UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE
BACHARELADO EM CIÊNCIAS BIOLÓGICAS

PREDIÇÃO DE INTERAÇÕES ENTRE PIRNAS E ELEMENTOS TRANSPONÍVEIS POR MEIO DE PREDICTIVE BI-CLUSTERING TREES

Hiago Freire Oliveira

São Carlos
2022

Hiago Freire Oliveira

PREDIÇÃO DE INTERAÇÕES ENTRE PIRNAS E ELEMENTOS TRANSPONÍVEIS POR MEIO DE PREDICTIVE BI-CLUSTERING TREES

Trabalho de Conclusão de Curso apresentado na Universidade Federal de São Carlos para obtenção do título de Bacharel em Ciências Biológicas.

Orientadora: Jane Eyre Gabriel

Coorientador: Ricardo Cerri

São Carlos

2022

PREDIÇÃO DE INTERAÇÕES ENTRE PIRNAS E ELEMENTOS TRANSPONÍVEIS POR MEIO DE PREDICTIVE BI-CLUSTERING TREES

FICHA CATALOGRÁFICA

Freire-Oliveira, Hiago

Título do trabalho: Predição de interações entre piRNAs e elementos transponíveis por meio de Predictive Bi-Clustering Trees.

Freire-Oliveira. -- São Carlos: UFSCar, 2022.

51p.

Trabalho de Conclusão de Curso -- Universidade Federal de São Carlos, 2022.

1. Bioinformática. 2. RNA não codificante. 3. Aprendizado de máquina.

FOLHA DE APROVAÇÃO

HIAGO FREIRE OLIVEIRA

PREDIÇÃO DE INTERAÇÕES ENTRE PIRNAS E ELEMENTOS TRANSPONÍVEIS POR MEIO DE PREDICTIVE BI-CLUSTERING TREES

Monografia apresentada junto ao curso de Ciências Biológicas para obtenção do título de Bacharel em Ciências Biológicas. Universidade Federal de São Carlos. São Carlos, 26 de abril de 2022.

_____

Orientadora:
Profa. Dra. Jane Eyre Gabriel
Universidade Federal de São Carlos

_____

Coorientador:
Prof. Dr. Ricardo Cerri
Universidade Federal de São Carlos

_____

Examinador:
Prof. Dr. Marco Antonio Del Lama
Universidade Federal de São Carlos

_____

Examinador:
Prof. Dr. Francis de Morais Franco Nunes
Universidade Federal de São Carlos

AGRADECIMENTOS

Agradeço, primeiro, a meus pais, pela confiança depositada em mim com seus investimentos, permitindo que eu morasse fora e estudasse aquilo que tenho paixão, sempre com suporte emocional, mas nunca sem a cobrança devida de esforço como retorno. À minha mãe, em especial, agradeço por todo o amor e interesse por cada novidade que eu trago quando volto pra casa. Ao meu pai, agradeço por nunca duvidar de minha capacidade e sempre demonstrar seu orgulho. Agradeço ao meu irmão que, não fosse seu interesse genuíno por computadores em uma época em que eu não tinha informações sobre isso em nenhum outro lugar, dificilmente teria escolhido o caminho desse trabalho. Agradeço à minha irmã que me inspira diariamente a sempre seguir em frente com quaisquer forças que ainda sobrem, sempre.

Agradeço a meus amigos da panelinha mais irritantemente brilhante do bacharel 2018, que me aceitaram e me apoiaram constantemente, em especial à encantadora Isabela, que se tornou a mais próxima de mim nesses anos e cuja parceria sempre tive a mais plena certeza. Apesar de tê-la mais perto, por todos os demais, Jacque e sua dedicação e foco implacáveis, Isra e sua paixão profunda pelo conhecimento, Príncia e sua admirável persistência na pesquisa desde o início da graduação, e Vitão e sua autenticidade incomparável, meu carinho e admiração é verdadeiro e duradouro.

Agradeço à Evelyn do Laboratório de Neurociências, que me guiou na minha primeira (e talvez única) experiência de pesquisa com animais e me ensinou muito com gentileza e atenção. Também não posso deixar de agradecer à Jé, igualmente importante no início da graduação, que me acolheu e acompanhou com afeição tanto em rolê quanto em estudo.

Agradeço a meu querido trio de amigos de Poços de Caldas que mantiveram o contato após esses anos e toparam os rolês que, embora poucos, ainda assim aconteceram. Agradeço imensamente ao meu corononalove, Laura, pelo companheirismo e carinho que me ajudaram a segurar as pontas nesses últimos anos tão esquisitos, além de suportar minhas conversas esquisitas sobre liberdade.

Agradeço ao professor Cerri por me incentivar a procurar por respostas por mim mesmo, acreditando, desde o início, no meu potencial de aprender algo tão fora da minha zona de conforto. Agradeço também ao professor Del Lama por me abrir

um horizonte de encanto pela biologia em sua disciplina de Genética de Populações, que ficou profundamente marcada em mim como a melhor disciplina que já cursei. Não só isso, por ter me enchido de honra ao aceitar o convite.

Por fim, agradeço a qualquer um que já tenha me estimulado intelectualmente de alguma forma. Cada discordância, discussão e argumentação objetiva me levou a reconstruir meu julgamento, me impelindo a arquitetar um indivíduo mais pleno. Obrigado por me conduzirem para mais perto da verdade por meio da racionalidade. A é A.

“Eu juro pela minha vida e pelo meu amor por ela que nunca irei viver em função de outro homem, nem vou pedir a outro homem que viva em função de mim.”

(RAND, 1957).

RESUMO

RNAs interagentes com PIWI (piRNAs) são uma classe de RNAs de interferência cujas ações variam de regulação da expressão gênica ao combate de infecções virais e silenciamento de elementos transponíveis, possuindo características únicas como tamanho de 21 a 35 nucleotídeos, viés de uracila na extremidade 5' e 2'-O-metilação na extremidade 3'. Os elementos transponíveis (TEs) são elementos genéticos com capacidade de se moverem por entre os genomas hospedeiros, sendo divididos entre retrotransposons e transposons de DNA. A replicação dos TEs pode promover eventos recombinatórios danosos pela geração de quebras nas duplas-fitas de DNA, além de interferências na expressão, uma vez que seus promotores podem levar a uma transcrição aberrante dos genes vizinhos. O silenciamento desses elementos pelos piRNAs ocorre na linhagem germinativa da maioria dos animais e é essencial à manutenção da integridade do genoma. O problema de predição *in silico* de interação entre piRNAs e TEs foi abordado por aprendizado de máquina por meio de um algoritmo de árvores de decisões do tipo *Predictive Bi-Clustering Trees* (PBCT). A fim de melhorar o desempenho do algoritmo, a matriz de interações entre os piRNAs e TEs foi reconstruída por meio de um algoritmo do tipo *Beta-distribution-rescored Neighborhood Regularized Logistic Matrix Factorization* (NRLMFβ). O PBCT foi aplicado em configurações de validação cruzada de 5-*folds* e de 10-*folds*, tanto para a matriz sem reconstrução (BICT) quanto para a matriz reconstruída por NRLMFβ (BICTR). De forma geral, o PBCT aplicado a esse conjunto de dados não foi capaz de predizer as interações positivas satisfatoriamente, comportando-se como um classificador aleatório. Comparativamente, no método BICT, o PBCT apresentou maiores valores de AUROC e AUPRC. Entretanto, no método BICTR, o PBCT foi capaz de predizer corretamente mais interações positivas, que são, de fato, o maior interesse neste estudo. Possíveis aplicações biológicas e caminhos para melhorar o desempenho do algoritmo também foram consideradas.

Palavras-chave: piRNA. Elemento transponível. Predição de interação. Árvore de decisão. Aprendizado de máquina.

ABSTRACT

PIWI-interacting RNAs (PiRNAs) are a class of interfering RNAs whose actions range from regulating gene expression to fighting viral infections and silencing transposable elements, possessing unique characteristics such as being from 21 to 35 nucleotides long, displaying an uracil bias at the end 5' and 2'-O-methylation at the 3' end. Transposable elements (TEs) are genetic elements capable of moving between host genomes, being split into retrotransposons and DNA transposons. The replication of TEs can promote harmful recombination events by generating breaks in DNA double strands, in addition to interference in expression, considering that their promoters can lead to aberrant transcription of neighboring genes. Silencing of these elements by piRNAs occurs in the germ line in the majority of animals and is essential for the maintenance of genome integrity. The problem of *in silico* prediction of interaction between piRNAs and TEs was addressed by machine learning using a decision tree algorithm, namely Predictive Bi-Clustering Trees (PBCT). In order to improve the algorithm's performance, the interaction matrix of piRNAs and TEs was reconstructed by means of an Beta-distribution-rescored Neighborhood Regularized Logistic Matrix Factorization (NRLMFβ) algorithm. PBCT was applied in 5-fold and 10-fold cross-validation configurations, both for the matrix without reconstruction (BICT) and for the matrix reconstructed by NRLMFβ (BICTR). In general, PBCT applied to this dataset was not able to predict positive interactions satisfactorily, behaving as a random classifier. Comparatively, in the BICT method, PBCT presented higher values of AUROC and AUPRC. However, in the BICTR method, PBCT was able to correctly predict more positive interactions, which are, in fact, the major interest in this study. Potential biological applications and ways to improve the algorithm's performance were also considered.

Keywords: piRNA. Transposable element. Interaction prediction. Decision tree. Machine learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AGO.	argonauta
AUPRC.	área sob a curva de precisão e revocação
AUROC.	área sob a curva característica de operação do receptor
BICT.	árvores de *bi-clustering*
BICTR.	árvores de *bi-clustering* com reconstrução do espaço de saída
CLASH.	reticulação, ligação e sequenciamento de híbridos
DNA.	ácido desoxirribonucleico
dsRNA.	RNA dupla-fita
LTR.	repetições terminais longas
miRNA.	microRNA
mRNA.	RNA mensageiro
NRLMF.	fatorização logística de matriz regularizada por vizinhança
PATC.	agrupamento periódico de A_n/T_n
PBCT.	árvores preditivas de *bi-clustering*
PCT.	árvores preditivas de *clustering*
piRNA.	pequenos RNAs associados à PIWI
PIWI.	*P-element-induced wimpy testis*
PR.	precisão-revocação
PRG-1.	*P53-responsive gene 1*
RISC.	complexo de silenciamento induzido por RNA
RNA.	ácido ribonucleico
RNAi.	RNA de interferência
ROC.	característica de operação do receptor
siRNA.	RNAs interferentes pequenos
TE.	elemento transponível
TIR.	repetições terminais invertidas
tRNA.	RNA transportador
WAGO.	AGO específica de verme

ÍNDICE

1. INTRODUÇÃO

1.1 Dos pequenos RNAs associados à PIWI

Nesta seção, serão apresentados os RNAs e os diversos tipos de RNAs de interferência, com ênfase em piRNAs e suas características mais relevantes para a interações com os elementos transponíveis, interações estas de suma importância ao presente trabalho.

A molécula de ácido ribonucleico (RNA) assume um papel protagonista no Dogma Central da Biologia Molecular, participando na transferência de informação da molécula de ácido desoxirribonucleico (DNA) até a síntese proteica (CRICK, 1970). Além de sua ação como molde no Dogma Central, o RNA mensageiro (ou mRNA) também apresenta uma ação catalítica amplamente discutida, em que ribozimas (enzimas de RNA) seriam capazes de agir como molde para serem copiadas, mas também atuar como catalisadoras da adição ou da remoção de bases, exemplificadas por elementos transponíveis e íntrons de auto-splicing (GILBERT, 1986).

Morris e Mattick (2014) descrevem um histórico de como essas concepções clássicas da função do RNA foram, em muito, incrementadas até o presente momento. Conforme notam, a noção “1 gene – 1 proteína” já é desacreditada desde a década de 70, dado o descobrimento do processamento de íntrons (*splicing* alternativo). Entretanto, com a descoberta dos íntrons nucleares e o RNA de interferência (RNAi), tais como os microRNAs e os RNAs interferentes pequenos, mesmo a ideia de proteína como sinônimo de produto gênico tem decaído, embora a passos pequenos.

Os RNAs *lin-4* e *let-7* foram os primeiros microRNAs (miRNAs) descobertos, os quais atuam na regulação temporal do desenvolvimento em *Caenorhabditis elegans*. Essas moléculas foram classificadas como RNAi regulatórios de aproximadamente 22 nucleotídeos. Posteriormente, muitos outros miRNAs foram descritos, com ação geral na inibição da tradução, mas também na aceleração da degradação dos mRNAs com os quais se pareiam (MORRIS; MATTICK, 2014). As sequências de miRNAs estão dispostas ao longo de todo o genoma humano, seja em regiões exônicas, intrônicas e mesmo intergênicas. De especial relevância na última década, os miRNAs foram associados a funções relevantes no

desenvolvimento do câncer, podendo agir tanto como supressores de tumor, tanto como oncomiR (reprimindo outros supressores de tumor) (ROMERO-CORDOBA *et al*, 2014).

Ainda, segundo Morris e Mattick (2014), a segunda classe de RNAi descoberta foi a dos siRNAs (*small interfering* RNAs, ou RNAs interferentes pequenos). Várias características os tornam similares aos miRNAs, como o tamanho aproximado de 22 nucleotídeos, biogênese a partir de dsRNA (*double-strand DNA*, isto é, DNA dupla-fita), o processamento pela enzima citoplasmática Dicer e o carregamento por enzimas AGO (Argonauta). Os siRNAs são carregados pelo componente AGO do complexo RISC (*RNA-induced silencing complex*, ou complexo de silenciamento induzido por RNA), o qual é guiado à molécula de mRNA complementar, resultando em clivagem e degradação do mesmo (ZENG; YI; CULLEN, 2003).

Os siRNAs são distintos uma vez que os miRNAs atuam, em geral, por pareamento parcial de bases com o mRNA alvo e restringem a expressão de diversos genes de sequências similares, enquanto os siRNAs atuam por pareamento perfeito e são específicos a regiões do mRNA alvo. Não somente, enquanto o miRNA tem um papel mais relacionado à regulação da expressão gênica, os siRNAs aparentam estar relacionados ao reconhecimento de ácidos nucleicos para fins imunológicos, podendo ser utilizados, por aplicação biotecnológica, contra todos os tipos de genomas virais, seja de DNA ou RNA, seja dupla ou simples fita (QURESHI *et al.*, 2018).

Uma classe de pequenos RNAs de características menos semelhantes é a dos RNAs interagentes com PIWI (piRNAs). Uma diferença notável em relação às outras classes é a especificidade de carregamento pelas proteínas PIWI (*P-element-induced wimpy testis*, nome derivado da descoberta inicial em testículos de *Drosophila melanogaster*), presentes predominantemente no núcleo celular (MORRIS; MATTICK, 2014). Outras diferenças compreendem o processamento a partir de transcritos de um longo precursor simples-fita, além de não requererem processamento pela Dicer (OZATA *et al*., 2019).

As famílias tipo Ago, tipo PIWI, WAGO (*worm-specific* AGO, ou AGO específica de verme) e Ago de *Trypanosoma* compreendem as principais classificações de proteínas Argonauta eucarióticas. A arquitetura das proteínas Argonauta compreendem um núcleo básico constituído pelos domínios N-terminal

(amino-terminal), PAZ (*PIWI-Argonaute-Zwille*), MID (*Middle*) e PIWI. As diferenças estruturais entre as proteínas Argonauta aparentam estar intimamente relacionadas não só aos tipos de pequeno RNA com o qual se ligam, como também nas estratégias adotadas para se ligarem a seus alvos (WU *et al.*, 2020).

Os piRNAs variam de 21 a 35 nucleotídeos, apresentam um viés de uracila na extremidade 5' e portam uma 2'-O-metilação na extremidade 3'. Os precursores de piRNA são transcritos de *loci* genômicos chamados de *clusters* (agrupamentos) de piRNA. As ações dos piRNAs variam de regulação da expressão gênica ao combate de infecções virais e silenciamento de elementos transponíveis. De modo similar à RISC, os piRNAs guiam as proteínas PIWI até o RNA-alvo para a clivagem. Além disso, promovem a montagem de heterocromatina e metilação do DNA (OZATA *et al*., 2019).

Ainda segundo Ozata *et al.* (2019), em *C. elegans*, os piRNAs são classicamente conhecidos como 21U-RNAs, dado ao viés de uracila e o comprimento de 21 nucleotídeos. Diferentemente do mecanismo de processamento do piRNA a partir de transcritos de um longo precursor fita-simples comentado anteriormente, os 21U-RNAs do tipo I são transcritos a partir de aproximadamente 12 mil mini-genes pelo fator de transcrição da família de proteínas *Forkhead*, que se liga ao motivo *Ruby* a montante de cada precursor de piRNA. Um transcrito precursor 5'-capeado de apenas 25-27 nucleotídeos é gerado a partir de cada mini-gene do tipo I. O precursor é carregado à proteína PRG-1 (*P53-Responsive Gene 1*) e processado a piRNA maduro. Já os 21U-RNAs do tipo II são transcritos a partir de promotores de genes produtores de proteínas de tamanho completo.

Um dos importantes papéis reconhecidos de piRNAs em *C. elegans* é a capacidade de reconhecimento entre transcritos próprios (*self*) e não próprios (*non-self)*. Isso permite o reconhecimento e silenciamento de transgenes e novas inserções transposônicas. Diferente das proteínas PIWI citoplasmáticas de outros animais, a capacidade incisiva da proteína PIWI de *C. elegans* (PRG-1) não é imperiosa, já que os piRNAs induzem a transcrição de siRNAs secundários por uma RNA polimerase dependente de RNA, utilizando o próprio alvo como molde. Os siRNAs resultantes são denominados 22G-RNA (novamente devido ao viés do primeiro nucleotídeo e o comprimento do transcrito). Os 22G-RNAs são carregados em proteínas WAGO, que enfim silenciam o transcrito não próprio (OZATA *et al*., 2019). É igualmente notável que certas AGO, tais como WAGO-9 e CSR-1

(*Chromosome Segregation and RNAi Deficient 1*) estejam envolvidas em processos de herança epigenética desse silenciamento de transcritos não próprios e proteção de transcritos próprios, respectivamente (SHIRAYAMA *et al.*, 2012).

Segundo Zhang *et al.* (2018), dos cerca de 15 mil piRNAs codificados por *C. elegans*, a grande maioria não apresenta ampla complementaridade a transposons. Muitos genes endógenos contêm sítios alvo de piRNAs, mas exibem resistência ao silenciamento, sobretudo devido à presença de PATCs (*Periodic* A_n/T_n *Clusters*, ou agrupamentos periódicos de A_n/T_n). Os PATCs são elementos não-codificantes espaçados por 10 pares de bases, de forma que podem licenciar a expressão do gene portador na linha germinativa em domínios de cromatina repressiva, representando assim a base do sistema imune celular (FRØKJÆR-JENSEN *et al*., 2016). Ainda segundo Zhang *et al.* (2018), a região semente de um piRNA vai do segundo ao sétimo nucleotídeo, e é distinta pelo pareamento ótimo aos sítios-alvo. Fora dessa região, até seis bases não pareadas tornam o transcrito um alvo potencial do piRNA.

1.2 Dos elementos transponíveis

Nesta seção, serão apresentados os elementos transponíveis, suas possíveis classificações e características mais relevantes para o estudo de suas interações com os piRNAs, interações estas de suma importância ao presente trabalho.

Os elementos transponíveis (TEs, do inglês *Transposable Elements*) são elementos genéticos com capacidade de se moverem por entre os genomas hospedeiros (MCCULLERS; STEINIGER, 2017). Cerca de 12% do genoma de *C. elegans* é composto desses elementos (BESSEREAU, 2006). Os TEs são classificados em duas grandes classes: classe 1, constituída de retrotransposons, e classe 2, composta de transposons de DNA. Os retrotransposons fazem a transposição a partir de uma cópia genômica que é transcrita a um intermediário de RNA, o qual é reversamente transcrito a DNA por uma transcriptase reversa (mecanismo de “cópia-e-colagem”) (WICKER *et al.,* 2007). Dessa classe, são exemplos notáveis os LTR (*Long Terminal Repeats*, ou repetições terminais longas), dentre os quais está a família *gypsy*, os quais são controlados pelo *cluster flamenco* em *Drosophila* (BRENNECKE *et al*., 2007), e o grupo *Cer* (composto por elementos

gypsy e *Bel*), o qual perfaz cerca de 0,4% de todo o genoma de *Caenorhabditis elegans* (GANKO *et al*., 2003).

Já os transposons de DNA são divididos em duas subclasses: subclasse 1, compreendendo os TEs possuidores de repetições terminais invertidas (TIR) e que utilizam o mecanismo “corte-e-colagem”, e subclasse 2, que compreende TEs que replicam sem clivagem da dupla-fita (“cópia-e-colagem”), diferindo dos retrotransposons pela ausência de intermediários de RNA, de tal maneira que uma única fita de DNA seja deslocada (WICKER *et al.,* 2007). Da subclasse 1, são exemplos notáveis os elementos P, associados à causa da disgenesia do híbrido em *Drosophila* (SILVA; KIDWELL, 2000) e os Tc1/*mariner*, primeiro descobertos em *C. elegans*, mas que provavelmente são os transposons de DNA mais pervasivos de toda a natureza (PLASTERK; IZSVÁK; IVICS, 1999). Da subclasse 2, uma subclasse bem menos diversa, são exemplos dignos de nota os *Helitrons*, que se transpõem via replicação por um círculo rolante e que, embora melhor caracterizados em plantas, constituem cerca de 2% do genoma do nemátodo modelo (KAPITONOV; JURKA, 2001).

Segundo Ozata *et al.* (2019), a replicação dos TEs pode promover eventos recombinatórios danosos pela geração de quebras nas duplas-fitas de DNA, além de interferências na expressão, uma vez que seus promotores podem levar a uma transcrição aberrante dos genes vizinhos. Todavia, os transposons devem se integrar ao DNA da célula germinativa para sobreviver. Na maioria dos animais, ao menos um subconjunto de piRNAs defendem o genoma da linhagem germinativa contra a mobilização de TEs. Uma vez silenciados, mutações podem enfim inativar proteínas codificadas por transposons, levando à sua derrocada.

A presença de piRNAs não é restrita à linhagem germinativa em todos os animais. A maioria dos artrópodes possui, em adição, piRNAs em células somáticas, apresentando mecanismos mais diversos do que o estrito silenciamento de TEs, como a defesa contra vírus (em adição aos siRNAs) em *Aedes aegypti.* Segundo Lewis *et al.* (2018), o papel de defesa contra transposons em artrópodes ancestrais provavelmente era dependente não de piRNAs, mas de siRNAs catalisados por uma RNA polimerase dependente de RNA (assim como observado em *C. elegans*), que fornecia precursores dsRNA diferentes dos da RNA polimerase II, isto é, uma nova gama de diversidade de substratos. Nesse sentido, a redefinição de função dos piRNAs na história evolutiva dos artrópodes (como as drosófilas) é evidenciada pela

diversidade de funções de piRNAs somáticos, os quais são apontados como ancestrais a todos os artrópodes.

1.3 Da predição

Nesta seção, serão apresentados os métodos para descoberta de interação piRNA-alvo *in vivo* por meio do protocolo CLASH, bem como ferramentas *web* para identificação de sítios-alvo de piRNAs e o algoritmo de PBCT para predição de interação piRNA-TE, que configura, de fato, o escopo deste trabalho.

O protocolo de reticulação, ligação e sequenciamento de híbridos (CLASH, ou *cross-linking, ligation and sequencing of hybrids)* fornece os métodos para a descoberta de interações RNA-RNA *in vivo* de forma que o trabalho seja amenizado e o conhecimento prévio dos pares não seja necessário, conhecimento este necessário em métodos como a cristalografia de raio-X e a ressonância magnética nuclear. O CLASH utiliza de métodos de reticulação por luz UV, seguido de purificação por afinidade dos complexos RNA-proteína e de ligação e sequenciamento dos híbridos RNA-RNA. Essas moléculas quiméricas sequenciadas são o produto final do CLASH, de modo que representam um evento de ligação entre o pequeno RNA e um fragmento de seu mRNA alvo (KUDLA *et al,* 2011).

A interação piRNA-mRNA foi verificada *in vivo* em *C. elegans* por meio de CLASH, de sorte que que foi observado que todos os mRNAs da linhagem germinativa são submetidos à vigilância dos piRNAs, isto é, todo o transcriptoma. Não somente, interações entre piRNAs e outros ncRNAs (*non-coding* RNAs), embora muito menos frequentes, foram verificadas, de especial relevância interações piRNA-tRNA (RNA transportador), uma vez que em *Drosophila* o acúmulo de tRNA não processado devido a uma mutação leva à falência do silenciamento de transposons mediado por PIWI (SHEN *et al,* 2018).

Tendo em mente as propriedades já discutidas dos alvos de piRNAs (como o limite de bases não pareadas e sequência semente) e o custo elevado de experimentos em laboratórios físicos, não é de surpreender que ferramentas computacionais tenham surgido de modo a prover dados acerca dos alvos de piRNAs em diferentes espécies. O pirScan, por exemplo, é uma ferramenta baseada em web adequada para a identificação de sítios-alvo de piRNAs de *Caenorhabditis elegans,* dado um mRNA ou uma sequência de DNA destituída dos íntrons (WU *et*

al., 2018). Já o piRTarBase é um banco de dados *online* interativo que fornece sítios de interação de piRNAs em *C. elegans* e *C. briggsae,* tanto preditos quanto experimentalmente verificados, a partir da entrada de genes. Não somente, pode-se obter os alvos de mRNA a partir da entrada de um piRNA (WU *et al*., 2019).

Não obstante, um modelo de predição *in silico* de interações entre piRNAs e TEs, exclusivamente, não foi encontrado. Contudo, a extração de conhecimento útil desses dados recai em um grande problema da biologia computacional, o qual requer o desenvolvimento de ferramentas precisas ao caso (LARRAÑAGA et al., 2006). Uma vez que os TEs são os elementos que, de fato, são assolados pelos piRNAs, o interesse em compreender como um piRNA reconhece um elemento transponível é definitivo, mesmo com conhecimento dos PATCs. Assim, a modelagem em computador pode ajudar a encontrar padrões difíceis de observar na experimentação direta.

Devido à natureza complexa desse mecanismo, o alinhamento de sequências pode não ser suficiente para predizer interações dos alvos. Assim, Larrañaga *et al.* (2006) demonstraram diversas circunstâncias biológicas em que o aprendizado de máquina se mostra uma ferramenta exímia na resolução. Segundo Mitchell (1997), aprendizado de máquina é a área do conhecimento interessada no estudo de algoritmos que permitam que programas computacionais possam automaticamente melhorar por meio de experiência, isto é, sem a programação explícita. Em um problema que utilize o modelo de predição, o critério de performance a ser otimizado é a acurácia.

O problema de predição da interação entre dois pares de objetos pode ser abordado por meio de árvores de decisão (PLIAKOS; GEURTS; VENS, 2018). Árvores de decisão correspondem a uma das técnicas mais simples, embora eficazes, de aprendizado de máquina. Elas têm como objetivo classificar exemplos desconhecidos, sendo construídas por meio da análise de um conjunto de exemplos de treinamento. A classificação se dá mediante perguntas sobre atributos associados aos itens. Um nó da árvore corresponde a uma pergunta, que pode seguir a nós inferiores (dois ou mais) a depender da resposta, isto é, do valor do atributo da observação. O nó final, aquele que não faz mais perguntas, é chamado de folha, que é a classe do objeto (KINGSFORD; SALZBERG, 2008).

Conforme Pliakos, Geurts e Vens (2018), uma abordagem popular na construção dessas árvores é o emprego de PCT (*Predictive Clustering Trees,* ou

Árvores Preditivas de *Clustering*), em que cada nó é considerado como um *cluster* (agrupamento) dos dados, construindo-se a árvore indutivamente de cima para baixo. Dessa forma, o conjunto total de dados é o *cluster* de nó raiz, a partir do qual novos nós são recursivamente divididos testando-se uma das características. A melhor divisão é obtida ao se considerar todos os pontos de divisão de todas as características, avaliando-os pelo critério de ganho de informação (medida derivada do cálculo da entropia). Um critério de parada é baseado na pureza do nó, de modo que, quando os dados contidos nesse mesmo nó são puros em relação ao alvo, a classificação, ou seja, a predição, é dada segundo a classe mais atribuída às instâncias de treinamento. Uma vez que um piRNA pode interagir com mais de um elemento transponível, e vice-versa, o problema pode ser abordado por classificação multirrótulo, tendo que um dado multirrótulo é aquele que pode ser associado a várias classes simultaneamente. Dessa forma, o PCT pode ser empregado para predizer múltiplos alvos simultaneamente, caracterizando assim uma árvore de decisão de múltiplas saídas..

As diversas interações de piRNAs com TEs podem ser consideradas como uma rede heterogênea, já que são dois conjuntos de itens que interagem entre si, mas são descritos por suas próprias características. Para essa estrutura, pode-se aplicar a aprendizagem pelas abordagens local e global. A abordagem local se baseia na divisão dos dados em N subconjuntos, em concordância com a quantidade de classes, de tal maneira que cada subconjunto possua exemplos associados a uma das classes, levando ao treinamento de N classificadores binários em cada subconjunto. A integração das N saídas leva à predição nos exemplos de teste. Já a abordagem global se baseia em lidar com um método de aprendizado ajustado de forma a ser capaz de lidar com todas as classes simultaneamente, sem dividir o problema em subproblemas, utilizando um classificador único (PLIAKOS, VENS; 2019).

De acordo com Madeira e Oliveira (2004), algoritmos *bi-clustering* são uma classe de algoritmos *clustering* (agrupamento) que fazem o agrupamento simultâneo de linhas e colunas. Assim, a escolha de um algoritmo PBCT (*Predictive Bi-Clustering Trees*, ou Árvores Preditivas de *Bi-Clustering*) que, diferente do algoritmo PCT que trabalha com apenas um conjunto de atributos (no presente caso, os atributos de piRNAs) e teria como classes os TEs com quem interagem, trabalham com os 2 conjuntos simultaneamente (atributos dos piRNAs e atributos do

TEs), visando classificar a interação ou não interações entre os mesmos (PLIAKOS, VENS; 2019), é mais ajustada ao propósito do presente estudo.

2. OBJETIVOS

2.1. Objetivo geral

O objetivo geral deste trabalho foi avaliar o desempenho de um modelo de aprendizagem de máquina do tipo árvore de decisão na predição de interação de piRNAs e TEs, a fim de contribuir na pesquisa de redes de interações entre biomoléculas e seus diversos potenciais fins.

2.2. Objetivos específicos

I. Aplicar um modelo de algoritmo PBCT a dados de interações reais entre piRNAs e TEs;
II. Avaliar o desempenho do modelo;
III. Identificar os potenciais fins biológicos da predição de interação de biomoléculas por meio de aprendizado de máquina.

3. MATERIAIS E MÉTODOS

3.1 Das interações *in vivo*

As interações piRNA-TE utilizadas no presente estudo foram previamente descritas por Shen *et al.* (2018), que destacaram instruções detalhadas acerca da descoberta de interações *in vivo* piRNA-alvo por meio de CLASH em *Caenorhabditis elegans*. De todas as interações, 19.092 leituras (totais, isto é, com redundância) apontam a interação piRNA-TE. Essas leituras indicam também os sítios precisos de interação das moléculas.

As informações adicionais das leituras piRNA-TE de CLASH foram removidas, de modo que apenas as bases foram mantidas. Uma tabela foi gerada a partir dessas interações, de modo que as sequências de piRNAs compuseram o índice das linhas e as sequências dos TEs compuseram o índice das colunas. A tabela foi populada com 0 e 1, de tal maneira que cada interação piRNA-TE foi representada por 1 e as não-interações por 0. Assim, uma matriz binária de interação foi gerada. Uma vez construída, os índices e a matriz foram exportados independentemente: a matriz como um arquivo CSV (*comma-separated values* - valores separados por vírgula) e os índices como 2 arquivos FASTA individuais. Foram contadas 5.218 interações (únicas) em uma matriz de 13.841.400 itens (aproximadamente 0.000377% do total).

3.2 Da geração dos atributos

Os arquivos FASTA foram utilizados para a extração de atributos das sequências pelo *software Pse-in-One* (LIU *et al.*, 2015). Uma vez que a tarefa de predição não pode lidar com cadeias de caracteres (*strings*), a sequência deve ser convertida em um vetor de dimensão fixa (18 itens, no caso deste estudo), o qual contenha os atributos chave. Esse vetor é denominado de vetor de atributos. O *Pse-in-One* possui 3 sub-servidores *web*, dentre os quais foi utilizado o *PseDAC-General* (*pseudo deoxyribonucleic acid compositions for DNA sequences* - composições de ácido pseudodesoxirribonucleico para sequências de DNA). O *PseDAC-General* contém 3 categorias, sendo composição de nucleotídeo, autocorrelação de nucleotídeo e composição de pseudonucleotídeo, de tal maneira

que 16 modos estão distribuídos entre essas categorias (LIU *et al.*, 2015). Os atributos calculados no presente estudo derivam do modo *PseDNC* (*Pseudo dinucleotide composition* - composição de pseudodinucleotídeo), o qual faz parte da categoria de composição de pseudonucleotídeo. O modo *PseDNC* leva em conta parâmetros estruturais locais da sequência, isto é, parâmetros angulares e translacionais (CHEN *et al*., 2013).

O *software* gerou, então, 2 novos arquivos CSV, sendo um deles o arquivo com os vetores de atributos extraídos de cada um dos piRNAs e outro deles o arquivo com atributos extraídos de cada um dos TEs, na exata ordem na matriz binária previamente gerada. Assim, com 3 arquivos CSV, isto é, um de atributos de piRNAs, um de atributos de TEs, e outro de informação de interações entre os pares de objetos (Figura 1), o algoritmo de PBCT, segundo Pliakos, Geurts e Vens (2018), para classificação multirrótulo estaria elegível para ser aplicado.

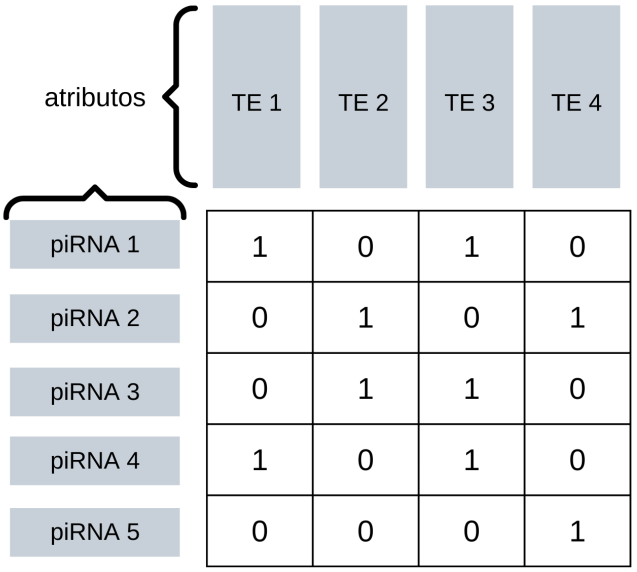

Figura 1. Matriz binária de interação entre piRNAs e TEs, representados por seus respectivos atributos. Adaptado de Pliakos, Geurts e Vens (2018).

Não obstante, de acordo com Pliakos e Vens (2020), a construção de modelos de aprendizagem por meio de árvores do tipo *ensemble* (isto é, a combinação de diversas árvores em um único classificador), quando elaborada a partir da reconstrução do espaço de saída, leva a melhores resultados de predição.

Dessa forma, os autores propõem o método BICTR (*Bi-Clustering Trees with output space Reconstruction* - árvores de *bi-clustering* com reconstrução do espaço de saída) para a aprendizagem. A reconstrução do espaço de saída é feita com a NRLMF (*Neighborhood Regularized Logistic Matrix Factorization* - fatorização logística de matriz regularizada por vizinhança).

3.3 Do NRLMF

Segundo Liu *et al.* (2016), a matriz de interação $Y \in \Re^{|D|\times|P|}$ é utilizada como entrada na predição de interação droga-alvo, em que, no caso do presente estudo, os pares D (drogas) são os piRNAs e os pares P (proteína) são os TEs. A matriz de entrada é utilizada para calcular duas matrizes, sendo $U \in \Re^{|D|\times k}$ e $V \in \Re^{|P|\times k}$, de tal forma que as matrizes U e V são consideradas representações latentes k-dimensionais dos piRNAs e dos TEs. O produto UV^T deve ser aproximadamente Y.

Os pares D e P, matrizes de entrada para o cálculo de U e V, são matrizes de similaridade, em que a matriz de similaridade de uma droga é definida por uma matriz real com valores entre 0 e 1. Assim, a pontuação normalizada de Smith-Waterman foi calculada entre as sequências, tanto de piRNAs quanto de TEs. A pontuação normalizada de Smith-Waterman é dada por $S_{sequência}(m, m') = \frac{SW(m,m')}{\sqrt{SW(m,m)SW(m',m')}}$ para as sequência *m* e *m'*, em que $SW(m, m')$ representa a pontuação de Smith-Waterman (BAN; OHUE; AKIYAMA, 2019). O resultado são duas matrizes quadradas (o que permite o cálculo do determinante na construção de U e V), uma para as sequências de piRNAs e outra para as sequências de TEs, exemplificada na figura 2 para os elementos transponíveis .

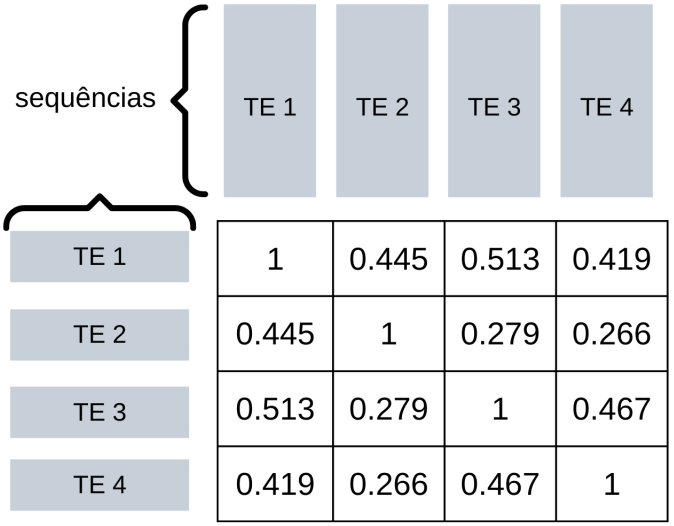

Figura 2. Matriz de similaridade das sequências dos TEs.

A probabilidade p_{ij}, isto é, a probabilidade que o piRNA d_i interaja com o TE p_j, é modelada conforme:

$$p_{ij} = \frac{exp(u_i v_j^T)}{1 + exp(u_i v_j^T)}, \quad (1)$$

em que u_i e v_j são vetores k-dimensionais e representações latentes de, respectivamente, d_i e p_j. Após o modelo de predição e regularização, uma função objetivo é obtida, levando em conta os hiperparâmetros de regularização e pesagem das observações no processo de otimização. Para compreensão integral do método e a derivação completa das equações, consultar Liu *et al.* (2016).

A reconstrução do espaço de saída é relevante, já que na tarefa de predição de interação droga-alvo (ou piRNA-TE, no presente caso), não há pares de interação verdadeiramente negativos (representados por 0): há interações positivas conhecidas (representadas por 1) e interações não rotuladas (já que simplesmente podem não ter sido observadas *ainda* em laboratório). Nesse sentido, essa configuração é denominada de aprendizado *Positive-Unlabeled* (ou positivo - não rotulado). A reconstrução, portanto, visa aliviar o desbalanceamento das classes (já que há muito mais não rotulados do que positivos), uma vez que os valores 0 não representam, de fato, interações negativas (PLIAKOS; VENS, 2020).

3.4 Do NRLMFβ

Conforme notam Ban, Ohue e Akiyama (2019), a predição com NRLMF não é precisa quando se há pouca informação sobre os pares de interações. Dessa forma, propõem o NRLMFβ (*Beta-distribution-rescored Neighborhood Regularized Logistic Matrix Factorization*, isto é, NRLMF com nova pontuação por distribuição beta). Essa nova pontuação se vale do fato de que o NRLMF é fundamentado em um modelo estatístico baseado na distribuição de Bernoulli, e de que a distribuição beta é a *priori* conjugada da distribuição de Bernoulli, de forma que a última pode refletir a quantidade de informação de interações para sua forma com base na inferência Bayesiana. A pontuação s_{ij} (eq. 2) do NRLMF é calculada conforme:

$$s_{ij} = \frac{exp(\tilde{u}_i \tilde{v}_j^T)}{1 + exp(\tilde{u}_i \tilde{v}_j^T)}, \quad (2)$$

a qual é utilizada para gerar a nova matriz, se torna muito baixa quando há pouca informação sobre a interação. A distribuição beta, descrita por:

$$Beta(x|a_{ij}, b_{ij}) = \frac{x^{a_{ij}-1}(1-x)^{b_{ij}-1}}{B(a_{ij}, b_{ij})} \quad (3)$$

corrige esse problema conforme a expressão relacional:

$$s'_{ij} = \frac{a_{ij}}{a_{ij}+b_{ij}} = \frac{s_{ij} \cdot (y_{ij} \cdot \eta_1 + \eta_2 - 2) + 1}{y_{ij} \cdot \eta_1 + \eta_2} \quad (4)$$

Assim, a reconstrução do espaço de saída, no seguinte estudo, foi feita com NRLFMβ, a fim de melhorar a acurácia de predição. Os hiperparâmetros η_1 e η_1 (constantes da equação de concentração da distribuição beta e que dizem respeito ao formato da distribuição), expressos na equação 4, bem como outros hiperparâmetros omitidos no texto, mas declarados na derivação completa que pode ser verificada em Liu *et al.* (2016) e Ban, Ohue e Akiyama (2019), foram definidos manualmente de acordo com a sugestão dos autores.

Com a reconstrução, uma nova matriz de interações é gerada, denotada **Ŷ**. O algoritmo recebe o espaço de atributos $\mathbf{X_d}$ dos piRNAs, o espaço de atributos $\mathbf{X_p}$ dos TEs e a matriz de interação original **Y**. Cada novo item da nova matriz, de tal forma que $\hat{y}_{ij} \in \hat{Y}$, é dado por:

$$\hat{y}_{ij} = \begin{cases} 1, & se\ y_{ij} = 1 \\ s'_{ij}, & caso\ contrário. \end{cases} \quad (5)$$

Com a matriz recalculada, o algoritmo PBCT foi aplicado, utilizando-se dos atributos gerados por *Pse-in-One* dos piRNAs (X1), dos atributos gerados por *Pse-in-One* dos TEs (X2), e da nova matriz de interações (Ŷ) calculada por NRLMFβ. O PBCT aplicado a essa nova matriz é o método denominado previamente de BICTR (Figura 3). O algoritmo de PBCT foi aplicado utilizando-se, também, da matriz de interações original Y, a fim de comparação de performance. Em contraste ao BICTR, o PBCT aplicado à matriz original foi denominado BICT (*Bi-Clustering Trees* - Árvores de *bi-clustering*).

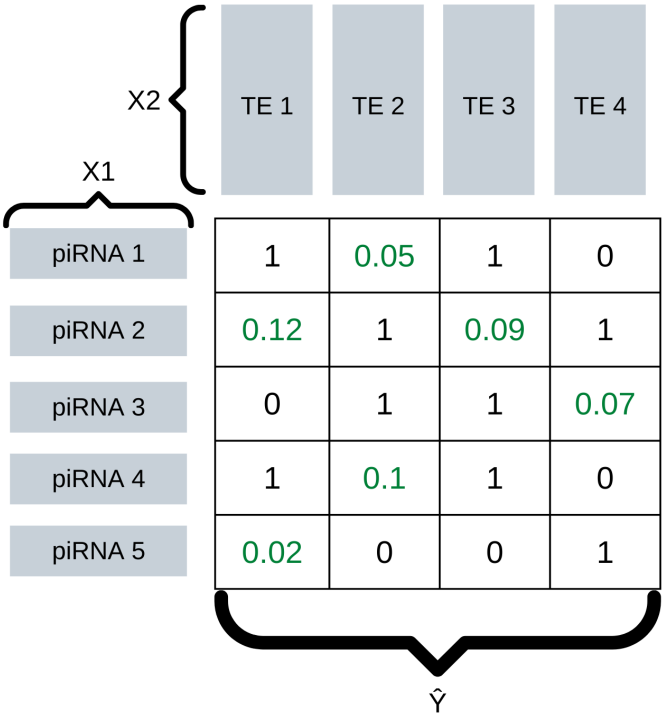

Figura 3. Dados de entrada para o algoritmo de PBCT na metodologia BICTR. Os dados sem a repopulação, isto é, os dados originais, são os dados de entrada para a metodologia BICT.

3.5 Da predição

A predição de interação, nesse caso, é uma inferência de rede DTI (*drug-target interaction*, ou interação de droga-alvo). A tarefa é do tipo classificação em pares de nós, o que significa que é uma aplicação de aprendizado supervisionado. O objetivo é obter a probabilidade de interação entre 2 pares de nodos. Um modelo de aprendizado é construído sobre um conjunto de treinamento de pares droga-alvo (ou, nesse caso, piRNA-TE). Uma vez completado o processo de aprendizado, o modelo pode realizar predição para pares de interação desconhecidos (PLIAKOS; VENS, 2020).

A estrutura de predição de uma rede de interação piRNA-TE é ilustrada na Figura 4. Os itens com *r* se referem às linhas (piRNAS), e os itens com *c* se referem às colunas (TEs). Já os itens com *T* se referem aos conjuntos de teste e os itens com *L* se referem aos conjuntos de treino. A maior porção da matriz de interação é submetida a treinamento na configuração piRNAs de treino - TEs de treino (Lr × Lc). Já as tarefas de predição DTI, segundo Pliakos e Vens (2020), são divididas em 3 configurações:

- piRNAs de teste - TEs de treino (Tr × Lc): interações entre TEs incluídos no procedimento de aprendizado e piRNAs desconhecidos
- piRNAs de treino - TEs de teste (Lr × Tc): interações entre piRNAs incluídos no procedimento de aprendizado e TEs desconhecidos
- piRNAs de teste - TEs de teste (Tr × Tc): interações entre piRNAs desconhecidos e TEs desconhecidos. Tal configuração é, naturalmente, a mais árdua de se obter.

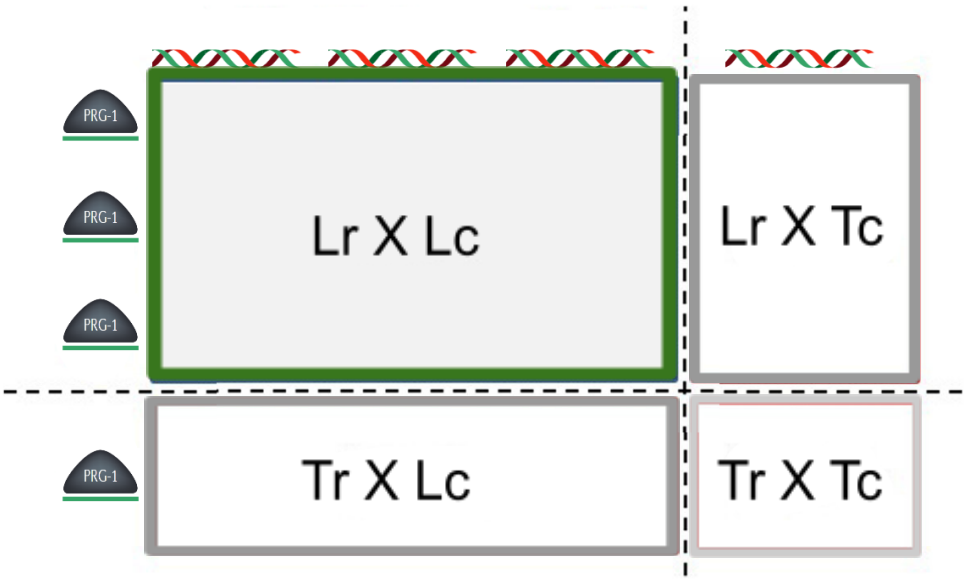

Figura 4. Estrutura de predição de interação. Adaptado de Pliakos e Vens (2020).

A abordagem global de múltiplas saídas, inerente ao método de *bi-clustering*, proposta por Pliakos, Geurts e Vens (2018) utiliza os conjuntos de atributos Xr e Xc (atributos das linhas e atributos das colunas), de forma que o objetivo é predizer interações entre esses conjuntos. O processo se dá pela construção de uma árvore de decisão que incorpora esses espaços de atributos, de modo que cada nó na árvore contenha instâncias que pertençam a ambos conjuntos de interações. A matriz de interações Y é particionada vertical e horizontalmente. Todos os atributos de ambos os conjuntos de instâncias são considerados candidatos à divisão a cada novo corte. A eleição do melhor corte se dá pela redução da impureza máxima (variância) de Y. Tendo que ϕ_r corresponde aos atributos de Xr, isto é, dos piRNAs, e que ϕ_c corresponde aos atributos de Xc, isto é, dos TEs, a redução da impureza é acumulada da redução de impureza das linha e da coluna. Caso o teste de divisão esteja nos atributos dos piRNAs, ou seja, em ϕ_r, a redução é calculada por $Var = \Sigma_j^M Var(Y_j)$, sendo M o número de colunas. Caso esteja em em ϕ_c, a redução é calculada por $Var = \Sigma_i^N Var(Y_i^T)$, sendo N o número de linhas e Y^T a transposta da matriz Y. O processo é ilustrado na Figura 5.

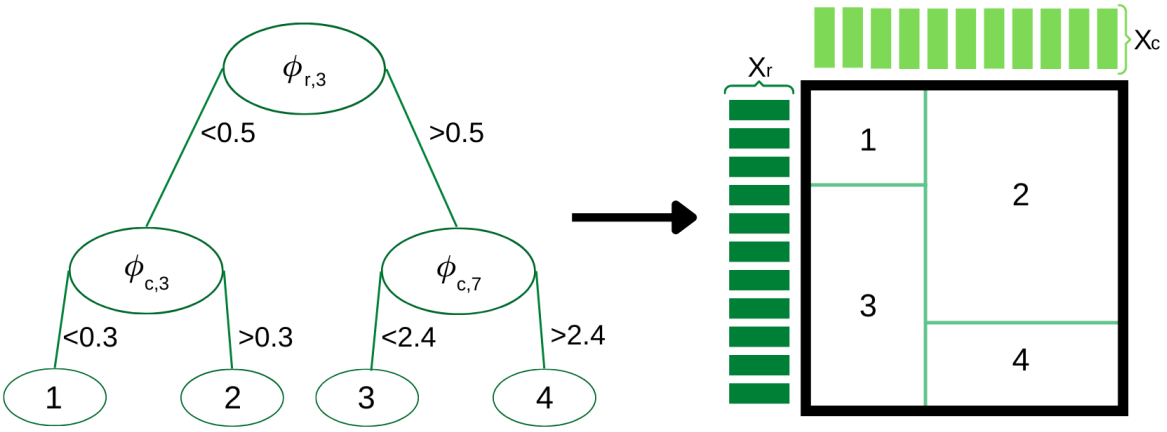

Figura 5. Exemplo do funcionamento da árvore de *bi-clustering*. À esquerda, a árvore de decisão com ϕ_r e ϕ_c, correspondentes aos os atributos das instâncias da linha e da coluna, respectivamente. À direita, a matriz de interação correspondente particionada pela mesma árvore. Adaptado de Pliakos, Geurts e Vens (2018), Pliakos e Vens (2019) e Pliakos e Vens (2020).

3.6 Da validação cruzada

De acordo com Refaeilzadeh, Tang e Liu (2016), validação cruzada (*cross-validation*) é um método estatístico de avaliação de algoritmos de aprendizagem. O método separa os dados em treino e teste em rodadas sucessivas, de tal maneira que cada ponto de dados tenha uma chance de ser validado (isto é, um dado deve compor o conjunto de treino pelo menos uma vez, e compor o conjunto de teste pelo menos uma vez). O método empregado neste estudo é a validação cruzada k-fold. Nesse método, os dados são divididos em k segmentos de tamanho igual, de modo que, a cada iteração, um segmento diferente seja utilizado para teste, enquanto o restante k-1 é utilizado para aprendizado (Figura 6). Em aprendizado de máquina, a validação cruzada com 10 repetições (ou seja, com k = 10) é a usual.

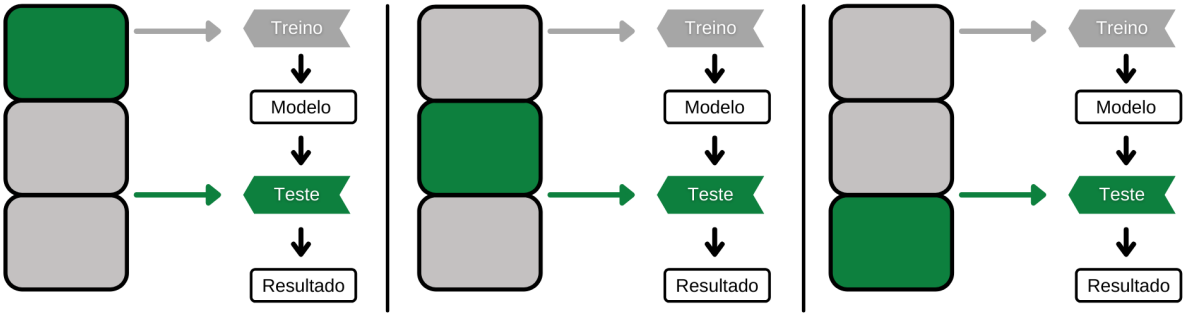

Figura 6. Exemplo de validação cruzada para k = 3. Adaptado de Refaeilzadeh, Tang e Liu (2016).

Contudo, considerando a abordagem de Pliakos, Geurts e Vens (2018), em que há, na verdade, 1 conjunto de treino e 3 conjuntos de teste, a validação cruzada gerou k estruturas de predição de interação (Figura 4) distintas, isto é, k conjuntos de treino e 3k conjuntos de teste. Além disso, os autores também recomendam a utilização de k = 5, já que, dada a esparsidade dos dados, a validação cruzada 10-fold pode ser um problema na configuração Tr × Tc, dada a probabilidade de haver segmentos com apenas zeros.

3.7 Das métricas de avaliação

Ao implementar um modelo de predição, os resultados são quantificados por métricas de avaliação que tornem as performances de quaisquer modelos comparáveis entre si. Segundo Baldi *et al*. (2000), a acurácia de algoritmos de predição em classificação, como é o caso do presente estudo, pode ser obtida de diferentes formas. Um dos métodos canônicos é o estabelecido por Rand (1971), denominado índice de Rand, mas chamado simplesmente de acurácia devido a seu uso ubíquo. O índice, no âmbito de decisões de um algoritmo, pode ser dado por $\frac{VP + VN}{VP + FP + VN + FN}$, em que VP são os verdadeiros positivos, VN são os verdadeiros negativos, FP os falsos positivos e FN os falsos negativos, isto é, o índice informa o percentual de acertos pelo total.

Todavia, em aprendizagem desbalanceada, como é o caso do presente estudo (0.000377% são interações e 0,999623% são não interações), a acurácia tende a ser em torno de 99%, já que os algoritmos tendem a classificar conforme a

classe majoritária, que é 0 nesse conjunto de dados. Dessa forma, é um método pouco informativo para avaliar a capacidade de discernimento do algoritmo BROWNLEE (2020). Assim, alternativas mais informativas foram consideradas para avaliar o modelo.

Conforme DAVIS e GOADRICH (2006), a curva ROC (*Receiver Operating Characteristic,* ou Característica de Operação do Receptor) é uma alternativa usual nos últimos anos, curva esta que mostra como o número de exemplos corretamente classificados como positivos varia com o número de exemplos incorretamente classificados como positivos. A taxa de verdadeiros positivos é dada por $\frac{VP}{VP + FN}$ e a taxa de falsos positivos é dada por $\frac{FP}{FP + VN}$. Porém, em casos de conjuntos de dados extremamente desbalanceados, a curva ROC pode apresentar uma visão mais otimista do que o algoritmo realmente é, já que uma grande mudança na quantidade de falsos positivos surte pouco efeito na taxa de falsos positivos.

Dadas as quantidades altamente desiguais de interações positivas e interações negativas, o uso da curva PR (*Precision-Recall*, ou Precisão-Revocação) é a métrica mais indicada. A revocação é dada por $\frac{VP}{VP + FN}$ (ou seja, é idêntica à taxa de verdadeiros positivos) e a precisão é dada por $\frac{VP}{VP + FP}$. A precisão, portanto, representa melhor as mudanças na quantidade de falsos positivos, já que compara falsos positivos com verdadeiros positivos, ao invés de verdadeiros negativos (DAVIS; GOADRICH, 2006).

Embora as curvas ROC e PR sejam boas representações visuais do comportamento do algoritmo, são a AUROC (*Area Under the Receiver Operating Characteristic Curve*, isto é, Área Sob a Curva Característica de Operação do Receptor) e a AUPRC (*Area Under the Precision-Recall Curve*, isto é, Área Sob a Curva de Precisão e Revocação) que, efetivamente, informam numericamente o desempenho do modelo. O cálculo da área se dá por integração numérica das áreas trapezoidais (regra dos trapézios) geradas por cada ponto da curva ROC ou da curva PR (DAVIS; GOADRICH, 2006).

3.8 Da matriz de confusão

Segundo Brownlee (2020), muitos algoritmos de aprendizado de máquina retornam probabilidades, e não classes, como no caso deste estudo. Nesse sentido, os valores não binários contínuos devem ser convertidos em valores binários (discretos) com base em limiares de decisão, de sorte que que todas as probabilidades iguais ou maiores que o limiar são convertidos a 1, e as menores que o limiar convertidos a 0. Para curva ROC, pode-se encontrar o melhor limiar por meio de *G-mean* (média geométrica). Para tal, utiliza-se a sensibilidade e especificidade do algoritmo. A sensibilidade é igual à taxa de verdadeiros positivos, conceituada no tópico anterior, e a especificidade é o elemento simétrico da taxa de falsos positivos ($1 - TFP$), também conceituada no tópico anterior. A média geométrica é dada por:

$$G - mean = \sqrt{Sensibilidade \times Especificidade}. \quad (6)$$

Já para curva PR, pode-se encontrar o melhor limiar pela *F-measure* (medida F), dada por:

$$F - measure = \frac{2 \times Precisão \times Revocação}{Precisão + Revocação}. \quad (7)$$

Os limiares com maiores valores de cada medida são os eleitos melhores limiares de decisão para o conjunto de dados em questão. Para cada limiar, pode-se construir matrizes de confusão. Uma matriz de confusão é meramente uma tabela que permite verificar quantitativamente os erros e acertos do algoritmo, comparando o valor predito com o valor real. Entretanto, visto que o número de limiares testados é muito grande, apenas a matriz de confusão para os melhores limiares foi calculada.

4. RESULTADOS

1.1 Da curva ROC

As AUROC e AUPRC de todos os *folds* de todos os testes foram obtidas. A Tabela 1 exibe os resultados da AUROC obtidos em validação cruzada de 5-*folds*. Já a Tabela 2 exibe os resultados da AUROC obtido em validação cruzada de 10-*folds*. A Curva ROC do método BICTR para um dos *folds* de Tr × Tc está representada no Gráfico 1. A curva ROC do método BICT para o mesmo fold está representada no Gráfico 2. As matrizes de confusão para os melhores limiares desse mesmo fold, tanto para o BICTR quanto para o BICT, estão representadas no Gráfico 3. O melhor limiar foi definido de acordo com o maior valor de *G-mean*.

Tabela 1. Resultados AUROC obtidos em validação cruzada de 5-*folds*

Dados	**BICTR**	**BICT**
	Tr x Lc	
Fold 1	0,468	0,539
Fold 2	0,474	0,532
Fold 3	0,454	0,535
Fold 4	0,456	0,533
Fold 5	0,467	0,521
Média	0,464	0,532
	Lr x Tc	
Fold 1	0,437	0,550
Fold 2	0,448	0,535
Fold 3	0,424	0,543
Fold 4	0,474	0,531
Fold 5	0,470	0,539
Média	0,451	0,540
	Tr x Tc	
Fold 1	0,476	0,543
Fold 2	0,478	0,511
Fold 3	0,442	0,516
Fold 4	0,479	0,517
Fold 5	0,470	0,515
Média	0,469	0,520

Tabela 2. Resultados AUROC obtidos em validação cruzada de 10-*folds*

Dados	**BICTR**	**BICT**
	Tr x Lc	
Fold 1	0,452	0,522
Fold 2	0,472	0,549
Fold 3	0,475	0,522
Fold 4	0,453	0,531
Fold 5	0,476	0,532
Fold 6	0,487	0,511
Fold 7	0,459	0,545
Fold 8	0,460	0,537
Fold 9	0,447	0,522
Fold 10	0,458	0,550
Média	0,464	0,532
	Lr x Tc	
Fold 1	0,460	0,528
Fold 2	0,419	0,551
Fold 3	0,422	0,548
Fold 4	0,446	0,522
Fold 5	0,463	0,539
Fold 6	0,478	0,542
Fold 7	0,427	0,556
Fold 8	0,500	0,551
Fold 9	0,456	0,537
Fold 10	0,453	0,531
Média	0,453	0,541
	Tr x Tc	
Fold 1	0,458	0,529
Fold 2	0,489	0,548
Fold 3	0,430	0,506
Fold 4	0,474	0,500
Fold 5	0,520	0,524
Fold 6	0,514	0,537
Fold 7	0,399	0,505
Fold 8	0,453	0,540
Fold 9	0,410	0,494
Fold 10	0,453	0,538
Média	0,460	0,522

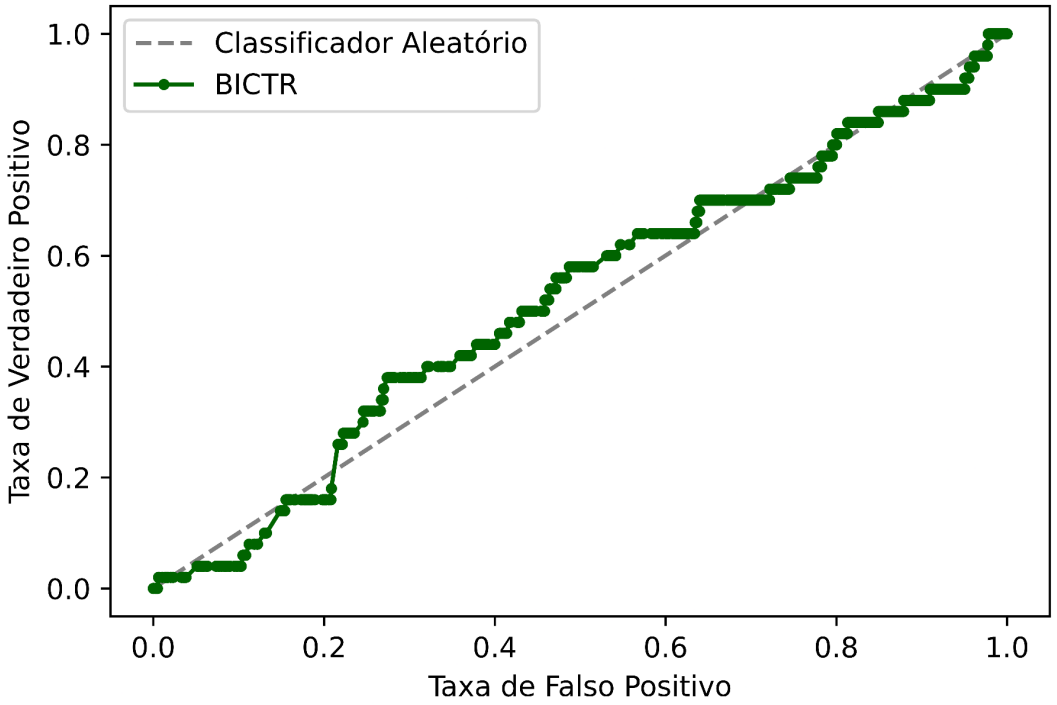

Gráfico 1. Curva ROC do BICTR para o *fold* 5 de Tr × Tc (10-*fold*).

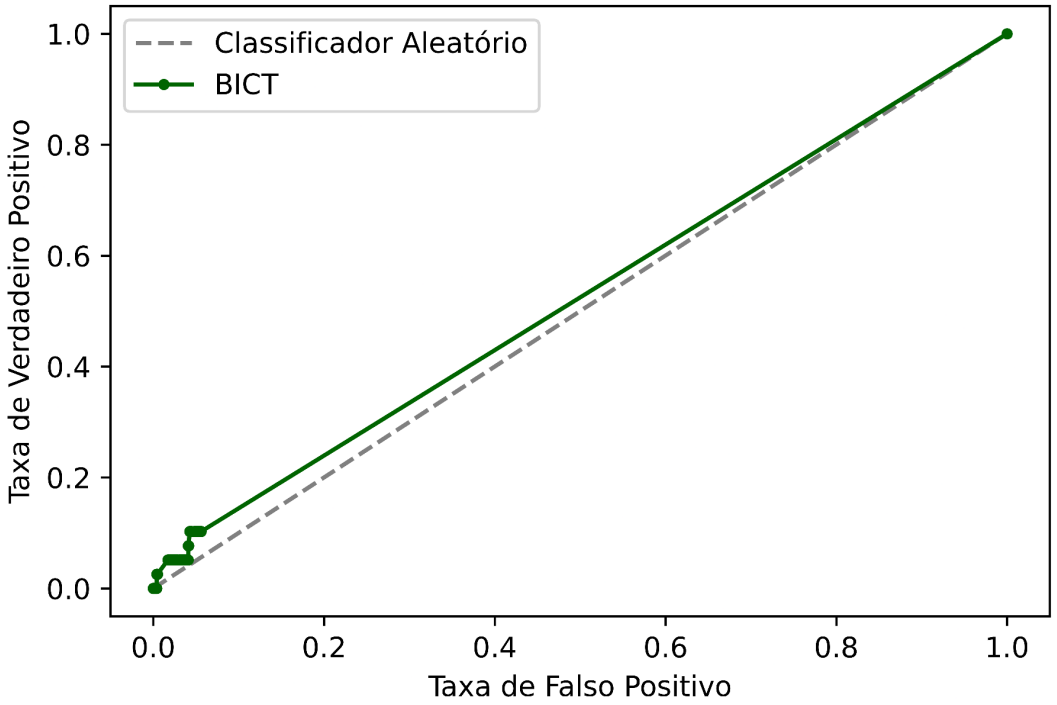

Gráfico 2. Curva ROC do BICT para o *fold* 5 de Tr × Tc (10-*fold*).

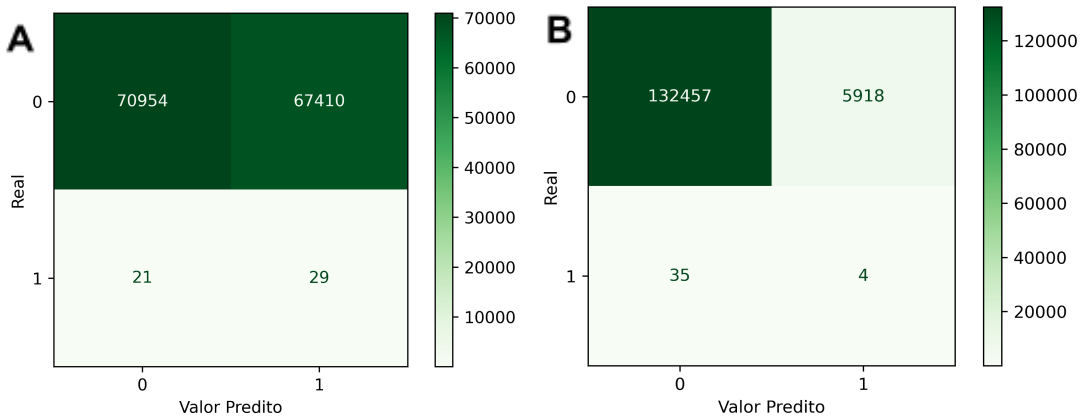

Gráfico 3. Matrizes de confusão para os melhores limiares da curva ROC para o *fold* 5 de Tr × Tc (10-*fold*). **A.** BICTR. **B.** BICT.

1.2 Da curva PR

A Tabela 3 exibe os resultados da AUPRC obtidos em validação cruzada de 5-*folds*. Já a Tabela 4 exibe os resultados da AUPRC obtido em validação cruzada de 10-*folds*. A Curva Precisão e Revocação (PR) do BICTR para um dos *folds* de Tr × Tc está representada no Gráfico 4. A PR do BICT para o mesmo fold está representada no Gráfico 5. As matrizes de confusão para os melhores limiares desse mesmo fold, tanto para o BICTR quanto para o BICT, estão representadas no Gráfico 6. O melhor limiar foi definido de acordo com o maior valor de *F-score*.

Tabela 3. Resultados AUPRC obtidos em validação cruzada 5-*fold*

Dados	**BICTR**	**BICT**
	Tr x Lc	
Fold 1	1,62E-03	7,03E-04
Fold 2	1,61E-03	8,05E-03
Fold 3	3,77E-04	2,37E-03
Fold 4	3,84E-04	1,39E-03
Fold 5	3,27E-04	5,82E-04
Média	8,62E-04	2,62E-03

	Lr x Tc	
Fold 1	3,07E-04	1,03E-03
Fold 2	9,15E-04	2,72E-03
Fold 3	2,85E-04	9,40E-04
Fold 4	1,54E-03	6,87E-04
Fold 5	3,83E-04	3,47E-03
Média	6,86E-04	1,77E-03
	Tr x Tc	
Fold 1	3,54E-04	1,02E-03
Fold 2	2,82E-03	5,23E-04
Fold 3	3,22E-04	5,22E-04
Fold 4	3,47E-04	4,85E-04
Fold 5	2,81E-03	6,01E-04
Média	1,33E-03	6,29E-04

Tabela 4. Resultados AUPRC obtidos em validação cruzada 10-*fold*

Dados	**BICTR**	**BICT**
	Tr x Lc	
Fold 1	3,46E-04	1,70E-03
Fold 2	3,32E-04	1,51E-03
Fold 3	3,72E-04	2,90E-03
Fold 4	3,41E-04	9,64E-04
Fold 5	3,60E-04	2,87E-03
Fold 6	3,85E-04	5,13E-04
Fold 7	1,46E-03	1,03E-03
Fold 8	3,41E-04	1,95E-03
Fold 9	3,07E-04	5,46E-04
Fold 10	3,54E-03	4,20E-03
Média	7,78E-04	1,82E-03
	Lr x Tc	
Fold 1	3,32E-04	7,45E-04
Fold 2	1,52E-03	8,72E-04
Fold 3	2,98E-04	1,15E-03
Fold 4	1,29E-03	6,14E-04
Fold 5	3,03E-04	7,51E-04
Fold 6	3,55E-04	1,01E-03
Fold 7	2,72E-04	2,53E-03

Fold 8	1,96E-03	3,64E-03
Fold 9	3,06E-04	2,65E-03
Fold 10	1,31E-03	7,33E-04
Média	7,95E-04	1,47E-03
	Tr x Tc	
Fold 1	3,55E-04	1,46E-03
Fold 2	3,67E-04	2,66E-03
Fold 3	4,28E-04	3,55E-04
Fold 4	4,63E-04	4,43E-04
Fold 5	3,77E-04	4,36E-04
Fold 6	3,58E-04	1,08E-03
Fold 7	3,51E-04	4,55E-04
Fold 8	6,28E-04	6,99E-04
Fold 9	2,93E-04	4,48E-04
Fold 10	4,19E-04	9,61E-04
Média	4,04E-04	9,00E-04

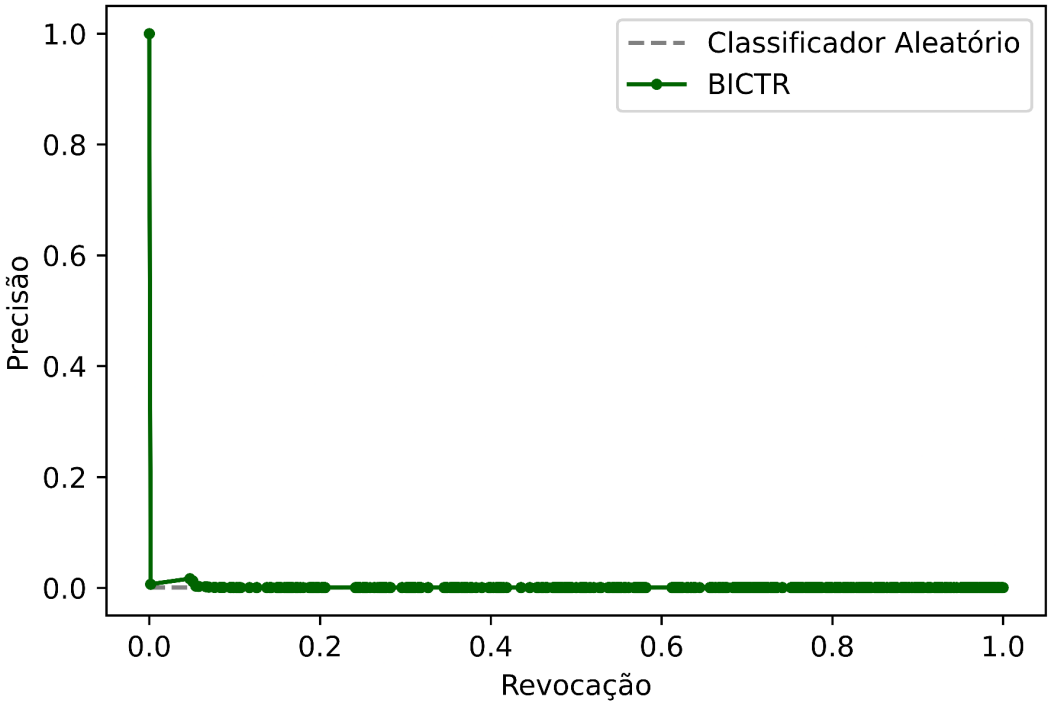

Gráfico 4. Curva de Precisão e Revocação do BICTR para o *fold* 8 de Lr × Tc (10-*fold*).

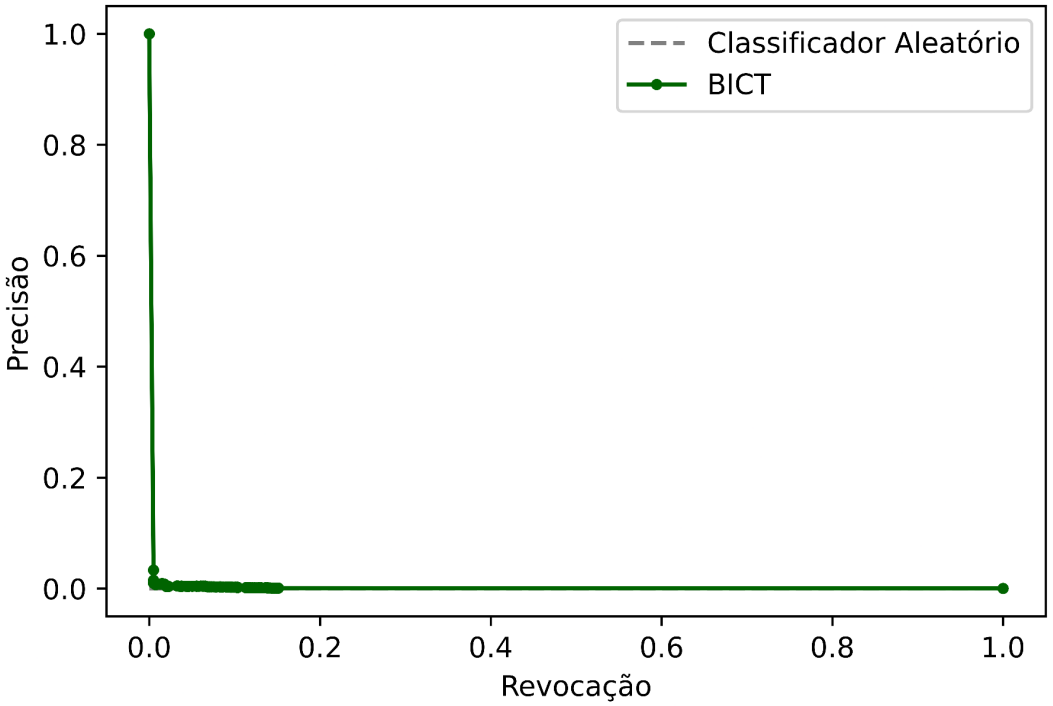

Gráfico 5. Curva de Precisão e Revocação do BICT para o *fold* 8 de Lr × Tc (10-*fold*).

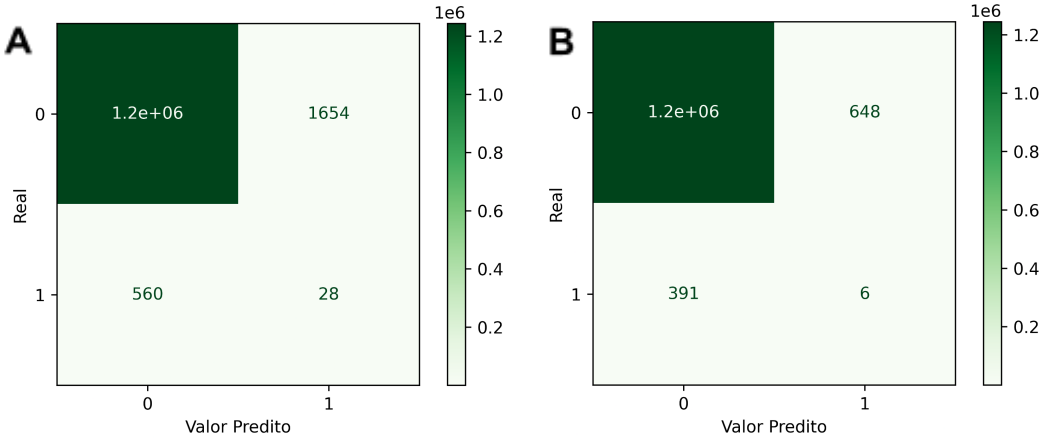

Gráfico 6. Matrizes de confusão para os melhores limiares da curva PR para o *fold* 8 de Lr × Tc (10-*fold*). **A.** BICTR. **B.** BICT.

5. DISCUSSÃO

Tendo-se em mente que, tanto para AUROC e AUPRC, tanto melhor é o modelo quanto mais próximas de 1 as áreas, é evidente, pelos valores exibidos nas tabelas, que o algoritmo não foi capaz de predizer de maneira satisfatória as interações verdadeiras, embora tenha sim predito corretamente algumas positivas, como evidenciado nas matrizes de confusão dos gráficos 3 e 6. Observando-se os gráficos ROC, pode-se perceber que o comportamento da curva traçada pelos pontos se assemelha muito ao comportamento de um classificador aleatório, oscilando em torno da linha que o representa. Observando-se os gráficos PR, é perceptível que a curva traçada também se assemelha à curva de um classificador aleatório. Por meio dessas observações, pode-se argumentar que o modelo, para os dados presentes, não obteve uma boa aprendizagem e se comportou como um classificador aleatório, não sendo capaz de predizer novas interações de tal maneira satisfatória. A explicação para isso se torna mais clara ao investigar as matrizes de confusão. A enorme quantidade de interações negativas torna o problema de predição um problema de classificação extremamente desbalanceado, o que dificulta, em muito, a aprendizagem do modelo.

Comparando-se os métodos BICT e BICTR, pode-se verificar que o BICTR, embora empregado com finalidade de melhorar o desempenho da predição, obteve menores áreas de AUPRC e AUROC. Entretando, uma vez que o interesse nesse problema é a investigação de interações positivas, verifica-se, pelas matrizes de confusão dos gráficos 3 e 6, que BICTR levou a um grande aumento de classificações como 1. Pode-se verificar, pelas matrizes de confusão, que BICT apresentou maiores áreas pelo forte viés de classificação como 0. Uma vez que, de fato, a vasta maioria dos itens são 0 (aproximadamente 0,999623%), o método BICT classificou corretamente muitos itens como 0, enquanto o BICTR classificou mais itens como 1. Em ambas as matrizes de confusão esse comportamento é observado, ainda que para a curva ROC isso seja particularmente notável, já que, para o melhor limiar do BICTR, houveram 61.492 falsos positivos a mais que para o melhor limiar de BICT.

É também evidente que a curva de BICTR, tanto em ROC quanto PR, é bem melhor distribuída pelo plano do que a curva de BICT. Isso ocorre porque os limiares de conversão são definidos de acordo com os próprios valores presentes na matriz

predita, ou seja, quanto mais populada a matriz predita, mais limiares de conversão e, consequentemente, mais pontos de precisão e revocação. Uma vez que, na metodologia BICT, o PBCT recebe os dados originais, que contêm muitos valores 0, muitos valores 0 são preditos, tornando a quantidade de limiares bastante limitada. Enquanto isso, na metodologia BICTR, o PBCT aprende com dados repopulados, de forma que não são preditos valores 0, tornando a quantidade de limiares excepcionalmente maior e, consequentemente, de pontos traçados nos gráficos.

Com isso em mente, vale notar que, embora o BICT apresente áreas maiores que as áreas de BICTR na grande maioria dos casos (Tabelas 1-4), a menor quantidade de pontos de BICT faz com que haja maiores ‘saltos’ entre os pontos, o que contribui para uma maior área, mas não é tão fiel quanto a curva traçada pelo BICTR, que é bem mais sutil.

Por mais que o algoritmo PBCT não tenha sido capaz de criar um modelo de predição satisfatório, a busca por algoritmos melhor ajustados a um problema de aprendizado de máquina não cessa. Há muitas formas diferentes de abordar o conjunto de dados desse estudo. Uma delas é pela chamada XMC (*Extreme multi-label classification* - Classificação Multirrótulo Extrema). Segundo Shen *et al.* (2020), XMC se refere à tarefa de encontrar, em um domínio imenso de possíveis rótulos, os rótulos relevantes. Outra forma de abordar esse problema é por métodos de *sampling* (amostragem) dos dados, tanto pelo *oversampling* (sobreamostragem) como pelo *undersampling* (subamostragem), já que boa parte dos algoritmos de classificação são construídos de modo a operar melhor em dados com números iguais de observações para cada classe (BROWNLEE, 2020). Um exemplo clássico de *oversampling* é o SMOTE (*Synthetic Minority Oversampling Technique* - Técnica de Sobreamostragem Minoritária Sintética), que utiliza dados sintéticos de uma vizinhança definida, dados esses criados pela interpolação de várias instâncias da classe minoritária, de sorte que que o algoritmo, ao invés de utilizar os pontos de dados originais, utiliza atributos dos dados e seus relacionamentos (FERNANDEZ *et al*., 2018). Considerando o *undersampling*, um clássico exemplo é o CNN (*Condensed Nearest Neighbor Rule* - Regra do Vizinho Mais Próximo Condensado), que classifica padrões no conjunto de treinamento, incluindo os padrões classificados incorretamente em um conjunto condensado, num processo iterativo que elimina todos os padrões mal classificados, tornando esse condensado final em um subconjunto do conjunto de treino original (KUMAR; VISWANATH; BINDU,

2017). Por fim, uma abordagem mais simples é o mero ajuste dos hiperparâmetros utilizados pelo NRLMFβ e o teste com outros extratores de atributos de sequências.

Considerado o exposto acima, é evidente que há um universo de formas para se explorar esses dados e construir modelos com bom desempenho, bem como não faltam motivos para prosseguir na investigação. Experimentos de descoberta de interações entre biomoléculas *in vitro* são bastante onerosos e laboriosos. Assim, uma vantagem óbvia é justamente a redução de trabalho e custo, assim como já acontece em pesquisas de descoberta de drogas. Não somente, a identificação *in vitro* está sujeita a contaminação, o que pode dificultar o reconhecimento dos padrões de ligação, os quais podem ser ocultos e de difícil caracterização, mas podem ser reconhecidos por modelos eficientes de aprendizado de máquina. Por fim, a caracterização e modelagem dessas interações contribui para a construção de conhecimento em biologia de sistemas e para todos os seus consequentes possíveis usos (LARRAÑAGA et al., 2006).

Como já descrito previamente, mas especialmente verificável em Ozata *et al.* (2019), os piRNAs interagem com transposons contribuindo em uma importante função: silenciá-los e, assim, defender o genoma da linhagem germinativa. Dessa forma, um vasto horizonte de aplicações embriológicas pode vir a ser explorado como aplicação biológica desse tipo de trabalho de predição de interações dessas biomoléculas. Em âmbito acadêmico, o conhecimento das estratégias de interação entre piRNAs e TEs nas diversas espécies pode trazer noções importantes sobre a história da evolução molecular, assim como já foram verificadas as relações ancestrais em Panarthropoda (LEWIS *et al.,* 2018). Já no âmbito mercadológico, uma possível aplicação é na manipulação genômica de linhagens animais, já que muitas malformações poderiam estar atribuídas a problemas na defesa contra transposons já que, ainda segundo Ozata *et al.* (2019), a transposição pode promover eventos recombinatórios danosos e interferências na expressão. Assim, o conhecimento preciso das interações em espécies de interesse comercial pode levar a métodos que ajudem a reduzir as chances de um evento de transposição acontecer em uma linhagem de interesse.

Entre as possibilidades há também uma, certamente mais distante, está nas intervenções farmacêuticas, já que, uma vez identificadas as transposições mais comumente associadas a determinadas doenças, vias de prevenção poderiam ser traçada, considerando que a terapia por meio de RNAi, sobretudo na terapia de

câncer, já é um assunto debatido e estudado amplamente pelos cientistas (WANG *et al.*, 2011). Não somente a identificação dos problemas causados pelas transposições, mas também a identificação dos problemas causados pelos piRNAs, particularmente em células cancerígenas, uma vez que, conforme Cheng *et al.* (2011), vias anormais de pIRNAs aumentam repetições de retrotransposons, os quais são componentes dos telômeros, além de que alguns piRNAs podem estar aberrantemente expressos em diversos tipos de células cancerígenas.

6. CONCLUSÕES

O estudo das moléculas biológicas é uma área extremamente excitante e em grande fervor, já que a compreensão das mesmas tem contribuído fortemente na busca humana por mais conforto e qualidade de vida. A biologia molecular vem fornecendo a fundação para o desenvolvimento acelerado da terapêutica, agricultura, pecuária e demais áreas de relevância suma ao ser humano que envolvam seres vivos. Não somente, a mesma tem trazido generosas contribuições para o entendimento da História Natural, não só nossa, como também de toda a teia de organismos que nos cercam. Contudo, a experimentação em laboratório, sozinha, pode não ser capaz de lidar com a complexidade das interações biológicas, já que as biomoléculas estão inseridas em um microuniverso de outras muitas biomoléculas, e as interações entre elas, que muitas vezes são extremamente informativas e são alvo majoritário de estudo pela biologia de sistemas, são difíceis de serem mapeadas *in vitro*. Soma-se a isso a grande quantidade de dados disponíveis na *web*, quantidade essa impossível de ser processada por um indivíduo. Nesse contexto, a bioinformática surge como meio de ajudar a sanar esse problema. Não somente, a predição de interação por aprendizado de máquina se sobressai para ajudar a sanar o problema específico da interação de biomoléculas. Dado o exposto, a relevância deste trabalho foi contribuir com a pesquisa dos métodos mais adequados para a predição *in silico* de interações entre biomoléculas, nomeadamente na conjuntura dos RNAs de interferência, os quais constituem a grande aposta de pesquisa em biologia molecular nos últimos anos.

Pode-se concluir, portanto, que a predição de interações entre piRNAs e TEs demonstra diversas potenciais aplicações biológicas, ainda a serem exploradas pelo mercado e pela academia. Conclui-se também que o algoritmo proposto não obteve um bom resultado no conjunto de dados proposto, comportando-se como um classificador aleatório, embora tenham sido levados em conta os fatores que levaram à baixa performance, sobretudo o viés extremo de interações negativas no conjunto de treino. Apesar disso, foram também propostas medidas futuras de mediação desse problema, abrindo-se um cenário fértil de investigações vindouras neste estimulante campo de investigação científica.

7. REFERÊNCIAS

BALDI, P.; BRUNAK, S.; CHAUVIN, Y.; ANDERSEN, C. A. F.; NIELSEN, H. Assessing the accuracy of prediction algorithms for classification: An overview. **Bioinformatics**, v. 16, n. 5, p. 412–424, 2000.

BAN, T.; OHUE, M.; AKIYAMA, Y. NRLMFβ: Beta-distribution-rescored neighborhood regularized logistic matrix factorization for improving the performance of drug–target interaction prediction. **Biochemistry and Biophysics Reports**, v. 18, n. January, p. 100615, 2019.

BESSEREAU, J.-L. Transposons in C. elegans. **WormBook**, p. 1–13, 2006.

BRENNECKE, J.; ARAVIN, A. A.; STARK, A.; DUS, M.; KELLIS, M.; SACHIDANANDAM, R.; HANNON, G. J. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. **Cell**, v. 128, n. 6, p. 1089–1103, 2007.

BROWNLEE, J. **Imbalanced classification with python: Better metrics, balance skewed classes, cost-sensitive learning.** In:_____. Probability Threshold Moving. Machine Learning Mastery, p. 245-262, 2020.

BROWNLEE, J. **Imbalanced classification with python: Better metrics, balance skewed classes, cost-sensitive learning.** In:_____. The Failure of Accuracy. Machine Learning Mastery, p. 48-56, 2020.

BROWNLEE, J. **Imbalanced classification with python: Better metrics, balance skewed classes, cost-sensitive learning.** In:_____. Tour of Data Sampling Methods. Machine Learning Mastery, p. 104-111, 2020.

CHEN, W.; FENG, P.-M.; LIN, H.; CHOU, K.-C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. **Nucleic Acids Research**, v. 41, n. 6, p. e68–e68, 2013.

CHENG, J.; GUO, J.-M.; XIAO, B.-X.; MIAO, Y.; JIANG, Z.; ZHOU, H.; LI, Q.-N. piRNA, the new non-coding RNA, is aberrantly expressed in human cancer cells. **Clinica Chimica Acta**, v. 412, n. 17–18, p. 1621–1625, 2011.

CRICK, F. Central Dogma of Molecular Biology. **Nature**, v. 227, n. 5258, p. 561–563, 1970.

DAVIS, J.; GOADRICH, M. The relationship between Precision-Recall and ROC curves. In: PROCEEDINGS OF THE 23RD INTERNATIONAL CONFERENCE ON

MACHINE LEARNING - ICML '06 2006, New York, New York, USA. **Anais**... New York, New York, USA: ACM Press, 2006.

FERNANDEZ, A.; GARCIA, S.; HERRERA, F.; CHAWLA, N. V. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. **Journal of Artificial Intelligence Research**, v. 61, p. 863–905, 2018.

FRØKJÆR-JENSEN, C.; JAIN, N.; HANSEN, L.; DAVIS, M. W.; LI, Y.; ZHAO, D.; REBORA, K.; MILLET, J. R. R. M.; LIU, X.; KIM, S. K.; DUPUY, D.; JORGENSEN, E. M.; FIRE, A. Z. An Abundant Class of Non-coding DNA Can Prevent Stochastic Gene Silencing in the C. elegans Germline. **Cell**, v. 166, n. 2, p. 343–357, 2016.

GANKO, E. W.; BHATTACHARJEE, V.; SCHLIEKELMAN, P.; MCDONALD, J. F. Evidence for the Contribution of LTR Retrotransposons to C. elegans Gene Evolution. **Molecular Biology and Evolution**, v. 20, n. 11, p. 1925–1931, 2003.

GILBERT, W. Origin of life: The RNA world. **Nature**, v. 319, n. 6055, p. 618–618, 1986.

KAPITONOV, V. V.; JURKA, J. Rolling-circle transposons in eukaryotes. **Proceedings of the National Academy of Sciences of the United States of America**, v. 98, n. 15, p. 8714–8719, 2001.

KAMINKER, J. S.; BERGMAN, C. M.; KRONMILLER, B.; CARLSON, J.; SVIRSKAS, R.; PATEL, S.; FRISE, E.; WHEELER, D. A.; LEWIS, S. E.; RUBIN, G. M.; ASHBURNER, M.; CELNIKER, S. E. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. **Genome biology**, v. 3, n. 12, 2002.

KINGSFORD, C.; SALZBERG, S. L. What are decision trees? **Nature Biotechnology**, v. 26, n. 9, p. 1011–1013, 2008.

KUDLA, G.; GRANNEMAN, S.; HAHN, D.; BEGGS, J. D.; TOLLERVEY, D. Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 24, p. 10010–10015, 2011.

KUMAR, R. R.; VISWANATH, P.; BINDU, C. S. Nearest neighbor classifiers: a review. **Int. J. Comput. Intell. Res**, v. 13, n. 2, p. 303–311, 2017.

LARRAÑAGA, P.; CALVO, B.; SANTANA, R.; BIELZA, C.; GALDIANO, J.; INZA, I.; LOZANO, J. A.; ARMAÑANZAS, R.; SANTAFÉ, G.; PÉREZ, A.; ROBLES, V. Machine learning in bioinformatics. **Briefings in Bioinformatics**, v. 7, n. 1, p. 86–112, 2006.

LEWIS, S. H.; QUARLES, K. A.; YANG, Y.; TANGUY, M.; FRÉZAL, L.; SMITH, S. A.; SHARMA, P. P.; CORDAUX, R.; GILBERT, C.; GIRAUD, I.; COLLINS, D. H.; ZAMORE, P. D.; MISKA, E. A.; SARKIES, P.; JIGGINS, F. M. Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. **Nature Ecology & Evolution**, v. 2, n. 1, p. 174–181, 2018.

LIU, B.; LIU, F.; WANG, X.; CHEN, J.; FANG, L.; CHOU, K.-C. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. **Nucleic Acids Research**, v. 43, n. W1, p. W65–W71, 2015.

LIU, Y.; WU, M.; MIAO, C.; ZHAO, P.; LI, X. L. Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. **PLoS Computational Biology**, v. 12, n. 2, p. 1–26, 2016.

LUO, S.; LU, J. Silencing of Transposable Elements by piRNAs in Drosophila : An Evolutionary Perspective. **Genomics, Proteomics & Bioinformatics**, v. 15, n. 3, p. 164–176, 2017.

MADEIRA, S. C.; OLIVEIRA, A. L. Biclustering algorithms for biological data analysis: A survey. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v. 1, n. 1, p. 24–45, 2004.

MCCULLERS, T. J.; STEINIGER, M. Transposable elements in Drosophila . **Mobile Genetic Elements**, v. 7, n. 3, p. 1–18, 2017.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997.

MORRIS, K. V.; MATTICK, J. S. The rise of regulatory RNA. **Nature Reviews Genetics**, v. 15, n. 6, p. 423–437, 2014.

OZATA, D. M.; GAINETDINOV, I.; ZOCH, A.; O'CARROLL, D.; ZAMORE, P. D. PIWI-interacting RNAs: small RNAs with big functions. **Nature Reviews Genetics**, v. 20, n. 2, p. 89–108, 2019.

PLASTERK, R. H. A.; IZSVÁK, Z.; IVICS, Z. Resident aliens the Tc1/mariner superfamily of transposable elements. **Trends in Genetics**, v. 15, n. 8, p. 326–332, 1999.

PLIAKOS, K.; GEURTS, P.; VENS, C. Global multi-output decision trees for interaction prediction. **Machine Learning**, v. 107, n. 8–10, p. 1257–1281, 2018.

PLIAKOS, K.; VENS, C. Drug-target interaction prediction with tree-ensemble learning and output space reconstruction. **BMC Bioinformatics**, v. 21, n. 1, p. 1V, 2020.

PLIAKOS, K.; VENS, C. Network inference with ensembles of bi-clustering trees. **BMC Bioinformatics**, v. 20, n. 1, p. 1–12, 2019.

QURESHI, A.; TANTRAY, V. G.; KIRMANI, A. R.; AHANGAR, A. G. A review on current status of antiviral siRNA. **Reviews in Medical Virology**, v. 28, n. 4, p. 1–11, 2018.

RAND, A. **Atlas shrugged**. New York: Signet, 1957.

RAND, W. M. Objective Criteria for the Evaluation of Clustering Methods. **Journal of the American Statistical Association**, v. 66, n. 336, p. 846–850, 1971.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-Validation. In: LIU, L.; ÖZSU, M. T. (Eds.). **Encyclopedia of Database Systems**. New York, NY: Springer New York, 2016. p. 1–7.

ROMERO-CORDOBA, S. L.; SALIDO-GUADARRAMA, I.; RODRIGUEZ-DORANTES, M.; HIDALGO-MIRANDA, A. miRNA biogenesis: Biological impact in the development of cancer. **Cancer Biology and Therapy**, v. 15, n. 11, p. 1444–1455, 2014.

ROUGET, C.; PAPIN, C.; BOUREUX, A.; MEUNIER, A.; FRANCO, B.; ROBINE, N.; LAI, E. C.; PELISSON, A.; SIMONELIG, M. Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. **Nature**, v. 467, n. 7319, p. 1128–1132, 2010.

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, v. 3, n. 3, p. 210–229, 1959.

SHEN, E. Z.; CHEN, H.; OZTURK, A. R.; TU, S.; SHIRAYAMA, M.; TANG, W.; DING, Y. H.; DAI, S. Y.; WENG, Z.; MELLO, C. C. Identification of piRNA Binding Sites Reveals the Argonaute Regulatory Landscape of the C. elegans Germline. **Cell**, v. 172, n. 5, p. 937- 951.e18, 2018.

SHEN, Y.; YU, H.-F.; SANGHAVI, S.; DHILLON, I. Extreme Multi-label Classification from Aggregated Labels. In: (H. D. III, A. Singh, Eds.)PROCEEDINGS OF THE 37TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING 2020, **Anais...** : PMLR, 2020.

SHIRAYAMA, M.; SETH, M.; LEE, H. C.; GU, W.; ISHIDATE, T.; CONTE, D.; MELLO, C. C. PiRNAs initiate an epigenetic memory of nonself RNA in the C. elegans germline. **Cell**, v. 150, n. 1, p. 65–77, 2012.

SILVA, J. C.; KIDWELL, M. G. Horizontal Transfer and Selection in the Evolution of P Elements. **Molecular Biology and Evolution**, v. 17, n. 10, p. 1542–1557, 2000.

SVENDSEN, J. M.; MONTGOMERY, T. A. piRNA Rules of Engagement. **Developmental Cell**, v. 44, n. 6, p. 657–658, 2018.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **Journal of Molecular Biology**, v. 147, n. 1, p. 195–197, 1981.

WANG, J.; ZHANG, P.; LU, Y.; LI, Y.; ZHENG, Y.; KAN, Y.; CHEN, R.; HE, S. PiRBase: A comprehensive database of piRNA sequences. **Nucleic Acids Research**, v. 47, n. D1, p. D175–D180, 2019.

WANG, Z.; RAO, D. D.; SENZER, N.; NEMUNAITIS, J. RNA Interference and Cancer Therapy. **Pharmaceutical Research**, v. 28, n. 12, p. 2983–2995, 2011.

WICKER, T.; SABOT, F.; HUA-VAN, A.; BENNETZEN, J. L.; CAPY, P.; CHALHOUB, B.; FLAVELL, A.; LEROY, P.; MORGANTE, M.; PANAUD, O.; PAUX, E.; SANMIGUEL, P.; SCHULMAN, A. H. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, n. 12, p. 973–982, 2007.

WU, W. S.; BROWN, J. S.; CHEN, T. Te; CHU, Y. H.; HUANG, W. C.; TU, S.; LEE, H. C. PiRTarBase: A database of piRNA targeting sites and their roles in gene regulation. **Nucleic Acids Research**, v. 47, n. D1, p. D181–D187, 2019.

WU, W. S.; HUANG, W. C.; BROWN, J. S.; ZHANG, D.; SONG, X.; CHEN, H.; TU, S.; WENG, Z.; LEE, H. C. PirScan: A webserver to predict piRNA targeting sites and to avoid transgene silencing in C. elegans. **Nucleic Acids Research**, v. 46, n. W1, p. W43–W48, 2018.

WU, J.; YANG, J.; CHO, W. C.; ZHENG, Y. Argonaute proteins: Structural features, functions and emerging roles. **Journal of Advanced Research**, v. 24, p. 317–324, 2020.

ZENG, Y.; YI, R.; CULLEN, B. R. MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. **Proceedings of the National Academy of Sciences of the United States of America**, v. 100, n. 17, p. 9779–9784, 2003.

ZHANG, P.; SI, X.; SKOGERBØ, G.; WANG, J.; CUI, D.; LI, Y.; SUN, X.; LIU, L.; SUN, B.; CHEN, R.; HE, S.; HUANG, D. W. PiRBase: A Web resource assisting piRNA functional study. **Database**, v. 2014, p. 1–7, 2014.

ZHANG, D.; TU, S.; STUBNA, M.; WU, W. S.; HUANG, W. C.; WENG, Z.; LEE, H. C. The piRNA targeting rules and the resistance to piRNA silencing in endogenous genes. **Science**, v. 359, n. 6375, p. 587–592, 2018.