CLEBIANO DA COSTA SÁ

Métodos de validação tradicional e temporal
aplicados à avaliação de classificadores de
RNAs codificantes e não codificantes

São Paulo

2018

CLEBIANO DA COSTA SÁ

Métodos de validação tradicional e temporal aplicados
à avaliação de classificadores de RNAs codificantes e
não codificantes

Texto de Exame de Mestrado apresentado ao
Instituto de Matemática e Estatística como
parte dos requisitos para obtenção do título
de Mestre em Ciências pelo Programa de
Mestrado em Bioinformática.

Área de concentração: Bioinformática -
Interunidades

Orientador: Profa. Dra. Arianne Machado
Lima

Coorientador: Prof. Dr. Marcelo de Souza
Lauretto

São Paulo

2018

FICHA CATALOGRÁFICA

S111 Sá, Clebiano da Costa
Métodos de validação tradicional e temporal aplicados à avaliação de classificadores de RNAs codificantes e não codificantes / Clebiano da Costa Sá [orient.] Ariane Machado Lima [coorient.] Marcelo de Souza Lauretto. São Paulo : 2018. 101 p.

Dissertação (Mestrado) - Universidade de São Paulo
Orientadora: Profa. Dra. Ariane Machado Lima
Coorientador: Prof. Dr. Marcelo de Souza Lauretto

Programa Interunidades de Pós-Graduação em Bioinformática
Área de concentração: Bioinformática

1. Bioinformática. 2. Aprendizado Computacional. 3. RNA. I. Lima, Ariane Machado, orientadora. II. Lauretto, Marcelo de Souza, coorientador. III. Universidade de São Paulo. IV. Título.

CDD - 572.8

Dissertação de autoria de Clebiano da Costa Sá, sob o título “**Métodos de validação tradicional e temporal aplicados à avaliação de classificadores de RNAs codificantes e não codificantes**”, apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Bioinformática, aprovada em 23 de março de 2018 pela comissão julgadora constituída pelos doutores:

Profa. Dra. Arianne Machado Lima

Presidente

Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP)

Prof. Dr. Ricardo Cerri

Universidade Federal de São Carlos (UFSCar)

Profa. Dra. Ana Carolina Tahira

Universidade de São Paulo (USP)

Dedico aos meus pais, irmãos e esposa.

Agradecimentos

Agradeço à minha orientadora Dr.^a Arianne Machado Lima e ao meu co-orientador Dr. Marcelo de Souza Laurretto.

Aos meus pais por todo o carinho, por acreditarem, por investirem suas energias na minha formação profissional e por serem as principais referências para a formação do meu caráter.

Ao Dr. Marcondes Albuquerque de Oliveira e ao Dr. José Alves de Siqueira Filho por terem acompanhado e guiado meus primeiros passos como pesquisador.

À Dr.^a Michely Correia Diniz pelo apoio e pelos constantes estímulos e cobranças, que foram determinantes em todas as minhas conquistas e no meu desempenho como pesquisador.

Ao CEMAFAUNA - Centro de Manejo de Fauna da Caatinga por financiar minha participação no curso de verão em que conheci este programa de pós-graduação.

À CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pela bolsa concedida.

À FAPESP grant # 2011/50761-2, CNPq, CAPES, NAP eScience - PRP - USP.

“O sucesso nasce do querer, da determinação e persistência em se chegar a um objetivo.

Mesmo não atingindo o alvo, quem busca e vence obstáculos, no mínimo fará coisas admiráveis.”

(José de Alencar)

Resumo

COSTA-SÁ, Clebiano. **Métodos de validação tradicional e temporal aplicados à avaliação de classificadores de RNAs codificantes e não codificantes**. 2018. 101 f. Dissertação (Mestrado em Bioinformática) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2018.

Os ácidos ribonucleicos (RNAs) podem ser classificados em duas classes principais: codificante e não codificante de proteína. Os codificantes, representados pelos RNAs mensageiros (mRNAs), possuem a informação necessária à síntese proteica. Já os RNAs não codificantes (ncRNAs) não são traduzidos em proteínas, mas estão envolvidos em várias atividades celulares distintas e associados a várias doenças tais como cardiopatias, câncer e distúrbios psiquiátricos. A descoberta de novos ncRNAs e seus papéis moleculares favorece avanços no conhecimento da biologia molecular e pode também impulsionar o desenvolvimento de novas terapias contra doenças. A identificação de ncRNAs é uma ativa área de pesquisa e um dos correntes métodos é a classificação de sequências transcritas utilizando sistemas de reconhecimento de padrões baseados em suas características. Muitos classificadores têm sido desenvolvidos com este propósito, especialmente nos últimos três anos. Um exemplo é o *Coding Potential Calculator* (CPC), baseado em Máquinas de Vetores de Suporte (SVM). No entanto, outros algoritmos robustos são também reconhecidos pelo seu potencial em tarefas de classificação, como por exemplo *Random Forest* (RF). O método mais utilizado para avaliação destas ferramentas tem sido a validação cruzada *k-fold*. Uma questão não considerada nessa forma de validação é a suposição de que as distribuições de frequências dentro do banco de dados, em termos das classes das sequências e outras variáveis, não se alteram ao longo do tempo. Caso essa premissa não seja verdadeira, métodos tradicionais como a validação cruzada e o *hold-out* podem subestimar os erros de classificação. Constata-se, portanto, a necessidade de um método de validação que leve em consideração a constante evolução dos bancos de dados ao longo do tempo, para proporcionar uma análise de desempenho mais realista destes classificadores. Neste trabalho comparamos dois métodos de avaliação de classificadores: *hold-out* temporal e *hold-out* tradicional (atemporal). Além disso, testamos novos modelos de classificação a partir da combinação de diferentes algoritmos de indução com características de classificadores do estado da arte e um novo conjunto de características. A partir dos testes das hipóteses, observamos que tanto a validação *hold-out* tradicional quanto a validação *hold-out* temporal tendem a subestimar os erros de classificação, que a avaliação por validação temporal é mais fidedigna, que classificadores treinados a partir de parâmetros calibrados por validação temporal não melhoram a classificação e que nosso modelo de classificação baseado em *Random Forest* e treinado com características de classificadores do estado da arte e mais um novo conjunto de características proporcionou uma melhora significativa na discriminação dos RNAs codificantes e não codificantes. Por fim, destacamos o potencial do algoritmo *Random Forest* e das características utilizadas, diante deste problema de classificação, e sugerimos o uso do método de validação *hold-out* temporal para a obtenção de estimativas de desempenho mais fidedignas para os classificadores de RNAs codificantes e não codificantes de proteína.

Palavras-chaves: *Hold-out* temporal. *Hold-out* tradicional. Classificação de RNAs. RNAs não codificantes. Reconhecimento de padrões. Aprendizado supervisionado.

Abstract

COSTA-SÁ, Clebiano. **Traditional and time validation methods applied to the evaluation of coding and non-coding RNA classifiers**. 2018. 101 p. Dissertation (Master in Bioinformatics) – Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2018.

Ribonucleic acids (RNAs) can be classified into two main classes: coding and non-coding of protein. The coding, represented by messenger RNAs (mRNAs), has the necessary information for protein synthesis. Non-coding RNAs (ncRNAs) are not translated into proteins but are involved in several distinct cellular activities associated with various diseases such as heart disease, cancer and psychiatric disorders. The discovery of new ncRNAs and their molecular roles favors advances in the knowledge of molecular biology and may also boost the development of new therapies against diseases. The identification of ncRNAs is an active area of research and one of the current methods is the classification of transcribed sequences using pattern recognition systems based on their characteristics. Many classifiers have been developed for this purpose, especially in the last three years. An example is the Coding Potential Calculator (CPC), based on Supporting Vector Machines (SVM). However, other robust algorithms are also recognized for their potential in classification tasks, such as Random Forest (RF). The most commonly used method for evaluating these tools has been cross-validation k-fold. An issue not considered in this form of validation is the assumption that frequency distributions within the database, in terms of sequence classes and other variables, do not change over time. If this assumption is not true, traditional methods such as cross-validation and hold-out may underestimate classification errors. The need for a validation method that takes into account the constant evolution of databases over time is therefore needed to provide a more realistic performance analysis of these classifiers. In this work we compare two methods of evaluation of classifiers: time hold-out and traditional hold-out (without considering the time). In addition, we tested new classification models from the combination of different induction algorithms with state-of-the-art classifier characteristics and a new set of characteristics. From the hypothesis tests, we observe that both the traditional hold-out validation and the time hold-out validation tend to underestimate the classification errors, that the time validation evaluation is more reliable, than classifiers trained from parameters calibrated by time validation did not improve classification and that our Random Forest-based classification model trained with state-of-the-art classifier characteristics and a new set of characteristics provided a significant improvement in the discrimination of the coding and non-coding RNAs. Finally, we highlight the potential of the Random Forest algorithm and the characteristics used, in view of this classification problem, and we suggest the use of the time hold-out validation method to obtain more reliable estimates of the protein coding and non-coding RNA classifiers.

Keywords: Time Hold-out. Traditional Hold-out. Classification of RNAs. Non-coding RNAs. Pattern Recognition. Supervised learning.

Lista de figuras

Figura 1 – Esquemas do fluxo da informação biológica: a) “Dogma Central” da biologia molecular e b) um novo fluxo da informação biológica	25
Figura 2 – Esquema da estrutura típica de um mRNA humano	26
Figura 3 – Exemplo de conjunto de treinamento	31
Figura 4 – Representação do processo de aprendizado supervisionado	32
Figura 5 – Representação do hiperplano ótimo, da margem e dos vetores de suporte em um problema de classificação binária	34
Figura 6 – Estrutura básica de uma árvore de decisão	36
Figura 7 – Esquema do processo de validação cruzada com $k = 3$	38
Figura 8 – Número de classificadores publicados em todos os anos	46
Figura 9 – Uso dos algoritmos de classificação pelos classificadores	46
Figura 10 – Uso das fontes dos transcritos que compõe os conjuntos de dados utilizados na validação ou no treinamento dos classificadores	47
Figura 11 – Total de citações dos artigos de 2006 a agosto de 2017	48
Figura 12 – Acumulado de citações dos artigos de 2016 a agosto de 2017	48
Figura 13 – Processo de extração e seleção de características para construção de modelos que simulam classificadores do estado da arte e de modelos novos	58
Figura 14 – Representação dos métodos de validação <i>hold-out</i> tradicional e temporal; pontos vermelhos=instâncias de treinamento, pontos azuis=instâncias de teste	62
Figura 15 – Representação dos métodos de validação <i>hold-out</i> tradicional e temporal aplicados em diferentes períodos	63
Figura 16 – AUC média para cada modelo de classificação sob o método de validação <i>hold-out</i> tradicional em cada período de treinamento: a) teste T1 e b) teste T2	68
Figura 17 – AUC média para cada modelo de classificação sob o método de validação <i>hold-out</i> temporal em cada período de treinamento: a) teste T1 e b) teste T2	70

Lista de tabelas

Tabela 1 – Representação da matriz de confusão para um problema de classificação binária	39
Tabela 2 – Sumário de todos os artigos analisados. Informações não mencionadas são descritas com NM.	42
Tabela 3 – Quantidade de RNAs codificantes (mRNA) e não codificantes (ncRNA) nos <i>folds</i> temporais	57
Tabela 4 – Características extraídas a partir do classificador PLEK	58
Tabela 5 – Características extraídas a partir do classificador CPAT	59
Tabela 6 – Características extraídas a partir do classificador CPC1	59
Tabela 7 – Características extraídas a partir do classificador CPC2	59
Tabela 8 – Características extraídas a partir do classificador CNCI	59
Tabela 9 – Características extraídas a partir de um novo extrator	60
Tabela 10 – Conjunto formado por todas as características não redundantes	61
Tabela 11 – AUC e ACC média das avaliações por período para cada modelo de classificação sob validação <i>hold-out</i> tradicional nos testes T1 e T2. Entre parênteses está a média dos desvios padrões obtidos a partir das avaliações por período	67
Tabela 12 – AUC e ACC média das avaliações por período para cada modelo de classificação sob validação <i>hold-out</i> temporal nos testes T1 e T2. Entre parênteses está a média dos desvios padrões obtidos a partir das avaliações por período	69
Tabela 13 – AUC e ACC média das avaliações por período para cada modelo de classificação sob os métodos de validação <i>hold-out</i> temporal e tradicional no teste T1. Entre parênteses está a média dos desvios padrões obtidos a partir das avaliações por período	71
Tabela 14 – AUC e ACC média das avaliações por período para cada modelo de classificação sob os métodos de validação <i>hold-out</i> temporal e tradicional no teste T2. Entre parênteses está a média dos desvios padrões obtidos a partir das avaliações por período	72
Tabela 15 – Características selecionadas com base no método RELIEF	74

Tabela 16 – AUC média calculada para as 100 amostras do teste T1 e desvio padrão entre parênteses. Na cor verde estão as maiores médias dentre os classificadores para cada método de validação e período de treinamento. Quando houver, a cor azul indica equivalência estatística entre este modelo e o de referência ($p\text{-valor} > 0,1$)	76
Tabela 17 – AUC média calculada para as 100 repetições do teste T2 e desvio padrão entre parênteses. Na cor verde estão as maiores médias dentre os classificadores para cada método de validação e período de treinamento. Quando houver, a cor azul indica equivalência estatística entre este modelo e o de referência ($p\text{-valor} > 0,1$)	77
Tabela 18 – ACC média calculada para as 100 amostras do teste T1 e desvio padrão entre parênteses. Na cor verde estão as maiores médias dentre os classificadores para cada método de validação e período de treinamento. Quando houver, a cor azul indica equivalência estatística entre este modelo e o de referência ($p\text{-valor} > 0,1$)	78
Tabela 19 – ACC média calculada para as 100 repetições do teste T2 e desvio padrão entre parênteses. Na cor verde estão as maiores médias dentre os classificadores para cada método de validação e período de treinamento. Quando houver, a cor azul indica equivalência estatística entre este modelo e o de referência ($p\text{-valor} > 0,1$)	79
Tabela 20 – Principais características em ordem decrescente de importância, calculada com base no decréscimo médio da acurácia durante a construção de um modelo de classificação baseado em Random Forest. O valor “Importância” corresponde ao Decréscimo Médio da Acurácia em porcentagem	98

Lista de abreviaturas e siglas

A	Adenina
AS	Antisenso
ACC	Acurácia
AUC	Área sob a curva ROC
BRF	Algoritmo <i>Random Forest</i> balanceado
C	Citosina
cDNA	DNA complementar
CDS	Sequência codificante
CNCI	<i>Coding-Non-Coding Index</i>
CNCTDiscriminator	<i>Coding and Noncoding Transcript Discriminator</i>
CONC	<i>Coding Or Non-Coding</i>
CPAT	<i>Coding Potential Assessment Tool</i>
CPC	<i>Coding Potential Calculator</i>
DDBT	<i>Dynamic Discriminant Bounds Tree</i>
DNA	Ácido Desoxirribonucleico
FN	Falso Negativo
FP	Falso Positivo
G	Guanina
H-InvDB	<i>H-Invitational Database</i>
IT	Intrônico
LINC	Intergênico
lncRNA	Longo ncRNA

miRNA	microRNA
MLCDS	<i>most-like</i> CDS
mRNA	RNA mensageiro
MTRY	Número de características testadas por nó
ncRNA	RNA não codificante
nt	Nucleotídeo
ORF	<i>Open Reading Frame</i>
piRNA	RNA <i>Piwi</i>
PLEK	<i>Predictor of long non-coding RNAs and messenger RNAs based on an improved k-mer scheme</i>
<i>profile</i> HMM	Perfis de modelos ocultos de <i>Markov</i>
RBF	<i>Radial Basis Function</i>
RF	<i>Random Forest</i>
RNA	Ácido Ribonucleico
ROSE	<i>Random Over Sampling Examples</i>
RPKM	<i>Read Per Kilobases Per Millions short reads</i>
rRNA	RNA ribossomal
siRNA	Pequeno RNA de interferência
SMOTE	<i>Synthetic Minority Over-sampling Technique</i>
snoRNA	Pequeno RNA nucleolar
snRNA	Pequeno RNA nuclear
sRNA	Pequeno ncRNA
SVM	Máquinas de Vetores de Suporte

TDIDT	<i>Top-Down Induction of Decision Tree</i>
tRNA	RNA de transferência
T1	Teste 1
T2	Teste 2
U	Uracila
UTR	<i>Untranslatable Region</i>
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

Lista de símbolos

$\in$	Pertence
$\mathbb{R}$	Conjunto dos números reais
ϕ	Letra grega Fi
γ	Letra grega minúscula gama

Sumário

1	Introdução	18
1.1	Organização deste documento	22
2	Objetivos	24
2.1	Objetivo geral	24
2.2	Objetivos específicos	24
3	Conceitos fundamentais	25
3.1	Diversidade estrutural e funcional do RNA	25
3.1.1	RNAs codificantes	26
3.1.2	RNAs não codificantes	26
3.2	Bioinformática: bancos de dados de RNA	28
3.3	Classificação e aprendizado supervisionado	30
3.3.1	Seleção de características	32
3.3.2	SVM linear e SVM não linear	33
3.3.3	Árvores de classificação	35
3.3.4	Random Forest	37
3.4	Avaliação e seleção de modelos	37
4	Estado da arte na área de pesquisa	41
4.1	Análise global	41
4.2	Análise qualitativa	48
4.2.1	Classificadores baseados em SVM	49
4.2.2	Classificadores baseados em Regressão Logística	51
4.2.3	Classificadores baseados em Random Forest	51
4.2.4	Classificadores baseados em Deep Learning	52
4.3	Discussões e considerações finais	52
5	Materiais e métodos	55
5.1	Conjunto de dados	55
5.2	Modelos de classificação	57
5.3	Métodos de validação <i>hold-out</i> tradicional e temporal	62

5.4	Consolidação dos resultados e teste das hipóteses	64
6	Resultados e discussões	66
6.1	Análise da validação <i>hold-out</i> tradicional	66
6.2	Análise da validação <i>hold-out</i> temporal	68
6.3	Análise comparativa dos métodos de validação <i>hold-out</i> tradicional e temporal	70
6.4	Avaliação dos modelos de classificação	73
7	Conclusão	80
7.1	Principais contribuições	80
7.2	Trabalhos futuros	82
	Referências ¹	83
	Apêndice A—Classificadores de RNAs codificantes e não codificantes baseados em apren- dizado supervisionado: Protocolo de Revisão	93
	Apêndice B—Importância das características para um modelo de classificação baseado em Random Forest	98

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

Até aproximadamente o ano 2000, os ácidos ribonucleicos (RNAs) eram considerados meros intermediários entre o genoma e as proteínas, como rege o “Dogma Central” da biologia molecular. Recentes descobertas envolvendo uma variedade de novos genes de RNAs não codificantes (ncRNAs), seus papéis biológicos e mecanismos de ação (LIU et al., 2005) têm mostrado que a diversidade e a importância dos ncRNAs foram subestimadas (HUTTENHOFER; SCHATTNER; POLACEK, 2005). Atualmente, sabe-se que RNAs que não codificam proteínas desempenham importantes funções.

As famílias de ncRNAs conhecidas até os anos de 1980 eram as de RNAs de transferência (tRNAs) (SPRINZL et al., 1998) e de RNAs ribossomais (rRNAs) (COLE et al., 2003). Atualmente milhões de ncRNAs são conhecidos. Até fevereiro de 2018 o banco de dados RNACentral versão 8 agrupou 26 bancos de ncRNAs, com cerca de 13,1 milhões de ncRNAs (RNACentral Consortium, 2015). Todos esses ncRNAs variam em função e tamanho, podendo possuir de 32 a 244.296 nucleotídeos (nt). Dentre as diversas espécies representadas, *Mus musculus* e *Homo sapiens* apresentam o maior número de resultados, com cerca de 181.400 e 178.700 ncRNAs, respectivamente.

Os RNAs não codificantes estão envolvidos em várias atividades celulares, como localização de proteínas (NAGAI et al., 2003), ativação de células T (TAM; BEN-YEHUDA; HAYWARD, 1997), demetilação de DNA (IMAMURA et al., 2004), regulação da localização subcelular de fatores de transcrição (WILLINGHAM et al., 2005), replicação do DNA (GREIDER, 1995), estabilidade cromossômica (BLACKBURN, 1993), processos de transporte e envolvimento com resistência a drogas (VAN ZON et al., 2003), *splicing* (VALADKHAN, 2005), estabilidade (NOVOA; WILLIAMS, 2004), translocação (KEENAN et al., 2001) e localização de proteínas (ROSENBLAD et al., 2003), processamento (PICCINELLI; ROSENBLAD; SAMUELSSON, 2005) e modificação (LOWE, 1999) de RNAs e regulação da expressão gênica (STORZ; ALTUVIA; WASSARMAN, 2005).

A regulação da expressão gênica, em particular, parece ser uma das principais atividades dos ncRNAs. O desenvolvimento normal de um organismo e a manutenção de seu bom funcionamento depende de níveis adequados de expressão gênica, sejam de genes codificantes de proteína ou de ncRNAs. Desta forma, uma desregulação dessa expressão pode ser a causa de muitas doenças ou, pelo menos, conferir susceptibilidade a elas (COSTA,

2005). É justamente no controle de expressão gênica que muitos ncRNAs exercem papéis fundamentais, seja no controle pré ou pós transcricional.

No nível pré-transcricional, os ncRNAs estão envolvidos no silenciamento de genes por meio de alterações da estrutura da cromatina (ZARATIEGUI; IRVINE; MARTIENSSSEN, 2007), o que ocorre, por exemplo, no *imprinting*¹ genômico e na inativação do cromossomo X em mamíferos (SLEUTELS; ZWART; BARLOW, 2000; ZARATIEGUI; IRVINE; MARTIENSSSEN, 2007; MOREIRA, 2015). No nível pós-transcricional, os ncRNAs estão envolvidos tanto no silenciamento completo como na modulação da expressão de outros genes. Três grandes grupos de ncRNAs envolvidos nestes processos são os microRNAs (miRNAs), pequenos RNAs de interferência (siRNAs) e RNAs antisenso (LIN; MILLER; YING, 2006; HANNON, 2002; HUTTENHOFER; SCHATTNER; POLACEK, 2005; REIS et al., 2004; REIS et al., 2005).

Diversos ncRNAs estão relacionados com doenças ou transtornos, tais como cânceres (COSTA, 2005; REIS et al., 2004; REIS et al., 2005; LIZ; ESTELLER, 2016), doenças coronárias (ISHII et al., 2006; BUSCH; EKEN; MAEGDEFESSEL, 2016), anomalias congênitas (ARTANDI, 2006; MATTICK; MAKUNIN, 2006; HERMANNNS et al., 2005; HUANG et al., 2015), síndromes neurológicas (COSTA, 2005; PETAZZI et al., 2013), transtornos psiquiátricos (JIN; ALISCH; WARREN, 2004; OSBORNE; THORNTON, 2006; POLESSKAYA et al., 2003; CAVAILLE et al., 2000; JI et al., 2015), entre outras.

Especula-se ainda que estas moléculas podem explicar melhor as diferenças na complexidade dos vários organismos do que as proteínas. Estudos apontam que há uma correlação entre o aumento do número e diversidade de ncRNAs com a complexidade do desenvolvimento de espécies (GUENNEWIG; COOPER, 2014). Neste novo cenário, é revelado que o “Dogma Central” da biologia molecular pode ser muito mais complexo, no qual ncRNAs desempenham relevantes papéis (COSTA, 2005).

A identificação computacional de ncRNAs é ainda um problema em aberto em bioinformática (ACHAWANANTAKUN et al., 2015; FAN; ZHANG, 2015; SUN et al., 2015; YOTSUKURA et al., 2016). Ao contrário dos mRNAs, que possuem uma longa ORF², ncRNAs não possuem uma característica em comum na sequência primária. Mesmo no

¹ *Imprinting* é um fenômeno no qual apenas um dos alelos de um gene é expresso. A diferenciação entre os dois alelos é realizada por marcas epigenéticas estabelecidas durante a gametogênese (WOOD; OAKEY, 2006).

² *Open Reading Frame*, um quadro aberto de leitura onde é possível decodificar a sequência de nucleotídeos em sequências de aminoácidos.

caso de ncRNAs de mesma função, principalmente quando a função está relacionada com sua estrutura tridimensional, a tendência de não conservação da sequência pode tornar ineficaz o uso de ferramentas baseadas em similaridade como BLAST ([ALTSCHUL et al., 1997](#)). Busca por sinais promotores de RNA polimerases em sequências genômicas também é mais difícil em ncRNAs, pois alguns deles são também transcritos pelas RNA polimerases I e III, cujos promotores são menos conhecidos ([BROWN, 2002](#)). Há ainda ncRNAs como os pequenos RNAs nucleolares (snoRNAs) de vertebrados que, ao invés de serem transcritos independentemente, são processados a partir de *íntrons* de outros RNAs ([EDDY, 2002](#)).

Sequenciadores de nova geração como *Ion PGM* e *Ion Proton*, baseados em metodologias muito mais simples e de menor custo ([MOREIRA, 2015](#)), viabilizam o sequenciamento em larga escala dos RNAs de uma amostra, permitindo a detecção tanto de genes conhecidos quanto de genes ainda não conhecidos, sejam eles codificantes ou não. Portanto, a classificação de transcritos a partir de suas características parece ser uma abordagem promissora e uma boa alternativa à estratégia de identificação de genes de ncRNAs em sequências genômicas.

O primeiro classificador de RNAs codificantes e não codificantes baseado em aprendizado supervisionado e em características de sequências transcritas, o CONC (*Coding Or Non-Coding*) ([LIU; GOUGH; ROST, 2006](#)), surgiu em 2006. A partir de então, outros classificadores foram desenvolvidos na tentativa de melhorar a discriminação destas classes ([FAN; ZHANG, 2015](#); [SUN et al., 2015](#); [ACHAWANANTAKUN et al., 2015](#); [LI; ZHANG; ZHOU, 2014](#)).

Todavia, a estratégia de classificação direta de transcritos baseada em características também impõe seus desafios, pois sabe-se atualmente que muitos ncRNAs se assemelham a RNAs codificantes. Por exemplo, são conhecidos ncRNAs longos e com longas ORFs, transcritos pela polimerase II (poliadenilados e com cap5') e que podem sofrer *splicing* ([RYMARQUIS et al., 2008](#)). Existem também pequenos peptídeos ([FRITH et al., 2006](#)), sendo alguns traduzidos de longos transcritos ([GALINDO et al., 2007](#)), que poderiam ser confundidos com ncRNAs. Há também ncRNAs que evoluíram de mRNAs ([ALLEN et al., 2004](#)), podendo ser confundidos com eles quando usado apenas um critério de similaridade. Longos ncRNAs intrônicos também poderiam ser, à primeira vista, considerados mRNAs imaturos ([LOURO; SMIRNOVA; VERJOVSKI-ALMEIDA, 2009](#)). Isso sem mencionar pseudogenes ([ZOU et al., 2009](#)) e regiões repetitivas ([HASLER; SAMUELSSON; STRUB,](#)

2007) que são expressos. Desta forma, há fatores confundidores que podem comprometer o resultado da classificação (DINGER et al., 2008).

Ainda se nota a carência de bons classificadores. A exemplo disso podemos citar a discrepância entre o fato de que, embora estudos indiquem que 80% do genoma humano seja transcrito (PENNISI, 2012) e que apenas aproximadamente 2% do genoma humano seja anotado como codificante (FRITH; PHEASANT; MATTICK, 2005), até fevereiro de 2018 pouco mais de 10% dos transcritos humanos depositados no banco de dados H-InvDB (*H-Invitational Database*) versão 9 são anotados como não codificantes (TAKEDA et al., 2013). A diferença é considerada uma “matéria escura”, cuja classificação e função ainda não é conhecida (JOHNSON et al., 2005).

Além disso, os classificadores têm sido avaliados com base no método de validação cruzada *k-fold*, uma estratégia atemporal baseada na premissa de que as sequências contidas no banco de dados são independentes e identicamente distribuídas, o que significa que a introdução de novas sequências nos bancos de dados é *um processo estacionário*, ou seja, sua distribuição não se altera ao longo do tempo.

Todavia, a premissa de estacionariedade na identificação e introdução de novas séries nos bancos de dados não é verificada na prática. O que normalmente ocorre é a identificação de uma determinada sequência e, a partir desta, a identificação de várias outras correlatas, seja por similaridade de sequência, de estrutura ou alguma outra característica marcante levada em consideração. Logo, o processo de evolução dos bancos envolve não só um acréscimo de mais exemplos correlatos aos pré-existentes mas também uma variação nas distribuições das sequências. Neste cenário, a validação *k-fold* pode não capturar adequadamente o desempenho dos classificadores ao longo do tempo.

Ao considerar tais variações ao longo do tempo, é importante verificar quão acurados os classificadores são para novas sequências e, conseqüentemente, a influência da evolução dessas bases de dados sobre o desempenho dos classificadores. Uma alternativa é realizar uma forma de validação temporal, similar à realizada para séries temporais, em que se treina o modelo com as instâncias depositadas até o instante t e testa-se com as instâncias depositadas no instante $t + 1$. A ideia desta proposta consiste em avaliar o desempenho de um classificador quando aplicado sobre novas instâncias ao longo do tempo.

Justifica-se, portanto, a busca por um método de avaliação que leve em consideração a constante evolução dos bancos de dados ao longo do tempo e o desenvolvimento de um

novo classificador que melhore, em algum aspecto, a classificação dos RNAs codificantes e não codificantes.

Considerando os problemas de pesquisa descritos, cinco hipóteses foram delineadas para essa pesquisa:

- Hipótese H1: A validação tradicional (atemporal) tende a subestimar os erros de classificação, ou seja, quando extraídos os conjuntos de treinamento e de teste sobre um mesmo grande período, sem levar em consideração o ano de descobrimento e rotulagem das sequências, as estimativas de erro tendem a ser menores do que os erros medidos sobre sequências mais novas (do último ano);
- Hipótese H2: A validação temporal tende a não subestimar os erros de classificação, no sentido de que, os erros de classificação obtidos sobre instâncias de testes em $t + 1$ tendem a ser parecidos com os erros de classificação em $t + 2$;
- Hipótese H3: A validação temporal tende a fornecer estimativas mais fidedignas, no sentido de que, quando utilizada a validação temporal os erros de classificação obtidos sobre instâncias de testes em $t + 1$ tendem a ser mais parecidos com os erros de classificação em $t + 2$;
- Hipótese H4: Os classificadores calibrados pelo método de validação temporal tendem a fornecer erros de classificação menores que classificadores calibrados pelo método tradicional;
- Hipótese H5: Um novo modelo de classificação, construído a partir da associação das características de classificadores do estado da arte com um conjunto de novas características, proporciona melhoras na discriminação dos RNAs codificantes e não codificantes.

1.1 Organização deste documento

O Capítulo 2 descreve o objetivo geral e os objetivos específicos deste trabalho. O Capítulo 3 descreve os conceitos fundamentais utilizados neste trabalho: RNAs (codificantes e não codificantes), bancos de dados de RNA, aprendizado supervisionado, algoritmos de classificação e seleção de características, bem como métodos de avaliação e seleção de classificadores. O Capítulo 4 apresenta o estado da arte na área de classificação computacional de RNAs levantado a partir de uma revisão sistemática. No Capítulo 5 são

apresentados os materiais e métodos utilizados. O Capítulo 6 apresenta os resultados e discussões desta pesquisa. Finalmente, o Capítulo 7 apresenta as conclusões a respeito dos resultados.

2 Objetivos

2.1 Objetivo geral

Comparar duas estratégias de avaliação de classificadores de RNAs: temporal e tradicional (atemporal). Além disso, testar um novo classificador proposto neste trabalho que inclua novas características.

2.2 Objetivos específicos

1. Realizar uma revisão sistemática dos classificadores de RNAs codificantes e não codificantes disponíveis na literatura;
2. Implementar os métodos de validação tradicional e temporal;
3. Avaliar o método de validação temporal, comparado ao tradicional, a partir da avaliação de diferentes classificadores de RNAs codificantes e não codificantes;
4. Aprimorar, se necessário, um programa extrator de características implementado previamente por Silva ([SILVA, 2009b](#); [SILVA, 2009a](#)) sob orientação da Prof^a. Arianne Machado Lima;
5. Avaliar o desempenho de diferentes métodos de seleção de características para o problema proposto;
6. Avaliar o desempenho de diferentes métodos de aprendizado supervisionado, a fim de identificar os métodos com desempenho significativamente superior;
7. Verificar se o melhor modelo de classificação (dentre os testados) é significativamente superior aos classificadores do estado da arte.

3 Conceitos fundamentais

3.1 Diversidade estrutural e funcional do RNA

As buscas pela molécula responsável por armazenar e transmitir a informação genética de uma célula ou de um indivíduo como um todo resultaram no estabelecimento do “Dogma Central” da biologia molecular, mais coerentemente denominado de “Fluxo da Informação Biológica” (Figura 1a). De acordo com este modelo, o Ácido Desoxirribonucleico (DNA) seria o detentor de toda a informação genética, as proteínas o produto final e o RNA um simples intermediário entre DNA e proteínas (MATIOLI; FERNANDES, 2012). No entanto, após a descoberta da grande diversidade estrutural e funcional dos RNAs, em especial dos ncRNAs, começou a emergir um novo modelo, mais completo, contemplando os principais papéis dos ncRNAs (Figura 1b) (COSTA, 2005).

Figura 1 – Esquemas do fluxo da informação biológica: a) “Dogma Central” da biologia molecular e b) um novo fluxo da informação biológica

Fonte: Adaptado de Fabrício F. Costa, 2005

Os ácidos ribonucleicos (RNAs) são moléculas formadas por um único filamento de nucleotídeos, que, por sua vez, são composto por um ou mais grupos fosfato, um açúcar ribose e uma das quatro bases nitrogenadas (adenina (A), uracila (U), guanina (G) ou citosina (C)) (SNUSTAD; SIMMONS, 2008). Tal estrutura possibilita ao RNA dobrar-se sobre si mesmo a partir do pareamento de segmentos complementares e da formação de

pares canônicos (G-C e A-U) e pares “não-convencionais” (geralmente G-U) ([ALBERTS et al., 2010](#)), refletindo em sua grande diversidade estrutural e funcional.

Embora sejam conhecidos vários tipos distintos de RNAs, estes podem ser classificados em duas grandes classes: os codificantes e os não codificantes de proteína. Enquanto os primeiros são representados pelos mRNAs, os não codificantes englobam todos os demais RNAs.

3.1.1 RNAs codificantes

Os RNAs codificantes, representados pelos RNAs mensageiros (mRNAs), são responsáveis por carrear a informação genética necessária à codificação das proteínas nos ribossomos. Em eucariotos a estrutura destas moléculas é tipicamente composta por uma cauda *cap5* na extremidade 5', uma região não traduzida (5' UTR - *Untranslatable Region*), um quadro aberto de leitura (ORF - *Open Reading Frame*) iniciado com um códon de início e finalizado com um códon de término, uma outra região não traduzida (3' UTR - *Untranslatable Region*) e uma cauda *poliA* na extremidade 3' (Figura 2) ([ALBERTS et al., 2010](#)).

Figura 2 – Esquema da estrutura típica de um mRNA humano

Fonte: Adaptado de Raphael Alves da Silva, 2009

3.1.2 RNAs não codificantes

Os RNAs não codificantes (ncRNAs) são moléculas funcionais que não serão traduzidas em proteínas, mas que apresentam diversas funções biológicas em várias redes

regulatórias (EDDY, 2001). Esta classe engloba quase toda a diversidade de RNAs, com exceção apenas dos mRNAs.

São conhecidos diversos ncRNAs com estrutura secundária conservada e/ou importante para sua função. No banco de dados Rfam (versão 12.1 de abril de 2016) estão anotadas 2.474 famílias (YOTSUKURA et al., 2016). Estas famílias estão organizadas em tipos e subtipos, tais como, microRNAs (miRNAs) (CULLEN, 2015; KIM; HAN; SIOMI, 2009), RNAs *Piwi* (piRNAs) (KIM; HAN; SIOMI, 2009), pequenos RNAs nucleares (snRNA) (MATERA; TERNES; TERNES, 2007), pequenos RNAs nucleolares (snoRNA) (HOLLEY; TOPKARA, 2011), RNAs de transferência (tRNAs) (SHEPHERD; IBBA, 2015), RNAs ribossomais (rRNAs) (COLE et al., 2003), ribozimas (DEROSE, 2002), RNAs antisenso (VILLEGAS; ZAPHIROPOULOS, 2015), além de íntrons catalíticos (LEHMANN; SCHMIDT, 2003) e de elementos regulatórios (WITTKOPP; KALAY, 2011).

microRNA são moléculas pequenas, com cerca de 20-24 nt. Conhecidos por desempenharem papéis regulatórios importantes durante a diferenciação e o desenvolvimento celular em mamíferos a partir da regulação da expressão gênica (CULLEN, 2015). Estas moléculas são complementares inversos a trechos de outros transcritos e muitas vezes podem inibir a expressão do gene alvo, em um processo conhecido como silenciamento gênico (KIM; HAN; SIOMI, 2009).

piRNA são pequenos RNAs (24 a 31 nt) associados a subfamília de proteínas *Piwi*. São altamente abundantes em células germinativas e pelo menos algumas estão envolvidas na formação de transposons através do silenciamento de heterocromatina ou desestabilização de RNA. Diferem de miRNA por serem maiores e pelo fato de não dependerem da enzima *Dicer* em sua biogênese (KIM; HAN; SIOMI, 2009).

snRNA são pequenos ncRNAs nucleares que unem-se a subunidades proteicas para a formação do complexo *spliceossomo*, que atua na remoção dos *íntrons* de mRNAs eucarióticos (MATERA; TERNES; TERNES, 2007; GRIFFITHS et al., 2008).

snoRNA são pequenos ncRNAs nucleolares (20 a 25 nt) localizados geralmente nos nucléolos (ENDER et al., 2008). Ao invés de serem transcritos independentemente, são processados a partir de *íntrons* de outros RNAs (EDDY, 2002). Estas moléculas exercem importantes papéis na maturação de outros ncRNAs tais como rRNA ou snRNA (HOLLEY; TOPKARA, 2011; ENDER et al., 2008).

tRNAs são moléculas em formato de um trevo de quatro folhas, responsáveis por carrear os aminoácidos para o processo de tradução nos ribossomos, além de exercerem

papel catalítico e estarem relacionados a regulação da morte celular (SHEPHERD; IBBA, 2015; XIAOLONG; ENDUO, 2013; HOU; YANG, 2013).

rRNAs são as moléculas constituintes dos complexos ribossomais, que guiam a montagem das cadeias de aminoácidos a partir de um mRNA e dos tRNAs (GRIFFITHS et al., 2008; COLE et al., 2003).

RNAs antisenso são definidos como transcritos derivados a partir da fita oposta do gene codificante de proteína. Estes ncRNAs operam baseados em duas formas, em *cis* ou *trans*. Em *cis*, o RNA antisenso interage com o gene alvo a partir da mesma região do DNA. Já em *trans*, o RNA antisenso interage com genes de regiões distantes ou até mesmo de outros cromossomos (VILLEGAS; ZAPHIROPOULOS, 2015).

Ribozimas são RNAs catalíticos, que catalizam diversas reações nas células. Apesar das diferenças globais em características químicas, ambas as enzimas RNA e proteína utilizam estratégias catalíticas semelhantes (FEDOR; WILLIAMSON, 2005; DEROSE, 2002).

Em termos de tamanho os ncRNAs podem ser divididos arbitrariamente em pequenos e longos (YOTSUKURA et al., 2016). Os pequenos ncRNAs (sRNAs) são definidos como menores que 200 nt. Exemplos incluem miRNAs, snRNAs e tRNAs. Já os longos ncRNAs (lncRNAs) são definidos como maiores que 200 nt. Estes são ainda sub-divididos em duas classes principais: (1) transcritos com função conhecida (por exemplo, XIST ou HOTAIR), e (2) transcritos sem anotação funcional, os quais são subdivididos de acordo com sua localização genômica, tais como: intrônicos (IT), antisenso (AS) ou intergênico (LINC) (PASCHOAL et al., 2012).

3.2 Bioinformática: bancos de dados de RNA

Há uma grande quantidade de bancos de dados públicos relacionados a RNAs, a grande maioria especializados em determinadas classes de ncRNAs ou em determinado tipo de informação armazenada (PASCHOAL et al., 2012; YOTSUKURA et al., 2016). Buscas a partir da ferramenta *web* NRDR v2.0 retornaram 144 bancos de dados disponíveis para ncRNAs (PASCHOAL et al., 2012). Dentre eles, 129 bancos são específicos para determinada classe de ncRNA e 15 são generalistas (com múltiplas classes). São armazenados 211 tipos de informações. Alguns exemplos incluem anotação, sequência,

expressão, gene alvo, gene e estrutura (armazenadas, respectivamente, por 72, 71, 41, 38, 37 e 33 bancos de dados).

Alguns exemplos de bancos de dados especializados incluem: lncRNAdb ([AMARAL et al., 2011](#); [QUEK et al., 2015](#)), LNCipedia ([VOLDERS et al., 2015](#)), LncRNAdisease ([CHEN et al., 2013](#)), DIANA-LncBase ([PARASKEVOPOULOU et al., 2016](#)) e miRBase ([KOZOMARA; GRIFFITHS-JONES, 2014](#)). Os quais armazenam distintas informações e distintas classes de ncRNAs. lncRNAdb foca em longos ncRNAs com funções descritas ([AMARAL et al., 2011](#); [QUEK et al., 2015](#)). LNCipedia deposita lncRNAs humanos e apresenta várias informações para cada entrada, tais como informações da estrutura secundária, potencial codificante e sítios de ligação a miRNAs ([VOLDERS et al., 2015](#)). LncRNAdisease apresenta lncRNAs associados a doenças ([CHEN et al., 2013](#)). DIANA-LncBase apresenta lncRNAs alvos de miRNAs ([PARASKEVOPOULOU et al., 2016](#)). Já o miRBase é dedicado aos miRNAs de diversas espécies ([KOZOMARA; GRIFFITHS-JONES, 2014](#)).

Dentre os bancos de dados generalistas podem ser citados exemplos como: Rfam ([NAWROCKI et al., 2015](#)), NONCODE ([ZHAO et al., 2016](#)), deepBase ([ZHENG et al., 2016](#)), fRNAdb ([KIN et al., 2007](#)), ncRNAdb ([SZYMANSKI; ERDMANN; BARCISZEWSKI, 2007](#)) e RNACentral ([RNACentral Consortium, 2015](#)).

Rfam é uma coleção de diversas famílias de ncRNAs (2.686 famílias na versão 13.0 em fevereiro de 2018) com estrutura secundária conservada representadas por alinhamentos múltiplos, modelos de covariância e estruturas secundárias consenso. Nele são armazenadas informações relacionadas a sequência, estrutura secundária, taxonomia, alinhamento e árvores filogenéticas de diferentes organismos ([NAWROCKI et al., 2015](#)).

NONCODE abrange todos os tipos de ncRNAs (exceto rRNAs e tRNAs) e tem como objetivo apresentar a mais completa coleção e anotação de ncRNAs, especialmente dos lncRNAs que já somam mais de 500 mil sequências em fevereiro de 2018. Os ncRNAs depositados foram extraídos a partir do Genbank, do PubMed e dos bancos de dados Ensembl, RefSeq e lncRNAdb com posterior curadoria manual para remoção de repetições e de falsos lncRNAs. Para calcular o potencial codificante das entradas NONCODE utiliza o classificador *Coding-Non-Coding Index* (CNCI) ([ZHAO et al., 2016](#)).

RNACentral é um banco de dados desenvolvido pelo *The RNACentral Consortium* com o propósito de coletar e integrar informações a partir de um consorcio internacional de bancos de dados de sequência estabelecidos. Até fevereiro de 2018 foram integrados

26 bancos, o que representa cerca de 13,1 milhões de ncRNAs ([RNAcentral Consortium, 2015](#)).

3.3 Classificação e aprendizado supervisionado

Um classificador é um conjunto de regras, comandos ou funções construídos com o objetivo de prever a classe de uma instância, com base em seus atributos ou características observadas. A construção ou indução do classificador pode ser realizada por aprendizado supervisionado, aprendizado semi-supervisionado ou por aprendizado não supervisionado ([FRANKLIN, 2005](#)).

O aprendizado supervisionado, que é o foco deste trabalho, consiste da saída de um algoritmo de aprendizado que, dada uma amostra de treinamento como entrada, infere um modelo (que depende do algoritmo de aprendizado - alguns deles descritos nas seções [3.3.2](#) a [3.3.4](#)) para as classes presentes na amostra de treinamento. Ou seja, os algoritmos de classificação buscam induzir classificadores que discriminem adequadamente as instâncias fornecidas segundo suas classes, de forma que, uma vez construídos, tais classificadores possam prever adequadamente as classes de novas instâncias ([VAPNIK, 2000](#)). Aqui o objetivo é aprender uma função a partir de uma entrada $\mathbf{x}$ (um vetor d -dimensional que representa os atributos da instância) para uma saída y , onde $y \in \{1, \dots, C\}$, sendo y a classe da instância e C o número de classes. Quando $C = 2$ diz-se que o classificador é binário e quando $C > 2$ diz-se que o classificador é do tipo multiclasse.

Atributos são variáveis observáveis e independentes que representam as propriedades ou características de um determinado elemento ([BASGALUPP, 2010](#)). Os atributos podem ser organizados em dois tipos: qualitativo (também conhecido como categórico) ou quantitativo ([MAIMON; ROKACH, 2010](#)). Mesmo quando representados na forma numérica os atributos qualitativos não podem ser submetidos às operações aritméticas (soma, subtração, média, etc.). Já os atributos quantitativos são de natureza numérica e representam contagens e medidas, sendo portanto aplicáveis as operações aritméticas. Os atributos quantitativos podem ainda ser organizados em discretos ou contínuos ([FRIZZARINI, 2013](#)).

- Atributos discretos: são valores que estão contidos em um conjunto finito e que podem ser contados, por exemplo: número de crianças na família.

- Atributos contínuos: são valores que estão contidos nos números reais, ou seja, podem assumir qualquer valor dentro de um intervalo. Os valores são obtidos através de medição ou observação, por exemplo: temperatura.

No contexto de classificação supervisionada, um conjunto de treinamento é um conjunto de N instâncias (observações), representadas por um vetor de atributos e sua classe. Um exemplo de conjunto de treinamento é apresentado na Figura 3. Nele são apresentadas como características a presença ou ausência de ORF, os conteúdos de guanina e citosina (GC) e de adenina e timina (AT) para sequências nucleotídicas pertencentes as classes codificante e não-codificante. Cada instância (observação) é representada por uma linha na tabela. Os atributos são representados como os valores que compõe as características.

Figura 3 – Exemplo de conjunto de treinamento

GC(%)	AT(%)	ORF	Tipo de RNA
40	60	presente	codificante
45	55	presente	codificante
44	56	ausente	não-codificante
60	40	presente	não-codificante
41	59	presente	codificante

→ **Exemplo**

Atributos
Classe

Fonte: Clebiano da Costa Sá, 2018

Na Figura 4 é representado o processo de construção de um classificador baseado em aprendizado supervisionado. A partir das instâncias coletadas é construído o conjunto de treinamento, o qual é utilizado pelo algoritmo de indução para construir o classificador. Esse classificador é então utilizado para prever a classe de cada nova instância (representada por seu vetor de atributos).

Figura 4 – Representação do processo de aprendizado supervisionado

Fonte: Clebiano da Costa Sá, 2018

Uma etapa primordial para construção de um classificador é a extração de características. É a partir deste processo que são obtidas as informações que representarão as instâncias. Geralmente, as características a serem extraídas são definidas a partir do conhecimento prévio que se tem das instâncias de estudo.

Uma vez extraídas, precisam ser selecionadas as características com maior poder discriminante para proporcionar uma classificação mais acurada e rápida. Os principais métodos de seleção de características são descritos na Seção 3.3.1.

Há vários modelos e técnicas distintas para a construção de classificadores supervisionados. Neste projeto, consideraremos inicialmente dois métodos principais de classificação: Máquinas de Vetores de Suporte (*Support Vector Machines-SVM*) e Floresta Aleatória (*Random Forest-RF*) (vide seções 3.3.2 a 3.3.4). A escolha das SVMs decorre do fato da maioria dos métodos de classificação de RNAs conhecidos serem baseados nessa abordagem (KONG et al., 2007; SUN et al., 2013; WANG et al., 2013; LI; ZHANG; ZHOU, 2014). A abordagem RF, por sua vez, é baseada em algoritmos conceitualmente simples e tem apresentado bom desempenho em comparação com outros métodos (CARUANA; NICULESCU-MIZIL, 2006).

3.3.1 Seleção de características

Em determinados problemas de classificação pode haver grandes quantidades de características, algumas das quais redundantes ou mesmo irrelevantes ao problema. Características irrelevantes aumentam o espaço de busca e dificultam a generalização. Além disso,

mesmo características relevantes, quando redundantes, acrescentam uma complexidade desnecessária ao modelo (GHEYAS; SMITH, 2010). Portanto, a etapa de seleção de características é fundamental para obtenção do subconjunto de características que fornece o melhor poder de discriminação entre as classes consideradas. Os principais objetivos desta estratégia são: melhorar o desempenho da predição do classificador; fornecer classificadores mais rápidos e de baixo custo computacional; produzir modelos de classificação simples (por exemplo, árvores de decisão pequenas); e permitir aos pesquisadores concentrar esforços em conjuntos de características relevantes, possibilitando uma melhor visão dos processos descritos pelos dados (SAEYS, 2004).

No contexto da classificação, as técnicas de seleção de características podem ser organizadas em três categorias, de acordo com o tipo de combinação entre a seleção de características e a construção do modelo de classificação: Filtro, Camada e Embutido. Técnicas de seleção do tipo Filtro são aplicadas antes e de forma separada da construção do classificador final. Técnicas do tipo Camada utilizam o classificador no processo de avaliação dos subconjuntos de características. Técnicas do tipo Embutido, como o próprio nome sugere, estão embutidas na construção do classificador (SAEYS; INZA; LARRANAGA, 2007; SAEYS, 2004).

3.3.2 SVM linear e SVM não linear

Máquinas de Vetores de Suporte (SVMs, do inglês *Support Vector Machines*) é uma técnica fundamentada na teoria da aprendizagem estatística (VAPNIK, 2000). Foi originalmente desenvolvida para classificação binária em problemas linearmente separáveis, buscando a separação máxima das instâncias a partir de um hiperplano, utilizado como superfície de decisão (BEN-HUR et al., 2008; BURGESS, 1998).

Uma SVM é construída a partir de um conjunto de treinamento rotulado $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)$, no qual $\mathbf{x}_i$ representa um vetor de características de dimensão p ($\mathbf{x}_i \in \mathbb{R}^p$) e $y_i \in \{+1, -1\}$ representa o rótulo da classe da i -ésima instância. As instâncias podem ser separadas por um hiperplano $f(x) = (\mathbf{w} \cdot \mathbf{x}) - b$, no qual $\mathbf{w}$ é o vetor p -dimensional perpendicular ao hiperplano e b é o termo “compensador”. A distância entre este hiperplano e as instâncias mais próximas pode ser definida como margem, sendo estas instâncias definidas como vetores de suporte (Figura 5) (VAPNIK, 2000).

Figura 5 – Representação do hiperplano ótimo, da margem e dos vetores de suporte em um problema de classificação binária

Fonte: Clebiano da Costa Sá, 2018

O hiperplano ótimo, $\mathbf{w} \cdot \mathbf{x} + b = 0$, é aquele que maximiza a margem (Figura 5). Em um problema de classificação binária este hiperplano particiona o espaço de características em duas regiões: $\mathbf{w} \cdot \mathbf{x} + b > 0$ e $\mathbf{w} \cdot \mathbf{x} + b < 0$. Estas regiões/classes assumem os respectivos rótulos $+1$ e -1 . Quanto maior a margem, menor será o erro de classificação esperado sobre novas instâncias. Portanto, ao treinar uma SVM o objetivo é obter os valores de $\mathbf{w}$ e b que maximizam a margem (TARCA et al., 2007; BEN-HUR et al., 2008).

Na maioria dos problemas reais, os elementos não podem ser separados por uma superfície de decisão linear. Uma alternativa para estes casos seria aplicar uma transformação $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^q$ no espaço de características, para tornar os dados, neste novo espaço de dimensão q ($q > p$), linearmente separáveis. No entanto, a aplicação direta da transformação ϕ pode tornar o treinamento das SVMs computacionalmente intratável se a dimensão do espaço transformado $\mathbb{R}^q$ cresce muito rapidamente com relação ao espaço $\mathbb{R}^p$ (por exemplo, na transformação polinomial, usualmente adotada com SVMs, $q = O(2^p)$). Um importante resultado na teoria subjacente às SVMs é que o problema de otimização utiliza as instâncias de treinamento somente para computar os produtos internos entre os vetores de características, $\mathbf{x}_i \cdot \mathbf{x}_j$, $i, j \in \{1 \dots N\}$. Logo, se for definida uma função *kernel* $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$, não é necessário aplicar explicitamente a transformação ϕ sobre cada instância individualmente, mas apenas computar a matriz de valores $K(\mathbf{x}_i, \mathbf{x}_j)$ para todos os pares de vetores $(\mathbf{x}_i, \mathbf{x}_j)$ do conjunto de treinamento *antes* de resolver o problema de otimização (BURGES, 1998).

Há vários *kernels* que podem ser utilizados, cada um com parâmetros a serem escolhidos (TARCA et al., 2007). Para a construção de classificadores de RNAs codificantes

e não codificantes tem sido utilizado predominantemente o *kernel Radial Basis Function* (RBF) (Equação 1).

Durante a criação de um classificador baseado em SVMs deve-se calibrar o parâmetro de custo (C), que controla a penalidade dos erros (TARCA et al., 2007), e os parâmetros relacionados ao *kernel*, a exemplo do parâmetro gama (γ), que controla o raio do *kernel* RBF. Ao treinar uma SVM com *kernel* RBF a escolha dos valores adequados para os parâmetros C e γ é crucial para a performance da SVM.

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (1)$$

3.3.3 Árvores de classificação

Árvores de classificação são árvores que classificam instâncias com base nos valores de suas características (QUINLAN, 1986; KOTSIANTIS, 2007). Estas árvores apresentam uma estrutura hierárquica que se inicia pelo nó raiz, passa pelos nós internos e termina nos nós resposta (folhas). Os nós raiz e internos são rotulados com atributos (características observadas no conjunto de treinamento). As folhas são rotuladas com classes. E, os ramos, responsáveis pela ligação entre os nós, são rotulados com valores (atributos categóricos) ou com intervalos (atributos numéricos). Quando um nó antecede a outro recebe o nome de nó pai, enquanto os que sucedem um determinado nó são denominados nós filhos (QUINLAN, 1986).

Uma vez construídas, as árvores de classificação formam conjuntos de regras de decisão para classificar novas instâncias. Para que o sistema atribua uma classe a estas instâncias é seguido o seguinte processo: a partir do vetor de características da instância é realizado o teste no atributo raiz, segue para o ramo correspondente, realiza o teste no segundo nó, segue para o próximo ramo correspondente, sucessivamente até atingir o nó folha. A classe atribuída à instância será o rótulo da referida folha. Na Figura 6 é apresentado a representação de um exemplo de uma árvore de classificação, construída a partir dos dados que representam o comportamento relacionado à leitura de e-mails de um dado usuário. O primeiro nó dessa árvore (nó raiz) testa o atributo **autor**, ou seja, se o autor do e-mail é ou não conhecido pelo usuário. Desse nó partem dois ramos correspondentes às duas respostas possíveis a esse teste: conhecido ou desconhecido. O

segundo nó (nó interno) testa o atributo **assunto**, indicando se a mensagem é referente a um assunto novo ou antigo. Os nós terminais são rotulados com as duas classes possíveis, correspondentes ao local em que o usuário lê o e-mail, se em casa (**Sim**) ou no próprio trabalho (**Não**) (MONARD; PRATI, 2005).

Figura 6 – Estrutura básica de uma árvore de decisão

Fonte: Maria Carolina Monard, 2005 (MONARD; PRATI, 2005)

A técnica mais utilizada para construção (ou indução) de árvores de decisão é a *TDIDT* (*Top-Down Induction of Decision Tree*). Normalmente os algoritmos baseados nesta técnica selecionam o atributo com maior poder de discriminação das classes; particionam os dados em dois ou mais subconjuntos disjuntos, de acordo com os valores desse atributo; verifica-se se todas as instâncias de cada subconjunto são da mesma classe, ou se um critério de parada foi atingido, ou ainda, se nenhum atributo foi encontrado, então, associa-se a esse subconjunto uma folha rotulada com a classe predominante entre as instâncias. Se a condição de parada não é satisfeita, então associa-se ao subconjunto um nó interno, e retorna-se ao primeiro passo, em que ocorre a seleção de atributos (QUINLAN, 1986).

Árvores muito gerais (muito pequenas) e excessivamente especializadas (muito grandes) apresentam, respectivamente, subajuste (*underfitting*) ou superajuste (*overfitting*) aos dados de treinamento, implicando em uma menor capacidade de prever corretamente novas instâncias. Este problema pode ser evitado a partir da pré-poda e pós-poda da árvore. A pré-poda consiste na interrupção do crescimento da árvore a partir do estabelecimento de regras de parada rígidas. Na pós-poda, uma vez que a árvore já tenha sido completamente

construída, é avaliada a confiabilidade de todas as suas sub-árvores, removendo (podando) os ramos considerados não confiáveis (QUINLAN, 1986).

3.3.4 Random Forest

Uma *Random Forest* (ou floresta aleatória) consiste em um conjunto de árvores de classificação, cuja combinação forma um preditor agregado que, a partir de um sistema de votação, é capaz de identificar a classe de novas instâncias não rotuladas (BREIMAN, 2001).

A robustez do método é baseada na construção de árvores com baixa correlação (similaridade) entre si, o que é possível através de dois mecanismos distintos (CRIMINISI; SHOTTON; KONUKOGLU, 2011). O primeiro consiste em construir cada árvore a partir de uma amostra *bootstrap*, ou seja, uma sub-amostra do conjunto de treinamento original (e de mesmo tamanho deste) obtida por sorteio simples com reposição. Pelas suas características, nas amostras *bootstrap* quase sempre há repetição de algumas instâncias e ausência de outras. O segundo mecanismo para garantir a baixa correlação entre as árvores consiste em, a cada nó a ser particionado, ao invés de se testar todos os atributos, testa-se apenas um subconjunto aleatório de atributos. O número de atributos sorteados em cada nó, denotado por *mtry*, é um parâmetro que deve ser adequadamente calibrado. Outra característica importante é que as árvores crescem sem interrupções, já que não há poda (ARCHER; KIMES, 2008).

3.4 Avaliação e seleção de modelos

Durante o desenvolvimento de um classificador é imprescindível a avaliação e seleção dos modelos. Este procedimento permite compreender o real potencial e limitações de cada modelo, bem como, estimar seu erro e definir o mais adequado ao problema.

Um algoritmo comumente utilizado é a validação cruzada, um método estatístico de avaliação e comparação de algoritmos de aprendizado. Neste método a capacidade de generalização do modelo, ou seja, a capacidade do classificador prever corretamente a classe de novas instâncias, é estimada pela divisão do conjunto de dados em subconjuntos de treinamento e de teste. Entre as diversas formas de validação cruzada, a mais comumente adotada é a validação cruzada *k-fold* (REFAEILZADEH; TANG; LIU, 2009).

Na validação cruzada k -vezes (ou k -fold), ilustrado na Figura 7, o conjunto original de dados é particionado aleatoriamente em k partes disjuntas, de tamanho aproximadamente igual (WONG, 2015). Estas partições são denominadas *folds*.

O classificador é treinado a partir de um conjunto de treinamento composto por $k - 1$ *folds*, e o *fold* restante compõe o conjunto de validação, que é submetido à classificação. Este processo é repetido k vezes até que todos os *folds* tenham sido utilizados no teste. A Figura 7 ilustra este método para $k = 3$.

Figura 7 – Esquema do processo de validação cruzada com $k = 3$

Fonte: Clebiano da Costa Sá, 2018

Um outro tipo de validação é o *hold-out*. Este método consiste em dividir os dados em duas partes, sendo que uma é utilizada para treinamento do classificador e a parte restante para teste. A proporção destes conjuntos pode variar de acordo com o problema de pesquisa.

Uma vez preditas as classes das instâncias de teste, o modelo é avaliado a partir de estimativas de desempenho. São exemplos a área sob a curva ROC (AUC), acurácia e demais medidas derivadas a partir de uma matriz de confusão, conforme descrito a seguir.

Os erros e acertos do classificador durante sua validação podem ser visualizados na forma de matriz de confusão M , na qual cada posição $M(i, j)$ representa a frequência de instâncias de classe verdadeira c_j que foram classificadas como pertencentes à classe c_i . A Tabela 1 representa uma matriz de confusão para um problema de classificação binária (com apenas duas classes). Neste caso, em especial, uma classe é designada como positiva (p) e a outra como negativa (n). A princípio quatro valores podem ser observados. Nas células $M(p, p)$ e $M(n, n)$ estão, respectivamente, os números de predições corretas para as classes positiva (Verdadeiros Positivos (VP)) e negativa (Verdadeiros Negativos (VN)); já nas células $M(p, n)$ e $M(n, p)$ estão, respectivamente, as predições incorretas para as

classes positiva (Falsos Positivos (FP)) e negativa (Falsos Negativos (FN)). A partir destes valores outras medidas podem ser derivadas, como as descritas a seguir:

- Acurácia: $(VP+VN)/(VP+VN+FP+FN)$. Mensura a proporção de acertos sobre o número total de instâncias testadas;
- Taxa de erro total: $(FP+FN)/(VP+VN+FP+FN)$. Mensura a proporção de erros sobre o número total de instâncias testadas;
- Sensibilidade = Taxa de verdadeiros positivos = *Recall*: $VP/(VP+FN)$. Mensura a proporção de positivos corretamente classificados;
- Especificidade = Taxa de verdadeiros negativos: $VN/(VN+FP)$. Mensura a proporção de negativos corretamente classificados;
- Valor preditivo positivo = Precisão: $VP/(VP+FP)$. Mensura a proporção de verdadeiros positivos sobre o número total de positivos preditos pelo modelo;
- Valor preditivo negativo: $VN/(VN + FN)$. Mensura a proporção de verdadeiros negativos sobre o número total de negativos preditos pelo modelo;
- Medida F1: $2 \cdot \text{Sensibilidade} \cdot \text{Precisão} / (\text{Sensibilidade} + \text{Precisão})$. Esta medida permite a análise combinada entre Sensibilidade e Precisão.

Tabela 1 – Representação da matriz de confusão para um problema de classificação binária

		Classe verdadeira	
		Positiva	Negativa
Classe predita	Positiva	$M(p,p)$	$M(p,n)$
	Negativa	$M(n,p)$	$M(n,n)$

Fonte: Clebiano da Costa Sá, 2018

Outra maneira de se quantificar a eficiência de uma técnica ou algoritmo de classificação é pela área sob a curva ROC (AUC, *Area Under the Curve*). Os gráficos ROC, ou simplesmente curva ROC, (*Receiver operating characteristics*) são gráficos bidimensionais nos quais a taxa de verdadeiros positivos (TVP) é plotada no eixo vertical (ordenada) e a taxa de falsos positivos (TFP) é plotada no eixo horizontal (abscissa), ao considerar-se diferentes valores de limiares ou de parâmetros. Diferentes limiares de classificação produzem diferentes pares (taxa de verdadeiros positivos, taxa de verdadeiros negativos), e cada par corresponde a um ponto na chamada curva ROC (FAWCETT, 2006). A melhor técnica ou algoritmo é aquele cuja área sob a curva ROC (AUC) seja maior. Em casos em que se possui apenas uma classificação (ou seja, quando o classificador

não é probabilístico), a AUC é estimada através da acurácia ponderada (SOKOLOVA; JAPKOWICZ; SZPAKOWICZ, 2006), $AUC = (\text{sensibilidade} + \text{especificidade})/2$.

4 Estado da arte na área de pesquisa

Neste capítulo é apresentada uma revisão sistemática dos classificadores baseados em aprendizado supervisionado desenvolvidos com o propósito de classificar os transcritos em codificante ou não codificante a partir de suas características. Na Seção 4.1 é apresentada uma análise global dos 12 artigos selecionados. Na Seção 4.2 os classificadores são descritos individualmente, organizados em categorias considerando o algoritmo de indução em que são baseados. E, na Seção 4.3 são apresentadas as discussões e considerações finais. O protocolo da revisão sistemática que descreve os métodos utilizados para a realização da mesma é apresentado no Apêndice A.

4.1 Análise global

Na Tabela 2 são apresentados os 12 classificadores obtidos nesta revisão. Esta sumarização aponta que em 2006 foi criado o primeiro classificador, CONC (*Coding Or Non-Coding*) (LIU; GOUGH; ROST, 2006), destinado a classificar os RNAs em codificante ou não codificante de proteína utilizando aprendizado supervisionado e as características da sequência transcrita. No decorrer dos anos outros classificadores foram desenvolvidos na tentativa de melhorar a discriminação destas classes em algum aspecto. De 2006 até 2013 foram publicados apenas classificadores baseados no algoritmo SVM (KONG et al., 2007; SUN et al., 2013). Até então, o algoritmo *Naive Bayes* foi testado, mas não superou o desempenho do SVM. Em 2013 surgiu o primeiro classificador baseado em Regressão Logística (WANG et al., 2013) e em 2015 os primeiros classificadores baseados em *Random Forest* (ACHAWANANTAKUN et al., 2015) e *Deep Learning* (FAN; ZHANG, 2015).

A Figura 8 mostra a evolução da quantidade de publicações desta área de pesquisa ao longo dos anos. Dos 12 classificadores publicados no período de 2006 a 2015, nove foram publicados nos três últimos anos, indicando um forte interesse neste tema mais recentemente (LI; ZHANG; ZHOU, 2014; ACHAWANANTAKUN et al., 2015; FAN; ZHANG, 2015).

Tabela 2 – Sumário de todos os artigos analisados. Informações não mencionadas são descritas com NM.

Classificador	Ano	Algoritmo indutor	Outros algoritmos testados	Dados para validação / treinamento	Acurácia da validação (%)	Tipos de características	Referências
CONC	2006	SVM	Naive Bayes	5.610 cDNAs codificantes eucarióticos do GenBank com <i>link</i> para proteínas anotadas em Swiss-Prot. 778 ncRNAs eucarióticos do RNAdb e 2.178 ncRNAs eucarióticos do NONCODE.	NM	Propriedades da proteína predita; Buscas por similaridade em bancos de dados; Composição nucleotídica	(LIU; GOUGH; ROST, 2006)
CPC	2007	SVM	NM	Mesmo conjunto de dados utilizado por CONC.	95.77	Estatísticas da ORF; Buscas por similaridade em bancos de dados	(KONG et al., 2007)
PORTRAIT	2009	SVM	Naive Bayes	234.112 cDNAs codificantes procarióticos e eucarióticos obtidos a partir do EMBL com <i>link</i> para proteínas anotadas com alta qualidade em Swiss-Prot. 213.849 ncRNAs obtidos a partir do RNAdb, NONCODE e Rfam. Todos esses cDNAs e ncRNAs formam um conjunto de dados principal, que foi dividido em outros dois conjuntos de acordo com a presença ou ausência de ORF predita, um para cada modelo de classificação.	92.1	Estatísticas da ORF; Composição nucleotídica; Propriedades da proteína predita; Tamanho do transcrito	(ARRIAL; TOGAWA; BRIGIDO, 2009)

Tabela 2 – (Continuação)

Classificador	Ano	Algoritmo indutor	Outros algoritmos testados	Dados para validação / treinamento	Acurácia da validação (%)	Tipos de características	Referências
CNCI	2013	SVM	NM	30.507 mRNAs humanos do RefSeq e 18.566 lncRNAs humanos do GENCODE	97.3	Predição da sequência codificante (CDS) utilizando matriz de <i>scores</i> das frequências de uso dos trinucleotídeos	(SUN et al., 2013)
iSeeRNA	2013	SVM	NM	2.583 mRNAs e 2.594 lincRNAs humanos e 424 lincRNAs e 465 mRNAs de ratos obtidos a partir do projeto HAVANA.	95.4 (humanos); 94.2 (ratos)	Estatísticas da ORF; Composição nucleotídica; Conservação baseada em resolução	(SUN et al., 2013)
CPAT	2013	Regressão Logística	NM	10.000 mRNAs selecionados aleatoriamente a partir do RefSeq. Todos com sequências de proteínas anotadas com alta qualidade pelo projeto <i>Consensus Coding Sequence</i> . 10.000 ncRNAs selecionados aleatoriamente a partir do GENCODE.	96.7	Estatísticas da ORF; Viés do uso de Hexamer; Linguística simples	(WANG et al., 2013)

Tabela 2 – (Continuação)

Classificador	Ano	Algoritmo indutor	Outros algoritmos testados	Dados para validação / treinamento	Acurácia da validação (%)	Tipos de características	Referências
CNCT-Discriminator	2013	SVM	NM	30.112 mRNAs e 13.876 ncRNAs do GENCODE, 6.749 mRNAs do RefSeq e 14.121 ncRNAs do Ensembl. Dados de expressão foram obtidos de experimentos de RNA-seq com cérebro e outras linhagens de células humanas	81.53	Estatísticas da ORF; Composição nucleotídica; Estrutura secundária do transcrito; Expressão do transcrito	(BISWAS et al., 2013)
PLEK	2014	SVM	NM	22.389 mRNAs humanos do RefSeq e 22.389 lncRNAs humanos do GENCODE	95.6	Frequências de uso de k-mer	(LI; ZHANG; ZHOU, 2014)
RNAcon	2014	SVM	NM	31.237 mRNAs obtidos a partir do RefSeq. 20.453 ncRNAs obtidos a partir do Rfam.	NM	Composição nucleotídica	(PANWAR; ARORA; RAGHAVA, 2014)
LncRNA-ID	2015	Random Forest	Regressão Logística	Conjunto H1: 15.308 mRNAs e 4586 lncRNAs humanos selecionados aleatoriamente a partir do GENCODE. Conjunto M1: 22.033 mRNAs e 2457 lncRNAs de ratos selecionados aleatoriamente a partir do GENCODE. Conjunto H2: mesmo dataset do CPAT (9.929 mRNAs e 9.066 ncRNAs humanos).	NM	Estatísticas da ORF; Interação com ribossomos; Alinhamentos baseados em perfis de modelos de cadeias ocultas de Markov (<i>profile HMM</i>)	(ACHAWANANTAKUN et al., 2015)

Tabela 2 – (Continuação)

Classificador	Ano	Algoritmo indutor	Outros algoritmos testados	Dados para validação / treinamento	Acurácia da validação (%)	Tipos de características	Referências
lncRScan-SVM	2015	SVM	NM	5.000 mRNAs e 5.000 lncRNAs humanos obtidos a partir do GENCODE. 25.000 mRNAs e 25.000 lncRNAs de ratos obtidos a partir do GENCODE.	NM	Composição nucleotídica; Predição de CDS; Estrutura do gene	(SUN et al., 2015)
lncRNA-MFDL	2015	Deep Learning	SVM	10.000 mRNAs humanos obtidos a partir do Refseq. 10.000 lncRNAs humanos obtidos a partir do GENCODE.	97.1	Estatísticas da ORF; Frequências de uso de k-mer; Estrutura secundária; Predição da região mais parecida com uma CDS (MLCDS)	(FAN; ZHANG, 2015)

Fonte: Clebiano da Costa Sá, 2018

Figura 8 – Número de classificadores publicados em todos os anos

Fonte: Clebiano da Costa Sá, 2018

A Figura 9 mostra que 75% (9) dos classificadores são baseados em SVM, 8% (1) em Regressão Logística, 8% (1) em *Random Forest* e 8% (1) em *Deep Learning*. Portanto, há uma predominância de classificadores baseados em SVM, principalmente de 2006 a 2013 (Tabela 2).

Figura 9 – Uso dos algoritmos de classificação pelos classificadores

Fonte: Clebiano da Costa Sá, 2018

A Figura 10 mostra que os bancos de dados RefSeq e GENCODE foram os mais utilizados para a construção dos conjuntos de validação ou de treinamento, sendo utilizados em sete classificadores. RNadb e NONCODE foram utilizados em três classificadores e GenBank e Rfam em dois classificadores. EMBL, Ensembl e HAVANA foram utilizados em apenas um classificador.

Figura 10 – Uso das fontes dos transcritos que compõe os conjuntos de dados utilizados na validação ou no treinamento dos classificadores

Fonte: Clebiano da Costa Sá, 2018

Todos os classificadores foram construídos a partir de múltiplas características e avaliados a partir do método de validação cruzada *k-fold* (10x). Apenas quatro artigos (SUN et al., 2013; PANWAR; ARORA; RAGHAVA, 2014; FAN; ZHANG, 2015; SUN et al., 2015) mencionam explicitamente o uso de métodos para seleção de características e apenas os classificadores CPC, PORTRAIT, CNCI, iSeeRNA, CPAT, PLEK e lncRNA-MFDL tiveram a acurácia geral de sua validação disponibilizada. Estes classificadores apresentaram altas acurácias relatadas pelos autores, variando de 92,1% a 97,3% (Tabela 2). No entanto, cabe destacar que a maioria destes valores não são comparáveis por terem sido obtidos utilizando diferentes amostras de treinamento e teste, sendo o CONC e CPC a única exceção.

Buscas no site www.researchgate.net, uma ferramenta para compartilhamento de artigos científicos e discussões, apontam que até o final de agosto de 2017 o artigo do classificador CPC foi o mais citado dentre aqueles aqui revisados (Figura 11). Nos dois últimos anos os artigos mais citados foram, respectivamente, CPC, CPAT, CNCI e PLEK (Figura 12).

Figura 11 – Total de citações dos artigos de 2006 a agosto de 2017

Fonte: Clebiano da Costa Sá, 2018

Figura 12 – Acumulado de citações dos artigos de 2016 a agosto de 2017

Fonte: Clebiano da Costa Sá, 2018

4.2 Análise qualitativa

Nesta Seção todos os classificadores são brevemente descritos, organizados pelo tipo de algoritmo de indução em que são baseados.

4.2.1 Classificadores baseados em SVM

Criado em 2006, CONC ([LIU; GOUGH; ROST, 2006](#)) foi o primeiro classificador de transcritos em codificante ou não codificante (Tabela 2), sendo baseado em SVM. Nele, os RNAs são classificados com base nas características da proteína predita a partir de sua sequência transcrita. Seu conjunto de características incluem o comprimento do peptídeo, a composição de aminoácidos, o conteúdo da estrutura secundária predita, a porcentagem predita de resíduos expostos, a entropia composicional, o número de homólogos a partir de pesquisas em banco de dados e a entropia de alinhamento. Há ainda características baseadas na frequência dos nucleotídeos.

No CPC (*Coding Potential Calculator*) ([KONG et al., 2007](#)) foi utilizado um conjunto reduzido de características com alto poder discriminante. Dentre elas, três características são obtidas diretamente da sequência nucleotídica e outras três características são provenientes de buscas por similaridade utilizando-se o programa BLASTX sobre um banco de proteínas. Desta forma, CPC proporcionou maior acurácia e rapidez. No entanto, a classificação com base na identificação de homólogos em bancos de proteínas pode errar a predição para transcritos codificantes de proteínas novas (ausentes em bancos de proteínas) e ncRNAs que facilmente se alinham com algumas proteínas hipotéticas mal anotadas nos bancos de proteínas. E, embora tenha avançado bastante na rapidez da classificação continua proibitivo para laboratórios com recursos limitados ([ARRIAL; TOGAWA; BRIGIDO, 2009](#); [WANG et al., 2013](#)).

Os classificadores CONC e CPC não são adequados para classificar sequências de espécies pouco estudadas, já que sua alta precisão depende da qualidade da informação de homologia e que ambos esperam sequências completas ([ARRIAL; TOGAWA; BRIGIDO, 2009](#)). Para contornar esse problema surgiu PORTRAIT ([ARRIAL; TOGAWA; BRIGIDO, 2009](#)), um classificador baseado em dois modelos: um dependente de proteína e outro independente de proteína. O conjunto de características do modelo dependente de proteína incluem composição nucleotídica, composição de aminoácidos, tamanho da ORF, ponto isoeletrico, entropia composicional e hidropatia média. Já o conjunto de características do modelo independente de proteína incluem apenas a composição nucleotídica e o tamanho do transcrito. Quando uma proteína putativa é encontrada na sequência alvo utiliza-se o primeiro modelo, caso contrário, utiliza-se o segundo modelo. Desta forma, mesmo que a identificação da ORF esteja comprometida a classificação pode ser realizada pelo modelo

independente de proteína. Portanto, este método não está propenso a classificar como ncRNAs aqueles transcritos que codificam proteínas novas, raras ou ausentes em bancos de dados, o que pode ocorrer em classificadores como CONC e CPC.

Ao contrário do que acontece nos classificadores CONC e CPC, iSeeRNA (SUN et al., 2013) foi desenvolvido utilizando uma coleção de linRNAs para representar a classe não codificante. Seu conjunto de características é composto por três categorias: conservação da sequência nucleotídica, estatísticas da ORF e composição nucleotídica. Quando comparado ao CPC, iSeeRNA apresentou alta acurácia para lincRNAs e foi várias ordens de magnitude mais rápido.

CNCTDiscriminator (*Coding and Noncoding Transcript Discriminator*) (BISWAS et al., 2013) é um sistema formado pela integração de quatro modelos. Cada modelo foi treinado a partir de uma categoria de características — estatísticas da ORF, composição da sequência nucleotídica, propriedades da estrutura secundária predita e uma quarta categoria inédita, baseada no nível de expressão dos mRNAs e ncRNAs. Esta última categoria é composta pelos *scores* RPKM (*Read Per Kilobases Per Millions short reads*), que são *scores* de expressão normalizados para um dado transcrito em várias linhagens celulares diferentes observadas em experimentos de RNA-seq.

O classificador CNCI (*Coding-Non-Coding Index*) (SUN et al., 2013) utiliza características baseadas apenas na composição intrínseca a sequência. Portanto, é potencialmente aplicável a uma variedade de espécies com informações pobremente anotadas. Seu conjunto de características é obtido a partir das regiões mais parecidas com um domínio codificante MLCDS (*most-like CDS*), preditas para todas as possíveis ORFs do transcrito. MLCDS são preditas utilizando um algoritmo de programação dinâmica e uma matriz de *score* (64x64), criada a partir da razão dos *logs* (*log-ratio*) das frequências de uso de todos os tipos de tripletes de nucleotídeos adjacentes pelas seis possíveis ORFs de mRNAs e ncRNAs. Quando comparado ao CPC, CNCI mostrou melhor desempenho, especialmente para a classificação de transcritos incompletos e pares de transcritos *sense-antisense*. Sua grande falha está na classificação de sequências com erros de sequenciamento (inserções ou deleções).

A classificação de sequências com erros de sequenciamento foi melhorada com PLEK (*predictor of long non-coding RNAs and messenger RNAs based on an improved k-mer scheme*) (LI; ZHANG; ZHOU, 2014). Um método livre de alinhamento e robusto a erros de sequenciamento, que distingue rapidamente lncRNAs e mRNAs, na ausência de sequência

genômica ou anotações. PLEK classifica os RNAs a partir das características obtidas por um *pipeline* computacional baseado em um esquema *k-mer* melhorado. Além de ser mais acurada e robusta, esta ferramenta é aproximadamente oito vezes mais rápida que CNCI e 244 vezes mais rápida que CPC.

Em RNAcon (PANWAR; ARORA; RAGHAVA, 2014) os RNAs são classificados com base em apenas uma categoria de características, a composição nucleotídica. Seu conjunto de características é composto pela frequência de ocorrência de 14 trinucleotídeos selecionados.

Já o classificador lncRScan-SVM (SUN et al., 2015) classifica RNAs pela integração de características obtidas a partir da conservação da sequência, da composição nucleotídica, da estrutura do gene (tais como, tamanho dos éxons e porcentagem dos éxons) e da ocorrência dos *stop códons* (por exemplo, desvio padrão da contagem de *stop códons* nas três *frames*).

4.2.2 Classificadores baseados em Regressão Logística

O classificador CPAT (*Coding Potential Assessment Tool*) (WANG et al., 2013) utiliza um modelo de regressão logística e apenas quatro características da sequência: tamanho da ORF, cobertura da ORF, viés do uso de *hexamer* e estatística *Fickett TESTCODE*. A estatística *Fickett TESTCODE* trata-se de uma característica linguística simples que distingue mRNAs e ncRNAs de acordo com o efeito combinatório da composição de nucleotídeos e o viés da utilização de códons. Este score foi selecionado porque captura simultaneamente o viés composicional e assimetria da posição.

4.2.3 Classificadores baseados em Random Forest

O classificador LncRNA-ID (ACHAWANANTAKUN et al., 2015) tem como base um algoritmo *Random Forest* balanceado (BRF). Este algoritmo balanceia as classes a partir de um conjunto de *Random Forests* induzidas por múltiplos subconjuntos balanceados pela sub-amostragem da classe majoritária. Esta sub-amostragem é realizada de tal forma que não há perda de informação. LncRNA-ID avalia o potencial codificante de um transcrito a partir das características de sequências de ORFs putativas, a partir da quantificação dos principais atributos da sequências nucleotídica relacionados a interação do ribossomo

com mRNAs e a partir da conservação contra perfis de modelos ocultos de *Markov* (*profile* HMM) de famílias proteicas caracterizadas. A aplicação de buscas por similaridade baseadas em perfis HMM de famílias proteicas possibilita algumas vantagens ao LncRNA-ID. Primeiro, o número de famílias é menor que o número de sequências de proteínas conhecidas, possibilitando buscas com maior rapidez. Trabalhos anteriores demonstraram que usar informações de famílias pode melhorar a sensibilidade de buscas por similaridade de proteínas remotas. Para os transcriptomas de organismos não modelos a sensibilidade de buscas por similaridade remota é especialmente importante para identificar possíveis novos homólogos.

4.2.4 Classificadores baseados em Deep Learning

Um dos classificadores mais recentes, o lncRNA-MFDL (FAN; ZHANG, 2015) foi desenvolvido utilizando múltiplas características e o algoritmo *Deep Learning* (LECUN; BENGIO; HINTON, 2015). Seu conjunto de características incluem as categorias: estatísticas da ORF, *k-mer*, estrutura secundária e MLCDS. Quando comparado ao CPC e CNCI, lncRNA-MFDL mostrou melhor desempenho sobre a classificação dos conjuntos de teste em humanos e em outras dez espécies. Embora a acurácia geral da validação relatada pelos autores do CNCI seja um pouco maior que a acurácia relatada pelos autores do lncRNA-MFDL (Tabela 2), estas não são comparáveis porque foram obtidas por validação cruzada em diferentes conjuntos de dados.

4.3 Discussões e considerações finais

A evolução desta área de pesquisa passa pela utilização de diferentes algoritmos de classificação, mas também pela formação de novos conjuntos de dados e principalmente pelo uso de novas características ou adaptações de características já existentes.

As análises global e qualitativa mostram que há uma predominância de classificadores baseados em SVM, mas que estão surgindo classificadores com grande potencial de classificação baseados em outros algoritmos, tais como Regressão Logística, *Random Forest* e *Deep Learning*.

Alguns classificadores como CNCI e PLEK podem ser utilizados na forma de espécies cruzadas. Ou seja, mesmo quando treinados com sequências humanas permitem a

classificação de transcritos de outras espécies de vertebrados com alta acurácia. Felizmente, muitos dos classificadores mais recentes disponibilizam programas para que os usuários construam modelos a partir do seu próprio conjunto de dados (ACHAWANANTAKUN et al., 2015; SUN et al., 2013; LI; ZHANG; ZHOU, 2014; SUN et al., 2015). Esta opção é fundamental para trabalhar com diferentes espécies e possibilita comparações mais justas já que estes classificadores podem ser avaliados por validação cruzada no mesmo conjunto de dados.

Todos os classificadores revisados trabalham com a combinação de múltiplas características. Esta prática proporciona um classificador melhor e mais robusto (com maior especificidade e sensibilidade), evitando erros de características individuais (por exemplo, alinhamento incorreto ou tamanho de peptídeo incorreto) (LIU; GOUGH; ROST, 2006). Há classificadores como RNAcon e CNCI baseados em apenas uma categoria de características. Contudo, a grande maioria dos classificadores são baseados na combinação de diferentes categorias de características.

Algumas das características mais utilizadas pelos classificadores são as características da ORF predita (tais como, tamanho e cobertura) e baseadas na composição nucleotídica (tais como, as frequências de mono, di e trinucleotídeos). Dentre as categorias de características mais recentes estão as características de MLCDS preditas a partir da frequência de uso dos tripletes de nucleotídeos adjacentes pelas seis possíveis ORFs, as frequências de uso *k-mer*, as características da sequência que permitem mensurar a interação do transcrito com o ribossomo, as características baseadas na estrutura do gene (tais como, tamanho dos éxons e porcentagem dos éxons) e na ocorrência dos *stop códons* (tais como, desvio padrão da contagem de *stop códons* nas três *frames*) e as características baseadas em buscas por similaridade a partir do alinhamento do transcrito com perfis HMM de famílias proteicas.

É importante notar também que os diversos conjuntos de dados criados durante o desenvolvimento e a avaliação dos classificadores analisados são uma ótima ferramenta para análises futuras. A partir deles podem ser realizadas análises confiáveis de uma gama de características, algoritmos de classificação e de seleção de características.

As comparações entre classificadores têm sido realizadas com base apenas na classificação de determinados conjuntos de teste. E embora tenham sido avaliados por validação cruzada, a maioria utilizaram diferentes conjuntos de dados, impossibilitando comparações. Portanto, há a necessidade de um trabalho mais abrangente, que avalie todos

esses classificadores utilizando as mesmas amostras em um sistema de validação cruzada. Também seria importante testá-los a partir de diferentes conjuntos de dados que simulem diferentes condições tais como presença de sequências parciais, proteínas raras, erros de sequenciamento, erros de anotação nos bancos, etc., sendo assim possível evidenciar os pontos fortes e fracos de cada ferramenta. Há também a necessidade de um método de avaliação que considere a constante evolução dos bancos de dados ao longo do tempo. Uma boa alternativa seria realizar uma forma de validação similar à de séries temporais, em que se treina o modelo até o instante t e testa-se com as instâncias em $t + 1$. Já o problema de alguns dos classificadores não disponibilizarem programas para a construção de novos modelos poderia ser contornado a partir da reprodução destes classificadores, criando um modelo a partir dos parâmetros e conjunto de treinamento informados no artigo. Desta forma, tais modelos poderiam ser avaliados com os métodos de validação de forma pareada, a partir dos mesmos conjuntos de sequências.

Também é importante o teste de uma maior diversidade de algoritmos de classificação e de seleção de características associados a essa grande diversidade de características de RNAs, no contexto da classificação dos transcritos em codificante ou não codificante de proteína, a fim de buscar resultados ainda melhores.

5 Materiais e métodos

Neste capítulo são apresentados os materiais e métodos utilizados para testar as hipóteses. Na Seção 5.1 é descrito como foi construído e agrupado em *folds* um grande e abrangente conjunto de transcritos humanos codificantes e não codificantes utilizando as datas de lançamento das versões dos bancos de dados. Na Seção 5.2 são descritos os modelos de classificação avaliados. Na Seção 5.3 são apresentados os métodos de validação tradicional e temporal, bem como, a motivação para o uso do método *hold-out*. A Seção 5.4 apresenta como os resultados foram consolidados e como foram realizados os testes das hipóteses.

5.1 Conjunto de dados

Um conjunto representativo e abrangente de sequências nucleotídicas de transcritos humanos, anotadas como codificante ou não codificante de proteína, foi criado de tal forma que representasse a evolução dos bancos de dados ao longo dos últimos quatorze anos.

A partir de buscas exploratórias na *internet* e na literatura científica foram encontrados três bancos de dados públicos que disponibilizam para *download* muitas de suas versões anteriores: RFAM¹(KALVARI et al., 2018), GENCODE²(HARROW et al., 2012) e ENSEMBL³(ZERBINO et al., 2018). O uso dessas versões e suas datas de lançamento possibilitou a obtenção das sequências nucleotídicas e sua informação temporal, representando a evolução do conhecimento acerca dos RNAs ao longo do tempo.

Durante a construção do conjunto de dados o ENSEMBL não foi utilizado por haver dúvidas quanto à confiabilidade das anotações (se codificante ou não) em suas primeiras versões. Além disso, o GENCODE já é um subconjunto de anotações confiáveis provenientes do ENSEMBL e que somado ao RFAM formam um conjunto expressivo, com cerca de 140400 sequências.

Todas as versões dos bancos de dados RFAM e GENCODE disponíveis para *download* foram baixadas automaticamente via *scripts* escritos em *Bash*. Para cada uma dessas versões foi construído, a partir de *scripts* escritos em *Perl*, uma tabela padronizada

¹ RFAM: <http://rfam.xfam.org/>

² GENCODE: <https://www.encodegenes.org/>

³ ENSEMBL: <https://www.ensembl.org/index.html>

contendo os atributos: sequência nucleotídica, nome do banco, versão do banco, data de início da versão, espécie, data de término da versão e se é codificante ou não.

Para construir as tabelas das versões do RFAM foi necessário obter as sequências nucleotídicas a partir de arquivos com extensão *seed* no formato *stockholm*, que são arquivos compostos por alinhamentos múltiplos de sequências que são conhecidamente pertencentes a uma determinada família de RNA. Para utilizar apenas sequências de transcritos de genes de ncRNAs, as tabelas foram construídas utilizando apenas as sequências das famílias RFAM que estão presentes no banco de dados RNACentral e que apresentam, em sua descrição no arquivo *seed*, o tipo de molécula “Gene”. Sequências pertencentes a famílias com a palavra “pseudogene” em sua descrição no arquivo *seed* não foram utilizadas.

Para construir as tabelas das versões do GENCODE foi necessário obter as sequências nucleotídicas a partir dos arquivos no formato *fasta* e a anotação no formato *gtf*. Também foi necessário filtrar apenas sequências confiáveis, as quais apresentam o *status* do transcrito como conhecido ou novo (KNOWN, NOVEL) e que passaram por anotação manual (níveis de anotação 1 e 2). Sequências com a palavra “pseudogene” em sua descrição no arquivo de anotação (no formato *gtf*) não foram utilizadas.

As tabelas de cada versão dos dois bancos foram importadas para um banco de dados criado no gerenciador PostgreSQL. Sequências menores que 80 nucleotídeos foram excluídas do banco de dados. A partir de consultas em SQL, utilizando a própria sequência nucleotídica como identificador, foi selecionada a primeira ocorrência de cada sequência ao longo dos anos. Tais sequências foram agrupadas em *folds* temporais de acordo com o ano em que surgiram pela primeira vez. As sequências dos primeiros 10 anos foram agrupadas devido ao fato destes primeiros anos possuírem individualmente poucas sequências e também por formarem uma base inicial de conhecimento de RNAs para os testes temporais. Por fim, foram definidos sete *folds* temporais (apresentados na Tabela 3). É possível observar uma variação significativa nas proporções das classes ao longo do tempo, a qual originou nossa preocupação inicial a respeito dos resultados obtidos pelas estratégias de avaliação tradicional (atemporal) dos classificadores.

Nas avaliações dos classificadores foram utilizadas 10000 sequências dos dois *folds* maiores (2002-2011 e 2012), selecionadas aleatoriamente. Dos demais *folds* foram utilizadas todas as sequências. Sendo excluídas das análises aquelas que não tiveram suas características extraídas pelos classificadores ou extrator.

Tabela 3 – Quantidade de RNAs codificantes (mRNA) e não codificantes (ncRNA) nos *folds* temporais

Fold	mRNA	ncRNA	Total
2002-2011	65964 (84.01%)	12559 (15.99%)	78523
2012	20570 (65.82%)	10684 (34.18%)	31254
2013	5739 (55.47%)	4608 (44.53%)	10347
2014	5494 (58.61%)	3880 (41.39%)	9374
2015	5866 (73.55%)	2109 (26.45%)	7975
2016	1265 (83.28%)	254 (16.72%)	1519
2017	1287 (87.14%)	190 (12.86%)	1477
Total	106185 (75.59%)	34284 (24.41%)	140469

Fonte: Clebiano da Costa Sá, 2018

Embora não tenham sido rastreadas as mudanças nas distribuições das características ao longo do tempo, a Tabela 3 mostra que as proporções das classes de RNAs codificantes e não codificantes variam entre os *folds* temporais.

5.2 Modelos de classificação

Foram avaliados oito modelos de classificação: quatro consistem nos classificadores de RNAs codificantes e não codificantes mais citados nos dois últimos anos (ver Capítulo 4), um é o CPC2 (KANG et al., 2017), que surgiu mais recentemente e após a revisão bibliográfica sistemática e os outros três são novos modelos propostos a partir da combinação de diferentes algoritmos de classificação e conjuntos de características.

Os classificadores CPAT (WANG et al., 2013) e PLEK (LI; ZHANG; ZHOU, 2014) disponibilizam programas para a construção de novos modelos. Por isso, os modelos *CPAT* e *PLEK* foram construídos a partir dos programas e indicações dos desenvolvedores.

Já os classificadores CPC (KONG et al., 2007), CPC2 (KANG et al., 2017) e CNCI (SUN et al., 2013) não disponibilizam programas para a construção (treinamento) de novos modelos a partir do conjunto de dados do usuário. Por essa razão, simulamos estes classificadores construindo os modelos a partir de suas características extraídas, a partir da calibração dos parâmetros com os métodos descritos na Seção 5.3 e com base nas informações descritas em seus artigos. Estes modelos são, respectivamente, *SVM_CPC1*, *SVM_CPC2* e *SVM_CNCI*.

Tanto para a simulação dos classificadores CPC, CPC2 e CNCI quanto para a construção dos modelos novos *RForestAllFeat*, *SVMrbfAllFeat* e *SVMrbfBestfeatRelief* foi

necessário extrair as características utilizadas pelos classificadores do estado da arte. Na Figura 13 é apresentado um esquema do processo de extração e seleção de características para construção de modelos que simulam classificadores do estado da arte e de modelos novos.

Figura 13 – Processo de extração e seleção de características para construção de modelos que simulam classificadores do estado da arte e de modelos novos

Fonte: Clebiano da Costa Sá, 2018

A primeira etapa deste processo foi a extração das características utilizadas pelos classificadores PLEK, CPAT, CPC1, CPC2 e CNCI. Os valores para estas características foram extraídos a partir da saída dos classificadores, quando realizada a predição da classe de sequências de RNAs transcritos. Sendo, em alguns casos, necessário a implementação de adaptações no código do programa para exportar os valores das características para um arquivo de saída.

Foram extraídas 1364 características do PLEK (Tabela 4), 4 do CPAT (Tabela 5), 9 do CPC1 (Tabela 6), 4 do CPC2 (Tabela 7) e 64 do CNCI (Tabela 8).

Tabela 4 – Características extraídas a partir do classificador PLEK

Origem	Descrição das características	Total de variáveis
PLEK	Frequências de uso de K-mer	1364

Tabela 5 – Características extraídas a partir do classificador CPAT

Origem	Descrição das características	Total de variáveis
CPAT	Tamanho da sequência nucleotídica	1
CPAT	Estatística Fickett TESTCODE	1
CPAT	Viés do uso de hexamers	1
CPAT	Tamanho da ORF	1

Tabela 6 – Características extraídas a partir do classificador CPC1

Origem	Descrição das características	Total de variáveis
CPC1	Hit seq count	1
CPC1	Hit HSP count	1
CPC1	Hit score	1
CPC1	Frame score	1
CPC1	Frame score2	1
CPC1	Tamanho da CDS	1
CPC1	Score	1
CPC1	Used	1
CPC1	Strict	1

Tabela 7 – Características extraídas a partir do classificador CPC2

Origem	Descrição das características	Total de variáveis
CPC2	Fickett score	1
CPC2	Ponto isoelétrico	1
CPC2	Integridade da ORF predita	1
CPC2	Tamanho do peptídeo	1

Tabela 8 – Características extraídas a partir do classificador CNCI

Origem	Descrição das características	Total de variáveis
CNCI	Tamanho da MLCDS	1
CNCI	S-score da MLCDS	1
CNCI	Razão do tamanho da melhor MLCDS pela soma dos tamanhos das seis MLCDS	1
CNCI	Diferença entre o S-score da melhor MLCDS e das outras cinco MLCDS dividida por 5	1
CNCI	Frequências dos trinucleotídeos únicos (có-dons) na MLCDS	61

Além destas, houve o interesse em testar outras características. Por essa razão, também foram extraídas 113 características (Tabela 9) a partir de um novo extrator escrito por Silva (SILVA, 2009b; SILVA, 2009a) em seu trabalho de conclusão de curso, orientado pela *Prof^a. Dr^a*. Ariane Machado Lima (orientadora do presente trabalho). Este extrator foi implementado em *Perl*, sendo executado via linha de comando pelo fato de ter sido criado com o objetivo de ser utilizado como módulo de um *pipeline*⁴. Este conjunto

⁴ *Pipeline* é um sistema formado por vários módulos, onde a saída de um módulo é a entrada de outro, normalmente sem interação com o usuário durante as execuções intermediárias (SILVA, 2009b; SILVA, 2009a).

de características incluem buscas por similaridade em um banco de proteínas utilizando BLASTP e BLASTX, tamanho da sequência, percentual de G e C na sequência, módulo da diferença entre a ocorrência percentual de A e ocorrência percentual de T/U, módulo da diferença entre a ocorrência percentual de G e ocorrência percentual de C, frequência dos possíveis dinucleotídeos na sequência (16 variáveis) e outras características baseadas em uma hipotética maior ORF, sendo elas o tamanho da 5' UTR, tamanho da 3' UTR, tamanho da maior ORF reconhecida, cobertura da ORF, quantidade de *start* códons na 5' UTR, quantidade de *start* códons fora da 5' UTR, distribuição dos aminoácidos (21 variáveis) e dos códons da maior ORF (64 variáveis). Algumas destas características já eram também obtidas por classificadores disponíveis na literatura e outras são provenientes da experiência da *Prof^a* Ariane Machado Lima em estudos com ncRNAs (SILVA, 2009b; SILVA, 2009a).

Tabela 9 – Características extraídas a partir de um novo extrator

Origem	Descrição das características	Total de variáveis
Novo Extrator	E-value retornado pelo Blastp	1
Novo Extrator	E-value retornado pelo Blastx	1
Novo Extrator	Porcentagem de GC	1
Novo Extrator	Diferença entre as porcentagens de A e de T	1
Novo Extrator	Diferença entre as porcentagens de G e de C	1
Novo Extrator	Cobertura da ORF sobre a sequência	1
Novo Extrator	Quantidade de códons de iniciação dentro da UTR 5	1
Novo Extrator	Quantidade de códons de iniciação fora da UTR 5	1
Novo Extrator	Tamanho da UTR 5	1
Novo Extrator	Tamanho da UTR 3	1
Novo Extrator	Frequência dos dinucleotídeos	16
Novo Extrator	Frequência dos trinucleotídeos	64
Novo Extrator	Distribuição dos aminoácidos	21
Novo Extrator	Tamanho da sequência nucleotídica	1
Novo Extrator	Tamanho da ORF	1

A soma de todas as características extraídas, provenientes do novo extrator e dos classificadores do estado da arte, corresponde a um total de 1558 características.

Após a remoção de características consideradas redundantes, com coeficiente de correlação de *Pearson* maior que 0.98, foram mantidas 1551 características (Tabela 10). Estas formam a base dos modelos novos *RForestAllFeat* e *SVMrbfAllFeat*. A diferença entre estes modelos está no algoritmo de indução aplicado. O modelo *RForestAllFeat* foi construído com o algoritmo *Random Forest*, já o modelo *SVMrbfAllFeat* tem como base o algoritmo SVM e o *Kernel* RBF.

Tabela 10 – Conjunto formado por todas as características não redundantes

Origem	Descrição das características	Total de variáveis
CPC1	Hit seq count	1
CPC1	Hit HSP count	1
CPC1	Hit score	1
CPC1	Frame score	1
CPC1	Frame score2	1
CPC1	Tamanho da CDS	1
CPC1	Score	1
CPC1	Used	1
CPC1	Strict	1
CPC2	Fickett score	1
CPC2	Ponto isoelétrico	1
CPC2	Integridade da ORF predita	1
CNCI	Tamanho da MLCDS	1
CNCI	S-score da MLCDS	1
CNCI	Razão do tamanho da melhor MLCDS pela soma dos tamanhos das seis MLCDS	1
CNCI	Diferença entre o S-score da melhor MLCDS e das outras cinco MLCDS dividida por 5	1
CNCI	Frequências dos trinucleotídeos únicos (có-dons) na MLCDS	61
CPAT	Tamanho da sequência nucleotídica	1
CPAT	Estatística Fickett TESTCODE	1
CPAT	Viés do uso de hexamers	1
PLEK	Frequências de uso de K-mer	1364
Novo Extrator	E-value retornado pelo Blastp	1
Novo Extrator	E-value retornado pelo Blastx	1
Novo Extrator	Porcentagem de GC	1
Novo Extrator	Diferença entre as porcentagens de A e de T	1
Novo Extrator	Diferença entre as porcentagens de G e de C	1
Novo Extrator	Cobertura da ORF sobre a sequência	1
Novo Extrator	Quantidade de códon de iniciação dentro da UTR 5	1
Novo Extrator	Quantidade de códon de iniciação fora da UTR 5	1
Novo Extrator	Tamanho da UTR 5	1
Novo Extrator	Tamanho da UTR 3	1
Novo Extrator	Frequência dos dinucleotídeos	12
Novo Extrator	Frequência dos trinucleotídeos	64
Novo Extrator	Distribuição dos aminoácidos	21

A partir das características não redundantes e do método *Relief* (KONONENKO, 1994) foram selecionadas as 27 características mais relevantes (Tabela 15), que serviram de base para o modelo novo *SVMrbfBestfeatRelief*. Este modelo foi construído com o algoritmo SVM e com o *Kernel* RBF.

A ideia chave do algoritmo *Relief* é estimar a relevância das características de acordo com o quão bom seus valores são para distinguir instâncias próximas uma das outras. Para cada instância sorteada são encontradas as duas instâncias vizinhas mais próximas: uma para a mesma classe (denominada *nearest hit*) e outra para a classe distinta

(denominada *nearest miss*). O algoritmo atualiza o peso das características baseado na ideia de que uma característica é mais relevante se esta distingue uma instância de sua *nearest miss* e menos relevante se esta distingue uma instância de sua *nearest hit*.

O procedimento de seleção consistiu nos seguintes passos:

- As relevâncias das características foram ordenadas em ordem decrescente, aqui denotadas $r^{(1)}, r^{(2)}, \dots, r^{(k)}$
- Dentro desse ranqueamento, foram calculadas as relevâncias médias sobre janelas móveis de tamanho 3: $m_i = (r^{(i)} + r^{(i+1)} + r^{(i+2)})/3$, $i = 1, 2, \dots, k - 2$
- Definiu-se como ponto de parada a quantidade de características a partir da qual o decremento relativo na relevância média fosse inferior a 10%:

$$i^* = \arg \min_{i=1, \dots, k-3} (m_i - m_{i+1})/m_i < 0.1$$

5.3 Métodos de validação *hold-out* tradicional e temporal

Foram definidos dois protocolos de validação: um *hold-out* tradicional (atemporal) e um *hold-out* temporal. A denominação “*hold-out* tradicional” é atribuída aqui ao método em que os subconjuntos de treinamento e de teste são sorteados a partir do mesmo conjunto de dados independente do tempo. Já a denominação “*hold-out* temporal” é atribuída aqui ao método em que o subconjunto de treinamento é sorteado entre as instâncias até o instante t e o subconjunto de teste sorteado entre as instâncias em $t + 1$, como representado na Figura 14.

Figura 14 – Representação dos métodos de validação *hold-out* tradicional e temporal; pontos vermelhos=instâncias de treinamento, pontos azuis=instâncias de teste

Fonte: Clebiano da Costa Sá, 2018

Na Figura 15 é apresentada a estratégia de avaliação dos modelos de classificação ao longo do tempo. Para cada instante t , de 2011 a 2015, no método temporal os classificadores

foram treinados com 2.800 instâncias sorteadas (sem reposição) no período de 2011 a t e testados com 1.200 instâncias sorteadas em $t + 1$ (teste T1). Já no método tradicional, os classificadores foram treinados e testados (teste T1) com instâncias sorteadas (sem reposição) dentro do mesmo período de 2011 a $t + 1$, sendo 2.800 instâncias para treinamento e 1.200 instâncias para teste. As instâncias para o teste T1 no método tradicional foram sorteadas de tal forma que mantem a distribuição original das classes no conjunto de dados em cada período. Note-se que, em ambos os métodos, os tamanhos dos conjuntos de treinamento e de testes foram sempre os mesmos, bem como as proporções entre as classes. Em ambos os métodos de validação os classificadores também foram testados com todas as instâncias em $t + 2$ (teste T2).

Figura 15 – Representação dos métodos de validação *hold-out* tradicional e temporal aplicados em diferentes períodos

Fonte: Clebiano da Costa Sá, 2018

Especificamente para a comparação dos métodos de validação temporal e tradicional a partir da avaliação de diferentes modelos de classificação, foram utilizadas 100 reamostragens por período. Em cada reamostragem foram calculadas a acurácia (ACC) e a área sob a curva (AUC). A AUC foi estimada pela acurácia balanceada (SOKOLOVA; JAPKOWICZ; SZPAKOWICZ, 2006), $AUC = (\text{sensibilidade} + \text{especificidade})/2$. Tomou-se o cuidado de treinar e testar os modelos de classificação usando os mesmos conjuntos (treinamento e teste) e de calibrar os principais parâmetros dos classificadores a partir dos métodos de validação *hold-out* tradicional e temporal. Na calibração de parâmetros foram utilizadas apenas 50 reamostragens por período devido ao fato deste processo apresentar um maior custo de processamento. Para cada método de validação, período de treinamento

e modelo de classificação foram determinados os valores dos parâmetros que, combinados, maximizam a AUC média.

Para as SVMs os melhores valores de C e γ foram encontrados aplicando a função de otimização por enxame de partículas (PSO) (EVERS; GHALIA, 2009) sobre uma *grid* bidimensional cujos valores de C e γ variaram em escala logaritma. Mais especificamente, $C = \exp(\log Cost)$ e $\gamma = \exp(\log Gamma)$ para valores de $\log Cost$ e $\log Gamma$ variando de -15 a 15 em um passo que varia de acordo com as otimizações da função PSO. Foi estabelecido como limite de parada uma estagnação máxima de 5 tentativas, em que não há acréscimo na AUC média. Nos demais parâmetros foram utilizados os valores padrões.

Para o modelo baseado em *Random Forest* o parâmetro *mtry* foi calibrado maximizando a AUC média a partir do teste de valores de uma sequência numérica com limite inferior $0.5\sqrt{nfeat}$, limite superior $6\sqrt{nfeat}$ e intervalo de 5, no qual *nfeat* corresponde ao total de características no conjunto de dados.

Por fim, faz-se útil aqui uma observação sobre a decisão de construção dos dois protocolos de validação serem baseados no método *hold-out*. Uma das perguntas de pesquisa desse trabalho era se, no contexto da classificação dos RNAs codificantes e não codificantes, as estimativas de desempenho obtidas por um método de validação baseado em *folds* temporais difere de um método de validação que ignora a informação de tempo. A estratégia inicial era fazer essa comparação através de técnicas de validação cruzada – usando o método *k-fold* tradicional para a validação atemporal, e aplicando uma adaptação para o caso temporal com base em Arlot (2010) (ARLOT; CELISSE, 2010). Todavia, não seria possível garantir que os tamanhos dos conjuntos treinamentos e testes fossem iguais sob os dois métodos de validação, nem tampouco as distribuições das classes. Isso poderia inviabilizar uma comparação direta entre os dois métodos de validação, uma vez que parte das eventuais diferenças nos resultados poderia ser atribuída às diferenças nas proporções das classes e nos tamanhos dos conjuntos de testes. Para contornar esse problema é que foram construídos os dois protocolos de validação com base no método *hold-out*.

5.4 Consolidação dos resultados e teste das hipóteses

Após a condução das validações *hold-out* tradicional e temporal dos classificadores ao longo do tempo, obteve-se uma tabela composta por oito atributos: método de validação, modelo de classificação, período de treinamento, ID da reamostragem, ID do teste (se

T1 ou T2), ID da sequência, classe predita e a classe real. Estes resultados brutos foram então processados para possibilitar os testes das hipóteses. Para cada método de validação, modelo de classificação, período de treinamento e reamostragem foram calculadas a área sob a curva ROC (AUC) e acurácia (ACC) sob os conjuntos de teste T1 e T2. A partir destas AUC e ACC é que foram testadas as hipóteses.

Para testar a validade da hipótese H1, de que a validação *hold-out* tradicional tende a subestimar os erros de classificação, e da hipótese H2, de que a validação *hold-out* temporal tende a não subestimar os erros de classificação, para cada método de validação, modelo de classificação e período de treinamento aplicou-se o teste t entre as AUC e entre as ACC dos testes T1 e T2. Neste caso, cada amostra corresponde às 100 observações de AUC e de ACC nas repetições por reamostragem dos dados.

Para testar a validade da hipótese H3, de que a validação *hold-out* temporal tende a fornecer estimativas mais fidedignas, e da hipótese H4, de que os classificadores calibrados pelo método de validação temporal tendem a fornecer erros de classificação menores que classificadores calibrados pelo método tradicional, para cada modelo de classificação, período de treinamento e teste (T1 ou T2) aplicou-se o teste t entre as AUC e entre as ACC obtidas a partir dos métodos de validação tradicional e temporal.

Para testar a validade da hipótese H5, de que um novo modelo de classificação, construído a partir da associação das características de classificadores do estado da arte com um conjunto de novas características, proporciona melhoras na discriminação dos RNAs codificantes e não codificantes, para cada método de validação, período de treinamento e teste (T1 ou T2) foi identificado o modelo de classificação com maior AUC e ACC média. Este modelo foi considerado a referência, a partir do qual foram realizados testes t entre as AUC e entre as ACC dos outros modelos.

6 Resultados e discussões

Neste capítulo são apresentados e discutidos os resultados obtidos para as cinco hipóteses testadas.

6.1 Análise da validação *hold-out* tradicional

Nesta seção são apresentados os resultados para os testes da hipótese H1, de que a validação *hold-out* tradicional tende a subestimar os erros de classificação, ou seja, quando extraídos os conjuntos de treinamento e de teste sobre um mesmo grande período, sem levar em consideração o ano de descobrimento e rotulagem das sequências, as estimativas de erro tendem a ser menores do que os erros medidos sobre sequências mais novas (do último ano).

Na Tabela 11 são apresentadas, respectivamente, as AUC e ACC médias para cada modelo de classificação, avaliados sob o método de validação *hold-out* tradicional nos testes T1 e T2. A apresentação das AUC e ACC por algoritmo de classificação, não especificamente por período, tem como propósito simplificar a apresentação e interpretação dos resultados. Na coluna “Diferenças” é apresentada a diferença da média de avaliação no teste T2 pela média de avaliação no teste T1. Na coluna “p-valor” são apresentados os p-valores obtidos a partir do teste estatístico t entre as medidas de avaliação obtidas nos testes T1 e T2.

Em todos os modelos de classificação foram observados p-valores abaixo de 0,001, indicando que os resultados das avaliações sobre os conjuntos de teste T1 e T2 não são estatisticamente equivalentes. Também é possível observar que as diferenças entre as avaliações AUC e ACC dos testes T1 e T2 apresentam valores negativos para todos os modelos de classificação. Isto significa que no método de validação *hold-out* tradicional os classificadores apresentam uma maior taxa de erros quando avaliados sobre sequências mais novas. Portanto, existem fortes indicativos de que o método de validação *hold-out* tradicional tende a subestimar os erros de classificação para o conjunto de sequências de RNAs codificantes e não codificantes utilizados neste trabalho, mantendo a hipótese H1.

A Figura 16 apresenta a AUC média dos modelos de classificação sob o método de validação *hold-out* tradicional em cada período de treinamento. Na Figura 16a são apresentadas as AUC médias sobre o teste T1 e na Figura 16b as AUC médias sobre o

Tabela 11 – AUC e ACC média das avaliações por período para cada modelo de classificação sob validação *hold-out* tradicional nos testes T1 e T2. Entre parênteses está a média dos desvios padrões obtidos a partir das avaliações por período

Modelo de classificação	T1	T2	Diferenças	p-valor
AUC				
RForestAllFeat	0.946 (± 0.007)	0.914 (± 0.045)	-0.032 (± 0.045)	< 0.001
SVMrbfAllFeat	0.923 (± 0.008)	0.906 (± 0.027)	-0.016 (± 0.030)	< 0.001
SVM_CPC1	0.931 (± 0.007)	0.904 (± 0.040)	-0.027 (± 0.041)	< 0.001
SVM_CPC2	0.869 (± 0.010)	0.860 (± 0.037)	-0.010 (± 0.038)	< 0.001
SVM_CNCI	0.888 (± 0.009)	0.870 (± 0.034)	-0.018 (± 0.037)	< 0.001
CPAT	0.882 (± 0.010)	0.865 (± 0.042)	-0.017 (± 0.043)	< 0.001
PLEK	0.845 (± 0.012)	0.807 (± 0.037)	-0.038 (± 0.040)	< 0.001
SVMrbfBestfeatRelief	0.911 (± 0.008)	0.892 (± 0.035)	-0.020 (± 0.036)	< 0.001
ACC				
RForestAllFeat	0.943 (± 0.007)	0.906 (± 0.060)	-0.037 (± 0.060)	< 0.001
SVMrbfAllFeat	0.913 (± 0.008)	0.896 (± 0.034)	-0.017 (± 0.036)	< 0.001
SVM_CPC1	0.926 (± 0.008)	0.894 (± 0.052)	-0.033 (± 0.053)	< 0.001
SVM_CPC2	0.850 (± 0.011)	0.841 (± 0.051)	-0.008 (± 0.052)	< 0.001
SVM_CNCI	0.874 (± 0.009)	0.864 (± 0.034)	-0.010 (± 0.036)	< 0.001
CPAT	0.884 (± 0.009)	0.879 (± 0.027)	-0.005 (± 0.029)	< 0.001
PLEK	0.837 (± 0.012)	0.817 (± 0.032)	-0.020 (± 0.036)	< 0.001
SVMrbfBestfeatRelief	0.906 (± 0.008)	0.881 (± 0.047)	-0.024 (± 0.048)	< 0.001

Fonte: Clebiano da Costa Sá, 2018

teste T2. É notável que as AUC médias no teste T1 são uniformes nos cinco períodos de treinamento, enquanto que no teste T2 são extremamente variáveis, principalmente nos últimos três anos. Esta análise reafirma a nossa hipótese de que a validação *hold-out* tradicional tende a subestimar os erros de classificação.

Figura 16 – AUC média para cada modelo de classificação sob o método de validação *hold-out* tradicional em cada período de treinamento: a) teste T1 e b) teste T2

Fonte: Clebiano da Costa Sá, 2018

6.2 Análise da validação *hold-out* temporal

Esta seção apresenta os resultados para os testes da hipótese H2, de que a validação *hold-out* temporal tende a não subestimar os erros de classificação, no sentido de que, os erros de classificação obtidos sobre instâncias de testes em $t + 1$ tendem a ser parecidos com os erros de classificação em $t + 2$.

Na Tabela 12 (estrutura descrita na Seção 6.1) são apresentadas, respectivamente, as AUC e ACC médias para cada modelo de classificação, avaliados sob o método de validação *hold-out* temporal nos testes T1 e T2.

Tabela 12 – AUC e ACC média das avaliações por período para cada modelo de classificação sob validação *hold-out* temporal nos testes T1 e T2. Entre parênteses está a média dos desvios padrões obtidos a partir das avaliações por período

Modelo de classificação	T1	T2	Diferenças	p-valor
AUC				
RForestAllFeat	0.929 (± 0.040)	0.912 (± 0.047)	-0.017 (± 0.035)	< 0.001
SVMrbfAllFeat	0.912 (± 0.023)	0.906 (± 0.025)	-0.007 (± 0.026)	< 0.001
SVM_CPC1	0.926 (± 0.022)	0.895 (± 0.051)	-0.031 (± 0.035)	< 0.001
SVM_CPC2	0.867 (± 0.034)	0.861 (± 0.033)	-0.006 (± 0.055)	0.008
SVM_CNCI	0.881 (± 0.025)	0.869 (± 0.034)	-0.012 (± 0.033)	< 0.001
CPAT	0.883 (± 0.018)	0.866 (± 0.044)	-0.017 (± 0.046)	< 0.001
PLEK	0.822 (± 0.029)	0.803 (± 0.035)	-0.020 (± 0.035)	< 0.001
SVMrbfBestfeatRelief	0.905 (± 0.024)	0.888 (± 0.037)	-0.017 (± 0.026)	< 0.001
ACC				
RForestAllFeat	0.911 (± 0.065)	0.896 (± 0.064)	-0.016 (± 0.066)	< 0.001
SVMrbfAllFeat	0.897 (± 0.033)	0.896 (± 0.031)	-0.001 (± 0.042)	0.620
SVM_CPC1	0.918 (± 0.027)	0.896 (± 0.051)	-0.023 (± 0.048)	< 0.001
SVM_CPC2	0.840 (± 0.052)	0.837 (± 0.051)	-0.002 (± 0.080)	0.451
SVM_CNCI	0.866 (± 0.034)	0.859 (± 0.037)	-0.007 (± 0.048)	0.004
CPAT	0.879 (± 0.028)	0.880 (± 0.027)	0.000 (± 0.047)	0.886
PLEK	0.815 (± 0.032)	0.808 (± 0.034)	-0.008 (± 0.053)	< 0.001
SVMrbfBestfeatRelief	0.892 (± 0.036)	0.881 (± 0.045)	-0.011 (± 0.046)	< 0.001

Fonte: Clebiano da Costa Sá, 2018

As comparações entre as AUC médias de todos os modelos de classificação apresentaram p-valores abaixo de 0,1, indicando que em termos de AUC os resultados das avaliações sobre os testes T1 e T2 não são estatisticamente equivalentes. Também é possível observar que as diferenças entre as AUC dos testes T1 e T2 apresentam valores negativos para todos os modelos de classificação. Isso mostra que as estimativas de erro para o teste T1 são menores que os erros de classificação sobre sequências novas (em T2). Desta forma, nota-se que a validação *hold-out* temporal também tende a subestimar os erros.

Considerando a ACC, três classificadores tiveram *p-valor* acima de 0,1, e as diferenças destes classificadores foram abaixo de 0,002. O CPAT inclusive nem sequer apresentou diferenças entre as ACC médias nos testes T1 e T2. O que é um indicativo da relevância da nossa hipótese.

Ao contrário da validação *hold-out* tradicional, na validação *hold-out* temporal os classificadores revelaram AUC médias similarmente variáveis ao longo do tempo (Figura 17). De maneira geral em alguns períodos as AUC médias são maiores, mas indica um decréscimo em ambos os teste T1 e T2, principalmente nos últimos períodos. Portanto,

embora não hajam evidências que sustentem a hipótese H2 existem indícios de que esta hipótese faz sentido.

Figura 17 – AUC média para cada modelo de classificação sob o método de validação *hold-out* temporal em cada período de treinamento: a) teste T1 e b) teste T2

Fonte: Clebiano da Costa Sá, 2018

6.3 Análise comparativa dos métodos de validação *hold-out* tradicional e temporal

Nesta seção são apresentados os resultados para os testes da hipótese H3, de que a validação *hold-out* temporal tende a fornecer estimativas mais fidedignas, no sentido de que, quando utilizada a validação temporal os erros de classificação obtidos sobre instâncias de testes em $t + 1$ tendem a ser mais parecidos com os erros de classificação em $t + 2$, e da hipótese H4, de que os classificadores calibrados pelo método de validação temporal

tendem a fornecer erros de classificação menores que classificadores calibrados pelo método tradicional.

Na Tabela 13 são apresentadas, respectivamente, as AUC e ACC médias para os modelos de classificação, avaliados sob os métodos de validação *hold-out* temporal e tradicional no teste T1.

Tabela 13 – AUC e ACC média das avaliações por período para cada modelo de classificação sob os métodos de validação *hold-out* temporal e tradicional no teste T1. Entre parênteses está a média dos desvios padrões obtidos a partir das avaliações por período

Modelo de classificação	Temporal	Tradicional	Diferenças	p-valor
AUC T1				
RForestAllFeat	0.929 (± 0.006)	0.946 (± 0.007)	0.017 (± 0.045)	0.005
SVMrbfAllFeat	0.912 (± 0.008)	0.923 (± 0.008)	0.010 (± 0.025)	< 0.001
SVM_CPC1	0.926 (± 0.007)	0.931 (± 0.007)	0.005 (± 0.024)	0.043
SVM_CPC2	0.867 (± 0.009)	0.869 (± 0.010)	0.002 (± 0.035)	< 0.001
SVM_CNCI	0.881 (± 0.008)	0.888 (± 0.009)	0.007 (± 0.026)	< 0.001
CPAT	0.883 (± 0.008)	0.882 (± 0.010)	0.000 (± 0.017)	< 0.001
PLEK	0.822 (± 0.011)	0.845 (± 0.011)	0.022 (± 0.031)	< 0.001
SVMrbfBestfeatRelief	0.905 (± 0.007)	0.911 (± 0.008)	0.006 (± 0.025)	< 0.001
ACC T1				
RForestAllFeat	0.911 (± 0.006)	0.943 (± 0.007)	0.032 (± 0.073)	< 0.001
SVMrbfAllFeat	0.897 (± 0.008)	0.913 (± 0.008)	0.016 (± 0.038)	< 0.001
SVM_CPC1	0.918 (± 0.007)	0.926 (± 0.008)	0.008 (± 0.029)	< 0.001
SVM_CPC2	0.840 (± 0.010)	0.850 (± 0.011)	0.010 (± 0.056)	< 0.001
SVM_CNCI	0.866 (± 0.009)	0.874 (± 0.009)	0.009 (± 0.038)	< 0.001
CPAT	0.879 (± 0.008)	0.884 (± 0.009)	0.005 (± 0.030)	< 0.001
PLEK	0.815 (± 0.011)	0.837 (± 0.011)	0.022 (± 0.037)	0.153
SVMrbfBestfeatRelief	0.892 (± 0.010)	0.906 (± 0.008)	0.014 (± 0.039)	< 0.001

Fonte: Clebiano da Costa Sá, 2018

No teste T1 (Tabela 13), com exceção apenas das ACC do PLEK, as AUC e ACC dos modelos de classificação não são estatisticamente equivalentes entre as validações tradicional e temporal. Além disso, as diferenças entre as AUC e entre as ACC médias são positivas, inclusive no PLEK, com uma diferença de 2,2%. Isso mostra que as estimativas de erro dos classificadores tendem a ser menores quando avaliados sob o método de validação *hold-out* tradicional. Portanto, existem evidências de que a avaliação dos classificadores sob o método de validação *hold-out* temporal tende a fornecer estimativas de desempenho mais fidedignas, mantendo a hipótese H3.

Na Tabela 14 são apresentadas, respectivamente, as AUC e ACC médias para os modelos de classificação, avaliados sob os métodos de validação *hold-out* temporal e tradicional no teste T2.

Tabela 14 – AUC e ACC média das avaliações por período para cada modelo de classificação sob os métodos de validação *hold-out* temporal e tradicional no teste T2. Entre parênteses está a média dos desvios padrões obtidos a partir das avaliações por período

Modelo de classificação	Temporal	Tradicional	Diferenças	p-valor
AUC T2				
RForestAllFeat	0.912 (± 0.002)	0.914 (± 0.004)	0.002 (± 0.003)	0.080
SVMrbfAllFeat	0.906 (± 0.004)	0.906 (± 0.004)	0.000 (± 0.004)	0.102
SVM_CPC1	0.895 (± 0.006)	0.904 (± 0.006)	0.009 (± 0.018)	0.155
SVM_CPC2	0.861 (± 0.003)	0.860 (± 0.003)	-0.002 (± 0.009)	0.001
SVM_CNCI	0.869 (± 0.005)	0.870 (± 0.005)	0.001 (± 0.001)	0.220
CPAT	0.866 (± 0.004)	0.865 (± 0.004)	-0.001 (± 0.005)	0.007
PLEK	0.803 (± 0.008)	0.807 (± 0.008)	0.004 (± 0.004)	0.162
SVMrbfBestfeatRelief	0.888 (± 0.004)	0.892 (± 0.004)	0.003 (± 0.007)	< 0.001
ACC T2				
RForestAllFeat	0.896 (± 0.002)	0.906 (± 0.004)	0.010 (± 0.021)	0.019
SVMrbfAllFeat	0.896 (± 0.004)	0.896 (± 0.004)	0.000 (± 0.005)	0.012
SVM_CPC1	0.896 (± 0.007)	0.894 (± 0.007)	-0.002 (± 0.007)	0.180
SVM_CPC2	0.837 (± 0.004)	0.841 (± 0.004)	0.004 (± 0.003)	0.015
SVM_CNCI	0.859 (± 0.004)	0.864 (± 0.005)	0.005 (± 0.005)	< 0.001
CPAT	0.880 (± 0.002)	0.879 (± 0.002)	0.000 (± 0.004)	0.083
PLEK	0.808 (± 0.007)	0.817 (± 0.007)	0.009 (± 0.005)	0.009
SVMrbfBestfeatRelief	0.881 (± 0.006)	0.881 (± 0.004)	0.001 (± 0.005)	< 0.001

Fonte: Clebiano da Costa Sá, 2018

No teste T2 (Tabela 14), em termos de AUC, dois modelos de classificação apresentaram diferenças negativas e *p-valor* abaixo de 0,1. No entanto, a diferença entre as AUC médias dos modelo *SVM_CPC2* e *CPAT* a favor da validação temporal é de apenas 0,2% e 0,1%, respectivamente. Além disso, o classificador *CPAT* não teve nenhum parâmetro calibrado. Desta forma, não temos indícios de que os classificadores apresentam maior desempenho quando treinados a partir de parâmetros calibrados via validação temporal. Dentre os demais modelos quatro são estatisticamente equivalentes (*p-valor* > 0,1) e dois apresentam *p-valor* abaixo de 0,1, embora com diferenças de apenas 0,2% e 0,3% a favor da validação tradicional. Portanto, não existem evidências suficientes para afirmar que classificadores treinados a partir de parâmetros calibrados via métodos de validação *hold-out* temporal ou tradicional apresentam melhor desempenho, refutando a hipótese H4.

Em termos de ACC, sete modelos apresentaram diferenças positivas e *p-valor* abaixo de 0,1, o que favorece ao método de validação tradicional. Isso mostraria que os classificadores com parâmetros calibrados via validação tradicional teriam melhor desempenho, mas a maioria destas diferenças são menores que 1% e a acurácia por si só não garante que os classificadores estão proporcionando uma melhor discriminação das classes. O único modelo a apresentar uma diferença negativa, o que favoreceria o método temporal, apresentou equivalência estatística entre as ACC (*p-valor* > 0,1).

Com base nestes resultados sugerimos o uso do método de validação *hold-out* temporal para a obtenção de estimativas mais fidedignas de desempenho dos classificadores. Por outro lado, os métodos de validação *hold-out* temporal e tradicional proporcionam classificadores com similar desempenho sobre sequências novas.

6.4 Avaliação dos modelos de classificação

Esta seção apresenta os resultados para os testes da hipótese H5, de que um novo modelo de classificação, construído a partir da associação das características de classificadores do estado da arte com um conjunto de novas características, proporciona melhoras na discriminação dos RNAs codificantes e não codificantes.

As tabelas 16, 17, 18 e 19 apresentam uma análise comparativa dos modelos de classificação por método de validação e período de treinamento. As AUC e ACC médias foram calculadas a partir das AUC e ACC das 100 subamostras aleatórias do conjunto de teste T1 e das 100 repetições do teste T2 em cada período de treinamento. Para cada método de validação e período de treinamento o modelo de classificação com maior AUC ou ACC média foi definido como referência (cédulas marcadas na cor verde). Os classificadores com equivalência estatística (*p-valor* > 0,1) em relação ao modelo de referência foram considerados como empates/equivalentes (cédulas marcadas na cor azul).

O nosso modelo de classificação *RForestAllFeat* apresentou uma melhor discriminação das classes nos testes T1 e T2 sob os métodos de validação tradicional e temporal em diferentes períodos de treinamento (tabelas 16, 17, 18 e 19), mantendo a hipótese H5. No Apêndice B é apresentado o nível de importância das características mais relevantes, de acordo com o algoritmo *Random Forest*. Dentre as características com maior nível de importância observamos que há uma predominância de características provenientes do classificador CPC1 e do novo extrator. Estas características são baseadas principalmente

na estrutura da ORF, em buscas por similaridade contra bancos de dados, no viés do uso de hexamers, no tamanho e score da MLCDS e no tamanho da sequência nucleotídica. Vale ressaltar que muitas das características no topo desta lista são novas em relação ao estado da arte. Exemplos incluem a cobertura da ORF sobre a sequência, o tamanho da UTR 3 e da UTR 5 e a quantidade de códons de iniciação dentro e fora da UTR 5.

Os outros modelos novos, *SVMrbfAllFeat* e *SVMrbfBestfeatRelief*, também apresentaram bons resultados. E dentre os modelos provenientes do estado da arte destacou-se o modelo *SVM_CPC1*.

Dentre as 27 características selecionadas pelo *Relief* e que compõem o modelo de classificação *SVMrbfBestfeatRelief* estão 18 do PLEK (LI; ZHANG; ZHOU, 2014), 2 do CPC (KONG et al., 2007), 1 do CPAT (WANG et al., 2013), 1 do CNCI (SUN et al., 2013) e 5 do novo conjunto de características (SILVA, 2009b; SILVA, 2009a) (Descritas na Tabela 15). As características novas selecionadas incluem as frequências de ocorrência dos códons CGT, TTC e GTA na maior ORF e a cobertura da ORF.

Tabela 15 – Características selecionadas com base no método RELIEF

Origem	Descrição das características	Total de variáveis
CPC1	Hit seq count	1
CPC1	Strict	1
CNCI	S-score da MLCDS	1
CPAT	Estatística Fickett TESTCODE	1
PLEK	Frequências de uso de K-mer	18
Novo Extrator	Cobertura da ORF sobre a sequência	1
Novo Extrator	Frequência dos trinucleotídeos CGT, TTC e GTA	3
Novo Extrator	Distribuição do aminoácido R	1

Todos os modelos apresentaram acurácias (ACC) acima de 75% e apenas o nosso modelo *RForestAllFeat* atingiu acurácias acima de 95%, isso em alguns períodos de treinamento, apesar do fato de que nos artigos dos classificadores do estado da arte CPC, CNCI, CPAT e PLEK as acurácias descritas serem, respectivamente, 95,77%, 97,3%, 96,7% e 95,6% (KONG et al., 2007; SUN et al., 2013; WANG et al., 2013; LI; ZHANG; ZHOU, 2014). Estes valores não são comparáveis com nossos resultados porque estes classificadores foram avaliados sobre diferentes conjuntos de sequências e a partir do método de validação cruzada *k-fold*. É importante ressaltar também que os classificadores CPC, CNCI, CPAT e PLEK foram desenvolvidos e avaliados a partir de RNAs não codificantes longos, formados por sequências de 200 ou mais nucleotídeos, enquanto neste trabalho utilizamos sequências

a partir de 80 nucleotídeos. Em futuras análises serão verificadas as possíveis causas dos erros de classificação ocorridos durante as avaliações dos modelos de classificação.

Tabela 16 – AUC média calculada para as 100 amostras do teste T1 e desvio padrão entre parênteses. Na cor verde estão as maiores médias dentre os classificadores para cada método de validação e período de treinamento. Quando houver, a cor azul indica equivalência estatística entre este modelo e o de referência ($p\text{-valor} > 0,1$)

Período	CPAT	PLEK	SVM_CPC1	SVM_CPC2	SVM_CNCI	SVMrbf- BestfeatRelief	SVMrbf- AllFeat	RForest- AllFeat
Validação hold-out temporal (AUC T1)								
2011	0.874 (± 0.009)	0.825 (± 0.011)	0.931 (± 0.007)	0.844 (± 0.009)	0.869 (± 0.008)	0.903 (± 0.008)	0.899 (± 0.008)	0.944 (± 0.007)
2012	0.893 (± 0.008)	0.838 (± 0.009)	0.937 (± 0.006)	0.886 (± 0.008)	0.896 (± 0.009)	0.922 (± 0.006)	0.926 (± 0.007)	0.954 (± 0.005)
2013	0.909 (± 0.008)	0.862 (± 0.009)	0.949 (± 0.006)	0.911 (± 0.008)	0.919 (± 0.008)	0.931 (± 0.007)	0.948 (± 0.006)	0.960 (± 0.005)
2014	0.862 (± 0.008)	0.792 (± 0.015)	0.926 (± 0.007)	0.817 (± 0.010)	0.868 (± 0.009)	0.906 (± 0.008)	0.899 (± 0.009)	0.937 (± 0.007)
2015	0.876 (± 0.007)	0.794 (± 0.012)	0.887 (± 0.007)	0.876 (± 0.008)	0.854 (± 0.008)	0.864 (± 0.007)	0.890 (± 0.007)	0.851 (± 0.005)
Média	0.883 (± 0.008)	0.822 (± 0.011)	0.926 (± 0.007)	0.867 (± 0.009)	0.881 (± 0.008)	0.905 (± 0.007)	0.912 (± 0.008)	0.929 (± 0.006)
Validação hold-out tradicional (AUC T1)								
2011	0.881 (± 0.010)	0.839 (± 0.011)	0.930 (± 0.007)	0.870 (± 0.009)	0.886 (± 0.008)	0.909 (± 0.008)	0.921 (± 0.007)	0.946 (± 0.007)
2012	0.883 (± 0.010)	0.845 (± 0.011)	0.931 (± 0.008)	0.869 (± 0.010)	0.887 (± 0.009)	0.911 (± 0.008)	0.920 (± 0.008)	0.946 (± 0.006)
2013	0.885 (± 0.010)	0.846 (± 0.010)	0.930 (± 0.007)	0.872 (± 0.009)	0.888 (± 0.008)	0.912 (± 0.009)	0.923 (± 0.007)	0.946 (± 0.007)
2014	0.881 (± 0.011)	0.850 (± 0.012)	0.930 (± 0.007)	0.868 (± 0.011)	0.889 (± 0.008)	0.914 (± 0.009)	0.925 (± 0.008)	0.945 (± 0.007)
2015	0.882 (± 0.010)	0.843 (± 0.011)	0.932 (± 0.008)	0.869 (± 0.009)	0.889 (± 0.010)	0.911 (± 0.008)	0.925 (± 0.009)	0.947 (± 0.007)
Média	0.882 (± 0.010)	0.845 (± 0.011)	0.931 (± 0.007)	0.869 (± 0.010)	0.888 (± 0.009)	0.911 (± 0.008)	0.923 (± 0.008)	0.946 (± 0.007)

Fonte: Clebiano da Costa Sá, 2018

Tabela 17 – AUC média calculada para as 100 repetições do teste T2 e desvio padrão entre parênteses. Na cor verde estão as maiores médias dentre os classificadores para cada método de validação e período de treinamento. Quando houver, a cor azul indica equivalência estatística entre este modelo e o de referência ($p\text{-valor} > 0,1$)

Período	CPAT	PLEK	SVM_CPC1	SVM_CPC2	SVM_CNCI	SVMrbf- BestfeatRelief	SVMrbf- AllFeat	RForest- AllFeat
Validação hold-out temporal (AUC T2)								
2011	0.901 (± 0.002)	0.826 (± 0.004)	0.935 (± 0.002)	0.879 (± 0.002)	0.894 (± 0.002)	0.918 (± 0.002)	0.919 (± 0.002)	0.953 (± 0.001)
2012	0.908 (± 0.002)	0.855 (± 0.004)	0.942 (± 0.003)	0.910 (± 0.002)	0.915 (± 0.002)	0.929 (± 0.003)	0.945 (± 0.002)	0.959 (± 0.001)
2013	0.859 (± 0.002)	0.788 (± 0.007)	0.925 (± 0.002)	0.815 (± 0.002)	0.864 (± 0.003)	0.905 (± 0.003)	0.901 (± 0.003)	0.937 (± 0.001)
2014	0.877 (± 0.004)	0.786 (± 0.012)	0.857 (± 0.014)	0.864 (± 0.004)	0.853 (± 0.007)	0.853 (± 0.006)	0.889 (± 0.004)	0.855 (± 0.003)
2015	0.787 (± 0.010)	0.757 (± 0.014)	0.814 (± 0.010)	0.838 (± 0.006)	0.819 (± 0.010)	0.837 (± 0.006)	0.874 (± 0.009)	0.857 (± 0.005)
Média	0.866 (± 0.004)	0.803 (± 0.008)	0.895 (± 0.006)	0.861 (± 0.003)	0.869 (± 0.005)	0.888 (± 0.004)	0.906 (± 0.004)	0.912 (± 0.002)
Validação hold-out tradicional (AUC T2)								
2011	0.893 (± 0.003)	0.836 (± 0.004)	0.936 (± 0.002)	0.883 (± 0.002)	0.895 (± 0.002)	0.920 (± 0.002)	0.926 (± 0.003)	0.953 (± 0.001)
2012	0.909 (± 0.002)	0.859 (± 0.004)	0.946 (± 0.003)	0.911 (± 0.002)	0.916 (± 0.003)	0.932 (± 0.002)	0.947 (± 0.002)	0.959 (± 0.001)
2013	0.861 (± 0.002)	0.790 (± 0.007)	0.926 (± 0.002)	0.816 (± 0.003)	0.866 (± 0.003)	0.903 (± 0.003)	0.897 (± 0.004)	0.937 (± 0.001)
2014	0.875 (± 0.005)	0.787 (± 0.010)	0.856 (± 0.012)	0.867 (± 0.003)	0.852 (± 0.008)	0.850 (± 0.006)	0.887 (± 0.005)	0.856 (± 0.002)
2015	0.790 (± 0.008)	0.762 (± 0.017)	0.855 (± 0.009)	0.821 (± 0.007)	0.820 (± 0.012)	0.852 (± 0.006)	0.874 (± 0.008)	0.865 (± 0.012)
Média	0.865 (± 0.004)	0.807 (± 0.008)	0.904 (± 0.006)	0.860 (± 0.003)	0.870 (± 0.005)	0.892 (± 0.004)	0.906 (± 0.004)	0.914 (± 0.004)

Fonte: Clebiano da Costa Sá, 2018

Tabela 18 – ACC média calculada para as 100 amostras do teste T1 e desvio padrão entre parênteses. Na cor verde estão as maiores médias dentre os classificadores para cada método de validação e período de treinamento. Quando houver, a cor azul indica equivalência estatística entre este modelo e o de referência ($p\text{-valor} > 0,1$)

Período	CPAT	PLEK	SVM_CPC1	SVM_CPC2	SVM_CNCI	SVMrbf- BestfeatRelief	SVMrbf- AllFeat	RForest- AllFeat
Validação hold-out temporal (ACC T1)								
2011	0.862 (± 0.009)	0.809 (± 0.012)	0.922 (± 0.008)	0.809 (± 0.011)	0.847 (± 0.010)	0.889 (± 0.009)	0.880 (± 0.010)	0.934 (± 0.007)
2012	0.895 (± 0.008)	0.836 (± 0.009)	0.937 (± 0.006)	0.882 (± 0.009)	0.893 (± 0.009)	0.921 (± 0.007)	0.924 (± 0.007)	0.954 (± 0.005)
2013	0.914 (± 0.008)	0.859 (± 0.009)	0.949 (± 0.006)	0.908 (± 0.008)	0.916 (± 0.008)	0.933 (± 0.007)	0.946 (± 0.006)	0.961 (± 0.005)
2014	0.837 (± 0.008)	0.770 (± 0.014)	0.907 (± 0.009)	0.764 (± 0.012)	0.833 (± 0.010)	0.885 (± 0.009)	0.870 (± 0.011)	0.923 (± 0.007)
2015	0.889 (± 0.005)	0.802 (± 0.010)	0.875 (± 0.008)	0.835 (± 0.010)	0.839 (± 0.006)	0.834 (± 0.017)	0.863 (± 0.007)	0.785 (± 0.006)
Média	0.879 (± 0.008)	0.815 (± 0.011)	0.918 (± 0.007)	0.840 (± 0.010)	0.866 (± 0.009)	0.892 (± 0.010)	0.897 (± 0.008)	0.911 (± 0.006)
Validação hold-out tradicional (ACC T1)								
2011	0.883 (± 0.010)	0.831 (± 0.011)	0.927 (± 0.007)	0.850 (± 0.009)	0.871 (± 0.009)	0.905 (± 0.007)	0.910 (± 0.007)	0.943 (± 0.006)
2012	0.885 (± 0.009)	0.835 (± 0.011)	0.927 (± 0.008)	0.850 (± 0.010)	0.875 (± 0.009)	0.905 (± 0.008)	0.911 (± 0.009)	0.943 (± 0.006)
2013	0.884 (± 0.009)	0.837 (± 0.009)	0.925 (± 0.008)	0.851 (± 0.011)	0.873 (± 0.009)	0.906 (± 0.008)	0.911 (± 0.008)	0.941 (± 0.007)
2014	0.884 (± 0.010)	0.844 (± 0.012)	0.925 (± 0.008)	0.848 (± 0.012)	0.876 (± 0.009)	0.906 (± 0.009)	0.915 (± 0.008)	0.942 (± 0.007)
2015	0.885 (± 0.009)	0.839 (± 0.011)	0.928 (± 0.008)	0.847 (± 0.011)	0.877 (± 0.009)	0.907 (± 0.008)	0.916 (± 0.009)	0.945 (± 0.007)
Média	0.884 (± 0.009)	0.837 (± 0.011)	0.926 (± 0.008)	0.850 (± 0.011)	0.874 (± 0.009)	0.906 (± 0.008)	0.913 (± 0.008)	0.943 (± 0.007)

Fonte: Clebiano da Costa Sá, 2018

Tabela 19 – ACC média calculada para as 100 repetições do teste T2 e desvio padrão entre parênteses. Na cor verde estão as maiores médias dentre os classificadores para cada método de validação e período de treinamento. Quando houver, a cor azul indica equivalência estatística entre este modelo e o de referência ($p\text{-valor} > 0,1$)

Período	CPAT	PLEK	SVM_CPC1	SVM_CPC2	SVM_CNCI	SVMrbf- BestfeatRelief	SVMrbf- AllFeat	RForest- AllFeat
Validação hold-out temporal (ACC T2)								
2011	0.900 (± 0.002)	0.823 (± 0.004)	0.933 (± 0.002)	0.872 (± 0.002)	0.891 (± 0.002)	0.916 (± 0.002)	0.915 (± 0.003)	0.951 (± 0.001)
2012	0.912 (± 0.002)	0.852 (± 0.004)	0.942 (± 0.004)	0.909 (± 0.002)	0.914 (± 0.002)	0.930 (± 0.003)	0.944 (± 0.002)	0.960 (± 0.001)
2013	0.838 (± 0.002)	0.750 (± 0.008)	0.909 (± 0.003)	0.761 (± 0.005)	0.823 (± 0.005)	0.886 (± 0.004)	0.870 (± 0.004)	0.924 (± 0.003)
2014	0.884 (± 0.003)	0.802 (± 0.010)	0.802 (± 0.023)	0.834 (± 0.006)	0.843 (± 0.005)	0.804 (± 0.009)	0.860 (± 0.005)	0.794 (± 0.003)
2015	0.864 (± 0.003)	0.811 (± 0.011)	0.891 (± 0.003)	0.812 (± 0.006)	0.825 (± 0.006)	0.867 (± 0.014)	0.890 (± 0.005)	0.850 (± 0.003)
Média	0.880 (± 0.002)	0.808 (± 0.007)	0.896 (± 0.007)	0.837 (± 0.004)	0.859 (± 0.004)	0.881 (± 0.006)	0.896 (± 0.004)	0.896 (± 0.002)
Validação hold-out tradicional (ACC T2)								
2011	0.894 (± 0.003)	0.835 (± 0.004)	0.936 (± 0.002)	0.878 (± 0.002)	0.892 (± 0.002)	0.919 (± 0.002)	0.923 (± 0.003)	0.953 (± 0.001)
2012	0.914 (± 0.001)	0.858 (± 0.004)	0.946 (± 0.003)	0.909 (± 0.001)	0.915 (± 0.002)	0.933 (± 0.002)	0.945 (± 0.002)	0.960 (± 0.001)
2013	0.835 (± 0.002)	0.764 (± 0.006)	0.908 (± 0.004)	0.762 (± 0.006)	0.830 (± 0.004)	0.882 (± 0.004)	0.864 (± 0.004)	0.922 (± 0.003)
2014	0.889 (± 0.003)	0.805 (± 0.009)	0.802 (± 0.019)	0.837 (± 0.006)	0.846 (± 0.005)	0.799 (± 0.009)	0.859 (± 0.005)	0.795 (± 0.003)
2015	0.864 (± 0.003)	0.822 (± 0.014)	0.878 (± 0.009)	0.819 (± 0.004)	0.838 (± 0.009)	0.874 (± 0.005)	0.888 (± 0.004)	0.896 (± 0.011)
Média	0.879 (± 0.002)	0.817 (± 0.007)	0.894 (± 0.007)	0.841 (± 0.004)	0.864 (± 0.005)	0.881 (± 0.004)	0.896 (± 0.004)	0.906 (± 0.004)

Fonte: Clebiano da Costa Sá, 2018

7 Conclusão

Neste trabalho comparamos duas estratégias de avaliação de classificadores de RNAs codificantes e não codificantes: *hold-out* temporal e *hold-out* tradicional (atemporal). Além disso, testamos novos modelos de classificação a partir da combinação de diferentes algoritmos geradores de classificadores com características de classificadores do estado da arte e um novo conjunto de características.

A partir dos testes das hipóteses observamos que, para o conjunto de dados de RNAs codificantes e não codificantes utilizados nesta pesquisa, tanto a validação *hold-out* tradicional quanto a validação *hold-out* temporal tendem a subestimar os erros de classificação, que a avaliação por validação temporal é mais fidedigna, que classificadores treinados a partir de parâmetros calibrados por validação temporal não melhoram a classificação e que nosso modelo de classificação baseado em *Random Forest* e treinado com características de classificadores do estado da arte somadas a um novo conjunto de características proporcionou uma melhora significativa na discriminação dos RNAs codificantes e não codificantes.

O grande e abrangente conjunto de dados temporal criado neste trabalho será de grande importância para pesquisas que envolvam o desenvolvimento e a avaliação de classificadores de RNAs codificantes e não codificantes. Os protocolos de validação temporal e tradicional produzidos poderão ser avaliados e aplicados sobre outros problemas de classificação. Destacamos também o potencial do algoritmo *Random Forest* e das características utilizadas, diante do problema de classificação dos RNAs em codificante ou não codificante de proteína.

Por fim, em função dos resultados aqui apresentados sugerimos o uso do método de validação *hold-out* temporal para a obtenção de estimativas de desempenho mais fidedignas para os classificadores de RNAs codificantes e não codificantes, embora classificadores treinados a partir de parâmetros calibrados via validação *hold-out* temporal e tradicional apresentem desempenhos similares quando aplicados sobre sequências novas.

7.1 Principais contribuições

As principais contribuições deste trabalho foram:

Protocolos de validação hold-out tradicional e temporal

Um dos principais enfoques deste trabalho foi a comparação de duas estratégias de validação: temporal e tradicional (atemporal). Além das conclusões apresentadas neste capítulo vale ressaltar que para estas análises construímos dois protocolos de validação *hold-out* tradicional e temporal baseados em subamostragem, escritos na linguagem R.

Novos modelos de classificação

Um outro enfoque deste trabalho foram os testes de novos modelos de classificação. Testamos três modelos: (1) formado pelo algoritmo *Random Forest* utilizando características de classificadores do estado da arte e mais um novo conjunto de características, (2) formado pelo algoritmo SVM e pelas mesmas características do modelo baseado em *Random Forest* e o (3) formado pelo algoritmo SVM e um subconjunto de características selecionadas a partir do *Relief*, um selecionador de características do tipo filtro. Destes modelos o primeiro, denominado *RForestAllFeat*, superou todos os demais modelos de classificação testados.

Revisão bibliográfica sistemática

Realizamos uma revisão sistemática dos classificadores baseados em aprendizado supervisionado desenvolvidos com o propósito de classificar os transcritos em codificante ou não codificante a partir de suas características. A partir desta revisão estabelecemos um panorama geral dos classificadores de RNAs codificantes e não codificantes por meio de análises quantitativas e qualitativas.

Criação de uma ampla amostra rotulada confiável de RNAs codificantes e não codificantes

Para construir o conjunto de dados foram necessárias buscas na *internet* e na literatura científica por bancos de dados que armazenam sequências nucleotídicas de transcritos de RNAs codificantes e não codificantes com anotação confiável. Ter um grande e abrangente conjunto de sequências nucleotídicas de RNAs codificantes e não codificantes com rotulação confiável é de grande importância para que a comunidade científica possa desenvolver e avaliar classificadores de RNAs codificantes e não codificantes de proteína. Estas sequências poderão ser disponibilizadas.

Conjunto de dados contendo a informação temporal

As análises em um contexto temporal demandaram a construção de um conjunto representativo e abrangente de sequências nucleotídicas de transcritos humanos, anotadas como codificante ou não codificante de proteína, de tal forma que representasse a evolução dos bancos de dados ao longo dos últimos quatorze anos.

7.2 Trabalhos futuros

Trabalhos futuros incluem uma análise das sequências classificadas incorretamente a fim de identificar as possíveis causas e soluções. Muitas características com grande potencial para a discriminação dos RNAs codificantes e não codificantes, descritas em nossa revisão sistemática, poderão ser incluídas no aperfeiçoamento dos modelos. Alguns exemplos incluem características relacionadas à interação com ribossomos, alinhamentos baseados em perfis de modelos de cadeias ocultas de *Markov* (*profile* HMM) e relacionadas à estrutura do gene.

Considerando o fato de muitos classificadores terem sido desenvolvidos com o propósito de prever a classe de RNAs transcritos a partir da sequência nucleotídica não é mais o suficiente ter um classificador com melhor classificação para justificar seu uso. Um bom classificador deve ser rápido e de fácil extração das características. Portanto, análises de desempenho computacional deverão ser realizadas. Características de difícil extração poderão ser removidas e outras características poderão ser integradas a fim de obter um classificador que seja rápido, de fácil uso e que melhore a predição da classe de sequências novas.

Referências¹

- ACHAWANANTAKUN, R. et al. LncRNA-ID: Long non-coding RNA IDentification using balanced random forests. *Bioinformatics*, v. 31, n. 24, p. 3897–3905, 2015. Citado 6 vezes nas páginas 19, 20, 41, 44, 51 e 53.
- ALBERTS, B. et al. *Biologia Molecular da Célula*. 5. ed. [S.l.]: Porto Alegre: Artmed, 2010. Citado na página 26.
- ALLEN, E. et al. Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nature Genetics*, v. 36, n. 12, p. 1282–1290, 2004. Citado na página 20.
- ALTSCHUL, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, v. 25, n. 17, p. 3389–3402, 1997. Citado na página 20.
- AMARAL, P. P. et al. lncRNadb: a reference database for long noncoding RNAs. *Nucleic Acids Research*, v. 39, n. 1, p. D146–D151, 2011. Citado na página 29.
- ARCHER, K. J.; KIMES, R. V. Empirical characterization of random forest variable importance measures. *Computational Statistics and Data Analysis*, v. 52, n. 4, p. 2249–2260, 2008. Citado na página 37.
- ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. *Statist. Surv.*, The American Statistical Association, the Bernoulli Society, the Institute of Mathematical Statistics, and the Statistical Society of Canada, v. 4, p. 40–79, 2010. Disponível em: <<https://doi.org/10.1214/09-SS054>>. Citado na página 64.
- ARRIAL, R. T.; TOGAWA, R. C.; BRIGIDO, M. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus *Paracoccidioides brasiliensis*. *BMC Bioinformatics*, v. 10, n. 1, p. 239, 2009. Citado 2 vezes nas páginas 42 e 49.
- ARTANDI, S. E. Telomeres, telomerase, and human disease. *New England Journal of Medicine*, v. 355, n. 12, p. 1195–1197, 2006. Citado na página 19.
- BASGALUPP, M. P. *LEGAL Tree: um algoritmo genérico multi-objetos lexicográfico para indução de árvores de decisão*. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação ICMC - USP, 2010. Citado na página 30.
- BEN-HUR, A. et al. Support vector machines and kernels for computational biology. *PLoS Computational Biology*, v. 4, n. 10, p. e1000173, 2008. Citado 2 vezes nas páginas 33 e 34.
- BISWAS, A. K. et al. CNCTDiscriminator: coding and noncoding transcript discriminator - an excursion through hypothesis learning and ensemble learning approaches. *Journal of Bioinformatics and Computational Biology*, v. 11, n. 05, p. 1342002, 2013. Citado 2 vezes nas páginas 44 e 50.
- BLACKBURN, E. H. Telomerase. In: _____. *The RNA World*. [S.l.]: Cold Spring Harbor Laboratory Press, 1993. cap. 22, p. 557–576. Citado na página 18.

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

- BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. Citado na página 37.
- BROWN, T. A. *Genomes*. 2. ed. [S.l.]: Wiley-Liss, 2002. Citado na página 20.
- BURGES, C. J. A tutorial on support vector machines for pattern recognition. In: . [S.l.: s.n.], 1998. v. 2, p. 121–167. Citado 2 vezes nas páginas 33 e 34.
- BUSCH, A.; EKEN, S. M.; MAEGDEFESSEL, L. Prospective and therapeutic screening value of non-coding RNA as biomarkers in cardiovascular disease. *Annals of Translational Medicine*, v. 4, n. 12, p. 236–236, 2016. Citado na página 19.
- CARUANA, R.; NICULESCU-MIZIL, A. An empirical comparison of supervised learning algorithms. In: *Proceedings of the 23rd International Conference on Machine Learning*. [S.l.]: ACM, 2006. p. 161–168. Citado na página 32.
- CAVAILLE, J. et al. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *PNAS*, v. 97, n. 26, p. 14311–14316, 2000. Citado na página 19.
- CHEN, G. et al. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Research*, v. 41, n. D1, p. 983–986, 2013. Citado na página 29.
- COLE, J. R. et al. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Research*, v. 31, n. 1, p. 442–443, 2003. Citado 3 vezes nas páginas 18, 27 e 28.
- COSTA, F. F. Non-coding RNAs: new players in eukaryotic biology. *Gene*, v. 357, n. 2, p. 83–94, 2005. Citado 2 vezes nas páginas 19 e 25.
- CRIMINISI, A.; SHOTTON, J.; KONUKOGLU, E. Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, v. 7, n. 2-3, p. 81–227, 2011. Citado na página 37.
- CULLEN, B. R. The virology–RNA biology connection. *RNA*, Cold Spring Harbor Lab, v. 21, n. 4, p. 592–594, 2015. Citado na página 27.
- DEROSE, V. J. Two decades of RNA catalysis. *Chemistry & Biology*, Elsevier, v. 9, n. 9, p. 961–969, 2002. Citado 2 vezes nas páginas 27 e 28.
- DINGER, M. E. et al. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *Plos Computational Biology*, v. 4, n. 11, p. e1000176, 2008. Citado na página 21.
- EDDY, S. R. Non-coding RNA genes and the modern RNA world. *Nature Reviews. Genetics*, v. 2, n. 12, p. 919–29, 2001. Citado na página 27.
- EDDY, S. R. Computational genomics of noncoding RNA genes. *Cell*, v. 109, n. 2, p. 137–140, 2002. Citado 2 vezes nas páginas 20 e 27.
- ENDER, C. et al. A human snoRNA with microRNA-like functions. *Molecular Cell*, v. 32, n. 4, p. 519–528, 2008. Citado na página 27.

EVERS, G. I.; GHALIA, M. B. Regrouping particle swarm optimization: A new global optimization algorithm with improved performance consistency across benchmarks. In: *2009 IEEE International Conference on Systems, Man and Cybernetics*. [S.l.: s.n.], 2009. p. 3901–3908. ISSN 1062-922X. Citado na página 64.

FAN, X. N.; ZHANG, S. W. lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning. *Mol. BioSyst.*, The Royal Society of Chemistry, v. 11, n. 3, p. 892–897, 2015. Citado 6 vezes nas páginas 19, 20, 41, 45, 47 e 52.

FAWCETT, T. An introduction to roc analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861 – 874, 2006. ISSN 0167-8655. ROC Analysis in Pattern Recognition. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S016786550500303X>>. Citado na página 39.

FEDOR, M. J.; WILLIAMSON, J. R. The catalytic diversity of RNAs. *Nature Reviews Molecular Cell Biology*, Nature Publishing Group, v. 6, n. 5, p. 399–412, 2005. Citado na página 28.

FRANKLIN, J. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, v. 27, n. 2, p. 83–85, Mar 2005. ISSN 0343-6993. Disponível em: <<https://doi.org/10.1007/BF02985802>>. Citado na página 30.

FRITH, M. C. et al. The abundance of short proteins in the mammalian proteome. *PLoS Genetics*, v. 2, n. 4, p. e52, 2006. Citado na página 20.

FRITH, M. C.; PHEASANT, M.; MATTICK, J. S. The amazing complexity of the human transcriptome. *European Journal of Human Genetics*, v. 13, n. 8, p. 894–897, 2005. Citado na página 21.

FRIZZARINI, C. *Algoritmo para indução de árvores de classificação para dados desbalanceados*. Dissertação (Mestrado) — Escola de Artes, Ciências e Humanidades - Universidade de São Paulo, 2013. Citado na página 30.

GALINDO, M. I. et al. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *Plos Biology*, v. 5, n. 5, p. e106, 2007. Citado na página 20.

GHEYAS, I. A.; SMITH, L. S. Feature subset selection in large dimensionality domains. *Pattern Recognition*, Elsevier, v. 43, n. 1, p. 5–13, 2010. Citado na página 33.

GREIDER, C. W. Telomerase Biochemistry and Regulation. In: _____. *Telomeres*. [S.l.]: Cold Spring Harbor Laboratory Press, 1995. cap. 3, p. 35–68. Citado na página 18.

GRIFFITHS, A. J. F. et al. *Introdução à Genética*. 9. ed. [S.l.]: Rio de Janeiro: Guanabara Koogan, 2008. Citado 2 vezes nas páginas 27 e 28.

GUENNEWIG, B.; COOPER, A. A. The Central Role of Noncoding RNA in the Brain. In: *International Review of Neurobiology*. 1. ed. [S.l.]: Elsevier Inc., 2014. v. 116, p. 153–194. Citado na página 19.

HANNON, G. J. RNA interference. *Nature*, v. 418, n. 6894, p. 244–251, 2002. Citado na página 19.

HARROW, J. et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, v. 22, n. 9, p. 1760–1774, September 2012. ISSN 1088-9051. Disponível em: <<http://europepmc.org/articles/PMC3431492>>. Citado na página 55.

HASLER, J.; SAMUELSSON, T.; STRUB, K. Useful ‘junk’: Alu RNAs in the human transcriptome. *Cellular and Molecular Life Sciences*, v. 64, n. 14, p. 1793–1800, 2007. Citado na página 21.

HERMANN, P. et al. Consequences of mutations in the non-coding RMRP RNA in cartilage-hair hypoplasia. *Human Molecular Genetics*, v. 14, n. 23, p. 3723–3740, 2005. Citado na página 19.

HOLLEY, C. L.; TOPKARA, V. K. An introduction to small non-coding RNAs: miRNA and snoRNA. *Cardiovascular Drugs and Therapy*, v. 25, n. 2, p. 151–159, 2011. Citado na página 27.

HOU, Y.-M.; YANG, X. Regulation of cell death by transfer RNA. *Antioxidants & Redox Signaling*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 19, n. 6, p. 583–594, 2013. Citado na página 28.

HUANG, W. et al. DDX5 and its associated lncRNA Rmrp modulate TH17 cell effector functions. *Nature*, v. 528, n. 7583, p. 517–522, 2015. Citado na página 19.

HUTTENHOFER, A.; SCHATTNER, P.; POLACEK, N. Non-coding RNAs: hope or hype? *Trends in Genetics*, v. 21, n. 5, p. 289–297, 2005. Citado 2 vezes nas páginas 18 e 19.

IMAMURA, T. et al. Non-coding RNA directed DNA demethylation of Sphk1 CpG island. *Biochemical and Biophysical Research Communications*, v. 322, n. 2, p. 593–600, 2004. Citado na página 18.

ISHII, N. et al. Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. *Journal of Human Genetics*, v. 51, n. 12, p. 1087–1099, 2006. Citado na página 19.

JI, B. et al. Over-expression of XIST, the master gene for X chromosome inactivation, in females with major affective disorders. *EBioMedicine*, v. 2, n. 8, p. 909–918, 2015. Citado na página 19.

JIN, P.; ALISCH, R. S.; WARREN, S. T. RNA and microRNAs in fragile X mental retardation. *Nature Cell Biology*, v. 6, n. 11, p. 1048–1053, 2004. Citado na página 19.

JOHNSON, J. M. et al. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics*, v. 21, n. 2, p. 93–102, 2005. Citado na página 21.

KALVARI, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding rna families. *Nucleic Acids Research*, v. 46, n. D1, p. D335–D342, 2018. Disponível em: <[+http://dx.doi.org/10.1093/nar/gkx1038](http://dx.doi.org/10.1093/nar/gkx1038)>. Citado na página 55.

KANG, Y.-J. et al. Cpc2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research*, v. 45, n. W1, p. W12–W16, 2017. Disponível em: <[+http://dx.doi.org/10.1093/nar/gkx428](http://dx.doi.org/10.1093/nar/gkx428)>. Citado na página 57.

KEENAN, R. J. et al. The signal recognition particle. *Annual Review of Biochemistry*, v. 70, n. 1, p. 755–775, 2001. Citado na página 18.

KIM, V. N.; HAN, J.; SIOMI, M. C. Biogenesis of small RNAs in animals. *Nature Reviews. Molecular cell biology*, v. 10, n. 2, p. 126–39, 2009. Citado na página 27.

KIN, T. et al. fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Research*, v. 35, n. 1, p. D145–D148, 2007. Citado na página 29.

KONG, L. et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, v. 35, n. 2, p. W345–W349, 2007. Citado 6 vezes nas páginas 32, 41, 42, 49, 57 e 74.

KONONENKO, I. Estimating attributes: Analysis and extensions of relief. In: BERGADANO, F.; RAEDT, L. D. (Ed.). *Machine Learning: ECML-94*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994. p. 171–182. ISBN 978-3-540-48365-6. Citado na página 61.

KOTSIANTIS, S. B. Supervised machine learning: a review of classification techniques. *Informatica*, v. 31, p. 249–268, 2007. Citado na página 35.

KOZOMARA, A.; GRIFFITHS-JONES, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, v. 42, n. D1, p. D68–D73, 2014. Citado na página 29.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, Springer Nature, v. 521, n. 7553, p. 436–444, may 2015. Disponível em: <<https://doi.org/10.1038/nature14539>>. Citado na página 52.

LEHMANN, K.; SCHMIDT, U. Group II introns: structure and catalytic versatility of large natural ribozymes. *Critical Reviews in Biochemistry and Molecular Biology*, v. 38, n. 3, p. 249–303, 2003. Citado na página 27.

LI, A.; ZHANG, J.; ZHOU, Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics*, v. 15, n. 1, p. 311, 2014. Citado 8 vezes nas páginas 20, 32, 41, 44, 50, 53, 57 e 74.

LIN, S. L.; MILLER, J.; YING, S. Y. Intronic microRNA (miRNA). *Journal of Biomedicine and Biotechnology*, v. 2006, p. 1–13, 2006. Citado na página 19.

LIU, C. et al. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Research*, v. 33, n. 1, p. D112–D115, 2005. Citado na página 18.

LIU, J.; GOUGH, J.; ROST, B. Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genetics*, v. 2, n. 4, p. e29, 2006. Citado 5 vezes nas páginas 20, 41, 42, 49 e 53.

LIZ, J.; ESTELLER, M. lncRNAs and microRNAs with a role in cancer development. *BBA - Gene Regulatory Mechanisms*, v. 1859, n. 1, p. 169–176, 2016. Citado na página 19.

- LOURO, R.; SMIRNOVA, A. S.; VERJOVSKI-ALMEIDA, S. Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics*, v. 93, n. 4, p. 291–298, 2009. Citado na página 20.
- LOWE, T. M. J. *Combining new computational and traditional experimental methods to identify tRNA and snoRNA gene families*. Dissertação (Mestrado) — Washington University, 1999. Citado na página 18.
- MAIMON, O.; ROKACH, L. *Data Mining and Knowledge Discovery Handbook*. 2. ed. [S.l.]: Springer New York, 2010. Citado na página 30.
- MATERA, A. G.; TERNS, R. M.; TERNS, M. P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol*, v. 8, n. 3, p. 209–220, 2007. Citado na página 27.
- MATIOLI, S. R.; FERNANDES, F. M. C. *Biologia Molecular e Evolução*. 2. ed. [S.l.]: Ribeirão Preto: Holos, Editora / Sociedade Brasileira de Genética, 2012. Citado na página 25.
- MATTICK, J. S.; MAKUNIN, I. V. Non-coding RNA. *Human Molecular Genetics*, v. 15, n. 1, p. 17–29, 2006. Citado na página 19.
- MONARD, M. C.; PRATI, R. C. Aprendizado de máquina simbólico para mineração de dados. In: *XIII Escola Regional de Informática da SBC - Santa Catarina*. 1. ed. Florianópolis: Sociedade Brasileira de Computação, 2005. p. 1–26. ISBN 8576690349. Disponível em: <<http://www.labic.icmc.usp.br/pub/mcmonard/eri05book.pdf>>. Citado na página 36.
- MOREIRA, L. M. (Ed.). *Ciências Genômicas: Fundamentos e Aplicações*. 1. ed. [S.l.]: Ribeirão Preto: Sociedade Brasileira de Genética, 2015. Citado 2 vezes nas páginas 19 e 20.
- NAGAI, K. et al. Structure, function and evolution of the signal recognition particle. *EMBO Journal*, v. 22, n. 14, p. 3479–3485, 2003. Citado na página 18.
- NAWROCKI, E. P. et al. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Research*, v. 43, n. D1, p. D130–D137, 2015. Citado na página 29.
- NOVOA, P. G.; WILLIAMS, K. P. The tmRNA website: reductive evolution of tmRNA in plastids and other endosymbionts. *Nucleic Acids Research*, v. 32, n. 1, p. D104–D108, 2004. Citado na página 18.
- OSBORNE, R. J.; THORNTON, C. A. RNA-dominant diseases. *Human Molecular Genetics*, v. 15, n. 2, p. 162–169, 2006. Citado na página 19.
- PANWAR, B.; ARORA, A.; RAGHAVA, G. P. Prediction and classification of ncRNAs using structural information. *BMC Genomics*, BMC Genomics, v. 15, n. 1, p. 127, 2014. Citado 3 vezes nas páginas 44, 47 e 51.
- PARASKEVOPOULOU, M. D. et al. DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Research*, v. 44, n. D1, p. D231–D238, 2016. Citado na página 29.

- PASCHOAL, A. R. et al. Non-coding transcription characterization and annotation: a guide and web resource for non-coding RNA databases. *RNA Biology*, v. 9, n. 3, p. 274–282, 2012. Citado na página 28.
- PENNISI, E. Encode project writes eulogy for junk dna. *Science*, American Association for the Advancement of Science, v. 337, n. 6099, p. 1159–1161, 2012. ISSN 0036-8075. Disponível em: <<http://science.sciencemag.org/content/337/6099/1159>>. Citado na página 21.
- PETAZZI, P. et al. Dysregulation of the long non-coding RNA transcriptome in a Rett syndrome mouse model. *RNA biology*, v. 10, n. 7, p. 1197–203, 2013. Citado na página 19.
- PICCINELLI, P.; ROSENBLAD, M. A.; SAMUELSSON, T. Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes. *Nucleic Acids Research*, v. 33, n. 14, p. 4485–4495, 2005. Citado na página 18.
- POLESSKAYA, O. O. et al. Novel putative nonprotein-coding RNA gene from 11q14 displays decreased expression in brains of patients with schizophrenia. *Journal of Neuroscience Research*, v. 74, n. 1, p. 111–122, 2003. Citado na página 19.
- QUEK, X. C. et al. lncRNADB v2.0: expanding the reference database for functional long noncoding RNAs. v. 43, n. D1, p. D168–D173, 2015. Citado na página 29.
- QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, n. 1, p. 81–106, Mar 1986. ISSN 1573-0565. Disponível em: <<https://doi.org/10.1007/BF00116251>>. Citado 3 vezes nas páginas 35, 36 e 37.
- REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-Validation. In: _____. *Encyclopedia of Database Systems*. Boston, MA: Springer US, 2009. p. 532–538. Citado na página 37.
- REIS, E. M. et al. As antisense RNA gets intronic. *OMICS*, v. 9, n. 1, p. 2–12, 2005. Citado na página 19.
- REIS, E. M. et al. Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene*, v. 23, n. 39, p. 6684–6692, 2004. Citado na página 19.
- RNAcentral Consortium. RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Research*, v. 43, n. D1, p. D123–D129, 2015. Citado 3 vezes nas páginas 18, 29 e 30.
- ROSENBLAD, M. A. et al. SRPDB: Signal Recognition Particle Database. *Nucleic Acids Research*, v. 31, n. 1, p. 363–364, 2003. Citado na página 18.
- RYMARQUIS, L. A. et al. Diamonds in the rough: mRNA-like non-coding RNAs. *Trends in Plant Science*, v. 13, n. 7, p. 329–334, 2008. Citado na página 20.
- SAEYS, Y. *Feature Selection for Classification of Nucleic Acid Sequences*. 171 p. Tese (Doutorado) — Faculty of Sciences - Universiteit Gent, 2004. Citado na página 33.
- SAEYS, Y.; INZA, I.; LARRANAGA, P. A review of feature selection techniques in bioinformatics. *Bioinformatics*, v. 23, n. 19, p. 2507–2517, 2007. Citado na página 33.

SHEPHERD, J.; IBBA, M. Bacterial transfer RNAs. *FEMS Microbiology Reviews*, The Oxford University Press, v. 39, n. 3, p. 280–300, 2015. Citado 2 vezes nas páginas 27 e 28.

SILVA, R. A. *Classificação de RNAs: expansão do módulo de extração de características*. Monografia (Graduação) — Escola de Artes, Ciências e Humanidades - Universidade de São Paulo, 2009. Citado 4 vezes nas páginas 24, 59, 60 e 74.

SILVA, R. A. *Classificação de RNAs: módulo de extração de características*. Monografia (Graduação) — Escola de Artes, Ciências e Humanidades - Universidade de São Paulo, 2009. Citado 4 vezes nas páginas 24, 59, 60 e 74.

SLEUTELS, F.; ZWART, R.; BARLOW, D. P. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, v. 415, n. 6873, p. 810–813, 2000. Citado na página 19.

SNUSTAD, D. P.; SIMMONS, M. J. *Fundamentos de Genética*. 4. ed. [S.l.]: Rio de Janeiro: Guanabara, 2008. Citado na página 25.

SOKOLOVA, M.; JAPKOWICZ, N.; SZPAKOWICZ, S. Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In: SATTAR, A.; KANG, B.-h. (Ed.). *AI 2006: Advances in Artificial Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 1015–1021. ISBN 978-3-540-49788-2. Citado 2 vezes nas páginas 40 e 63.

SPRINZL, M. et al. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research*, v. 26, n. 1, p. 148–153, 1998. Citado na página 18.

STORZ, G.; ALTUVIA, S.; WASSARMAN, K. M. An abundance of RNA regulators. *Annual Review of Biochemistry*, v. 74, n. 1, p. 199–217, 2005. Citado na página 18.

SUN, K. et al. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics*, BioMed Central Ltd, v. 14, n. 2, p. S7, 2013. Citado 4 vezes nas páginas 43, 47, 50 e 53.

SUN, L. et al. lncRScan-SVM: a tool for predicting long non-coding RNAs using support vector machine. *PLoS ONE*, Public Library of Science, v. 10, n. 10, p. 1–16, 2015. Citado 6 vezes nas páginas 19, 20, 45, 47, 51 e 53.

SUN, L. et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research*, v. 41, n. 17, 2013. Citado 6 vezes nas páginas 32, 41, 43, 50, 57 e 74.

SZYMANSKI, M.; ERDMANN, V. A.; BARCISZEWSKI, J. Noncoding RNAs database (ncRNAdb). *Nucleic Acids Research*, v. 35, n. 1, p. D162–D164, 2007. Citado na página 29.

TAKEDA, J.-i. et al. H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery. *Nucleic Acids Research*, v. 41, n. D1, p. D915–D919, 2013. Citado na página 21.

TAM, W.; BEN-YEHUDA, D.; HAYWARD, W. bic, a novel gene activated by proviral insertions in avian leukosis virus-induced lymphomas, is likely to function through its noncoding RNA. *Molecular and Cellular Biology*, v. 17, n. 3, p. 1490–1502, 1997. Citado na página 18.

- TARCA, A. L. et al. Machine learning and its applications to biology. *PLoS Computational Biology*, v. 3, n. 6, p. e116, 2007. Citado 2 vezes nas páginas 34 e 35.
- VALADKHAN, S. snRNAs as the catalysts of pre-mRNA splicing. *Current Opinion in Chemical Biology*, v. 9, n. 6, p. 603–608, 2005. Citado na página 18.
- VAN ZON, A. et al. The vault complex. *Cell Molecular Life Sciences*, v. 60, n. 9, p. 1828–1837, 2003. Citado na página 18.
- VAPNIK, V. N. *The Nature of Statistical Learning Theory*. 2. ed. New York, NY: Springer New York, 2000. Citado 2 vezes nas páginas 30 e 33.
- VILLEGAS, V. E.; ZAPHIROPOULOS, P. G. Neighboring gene regulation by antisense long non-coding RNAs. *International Journal of Molecular Sciences*, v. 16, n. 2, p. 3251, 2015. Citado 2 vezes nas páginas 27 e 28.
- VOLDERS, P. J. et al. An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Research*, v. 43, n. D1, p. D174–D180, 2015. Citado na página 29.
- WANG, L. et al. CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Research*, v. 41, n. 6, p. e74–e74, 2013. Citado 7 vezes nas páginas 32, 41, 43, 49, 51, 57 e 74.
- WILLINGHAM, A. T. et al. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science*, v. 309, n. 5740, p. 1570–1573, 2005. Citado na página 18.
- WITTKOPP, P. J.; KALAY, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, Nature Publishing Group, v. 13, n. 1, p. 59–69, 2011. Citado na página 27.
- WONG, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, Elsevier, v. 48, n. 9, p. 2839–2846, 2015. Citado na página 38.
- WOOD, A. J.; OAKEY, R. J. Genomic imprinting in mammals: emerging themes and established theories. *PLoS Genetics*, v. 2, n. 11, p. 1677–1685, 2006. Citado na página 19.
- XIAOLONG, Z.; ENDUO, W. Transfer RNA: a dancer between charging and mis-charging for protein biosynthesis. *Science China Life Sciences*, v. 56, n. 10, p. 921–932, 2013. Citado na página 28.
- YOTSUKURA, S. et al. Computational recognition for long non-coding RNA (lncRNA): software and databases. *Briefings in Bioinformatics*, p. bbv114, 2016. Citado 3 vezes nas páginas 19, 27 e 28.
- ZARATIEGUI, M.; IRVINE, D. V.; MARTIENSSSEN, R. A. Noncoding RNAs and gene silencing. *Cell*, v. 128, n. 4, p. 763–776, 2007. Citado na página 19.
- ZERBINO, D. R. et al. Ensembl 2018. *Nucleic Acids Research*, v. 46, n. D1, p. D754–D761, 2018. Disponível em: <[+http://dx.doi.org/10.1093/nar/gkx1098](http://dx.doi.org/10.1093/nar/gkx1098)>. Citado na página 55.
- ZHAO, Y. et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Research*, v. 44, n. D1, p. D203–D208, 2016. Citado na página 29.

ZHENG, L.-L. et al. deepBase v2.0: identification, expression, evolution and function of small RNAs, lncRNAs and circular RNAs from deep-sequencing data. *Nucleic Acids Research*, v. 44, n. D1, p. D196–D202, 2016. Citado na página [29](#).

ZOU, C. et al. Evolutionary and expression signatures of pseudogenes in *Arabidopsis thaliana* and Rice. *Plant Physiology*, v. 151, n. 1, p. 3–151, 2009. Citado na página [20](#).

Apêndice A – Classificadores de RNAs codificantes e não codificantes baseados em aprendizado supervisionado: Protocolo de Revisão

1 Objetivo

Identificar e analisar os classificadores baseados em aprendizado supervisionado, que sejam aplicados à classificação dos RNAs em codificantes ou não codificantes de proteína.

2 Questões de pesquisa

2.1 Questão primária

Quais classificadores baseados em aprendizado supervisionado são aplicados à classificação de RNAs transcritos, conhecidos ou novos, em codificante ou não codificante de proteína?

- **Intervenção:** Identificação e compreensão dos classificadores de RNAs codificantes e não codificantes.
- **Controle:** São utilizados como controle artigos científicos obtidos a partir de buscas exploratórias no google acadêmico e no PubMed. Foram incluídos também artigos sugeridos pelo orientador.
- **População:** Projetos para o desenvolvimento de classificadores de RNAs codificantes e não codificantes.
- **Resultados:** Espera-se obter uma visão profunda e abrangente dos classificadores baseados em aprendizado supervisionado aplicados à classificação de RNAs em codificantes e não codificantes.
- **Aplicação:** Contribuirá com pesquisadores das áreas biológicas e médicas.

2.2 Questões secundárias

- Quais são os algoritmos de classificação utilizados pelos classificadores de RNA?

- Quais são os métodos ou algoritmos utilizados para a seleção de características durante a construção dos classificadores de RNA?
- Quais características são utilizadas pelos classificadores de RNA?
- Qual tipo de sequência foi utilizado para treinar o classificador?
- Qual base de dados foi utilizada para obtenção das sequências?
- Quais tipos de RNA compõem as classes positivas e negativas?
- Qual o método ou algoritmo utilizado para seleção das instâncias que compõe os conjuntos de treinamento e teste?
- Qual o método ou algoritmo utilizado para validação do classificador?
- Quais as medidas utilizadas para avaliar o classificador?
- Qual o método utilizado para comparar o classificador aos classificadores disponíveis na literatura?
- Quais são os pontos fortes e fracos do classificador?
- Para quais espécies o autor sugere um bom desempenho do classificador?

3 Seleção de fontes

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

4 Palavras-chaves

- **Grupo 1:** *“supervised learning”; “pattern recognition”; “machine learning”; “coding potential”; discriminates; classifying; classify; classification; discrimination; distinguish; discriminator; classifier; classifies; distinguishing;*
- **Grupo 2:** *protein-coding; “protein coding”; coding; mrna; mrnas; “messenger rna”; “messenger rnas”;*
- **Grupo 3:** *ncrna; ncarnas; non-coding; noncoding; non-protein-coding; “functional rnas”; lncrna; lncarnas;*
- **Grupo 4:** *review; reviews; virus; dna; patients; biomarkers; biomarker; markers; prognosis; prognostic; diagnosing; diagnosis; diagnostic*

5 *String* de busca

(“supervised learning”[Title/Abstract] OR “pattern recognition”[Title/Abstract] OR “machine learning”[Title/Abstract] OR “coding potential”[Title/Abstract] OR discriminates[Title/Abstract] OR classifying[Title/Abstract] OR classify[Title/Abstract] OR classification[Title/Abstract] OR discrimination[Title/Abstract] OR distinguish[Title/Abstract] OR discriminator[Title/Abstract] OR classifier[Title/Abstract] OR classifies[Title/Abstract] OR distinguishing[Title/Abstract]) AND (protein-coding[Title/Abstract] OR “protein coding”[Title/Abstract] OR coding[Title/Abstract] OR mrna[Title/Abstract] OR mrnas [Title/Abstract] OR “messenger rna”[Title/Abstract] OR “messenger rnas”[Title/Abstract]) AND (ncrna[Title/Abstract] OR ncrnas[Title/Abstract] OR non-coding[Title/Abstract] OR noncoding[Title/Abstract] OR non-protein-coding[Title/Abstract] OR “functional rnas”[Title/Abstract] OR lncrna[Title/Abstract] OR lncrnas[Title/Abstract]) NOT (review[Title/Abstract] OR reviews[Title/Abstract] OR virus[Title/Abstract] OR dna [Title/Abstract] OR patients[Title/Abstract] OR biomarkers[Title/Abstract] OR biomarker[Title/Abstract] OR markers[Title/Abstract] OR prognosis[Title/Abstract] OR prognostic[Title/Abstract] OR diagnosing[Title/Abstract] OR diagnosis[Title/Abstract] OR diagnostic[Title/Abstract])

6 Lista de fontes

PUBMED (<http://www.ncbi.nlm.nih.gov/pubmed>)

7 Tipo dos artigos

Serão considerados artigos de periódicos que atendam aos requisitos definidos nos critérios de inclusão, exclusão e de qualidade.

8 Idioma dos artigos

Inglês e português.

9 Critérios de inclusão e exclusão dos trabalhos

9.1 Critérios de inclusão

- (I1) Trabalho que propõe um classificador baseado em aprendizado supervisionado para a classificação de RNAs em codificante ou não codificante de proteína.

9.2 Critérios de exclusão

- (E1) Trabalho que NÃO propõe um classificador baseado em aprendizado supervisionado para a classificação de RNAs em codificante ou não codificante de proteína.
- (E2) Classificador específico para pequenos ncRNAs.
- (E3) Trabalho não disponível na íntegra.
- (E4) Trabalho que não esteja em inglês ou português.

10 Critério de qualidade dos estudos primários

Os trabalhos serão incluídos de acordo com os critérios de inclusão e exclusão. Além disso, a depender da necessidade ou do número de resultados critérios de qualidade poderão ser aplicados para facilitar a exclusão de artigos.

11 Processo de seleção dos estudos primários

Palavras-chave serão definidas a partir da análise dos títulos, resumos e palavras-chave dos artigos controle. *Strings* de busca, construídas a partir das palavras-chave definidas e de seus sinônimos, serão calibradas de forma que possibilite a recuperação do máximo possível de artigos controle. Uma vez calibrada, a *string* será submetida às máquinas de busca. Durante o procedimento de recuperação das informações serão priorizadas as ocorrências de palavras-chave em títulos e resumos.

A partir da leitura do título e resumo dos artigos serão atribuídos critérios de exclusão e inclusão. Se ao menos um critério de exclusão for atribuído o artigo será excluído. O trabalho será selecionado se confirmada sua relevância pelo principal revisor (aluno). Se houver dúvida da relevância os demais revisores serão consultados. Uma vez

confirmado sua relevância o trabalho será lido em sua totalidade e terá suas informações extraídas.

12 Estratégia de extração de informação

Após definidos os trabalhos definitivamente incluídos, estes serão lidos na íntegra. O revisor fará um resumo de cada um deles, destacando os métodos utilizados para a avaliação e parâmetros considerados, quando for o caso.

Serão preenchidos “formulários de extração de dados” para cada artigo incluído. Além das informações básicas (dados bibliográficos, data de publicação, *abstract*, entre outros), esses formulários deverão conter a síntese do trabalho, redigida pelo pesquisador que conduzirá a Revisão Sistemática e reflexões pessoais do mesmo a respeito do conteúdo e das conclusões do estudo.

13 Sumarização dos resultados

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

Apêndice B – Importância das características para um modelo de classificação baseado em Random Forest

Tabela 20 – Principais características em ordem decrescente de importância, calculada com base no decréscimo médio da acurácia durante a construção de um modelo de classificação baseado em Random Forest. O valor “Importância” corresponde ao Decréscimo Médio da Acurácia em porcentagem

Característica	Descrição	Importância
CPC1.frame_score	Frame score	9.86%
CPC1.hit_seq_count	Hit seq count	6.40%
CPC1.hit_score	Hit score	6.23%
CPC1.Score	Score	5.82%
CPC1.hit_HSP_count	Hit HSP count	3.65%
CPC1.frame_score2	Frame score2	3.64%
CPAT.Hexamer	Viés do uso de hexamers	2.85%
CPC1.CDSLength	Tamanho da CDS	2.03%
NEW.blastx	E-value retornado pelo Blastx	1.48%
NEW.X_	Distribuição de aminoácido	1.19%
CNCI.M_length	Tamanho da MLCDS	1.11%
CPC1.Used	Used	0.96%
NEW.orf.cover	Cobertura da ORF sobre a sequência	0.77%
NEW.blastp	E-value retornado pelo Blastp	0.68%
CPAT.mRNA	Tamanho da sequência nucleotídica	0.47%
CNCI.M_score	S-score da MLCDS	0.37%
NEW.utr5length	Tamanho da UTR 5	0.32%
CNCI.score_distance	Diferença entre o S-score da melhor MLCDS e das outras cinco MLCDS dividida por 5	0.30%
NEW.start.utr5out	Quantidade de códons de iniciação fora da UTR 5	0.28%
NEW.utr3length	Tamanho da UTR 3	0.20%
NEW.start.utr5in	Quantidade de códons de iniciação dentro da UTR 5	0.20%
NEW.gcfreq	Porcentagem de GC	0.19%
CPAT.Fickett	Estatística Fickett TESTCODE	0.18%
CNCI.atg	Frequência do trinucleotídeo ATG na MLCDS	0.18%
CNCI.acg	Frequência do trinucleotídeo ACG na MLCDS	0.16%
PLEK.14	* Frequências de uso de K-mer	0.15%
CNCI.gcc	Frequência do trinucleotídeo GCC na MLCDS	0.14%
PLEK.747	* Frequências de uso de K-mer	0.12%
PLEK.58	* Frequências de uso de K-mer	0.12%
PLEK.11	* Frequências de uso de K-mer	0.11%
CPC2.Fickett_score	Fickett score	0.10%
PLEK.942	* Frequências de uso de K-mer	0.10%
PLEK.186	* Frequências de uso de K-mer	0.08%
CNCI.gct	Frequência do trinucleotídeo GCT na MLCDS	0.08%
CNCI.tcc	Frequência do trinucleotídeo TCC na MLCDS	0.08%

Tabela 20 – (Continuação)

Característica	Descrição	Importância
CNCI.agc	Frequência do trinucleotídeo AGC na MLCDS	0.07%
PLEK.235	* Frequências de uso de K-mer	0.07%
PLEK.62	* Frequências de uso de K-mer	0.06%
CNCI.aca	Frequência do trinucleotídeo ACA na MLCDS	0.05%
CNCI.ctg	Frequência do trinucleotídeo CTG na MLCDS	0.05%
PLEK.46	* Frequências de uso de K-mer	0.05%
CNCI.cag	Frequência do trinucleotídeo CAG na MLCDS	0.04%
PLEK.433	* Frequências de uso de K-mer	0.04%
PLEK.853	* Frequências de uso de K-mer	0.04%
PLEK.17	* Frequências de uso de K-mer	0.04%
PLEK.59	* Frequências de uso de K-mer	0.04%
CNCI.length_precent	Razão do tamanho da melhor MLCDS pela soma dos tamanhos das seis MLCDS	0.04%
CNCI.gca	Frequência do trinucleotídeo GCA na MLCDS	0.04%
PLEK.78	* Frequências de uso de K-mer	0.04%
PLEK.500	* Frequências de uso de K-mer	0.04%
PLEK.958	* Frequências de uso de K-mer	0.03%
CNCI.aac	Frequência do trinucleotídeo AAC na MLCDS	0.03%
NEW.TTT	Frequência do trinucleotídeo TTT	0.03%
PLEK.8	* Frequências de uso de K-mer	0.03%
PLEK.69	* Frequências de uso de K-mer	0.03%
CNCI.aat	Frequência do trinucleotídeo AAT na MLCDS	0.03%
NEW.I	Distribuição do aminoácido I	0.03%
NEW.A	Distribuição do aminoácido A	0.03%
PLEK.771	* Frequências de uso de K-mer	0.03%
CPC2.isoelectric_point	Ponto isoelétrico	0.03%
PLEK.692	* Frequências de uso de K-mer	0.03%
CNCI.ggc	Frequência do trinucleotídeo GGC na MLCDS	0.03%
PLEK.1359	* Frequências de uso de K-mer	0.03%
PLEK.1364	* Frequências de uso de K-mer	0.03%
PLEK.190	* Frequências de uso de K-mer	0.03%
PLEK.454	* Frequências de uso de K-mer	0.03%
PLEK.55	* Frequências de uso de K-mer	0.03%
PLEK.2	* Frequências de uso de K-mer	0.03%
NEW.E	Distribuição do aminoácido E	0.03%
PLEK.174	* Frequências de uso de K-mer	0.03%
PLEK.24	* Frequências de uso de K-mer	0.03%
CNCI.ata	Frequência do trinucleotídeo ATA na MLCDS	0.03%
PLEK.698	* Frequências de uso de K-mer	0.03%
PLEK.239	* Frequências de uso de K-mer	0.03%
PLEK.13	* Frequências de uso de K-mer	0.03%
PLEK.763	* Frequências de uso de K-mer	0.02%
PLEK.870	* Frequências de uso de K-mer	0.02%
PLEK.339	* Frequências de uso de K-mer	0.02%
PLEK.1	* Frequências de uso de K-mer	0.02%
CNCI.atc	Frequência do trinucleotídeo ATC na MLCDS	0.02%
PLEK.251	* Frequências de uso de K-mer	0.02%
PLEK.5	* Frequências de uso de K-mer	0.02%
CNCI.aga	Frequência do trinucleotídeo AGA na MLCDS	0.02%
PLEK.875	* Frequências de uso de K-mer	0.02%
NEW.ACA	Frequência do trinucleotídeo ACA	0.02%

Tabela 20 – (Continuação)

Característica	Descrição	Importância
CNCl.agt	Frequência do trinucleotídeo AGT na MLCDS	0.02%
NEW.F	Distribuição do aminoácido F	0.02%
NEW.G	Distribuição do aminoácido G	0.02%
PLEK.1022	* Frequências de uso de K-mer	0.02%
PLEK.954	* Frequências de uso de K-mer	0.02%
NEW.W	Distribuição do aminoácido W	0.02%
NEW.TT	Frequência do dinucleotídeo TT	0.02%
CNCl.aaa	Frequência do trinucleotídeo AAA na MLCDS	0.02%
CNCl.caa	Frequência do trinucleotídeo CAA na MLCDS	0.02%
PLEK.1079	* Frequências de uso de K-mer	0.02%
PLEK.250	* Frequências de uso de K-mer	0.02%
CNCl.cca	Frequência do trinucleotídeo CCA na MLCDS	0.02%
PLEK.340	* Frequências de uso de K-mer	0.02%
PLEK.1136	* Frequências de uso de K-mer	0.02%
CNCl.ttt	Frequência do trinucleotídeo TTT na MLCDS	0.02%
CNCl.tct	Frequência do trinucleotídeo TCT na MLCDS	0.02%
PLEK.828	* Frequências de uso de K-mer	0.02%
PLEK.594	* Frequências de uso de K-mer	0.02%
NEW.CGC	Frequência do trinucleotídeo CGC	0.02%
PLEK.57	* Frequências de uso de K-mer	0.02%
NEW.D	Distribuição do aminoácido D	0.02%
CNCl.cta	Frequência do trinucleotídeo CTA na MLCDS	0.02%
PLEK.416	* Frequências de uso de K-mer	0.02%
PLEK.175	* Frequências de uso de K-mer	0.02%
CNCl.tgg	Frequência do trinucleotídeo TGG na MLCDS	0.02%
PLEK.1248	* Frequências de uso de K-mer	0.02%
PLEK.87	* Frequências de uso de K-mer	0.02%
CNCl.cgt	Frequência do trinucleotídeo CGT na MLCDS	0.02%
PLEK.1107	* Frequências de uso de K-mer	0.02%
PLEK.402	* Frequências de uso de K-mer	0.02%
PLEK.1224	* Frequências de uso de K-mer	0.02%
PLEK.404	* Frequências de uso de K-mer	0.02%
PLEK.165	* Frequências de uso de K-mer	0.02%
CNCl.cga	Frequência do trinucleotídeo CGA na MLCDS	0.02%
PLEK.740	* Frequências de uso de K-mer	0.02%
PLEK.794	* Frequências de uso de K-mer	0.02%
PLEK.3	* Frequências de uso de K-mer	0.02%
PLEK.940	* Frequências de uso de K-mer	0.02%
CNCl.gta	Frequência do trinucleotídeo GTA na MLCDS	0.02%
CNCl.tca	Frequência do trinucleotídeo TCA na MLCDS	0.02%
PLEK.39	* Frequências de uso de K-mer	0.02%
CNCl.att	Frequência do trinucleotídeo ATT na MLCDS	0.02%
PLEK.331	* Frequências de uso de K-mer	0.02%
PLEK.749	* Frequências de uso de K-mer	0.02%
PLEK.53	* Frequências de uso de K-mer	0.02%
CNCl.gtc	Frequência do trinucleotídeo GTC na MLCDS	0.02%
PLEK.43	* Frequências de uso de K-mer	0.01%
PLEK.1198	* Frequências de uso de K-mer	0.01%
PLEK.644	* Frequências de uso de K-mer	0.01%
PLEK.233	* Frequências de uso de K-mer	0.01%
PLEK.581	* Frequências de uso de K-mer	0.01%

Tabela 20 – (Continuação)

Característica	Descrição	Importância
PLEK.469	* Frequências de uso de K-mer	0.01%
CNCL.cgc	Frequência do trinucleotídeo CGC na MLCDS	0.01%
PLEK.20	* Frequências de uso de K-mer	0.01%
PLEK.224	* Frequências de uso de K-mer	0.01%
NEW.TTA	Frequência do trinucleotídeo TTA	0.01%
PLEK.9	* Frequências de uso de K-mer	0.01%
PLEK.980	* Frequências de uso de K-mer	0.01%
PLEK.801	* Frequências de uso de K-mer	0.01%

** Não foi possível obter os rótulos das características a partir da saída dos módulos do classificador PLEK. Por esta razão, atribuímos um identificador numérico para estas características*

Fonte: Clebiano da Costa Sá, 2018