UNIVERSIDADE FEDERAL DE SÃO CARLOS

PRÓ-REITORIA DE PESQUISA

COORDENADORIA DE INICIAÇÃO CIENTÍFICA E TECNOLÓGICA

Titulo do Projeto de Pesquisa:

Predição de Alvos de MicroRNAs Utilizando Aprendizado Multirrótulo e Regras de Associação

2019

Resumo

MicroRNAs (miRNAs) constituem uma família de RNAs de 21-25 nucleotídeos que regulam negativamente a expressão gênica no nível pós-transcricional de alguns RNAs mensageiros (mRNAs) codificantes. Essa regulação ocorre por meio da interação do miRNA com características específicas dos mRNAs transcritos. Todavia, a interação RNA-miRNA apresenta um grande número de características e complexidade, dificultando sua predição. O fato de um miRNA possuir múltiplos alvos simultaneamente dificulta a aplicação de métodos de classificação convencionais em Aprendizado de Máquina. Assim, o objetivo deste projeto é utilizar Aprendizado Multirrótulo, possibilitando a predição de múltiplos rótulos (alvos) de uma única amostra (miRNAs). Devido ao fato de haver muito poucos exemplos positivos para cada rótulo (baixa cardinalidade de rótulos), a abordagem de Regras de Associação será utilizada como uma etapa complementar de classificação, para auxiliar predições de rótulos menos frequentes no conjunto de dados. Por fim, este projeto de Iniciação Científica (IC) tem como objetivo o desenvolvimento de uma ferramenta capaz de classificar se um mRNA é alvo de interação de um miRNA ou não.

1 Introdução e Justificativa

MicroRNAs (miRNAs) são uma família de RNAs de 21-25 nucleotídeos que regulam negativamente expressões gênicas de RNAs mensageiros (mRNAs) no nível pós-transcricional (Bartel, 2009). Nos mamíferos, os miRNAs são responsáveis por controlar a atividade de aproximadamente 50% dos genes codificantes de proteínas (He & Hannon, 2004). Estudos funcionais indicam que os miRNA participam na regulação de quase todos os processos celulares investigados até agora, e mudanças em sua expressão estão associadas a muitas patologias humanas (Krol, Loedige, & Filipowicz, 2010).

Um alvo de miRNA é definido como um mRNA que sofre regulação gênica por tal miRNA. Também se tornou evidente que muitos, se não a maioria dos transcritos de codificação de proteínas, são alvos da regulação de miRNAs e que os miRNAs podem, em alguns casos, regular um grande número de mRNAs alvo e, reciprocamente, que muitos mRNAs contêm sítios alvo para muitos miRNAs (Morris & Mattick, 2014).

Aprendizado de Máquina é definido como a área que realiza estudos de algoritmos de computador que melhoram automaticamente através da experiência (Mitchell et al., 1997). Tais algoritmos aprendem com um conjunto de dados de entrada e conseguem realizar a classificação de novos dados em diferentes classes.

O experimento de teste de interação entre mRNA-miRNA pode ser feito manualmente em laboratórios, porém tais técnicas levam um tempo excessivo diante o número de RNAs e miRNAs existentes. Considerando ainda que um único miRNA pode possuir múltiplos alvos, o esforço manual para a validação experimental destas interações aumenta. Para isso, é vantajoso o uso de Aprendizado Multirrótulo para a classificação de múltiplas interações entre os miRNAs e todos os candidatos à alvos. O fato de existirem muitos candidatos a alvos aumenta a dimensionalidade do problema multirrótulo (um miRNA é possivelmente associado a centenas ou milhares de alvos). Isso faz com que existam muitos rótulos com poucos exemplos positivos. Para amenizar esse problema, serão utilizadas Regras de Associação, com o intuito de remover miRNAs que contenham poucos alvos validados experimentalmente do problema multirrótulo original, melhorando sua confiabilidade, sem perder o resultado da predição. Assim, classificadores multirrótulo serão utilizados no problema modificado, e posteriormente a classificação de alvos com poucos exemplos positivos será feita com o apoio de regras de associação previamente criadas no problema completo original.

O objetivo deste projeto de pesquisa é desenvolver uma ferramenta capaz de realizar a predição da interação entre vários miRNAs e candidatos à alvos.

2 Síntese da Bibliografia Fundamental

2.1 Classificação de Dados

A Classificação de Dados é a tarefa de atribuir categorias pré-definidas a objetos (Tan, Steinbach, & Kumar, 2005). Dadas certas características de um objeto, é possível associá-las à um rótulo. Algoritmos de Aprendizado podem aprender estes padrões e categorizar novos objetos automaticamente. Os métodos mais comumente usados para a classificação de dados são árvores de decisão, metodos baseados em regras, métodos probabilísticos e redes neurais (Aggarwal, 2014).

Uma Árvore de Decisão é definida como um procedimento de classificação que recursiva-

mente particiona um conjunto de dados em subconjuntos menores. Cada nó interno da árvore de decisão possui testes que são aplicados para recursivamente dividir os dados em sucessivos grupos menores (Friedl & Brodley, 1997). Os rótulos em cada nó folha referem-se ao rótulo de classe atribuído a cada objeto com base num conjunto de testes definidos em cada ramo (ou nó) na árvore.

Tendo como inspiração o comportamento dos neurônios biológicos e as suas funções exercidas em conjunto, as Redes Neurais Artificiais são sistemas paralelamente distribuídos constituídos de unidades de processamento (neurônios artificiais) massivamente interconectados, cuja tarefa é processar informações via cálculos de determinadas funções matemáticas (Costa, de Pádua Braga, & de Menezes, 2007).

Random Forest é um conjunto de árvores de decisão onde cada árvore é construída utilizando conjuntos de amostras distintas. O processo de classificação se dá quando cada árvore do conjunto faz sua predição individualmente, e a classe mais recorrente entre as predições se torna a saída da *Random Forest*.

2.2 Aprendizado Multirrótulo

Na classificação multirrótulo, os exemplos estão associados a um conjunto de rótulos Y contido no conjunto L de todos os possíveis rótulos (Tsoumakas & Katakis, 2007). Um exemplo de problema multirrótulo é a categorização de músicas em gêneros musicais. Uma música pode tanto ser categorizada com *Rock*, como *Blues*. Uma outra pode ser categorizada como clássica e trilha sonora.

Amostra	$Atributo_i, ..., Atributo_{\|X\|}$	**Rock**	**Blues**	**Clássica**	**Trilha Sonora**
A	...	0	0	0	1
B	...	1	0	0	1
C	...	1	1	0	0
D	...	0	0	1	1

Tabela 1: Exemplo do espaço de rótulos em problemas multirrótulo

Note que a Tabela 1 faz a representação do problema multirrótulos citado. As amostras são classificadas em diferentes rótulos não relacionados, podendo também serem classificados positivamente (1) em mais de um rótulo.

Uma maneira comum de se representar as predicões retornadas por um classificador multirrótulo é por meio de um vetor binário, sendo o tamanho do vetor dado pelo número de rótulos presentes no problema. Cada posição do vetor é associada a uma classe do problema, e possui o valor numérico 0/1, representando se o exemplo é rotulado (1) ou não é rotulado (0) pela respectiva classe.

Dado um problema multirrótulo, a *Cardinalidade* do conjunto de dados é definida como a média do número de rótulos das amostras no conjunto de dados. A *Densidade* do conjunto de dados é definida como a média do número de rótulos das amostras no conjunto de dados dividido pelo número de classes (Tsoumakas, Katakis, & Vlahavas, 2009). As Equações 1 e 2 representam a fórmula da Cardinalidade e Densidade respectivamente, sendo D o conjunto de dados analisado, $|D|$ o número de amostras no conjunto de dados, Y_i os verdadeiros rótulos da amostra i e $|L|$ o número de rótulos.

$$Card(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} Y_i \tag{1}$$

$$Dens(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{Y_i}{|L|} \tag{2}$$

Diferentes abordagens foram propostas em (Tsoumakas & Katakis, 2007) para lidar com problemas multirrótulo. Tais abordagens foram organizadas em dois grupos distintos: *Transformação de problema* e *Adaptação de algoritmo*.

2.2.1 Transformação de Problema

Existem algumas estratégias para a transformação de um problema multirrótulo em um ou mais problemas de classificação convencionais. A estratégia *Binary Relevance (BR)* transforma o problema multirrótulo em vários problemas de rótulo único, treinando vários modelos para a classificação de cada uma das classes exclusivamente. Seguindo o exemplo da classificação de gêneros musicais, proposto anteriormente, temos que a transformação *BR* nos daria 4 diferentes problemas único rótulo. As tabelas 2, 3, 4 e 5 representam os 4 problemas de único rótulo gerados pela transformação *BR*.

Amostra	$Atributo_i, ..., Atributo_{\|X\|}$	**Rock**
A	...	0
B	...	1
C	...	1
D	...	0

Tabela 2: Exemplo da abordagem BR para o rótulo *Rock*

Amostra	$Atributo_i, ..., Atributo_{\|X\|}$	**Blues**
A	...	0
B	...	0
C	...	1
D	...	0

Tabela 3: Exemplo da abordagem BR para o rótulo *Blues*

Amostra	$Atributo_i, ..., Atributo_{\|X\|}$	**Clássica**
A	...	0
B	...	0
C	...	0
D	...	1

Tabela 4: Exemplo da abordagem BR para o rótulo *Clássica*

Amostra	$Atributo_i, ..., Atributo_{\|X\|}$	**Trilha Sonora**
A	...	1
B	...	1
C	...	0
D	...	1

Tabela 5: Exemplo da abordagem BR para o rótulo *Trilha Sonora*

A estratégia *Label Powerset (LP)* encontra todas as possíveis combinações de rótulos das amostras do conjunto de dados e os transforma em uma nova classe, criando assim um problema multi-classe. A Tabela 6 representa a transformação LP associando os conjuntos de rótulos de cada exemplo à uma classe utilizando o conjunto de dados dos gêneros musicais.

Amostra	$Atributo_i, ..., Atributo_{\|X\|}$	**Trilha Sonora**	**Trilha Sonora & Rock**	**Rock & Blues**	**Clássica & Trilha Sonora**
A	...	1	0	0	0
B	...	0	1	0	1
C	...	0	0	1	0
D	...	0	0	0	1

Tabela 6: Demonstração das classes obtidas após a transformação LP

A estratégia *Classifier Chain* adota a criação de $|L|$ classificadores único rótulo. Sendo C o conjunto de classificadores, temos que o classificador c_i é associado à predição do rótulo l_i. O espaço de atributos de cada classificador em C é extendido pelos valores dos rótulos dos classificadores passados. O processo de classificação começa em c_1 e se propaga pelo conjunto C. c_1 determina a $Pr(l_1|x)$, c_2 determina a $Pr(l_2|x, l_1)$ e c_i determina a $Pr(l_i|x, l_1, l_2, ..., l_{i-1})$, sendo x o espaço de atributos de uma dada amostra (Read, Pfahringer, Holmes, & Frank, 2009). Esta estratégia mantém a correlação entre os rótulos.

As tabelas 7, 8, 9 e 10 representam a disposição dos dados para C_i no problema da classificação dos gêneros musicais.

Amostra	$Atributo_i, ..., Atributo_{\|X\|}$	**Rock**
A	...	1
B	...	0
C	...	0
D	...	0

Tabela 7: Conjunto de dados analisado pelo C_1

Amostra	$Atributo_i, ..., Atributo_{\|X\|}$	Rock	**Blues**
A	...	1	0
B	...	0	0
C	...	0	1
D	...	0	0

Tabela 8: Conjunto de dados analisado pelo C_2

Amostra	$Atributo_i, ..., Atributo_{\|X\|}$	Rock	Blues	**Clássica**
A	...	1	0	0
B	...	0	0	0
C	...	0	1	0
D	...	0	0	1

Tabela 9: Conjunto de dados analisado pelo C_3

Amostra	$Atributo_i, ..., Atributo_{\|X\|}$	Rock	Blues	Clássica	**Trilha Sonora**
A	...	0	0	0	1
B	...	1	0	0	1
C	...	1	1	0	0
D	...	0	0	1	1

Tabela 10: Conjunto de dados analisado pelo C_4

2.2.2 Adaptação de Algoritmo

Uma outra abordagem para lidar com problemas multirrótulo é a Adaptação de Algoritmos de aprendizado que, primeiramente tratavam problemas único rótulo, para os casos multirrótulo.

Em (Clare & King, 2001), o algoritmo da árvore C4.5 foi modificado para aceitar múltiplos rótulos nas folhas da árvore, assim como uma modificado função de entropia.

Em (Schapire & Singer, 2000) é proposto duas versões do algoritmo AdaBoost: AdaBoost.MR e AdaBoost.MH, ambas multirrótulo. A organização do AdaBoost.MH faz com que ele encontre uma minimização da *Hamming Loss*, enquanto o AdaBoost.MR busca criar hipóteses sobre os atributos que ficariam no topo do ranking (Tsoumakas et al., 2009).

O algoritmo *Back-Propagation* para o treino de Redes Neurais foi modificado para suportar dados multirrótulo em (Zhang & Zhou, 2006).

2.3 Mineração de Regras de Associação

Regras de associação é um método do Aprendizado de Máquina para a descoberta de relações entre variáveis de um conjunto de dados.

Por exemplo, uma empresa de marketing pode querer perguntar: “Qual é a porcentagem de pessoas que compraram X também comprou Y?". Outra pergunta poderia ser: “Quais são os dois itens mais populares entre as pessoas de idade entre 18 e 25?” (Agrawal, Srikant, et al., 1994).

Para isso, chamamos todo conjunto de elementos que buscamos uma relação de *itemset*. Durante o processo de mineração, todas as possíveis permutações de itens são usadas para constituir itemsets. Em muitos casos, uma grande maioria destes itemsets não são úteis devido à correlação entre seus itens ser rara. É possível medir a relação e raridade de um itemset usando as medidas de *Suporte* e *Confiança*.

Suporte de um itemset I é definido como o número de vezes que o itemset aparece nas amostras x_i do conjunto de dados D.

$$Suporte(I, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} f(I, x_i) \tag{3}$$

$$f(I, x) = \begin{cases} 1, & x \bigcap I = I \\ 0, & cc. \end{cases} \tag{4}$$

Um limiar de aceitação pode ser estabelecido, limitando a geração de itemsets aos quais o suporte não o excede. A Figura 1 dá o exemplo da extração de itemsets com suporte maior ou igual à 0.5 de um conjunto de amostras.

A = {bermuda, calça, camiseta, sandália, tênis}

Suporte Mínimo = 0.5

Amostra	Itens Comprados
1	Calça, Camiseta, Tênis
2	Camiseta, Tênis
3	Bermuda, Tênis
4	Calça, Sandália

Itemset	Suporte
{Tênis}	0.75
{Calça}	0.5
{Camiseta}	0.5
{Camiseta, Tênis}	0.5

Figura 1: Exemplo de Regras de Associação. À esquerda, o conjunto ao qual as regras serão extraídas. À direita, o conjunto de itemsets minerados com Suporte >= 0.5

À partir de itemsets com mais do que um elemento, montamos as regras de associação. Uma regra de associação R é formada por uma inferência $X \implies y$, onde y é um elemento de um itemset I e X é o itemset I sem o elemento y.

Uma nova medida de análise é introduzida em Regras de Associação: a medida de *Confiança*. Esta medida define a probabilidade de y acontecer sendo que X aconteceu.

A Figura 2 representa as regras de associação geradas à partir do itemset do exemplo passado. Note

$$Confianca(R, D) = \frac{Suporte(I)}{Suporte(X)} \tag{5}$$

Neste projeto, a exploração das Regras de Associação se tornou necessária devido ao alto número de rótulos a serem preditos. Ao criar fortes associações entre rótulos, é possível excluir do conjunto de dados rótulos que possuem baixa cardinalidade. Esta exclusão é temporária e tem o objetivo de esconder rótulos que não garantiríam um aprendizado efetivo de um modelo classificador. Após a remoção destes rótulos, o modelo de predição multirrótulo fará a predição em rótulos que possuem alta cardinalidade, portanto, terão um grau de confiabilidade maior. Assim que o classificador fizer suas predições, as mesmas serão complementadas pelas Regras de Associação mineradas.

Regras de Associação	Suporte	Confiança
{Camiseta} -> {Tênis}	0.5	1.0
{Tênis} -> {Camiseta}	0.5	1.0

Figura 2: Construção de Regras de Associação à partir de itemsets frequentes

Um algoritmo famoso para minerar Regras de Associação é o algoritmo *Apriori* (Hornik, Grün, & Hahsler, 2005). Esse algoritmo será utilizado nesta pesquisa.

2.4 Trabalhos na literatura de predição de alvos de miRNAs

Um dos trabalhos mais conhecidos e recentes de predição de alvos de miRNAs utilizando métodos de Aprendizado de Máquina (AM) foi desenvolvido por (Agarwal, Bell, Nam, & Bartel, 2015). Nesse trabalho, a ferramenta TargetScan foi criada. Ela é usada para a predição de sítios de interação em alvos de miRNAs exclusivamente em mamíferos.

Outro trabalho mais antigo foi desenvolvido por (Sturm, Hackenberg, Langenberger, & Frishman, 2010), onde foi feita a predição de sítios em alvos de miRNAs sem a utilização de algum atributo relativo à pareamento perfeito da interação.

Em (Liu, Yue, Chen, Gao, & Huang, 2010) foi desenvolvida uma ferramenta para a predição de alvos de miRNAs utilizando *Support Vector Machines (SVMs)*, outro classificador muito utilizado em Aprendizado de Máquina.

A ideia de usar Regras de Associação para reduzir a dimensionalidade do espaço de rótulos foi introduzida em (Charte, Rivera, del Jesus, & Herrera, 2012). A abordagem usada foi justamente a proposta neste projeto: a mineração de Regras de Associação referentes ao espaço de rótulos de um conjunto de dados, a eliminação temporária dos rótulos de baixa cardinalidade, a predição feita por classificadores e a complementação da predição utilizando as associações das regras. Esta abordagem foi utilizada com conjuntos multirrótulo transformados usando BR, LP e também com algoritmos adaptados para classificação multirrótulo. Os resultados garantiram uma taxa de acerto maior para as predições complementadas por regras de associação.

3 Materiais e Métodos

Esta sessão apresenta como foi criado o conjunto de dados utilizado nos experimentos. São apresentadas também as medidas de avaliação utilizadas para a validação do modelo de classificação proposto, e as ferramentas para a construção do mesmo.

3.1 Conjuntos de dados

Para os experimentos, foi criado um Conjunto de dados de miRNAs com alvos em humanos. Esse conjunto de dados possui 11951 amostras e 2606 colunas. A primeira coluna corresponde ao identificador de cada amostra, seis correspondentes aos atributos, e 11944 aos rótulos (alvos).

Dados referentes a interações de miRNAs e alvos foram obtidos do *mirTarBase* (Griffiths-Jones, 2006), um banco de dados de interações de miRNAs e alvos validadas experimentalmente. Os dados referentes aos genes foram obtidos do banco de dados *NCBI* (Coordinators, 2016). Os dados coletados foram filtrados para seleção apenas da taxonomia *Homo Sapiens*. Foi desenvolvido um script de extração das atributos necessárias para o processo de AM. Esse estágio foi essencial para conhecer os dados que serão utilizados, e selecionar atributos que têm um impacto maior no processo de predição.

A partir do conjunto de dados filtrado, as interações de miRNAs e mRNAs de humanos foram mapeadas. Foi então construído um histograma de cardinalidade de interações (Figura 3). É importante notar que o número de interações de cada miRNA pode variar, chegando a até 2300 interações, nos casos mais raros. Esse fato demonstra que a predição das interações é uma tarefa desafiadora, considerando que não é todo miRNA que interage com todo mRNA.

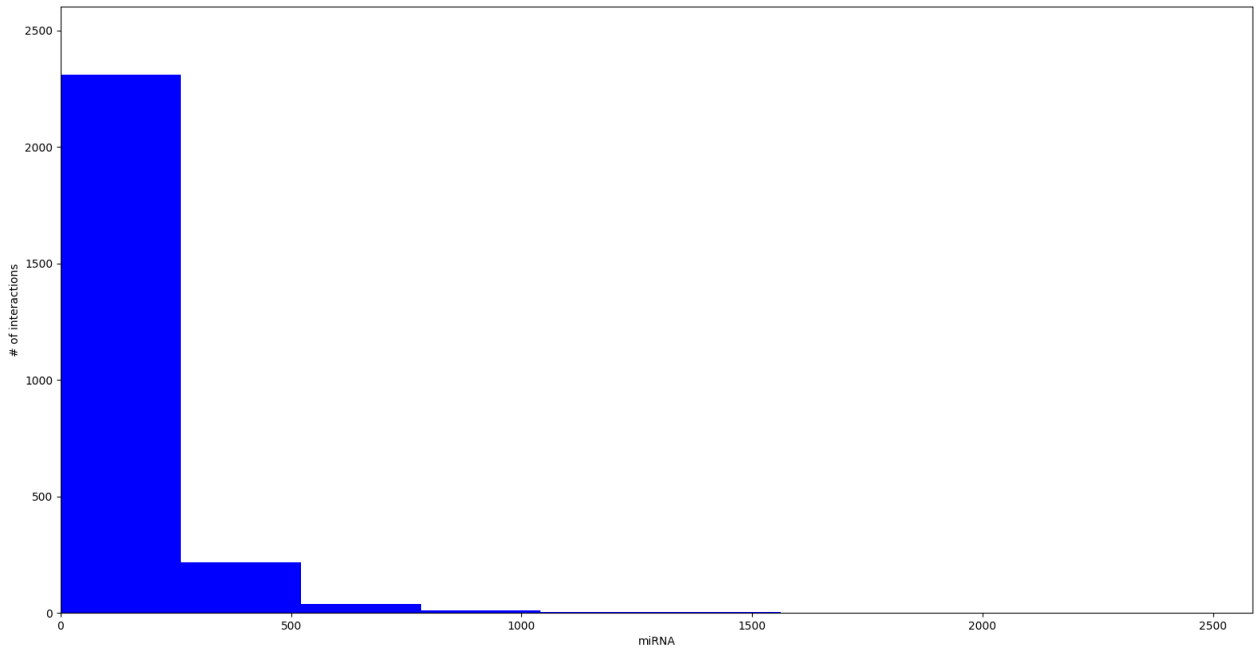

Figura 3: Cardinalidade do conjunto de dados

Por meio do histograma da Figura 3 é possível observar que além de alta dimensionalidade, o conjunto de dados é esparso. Apenas o modelo de Aprendizado Multirrótulo não terá o desempenho necessário para que as predições sejam confiáveis, dada a existência de muitos rótulo com poucos exemplos positivos. Por este motivo que é proposto a eliminação temporária de rótulos de baixa cardinalidade para que o classificador multirrótulo tenha maior confiabilidade.

3.2 Extração de Atributos

Em (Agarwal et al., 2015) foram estudados atributos de alvos e as interações entre miRNAs. Foram levantadas 26 atributos, sendo 14 destes dados como relevantes para a predição. Tomando esse estudo como base, serão utilizadas seis atributos selecionados dos 14 relevantes para o aprendizado multirrótulo. Eles são listadas a seguir.

- Comprimento da ORF;
- Comprimento da 3'-UTR;
- Comprimento da 5'-UTR;
- Conteúdo de AU na ORF;
- Conteúdo de AU na 3'-UTR;
- Conteúdo de AU na 5'-UTR;

Estes 6 atributos foram selecionados para o problema pelos dados serem de rápida obtenção.

3.3 Medidas de Avaliação

Medidas de avaliação convencionais não são adequadas para avaliar o desempenho de classificadores multirrótulo. Assim, medidas específicas foram propostas. Algumas delas são apresentadas a seguir.

Duas medidas muito utilizadas são a Acurácia (Equação 6), juntamente com o *Hamming Loss* (Equação 7) para avaliar a confiabilidade do modelo. Serão avaliadas também a Precisão e Recall obtidas pelo preditor. Nas seguintes equações, *H* representa o modelo de classificação, *D* o conjunto de dados, Y_i o conjunto de rótulos verdadeiros e Z_i o conjunto de rótulos preditos por *H*.

$$Accuracy(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \bigcap Z_i|}{|Y_i \bigcup Z_i|} \tag{6}$$

$$HammingLoss(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \Delta Z_i|}{|L|} \tag{7}$$

$$Precision(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \bigcap Z_i|}{|Z_i|} \tag{8}$$

$$Recall(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \bigcap Z_i|}{|Y_i|} \tag{9}$$

3.4 Ferramentas de desenvolvimento

As implementações serão realizadas com a linguagem de programação *Python*, em conjunto com a biblioteca de Aprendizado de Máquina Scikit-learn (Pedregosa et al., 2011). Tal biblioteca é amplamente utilizada na comunidade Python para o uso de algoritmos e estruturas de Aprendizado de Máquina. Todos os métodos e algoritmos para o treino e implementação dos modelos de classificação já estão inclusos na biblioteca *Scikit-learn*. Para a mineração de Regras de Associação, será utilizada a biblioteca *ARules* (Hahsler, Chelluboina, Hornik, & Buchta, 2011) da linguagem de programação *R*.

4 Resultados Preliminares

Um planejamento e desenvolvimento prévio já vem sendo realizado. A Figura 4 representa os estágios do desenvolvimento em alto nível.

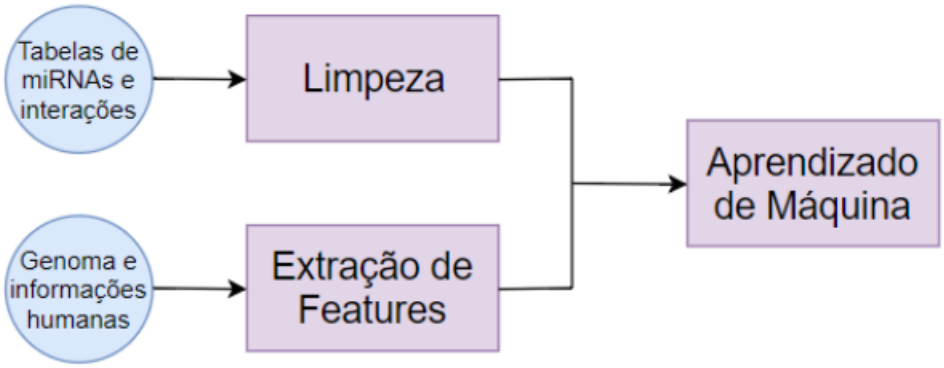

Figura 4: Estágios do desenvolvimento do projeto

A etapa de limpeza é responsável pela obtenção e filtragem das amostras de dados. A etapa de extração de atributos é responsável pela obtenção dos atributos e rótulos das amostras, e prepara o conjunto de dados para o estágio de Aprendizado de Máquina.

A Figura 5 é um diagrama dos subprocessos de cada estágio. Os quadrados preenchidos em verde são os processos já realizados previamente.

A seguir, são descritos todos os processos dentro do desenvolvimento do projeto.

- **Cruzamento de bancos de dados:** como dois bancos de dados de alvos de miRNAs estão sendo usados, um contendo informação dos alvos e outro contendo as interações, foi feito o cruzamento dos bancos e tirado a intersecção de ambos;
- **Download de informações dos alvos:** visto os alvos que sobraram, a ferramenta se conecta com o National Center for Biotechnology Information (NCBI) e faz o download automático das informações do alvos. Tais informações são importantes pois alvos podem evoluir e receber nova identificação, ou até mesmo serem desconsiderados. Estes casos extremos devem ser removidos dos nossos processos para evitar *outliers*;

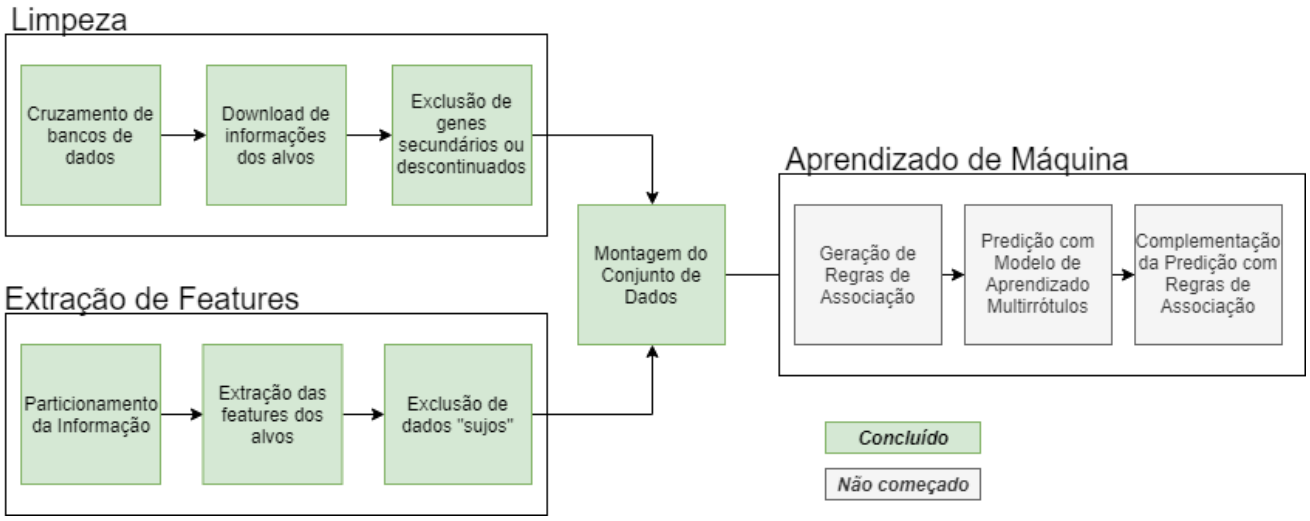

Figura 5: Processos ordenados de cada estágio

- **Exclusão de genes secundários ou descontinuados:** alvos vistos como “descontinuados” ou “secundários” são casos extremos que não precisam estar no conjunto de dados. A filtragem dos alvos foi feita neste processo;
- **Particionamento da informação:** a obtenção das atributos é um processo intenso e computacionalmente demorado. Para diminuir o tempo de execução, técnicas de otimização foram aplicadas, e um algoritmo de extração foi desenvolvido especificamente para a extração dos comprimentos dos alvos;
- **Extração das atributos dos alvos:** o algoritmo de extração é executado e um protótipo do conjunto de dados final é gerado;
- **Exclusão de dados “sujos”:** as amostras tidas como “descontinuadas” ou “secundárias” no estágio de limpeza são removidas do protótipo do conjunto de dados;
- **Montagem do conjunto de dados:** o conjunto de dados protótipo é incrementado com os seus devidos rótulos, gerando assim o conjunto de dados final e pronto para o Aprendizado de Máquina;
- **Geração das Regras de Associação:** visto que a dimensionalidade do conjunto de dados é muito grande, foram analisadas diferentes possibilidades de redução. A proposta mais efetiva foi minerar Regras de Associação apenas no espaço de rótulos. A mineração salvará as Regras que tenham suporte baixo e alta confiabilidade e removerão temporariamente os rótulos inferidos do conjunto de dados, de forma que possam ser acrescentados novamente, após o classificador fazer a predição das interações dos rótulos de alta cardinalidade. O limiar para definir quais rótulos serão removidos está para ser definido.
- **Predição com Modelo de Aprendizado Multirrótulos:** visto que a dimensionalidade resultante se mantém alta, novas técnicas estão sendo estudadas e testadas com o intuito de criar um modelo de classificação multirrótulo confiável;

- **Complementação da predição com Regras de Associação:** após a predição com o modelo multirrótulos, serão aplicadas as regras de associação extraídas previamente para realizar a predição dos rótulos reduzidos.

Os resultados preliminares e ferramenta de construção do conjunto de dados serão disponibilizados livremente após a conclusão da pesquisa.

5 Plano de Trabalho e Cronograma

O plano de trabalho deste projeto é composto de 3 etapas. Elas são listadas a seguir.

- **Mineração de Regras de Associação:** neste estágio será usada a biblioteca *ARules* da linguagem *R* para gerar as regras de associação;
- **Treinamento e Experimentação:** nesta etapa, iremos estudar modelos de classificação multirrótulos que tenham confiabilidade ao trabalhar com dados esparsos de alta dimensionalidade. O treino e avaliação desses modelos e a possível melhoria nos dados que podem aumentar a eficiência e confiabilidade do preditor também serão investigados;
- **Comparação com a Literatura:** por fim, será feita a comparação dos resultados obtidos nos experimentos com os resultados de estratégias existentes na literatura.

6 Forma de Análise dos Resultados

Após os resultados serem obtidos, comparações serão feitas entre o método desenvolvido e outros trabalhos da literatura. Serão avaliados a velocidade de execução, confiabilidade das predições, e número de atributos utilizadas para atingir os resultados.

Confiabilidade de predições é a métrica mais importante. Uma ferramenta melhor é capaz de classificar um conjunto maior de alvos de miRNAs corretamente.

Velocidade de execução é a métrica menos importante, dado que não trata-se do foco dessa pesquisa. Não apenas serão analisadas técnicas de AM mais eficientes, mas também maneiras de otimizar sua execução.

Número de atributos é uma métrica puramente comparativa. Caso o método desenvolvido tiver a mesma confiabilidade do que outros métodos da literatura que usam mais atributos para predizer seus alvos, haverá mais evidências de que o método desenvolvido é superior.

7 Resultados Esperados

Como resultados, espera-se a obtenção de um método eficiente e confiável para a predição de interações de alvos de miRNAs. É esperado que essa pesquisa auxilie em futuras pesquisas que envolvam a predição de alvos de miRNAs utilizando Aprendizado de Máquina. Os dados gerados também podem auxiliar outros pesquisadores no desenvolvimento de suas pesquisas. A utilização de aprendizado mutirrótulo é inovadora para tal tarefa.

A ferramenta de geração dos conjuntos de dados *atributos/interações* será disponibilizada juntamente com o método de predição, para auxiliar outros pesquisadores que venham a precisar de um conjunto de dados de interações.

8 Cronograma de Trabalho

A Tabela 11 apresenta a distribuição das tarefas restantes em função do tempo em bimestres. É importante notar que o estudo de modelos multirrótulo é o processo inicial, e que o estudo da literatura se dará durante todo o desenvolvimento da pesquisa. Não estão incluídas no cronograma as atividades referentes à escrita de artigos científicos e relatórios.

		Bimestre					
		1	2	3	4	5	6
ATIVID.	Estudo da literatura	X	X	X	X	X	X
	Estudo e Implementação do modelo multirrótulo	X	X	X			
	Predição com Regras de Associação			X	X	X	
	Comparação da Literatura					X	X

Tabela 11: Cronograma de atividades

Referências

Agarwal, V., Bell, G. W., Nam, J.-W., & Bartel, D. P. (2015). Predicting effective microrna target sites in mammalian mrnas. *elife*, *4*, e05005.

Aggarwal, C. C. (2014). *Data classification: algorithms and applications*. CRC press.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, vldb* (Vol. 1215, pp. 487–499).

Bartel, D. P. (2009). Micrornas: target recognition and regulatory functions. *cell*, *136*(2), 215–233.

Charte, F., Rivera, A., del Jesus, M. J., & Herrera, F. (2012). Improving multi-label classifiers via label reduction with association rules. In *International conference on hybrid artificial intelligence systems* (pp. 188–199).

Clare, A., & King, R. D. (2001). Knowledge discovery in multi-label phenotype data. In *European conference on principles of data mining and knowledge discovery* (pp. 42–53).

Coordinators, N. R. (2016). Database resources of the national center for biotechnology information. *Nucleic acids research*, *44*(Database issue), D7.

Costa, M. A., de Pádua Braga, A., & de Menezes, B. R. (2007). Improving generalization of mlps with sliding mode control and the levenberg–marquardt algorithm. *Neurocomputing*, *70*(7), 1342–1347.

Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment*, *61*(3), 399–409.

Griffiths-Jones, S. (2006). mirbase: the microrna sequence database. In *Microrna protocols* (pp. 129–138). Springer.

Hahsler, M., Chelluboina, S., Hornik, K., & Buchta, C. (2011). The arules r-package ecosystem: analyzing interesting patterns from large transaction data sets. *Journal of Machine Learning Research*, *12*(Jun), 2021–2025.

He, L., & Hannon, G. J. (2004). Micrornas: small rnas with a big role in gene regulation. *Nature Reviews Genetics*, *5*(7), 522.

Hornik, K., Grün, B., & Hahsler, M. (2005). arules-a computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, *14*(15), 1–25.

Krol, J., Loedige, I., & Filipowicz, W. (2010). The widespread regulation of microrna biogenesis, function and decay. *Nature Reviews Genetics*, *11*(9), 597.

Liu, H., Yue, D., Chen, Y., Gao, S.-J., & Huang, Y. (2010). Improving performance of mammalian microrna target prediction. *BMC bioinformatics*, *11*(1), 476.

Mitchell, T. M., et al. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, *45*(37), 870–877.

Morris, K. V., & Mattick, J. S. (2014). The rise of regulatory rna. *Nature Reviews Genetics*, *15*(6), 423.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009). Classifier chains for multi-label classification. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 254–269).

Schapire, R. E., & Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. *Machine learning*, *39*(2-3), 135–168.

Sturm, M., Hackenberg, M., Langenberger, D., & Frishman, D. (2010). Targetspy: a supervised machine learning approach for microrna target prediction. *BMC bioinformatics*, *11*(1), 292.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining* (US ed ed.). Addison Wesley. Hardcover.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, *3*(3), 1–13.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2009). Mining multi-label data. In *Data mining and knowledge discovery handbook* (pp. 667–685). Springer.

Zhang, M.-L., & Zhou, Z.-H. (2006). Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering*, *18*(10), 1338–1351.