Specificity of microRNA target selection in translational repression. Genes Dev. 2004 Mar 1;18(5):504-11.

Dweep H, Sticht C, Pandey P, Gretz N. miRWalk--database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. J Biomed Inform. 2011 Oct;44(5):839-47.

Dweep H, Sticht C, Gretz N. In-Silico Algorithms for the Screening of Possible microRNA Binding Sites and Their Interactions. Curr Genomics. 2013 Apr;14(2):127-36.

Dweep H, Gretz N, Sticht C. miRWalk database for miRNA-target interactions. Methods Mol Biol. 2014;1182:289-305.

Elemento O, Tavazoie S. Fastcompare: a nonalignment approach for genome-scale discovery of DNA and mRNA regulatory elements using network-level conservation. Methods Mol Biol. 2007;395:349-66.

Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila.Genome Biol.2003;5(1):R1.

Fan X, Kurgan L. Comprehensive overview and assessment of computational prediction of microRNA targets in animals. Brief Bioinform. 2015 Sep;16(5):780-94.

França NR, Júnior DM, Lima AB, Pucci FVC, Andrade LEC, Silva NP. Interferência por RNA: uma nova alternativa para terapia nas doenças reumáticas. Rev Bras Reumatol. 2010;50(6):695-709.

Fransen K, Visschedikj MC, van Sommeren S, Fu JY, Franke L, Festen EA et al. Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. Hum Mol Genet. 2010 sep 1;19(17):3482-8.

Friedman, R. C., Farh, K. K., Burge, C. B., Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Research 2009 jan;19(1):92–105.

Ghoshal A, Shankar R, Baqchi S, Grama a, Chaterji S. MicroRNA target prediction using thermodynamic and sequence curves. BMC Genomics. 2015 Nov 25;16:999.

Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, et al.. Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. PLoS Genet. 2010 May 13;6(5):e1000952.

Gong, J., Tong, Y., Zhang, H. M. & et al. Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. Human Mutation 2012;33:254–63.

Gong J, Liu C, Liu W, Wu Y, Ma Z, Chen H et al.. An update of miRNASNP database for better SNP selection by GWAS data, miRNA expression and onlinetools. Database (Oxford). 2015 Apr 15;2015:bav029.

Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele P, Lim LP, Bartel DP. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. Molecular Cell 2007;27(1):91–105.

Guildiyal M, Zamore P. Small silencing RNAs: an expanding universe. Nat Rev Genet 2009 feb;10(2):94-108.

Gumienny R, Zavolan M. Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. Nucleic Acids Res. 2015 Oct 15;43(18):9095.

Guo L, Du Y, Chang S, Zhang K, Wang J. rSNPBase: a database for curated regulatory SNPs. Nucleic Acids Res. 2014 Jan;42(Database issue):D1033-9.

Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P et al.. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell. 2010 Apr 2;141(1):129-41.

Hansen BG, Halkier BA, Kliebenstein DJ. Identifying the molecular basis of QTLs: eQTLs add a new dimension. Trends Plant Sci. 2008 Feb;13(2):72-7.

He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet. 2004 Jul;5(7):522-31.

Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell. 2013 Apr 25;153(3):654-65.

Hiard S, Charlier C, Coppieters W, Georges M, Baurain D. Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. Nucleic Acids Res. 2010 Jan;38(Database issue):D640-51.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S., Tacker M, Schuster P. .Fast folding and comparison of RNA secondary structures. Monatshefte f. Chemie. 1994 125(2):167–188.

Hofacker IL. Vienna RNA secondary structure server. Nucleic Acids Res 2003 31(13):3429–3431.

Holte RC. Very Simple Classication Rules Perform Well on Most Commonly Used Datasets. Machine Learning 1993;11:63-91.

Hsu SD1, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, et al.. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res. 2011 Jan;39(Database issue):D163-9.

Hsu SD[1], Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res.2014 Jan;42(Database issue):D78-85.

Huang Y, Zou Q, Song H, Song F, Wang L, Zhang G, Shen X. A study of miRNAs targets prediction and experimental validation. Protein Cell.2010 Nov;1(11):979-86.

Huang, V., Place, R. F., Portnoy, V. & et al. Upregulation of cyclin B1 by miRNA and its implications in cancer. Nucleic Acids Research 2012; 40:1695–1707.

Incarnato D, Neri F, Diamanti D, Oliviero S. MREdictor: a two-step dynamics interaction model that accounts for mRNA accessibility and Pumilio binding accurately predicts microRNA targets. Nucleic Acids Res. 2013 Oct;41(18):8421-33.

Iwasaki S, Tomari Y. Argonaute-mediated translational repression (and activation). Fly 2009 Jul-Sep;3(3):204-6.

Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X et al.. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. 2009 Jan;37(Database issue):D98-104.

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets.PLoS Biol. 2004 nov;2(11):1862-79.

Johnson JM1, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, et al.. Genome-wide survey of human alternative pre-mRNAsplicing with exon junction microarrays.Science. Science. 2003 Dec 19;302(5653):2141-4.

Johnson A, Raff Lewis, Walter R. Biologia Molecular da Célula. 5.ed. Porto Alegre: Artmed; 2010.

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, et al.. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science. 2003 Dec 19;302(5653):2141-4.

Jubb HC, Pandurangan AP, Turner MA, Ochoa-Montaño B, Blundell TL, Ascher DB. Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol. 2016 Nov:1-11.

Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet. 2007 ct;39(10):1278-84.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al.. Human Protein Reference Database--2009 update. Nucleic Acids Res. 2009 Jan;37(Database issue):D767-72.

Khorshid M, Rodak C, Zavolan M. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. Nucleic Acids Res. 2011 Jan;39(Database issue):D245-52.

Khorshid M, Hausser J, Zavolan M, van Nimwegen E. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. Nat Methods. 2013 Mar;10(3):253-5.

Kim S, Cho H, Lee D, Webster MJ. Association between SNPs and gene expression in multiple regions of the human brain. Transl Psychiatry. 2012 May 8;2:e113.

Kim SK, Nam JW, Rhee JK, Lee WJ, Zhang BT. miTarget: microRNA target gene prediction using a support vector machine. BMC Bioinformatics. 2006 Sep 18;7:411.

Kim VN, Han J, Siomi MC. Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol. 2009 feb;10(2):126-39.

Kiriakidou, M., Nelson, P. T., Kouranov, A, Fitziev P, Bouyioukos C, Mourelatos Z, Hatzigeorgiou A. A combined computational experimental approach predicts human microRNA targets. Genes Development 2004;18(10):1165–1178.

Knuth DE. The Art of Computer Programming: Seminumerical Algorithms II, Volume 2. Boston, MA: Addison-Wesley; 2014.

Krek, A., Gr¨un, D., Poy, M. N, Wolf R, Rosenberg L, Epstein EJ et al. Combinatorial microRNA target predictions. Nature genetics 2005;37(5):495–500.

Krol J, Loedige I, Filipowicz W. The widespread regulation of microRNA biogenesis, function and decay.Nat RevGenet. 2010 Sep;11(9):597-610.

Krüger J, Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast, and flexible. Nucleic Acids Research 2006;34(Web Server issue):W451-4.

Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. Gene expression levels are a target of recent natural selection in the human genome. Mol Biol Evol. 2009 Mar;26(3):649-58.

Kwok PY. SNPs: Why do we care? In: Single Nucleotide Polymorphisms Methods and Protocols. Methods in Molecular Biology 2003;212:1-14.

Lall, S., Gr¨un, D., Krek, A., Chen K, Wang YL, Dewey CN, et al. A Genome-Wide Map of Conserved MicroRNA Targets in C. elegans. Current Biology 2006;16(5):460–471.

Lenz G. Métodos Imunológicos [acesso em 4 nov 2016]. Disponível em: http://www.ufrgs.br/biofisica/Bio10003/MIMUNO.pdf.

Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell. 2003 Dec 26;115(7):787-98.

Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005;120: 15–20.

Li JH1, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2014 Jan;42(Database issue):D92-7.

Li L, Xu J, Yang D, Tan X, Wang H. Computational approaches for microRNA studies: a review. Mamm Genome. 2010 Feb;21(1-2):1-12.

Li L, Gao Q, Mao X, Cao Y. New support vector machine-based method for microRNA target prediction. Genet Mol Res. 2014 jun 9;13(2):4165-76.

Liang L, Morar N, Dixon AL, Lathrop GM, Abecasis GR, Moffatt MF et al.. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid celllines. Genome Res. 2013 Apr;23(4):716-26.

Liang L, Zhang Q, Luo LL, Yue J, Zhao YL, Han M et al. Polymorphisms in the prostaglandin receptor EP2 gene confers susceptibility to tuberculosis. Infect Genet Evol. 2016 oct 22;46:23-27.

Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, et al.. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012 Jan;40(Database issue):D857-61.

Liu, C., Zhang, F., Li, T. & et al. MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. BMC Genomics 2012;13(1):661.

Liu H, Yue D, Chen Y, Gao SJ, Huang Y. Improving performance of mammalian microRNA target prediction. BMC Bioinformatics. 2010 Sep 22;11:476.

Mariaselvam CM, Tamouza R, Krishnamoorthy R, Charron D, Misra DP, Jain VK, Negi VS. Association of NKG2D gene variants with susceptibility and severity of Rheumatoid Arthritis. Clin Exp Immunol. 2016 oct 26.

Markham NR, Zuker M. DINAMelt Web Server for Nucleic Acid Melting Prediction. Nucleic Acids Res 2005;33(Web Server issue):W577-81.

McDowall MD, Scott MS, Barton GJ. PIPs: human protein-protein interaction prediction database. Nucleic Acids Res. 2009 Jan;37(Database issue):D651-6.

Meister G. Argonaute proteins: funcional insights and emerging roles. Nat Rev Genet. 2013 Jul;14(7):447-59.

Menor M, Ching T, Zhu X, Garmire D,Garmire LX. mirMark: a site-level and UTR-level classifier for miRNA target prediction. Genome Biol. 2014;15(10):500.

Mez J, Chung J, Jun G, Kriegel J, Bourlas AP, Sherva R et al. Two novel loci, COBL and SLC10A2, for Alzheimer's disease in African Americans. Alzheimer's Dement. 2016 oct 19;1-11.

Min H, Yoon S. Got target?:computational methods for microRNA target prediction and their extension. ExpMol Med. 2010 Apr 30;42(4):233-44.

Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM et al. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell 2006;126(6):1203–1217.

Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, et al.. Human protein reference database--2006 update. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D411-4.

Mitra R, Bandyopadhyay S. MultiMiTar: A Novel Multi Objective Optimization based miRNA-Target Prediction Method. PLoS One. 2011;6(9):e24583.

Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLOS Medicine 2009 Jul 21;6(7):e1000097.

Moore MJ, Scheel TK, Luna JM, Park CY, Fak JJ, Nishiuchi E et al. miRNA-target chimeras reveal miRNA 3'end pairing as a major determinant of Argonaute target specificity. Nat Commun. 2015 nov 25;6:8864.

Moxon S, Moulton V, Kim JT. A scoring matrix approach to detecting miRNA target sites. Algorithms Mol Biol. 2008 Mar 31;3:3.

Myers AJ1, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, et al.. A survey of genetic human cortical gene expression. Nat Genet. 2007 Dec;39(12):1494-9.

Nicoloso MS, Sun H, Spizzo R, Kim H, Wickramasinghe P, Shimizu M, et al. Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. Cancer Res. 2010 Apr 1;70(7):2789-98.

Oğul H1, Umu SU, Tuncel YY, Akkaya MS. A probabilistic approach to microRNA-target binding. Biochem Biophys Res Commun. 2011 Sep 16;413(1):111-5.

Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein–protein interactions. J Med Genet. 2006 Aug;43(8):691-8.

Oulas A, Karathanasis N, Louloupi A, Iliopoulos I, Kalantidis K, Poirazi P. A new microRNA target prediction tool identifies a novel interaction of a putative miRNA with CCND2. RNA Biol. 2012 Sep;9(9):1196-207.

Panoutsopoulou K, Tachmazidou I, Zeggini E. In search of low-frequency and rare variants affecting complex traits. Hum Mol Genet. 2013 Oct 15;22(R1):R16-21.

Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG. The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Res 2009;37(Database issue):D155-8.

Papatsenko D. ClusterDraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors. Bioinformatics. 2007 Apr 15;23(8):1032-4.

Paraskevopoulou MD1,2, Vlachos IS1,2, Hatzigeorgiou AG1,2. DIANA-TarBase and DIANA Suite Tools: Studying Experimentally Supported microRNA Targets. Curr Protoc Bioinformatics. 2016 Sep 7;55:12

Park K, Kim KB. miRTarHunter: A Prediction System for Identifying Human microRNA Target Sites. Mol Cells. 2013 Mar;35(3):195-201.

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, et al.. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 2003 Oct;13(10):2363-71.

Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, et al.. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D497-501.

Piovezani AR. SIMTar: Uma ferramenta para predição de SNPs interferindo em sítios alvos de microRNAs[dissertação]. São Paulo: Universidade de São Paulo, Programa de Interunidades em Bioinformática; 2013.

Radfar H, Wong W, Morris Q. BayMiR: inferring evidence for endogenous miRNA-induced gene repression from mRNA expression profiles. BMC Genomics. 2013 Aug 30;14:592.

Ragan C, Cloonan N, Grimmond SM, Zuker M, Ragan MA. Transcriptome-Wide Prediction of miRNA Targets in Human and Mouse Using FASTH. PLoS One. 2009 May 29;4(5):e5745.

Rajewsky, N. and Socci, N.D.. Computational identification of microRNA targets. Dev. Biol. 2004;267:529-35.

Reczko M, Maragkakis M, Alexiou P, Papadopoulos GL, Hatzigeorgiou AG. Accurate microRNA Target Prediction Using Detailed Binding Site Accessibility and Machine Learning on Proteomics Data. Front Genet. 2012 jan 18;2:103.

Rehmsmeier, M., Steffen, P., Hochsmann, Giegerich R. Fast and effective prediction of microRNA/target duplexes. RNA 2004;10(10):1507–1517.

Ritchie W, Rasko JE.Refining microRNA target predictions: sorting from the wheat to the chaff. BiochemBiophys Res Commun. 2014 Feb 10.

Rusinov V, Baev V, Minkov IN, Tabler M. MicroInspector: a web tool for detection of miRNa binding sites in an RNA sequence. Nucleic Acids Res. 2005 jul 1;33(Web Server issue):W696-700.

Saetrom O, Snove O Jr, Saetrom P. Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. RNA. 2005 Jul;11(7):995-1003.

Saetrom, P., Heale, B. S. E., Snove, O. & et al. Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Research 2007;35(7):2333–2342.

Saito T, Saetrom P. A two-step site and mRNA-level model for predicting microRNA targets. BMC Bioinformatics. 2010 Dec 31;11:612.

Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, et al.. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008 May 6;6(5):e107.

Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. Nature 2008;455(7209):58-63.

Sethupathy P1, Corda B, Hatzigeorgiou AG. TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA. 2006 Feb;12(2):192-7.

Sethupathy, P., Megraw, M., Hatzigeorgiou, A.G. A guide through present computational approaches for the identification of mammalian microRNA targets. Nat Methods 2006;3:881–886.

Shahi P, Loukianiouk S, Bohne-Lang A, Kenzelmann M, Küffer S, Maertens S et al.. Argonaute--a database for gene regulation by mammalian microRNAs. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D115-8.

Shin C, Nam JW, Farh KK, Chiang HR, Shkumatava A, Bartel DP. Expanding the microRNA targeting code: functional sites with centered pairing. Mol Cell. 2010 jun 25;38(6):789-802.

Shoemaker BA1, Zhang D, Thangudu RR, Tyagi M, Fong JH, Marchler-Bauer A, et al.. Inferred Biomolecular Interaction Server--a web server to analyze and predict protein interacting partners and binding sites. Nucleic Acids Res. 2010 Jan;38(Database issue):D518-24.

Shoemaker BA1, Zhang D, Tyagi M, Thangudu RR, Fong JH, Marchler-Bauer A, et al.. IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. Nucleic Acids Res. 2012 Jan;40(Database issue):D834-40.

Smalheiser NR, Torvik VI. A population-based statistical approach identifies parameters characteristic of human microRNA-mRNA interactions. BMC Bioinformatics. 2004 Sep 28;5:139.

Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981, 147:195-197.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, et al.. The Human Gene Mutation Database: 2008 update. Genome Med. 2009 Jan 22;1(1):13.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al.. Population genomics of human gene expression. Nat Genet. 2007 Oct;39(10):1217-24.

Sturm M, Hackenberg M, Langenberger D, Frishman D. TargetSpy: a supervised machine learning approach for microRNA target prediction. BMC Bioinformatics. 2010 May 28;11:292.

Thadani R, Tammi MT. MicroTar: predicting microRNA targets from RNA duplexes. BMC Bioiformatics. 2006 dec 18;7(5):S20.

Thomas M, Lieberman J, Lal A.Desperately seeking microRNA targets.Nat Struct Mol Biol. 2010 Oct;17(10):1169-74.

Thomas LF, Saito T, Sætrom P. Inferring causative variants in microRNA target sites. Nucleic Acids Res. 2011 Sep 1;39(16):e109.

Toscano-Garibay, J. D. & Aquino-Jarquin, G. Regulation exerted by miRNAs in the promoter and UTR sequences: MDR1/P-gp expression as a particular case. DNA Cell Biology 2012;31:1358–64.

van Dongen S, Abreu-Goodger C, Enright AJ. Detecting microRNA binding and siRNA off-target effects from expression data. Nat Methods. 2008 dec;5(12):1023-5.

Vergoulis T1, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M et al.. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Res. 2012 Jan;40(Database issue):D222-9.

Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I et al.. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Res. 2015 Jan;43(Database issue):D153-9.

Wang P, Ning S, Wang Q, Li R, Ye J, Zhao Z, Li Y, Huang T, Li X. mirTarPri: Improved Prioritization of MicroRNA Targets through Incorporation of Functional Genomics Data. PLoS One 2013;8(1):e53685.

Wang T, Xiao G, Chu Y, Zhang MQ, Corey DR, Xie Y. Design and bioinformatics analysis of genome-wide CLIP experiments. Nucleic Acids Res. 2015 Jun 23;43(11):5263-74.

Webb GI: MultiBoosting: A Technique for Combining Boosting and Wagging. Machine Learning 2000;40(2):159-196.

Westra HJ1, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al.. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013 Oct;45(10):1238-43.

Witkos TM, Koscianska E, Krzyzosiak WJ. Practical aspects of microRNA target prediction. Curr Mol Med. 2011 Mar;11(2):93-109.

Wu L, Belasco JG. Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. Mol Cell.2008 Jan 18;29(1):1-7.

Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete Suboptimal Folding of RNA and the Stability ofSecondary Structures. Biopolymers. 1999 Feb;49(2):145-65.

Xia K, Shabalin AA, Huang S, Madar V, Zhou YH, Wang W, et al.. seeQTL: a searchable database for human eQTLs. Bioinformatics. 2012 Feb 1;28(3):451-2.

Xiao, F., Zuo, Z., Cai, G. & et al. miRecords: an integrated resource for microRNA-target inter - actions. Nucleic acids research 2009;37:D105–D110.

Xu W, San Lucas A, Wang Z, Liu Y. Identifying microRNA targets in different gene regions. BMC Bioinformatics. 2014;15(7):S4.

Yang HC, Lin CW, Chen CW, Chen JJ. Applying genome-wide gene-based expression quantitative trait locus mapping to study population ancestry and pharmacogenetics. BMC Genomics. 2014 Apr 29;15:319.

Yang JH, Li JH, Shao P, Zhou H, Chen YQ, Qu LH. starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic Acids Res. 2011 Jan;39(Database issue):D202-9.

Yang Y, Wang YP, Li KB. MiRTif: a support vector machine-based microRNA target interaction filter. BMC Bioinformatics. 2008 Dec 12;9 Suppl 12:S4.

Yang YC, Di C, Hu B, Zhou M, Liu Y, Song N et al.. CLIPdb: a CLIP-seq database for protein-RNA interactions. BMC Genomics. 2015 Feb 5;16:51.

Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTeraction database. FEBS Lett. 2002 Feb 20;513(1):135-40.

Zhang B, Wang Q, Pan X. MicroRNAs and their regulatory roles in animals and plants. J Cell Physiol. 2007 feb;210(2):279-89.

Ziebarth, J. D., Bhattacharya, A., Chen, A. & et al. PolymiRTS Database 2.0: linking polymorphisms in microRNA target sites with human diseases and complex traits. Nucleic Acids Research 2012;40(D1), D216–D221.

Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Estatist Soc B. 2005;67(Part 2):301-20.

Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. Nucleic Acids Res. 1981 Jan 10;9(1):133-48.

Ørom UA, Nielsen FC, Lund AH. MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. Mol Cell. 2008 May 23;30(4):460-71.

APÊNDICE A - PROTOCOLO DA REVISÃO SISTEMÁTICA

“Análise Comparativa de Ferramentas Computacionais de Predição de Alvos de microRNAs ”

OBJETIVO:

Analisar e comparar as ferramentas existentes e disponiveis para a predição de alvos de microRNAs.

QUESTÕES DE PESQUISA:

Quais são os algoritmos utilizados em predição de alvos de microRNAs?

Quais ferramentas estão disponíveis para teste?

SELEÇÃO DE FONTES:

As fontes deverão estar disponíveis via web, preferencialmente em bases de dados científicas da área. Poderão ser selecionados também trabalhos disponíveis em outros meios, desde que atendam aos requisitos da Revisão Sistemática.

PALAVRAS-CHAVES:

(microRNA*[Title] OR miRNA*[Title]) AND target*[Title] AND (prediction[Title/Abstract] OR identification[Title/Abstract] OR detecting[Title/Abstract] OR detection[Title/Abstract]) AND (tool[Title/Abstract] OR approach[Title/Abstract] OR method[Title/Abstract] OR algorithm[Title/Abstract] OR program[Title/Abstract])

LISTAGEM DE FONTES:

· *PubMed (http://www.ncbi.nlm.nih.gov/pubmed)*

· *CAPES - Banco de Teses (http://www.capes.gov.br/serviços/banco-de-teses*

TIPO DOS ARTIGOS:

Serão considerados dissertações de mestrado, teses de doutorado e artigos que abordem o tema predição de alvos de microRNAs.

IDIOMA(S) DOS ARTIGOS:

Inglês e português.

CRITÉRIOS DE INCLUSÃO E EXCLUSÃO DOS TRABALHOS:

Critérios de inclusão:

(a) Serão incluídos trabalhos publicados e disponíveis integralmente em bases de dados científicas ou em versões impressas.

(b) Serão incluídos os trabalhos que proponham uma ferramenta computacional de predição de alvos de microRNAs em animais.

Critérios de exclusão:

(a) Serão excluídos trabalhos que tratem de ferramentas já analisadas.

(b) *Serão excluídos trabalhos que não disponibilizem a ferramenta para download ou não disponilizem a versão* web *da ferramenta.*

(c) Serão excluídos trabalhos que não sejam disponibilizados na íntegra.

(d) Serão excluídos trabalhos escritos em outro idioma que não em português ou inglês.

(e) Serão excluídos trabalhos referentes a alvos de microRNAs em plantas.

CRITÉRIOS DE QUALIDADE DOS ESTUDOS PRIMÁRIOS:

O trabalho deverá ter sido publicado em periódico ou anais de eventos com revisão por pares quando se referir a artigos ou aprovado por banca examinadora quando se referir a trabalhos de conclusão de curso, mestrado ou doutorado.

AVALIAÇÃO DA QUALIDADE DOS ESTUDOS PRIMÁRIOS:

Serão considerados os artigos que atinjam os dois critérios de inclusão e nenhum dos critérios de exclusão.

PROCESSO DE SELEÇÃO DOS ESTUDOS PRIMÁRIOS:

A strings de busca definida na seção "palavras-chave", e possíveis variações, serão submetidas às máquinas de busca. A busca, no entanto, será restrita aos campos Título e Resumo (Title/Abstract). Após remoção de redundâncias (mesmo artigo retornado por diferentes buscas, a leitura do resumo e aplicação dos critérios de inclusão e exclusão, o trabalho será selecionado se confirmada a sua relevância pelo principal revisor (aluno). Se houver dúvida da relevância a orientadora será consultada.

Se artigos de revisão forem retornados na busca e tragam uma lista de ferramentas computacionais de predição de miRNAs, as listas serão avaliadas no sentido de verificar se alguma ferramenta não foi incluída na revisão, e neste caso ela será incluída.

ESTRATÉGIA DE EXTRAÇÃO DE INFORMAÇÃO:

Após definidos os trabalhos definitivamente incluídos, este serão lidos na íntegra, para que possa ser retirado o método que cada ferramenta utiliza para predizer os alvos do microRNA.O revisor fará um resumo de cada um deles, destacando os métodos utilizados para a predição e parâmetros considerados, quando for o caso.

Serão preenchidos "formulários de extração de dados" para cada texto considerado válido para a revisão. Além das informações básicas (dados bibliográficos, data de publicação, abstract, entre outros), esses formulários deverão conter a síntese do trabalho, destacando não só os métodos e as técnicas utilizadas, mas também os resultados obtidos, redigida pelo pesquisador que conduzirá a revisão e reflexões pessoais do mesmo a respeito do conteúdo e das conclusões do estudo.

SUMARIZAÇÃO DOS RESULTADOS:

Após a leitura e o resumo dos trabalhos selecionados, será elaborado um relatório técnico com uma análise quantitativa dos trabalhos. Também será elaborada uma análise qualitativa a fim de definir as vantagens e desvantagens de cada método. Para auxiliar na análise qualitativa será elaborado um checklist com itens importantes a serem observados em cada método apresentado.

APÊNDICE B - SNPS COM INTERFERÊNCIA VALIDADA EM SÍTIOS DE MIRNAS

B.1- Protocolo da revisão

“SNPs que interferem em sítios alvos de microRNAs”

OBJETIVO:

Obter SNPs que interferem em sítios alvos de microRNAs.

QUESTÕES DE PESQUISA:

Quais são os SNPs que interferem em sítios alvos de microRNAs?

SELEÇÃO DE FONTES:

As fontes deverão estar disponíveis via web, preferencialmente em bases de dados científicas da área. Poderão ser selecionados também trabalhos disponíveis em outros meios, desde que atendam aos requisitos da Revisão Sistemática.

PALAVRAS-CHAVES:

(miRNA* [Title/Abstract] OR microRNA* [Title/Abstract] OR miR [Title/Abstract]) AND (target [Title/Abstract] OR "binding site" [Title/Abstract]) AND (SNP [Title] OR variation[Title] OR variant*[Title] OR allele [Title] OR polymorphism*)

LISTAGEM DE FONTES:

- ***PubMed (http://www.ncbi.nlm.nih.gov/pubmed)***
- ***CAPES - Banco de Teses (http://www.capes.gov.br/serviços/banco-de-teses***

TIPO DOS ARTIGOS:

Serão considerados dissertações de mestrado, teses de doutorado e artigos que apresentem SNPs que interferem em sítios alvos de microRNAs.

IDIOMA(S) DOS ARTIGOS:

Inglês e português.

CRITÉRIOS DE INCLUSÃO E EXCLUSÃO DOS TRABALHOS:

Critérios de inclusão:

(a) Serão incluídos trabalhos apresentando SNPs que interferem em sítios alvos de microRNAs em humanos.

(b) Serão incluídos os trabalhos que apresentem interações alvo-miRNA validadas relacionadas a SNPs.

Critérios de exclusão:

(a) Serão excluídos trabalhos que apresentem interações alvo-miRNA preditas.

(b) Serão excluídos trabalhos que apresentem SNPs em sítios alvos de miRNA mas não interferem na ligação.

(c) Serão excluídos trabalhos escritos em outro idioma que não em português ou inglês.

(d) Serão excluídos trabalhos referentes a interações alvo-microRNA em plantas e animais que não seja o homem.

CRITÉRIOS DE QUALIDADE DOS ESTUDOS PRIMÁRIOS:

O trabalho deverá ter sido publicado em periódico ou anais de eventos com revisão por pares quando se referir a artigos ou aprovado por banca examinadora quando se referir a trabalhos de conclusão de curso, mestrado ou doutorado.

AVALIAÇÃO DA QUALIDADE DOS ESTUDOS PRIMÁRIOS:

Serão considerados os artigos que atinjam os dois critérios de inclusão e nenhum dos critérios de exclusão.

PROCESSO DE SELEÇÃO DOS ESTUDOS PRIMÁRIOS:

A strings de busca definida na seção "palavras-chave", e possíveis variações, serão submetidas às máquinas de busca. A busca, no entanto, será restrita aos campos Título (Title). Após remoção de redundâncias (mesmo artigo retornado por diferentes buscas, a leitura do resumo e aplicação dos critérios de inclusão e exclusão, o trabalho será selecionado se confirmada a sua relevância pelo principal revisor (aluno). Se houver dúvida da relevância a orientadora será consultada.

Se artigos de revisão forem retornados na busca e tragam uma lista de SNPs que interferem em sítios alvos de miRNAs, as listas serão avaliadas no sentido de verificar se algum SNP não foi incluído na revisão, e neste caso ele será incluído.

ESTRATÉGIA DE EXTRAÇÃO DE INFORMAÇÃO:

Os títulos e os resumos dos artigos serão lidos em busca das seguintes informações: nome do SNP, nome do gene e nome do microRNA. Se não houver informação suficiente no resumo, o artigo completo será consultado.

SUMARIZAÇÃO DOS RESULTADOS:

Após a leitura do título e do resumo dos trabalhos selecionados, será elaborada uma tabela com o nome do SNP, nome do gene e nome do microRNA.

B.2 - Lista de SNPs

rs15869
rs8946
rs1048638
rs3218073
rs1054564
rs5889

rs78790512
rs17228616
rs56292801
rs818708
rs1071738
rs1051312
rs4909237
rs1042752
rs4225
rs5225
rs8113500
rs12532
rs1131445
rs2266788
rs13385
rs11169571
rs4819388
rs78544189
rs7431
rs187729
rs465646
rs1054204
rs17813964
rs72558377
rs61764370
rs1049253
rs5186
rs7747909
rs115785973
rs6573
rs241456
rs2297136
rs5534
rs7201
rs35717904
rs12904
rs6631
rs1425486
rs4647940
rs11064
rs28382751
rs16917496
rs7213430
rs2929970
rs13306046
rs13347
rs41391245
rs978906
rs759330
rs10485058
rs183204610

rs4151672
rs417309
rs3203358
rs1057147
rs2240688
rs1573613
rs6265
rs662702
rs334348
rs7079
rs10773771
rs3742106
rs1062044
rs7963551
rs4143815
rs1050286
rs6435156
rs11174811
rs10065172
rs1042157
rs56288038
rs3803012
rs56109847
rs1054191
rs461155
rs11655237
rs4585
rs1434536
rs1057035
rs8126
rs11466537
rs9291296
rs4702
rs1051296
rs10759
rs1062980
rs10749571
rs1056629
rs1805672
rs1057233
rs12373
rs2790
rs2057482
rs896
rs3742943
rs2147578
rs4790522
rs9341070
rs10234329
rs184456571
rs3732360

rs4790521
rs3134615
rs3814058
rs3217992
rs2229295
rs12915554
rs1054190
rs2168518
rs11895168
rs5848
rs3660
rs1045411
rs9299
rs712
rs4742098
rs1836724
rs1437134
rs3115758
rs7646
rs1063320
rs9909
rs12976445
rs2229591
rs28674628
rs1046322
rs10204525
rs137853044
rs78378222
rs739837
rs113810300
rs12190287
rs111638916
rs12720208
rs2738464
rs7194256
rs2279398
rs531564
rs1049337
rs1048201
rs3733336
rs115214213
rs111681798
rs9457
rs1050347
rs868
rs884225
rs799917
rs141178472
rs7143400
rs1057317
rs2239680

rs6774494
rs1050283
rs1060120
rs2278414
rs6976789
rs9479
rs11473
rs231253
rs8259
rs1799782
rs35592567
rs17592236
rs8506
rs3208684
rs17737058
rs4846049
rs1042538
rs2735383
rs201253747
rs1044129
rs550067317
rs114673809
rs16958754
rs2677743
rs111904020
rs10719
rs546782
rs1063192
rs1063611
rs6489956
rs108621
rs78195212
rs7930
rs1056628
rs3088440
rs3739497
rs4245739
rs11574744
rs1045385
rs3735590
rs17143818
rs1599795
rs1049216
rs2229901
rs12731181
rs535860
rs3746544
rs696
rs3811463