Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Classificação de RNAs não-codificadores Longos Intergênicos usando Máquina de Vetores de Suporte: um Estudo de Caso para a Cana-de-açúcar

Lucas Maciel Vieira

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientadora
Prof.ª Dr.ª Maria Emilia M. T. Walter

Brasília
2016

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Classificação de RNAs não-codificadores Longos Intergênicos usando Máquina de Vetores de Suporte: um Estudo de Caso para a Cana-de-açúcar

Lucas Maciel Vieira

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof.ª Dr.ª Maria Emilia M. T. Walter (Orientadora)
CIC/UnB

Prof.ª Dr.ª Célia Ghedini Ralha
CIC/UnB

MsC. João Victor de Araújo Oliveira
CIC/UnB

Prof. Dr. Ricardo Zelenovski
Coordenador do Curso de Engenharia da Computação

Brasília, 12 de janeiro de 2016

Dedicatória

Dedico esse trabalho, primeiramente à meus pais, por sempre me incentivarem nos estudos e por me apoiarem ao longo dessa trajetória, sempre dando o suporte necessário. Aos meus amigos, principalmente meus colegas de faculdade e de intercâmbio, que trilharam esse caminho acadêmico comigo e propiciaram grandes momentos de alegria nessa jornada.

"We can only see a short distance ahead, but we can see plenty there that needs to be done." Alan Turing

Agradecimentos

Agradeço a todas as pessoas que encontrei durante minha jornada acadêmia, pois mesmo que em um breve momento, todas elas contribuíram para chegar ao término dessa estapa. Principalmente a minha orientadora Maria Emília que me ajudou no desenvolvimento do projeto e por me apresentar essa área de estudo tão interessante.

Resumo

Dentre os RNAs, temos os que expressam proteínas, e aqueles que, embora não participando da síntese proteica, realizam funções importantes nas células, sendo denominados de RNAs não-codificadores (ncRNAs). Dentre os ncRNAs, existem os RNAs não-codificadores longos intergênicos (*long intergenic ncRNAs* - lincRNAs), que estão localizados em regiões intergênicas, e podem desempenhar importantes papéis na regulação gênica e em doenças. Embora existam vários projetos relacionados a lincRNAs, tanto na parte da biologia molecular quanto computacional, não há métodos amplamente usados para sua predição. Neste contexto, validando características obtidas na literatura, criamos um modelo baseado em máquinas de vetores de suporte (*Support Vector Machine* - SVM) para predizer lincRNAs. Desenvolvemos dois estudos de caso, um para calcular o desempenho do modelo usando *Mus musculus* (camundongo) e *Homo sapiens* (humano) e outro para predizer lincRNAs em *Saccharum officinarum* (cana-de-açúcar). Os experimentos mostraram que o modelo tem boa acurácia, em camundongos 90%, em humanos 99% e em ambos simultaneamente 91%, que são melhores resultados, quando comparados ao iSeeRNA. Para a cana-de-açúcar, o método predisse 67 lincRNAs, usando um *pipeline* construído especialmente para predizer lincRNAs, que inclui o modelo SVM treinado com características extraídas de plantas.

Palavras-chave: RNAs não-codificadores longos intergênicos, RNAs não-codificadores longos, RNAs não-codificadores, Aprendizagem de Máquina, Máquinas de Vetores de Suporte

Abstract

Among RNAs, some are involved in protein expression, and some other, although not participating in protein synthesis, perform important functions in cells, called non-coding RNAs (ncRNAs). Some functions of ncRNAs are: to catalyze chemical reactions and act in regulation of other RNAs. Generically, we can classify ncRNAs into two classes: small (*small ncRNAs*), having sizes between 20 and 300 nucleotides and presenting known features; and longs (*long ncRNAs* - lncRNAs), which have sizes larger than 200 nucleotides and small protein synthesis capacity, today not entirely known. Among the lncRNAs, there are the so called long intergenic non-coding RNAs (lincRNAs), those located in intergenic regions, which play important roles in gene regulation and diseases. Although there are many projects related to lincRNAs, both in molecular biology and in computational systems, there are no methods broadly used to predict lincRNAs. In this context, validating features extracted from literature, we created a model based on Support Vector Machine (SVM) to predict lincRNAs. Two case studies were developed, the first one to verify the performance of the model, using *Mus musculus* (mouse) and *Homo sapiens* (human), and the other one to predict lincRNAs in *Saccharum officinarum* (sugarcane). The experiments showed that the model presented good accuracy, in mouse 90%, humans 99%, and in both simultaneously 91%, which were better when compared to iSeeRNA. For sugarcane, the method predicted 67 lincRNAs, using a specially designed pipeline to predict lincRNAs, including the SVM model trained with features extracted from plants.

Keywords: long intergenic non-coding RNAs, long non-coding RNAs, non-coding RNAs, machine learning, Support Vector Machine

Sumário

1 Introdução **1**

1.1 Motivação 2
1.2 Problema 2
1.3 Objetivos 3
1.4 Descrição dos capítulos 3

2 RNAs não-codificadores **4**

2.1 Biologia Molecular 4
2.1.1 DNA 4
2.1.2 RNA 7
2.1.3 Dogma Central da Biologia Molecular 8
2.2 Sequenciamento e Bioinformática 13
2.3 Classificação de RNAs 19
2.3.1 NcRNAs pequenos 19
2.3.2 NcRNAs longos 22
2.4 LincRNAs 23
2.4.1 Métodos de identificação biológicos e computacionais 24
2.4.2 Bancos de dados 29

3 Aprendizagem de Máquina **31**

3.1 Abordagens 31
3.1.1 Aprendizagem não-supervisionada 31
3.1.2 Aprendizagem supervisionada 32
3.1.3 Aprendizagem por reforço 34
3.1.4 Aprendizagem semi-supervisionada 35
3.2 *SVM* 35
3.2.1 Conceitos básicos 35
3.2.2 Descrição 35
3.2.3 *Kernel* 38

3.2.4 *K-fold cross-validation* 39

4 Método de predição de lincRNAs **41**
4.1 Descrição geral 41
4.2 Estudo de Caso 1: humanos e camundongos 42
4.2.1 Dados 42
4.2.2 Características 42
4.2.3 Extração das características 43
4.2.4 Opções de Treinamento 43
4.2.5 Testes 44
4.3 Estudo de caso 2: Cana-de-açúcar 44
4.3.1 Informações sobre cana-de-açúcar 45
4.3.2 LincRNAs na cana-de-açúcar 45

5 Resultados **48**
5.1 Performance 48
5.1.1 Camundongo 48
5.1.2 Humano 51
5.1.3 Humano + Camundongo 54
5.1.4 Comparação de performance com o iSeeRNA 57
5.2 LincRNAs em cana-de-açúcar 59

6 Conclusão **62**
6.1 Contribuições 62
6.2 Trabalhos Futuros 63

Referências **64**

Anexo **70**

I Informações detalhadas dos lincRNAs da cana-de-açúcar **71**

Lista de Figuras

2.1 Os quatro tipos de nucleotídeos presentes no DNA [10]. 5
2.2 Desoxirribose 5
2.3 Cadeia de nucleotídeos formada pela ligação dos grupos fosfatos [10]. . . . 6
2.4 Dupla fita de DNA, ligada por bases complementares [10]. 6
2.5 Fita de DNA, onde as regiões mais escuras indicam os genes [28]. 7
2.6 Ribose 7
2.7 Os quatro tipos de nucleotídeos presentes no RNA [22]. 8
2.8 Fitas de RNA ligadas na mesma molécula 8
2.9 Dogma Central da Biologia Molecular, que explica o processo de síntese de proteínas a partir das informações armazenadas no DNA por diferentes tipos de RNA [19]. 9
2.10 Estruturas de um gene procarioto e de um gene eucarioto [11]. 10
2.11 Processo de *splicing* em eucarioto [25] 11
2.12 Processo de tradução [24]. 11
2.13 Código genético 12
2.14 Processo de sequenciamento utilizado pelo sequenciador *Illumina* [39]. . . . 14
2.15 Exemplo de *pipeline*. 15
2.16 Exemplo de arquivo *fasta*. 15
2.17 Exemplo de arquivo *fastq*. 16
2.18 Qualidade das sequências 16
2.19 Montagem por referência 17
2.20 Montagem *de novo* 17
2.21 Visão geral do processo de anotação 18
2.22 Estrutura secundária do tRNA [27]. 20
2.23 Estrutura secundária do rRNA [23]. 20
2.24 Estrutura secundária do snoRNA [15]. 21
2.25 Estrutura secundária do miRNA [23]. 21
2.26 Categorias de lncRNA 22
2.27 Modelo de funções propostas a lincRNAs [90]. 23

2.28 Métodos utilizados para classificação de lincRNAs 25
2.29 Modelo para classificação de lincRNAs [90]. 26
2.30 Uso de microarranjo para identificar lincRNAs 27
2.31 *Pipeline* do iSeeRNA [83]. 27
2.32 Ferramenta *online* do iSeeRNA [14]. 28
2.33 Banco Rfam [21]. 29
2.34 Exemplo de distribuição de espécie no Banco Rfam [21]. 29
2.35 Banco lncRNADisease [17]. 30

3.1 *Clustering* hierárquico 32
3.2 Exemplo do *k-means*, com dois grupos, cada um com seu centróide [16]. . 33
3.3 Ciclo de iteração de aprendizagem por reforço (adaptado de [84]). 34
3.4 Exemplo de vetores de suporte de dimensão 2 [1]. 36
3.5 SVM- Margem de separação 36
3.6 Exemplo de SVM com separador em três dimensções [26]. 37
3.7 Espaço de características 38
3.8 Exemplo de *K-fold cross-validation* com $k = 5$ [3]. 40

4.1 *Pipeline* para classificação de lincRNAs. 41
4.2 Processo de extração de características dos transcritos de entrada. 43
4.3 *Pipeline* para classificação de lincRNAs em humanos e camundongos. . . . 44
4.4 Árvore filogenética da cana-de-açúcar 45
4.5 *Pipeline* para classificação de lincRNAS na cana-de-açúcar. 47

5.1 Performance do modelo SVM para o teste 1. 49
5.2 Performance do modelo SVM para os testes 2 e 3. 51
5.3 Performance do modelo SVM para o teste 1. 52
5.4 Performance do modelo SVM para os testes 2 e 3. 54
5.5 Performance do modelo SVM para o teste 1. 56
5.6 Performance do modelo SVM para os testes 2 e 3. 58

Lista de Tabelas

2.1 Aminoácidos 12
2.2 Alguns tipos de RNAs não-codificadores pequenos 21
2.3 Bancos de dados 30

3.1 Tabela de contigência 33
3.2 Tabela de *kernels* mais utilizados. 39

5.1 Tabela de contingência de Conservação 48
5.2 Tabela de contingência de ORFs 49
5.3 Tabela de contingência de Frequências de nucleotídeos 49
5.4 Teste 1 - Performance do modelo SVM para o camundongo 49
5.5 Tabela de contingência de Conservação + ORFs 50
5.6 Tabela de contingência conjunto Conservação + Frequências 50
5.7 Tabela de contingência conjunto ORFs + Frequências 50
5.8 Teste 2 - Performance modelo SVM camundongo 50
5.9 Tabela de contingência do camundongo 51
5.10 Teste 3 - Performance do modelo SVM para o camundongo 51
5.11 Tabela de contingência de Conservação 51
5.12 Tabela de contingência de ORFs 52
5.13 Tabela de contingência de Frequências de nucleotídeos 52
5.14 Teste 1 - Performance do modelo SVM para humanos 52
5.15 Tabela de contingência de Conservação + ORFs 53
5.16 Tabela de contingência de Conservação + Frequências 53
5.17 Tabela de contingência de ORFs + Frequências 53
5.18 Teste 2 - Performance do modelo SVM para humanos 53
5.19 Tabela de contingência de humanos 54
5.20 Teste 3 - Performance do modelo SVM para humanos 54
5.21 Tabela de contingência de Conservação 55
5.22 Tabela de contingência de ORFs 55
5.23 Tabela de contingência de Frequências de nucleotídeos 55

5.24 Teste 1 - Performance do modelo SVM para humanos + camundongos . . . 55
5.25 Tabela de contingência de Conservação + ORFs 56
5.26 Tabela de contingência de Conservação + Frequências 56
5.27 Tabela de contingência de ORFs + Frequências 56
5.28 Teste 2 - Performance do modelo SVM para humanos + camundongos . . . 57
5.29 Tabela de contingência Camundongo 57
5.30 Teste 3 - Performance do modelo SVM para humanos + camundongos . . 57
5.31 Tabela de contingência do iSeeRNA para humanos 58
5.32 Tabela de contingência do iSeeRNA para camundongos 59
5.33 Tabela de contingência modelo SVM Cana 60
5.34 Performance modelo SVM cana 60
5.35 Dados do estudo de caso da cana-de-açúcar 61

I.1 Transcritos não-codificadores mapeados em regiões intergênicas 71
I.2 LincRNAs preditos pelo modelo SVM 72
I.3 LincRNAs anotados como lncRNAs pelo BLAST 73
I.4 Predição de lincRNAs da cana-de-açúcar 74

Capítulo 1

Introdução

Desde que a estrutura de dupla hélice da molécula de DNA foi proposta por Watson e Crick [94], diversos campos de estudo relacionados à investigação dessa molécula receberam grandes avanços. A Biologia Molecular, um desses campos, é o ramo da Biologia que busca entender as estruturas e funções de proteínas e ácidos nucléicos [78].

Os ácidos nucléicos têm a função principal de armazenar informação necessária e prover mecanismos para a criação de proteínas, e também de possibilitar a transferência desta informação para outros organismos, por meio de processos de reprodução celular [78]. Na natureza encontramos dois tipos de ácidos nucléicos: o DNA (ácido desoxirribonucléico) e o RNA (ácido ribonucléico). O DNA possui informações suficientes para gerar aminoácidos e diversas moléculas de RNA. Dentre os RNAs, temos aqueles que expressam proteínas e outros que não originam proteínas, mas realizam funções importantes nas células, denominados de RNAs não-codificadores (ncRNAs).

Proteína é uma cadeia de moléculas (aminoácidos) geradas por RNAs codificadores, que desempenha diferentes papéis nos seres vivos, tais como transporte de nutrientes, aceleração de reações químicas e construção de estuturas nas células [32].

Os ncRNAS agem diretamente na célula, por exemplo, como catalisadores de reações químicas e em diversos papéis regulatórios [95]. Podemos encontrar na literatura [44] uma classificação genérica de ncRNAs em: pequenos (*small ncRNAs*), que possuem características conhecidas e tamanhos pequenos (20 a 300 nucleotideos); e longos (*long ncRNAs* - lncRNAs), que apresentam tamanhos maiores do que 200 nucleotideos e pouca capacidade de síntese de proteínas, sendo os transcritos menos conhecidos atualmente [68, 67]. Dentre as classes de lncRNAs, existem os RNAs não-codificadores longos intergênicos (*long intergenic ncRNAs* - lincRNAs), que são transcritos localizados na região intergênica, e podem desempenhar vários papéis na regulação gênica e em outros processos celulares [90].

Pesquisas mostram que, no genoma humano, menos de 2% do material genético é transcrito em RNAs codificadores de proteínas, sendo que uma significativa parcela do

material genético é transcrito em diversos tipos de ncRNAs [89]. Por outro lado, em plantas, ncRNAs, como os lncRNAs, são pouco conhecidos, embora eles sejam importantes componentes em seus transcritomas [85].

Com o avanço dos estudos e das tecnologias na área da Biologia Molecular, projetos que possuem o intuito de analisar DNAs, RNAs e proteínas, característicos de diversos organismos pelo mundo, criaram um grande volume de dados biológicos [36, 73]. Projetos que buscam analisar o conjunto completo de RNAs em um dado organismo são chamados projetos transcritoma, e os que buscam analisar o conjunto de cadeias de DNA (cromossomos) são chamados projetos genoma. Ambos os tipos de projetos utilizam máquinas conhecidas como sequenciadores, que identificam a ordem dos nucleotídeos em uma dada sequência de DNA/RNA produzidas nos laboratórios de Biologia Molecular. Exemplos de sequenciadores de alto desempenho muito utilizados são: *Illumina* [35], *DNA nanoball sequencing* [69] e *Helioscope single molecule sequencing* [88].

No Instituto de Ciências Biomédicas da UFRJ, o Prof. Paulo Cavalcanti Gomes Ferreira vem trabalhando com plantas, em particular com a cana-de-açúcar. A cana-de-açúcar é importante em diversos aspectos, por exemplo, é usada para produção de açúcar, como matéria prima para fabricação de papel e também na alimentação de animais [59, 41]. A cana-de-açúcar possui um dos genomas conhecidos mais complexos em plantas [53]. Diversos estudos que buscam evidenciar mecanismos de regulação nesse organismo foram feitos, revelando papéis importantes de ncRNAs, tais como *micro RNAs* (miRNAs) e *small interfiring RNAs* (siRNAs) [37, 86, 87].

1.1 Motivação

Pesquisas em lncRNAs vem ocorrendo cada vez em maior quantidade, devido à participação de lncRNAs em vários processos importantes nas células, como regulação de expressão gênica [65]. Estudos recentes apontam importantes papéis funcionais a transcritos de DNA que não expressam proteínas, presentes em regiões intergênicas, lincRNAs. Entretanto, não há métodos amplamente usados para identificação de lincRNAs, embora existam algoritmos [83] e bancos de dados [7, 12, 17] de lincRNAs. Além disso, para o caso específico da cana-de-açúcar, a classificação de lincRNAs permitiria descobrir possíveis mecanismos de regulação celular e expressão diferencial.

1.2 Problema

Não há métodos amplamente utilizados no mundo para classificação de lincRNAs, e em particular não há classificação de lincRNAs em cana-de-açúcar.

1.3 Objetivos

Os objetivo principal deste trabalho é desenvolver um método baseado em Máquinas de Vetores de Suporte (*Support Vector Machine* - SVM) para classificar lincRNAs em humanos, camundongos e em cana-de-açúcar. Os objetivos específicos são:

- Identificar características que possam ser usadas no SVM;
- Implementar o método baseado em SVM para identificar lincRNAs;
- Criar banco de dados de lincRNAs de boa qualidade, para gerar o modelo SVM;
- Realizar um estudo de caso em humanos e camundongos para analisar o desempenho do modelo SVM;
- Realizar um estudo de caso com transcritos da cana-de-açúcar para classificar lincRNAs.

1.4 Descrição dos capítulos

No Capítulo 2, inicialmente serão apresentados conceitos básicos de Biologia Molecular e de Bioinformática. Em seguida, são descritos RNAs não-codificadores, suas classificações, funções e métodos de classificação computacionais, além de bancos de dados que contêm dados de ncRNAs.

No Capítulo 3, discutiremos noções básicas de Aprendizagem de Máquina e em seguida mostraremos o método SVM, que será usado neste projeto.

No Capítulo 4, será proposto o método para classificar lincRNAs, baseado em SVM, usando características obtidas da literatura.

No Capítulo 5, serão utilizados dados de humanos e camundongos para treinar e testar o método proposto e medir sua performance. Em seguida, será discutido um estudo de caso com transcritos da cana-de-açúcar.

Finalmente, no Capítulo 6, concluiremos este trabalho e sugeriremos trabalhos futuros.

Capítulo 2

RNAs não-codificadores

Neste capítulo, conceitos básicos de Biologia Molecular serão apresentados, com foco em RNAs não-codificadores. Na Seção 2.1, descreveremos os ácidos nucleicos, as proteínas e o Dogma Central da Biologia Molecular. Na Seção 2.2, serão descritos brevemente sequenciadores automáticos e *pipelines* de Bioinformática. Na Seção 2.3, algumas classificações de ncRNAs serão mostradas, explicitando as diferenças entre as diversas classes. Na Seção 2.4, serão detalhados lincRNAs, métodos computacionais para predizê-los e bancos de dados contendo lincRNAs.

2.1 Biologia Molecular

A Biologia Molecular é o ramo da Biologia que busca entender as estruturas e funções de proteínas e ácidos nucléicos [78]. Os ácidos nucléicos têm a função principal de armazenar informação necessária e prover mecanismos para a produção de proteínas, e também de possibilitar a transferência desta informação para os descendentes, por meio de processos de reprodução celular [78]. Na natureza, encontramos dois tipos de ácidos nucléicos: o DNA (ácido desoxirribonucléico) e o RNA (ácido ribonucléico).

2.1.1 DNA

O DNA é composto por nucleotídeos, formados por moléculas de fosfato, desoxirribose e uma base nitrogenda. Existem quatro tipos diferentes de bases nitrogenadas: Adenina-A, Guanina-G, Timina-T e Citosina-C. Podemos ver exemplos dos nucleotídeos do DNA na Figura 2.1.

A desoxirribose presente no nucleotídeo possui 5 átomos de carbono (1' a 5'), com o carbono 2' ligado a um hidrogênio (Figura 2.2).

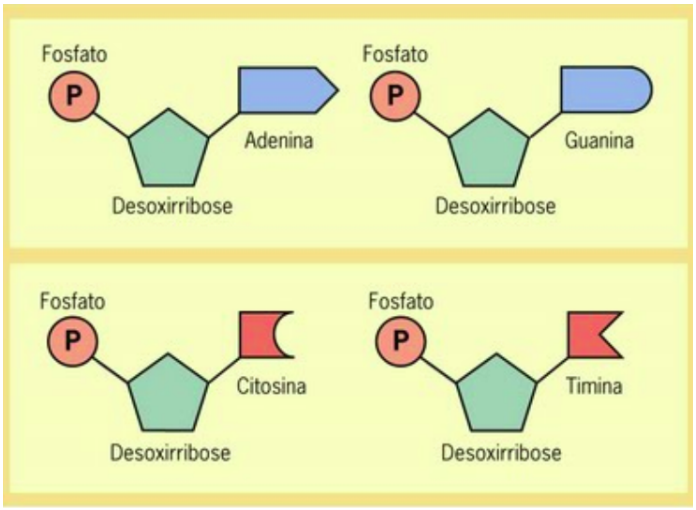

Figura 2.1: Os quatro tipos de nucleotídeos presentes no DNA [10].

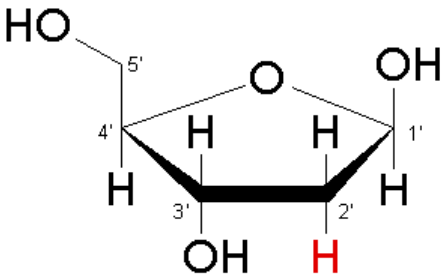

Figura 2.2: Molécula de desoxirribose, com cinco átomos de carbono (1' a 5'). Observe que o carbono 2' liga-se com um átomo de hidrogênio (H) [2].

Dois nucleotídeos unem-se por meio de ligações dos grupos fosfatos da seguinte forma: o carbono 3' do primeiro nucleotídeo liga-se a um grupo fosfato, que se liga ao carbono 5' do próximo nucleotídeo. A ligação dos nucleotídeos cria uma cadeia, como mostrado na Figura 2.3.

O DNA forma uma dupla fita, sendo uma fita complementar à outra. A complementaridade se dá por meio das bases complementares, A-T e C-G (Figura 2.4).

O DNA tem função de armazenar as informações necessárias para formar RNAs e proteínas. Um genoma de um organismo é formado por seus cromossomos e seu material genético. O genoma é transmitido, com variações individuais, de geração em geração, e determina a espécie do ser vivo [81]. Na molécula de DNA, existem regiões que contêm informações específicas para formar proteínas, chamadas de genes (Figura 2.5).

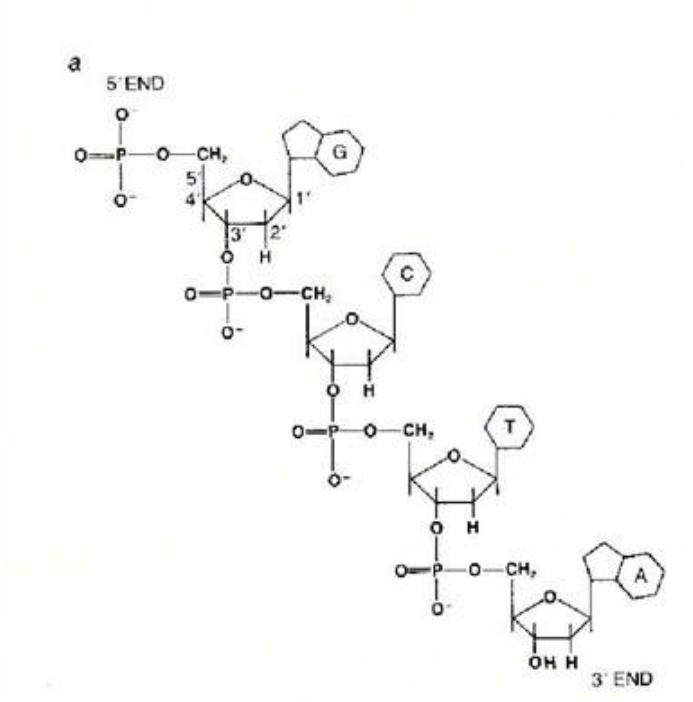

Figura 2.3: Cadeia de nucleotídeos formada pela ligação dos grupos fosfatos [10].

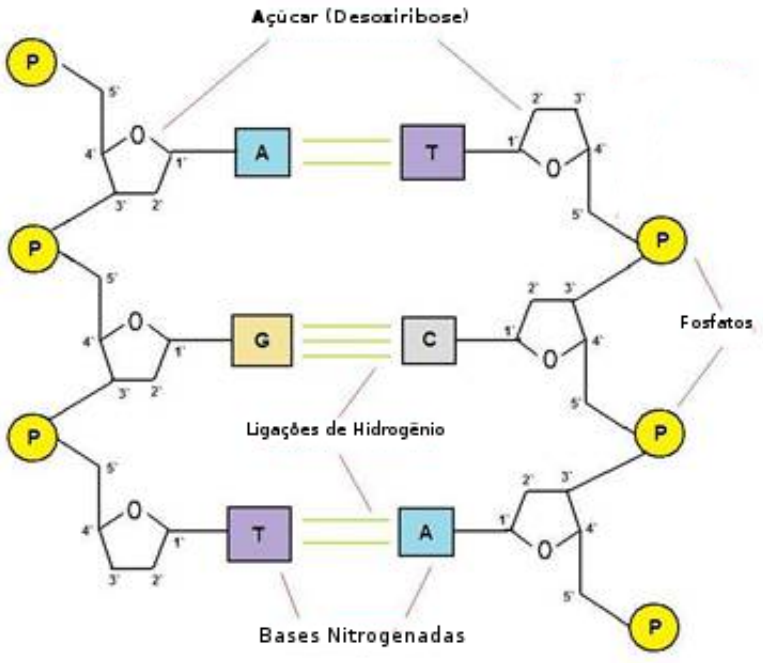

Figura 2.4: Dupla fita de DNA, ligada por bases complementares [10].

Figura 2.5: Fita de DNA, onde as regiões mais escuras indicam os genes [28].

2.1.2 RNA

O RNA é composto por nucleotídeos formados por moléculas de fosfato, ribose e uma base nitrogenda (Figura 2.6).

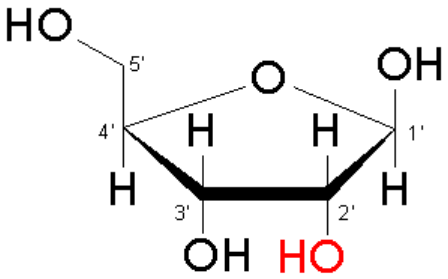

Figura 2.6: Molécula de ribose, com cinco átomos (pentose) de carbono (1' a 5'). Observe que o carbono 2' é ligado a uma molécula OH [2].

Assim como no DNA, existem quatro tipos de bases nitrogenadas que formam o RNA: A, G, C e Uracila-U (no lugar da Timina) [81], e os seus nucleotídeos também se relacionam por meio de ligações dos grupos fosfatos (Figura 2.7).

A uracila também se liga à adenina, mas há uma mudança na estrutura do RNA, pois o RNA nem sempre forma uma dupla fita, e às vezes vemos hélices híbridas de RNA-DNA. Bases de uma molécula de RNA podem ligar-se a outras bases da mesma molécula por complementaridade [82] (Figura 2.8).

Além disso, diferentemente do DNA, encontramos vários tipos de moléculas de RNA, cada qual executando uma função diferente [62].

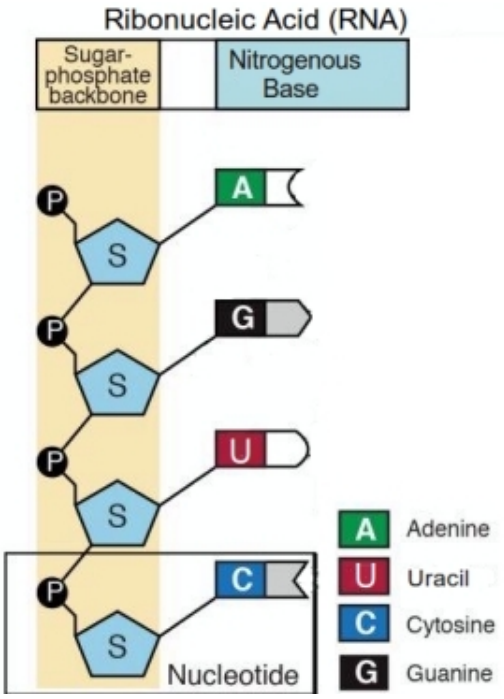

Figura 2.7: Os quatro tipos de nucleotídeos presentes no RNA [22].

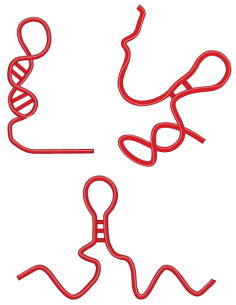

Figura 2.8: Fitas de RNA podem apresentar bases ligadas a outras bases complementares, na mesma molécula [9].

2.1.3 Dogma Central da Biologia Molecular

O Dogma Central da Biologia Molecular é a teoria que relaciona DNA, RNA e proteína: processo de replicação, no qual uma fita de DNA é duplicada; processo de transcrição, no qual parte do DNA é transcrito em RNA; e o processo de tradução, no qual o RNA, proveniente do processo de transcrição, é interpretado em uma proteína (Figura 2.9).

Na replicação, a hélice de DNA é separada em duas fitas pela enzima helicase, a qual se liga à cadeia de DNA e quebra as ligações de hidrogênio entre as fitas. Enquanto a

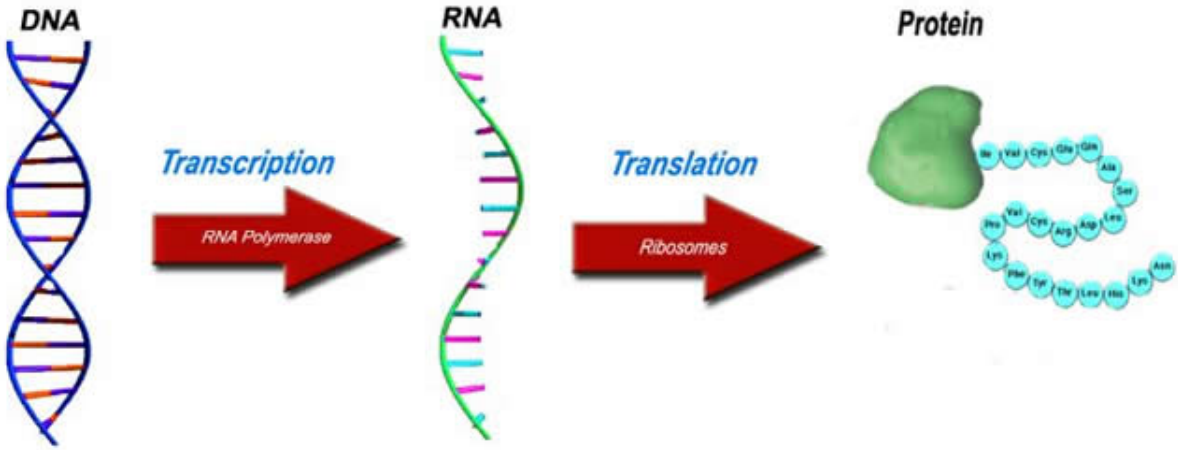

Figura 2.9: Dogma Central da Biologia Molecular, que explica o processo de síntese de proteínas a partir das informações armazenadas no DNA por diferentes tipos de RNA [19].

helicase vai abrindo a dupla fita, outra enzima chamada DNA polimerase encarrega-se de ligar os nucleotídeos das fitas quebradas com novas fitas complementares.

A transcrição também é iniciada pela separação da dupla fita de DNA pela helicase. Quando a fita é separada, a enzima RNA polimerase identifica a fita molde ($5' \rightarrow 3'$) na região de um gene (explicado antes). A RNA polimerase reconhece essa região, que é normalmente precedida por uma sequência de TA (chamada de *TATA box*) [42]. Ao identificar essa região promotora, a RNA polimerase conduz o processo de transcrição do DNA em um RNA mensageiro não maduro (pré-mRNA) nos eucariotos e em RNA mensageiro (mRNA) nos procariotos, sendo que, neste processo de conversão de DNA para RNA, a transcrição ocorre no sentido $5' \rightarrow 3'$ e converte as bases da fita molde para suas bases complementares no RNA gerado. Na Figura 2.10 podemos ver a diferença dos genes nos eucariotos e nos procariotos.

No caso dos organismos eucariotos, o pré-mRNA gerado pela transcrição passa por um processo conhecido como *splicing* (Figura 2.11). Este processo ocorre com o intuito de remover algumas regiões (íntrons) do pré-mRNA e ligar outras (éxons), formando assim o mRNA maduro.

Após o processo de transcrição e do *splicing*, inicia-se a tradução, onde o mRNA é sintetizado em uma proteína. A cadeia de aminoácidos de uma proteína é formada nos ribossomos, que são compostos por RNAs ribossomais (rRNA), por meio de RNA transportadores (tRNAs). Cada tRNA liga triplas de nucleotídeos (códons) em uma ponta com o aminoácido correspondente em outra (Figura 2.12).

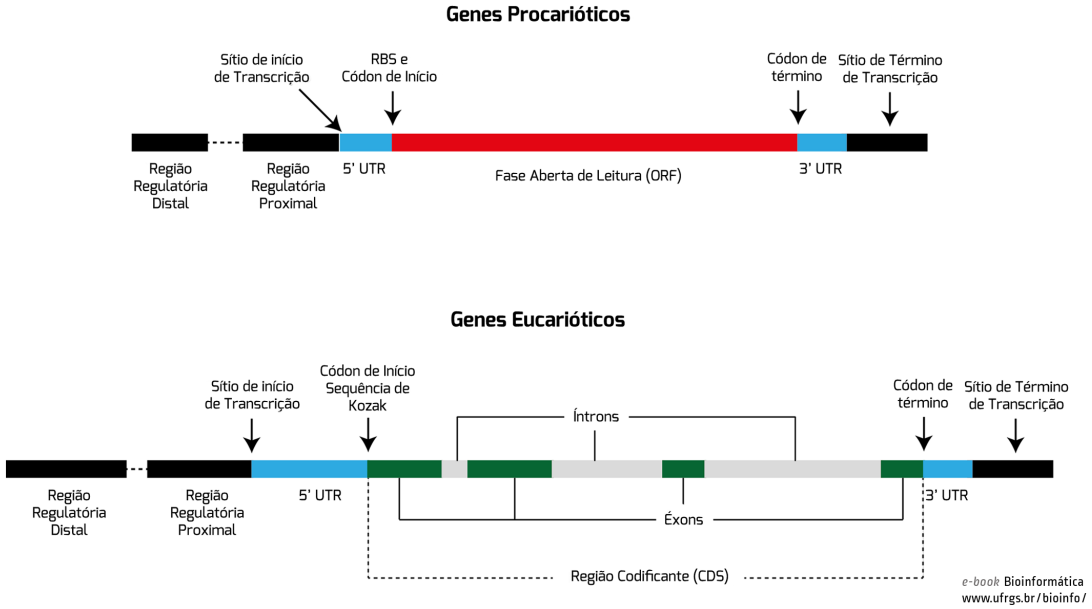

Figura 2.10: Estruturas de um gene procarioto e de um gene eucarioto [11].

A Figura 2.13 mostra a correspondência de cada tripla de bases com o aminoácido correspondente, enquanto a Tabela 2.1 mostra os 20 aminoácidos mais comumente encontrados na natureza.

Dado o código genético, as possíveis sequências de nucleotídeos capazes de serem traduzidas em proteínas, a partir de um códon de iniciação (*start codon* - Metionina - AUG) de tal modo que acabe em um *stop* códon, são chamadas de ORFs (*Open Reading Frames* [78].

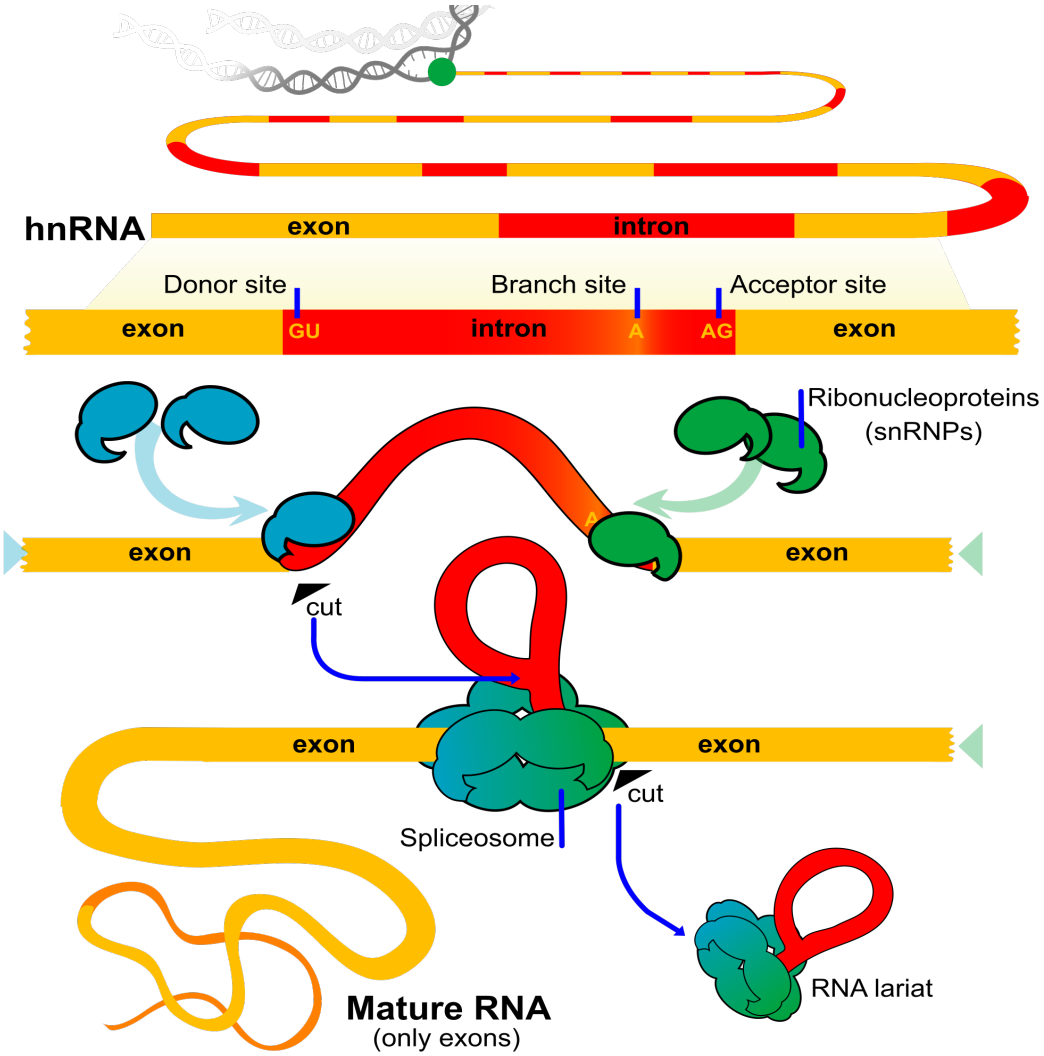

Figura 2.11: Processo de *splicing* em eucarioto [25]

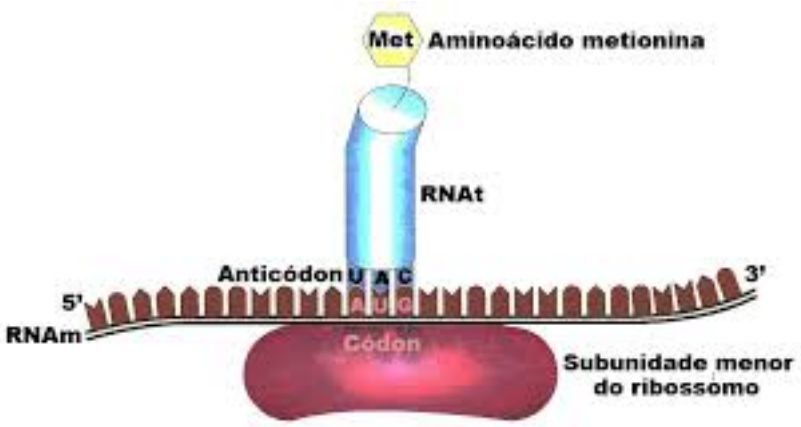

Figura 2.12: Processo de tradução [24].

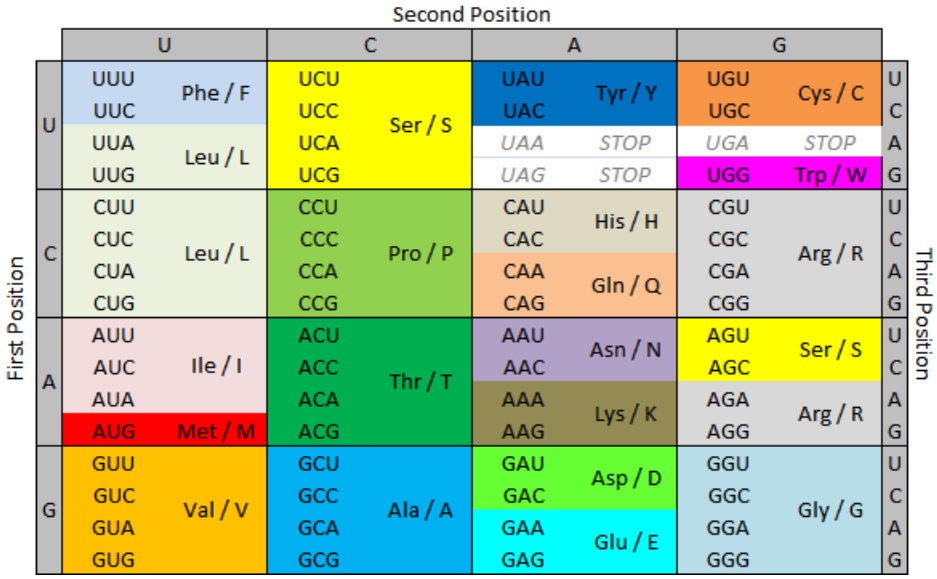

Figura 2.13: Triplas de RNA formam aminoácidos, sendo esta tabela conhecida como código genético [29].

Tabela 2.1: Os vinte aminoácidos mais comumente encontrados na natureza e suas estruturas químicas correspondentes [78].

Abreviatura	Nome
Ala	Alanina
Cys	Cisteína
Asp	Aspartato
Glu	Glutamato
Phe	Fenilalanina
Gly	Glicina
His	Histidina
Ile	Isoleucina
Lys	Lisina
Leu	Leucina
Met	Metionina
Asn	Asparagina
Pro	Prolina
Gin	Glutamina
Arg	Arginina
Ser	Serina
Thr	Treonina
Val	Valina
Trp	Triptofano
Tyr	Tirosina

2.2 Sequenciamento e Bioinformática

Sequenciamento é o processo de obter a sequência de nucleotídeos que formam uma porção de DNA ou de RNA. As novas tecnologias, conhecidas como sequenciamento de nova geração, estão evoluindo rapidamente. Essas tecnologias realizam o sequenciamento de DNA em plataformas capazes de gerar milhões de bases, em pouco tempo. Atualmente, o *Illumina* [35], que se baseia em um sequenciamento por síntese, é um dos sequenciadores mais utilizados.

O processo de sequenciamento do *Illumina* começa quando ele recebe o DNA a ser sequenciado. Primeiramente, o DNA recebido é fragmentado e ligado a adaptadores em suas extremidades 5' e 3'. Então essas moléculas de DNA são ligadas a um suporte sólido, onde estão presentes oligonucleotídeos complementares aos adaptadores, nas extremidades das moléculas.

Quando ligados nos suportes, uma etapa para amplificação do DNA acontece, usando a técnica de *Polymerase Chain Reaction* (PCR). O PCR usa uma enzima conhecida como Taq DNA polimerase para replicar fitas de DNA, sendo as moléculas aderidas ao suporte amplificadas. Esse processo de amplificação das fitas de DNA se repete, até que grupos de várias moléculas idênticas sejam formadas no suporte.

Agora, com moléculas de DNA suficientes e a incorporação de terminadores marcados[1], ocorre uma excitação a laser para que se gere um sinal luminoso, que difere de terminador para terminador. Esse sinal é captado por um dispositivo de leitura e interpretado como um dos quatro possíveis nucleotídeos componentes da molécula.

O processo de incorporação de terminadores, excitação e leitura é repetido para cada nucleotídeo que compõe a sequência até obter o sequenciamento final [39]. A Figura 2.14 mostra o processo de sequenciamento do *Illumina*.

Encontramos também tecnologias especializadas no sequenciamento de RNA, como é o caso do *RNA-seq*. Estudos utilizando este método vem contribuindo para entender e aprofundar estudos em transcritomas de eucariotos. O *RNA-seq* também fornece uma medida muito mais precisa de níveis de transcrição de RNAs e as suas isoformas[2], quando comparada a outros métodos [93].

Na decáda de 1990, foram desenvolvidos vários projetos genoma, dentre eles, o genoma humano [92, 64], que criaram uma nova área de conhecimento, a Bioinformática. A Bioinformática tem por objetivo criar e aplicar técnicas computacionais e matemáticas para analisar informações geradas pelos projetos de sequenciamento [34].

[1] Uma sequência de nucleotídeos pré-determinada.

[2] Formas distintas de uma proteína que são produzidas a partir de genes diferentes ou do processo de *splicing*.

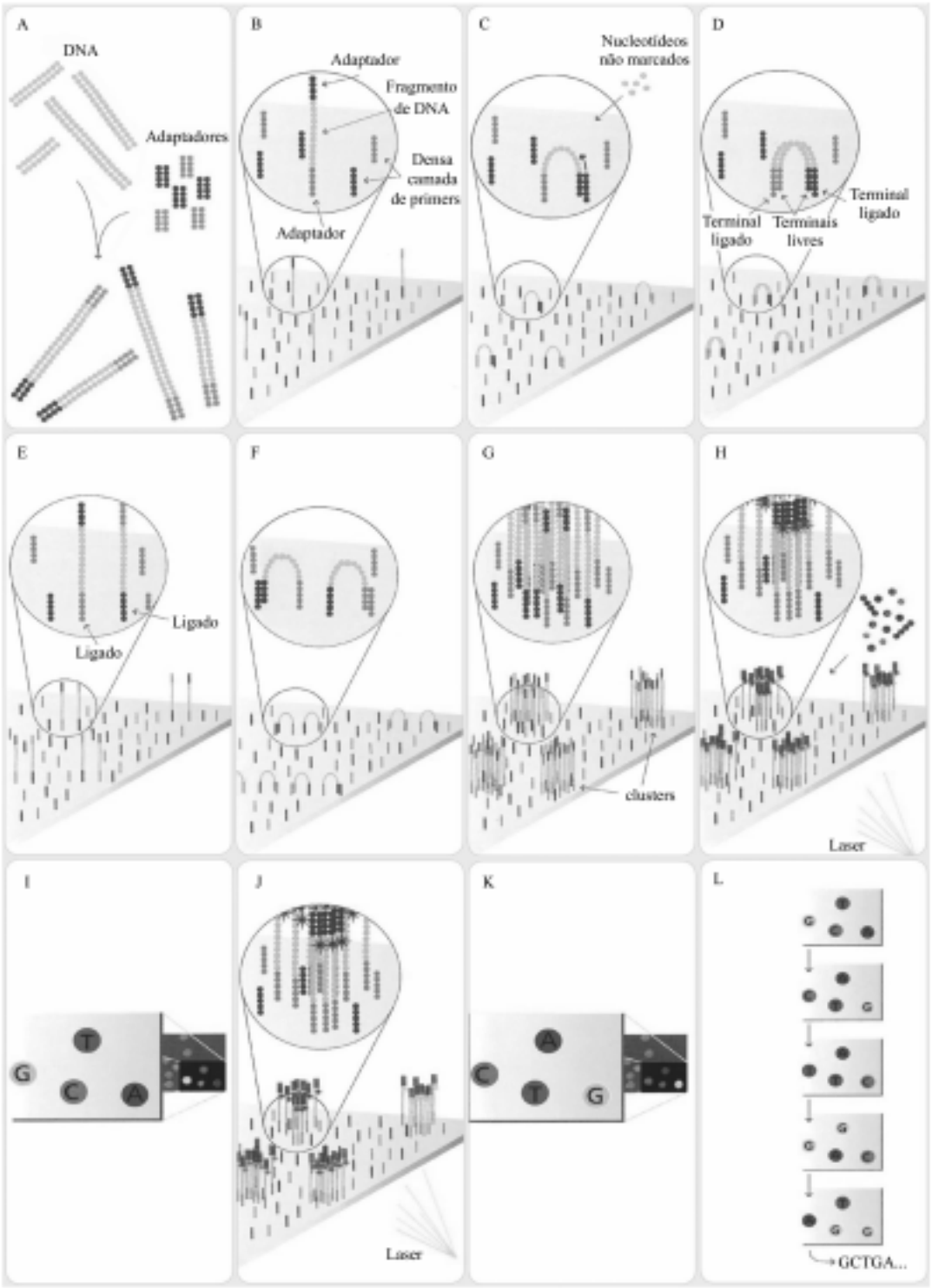

Figura 2.14: Processo de sequenciamento utilizado pelo sequenciador *Illumina* [39].

Para analisar as sequências de DNA e RNA, são criados os *pipelines*. Um *pipeline* é definido por uma sequência de métodos computacionais utilizados no tratamento de dados gerados em um projeto genoma ou transcritoma. Um exemplo desse *pipeline* é mostrado

na Figura 2.15.

Figura 2.15: Exemplo de *pipeline*.

Conforme apresentado, nos sequenciadores, sequências de DNA/RNA são transformadas em cadeias de caracteres, sendo essas cadeias formadas a partir do alfabeto $\sum = \{A, C, G, T/U\}$. Essas sequências são armazenadas em arquivos com formato conhecido, como o *fasta* e o *fastq* por exemplo. O *fasta* é um dos formatos mais utilizados na área de Bioinformática, tendo na primeira linha um identificador de sequência após o caractere '>', e nas próximas linhas a sequência do genoma/transcrito (Figura 2.16).

```
> PBDMB-M1-001t_A01

TAGTCCCGGGCTGAGGAATTCGGCACGAGGCCTAGATGAGAGCTTGTCTC
GTGAGTATGACCCTCACAGACGGCACAGACCTGAGCCAAGCTGTCTTGGA
AAATAAGAGGAGAGATAACGAGAACACCTGGGTTCAGGAGTGGACTTGGG
AACGGATTGAGGAGCAGAGATTGAAGGGTCTAGATGTTGTCAAGGCGTTT
ATTGGACTTGATGCGAAGCTTCTCCAGGNAGCAGAGTTGTAGGGCTTCAC
AGACGTCATGAGTTATGCTGGTTTCTTTTTGGGATGTAGGGGTTTTCTTC
TCTCATGAGGTTTGATGATTCTTCTGTGCCTACAGGATTGGTGTTGGGCT
TTCTATTATTATTTCTTAGCTTGAGTGTTTGTGTTTGTCATTATCATTCA
TCTTCAATACCCCTTCTTGTTTACCCCATCAAACTATTTCACGTAAGAGT
CCTTAATTCCTCTTTTTCTAGATTTTTATATCTCATATAGATGTNTCCAG
TTACTTGTAAAAACAAAAAAAAAAAAAACTTGGGGGGGGGGCCGGGTACC
AATTTCGCCTTTTTGGTTCGTTTCTAACGGGCGAGGATGGAGAGAGAGAG
AAGAGAGGAGGGAGAGCGAGGACGAAGAGAAGAGAGAGGGAACGGCAGGG
GAGAAGCAAGGATGAGTGACGGAGCAAGAGCAAGAAGGGAGCGAACAAGA
AAAGGAGAAGAGAAAACGAAGGTAGAGAAAACAACGAAAAGCAACAGGAA
CGAGCAGAGGAGAGCACGGAGAGAGATGAGCAGGACGGCAAAAGAACCGA
CACAAG
```

Figura 2.16: Exemplo de arquivo *fasta*.

Um outro formato de arquivo é o *fastq*, que além de possuir a sequência do genoma/transcrito, seu identificador e sua descrição, também possui informações das qualidades de cada nucleotídeo em código ASCII, como mostrado na Figura 2.17.

Como alguns erros podem ocorrer no sequenciamento, é necessário que os arquivos gerados passem por um processo de filtragem para garantir a qualidade das sequências que serão utilizadas nas outras etapas do *pipeline*. A filtragem é a primeira etapa do *pipeline*, que usa programas como o *prinseq* [75], que permite filtrar as sequências de acordo com a qualidade desejada. Por fim, programas como o FastQC [8] recebem arquivos *fastq* como entrada, e produzem visualização das qualidades das sequências (Figura 2.18),

Após a filtragem, é necessário agrupar as sequências e obter fragmentos mais próximos das sequências genômicas ou dos transcritos, o que é realizado na etapa de montagem. Existem dois tipos de montagem: com genoma de referência e *de novo*. Na primeira, utiliza-se o genoma de um organismo evolutivamente próximo para guiar o agrupamento.

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

@title and optional description
sequence line(s)
+optional repeat of title line
quality line(s)

Figura 2.17: Exemplo de arquivo *fastq*.

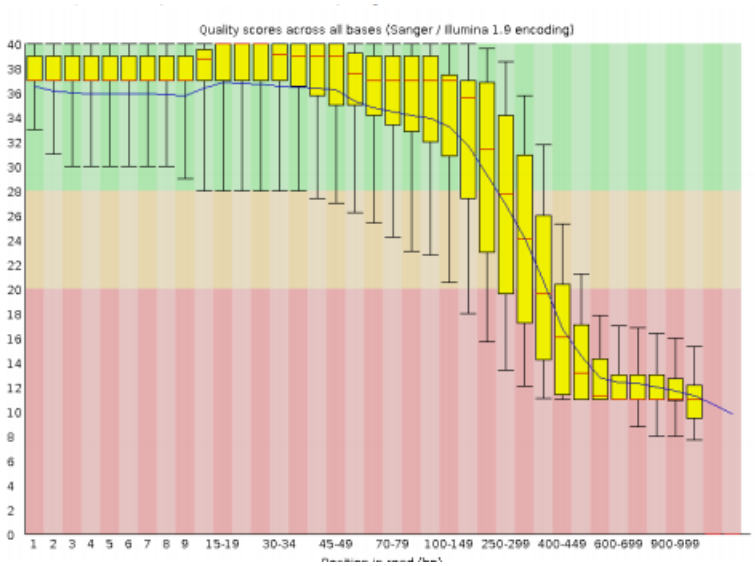

Figura 2.18: Gráfico que mostra a qualidade das sequências geradas usando o programa FastQC.

O fato de usar um genoma de referência torna a montagem mais rápida e precisa, porém nem sempre temos um genoma apropriado, o que pode dificultar na descoberta de sequências específicas do organismo sendo mapeado. A Figura 2.19 mostra um exemplo de montagem com genoma de referência.

Já na montagem *de novo* os grupos são obtidos a partir da sobreposição das sequências resultantes do sequenciamento. Somente grupos que possuem um número suficiente de sequências (boa cobertura) garantem que o grupo é confiável. Como não possui nenhuma sequencia de referência, a montagem pode ser mais demorada. A Figura 2.20 mostra um exemplo de montagem *de novo*.

Após a etapa de montagem, podemos começar a última etapa do *pipeline*, chamada anotação, cujo principal objetivo é atribuir funções biológicas às sequências analisadas.

Figura 2.19: Exemplo de montagem com referência, sendo a *Reference sequence* o genoma de um organismo evolutivamente próximo e $s1, s2, s3, s4, s5$ e $s6$ sequências do organismo sequenciado [20].

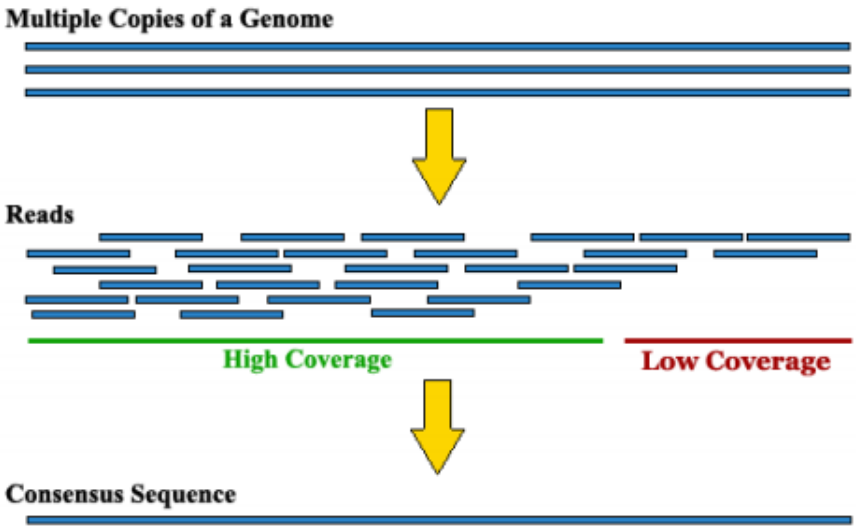

Figura 2.20: Exemplo de montagem *de novo*, com exemplos de áreas com *high coverage* (boa cobertura) e áreas com *low coverage* (cobertura ruim) de acordo com a quantidade de sequências presentes para formar o grupo [20].

Os objetivos da fase de anotação mudam de acordo com o objetivo do projeto. Em projetos transcritoma, a anotação busca descrever genes expressos ou porções deles e suas isoformas, as proteínas expressas pelos transcritos, e possivelmente suas funções no metabolismo celular do organismo. Já em projetos genoma, pode-se tentar identificar regiões não-codificadoras e descobrir genes, buscando descrever suas características e funções no metabolismo. Para tais tarefas, bancos de dados contendo informações de moléculas com funções biológicas já conhecidas e ferramentas para análise de similaridades são utilizados.

Uma das ferramentas mais utilizadas no processo de anotação é o *Basic Local Alignment Search Tool* (BLAST) [31], que encontra regiões similares em sequências, computando alinhamentos locais. O BLAST pode ser aplicado para várias tarefas, incluindo

descoberta de função por busca de sequências de DNA e de proteínas em bancos de dados, identificação de genes e análise de similaridade em sequências longas de DNA. Além de tantas funcionalidades, o BLAST, quando comparada com outras ferramentas, apresenta um menor tempo de execução [31].

Há diversas variações dessa ferramenta:

- o blastn, que recebe nucleotídeos de entrada e busca similaridade com nucleotídeos contidos em um banco de dados;
- o blastp, que recebe aminoácidos de entrada e busca similaridade com aminoácidos contidos em um banco de dados;
- o blastx, que recebe nucleotídeos traduzidos em aminoácidos de entrada e busca similaridade com aminoácidos contidos em um banco de dados;
- o tblastn, que recebe aminoácidos de entrada e busca similaridade com nucleotídeos traduzidos em aminoácidos contidos em um banco de dados;
- o tblastx, que recebe nucleotídeos traduzidos em aminoácidos de entrada e busca similaridade com nucleotídeos traduzidos em aminoácidos contidos em um banco de dados.

Na Figura 2.21 podemos ver como funciona o processo de anotação.

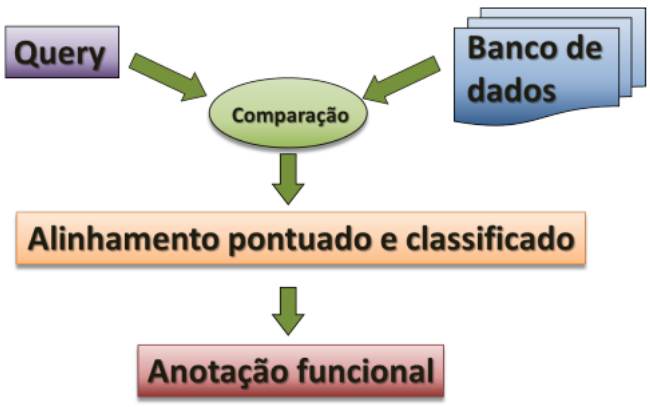

Figura 2.21: Visão geral do processo de anotação. *Query* é a sequência de entrada, e o *alinhamento pontuado e classificado* é realizado pelo BLAST. Sequências similares indicam conservação de funções, assim se uma sequência similar à *Query* for encontrada no banco de dados, essa *Query* pode ser anotada funcionalmente.

2.3 Classificação de RNAs

Os mRNAs são RNAs codificadores de proteínas, já os tRNAs e os rRNAs, embora envolvidos no processo de síntese de proteínas, não codificam proteínas. Acredita-se que menos de 2% dos RNAs encontrados no genoma dos mamíferos são traduzidos em proteínas [81]. Uma porção significativa do restante de RNAs pode ser classificada em diversas famílias de ncRNAs.

Os ncRNAs são transcritos de genes que exercem papéis celulares importantes [82]. Os ncRNAs agem diretamente na célula em funções estruturais, catalíticas ou regulatórias [44, 95] e têm papel fundamental no controle da expressão de genes em proteínas [55]. As linhas de pesquisa atuais apontam relações extensas entre ncRNAs e diversos processos de um organismo, mas muito pouco ainda é conhecido sobre essas moléculas, principalmente pela grande difculdade em verificar experimentalmente qual é exatamente a funcionalidade do determinado gene não codificador no organismo [62].

Podemos encontrar na literatura [44] uma classificação genérica de ncRNAs em: pequenos (*small ncRNAs*), que possuem características conhecidas e tamanhos pequenos (20 a 300 nucleotídeos); e longos (*long ncRNAs* - lncRNAs), que apresentam tamanhos maiores do que 200 nucleotídeos e pouca capacidade de síntese de proteínas, sendo os transcritos menos conhecidos atualmente [68].

2.3.1 NcRNAs pequenos

Diversos ncRNAs são considerados pequenos devido ao pequeno número, de 20 a 30 bases, que formam sua molécula. Embora estes RNAs não sejam codificados em proteínas, eles possuem os mais diversos papéis, sendo esses papeis intimamente ligados à sua estrutura secundária. E diversos desses RNAs já possuem estruturas secundárias bem conhecidas.

Um dos ncRNAs pequenos mais conhecidos, o tRNA, é responsável pelo transporte de aminoácidos que serão usados na síntese de proteínas. Sua estrutura secundária se assemelha a um 'T', tendo em uma de suas extremidades um aminoácido e na outra o anticodon equivalente a este aminoácido (Figura 2.22).

Outro ncRNA bastante conhecido é o rRNA. O rRNA compõe os ribossomos e é responsável pela catálise de síntese proteíca [63] (Figura 2.23).

Ainda temos diversos outros ncRNAs pequenos conhecidos, como os snoRNAs (Figura 2.24) e os miRNAs (Figura 2.25), que podem modificar outros ncRNAs (como rRNAs) e agir como reguladores no processo de tradução, respectivamente.

Na Tabela 2.2 podemos ver diversos ncRNAs pequenos e seus papéis nos mecanismos celulares.

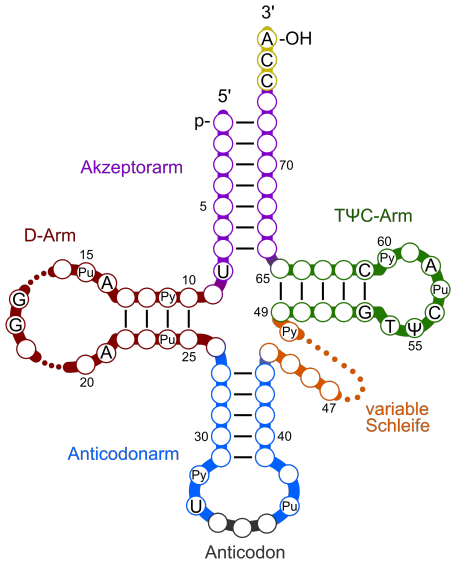

Figura 2.22: Estrutura secundária do tRNA [27].

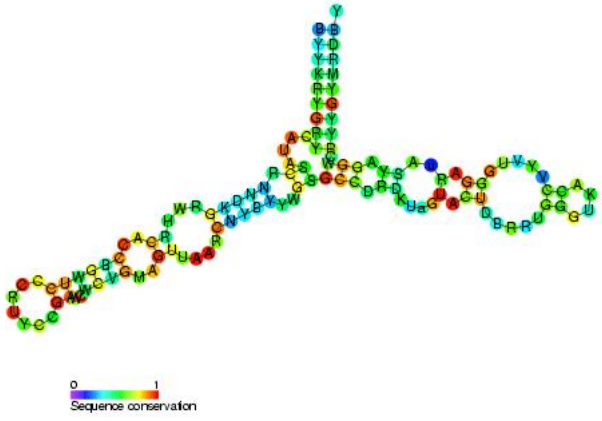

Figura 2.23: Estrutura secundária do rRNA [23].

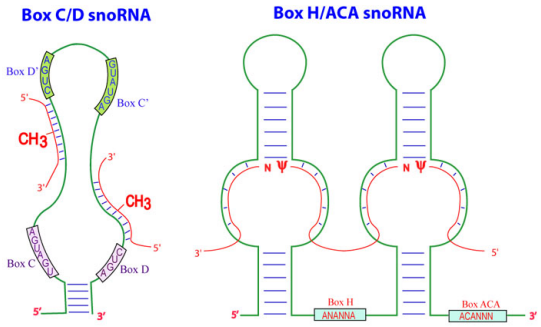

Figura 2.24: Estrutura secundária do snoRNA [15].

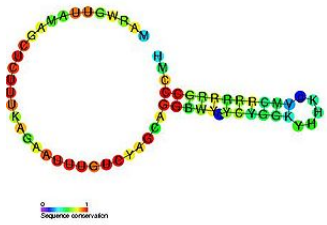

Figura 2.25: Estrutura secundária do miRNA [23].

Tabela 2.2: Alguns tipos de RNAs não-codificadores pequenos

Sigla	Nome	Função
rRNA	RNA ribossomal	Envolvidos com a síntese de proteínas
tRNA	RNA transportador	Realizam transporte de aminoácidos
snoRNA	*Small nucleolar* RNA	Modificação de partes do rRNAs, tRNAs e snRNAs
snRNA	*Small nuclear* RNA	Envolvidos na excisão dos introns no processo de *splicing*
siRNA	*Small interfiring* RNA	Interferência na tradução de proteínas separando e promovendo a degradação de trechos de mRNAs
snmRNA	*Small non-messenger* RNA	Pequenos ncRNAs com função regulatória
rasiRNA	*Repeat-associated* RNA	Silenciamento da transcrição de genes via remodelagem da cromatina
miRNA	*Micro* RNA	Regulam a tradução
piRNA	*Piwi-interacting* RNA	Regulação de tradução e estabilidade de mRNA, dentre outras funções
stRNA	*Small temporal* RNA	Interrupção da tradução de mRNA

2.3.2 NcRNAs longos

LncRNAs são moléculas de ncRNAs que podem chegar ao tamanho de centenas a milhares de bases, com um mínimo 200 bases. Atualmente, ainda não se sabe muito a respeito dos papéis exercidos pelos lncRNAs [68], mas sabemos que muitos transcritos são associados a lncRNAs e possuem um baixo poder de síntese de proteínas [67, 68].

Os lncRNAs podem ser classificados em cinco categorias [68]: (a) senso: quando o lncRNA se sobrepõe a um gene na mesma fita; (b) antisenso: quando o lncRNA se sobrepõe a um gene na fita oposta; (c) bidirecional: quando o lncRNA e o gene são expressos juntos e estão em fitas opostas; (d) intrônico: quando o lncRNA está localizado dentro de uma região intrônica; (e) intergênico (*long intergenic ncRNA* - lincRNA): quando o lncRNA situa-se entre dois genes. A Figura 2.26 mostra essas classificações.

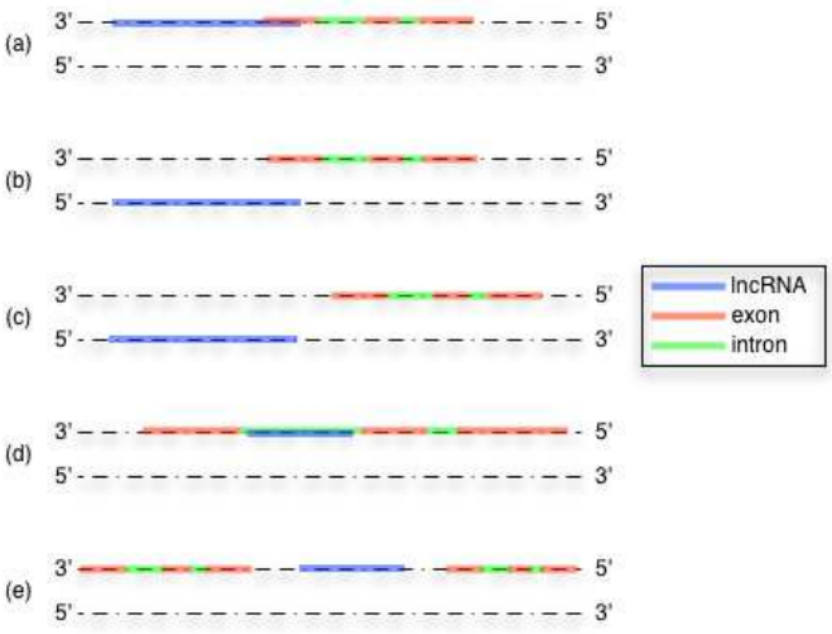

Figura 2.26: Cinco categorias de lncRNA: (a) senso; (b) antisenso; (c) bidirecional; (d) intrônico; e (e) intergênico [76].

A identificação de lncRNAs é dificultada pelo fato deles serem confundidos com genes codificadores de proteínas [65]. Este trabalho tem como foco os lincRNAs, que serão detalhados na Seção 2.4.

2.4 LincRNAs

Conforme apresentado, os lincRNAs são transcritos localizados na região intergênica. Embora pouco se saiba a respeito dos papéis biológicos dos lincRNAs, acredita-se que eles possuem papéis regulatórios, participam em fatores importantes ligados ao câncer [49, 51]. Diversas funções foram propostas aos lincRNAs, como mostrado na Figura 2.27.

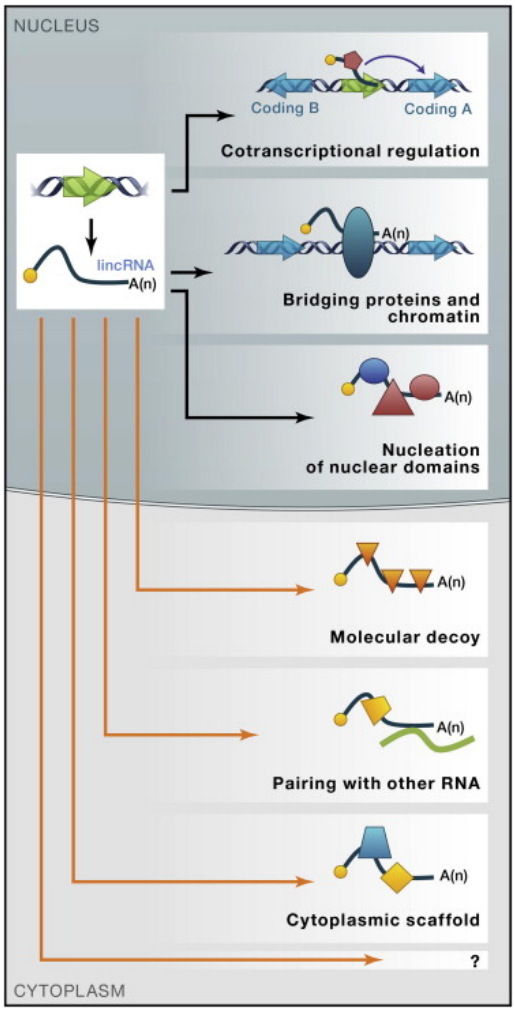

Figura 2.27: Modelo de funções propostas a lincRNAs [90].

A classificação de lincRNAs diferencia-se bastante da classificação de outros ncRNAs, pois não possuem uma estrutura secundária bem definida. Os lincRNAs são atualmente o tipo de lncRNA mais estudado, pois eles não se sobrepõem a nenhum gene, o que facilita a identificação de suas características próprias [90, 74, 79].

Diversos lincRNAs já foram descobertos por apresentar papéis importantes em diversos organismos, dentre esses podemos citar: o *H19*, que pode limitar o crescimento da placenta em mamíferos [57]; o Cyrano e o *megamind*, que são necessários para um

bom desenvolvimento embrionário [91]; o HotairM1, que regula o ciclo de desenvolvimento na maturação da medula óssea [96]; e o Tug1, que pode agir como um supressor de tumores [60];

2.4.1 Métodos de identificação biológicos e computacionais

LincRNAs foram catalogados para humanos, camundongos, peixe-zebra, sapos e outras espécies [90]. Essa catalogação se dá pelo uso de modelos que consideram diversas características como: posição de início no genoma, posições de *splicing* e posição da cauda poli-A de cada transcrito. Na Figura 2.28, podemos ver diversos métodos utilizados para identificar lincRNAs em humanos e camundongos.

Métodos biológicos como clonagem de cDNA e *tilling array*[3], juntamente com uma pós-análise, podem gerar resultados satisfatórios na identificação de lincRNAs. Na Figura 2.29, podemos ver alguns desses métodos.

Alguns experimentos utilizando *microarrays* também evidenciaram lincRNAs em células renais com câncer [45], como mostrado na Figura 2.30.

Existem outros métodos para identificar os lincRNAs, como: combinação de ChIP-seq e análise *in silico* [52]; identificação de estruturas secundárias comuns aos lincRNAs [5]; e ainda métodos que usam características extraídas das sequências consenso obtidas na montagem [90].

Apesar das pesquisas, não há métodos amplamente utilizados para identificação e classificação de lincRNAs. Diversos estudos apontam o uso de aprendizagem de máquina como uma boa alternativa para diferenciar RNAs codificadores de proteínas (*protein coding transcripts* - PCTs) de ncRNAs, como CONC (Coding Or Non-Coding) [61], CPC (Coding Potential Calculator) [58] e PORTRAIT [33]. CONC é demorado para analisar arquivos grandes [58] e o CPC funciona bem com PCTs conhecidos, mas tende a classificar PCTs novos em ncRNAs, caso eles não tenham sido ainda anotados como proteínas nos bancos de dados [58].

Citamos ainda o iSeeRNA [83] como uma ferramenta que pode identificar lincRNAs em humanos e camundongos, sendo utilizado para isso o *pipeline* da Figura 2.31 para construção de seu modelo.

No iSeeRNA, foi disponibilizada uma ferramenta para consulta *online* e uma outra para *download* (Figura 2.32).

[3]Um tipo de *chip* de microarranjo especializado em identificar sequências conhecidas

Reference	Data for Transcript Reconstruction	Genomic Features and Filters	Coding-Potential Filters	Number of lincRNAs
Mouse				
Ravasi et al., 2006	cDNAs		Manual curation, ORF length, CRITICA	13,502 transcripts
Ponjavic et al., 2007	cDNAs, CAGE		Manual curation, ORF length, BLAST, CRITICA	3,122 transcripts
Guttman et al., 2009	Chromatin marks, tiling arrays	Collection of approximate exonic regions, chromatin domain ≥5 kb	CSF	1,675 loci (1,250 conservatively defined)
Guttman et al., 2010	RNA-seq	Multi-exon only	CSF	1,140 lincRNA transcripts
Sigova et al., 2013	RNA-seq, cDNAs, chromatin marks,	Antisense overlap with mRNA introns allowed, ≥100 nt mature length	CPC	1,664 loci
Human				
Khalil et al., 2009	Chromatin marks, tiling arrays	Collection of approximate exonic regions, chromatin domain ≥ 5 kb	CSF	3,289 loci
Jia et al., 2010	cDNAs	Overlap with mRNAs allowed		5,446 transcripts
Ørom et al., 2010	cDNAs	Restricted to loci >1 kb away from known protein-coding genes, ≥200 nt mature length	Manual curation based on length, conservation and other characteristics of the ORFs	3,019 transcripts from 2,286 loci
Cabili et al., 2011	RNA-seq	Multi-exon only, ≥200 nt mature length	PhyloCSF, Pfam	8,195 transcripts (4,662 in the stringent set)
Derrien et al., 2012	cDNAs	Overlap with mRNAs allowed (intergenic transcripts reported separately), ≥200 nt mature length	Manual curation based on length, conservation and other characteristics of the ORFs	14,880 transcripts from 9,277 loci, including 9,518 intergenic transcripts
Sigova et al., 2013	RNA-seq, cDNAs, chromatin marks,	Antisense overlap with mRNA introns allowed, ≥100 nt mature length	CPC	3,548 loci from embryonic stem cells, and 3,986 loci from endodermal cells

Figura 2.28: Métodos utilizados para classificação de lincRNAs em humanos e camundongo [90].

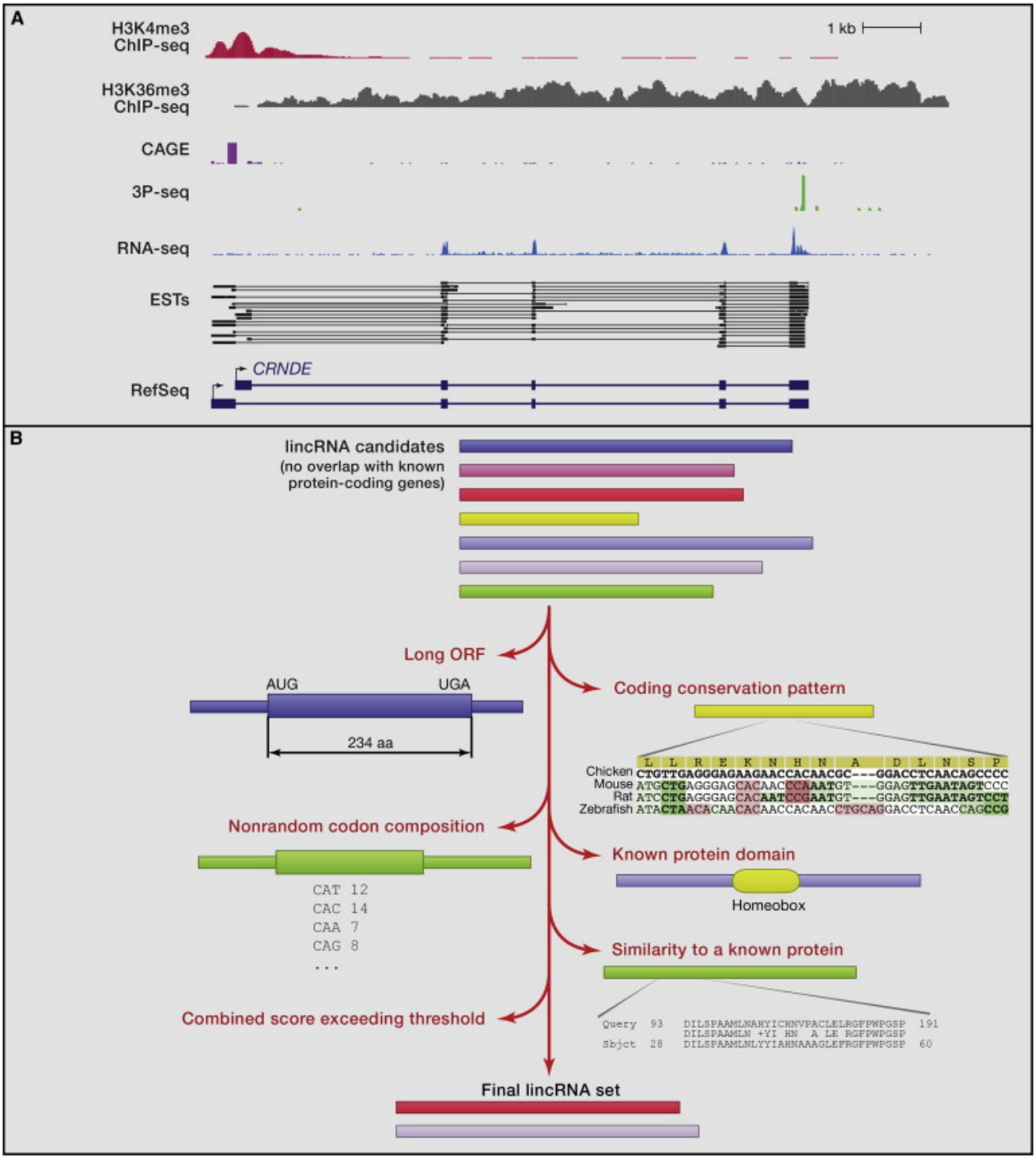

Figura 2.29: Modelo para classificação de lincRNAs [90].

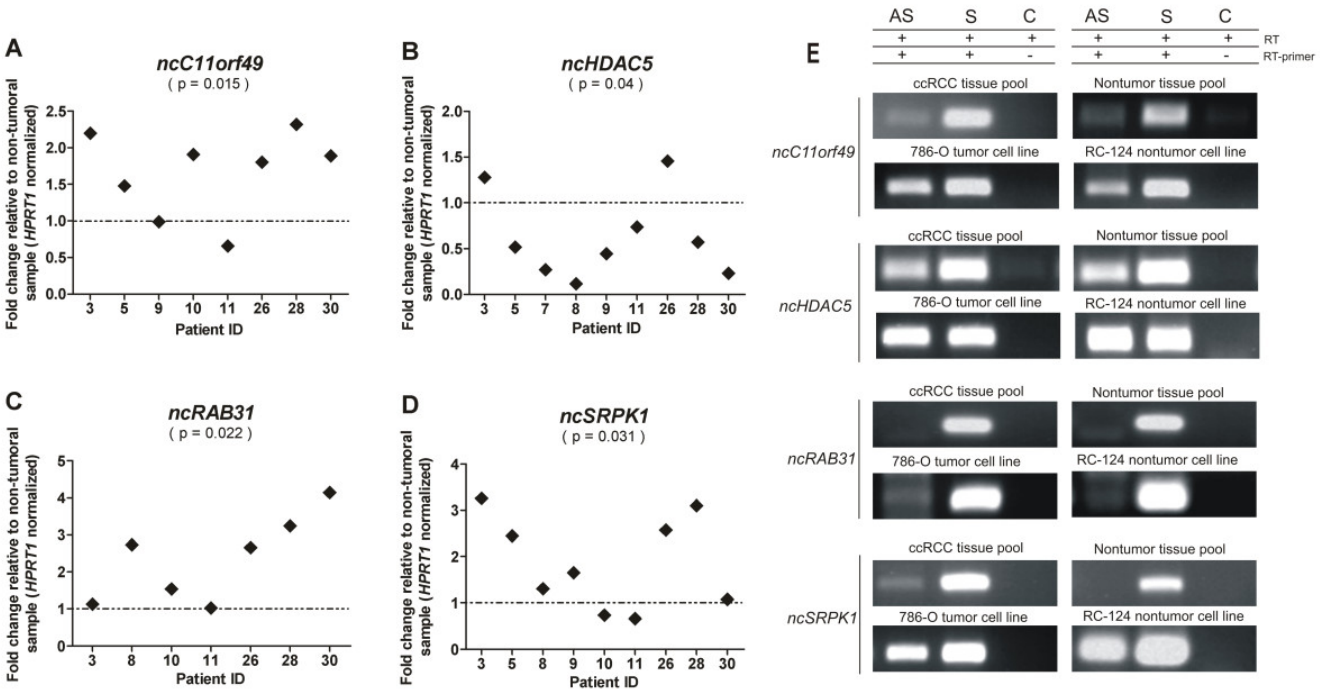

Figura 2.30: Uso de microarranjo para identificar lincRNAs em células renais com câncer [45].

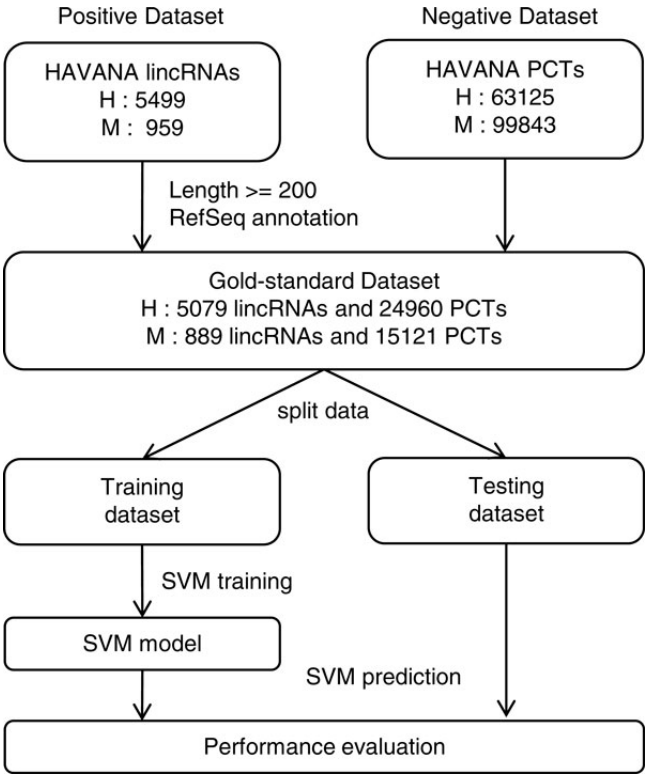

Figura 2.31: *Pipeline* do iSeeRNA [83].

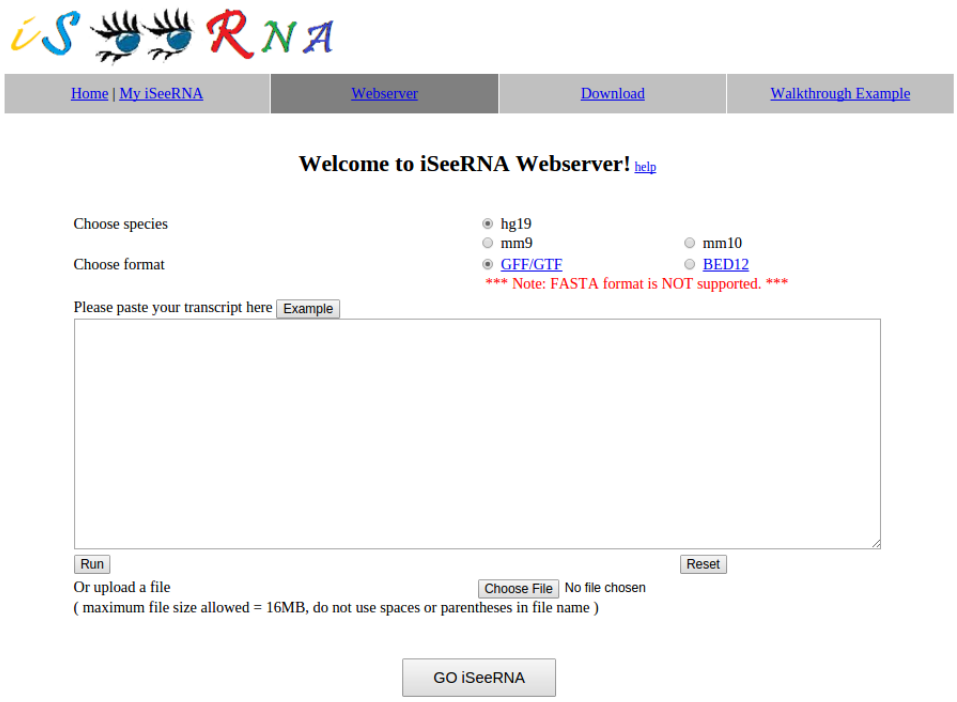

Figura 2.32: Ferramenta *online* do iSeeRNA [14].

2.4.2 Bancos de dados

Na literatura, encontramos diversos bancos de dados com informações de RNAs, sendo os mais relevantes descritos na sequência.

O Rfam [21] é um banco de dados especializado em ncRNAs, principalmente os pequenos. Além de disponibilizar informações da sequência do transcrito, também é possível obter sua descrição e visualizar sua estrutura secundária (Figura 2.33) e visualizar a distribuição entre espécies (Figura 2.34).

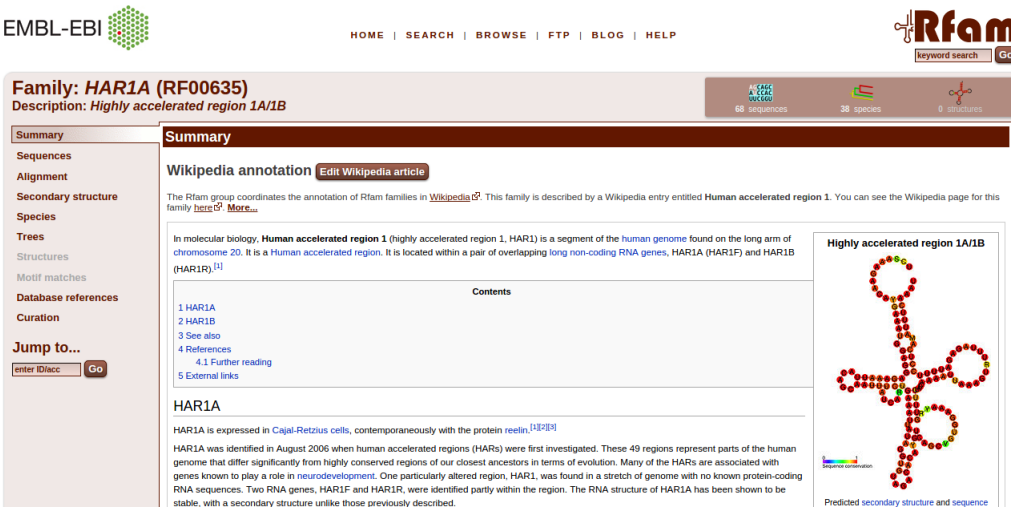

Figura 2.33: Banco Rfam [21].

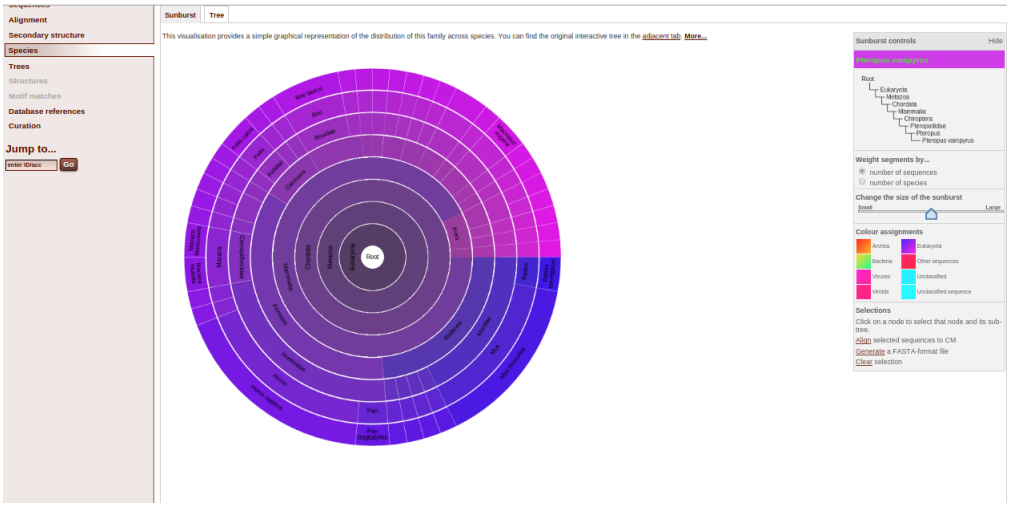

Figura 2.34: Exemplo de distribuição de espécie no Banco Rfam [21].

Outro banco especializado em ncRNAs é o *DIANA Tools* [6], que possui dados de mRNAs e suas relações com lncRNAs.

Podemos encontrar, também, bancos de dados especializados em lncRNAs, como é o caso do lncRNADisease [17], que disponibiliza informações, comprovadas experimentalmente, de lncRNAs que estão envolvidos em doenças, mostrando também o relacionamento desses com outros RNAs, DNAs e proteínas (Figura 2.35).

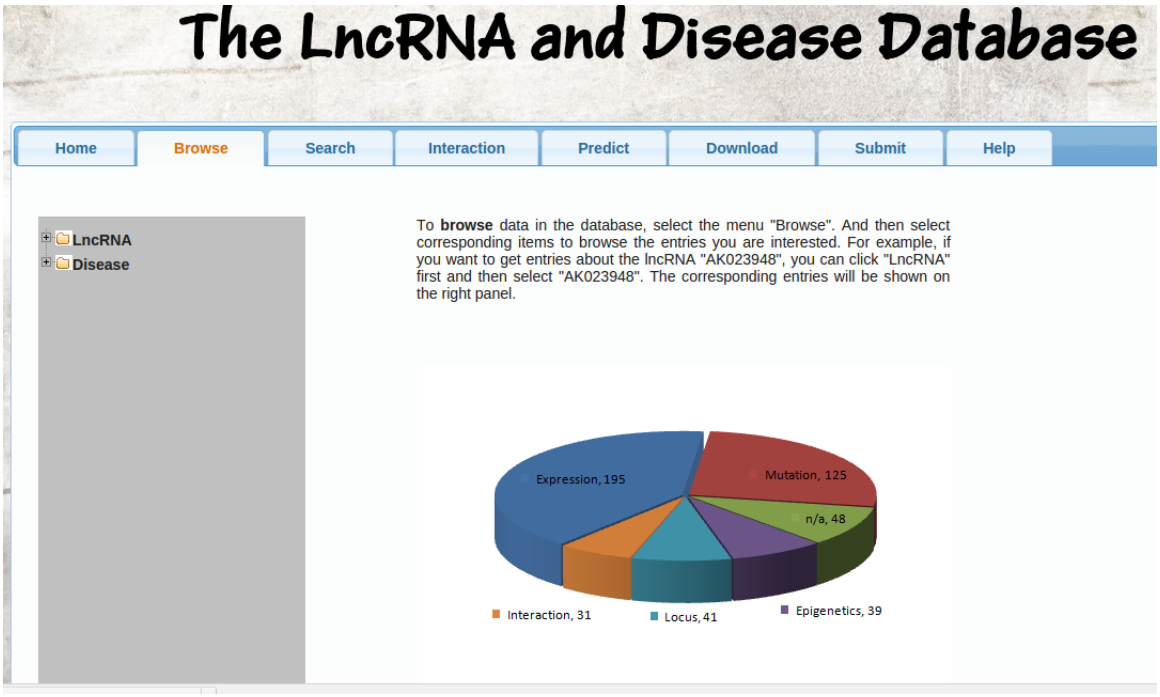

Figura 2.35: Banco lncRNADisease [17].

Bancos de dados como o Ensembl [7] não são tão curados, mas contêm um volume de dados maior. Isso é bom em casos como os lincRNAs, sobre os quais não são conhecidas tantas informações. O Ensembl possui diversas informações de vertebrados e outras espécies de eucariotos.

Por fim, temos o HAVANA [12], que é um banco de dados similar ao Ensembl, mas com dados obtidos de anotação manual, o que os torna mais confiáveis.

A Tabela 2.3 apresenta uma compilação dos bancos de dados detalhados acima.

Tabela 2.3: Bancos de dados

Banco de Dados	Conteúdo
Rfam	NcRNAs, especialmente ncRNAs pequenos
DIANA Tools	MRNAs e interações mRNAs-lncRNAs
lncRNADisease	LncRNAs verificados experimentalmente, que possuem relação com doenças
Ensembl	Dados de vertebrados e outras espécies de eucariotos
HAVANA	Dados anotados manualmente

Capítulo 3

Aprendizagem de Máquina

Neste capítulo, serão descritos conceitos de Aprendizagem de Máquina utilizados neste projeto. Na Seção 3.1, conceitos básicos e técnicas serão apresentados, de forma breve. Na Seção 3.2, o método de Máquina de Vetores de Suporte será descrito com mais detalhes.

3.1 Abordagens

Aprendizagem de Máquina é uma subárea da Inteligência Artificial, que tem como foco o desenvolvimento de programas que detectam padrões e aprendem por experiência [72]. Existem quatro paradigmas de aprendizagem: não-supervisionada, supervisionada, por reforço e semi-supervisionada, descritos em seguida.

3.1.1 Aprendizagem não-supervisionada

O método de aprendizagem não-supersionada busca reconhecer padrões em dados que não foram previamente classificados. Ao reconhecer esses padrões, cada dado de entrada é agrupado em um conjunto específico de dados. Um programa que somente utiliza técnicas de aprendizagem não-supersionada agrupa dados em classes, já que não tem informação de qual ação deve tomar e de qual estado é o desejado. *Clustering* hierárquico [18] e *k-means* [16] são exemplos de algoritmos que usam de aprendizagem não-supersionada e serão descritos a seguir.

O *clustering* hierárquico não encontra m classes para um conjunto de dados de entrada em um único passo. Em vez disso, ele tenta achar o melhor número de classes dentre os agrupamentos $d-1, d-2, ..., 1$, sendo d a quantidade de objetos no conjunto. Neste método, cada objeto começa no seu próprio conjunto de dados e a cada iteração, o algoritmo une os dois conjuntos mais similares. Como nem sempre é fácil achar o número m de classes de um determinado conjunto de dados, essa técnica pode ser bem útil, mas

depende do tamanho do conjunto de dados. Assim, este método pode ser bem custoso. Na Figura 3.1 um exemplo de *clustering* hierárquico é mostrado.

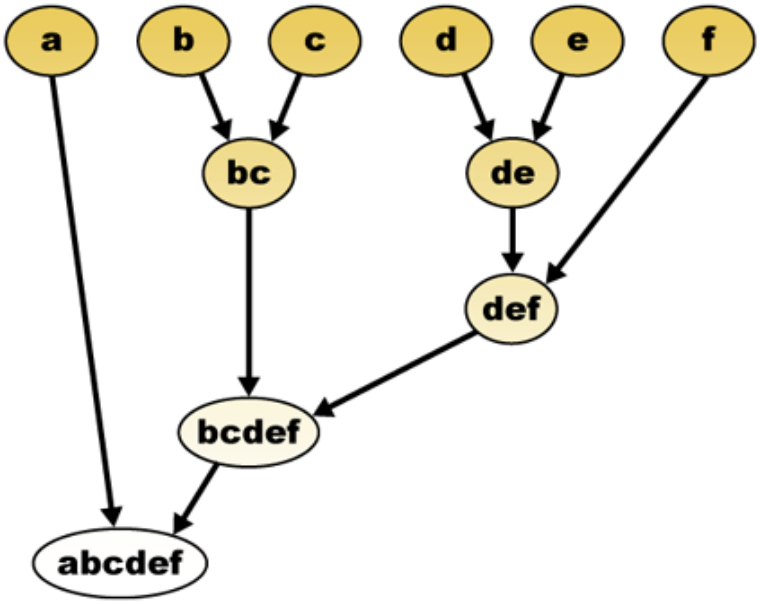

Figura 3.1: Exemplo de *clustering* hierárquico. Note que começamos com o número de agrupamentos igual ao número de objetos informados na entrada e a cada iteração os agrupamentos vão se unindo [18].

O *k-means*, assim como o *clustering* hierárquico, é um algoritmo iterativo, mas em vez de começar com o número de agrupamentos igual ao número de objetos, começamos com um número k de classes pré-definido. Cada uma dessas classes possui um centróide, que é o objeto mais ao centro do agrupamento. A cada iteração, cada objeto é incluído no agrupamento com o centróide mais próximo, sendo isso calculado pela distância entre o objeto e os centróides. No total, são efetuadas o número de iterações necessárias até que os dados não mudem de agrupamentos. A Figura 3.2 mostra um exemplo do *k-means*.

3.1.2 Aprendizagem supervisionada

Diferentemente do aprendizagem não-supersionada, a supervisionada busca identificar características que podem classificar dados como pertencentes a diferentes classes já pré-definidas, utilizando conjuntos de treinamento para construção do modelo e conjuntos de teste para validação.

Basicamente, a aprendizagem supervisionada tenta construir uma função h (hipótese) que classifica objetos do conjunto de teste em uma das classes criadas na fase de treinamento. A performance é calculada de acordo com o número de objetos do conjunto de teste classificados corretamente, levando em consideração os verdadeiros positivos (VP),

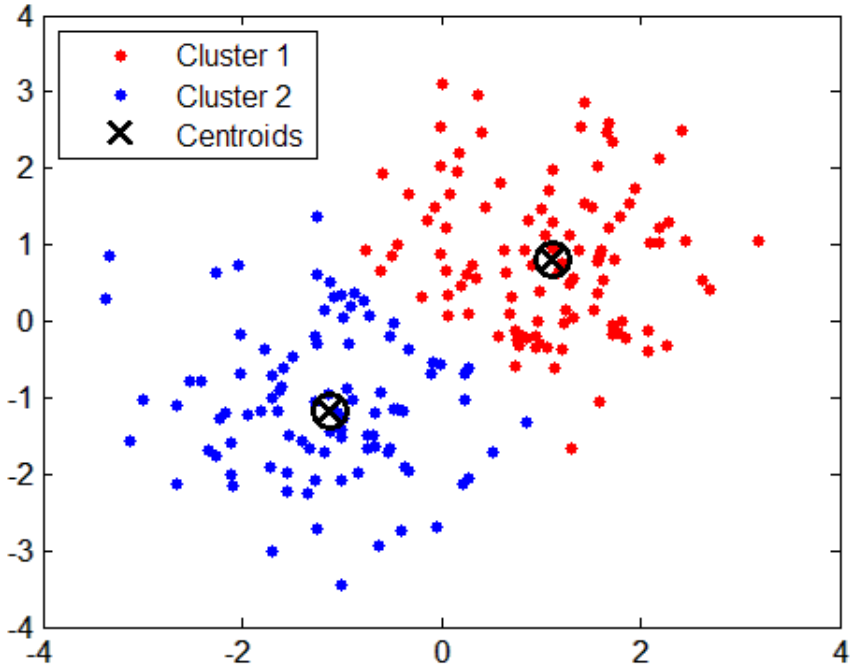

Figura 3.2: Exemplo do *k-means*, com dois grupos, cada um com seu centróide [16].

verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN). A Tabela 3.1 mostra como a tabela de contigência, que é usada para facilitar o cálculo da performance dessa hipotése.

Tabela 3.1: Tabela de contigência

Classes	Objetos preditos como verdadeiros pelo modelo	Objetos preditos como falsos pelo modelo
Verdadeira	Número de verdadeiros Positivos (VP)	Número de falsos Negativos(FN)
Falsa	Número de falsos Positivos (FP)	Número de verdadeiros Negativos (VN)

Com as informações de VP, FN, FP e VN, pode-se calcular diversas medidas do modelo construído, tais como: sensibilidade, especificidade e acurácia.

A sensibilidade evidencia a taxa de verdadeiros positivos encontrada pelos modelo e é calculada por:

$$\frac{VP}{VP + FN}$$

Em aplicações como uma que decide se a pessoa tem ou não doenças infecciosas, por exemplo, a sensibilidade precisa ser alta, pois uma sensibilidade baixa pode evidenciar um número elevado de falsos negativos, classificando assim uma pessoa que necessita entrar em quarentana como uma pessoa saudável, que poderá sair e infectar as outras pessoas.

A especificidade evidencia a taxa de verdadeiros negativos encontradas pelo modelo e é calculada por:

$$\frac{VN}{VN + FP}$$

Em aplicações como uma que decide amputações, por exemplo, a especificidade precisa ser alta, pois uma especificidade baixa pode evidenciar um número elevado de falsos positivos, classificando assim uma pessoa que não necessita de ser amputada como uma que necessita, fazendo-a perder um membro sem necessidade.

Por fim, temos a acurácia, que calcula a taxa de eficiência em geral do modelo gerado, e ela é calculada da seguinte forma:

$$\frac{VP + VN}{VP + VN + FP + FN}$$

. Em conjunto com boas taxas de especificidade e de sensibilidade, dependendo da aplicação, uma porcentagem alta de acurácia evidencia um bom modelo.

3.1.3 Aprendizagem por reforço

A aprendizagem por reforço baseia-se no conceito de aprender a cada interação, para atingir determinado resultado. O programa é o elemento que faz as decisões, e também aprende. O programa percebe e interage com o ambiente, o qual é caracterizado por todos os outros elementos, exceto o programa. As ações tomadas pelo programa geram recompensas, sendo que essas recompensas dizem qual a melhor ação a ser tomada, dados os possíveis estados do ambiente conhecidas [84]. A Figura 3.3 mostra um ciclo de iteração de aprendizagem por reforço.

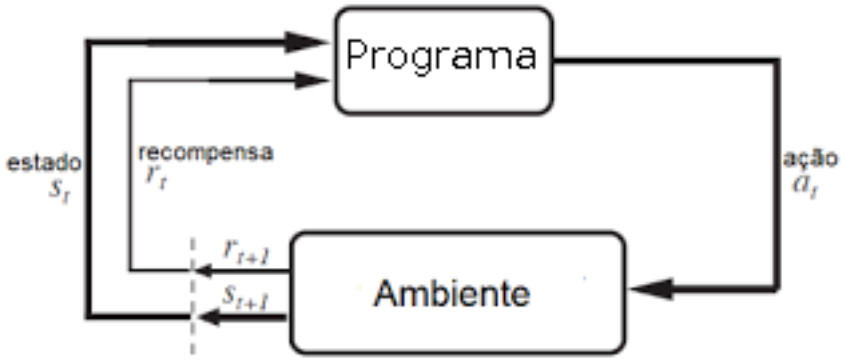

Figura 3.3: Ciclo de iteração de aprendizagem por reforço (adaptado de [84]).

A aprendizagem por reforço pode ser usada quando se necessitam de programas com maior autonomia, inseridos em um ambiente que possui exemplos de ações com recompen-

sas que sirvam como parâmetros para determinar a próxima ação a ser tomada. Neste caso, recompensas pode ser boas ou ruins, dependendo da ação tomada pelo programa. Tendo isso em mente, o papel do aprendizagem por reforço é usar recompensas obtidas para aprender qual ação é ótima, ou próxima da ação ótima, em determinado ambiente [66].

3.1.4 Aprendizagem semi-supervisionada

O aprendizagem semi-supervisionada é uma metodologia que busca estender a aprendizagem supervisionada usando técnicas de aprendizagem não-supervisionada. Em alguns casos, esses algoritmos superam a performance dos dois métodos, se fossem utilizados sozinhos.

Neste método, os dados de entrada normalmente são constituídos por um grupo de dados $X = \{x_1, ..., x_{i \in \mathbb{N}}\}$ divididos em dois subgrupos: (i) um grupo $X_l = \{x_1, ..., x_l\}$, em que cada x possui um dado em $Y_l = \{y_1, ..., y_l\}$ correspondente, que representa sua classe; e (ii) um grupo $X_u = \{x_1, ..., x_u\}$ de dados sem classificação previamente conhecida [40].

3.2 *SVM*

Neste projeto, o método de aprendizagem supervisionado SVM foi escolhido para a construção do modelo de classificação de lincRNAs.

3.2.1 Conceitos básicos

Os seguintes conceitos são importantes para entender o SVM:

- Classe: é a classificação dada a um objeto;
- Conjunto de Exemplos: Os conjuntos de exemplos são divididos em conjunto de treinamento e conjunto de teste. O conjunto de treinamento, como o próprio nome sugere, é utilizado para treinar o método, enquanto o conjunto de teste, é usado para validar o método;
- *Kernel*: função que possibilita operações em dimensões mais altas;
- *Overfitting*: ocorre quando o algoritmo se ajusta para um conjunto de dados muito específico, sendo assim ineficaz para lidar com um conjunto de dados mais geral.

3.2.2 Descrição

Um dos métodos supervisionados mais usados atualmente [72], o SVM classifica grupos com base na criação de margens de separação dos dados. Essas margens, delineadas por

uma fração dos dados de treinamento, são denominadas vetores de suporte [30], e separam conjuntos de dados em classes (Figura 3.4).

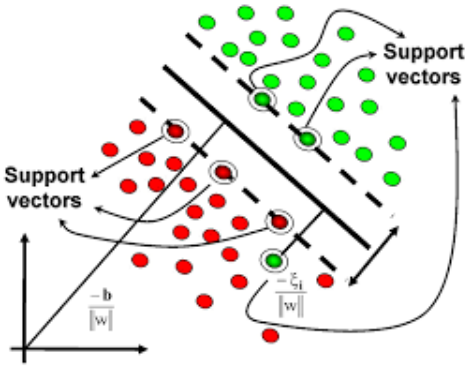

Figura 3.4: Exemplo de vetores de suporte de dimensão 2 [1].

O SVM é um método não-paramétrico, pelo fato de não se limitar apenas ao tamanho do conjunto de treinamento, e busca construir um hiperplano como superfície de decisão, de tal modo que a margem de separação entre as classes seja maximizada [48]. Na fase de treinamento, as classes serão separadas por uma função. Na fase de testes, dados não classificados terão suas classes preditas pelo modelo SVM construído na fase de treinamento. Na Figura 3.5, é mostrado um conjunto de classificadores lineares (hiperplanos) separando duas classes distintas, e também podemos visualizar o classificador que maximiza a margem de separação entre essas classes.

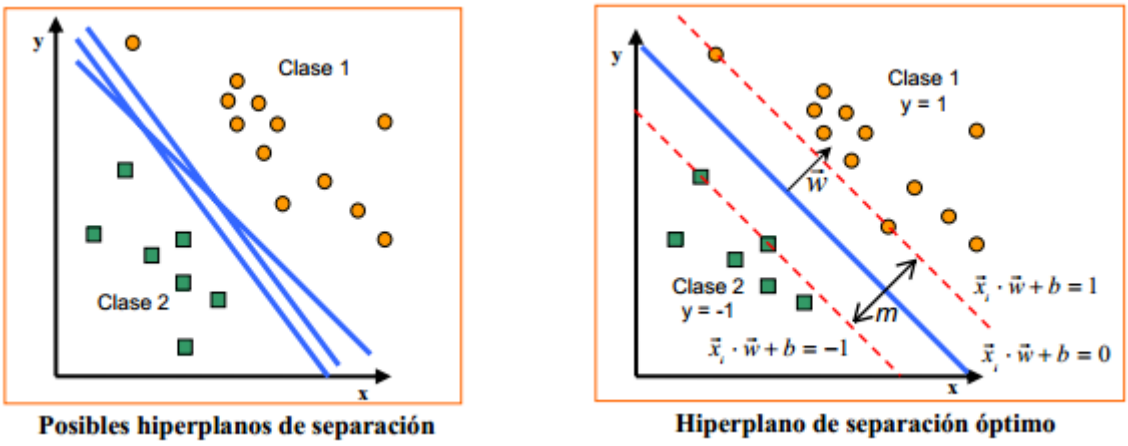

Figura 3.5: À esquerda, os classificadores lineares separando as classes do conjunto de dados e à direita o classificador ótimo que maximiza a margem de separação entre as classes [4].

Como visto, vários hiperplanos de separação podem ser gerados, mas para diminuir as chances de erro de classificação, tornar a classificação menos suscetível ao *overfitting* e por questões de performance [54], sempre se deve maximizar a margem de separação. Esses erros de classificação podem ocorrer quando objetos de uma classe se encontram dentro da margem de separação máxima. Quando utilizada outra margem de separação, esses objetos podem ser classificados erroneamente. Além disso, ao maximizar a margem de separação, a chance de ocorrer mínimos locais diminui, melhorando assim a classificação [54].

Como o SVM busca encontrar um separador linear para dividir os dados em subgrupos, pode-se ter dificuldade usando métodos de separação lineares simples, quando os dados não são linearmente separáveis. Para contornar esse problema, pode-se usar uma função *kernel* para elevar a dimensão do espaço em questão, tornando os componentes linearmente separáveis em dimensões mais elevadas (Figura 3.6).

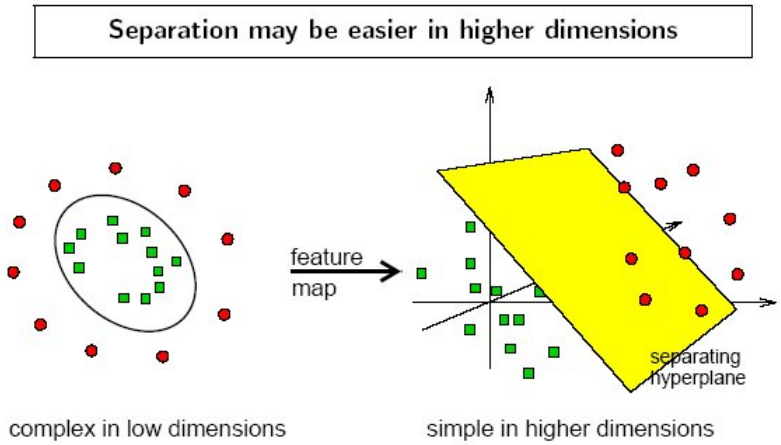

Figura 3.6: Exemplo de SVM com separador em três dimensções [26].

A escolha do método SVM foi baseada nas seguintes justificativas:

- Construção de uma margem separadora máxima, para diminuir erros de classificação;
- Criação de hiperplanos, mesmo quando as classes não são separáveis linearmente, usando *kernels*;
- Como o método é não-paramétrico, a capacidade de generalização do modelo construído é maior.

A seguir, algumas funções *kernel* e seu funcionamento serão brevemente descritos.

3.2.3 *Kernel*

Seguindo Haussler [47], dado um conjunto X e uma função $K : X \times X \to \mathbb{R}$, dizemos que K é uma função *kernel* em $X \times X$ se K é simétrico, por exemplo, se para todo x e $y \in XK(x, y) = K(y, x)$; e se K é positivo, por exemplo, se para todo $N \geq 1$ e para todo $x_1, ..., x_N \in X$, a matriz K definida por $K_{ij} = K(x_i, y_j)$ é positiva, ou seja, a equação $\sum_{ij} c_i c_j K_{ij} \geq 0$ é positiva para todo $c_1, ..., c_N \in \mathbb{R}$.

A função kernel denota um produto interno em um espaço de características e é denotada por $K(x, y) = (\phi(x), \phi(y))$. Esse espaço de características tem dimensão mais alta e é usado com o intuito de que os objetos do espaço de entrada sejam transformados para esse espaço de características de tal forma que possam ser separados mais facilmente. Na Figura 3.7, podemos verificar o mapeamento de um objeto no espaço de entrada sendo mapeado para o espaço de características dado uma função ϕ.

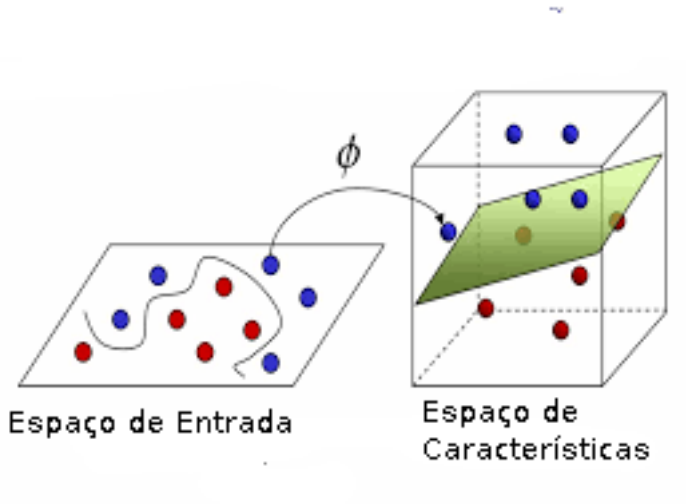

Figura 3.7: Exemplos de um objeto em um espaço de entrada de duas dimensões, sendo mapeado para o espaço de características de terceira dimensão, por uma função ϕ, para que se obtenha um melhor separador [13].

Resumindo, um *kernel* é uma função de produto interno aplicado a objetos de dados para mapea-los em um espaço de dimensão mais elevada através de uma transformação ϕ. A Tabela 3.2 mostras as funções *kernel* mais utilizadas.

Na Tabela 3.2 podemos notar que algumas funções de *kernel*, tal como o RBF, possui um parâmetro γ. Esse parâmetro é ajustável e ele é, recorrentemente, usado no SVM com um valor ótimo para obter melhores resultados de classificação. Além da variação de valor do γ podemos aplicar diversas técnicas para tentar obter um modelo SVM com melhor acurácia, como a alteração de um outro paramêtro, o C (custo), que afeta a penalidade

Tabela 3.2: Tabela de *kernels* mais utilizados.

Kernel	Fórmula
Linear	$X_i \bullet X_j$
Polinomial	$(\gamma X_i \bullet X_j + C)^d$
RBF (radial)	$exp(-\gamma \mid X_i - X_j \mid^2)$
Sigmoid	$tanh(\gamma X_i \bullet X_j + C)$

em aceitar objetos no lado errado da margem para obter um melhor modelo, e o *K-fold cross-validation*, detalhado a seguir, pois foi usado nesse projeto.

Uma observação importante é que os valores atribuidos a C podem influenciar no problema de *overfitting*, pois quanto maior o valor de C mais restritos são os vetores de suporte as respectivas classes. Isso significa que o modelo pode perder a capacidade de generalização e classificar incorretamente novos dados.

3.2.4 *K-fold cross-validation*

Algumas técnicas de validação cruzada propiciam a otimização quando utilizadas em algoritmos de aprendizagem de máquina, sendo que, neste trabalho, utilizou-se o *k-fold cross-validation*.

No *k-fold cross-validation* os dados são particionados em k segmentos (*fold*) de mesmo tamanho. Após essa divisão, k iterações de treinamento e validação são realizadas de modo que, a cada iteração um segmento dos dados é usado como validação enquanto os outro $k-1$ segmentos são utilizados como treinamento. Os dados normalmente são estratificados ao serem particionados, ou seja, eles são rearranjados de forma a assegurar uma boa representatividade para cada segmento [71]. A Figura 3.8 mostra um exemplo da utilização do *k-fold crossvalidation*.

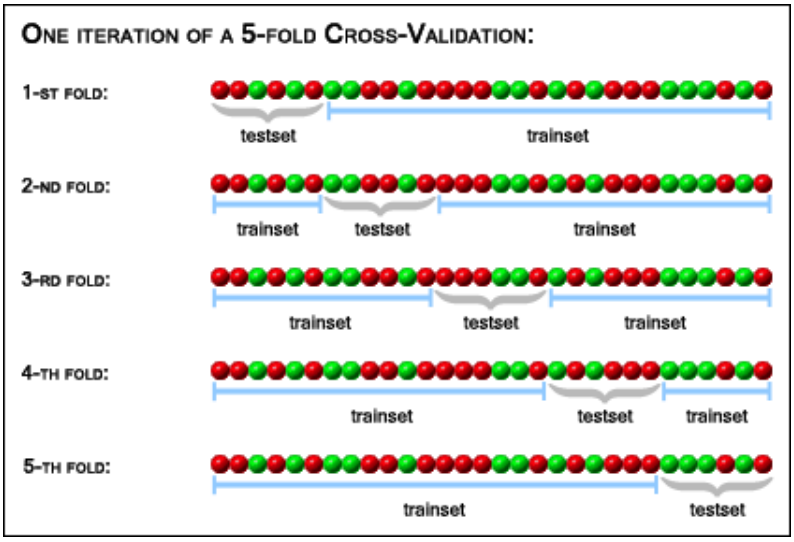

Figura 3.8: Exemplo de *K-fold cross-validation* com $k = 5$ [3].

Capítulo 4

Método de predição de lincRNAs

Este capítulo apresentará o método proposto nessa monografia. Na Seção 4.1, o método genérico de classificação de lincRNAs será descrito. Na Seção 4.2, um estudo de caso para classificação de lincRNAs em humanos e camundongos, focando na análise de performance do modelo SVM construído neste projeto será descrito. Na Seção 4.3, um estudo de caso para classificação de lincRNAs na cana-de-açúcar será descrito.

4.1 Descrição geral

As fases genéricas do método de classificação de lincRNAs são descritos como segue. Primeiramente, deve-se criar um banco de dados com lincRNAs já conhecidos, que serão utilizado na SVM. Esses dados serão filtrados para remover sequências com potencial alto de codificar proteínas. O próximo passo é definir características de lincRNAs, ainda pouco conhecidas, que serão utilizadas. Após obter as características, serão usados os dados filtrados como conjuntos de treinamento e teste da SVM. Na Figura 4.1, é mostrado o *pipeline* deste método.

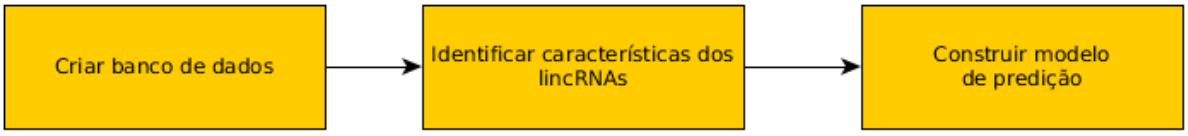

Figura 4.1: *Pipeline* para classificação de lincRNAs.

4.2 Estudo de Caso 1: humanos e camundongos

4.2.1 Dados

Neste trabalho, foram utilizados dois bancos de dados: Ensembl [7] e HAVANA [12], pois são as únicas bases com dados suficientes de lincRNAs para gerar o modelo de classificação. Transcritos codificadores de proteínas (*protein coding transcripts* - PCTs) foram usados como dados de treinamento negativos e lincRNAs como dados de treinamento positivos. Os dados (PCTS e lincRNAs) foram restritos a duas espécies, pois são as únicas com dados de lincRNAs suficientes para a construção de um modelo com boa generalização. Então foram selecionados dois genomas: o GRCm38 do *Mus musculus* [7, 12] e o GRCh38 do *Homo sapiens* [7, 12].

Embora os bancos de dados selecionados possuam informação de boa qualidade, foram ainda filtradas para confirmar sua qualidade. Nessa etapa de filtragem, os transcritos foram escolhidos por tamanho, taxa de conservação e ORFs. Além disso, utilizamos o Blast com o banco de dados do Refseq [70] para eliminar as sequências codificadoras de proteínas.

4.2.2 Características

Foram escolhidas 9 características, agrupadas em 3 conjuntos seguindo Sun et al [83]. Diversos estudos demonstraram que lincRNAs possuem uma taxa de conservação menor do que PCTs [38]. Por isso, o primeiro conjunto é composto pela taxa de conservação. Observamos que conservação indica quanto o transcrito é similar a transcritos de outros organismos. Para calcular a taxa de conservação de cada transcrito, primeiramente usamos o arquivo que possui índices de conservação do organismo (phastCons [80]) da página UCSC [28]. Esses índices de conservação são atribuídos a partes específicas do genoma. Para calcular a conservação do transcrito, deve-se verificar se a posição do transcrito no genoma encontra-se nessas partes específicas, e o transcrito deve pertencer a um genoma conhecido. Dadas essas duas informações, a média da pontuação de cada nucleotídeo é calculada e a pontuação total é gerada para cada transcrito. Por exemplo, suponha que a taxa de conservação do genoma humano nas posições 1 à 10 seja 25, caso tenhamos um transcrito, também humano, na posição 1 a 5, sua taxa vai ser $(25 \div 10) \times (5)$, ou seja, (média de pontuação de conservação por base das posições 1 a 10) $\times$ (número de bases do transcrito nas posições 1 a 10).

O segundo conjunto de dados leva em conta o potencial de codificação de cada transcrito. Esse potencial é calculado usando duas características relacionadas às ORFs. A primeira característica é o tamanho da ORF e a segunda é a proporção dada pelo tama-

nho da ORF dividido pelo tamanho da sequência. Quanto menor o valor de ambas as características, mais provável que o transcrito seja de fato lincRNA, pois possui um baixo potencial de codificação. Para calcular as ORFs de cada transcrito, foi usado o programa txCdsPredict da UCSC [28].

O terceiro conjunto de dados agrupa as 6 características restantes, que são a frequência de di e tri-nucleotídeos específicos (GC, CT, TAG, TGT, ACG e TCG) em cada transcrito. De acordo com Sun et al [83], esses di e tri-nucleotídeos aparecem com alta frequência em lincRNAS de humanos e camundongos.

4.2.3 Extração das características

Para a extração das características (Figura 4.2), foram usados o programa txCdsPredict para localizar ORFs nos transcritos, além de scripts Perl para determinar os conjuntos de características conforme descritos na seção anterior.

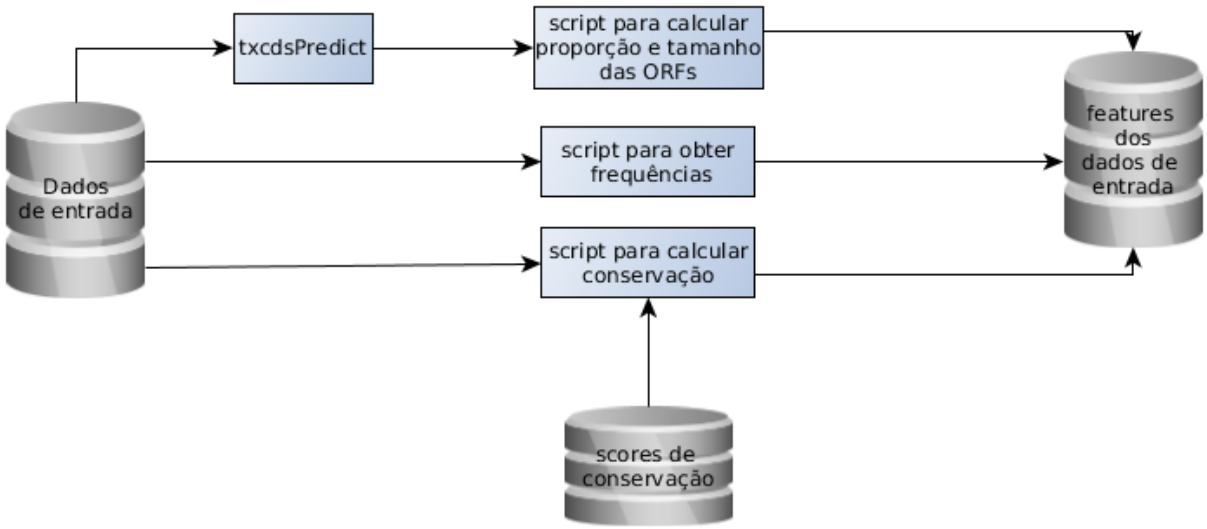

Figura 4.2: Processo de extração de características dos transcritos de entrada.

4.2.4 Opções de Treinamento

Para a criação de um modelo SVM com boa performance, os seguintes parâmetros de treinamento foram levados em conta:

- *10-fold cross-validation*;
- *Kernel* radial;

- lincRNAs como conjuntos de treinamento e teste positivos;
- PCTs como conjuntos de treinamento e teste negativos;

A Figura 4.3 mostra o método de SVM utilizado.

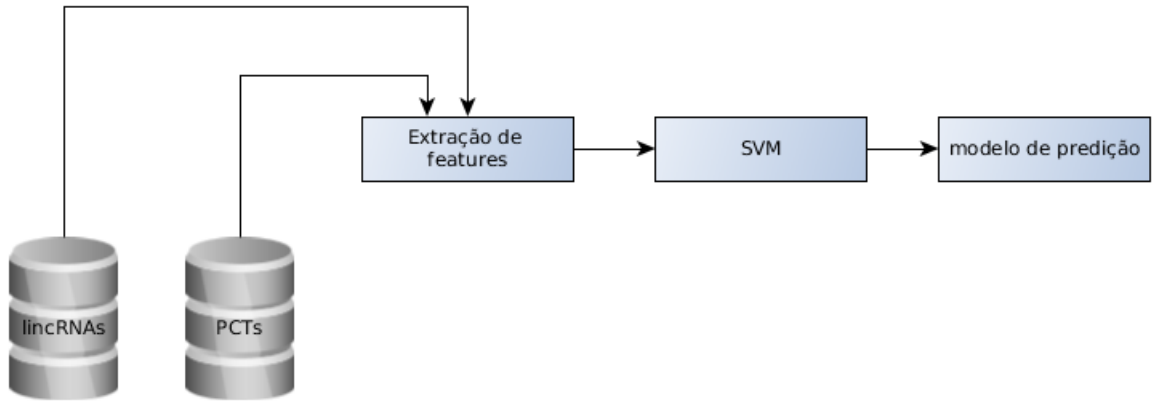

Figura 4.3: *Pipeline* para classificação de lincRNAs em humanos e camundongos.

4.2.5 Testes

Para estudar as características escolhidas, e verificar se elas poderiam de fato construir um bom modelo, foram realizados vários testes. Cada teste foi aplicado em três bancos de dados: Humanos, Camundogo e o terceiro combinando Humanos + Camundongos. As abordagens de cada teste foram aplicados tanto nos dados de treinamento quanto nos de testes.

Os testes foram realizados da seguinte forma:

- Teste 1: cada um dos três conjuntos de características foi usado separadamente pela SVM;
- Teste 2: dois conjuntos de características foram usados ao mesmo tempo pela SVM;
- Teste 3: os três conjuntos de características foram usados simultaneamente pela SVM;

4.3 Estudo de caso 2: Cana-de-açúcar

Como dito antes, em colaboração com o Prof. Paulo Cavalcanti Gomes Ferreira do Instituto de Ciências Biomédicas da UFRJ, foi feito um estudo de caso para identificar lincRNAs de cana-de-açúcar.

4.3.1 Informações sobre cana-de-açúcar

A cana-de-açúcar (*Saccharum officinarum*) é um membro da subfamília Panicoideae e é uma das plantas mais importantes no aspecto industrial, devido a grande quantidade de sacarose que contém (10 a 15% do peso do seu caule) [46]. Cerca de 70% do açúcar em todo mundo é derivado da cana-de-açúcar [77], e o bagaço da cana-de-açúcar pode ser usado como matéria prima para fabricar papel e alimentar animais [59, 41].

Apesar da importância da cana-de-açúcar, poucos dados genómicos relacionados a ela são encontrados na literatura. Então, para encontrar funções em transcritos da cana-de-açúcar, são usados organismos próximos evolutivamente. A Figura 4.4 mostra uma árvore filogenética, que mostra organismos próximos evolutivamente à cana-de-açúcar.

Figura 4.4: Arvoré filogenética da cana-de-açúcar. Note que o arroz (*Oriza sativa*), o Sorgo (*Sorghum bicolor*) e o milho (*Zea mays*) são os organismos mais próximos evolutivamente, que tem dados de lncRNAs [46].

4.3.2 LincRNAs na cana-de-açúcar

Para classificar os transcritos que poderiam ser lincRNAs, foi desenvolvido um novo *pipeline*, que usa o modelo SVM para dar suporte aos resultados. Como ainda não existe nenhum genoma da cana-de-açúcar, os passos do *pipeline* utilizaram organismos próximos evolutivamente, a fim de dar um maior suporte na parte de anotação. O *pipeline* utilizado será descrito a seguir.

Primeiramente foi feita a filtragem de regiões codificadoras. Usando o BLAST e os bancos de dados Repbase [56] e PlantGDB [43], regiões codificadoras da cana-de-açúcar serão descartados.

Após descartadas as regiões codificadoras, usando o BLAST e o banco de dados CantataDB [85], os transcritos da cana-de-açúcar que forem anotados como lncRNAs serão separados.

Para dar suporte a classificação de lincRNAs, o próximo passo mapeia os transcritos da cana-de-açúcar a um genoma de referência evolutivamente próximo, para descobrir

quais lncRNAs estariam em uma posição intergênica. O genoma de referência utilizado foi o do Sorgo, presente no banco de dados do PlantGDB [43]. Neste passo, os transcritos da cana-de-açúcar classificados como codificadores de proteína e como lncRNAs foram mapeados, sendo que os classificados como lncRNAs, mapeados entre os classificados como codificadores de proteína, provavelmente estão em uma região intergênica. O mapeamento foi feito utilizando o montador Segemehl [50].

O último passo do *pipeline* foi treinar um modelo SVM para cana-de-açúcar e classificar quais dos transcritos resultantes dos passos anteriores são lncRNAS. Para criar esse modelo, o conjunto de dados de treinamento positivo foi formado pelos lncRNAS do banco de dados CantataDB [85], pois não temos informações de lncRNAs nem de lincRNAs da cana-de-açúcar. Como conjunto de dados de treinamento negativo, foram usados os transcritos da cana-de-açúcar classificados como codificadores de proteína, sendo utilizado um número suficiente de dados para dar suporte à criação do modelo SVM.

Na Figura 4.5, é mostrado o *pipeline* descrito acima.

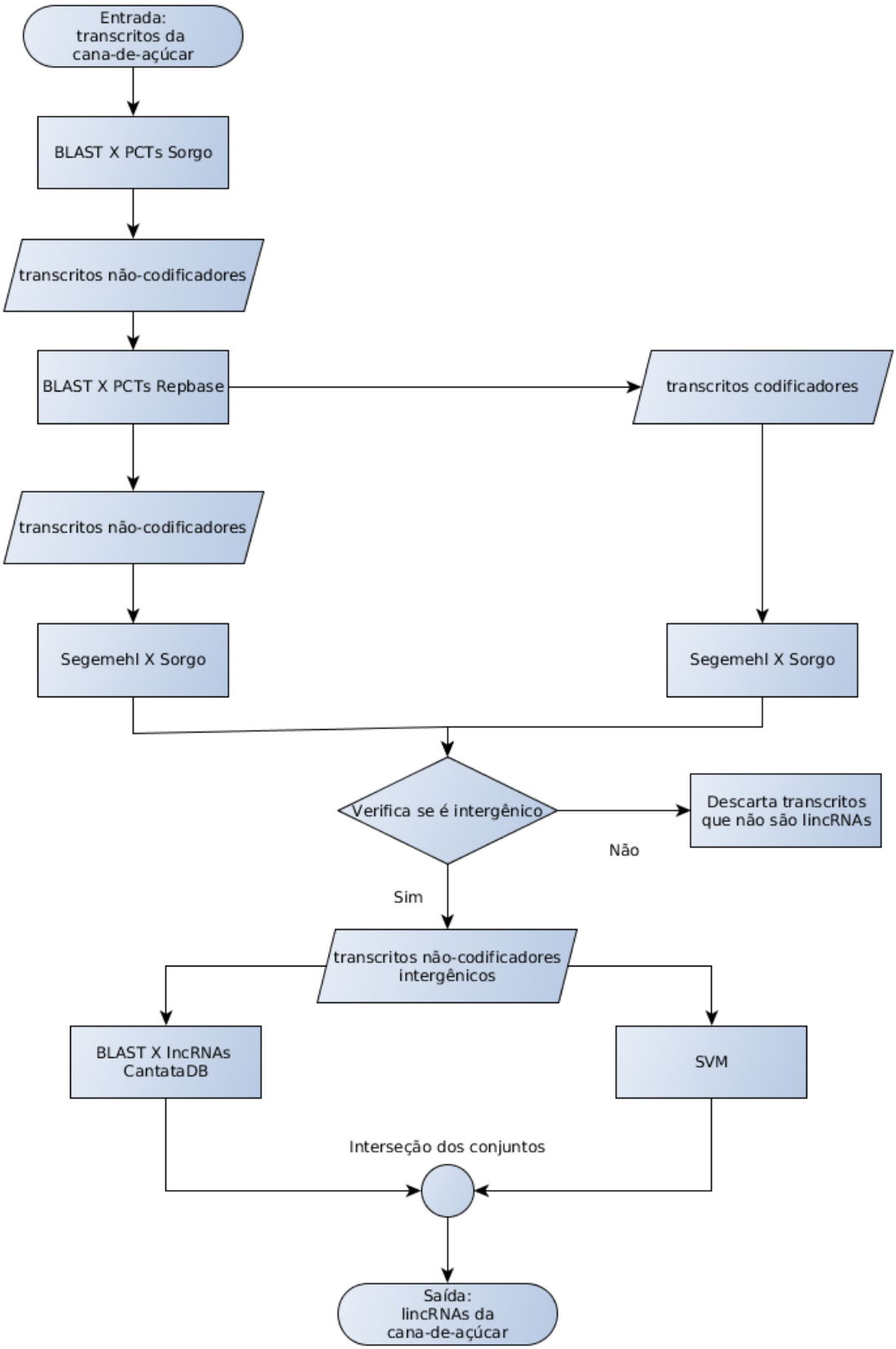

Figura 4.5: *Pipeline* para classificação de lincRNAS na cana-de-açúcar.

Capítulo 5

Resultados

Neste capítulo, serão discutidos os resultados obtidos a partir do método implementado. Na Seção 5.1, será apresentada a performance do método usando os dados de humano e camundongo. Na Seção 5.2, discutimos um estudo de caso para a classificação de lincRNAs em transcritos de cana-de-açúcar.

5.1 Performance

Nas subseções seguintes, os resultados de performance dos modelos SVM construídos para humanos e camundongos serão apresentados.

5.1.1 Camundongo

Neste modelo, foram usados 400 lincRNAs e 400 PCTs para treinamento e 400 lincRNAs e 400 PCTs para teste, usando as características de conservação, ORFs e frequência.

Teste 1: cada um dos três conjuntos de características testados separadamente

As Tabelas de contingência 5.1, 5.2 e 5.3 mostram os resultados dos testes para os três conjuntos de características descritos na Seção 4.2.2.

Tabela 5.1: Tabela de contingência de Conservação

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	370	272
PCTs	30	128

Tabela 5.2: Tabela de contingência de ORFs

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	380	79
PCTs	20	321

Tabela 5.3: Tabela de contingência de Frequências de nucleotídeos

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	366	147
PCTs	34	253

Na Tabela 5.4 e na Figura 5.1 é mostrada a compilação de resultados acima. Podemos notar que as ORFs são as características mais relevantes para o modelo, seguida pelas frequências e pela conservação.

Tabela 5.4: Teste 1 - Performance do modelo SVM para o camundongo

Conjunto	Sensibilidade	Especificidade	Acurácia
Conservação	92%	32%	62%
ORFs	95%	80%	87%
Frequências	91%	63%	77%

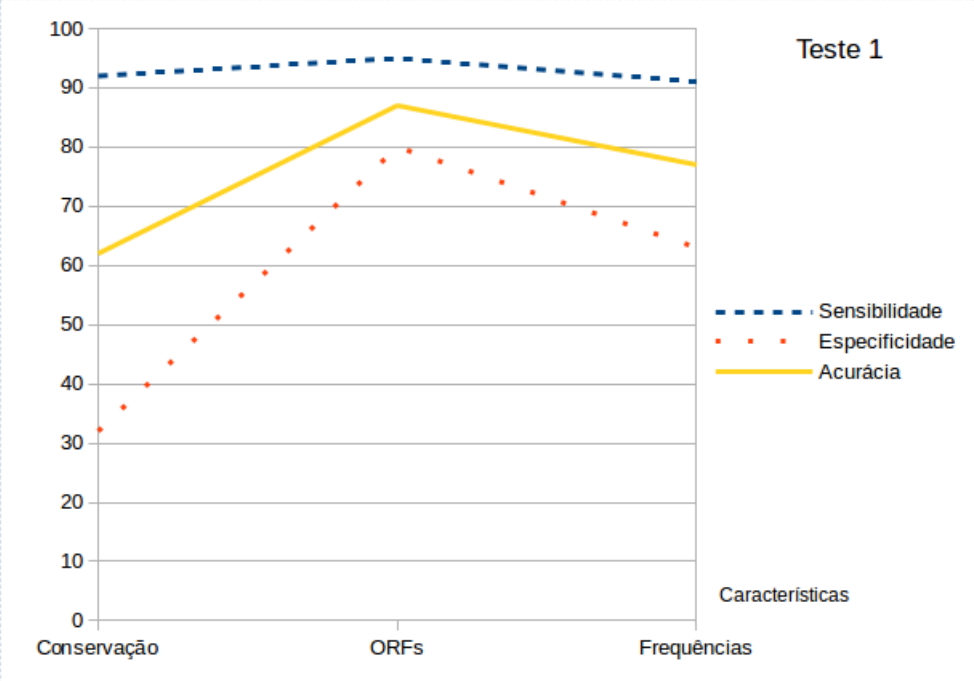

Figura 5.1: Performance do modelo SVM para o teste 1.

Teste 2: dois conjuntos de características testados simultaneamente

As Tabelas de contingência 5.5, 5.6 e 5.7 mostram os resultados dos testes para combinações dois a dois, dos três conjuntos de características descritos anteriormente.

Tabela 5.5: Tabela de contingência de Conservação + ORFs

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	378	72
PCTs	22	328

Tabela 5.6: Tabela de contingência conjunto Conservação + Frequências

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	361	133
PCTs	39	267

Tabela 5.7: Tabela de contingência conjunto ORFs + Frequências

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	381	76
PCTs	19	324

Na Tabela 5.8, podemos notar que qualquer característica combinada com ORFs gera modelos mais acurados.

Tabela 5.8: Teste 2 - Performance modelo SVM camundongo

Conjuntos	Sensibilidade	Especificidade	Acurácia
Conservação + ORFs	94%	82%	88%
Conservação + Frequências	90%	66%	78%
ORFs + Frequências	95%	81%	88%

Teste 3: três conjuntos de características testados simultaneamente

A Tabela de contingência 5.9 mostra os resultados do teste para os três conjuntos de características juntos.

A Tabela 5.10 mostra a performance do modelo SVM testado com os três conjuntos de parâmetros ao mesmo tempo. Quando comparada a Tabela 5.8, podemos verificar que a sensibilidade decresceu e a especificidade aumentou, melhorando a acurácia do método.

A Figura 5.2 mostra a sensibilidade, a especificidade e a acurácia, para dois conjuntos de parâmetros, e os três conjuntos simultaneamente.

Tabela 5.9: Tabela de contingência do camundongo

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	375	59
PCTs	25	341

Tabela 5.10: Teste 3 - Performance do modelo SVM para o camundongo

Conjunto	Sensibilidade	Especificidade	Acurácia
Conservação + ORFs + frequências	93%	85%	89%

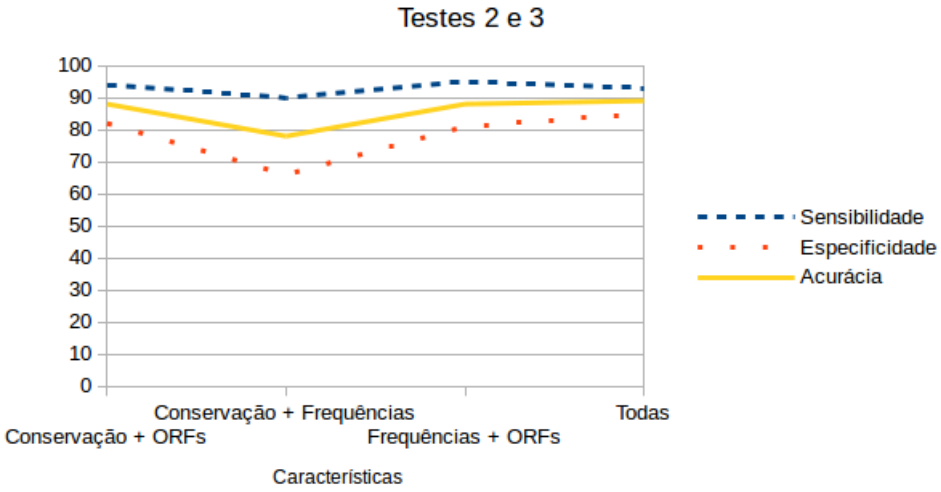

Figura 5.2: Performance do modelo SVM para os testes 2 e 3.

5.1.2 Humano

Neste modelo, foram usados 400 lincRNAs e 400 PCTs para treinamento e 400 lincRNAs e 400 PCTs para teste, com as mesmas características: Conservação, ORFs e frequências.

Teste 1: cada um dos três conjuntos de características usado separadamente

As Tabelas de contingência 5.11, 5.12 e 5.13 mostram os resultados dos testes para os três conjuntos de características descritos na Seção 4.2.2.

Tabela 5.11: Tabela de contingência de Conservação

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	340	219
PCTs	60	181

Tabela 5.12: Tabela de contingência de ORFs

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	398	2
PCTs	2	398

Tabela 5.13: Tabela de contingência de Frequências de nucleotídeos

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	331	171
PCTs	69	229

Na Tabela 5.14 e na Figura 5.3, mostramos uma compilação dos resultados acima. Assim como em camundongos, podemos notar que as ORFs são as características individuais mais relevantes para o modelo. Notamos também que, ao contrário de camundongos, a conservação mostra-se mais relevante em comparação com as frequências de nucleotídeos.

Tabela 5.14: Teste 1 - Performance do modelo SVM para humanos

Conjunto	Sensibilidade	Especificidade	Acurácia
Conservação	85%	45%	65%
ORFs	99%	99%	99%
Frequências	82%	57%	70%

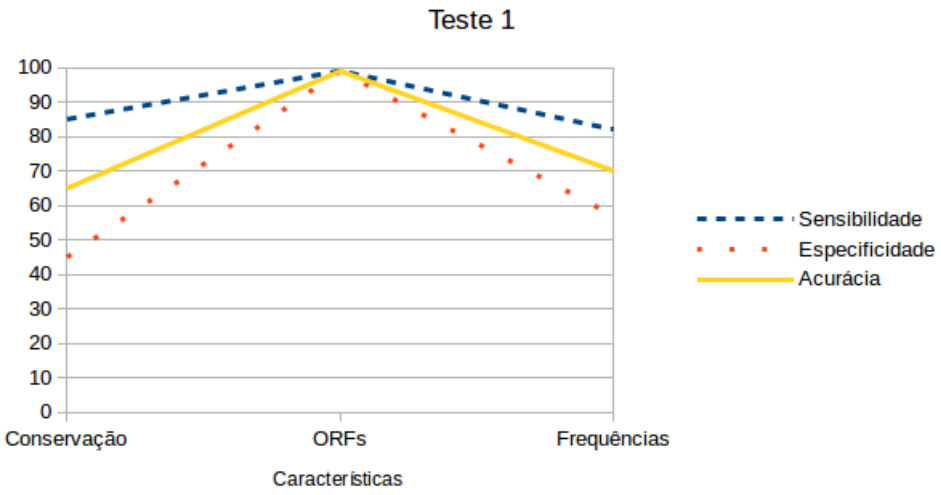

Figura 5.3: Performance do modelo SVM para o teste 1.

Teste 2: dois conjuntos de características testados simultaneamente

As Tabelas de contingência 5.15, 5.16 e 5.17 mostram os resultados dos testes para combinações dois a dois, dos três conjuntos de características descritos anteriormente.

Tabela 5.15: Tabela de contingência de Conservação + ORFs

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	398	2
PCTs	2	398

Tabela 5.16: Tabela de contingência de Conservação + Frequências

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	336	125
PCTs	64	275

Tabela 5.17: Tabela de contingência de ORFs + Frequências

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	395	2
PCTs	5	398

Na Tabela 5.18, podemos notar que as características combinadas com as ORFs geram os modelos mais acurados, com quase 100% de acurácia. Isso levanta a questão sobre se as outras características, além das ORFs, são de fato relevantes para construir modelos SVM de classificação de lincRNAs para humanos. Uma outra explicação poderia ser do modelo apresentar *overfitting*.

Tabela 5.18: Teste 2 - Performance do modelo SVM para humanos

Conjuntos	Sensibilidade	Especificidade	Acurácia
Conservação + ORFs	99%	99%	99%
Conservação + Frequências	84%	68%	76%
ORFs + Frequências	98%	99%	99%

Teste 3: três conjuntos de características testados simultaneamente

A Tabela de contingência 5.19 mostra os resultados do teste para os três conjuntos de características juntos.

Tabela 5.19: Tabela de contingência de humanos

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	375	59
PCTs	25	341

A Tabela 5.20 mostra a performance do modelo SVM testado com os três conjuntos de parâmetros ao mesmo tempo. Quando comparada a Tabela 5.18, podemos verificar que a sensibilidade, a especificidade e a acurácia são iguais ao modelo gerado pela combinação de ORFs e frequências.

Tabela 5.20: Teste 3 - Performance do modelo SVM para humanos

Conjunto	Sensibilidade	Especificidade	Acurácia
Conservação + ORFs + Frequências	98%	99%	99%

A Figura 5.4 mostra a sensibilidade, a especificidade e a acurácia, para dois conjuntos de parâmetros, e os três conjuntos simultaneamente.

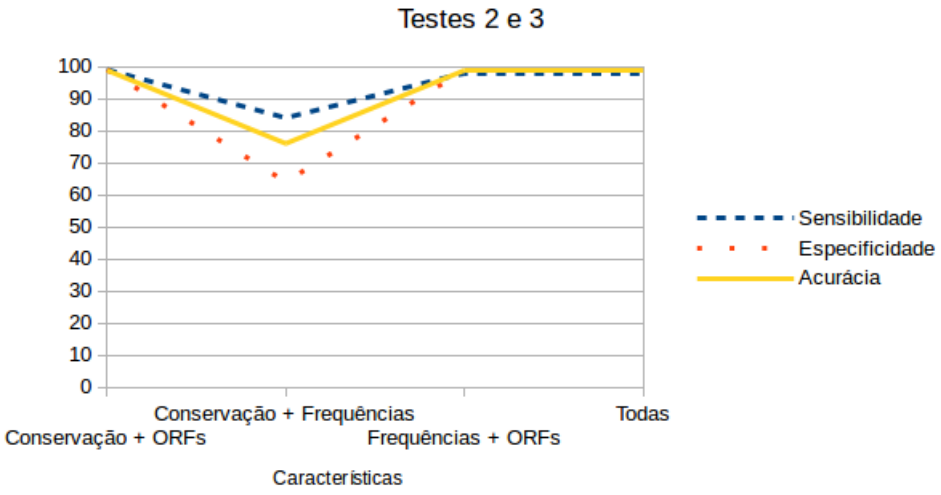

Figura 5.4: Performance do modelo SVM para os testes 2 e 3.

5.1.3 Humano + Camundongo

Neste modelo, foram usados 800 lincRNAs e 800 PCTs para treinamento e 800 lincRNAs e 800 PCTs para teste, com as mesmas características.

Teste 1: cada um dos três conjuntos de características usado separadamente

As Tabelas de contingência 5.21, 5.22 e 5.23 mostram os resultados dos testes para os três conjuntos de características descritos na Seção 4.2.2.

Tabela 5.21: Tabela de contingência de Conservação

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	710	482
PCTs	90	318

Tabela 5.22: Tabela de contingência de ORFs

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	784	113
PCTs	16	687

Tabela 5.23: Tabela de contingência de Frequências de nucleotídeos

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	684	315
PCTs	116	485

Na Tabela 5.24 e na Figura 5.5, mostramos uma compilação dos resultados acima. Podemos notar que as ORFs, assim como nos modelos individuais de camundongos e do humanos, são as características predominantes na combinação dos dados.

Tabela 5.24: Teste 1 - Performance do modelo SVM para humanos + camundongos

Conjunto	Sensibilidade	Especificidade	Acurácia
Conservação	88%	39%	64%
ORFs	98%	85%	91%
Frequências	85%	60%	73%

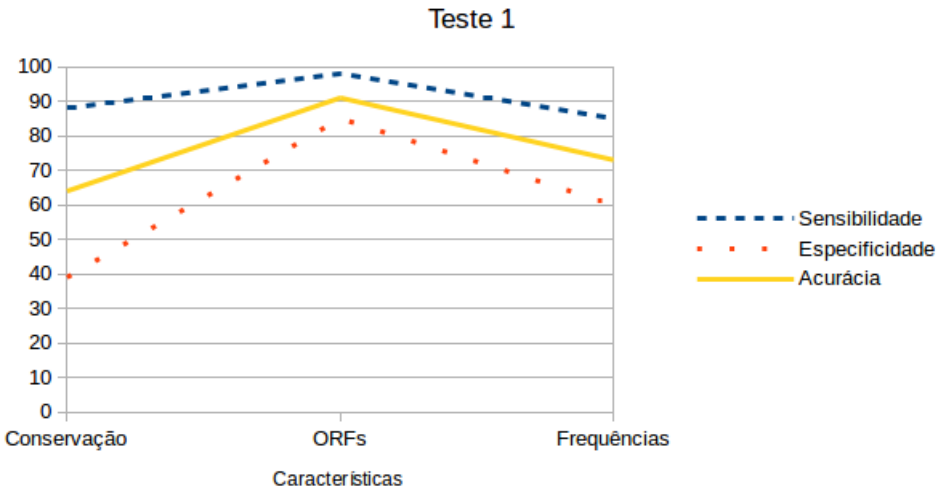

Figura 5.5: Performance do modelo SVM para o teste 1.

Teste 2: dois conjuntos de características testados simultaneamente

As Tabelas de contingência 5.25, 5.26 e 5.27 mostram os resultados dos testes para combinações dois a dois, dos três conjuntos de características descritos anteriormente.

Tabela 5.25: Tabela de contingência de Conservação + ORFs

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	781	109
PCTs	19	691

Tabela 5.26: Tabela de contingência de Conservação + Frequências

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	686	248
PCTs	114	552

Tabela 5.27: Tabela de contingência de ORFs + Frequências

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	774	105
PCTs	26	695

Na Tabela 5.28, podemos notar que as características combinadas com as ORFs geram os modelos mais acurados. Quando comparados com os testes só de humanos, perdeu-se

um pouco de acurácia, mas isso talvez indique maior capacidade de generalização, o que precisa ser confirmado com testes de validação.

Tabela 5.28: Teste 2 - Performance do modelo SVM para humanos + camundongos

Conjuntos	Sensibilidade	Especificidade	Acurácia
Conservação + ORFs	97%	86%	92%
Conservação + Frequências	85%	69%	77%
ORFs + Frequências	96%	86%	91%

Teste 3: três conjuntos de características testados simultaneamente

A Tabela de contingência 5.29 mostra os resultados do teste para os três conjuntos de características juntos.

Tabela 5.29: Tabela de contingência Camundongo

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	775	106
PCTs	25	694

A Tabela 5.30 mostra a performance do modelo SVM testado com os três conjuntos de parâmetros ao mesmo tempo. Quando comparada a Tabela 5.20, podemos verificar que a acurácia decresceu, mas em compensação, o modelo constuído é mais genérico.

Tabela 5.30: Teste 3 - Performance do modelo SVM para humanos + camundongos

Conjuntos	Sensibilidade	Especificidade	Acurácia
Conservação + ORFs	91%	86%	91%

Na Figura 5.6, mostramos a compilação de resultados dos testes 2 e 3 para humanos + camundongos. Podemos notar que quando o modelo é treinado somente com as ORFs (teste 1), obtemos a mesma acurácia de quando o modelo é treinado com todas as características juntas. Embora tenha a mesma acurácia, houve alteração nos valores de sensibilidade e especificidade.

5.1.4 Comparação de performance com o iSeeRNA

Como dito antes, o iSeeRNA [83] é uma ferramenta que pode identificar lincRNAs em humanos e camundongos. Para poder testar a performance dos modelos construídos por nosso método, foram feitas comparações dos resultados obtidos do nosso modelo SVM com o iSeeRNA.

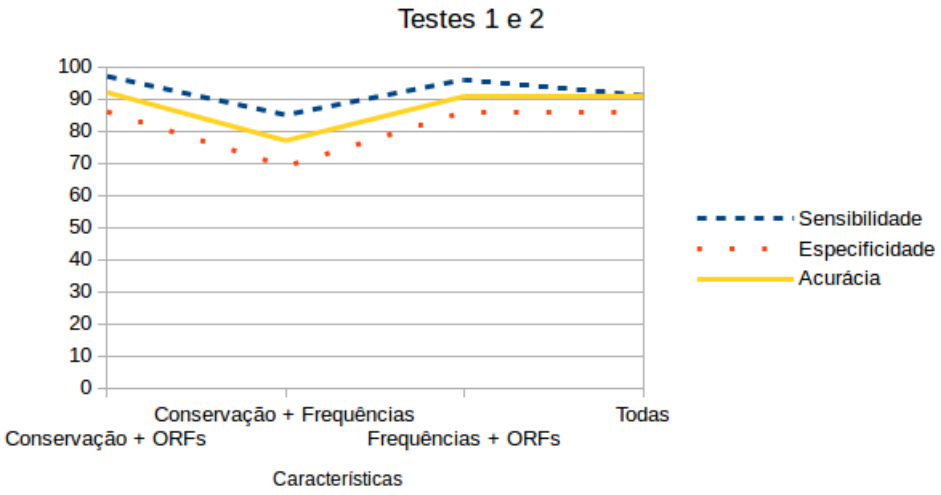

Figura 5.6: Performance do modelo SVM para os testes 2 e 3.

Como o iSeeRNA só aceita transcritos dos organismos camundongo (GRCm37 e GRCm38) [12] e humano (GRCh37) [12], tivemos alguns problemas na comparação de performance com os humanos, já que os dados de teste utilizados para testar nosso modelo são do genoma humano GRCh38.

A Tabela 5.31 mostra que, dos 800 transcritos de teste, 76 não foram classificados, devido a diferença de montagens dos genomas citadas acima, logo foram usados 724 transcritos para teste.

Tabela 5.31: Tabela de contingência do iSeeRNA para humanos

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	345	22
PCTs	333	24

Podemos notar que a acurácia do modelo para humanos do iSeeRNA usando os nossos dados de teste foi de 50%. A acurácia obtida foi ruim, comparada com a de 99% obtida por nosso modelo treinado com todas as características. Isso mostra que o iSeeRNA ou o nosso modelo SVM provavelmente está com *overfitting*. Podemos notar também que a classificação do iSeeRNA gera um grande número de falsos positivos. Provavelmente essa baixa performance do iSeeRNA tenha ocorrido devido ao fato das montagens dos genomas serem diferentes.

A Tabela 5.32 mostra os resultados da classificação do iSeeRNA de lincRNAs para camundongos.

Tabela 5.32: Tabela de contingência do iSeeRNA para camundongos

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lincRNAs	369	31
PCTs	340	60

A acurácia do modelo do iSeeRNA para camundongos foi de 53%. A sua acurácia é ruim, comparada com a de 89% obtida por nosso modelo treinado com todas características. Como os genomas foram os mesmos (camundongo GRCm38), podemos dizer que o modelo do iSeeRNA não obteve uma boa performance e gerou um grande número de falsos positivos.

A seguir, apresentamos uma comparação entre os resultados do iSeeRNA e nosso modelo:

- humanos:
  - conforme Sun et al [83]: 95.4%
  - nossos testes: 50%
- camundongo:
  - conforme Sun et al [83]: 94.2%
  - nossos testes: 53%

5.2 LincRNAs em cana-de-açúcar

Foram recebidos um total de 168.767 transcritos da cana-de-açúcar. Destes, 65.419 transcritos não foram identificados como codificadores de proteína, quando feito um BLAST entre os transcritos da cana-de-açúcar e transcritos codificadores de proteína do Sorgo obtidos no PlantGDB [43].

Dos 65.419 transcritos resultantes, 63.389 não foram identificados como codificadores de proteína, quando feito um BLAST entre eles e transcritos da subfamília Panicoideae (milho, arroz, sorgo) obtidos no Repbase [56].

Desses 63.389, 1.446 transcritos foram anotados como lncRNAs de acordo com o resultado do BLAST contra os lncRNAs do arroz e do milho, na base CantataDB [85].

Na etapa de montagem, dos 105.378 transcritos que foram classificados como codificadores de proténas, 10.499 foram mapeados contra o genoma de referência do Sorgo. Dos 63.389 transcritos não-codificadores restantes, 4.425 foram mapeados contra o genoma de referência do Sorgo.

Na etapa da identificação da posição dos transcritos obtivemos: 9.488 transcritos, dos 10.499 codificadores de proteína foram mapeados em região gênica; e 2.687, dos 4.425 transcritos não codificadores foram mapeados em regiões intergênicas dos 9.488.

Por outro lado, na etapa de classificação de lincRNAs, foi criado um modelo SVM com 13 características: tamanho das ORFs, proporção das ORFs e frequência dos dinucleotídeos AA, AT, CA, CC, CG, GA, GC, GG, TG, TT, obtidos através do uso da análise de componentes principais (*Principal Component Analyses* - PCA) [76]. As características usadas para treinar o modelo SVM da cana-de-açúcar diferem do modelo dos humanos e camundongos, pois não existe score de conservação para cana-de-açúcar e porque as combinações de nucleotídeos comuns nos lincRNAs desses organismos são diferentes dos presentes na cana-de-açúcar. Para treinar esse modelo, foi montado um conjunto de treinamento composto por 1000 lncRNAS de arroz e milho obtidos do CantataDB [85] e 1000 PCTs da cana, dos 105.378 transcritos que foram classificados como codificadores de proténas, e um conjunto de teste com o mesmo número de dados. A performance desse modelo é mostrada na Tabela 5.34 e a Tabela de contingência na Tabela 5.33.

Tabela 5.33: Tabela de contingência modelo SVM Cana

Amostras	Modelo reconheceu transcrito como lincRNA	Modelo não reconheceu transcrito como lincRNA
lncRNAs	813	284
PCTs	187	716

Tabela 5.34: Performance modelo SVM cana

Sensibilidade	81%
Especificidade	71%
Acurácia	76%

Pelo modelo SVM, dos 2.432 dos transcritos mapeados em região intergênica, 1.689 foram classificados como lncRNAS. Desses 2.432, 97 foram classificados como lncRNAs na etapa de anotação, e a intersecção entre os resultados do modelo SVM e da anotação evidencia 67 transcritos, os quais têm grandes chances de serem lincRNAs.

É importante ressaltar que o modelo SVM foi treinado com lncRNAs, pois não há dados suficientes de lincRNAs da cana-de-açúcar e organismos próximos evolutivamente para montar um modelo. Porém, a etapa de montagem do *pipeline* executado, asseguran a posição intergênica desses lncRNAs, como lincRNAs. A Tabela 5.35 mostra uma compilação dos resultados obtidos no estudo de caso da cana-de-açúcar.

Tabela 5.35: Dados do estudo de caso da cana-de-açúcar

Entrada Bruta	168.767
Entrada filtrada (sem codificadores de proteína)	63.389
Transcritos codificadores mapeados em região gênica do Sorgo	9.488
Transcritos não-codificadores mapeados no Sorgo	4.425
Transcritos não-codificadores mapeados em região intergênica (Tabela I.1 em anexo)	2.432
lincRNAs preditos pelo SVM (Tabela I.2 em anexo)	1.689
lincRNAs anotados como lncRNAs pelo BLAST (Tabela I.3 em anexo)	97
lincRNAs anotados como lncRNAs pelo BLAST e preditos pelo SVM (Tabela I.4 em anexo)	67

Capítulo 6

Conclusão

Neste trabalho, propusemos um método baseado em Máquinas de Vetores de Suporte (*Support Vector Machine* - SVM) para classificar lincRNAs. Foram desenvolvidos dois estudos de caso usando o método. O primeiro tratou de classificar lincRNAs em humanos e camundongos. Foram usadas características conforme proposto por Sun et al [83]. Essas características foram testadas, vários testes foram aplicados, que permitiram identificar a importância das ORFs (tamanho e proporção) na classificação de lincRNAs. O método baseado em SVM para identificar lincRNAs foi implementado e apresentou uma boa performance em transcritos de humanos e camundongos. Nesses testes a acurácia usando todas as características e os dois organismos simultaneamente foi de 91%. As acurácias de humanos foi de 99%, a de camundongos 89%, próximas da acurácia informada no artigo do iSeeRNA [83], de 95.4% e 94.2%, respectivamente. No entanto, quando testamos os nossos dados com o iSeeRNA, obtivemos acurácias de 50% e 53%, respectivamente, mostrando que possivelmente o iSeeRNA esteja com *overfitting*. Observamos que a diferença de performance pode ser explicada pelo fato de termos construído um banco de dados de lincRNAs de boa qualidade, com dados do HAVANA e do Ensembl, que difere dos dados utilizados para construir o modelo SVM do iSeeRNA.

O segundo estudo de caso, com transcritos da cana-de-açúcar, foi realizado por um *pipeline* específico para classificar lincRNAs nesse organismo, que usa o nosso modelo SVM. Esse *pipeline*, a partir de 168.767 transcritos de entrada, classificou 67 transcritos candidatos a lincRNAs, que serão testados experimentalmente pelo grupo do Prof. Paulo Cavalcanti da UFRJ.

6.1 Contribuições

Neste projeto, fizemos duas contribuições relevantes:

- Um modelo de classificação de lincRNAs baseado em SVM para humanos e camundongos que apresentou melhores resultados quando comparados com o iSeeRNA;
- Um modelo de classificação de lincRNAs da cana-de-açúcar, que usa um modelo SVM, em colaboração com um grupo de Biologia Molecular da UFRJ.

6.2 Trabalhos Futuros

Os nossos próximos passos são:

- Calcular valores de conservação de uma maneira mais genérica, que poderão ser utilizados para organismos que não possuem genoma sequenciado;
- Otimização de parâmetros C e γ da SVM para gerar modelos com maior acurácia;
- Escrever um artigo com o modelo SVM criado neste projeto;
- Disponibilizar uma ferramenta online com o método de classificação de lincRNAs baseado em SVM;
- Validar em laboratório os 67 lincRNAs da cana-de-açúcar preditos pelo nosso método;
- Refinar a criação do conjunto negativo, incluindo transcritos intergênicos que não sejam lincRNAs;
- Realizar testes de validação, com outros mamíferos, para testar a capacidade de generalização do modelo SVM de camundongos e humanos, e de outras plantas diferentes de cana-de-açúcar.

Referências

[1] Big data optimization at SAS. http://www.maths.ed.ac.uk/~prichtar/Optimization_and_Big_Data/slides/Polik.pdf. Accessed: 2016-01-06. x, 36

[2] Composição dos nucleotídeos. http://geneticavirtual.webnode.com.br. Accessed: 2016-01-26. 5, 7

[3] Cross-validation. http://stats.stackexchange.com/questions/1826/cross-validation-in-plain-english. Accessed: 2016-01-06. x, 40

[4] Detección de anomalías cardíacas con aprendizaje automático. http://samuelabad1991.blogspot.com.br/2014/02/analisis-con-maquinas-de-vectores.html. Accessed: 2016-01-24. 36

[5] Detecting structural elements of lincRNAs using RNAz. http://www.bioinf.uni-freiburg.de/Lehre/Theses/TP_KRAIBOOJ_slides.pdf. Accessed: 2015-10-24. 24

[6] Diana tools. http://diana.imis.athena-innovation.gr/DianaTools/index.php. Accessed: 2016-01-21. 29

[7] Ensembl. http://www.ensembl.org/index.html. Accessed: 2016-01-21. 2, 30, 42

[8] FastQC: A quality control application for FastQ data. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc. Accessed: 2016-01-23. 15

[9] Five questions for David Root: RNA Interference explained. https://www.broadinstitute.org/blog/five-questions-david-root-rna-interference-explained. Accessed: 2016-01-23. 8

[10] Genética Molecular. http://www.ufv.br/dbg/genetica/cap10.htm. Accessed: 2016-01-22. ix, 5, 6

[11] Grupo de Biooinformática Estrutural - UFRGS. http://www.ufrgs.br/bioinfo/th_gallery/capitulo-4/. Accessed: 2016-01-23. ix, 10

[12] Havana. http://vega.sanger.ac.uk/index.html. Accessed: 2016-01-21. 2, 30, 42, 58

[13] Hidden markov support vector machines. http://www.cs.helsinki.fi/group/smart/teaching/58308109/niissaloPrint.pdf. Accessed: 2016-01-26. 38

[14] IseeRNA. http://137.189.133.71/iSeeRNA/. Accessed: 2015-11-15. x, 28

[15] It's an snoRNA world. https://biochem.ncsu.edu/faculty/maxwell/Research.htm. Accessed: 2015-11-15. ix, 21

[16] K-means. http://mines.humanoriented.com. Accessed: 2015-12-05. x, 31, 33

[17] lncRNADisease. http://www.cuilab.cn/lncrnadisease. Accessed: 2015-11-15. x, 2, 30

[18] Mining for groups using clustering algorithms. http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/mvoget/cluster/cluster.html. Accessed: 2016-01-24. 31, 32

[19] Mitochondrial diseases at the molecular level. http://http://www.lhsc.on.ca/Patients_Families_Visitors/Genetics/Inherited_Metabolic/Mitochondria/DiseasesattheMolecularLevel.htm. Accessed: 2015-11-15. ix, 9

[20] Montagem. http://compbio.davidson.edu/phast/. Accessed: 2016-01-26. 17

[21] Rfam. http://rfam.xfam.org/. Accessed: 2015-11-15. x, 29

[22] RNA. http://biology.about.com/od/molecularbiology/ss/rna.htm. Accessed: 2016-01-23. ix, 8

[23] RNA. https://en.wikipedia.org/wiki/5S_ribosomal_RNA. Accessed: 2016-01-26. ix, 20, 21

[24] Sintese proteica. https://djalmasantos.wordpress.com/2014/07/16/sintese-proteica/. Accessed: 2016-01-26. ix, 11

[25] Splicing. https://pt.wikipedia.org/wiki/Splicing. Accessed: 2015-11-15. ix, 11

[26] SVM- Support Vector Machines. https://www.dtreg.com/solution/view/20. Accessed: 2016-01-06. x, 37

[27] The tRNA cloverleaf genera. https://commons.wikimedia.org/wiki/File:The_tRNA_cloverleaf_general.svg. Accessed: 2016-01-23. ix, 20

[28] UCSC genome bioinformatics. https://genome.ucsc.edu/. Accessed: 2016-01-26. ix, 7, 42, 43

[29] The warak warak method. https://biologywarakwarak.wordpress.com. Accessed: 2015-11-15. 12

[30] Why are Support Vectors Machines called so? https://onionesquereality.wordpress.com/2009/03/22/why-are-support-vectors-machines-called-so/. Accessed: 2016-01-26. 36

[31] S. Altschul, W. Gish, W. Miller, E. Myers, e D. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990. 17, 18

[32] P. Alvarez. Pipelines para transcritomas obtidos por sequenciadores de alto desempenho. *Monografia de Graduação. Departamento de Ciência da Computação. Universidade de Brasília*, 2009. 1

[33] R. Arrial, R. Togawa, e M. Brígido. Outlining a Strategy for Screening Non-coding RNAs on a Transcriptome Through Support Vector Machines. 4643:149–152, 2007. 10.1007/978-3-540-73731-5_14. 24

[34] T. Attwood, A. Gisel, E. Bongcam-Rudloff, e N. Eriksson. *Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective*. INTECH Open Access Publisher, 2011. 13

[35] S. Bennett. Solexa ltd. *Pharmacogenomics*, 5(4):433–438, 2004. 2, 13

[36] A. Bernal, U. Ear, e N. Kyrpides. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Research*, 29(1):126–127, 2001. 2

[37] M. Bottino, S. Rosario, C. Grativol, F. Thiebaut, C. Rojas, L. Farrineli, A. Hemerly, e P. Ferreira. High-throughput sequencing of small rna transcriptome reveals salt stress regulated micrornas in sugarcane. *PloS one*, 8(3):e59423, 2013. 2

[38] M. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, e J. Rinn. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development*, 25(18):1915–1927, 2011. 42

[39] M. Carvalho, D. Silva, et al. Sequenciamento de dna de nova geração e suas aplicações na genômica de plantas. *Ciência Rural*, 40(3):735–744, 2010. ix, 13, 14

[40] O. Chapelle, B. Schölkopf, e A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006. 35

[41] V. Chiang, R. Puumala, H. Takeuchi, e R. Eckert. Comparison of softwood and hardwood kraft pulping. *Tappi journal*, 71(9):173–176, 1988. 2, 45

[42] P. Clote e R. Backofen. *Computational molecular biology: an introduction a self contained approach to bioinformatics*. Chichester Wiley, 2000. 9

[43] J. Duvick, A. Fu, U. Muppirala, M. Sabharwal, M. Wilkerson, C. Lawrence, C. Lushbough, e V. Brendel. Plantgdb: a resource for comparative plant genomics. *Nucleic Acids Research*, 36(suppl 1):D959–D965, 2008. 45, 46, 59

[44] S. Eddy. Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929, 2001. 1, 19

[45] A. Fachel, A. Tahira, S. Vilella-Arias, V. Maracaja-Coutinho, E. Gimba, G. Vignal, F. Campos, E. Reis, e S. Verjovski-Almeida. Expression analysis and in silico characterization of intronic long noncoding rnas in renal cell carcinoma: emerging functional associations. *Mol Cancer*, 12(140):10–1186, 2013. 24, 27

[46] M. Hashemitabar, M. Kolahi, M. Tabandeh, P. Jonoubi, e A. Majd. cDNA cloning, phylogenic analysis and gene expression pattern of phenylalanine ammonia-lyase in sugarcane (saccharum officinarum l.). *Brazilian Archives of Biology and Technology*, 57(4):456–465, 2014. 45

[47] D. Haussler. Convolution kernels on discrete structures. Technical report, Citeseer, 1999. 38

[48] S. Haykin. A comprehensive foundation. *Neural Networks*, 2(2004), 2004. 36

[49] J.-H. He, Z.-P. Han, e Y.-G. Li. Association between long non-coding RNA and human rare diseases (review). *Biomedical reports*, 2(1):19–23, 2014. 23

[50] S. Hoffmann, C. Otto, S. Kurtz, C. Sharma, P. Khaitovich, Jörg V., P. Stadler, e J. Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9):e1000502, 2009. 46

[51] M. Huarte e J. Rinn. Large non-coding RNAs: missing links in cancer? *Human Molecular Genetics*, 19(R2):R152–R161, 2010. 23

[52] M. Idogawa, T. Ohashi, Y. Sasaki, R. Maruyama, L. Kashima, H. Suzuki, e T. Tokino. Identification and analysis of large intergenic non-coding RNAs regulated by p53 family members through a genome-wide analysis of p53 binding sites. *Human Molecular Genetics*, page ddt673, 2014. 24

[53] I. Ingelbrecht, J. Irvine, e T. Mirkov. Posttranscriptional gene silencing in transgenic sugarcane. dissection of homology-dependent virus resistance in a monocot that has a complex polyploid genome. *Plant physiology*, 119(4):1187–1198, 1999. 2

[54] V. Jakkula. Tutorial on Support Vector Machine (SVM). *School of EECS, Washington State University*, 2006. 37

[55] F. Jossinet, T. Ludwig, e E. Westhof. RNA structure: bioinformatics analysis. Science Direct. 2007. 19

[56] J. Jurka, V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, e J. Walichiewicz. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110(1-4):462–467, 2005. 45, 59

[57] A. Keniry, D. Oxley, P. Monnier, M. Kyba, L. Dandolo, G. Smits, e W. Reik. The h19 lincRNA is a developmental reservoir of mir-675 that suppresses growth and igf1r. *Nature cell biology*, 14(7):659–665, 2012. 23

[58] L. Kong, Y. Zhang, Z.-Q. Ye, X.-Q. Liu, S.-Q. Zhao, L. Wei, e G. Gao. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, 35:W345–9, Jul 2007. 24

[59] R. Leng et al. *Application of biotechnology to nutrition of animals in developing countries.* FAO, 1991. 2, 45

[60] J. Li, M. Zhang, G. An, e Q. Ma. Lncrna tug1 acts as a tumor suppressor in human glioma by promoting cell apoptosis. *Experimental Biology and Medicine*, page 1535370215622708, 2016. 24

[61] J. Liu, J. Gough, e B. Rost. Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genet*, 2(4):e29, Apr 2006. 24

[62] A. Machado-Lima, H. Del Portillo, e A. Durham. Computational methods in non-coding RNA research. *Journal of Mathematical Biology*, 56(1-2):15–49, 2008. 7, 19

[63] B. Maidak, J. Cole, T. Lilburn, C. Parker Jr, P. Saxman, R. Farris, G. Garrity, G. Olsen, T. Schmidt, e J. Tiedje. The rdp-ii (ribosomal database project). *Nucleic Acids Research*, 29(1):173–174, 2001. 19

[64] J. McPherson, M. Marra, L. Hillier, R. Waterston, A. Chinwalla, J. Wallis, M. Sekhon, K. Wylie, E. Mardis, R. Wilson, et al. A physical map of the human genome. *Nature*, 409(6822):934–941, 2001. 13

[65] T. Mercer, M. Dinger, e J. Mattick. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*, 10(3):155–159, 2009. 2, 22

[66] T. Mitchell. *Machine Learning*. McGraw-Hill, New York, 2 edition, 1997. 35

[67] U. Ørom e R. Shiekhattar. Noncoding RNAs and enhancers: complications of a long-distance relationship. *Trends in Genetics*, 27(10):433–439, 2011. 1, 22

[68] C. Ponting, P. Oliver, e W. Reik. Evolution and functions of long noncoding RNAs. *Cell*, 136(4):629–641, February 2009. 1, 19, 22

[69] G. Porreca. Genome sequencing on nanoballs. *Nature biotechnology*, 28(1):43–44, 2010. 2

[70] K. Pruitt, T. Tatusova, e D. Maglott. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(suppl 1):D501–D504, 2005. 42

[71] P. Refaeilzadeh, L. Tang, e H. Liu. Cross-validation. In *Encyclopedia of Database Systems*, pages 532–538. Springer, 2009. 39

[72] S. Russell e P. Norvig. *AI a modern approach*, volume 3. Pearson, 2010. 31, 35

[73] L. Sabin, M. Delás, e G. Hannon. Dogma derailed: The many influences of RNA on the genome. *Molecular Cell*, 49(5):783–794, 2013. 2

[74] M. Sauvageau, L. Goff, S. Lodato, B. Bonev, A. Groff, C. Gerhardinger, D. Sanchez-Gomez, E. Hacisuleyman, E. Li, M. Spence, et al. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife*, 2:e01749, 2013. 23

[75] R. Schmieder e R. Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, 2011. 15

[76] H. Schneider. Identificação de RNA não-codificador utilizando SVM. Qualificação para o doutorado. em preparação. Departamento de Ciência da Computação. Universidade de Brasília. 2015. 22, 60

[77] G. Selman-Housein, M. Lopez, D. Hernandez, L. Civardi, F. Miranda, J. Rigau, e P. Puigdomenech. Molecular cloning of cdnas coding for three sugarcane enzymes involved in lignification. *Plant Science*, 143(2):163–171, 1999. 45

[78] J. Setubal, J. e Meidanis. *Introduction to Computational Molecular Biology*. PWS Pub., 1997. 1, 4, 10, 12

[79] P. Shuai, D. Liang, S. Tang, Z. Zhang, C.-Y. Ye, Y. Su, X. Xia, e W. Yin. Genome-wide identification and functional prediction of novel and drought-responsive lincRNAs in populus trichocarpa. *Journal of Experimental Botany*, page eru256, 2014. 23

[80] A. Siepel, G. Bejerano, J. Pedersen, A. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. Hillier, S. Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005. 42

[81] T. Silva. SOM-Portrait: um método para identificar RNAs não codificadores utilizando Mapas Auto Organizáveis. Monografia de Graduação. Departamento de Ciência da Computação. Universidade de Brasília. 2009. 5, 7, 19

[82] T. Silva. Identificação de RNA não-codificador utilizando redes neurais artificiais de treinamento não supervisionado. Monografia de conclusão de mestrado. Departamento de Ciência da Computação. Universidade de Brasília. 2012. 7, 19

[83] K. Sun, X. Chen, P. Jiang, X. Song, H. Wang, e H. Sun. ISeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics*, 14 Suppl 2:S7, 2013. x, 2, 24, 27, 42, 43, 57, 59, 62

[84] R. Sutton e A. Barto. Reinforcement learning: an introduction. The MIT Press. *Cambridge, MA*, 1998. x, 34

[85] M. Szcześniak, W. Rosikiewicz, e I. Makałowska. Cantatadb: A collection of plant long non-coding RNAs. *Plant and Cell Physiology*, 57(1):e8–e8, 2016. 2, 45, 46, 59, 60, 73

[86] F. Thiebaut, C. Grativol, M. Tanurdzic, M. Carnavale-Bottino, T. Vieira, M. Motta, C. Rojas, R. Vincentini, A. Chabregas, S. Hemerly, et al. Differential srna regulation in leaves and roots of sugarcane under water depletion. *PloS One*, 9(4):e93822, 2014. 2

[87] F. Thiebaut, C. Rojas, K. Almeida, C. Grativol, G. Domiciano, C. Lamb, J. Engler, A. Hemerly, e P. Ferreira. Regulation of mir319 during cold stress in sugarcane. *Plant, cell & environment*, 35(3):502–512, 2012. 2

[88] J. Thompson e K. Steinmann. Single molecule sequencing with a heliscope genetic analysis system. *Current Protocols in Molecular Biology*, pages 7–10, 2010. 2

[89] H. Timmers e L. Tora. The spectacular landscape of chromatin and ncRNAs under the tico sunlight. *EMBO reports*, 11(3):147–149, 2010. 2

[90] I. Ulitsky e D. Bartel. lincRNAs: genomics, evolution, and mechanisms. *Cell*, 154(1):26–46, 2013. ix, x, 1, 23, 24, 25, 26

[91] I. Ulitsky, A. Shkumatava, C. Jan, H. Sive, e D. Bartel. Conserved function of lincrnas in vertebrate embryonic development despite rapid sequence evolution. *Cell*, 147(7):1537–1550, 2011. 24

[92] J. Venter, M. Adams, E. Myers, P. Li, R. Mural, G. Sutton, H. Smith, M. Yandell, C. Evans, R. Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001. 13

[93] Z. Wang, M. Gerstein, e M. Snyder. RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009. 13

[94] J. Watson e F. Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953. 1

[95] J. Wu, D. Delneri, R. O'Keefe, et al. Non-coding RNAs in saccharomyces cerevisiae: what is the function? *Biochemical Society Transactions*, 40(4):907, 2012. 1, 19

[96] X. Zhang, S. Weissman, e P. Newburger. Long intergenic non-coding rna hotairm1 regulates cell cycle progression during myeloid maturation in nb4 human promyelocytic leukemia cells. *RNA Biology*, 11(6):777–787, 2014. 24

Anexo I

Informações detalhadas dos lincRNAs da cana-de-açúcar

Neste anexo, apresentamos parte das informações obtidas no estudo de caso da cana-de-açúcar.

A Tabela I.1 mostra parte dos 2.432 transcritos mapeados em regiões intergênicas.

Tabela I.1: Transcritos não-codificadores mapeados em regiões intergênicas

id	cromossomo	posicao	tamanho	long no blast	predito pelo svm	gene a esquerda	gene a direita
Locus_3_Transcript_18_Confidence_0.095_Length_278	3	10008226	278	N	S	9975249	10015850
Locus_5_Transcript_2503585_Confidence_1.000_Length_1088	3	69034186	1088	N	S	68997952	69036333
Locus_5_Transcript_4193585_Confidence_1.000_Length_212	1	32912529	212	N	S	30563752	43093433
Locus_5_Transcript_7083585_Confidence_1.000_Length_237	4	60223431	237	N	N	60175745	60223620
Locus_5_Transcript_7153585_Confidence_1.000_Length_237	4	60223431	237	N	N	60175745	60223620
Locus_5_Transcript_7213585_Confidence_1.000_Length_312	10	56261706	312	N	S	56261023	56261966
Locus_5_Transcript_7323585_Confidence_1.000_Length_312	10	56261706	312	N	S	56261023	56261966
Locus_5_Transcript_7633585_Confidence_1.000_Length_237	4	60223431	237	N	N	60175745	60223620
Locus_5_Transcript_7703585_Confidence_1.000_Length_237	4	60223431	237	N	N	60175745	60223620
Locus_5_Transcript_10073585_Confidence_1.000_Length_205	4	62377926	205	N	N	62361285	62378133
Locus_5_Transcript_10083585_Confidence_1.000_Length_205	4	62377926	205	N	N	62361285	62378133
Locus_5_Transcript_10293585_Confidence_1.000_Length_205	4	62377926	205	N	S	62361285	62378133
Locus_5_Transcript_10303585_Confidence_1.000_Length_205	4	62377926	205	N	S	62361285	62378133
Locus_5_Transcript_11193585_Confidence_1.000_Length_211	1	5965295	211	N	N	5936642	5989645
Locus_5_Transcript_11523585_Confidence_1.000_Length_284	2	67426781	284	N	S	67407524	67479152
Locus_5_Transcript_11603585_Confidence_1.000_Length_284	2	67426781	284	N	S	67407524	67479152
Locus_5_Transcript_11683585_Confidence_1.000_Length_201	5	6526699	201	N	N	6497103	6790106
Locus_5_Transcript_11703585_Confidence_1.000_Length_201	5	6526699	201	N	N	6497103	6790106
Locus_5_Transcript_11863585_Confidence_1.000_Length_201	5	6526699	201	N	S	6497103	6790106
Locus_5_Transcript_11913585_Confidence_1.000_Length_201	5	6526699	201	N	S	6497103	6790106
Locus_5_Transcript_12553585_Confidence_1.000_Length_246	3	66243735	246	N	S	66074904	66277077
Locus_5_Transcript_12723585_Confidence_1.000_Length_224	2	69482934	224	S	S	69425409	69483162
Locus_5_Transcript_13033585_Confidence_1.000_Length_217	4	41856064	217	S	N	41689525	41856233
Locus_5_Transcript_13293585_Confidence_1.000_Length_217	4	41856064	217	S	N	41689525	41856233
Locus_5_Transcript_14713585_Confidence_1.000_Length_492	10	3185331	492	N	N	3163535	3209006
Locus_5_Transcript_14733585_Confidence_1.000_Length_233	10	3186135	233	N	N	3163535	3209006
Locus_5_Transcript_16623585_Confidence_1.000_Length_209	4	53554780	209	N	S	53530232	53607756
Locus_5_Transcript_16763585_Confidence_1.000_Length_209	4	53554780	209	N	S	53530232	53607756
Locus_5_Transcript_17103585_Confidence_1.000_Length_212	1	32912529	212	N	N	30563752	43093433
Locus_5_Transcript_17393585_Confidence_1.000_Length_374	3	65507359	374	N	S	65498526	65523533
Locus_5_Transcript_17613585_Confidence_1.000_Length_374	3	65507359	374	N	S	65498526	65523533

A Tabela I.2 mostra parte dos 1.689 lincRNAs preditos pelo modelo SVM desenvolvido neste trabalho.

Tabela I.2: LincRNAs preditos pelo modelo SVM

id	cromossomo	posicao	tamanho	long no blast	predito pelo svm	gene a esquerda	gene a direita
Locus_3_Transcript_18_Confidence_0.095_Length_278	3	10008226	278	N	S	9975249	10015850
Locus_5_Transcript_2503585_Confidence_1.000_Length_1088	3	69034186	1088	N	S	68997952	69036333
Locus_5_Transcript_4193585_Confidence_1.000_Length_212	1	32912529	212	N	S	30563752	43093433
Locus_5_Transcript_7213585_Confidence_1.000_Length_312	10	56261706	312	N	S	56261023	56261966
Locus_5_Transcript_7323585_Confidence_1.000_Length_312	10	56261706	312	N	S	56261023	56261966
Locus_5_Transcript_10293585_Confidence_1.000_Length_205	4	62377926	205	N	S	62361285	62378133
Locus_5_Transcript_10303585_Confidence_1.000_Length_205	4	62377926	205	N	S	62361285	62378133
Locus_5_Transcript_11523585_Confidence_1.000_Length_284	2	67426781	284	N	S	67407524	67479152
Locus_5_Transcript_11603585_Confidence_1.000_Length_284	2	67426781	284	N	S	67407524	67479152
Locus_5_Transcript_11863585_Confidence_1.000_Length_201	5	6526699	201	N	S	6497103	6790106
Locus_5_Transcript_11913585_Confidence_1.000_Length_201	5	6526699	201	N	S	6497103	6790106
Locus_5_Transcript_12553585_Confidence_1.000_Length_246	3	66243735	246	N	S	66074904	66277077
Locus_5_Transcript_12723585_Confidence_1.000_Length_224	2	69482934	224	S	S	69425409	69483162
Locus_5_Transcript_16623585_Confidence_1.000_Length_209	4	53554780	209	N	S	53530232	53607756
Locus_5_Transcript_16763585_Confidence_1.000_Length_209	4	53554780	209	N	S	53530232	53607756
Locus_5_Transcript_17393585_Confidence_1.000_Length_374	3	65507359	374	N	S	65498526	65523533
Locus_5_Transcript_17613585_Confidence_1.000_Length_374	3	65507359	374	N	S	65498526	65523533
Locus_5_Transcript_18173585_Confidence_1.000_Length_324	6	56083352	324	N	S	56083199	56236348
Locus_5_Transcript_19253585_Confidence_1.000_Length_233	10	3186135	233	N	S	3163535	3209006
Locus_5_Transcript_20953585_Confidence_1.000_Length_254	2	54591825	254	N	S	54577027	54592175
Locus_5_Transcript_21913585_Confidence_1.000_Length_254	2	54591825	254	N	S	54577027	54592175
Locus_5_Transcript_22063585_Confidence_1.000_Length_246	3	66243735	246	N	S	66074904	66277077
Locus_5_Transcript_22253585_Confidence_1.000_Length_224	2	69482934	224	S	S	69425409	69483162
Locus_5_Transcript_22783585_Confidence_1.000_Length_397	1	23154645	397	N	S	23117644	23314890
Locus_5_Transcript_22803585_Confidence_1.000_Length_759	1	23157070	759	N	S	23117644	23314890
Locus_5_Transcript_22823585_Confidence_1.000_Length_644	1	23155905	644	N	S	23117644	23314890
Locus_5_Transcript_23613585_Confidence_1.000_Length_397	1	23154645	397	N	S	23117644	23314890
Locus_5_Transcript_23633585_Confidence_1.000_Length_644	1	23155905	644	N	S	23117644	23314890
Locus_5_Transcript_23643585_Confidence_1.000_Length_759	1	23157070	759	N	S	23117644	23314890
Locus_5_Transcript_24063585_Confidence_1.000_Length_211	1	5965295	211	N	S	5936642	5989645
Locus_5_Transcript_24753585_Confidence_1.000_Length_461	6	46406774	461	N	S	46227790	46426398
Locus_5_Transcript_27003585_Confidence_1.000_Length_461	6	46406774	461	N	S	46227790	46426398
Locus_5_Transcript_29463585_Confidence_1.000_Length_270	1	63015878	270	N	S	62823097	63048163
Locus_5_Transcript_30123585_Confidence_1.000_Length_270	1	63015878	270	N	S	62823097	63048163
Locus_5_Transcript_30293585_Confidence_1.000_Length_416	7	1102840	416	N	S	1027858	1104989
Locus_5_Transcript_30313585_Confidence_1.000_Length_265	7	1103328	265	N	S	1027858	1104989
Locus_5_Transcript_30343585_Confidence_1.000_Length_507	7	1103654	507	N	S	1027858	1104989
Locus_5_Transcript_30373585_Confidence_1.000_Length_201	7	1104839	201	N	S	1027858	1104989
Locus_5_Transcript_30853585_Confidence_1.000_Length_520	2	21631509	520	N	S	21533969	21631988
Locus_5_Transcript_31123585_Confidence_1.000_Length_416	7	1102840	416	N	S	1027858	1104989
Locus_5_Transcript_31143585_Confidence_1.000_Length_265	7	1103328	265	N	S	1027858	1104989
Locus_5_Transcript_31163585_Confidence_1.000_Length_507	7	1103654	507	N	S	1027858	1104989
Locus_5_Transcript_31233585_Confidence_1.000_Length_201	7	1104839	201	N	S	1027858	1104989
Locus_5_Transcript_31633585_Confidence_1.000_Length_520	2	21631509	520	N	S	21533969	21631988
Locus_5_Transcript_32243585_Confidence_1.000_Length_1088	3	69034186	1088	N	S	68997952	69036333
Locus_5_Transcript_34493585_Confidence_1.000_Length_201	4	59972280	201	N	S	59972171	60006077
Locus_5_Transcript_35483585_Confidence_1.000_Length_324	6	56083352	324	N	S	56083199	56236348
Locus_78_Transcript_40633_Confidence_1.000_Length_630	1	70001340	630	N	S	69935166	70002746
Locus_78_Transcript_73633_Confidence_1.000_Length_263	9	58840125	263	N	S	58793162	58889249

A Tabela I.3 mostra parte dos 97 lincRNAs anotados como lncRNAs pelo BLAST com o banco de dados CantataDB [85].

Tabela I.3: LincRNAs anotados como lncRNAs pelo BLAST

id	cromossomo	posicao	tamanho	long no blast	predito pelo svm	gene a esquerda	gene a direita
Locus_5_Transcript_12723585_Confidence_1.000_Length_224	2	69482934	224	S	S	69425409	69483162
Locus_5_Transcript_13033585_Confidence_1.000_Length_217	4	41856064	217	S	N	41689525	41856233
Locus_5_Transcript_13293585_Confidence_1.000_Length_217	4	41856064	217	S	N	41689525	41856233
Locus_5_Transcript_22253585_Confidence_1.000_Length_224	2	69482934	224	S	S	69425409	69483162
Locus_78_Transcript_226633_Confidence_1.000_Length_300	1	66196799	300	S	N	66194085	66198451
Locus_78_Transcript_573633_Confidence_1.000_Length_300	1	66196799	300	S	S	66194085	66198451
Locus_87_Transcript_11_Confidence_1.000_Length_244	1	7679728	244	S	N	7664553	7858366
Locus_223_Transcript_111_Confidence_1.000_Length_210	3	38367175	210	S	N	28884580	38393488
Locus_400_Transcript_57_Confidence_0.471_Length_1847	10	58753541	1847	S	S	58692467	58829738
Locus_725_Transcript_12_Confidence_0.667_Length_490	3	60742312	490	S	S	60626774	60752312
Locus_725_Transcript_22_Confidence_0.667_Length_494	3	60742312	494	S	S	60626774	60752312
Locus_857_Transcript_111_Confidence_1.000_Length_310	2	69071348	310	S	S	69031469	69098899
Locus_892_Transcript_1224_Confidence_1.000_Length_293	2	68496737	293	S	S	68446795	68533640
Locus_1520_Transcript_1132_Confidence_1.000_Length_224	3	263726	224	S	N	263147	308508
Locus_1520_Transcript_1532_Confidence_1.000_Length_337	3	263168	337	S	S	263147	308508
Locus_1520_Transcript_2032_Confidence_1.000_Length_224	3	263726	224	S	N	263147	308508
Locus_4256_Transcript_45_Confidence_0.308_Length_949	3	28861325	949	S	S	26989479	28883488
Locus_4334_Transcript_69_Confidence_0.160_Length_373	4	64747906	373	S	S	64516494	64775108
Locus_5872_Transcript_46_Confidence_0.417_Length_874	6	39392390	874	S	S	39369644	39397394
Locus_11102_Transcript_46_Confidence_0.385_Length_910	1	48167525	910	S	S	47446334	48233734
Locus_11481_Transcript_57_Confidence_0.125_Length_382	9	54606830	382	S	N	54589559	54706133
Locus_12274_Transcript_33_Confidence_0.714_Length_259	4	2615104	259	S	S	2562877	2665436
Locus_14392_Transcript_1526_Confidence_1.000_Length_291	5	56982166	291	S	S	56980459	56984395
Locus_14392_Transcript_1626_Confidence_1.000_Length_291	5	56982166	291	S	S	56980459	56984395
Locus_14553_Transcript_313_Confidence_1.000_Length_237	7	5472556	237	S	S	5390360	5483626
Locus_14637_Transcript_26_Confidence_0.294_Length_401	9	81096	401	S	S	NULL	81471
Locus_14816_Transcript_36_Confidence_0.353_Length_384	9	1609405	384	S	S	1482649	1611112
Locus_16281_Transcript_11_Confidence_1.000_Length_409	4	5931106	409	S	N	5887954	5931492
Locus_18849_Transcript_13_Confidence_0.571_Length_665	2	57728506	665	S	S	57664087	57846539
Locus_18849_Transcript_23_Confidence_0.714_Length_666	2	57728506	666	S	S	57664087	57846539
Locus_18849_Transcript_33_Confidence_0.714_Length_661	2	57728506	661	S	S	57664087	57846539
Locus_21131_Transcript_12_Confidence_0.667_Length_329	1	14190661	329	S	S	14013853	14250582
Locus_21405_Transcript_312_Confidence_0.212_Length_234	4	17410403	234	S	N	17399105	17763568
Locus_21405_Transcript_1012_Confidence_0.333_Length_231	4	17410403	231	S	N	17399105	17763568
Locus_24127_Transcript_11_Confidence_1.000_Length_217	3	38366810	217	S	N	28884580	38393488
Locus_28504_Transcript_16_Confidence_0.474_Length_749	7	10090786	749	S	S	9657440	10181006
Locus_28504_Transcript_26_Confidence_0.316_Length_505	7	10091018	505	S	S	9657440	10181006
Locus_28504_Transcript_46_Confidence_0.421_Length_565	7	10090963	565	S	S	9657440	10181006
Locus_31422_Transcript_11_Confidence_1.000_Length_257	2	41908394	257	S	N	41614076	45011166
Locus_36760_Transcript_11_Confidence_1.000_Length_257	9	57612282	257	S	N	57561522	57616818
Locus_36788_Transcript_11_Confidence_1.000_Length_344	2	65799814	344	S	S	65762431	65815815
Locus_37761_Transcript_11_Confidence_1.000_Length_301	8	44858123	301	S	S	44796827	44863499
Locus_39389_Transcript_11_Confidence_1.000_Length_238	3	60394555	238	S	S	60346653	60416947
Locus_40443_Transcript_11_Confidence_1.000_Length_321	6	60038634	321	S	S	60017927	60138003
Locus_42282_Transcript_48_Confidence_0.200_Length_493	9	58812937	493	S	N	58793162	58889249
Locus_42575_Transcript_11_Confidence_1.000_Length_427	1	57592388	427	S	N	57556877	57594301
Locus_43699_Transcript_12_Confidence_0.667_Length_287	6	59494226	287	S	S	59372757	59528894
Locus_43699_Transcript_22_Confidence_0.889_Length_287	6	59494226	287	S	S	59372757	59528894
Locus_44244_Transcript_11_Confidence_1.000_Length_232	6	60669538	232	S	S	60582988	60671012
Locus_46732_Transcript_11_Confidence_1.000_Length_294	3	8176104	294	S	N	8110185	8250250
Locus_49638_Transcript_11_Confidence_1.000_Length_215	1	55636148	215	S	N	55328105	55685998
Locus_50255_Transcript_11_Confidence_1.000_Length_339	6	49172865	339	S	S	49139928	49249627
Locus_52242_Transcript_11_Confidence_1.000_Length_255	1	57370131	255	S	S	57333036	57443445
Locus_52859_Transcript_22_Confidence_0.500_Length_551	6	60045586	551	S	S	60017927	60138003
Locus_53220_Transcript_26_Confidence_0.500_Length_330	9	58112509	330	S	S	58108544	58181612
Locus_54955_Transcript_12_Confidence_0.833_Length_312	4	3323773	312	S	S	3321824	3383740
Locus_54955_Transcript_22_Confidence_0.667_Length_305	4	3323773	305	S	S	3321824	3383740
Locus_55822_Transcript_11_Confidence_1.000_Length_200	3	63803445	200	S	N	63802988	63810784
Locus_55826_Transcript_11_Confidence_1.000_Length_200	4	880309	200	S	S	835652	981672
Locus_58190_Transcript_11_Confidence_1.000_Length_508	1	71295162	508	S	S	71259385	71295765
Locus_59686_Transcript_11_Confidence_1.000_Length_468	3	57810837	468	S	S	57810616	57860000
Locus_59777_Transcript_11_Confidence_1.000_Length_462	2	306228	462	S	S	196763	373026
Locus_61543_Transcript_12_Confidence_0.667_Length_878	5	8921233	878	S	N	8862156	9157041
Locus_61543_Transcript_22_Confidence_0.667_Length_878	5	8921233	878	S	N	8862156	9157041
Locus_61934_Transcript_11_Confidence_1.000_Length_370	4	64747583	370	S	S	64516494	64775108
Locus_62723_Transcript_11_Confidence_1.000_Length_1455	3	12299657	1455	S	S	12296148	12301246
Locus_64490_Transcript_11_Confidence_1.000_Length_262	10	43958550	262	S	S	43505954	43970672
Locus_65513_Transcript_11_Confidence_1.000_Length_504	9	53078545	504	S	N	53077866	53125093
Locus_66397_Transcript_11_Confidence_1.000_Length_739	1	47668315	739	S	S	47446334	48233734
Locus_67674_Transcript_11_Confidence_1.000_Length_552	4	49644603	552	S	S	49596847	49788988
Locus_69038_Transcript_11_Confidence_1.000_Length_204	1	64054655	204	S	S	64001171	64056769
Locus_70267_Transcript_11_Confidence_1.000_Length_375	3	2254615	375	S	S	2152142	2464782
Locus_70463_Transcript_11_Confidence_1.000_Length_470	2	71947099	470	S	S	71945345	72010217
Locus_71706_Transcript_11_Confidence_1.000_Length_231	10	52909219	231	S	N	52625103	52910095
Locus_72924_Transcript_11_Confidence_1.000_Length_449	1	53808301	449	S	S	53807837	53833141
Locus_72937_Transcript_11_Confidence_1.000_Length_250	6	61633986	250	S	S	61631391	61654729
Locus_73116_Transcript_11_Confidence_1.000_Length_267	10	53521026	267	S	S	53353901	53546143
Locus_73167_Transcript_11_Confidence_1.000_Length_507	3	9656122	507	S	S	9582282	9693517

A Tabela I.4 mostra os 67 lincRNAs anotados pelo BLAST e preditos pelo SVM.

Tabela I.4: Predição de lincRNAs da cana-de-açúcar

id	cromossomo	posicao	tamanho	long no blast	predito pelo svm	gene a esquerda	gene a direita
Locus_5_Transcript_12723585_Confidence_1.000_Length_224	2	69482934	224	S	S	69425409	69483162
Locus_5_Transcript_22253585_Confidence_1.000_Length_224	2	69482934	224	S	S	69425409	69483162
Locus_78_Transcript_573633_Confidence_1.000_Length_300	1	66196799	300	S	S	66194085	66198451
Locus_400_Transcript_57_Confidence_0.471_Length_1847	10	58753541	1847	S	S	58692467	58829738
Locus_725_Transcript_12_Confidence_0.667_Length_490	3	60742312	490	S	S	60626774	60752312
Locus_725_Transcript_22_Confidence_0.667_Length_494	3	60742312	494	S	S	60626774	60752312
Locus_857_Transcript_111_Confidence_1.000_Length_310	2	69071348	310	S	S	69031469	69098899
Locus_892_Transcript_1224_Confidence_1.000_Length_293	2	68496737	293	S	S	68446795	68533640
Locus_1520_Transcript_1532_Confidence_1.000_Length_337	3	263168	337	S	S	263147	308508
Locus_4256_Transcript_45_Confidence_0.308_Length_949	3	28861325	949	S	S	26989479	28883488
Locus_4334_Transcript_69_Confidence_0.160_Length_373	4	64747906	373	S	S	64516494	64775108
Locus_5872_Transcript_46_Confidence_0.417_Length_874	6	39392390	874	S	S	39369644	39397394
Locus_11102_Transcript_46_Confidence_0.385_Length_910	1	48167525	910	S	S	47446334	48233734
Locus_12274_Transcript_33_Confidence_0.714_Length_259	4	2615104	259	S	S	2562877	2665436
Locus_14392_Transcript_1526_Confidence_1.000_Length_291	5	56982166	291	S	S	56980459	56984395
Locus_14392_Transcript_1626_Confidence_1.000_Length_291	5	56982166	291	S	S	56980459	56984395
Locus_14553_Transcript_313_Confidence_1.000_Length_237	7	5472556	237	S	S	5390360	5483626
Locus_14637_Transcript_26_Confidence_0.294_Length_401	9	81096	401	S	S	NULL	81471
Locus_14816_Transcript_36_Confidence_0.353_Length_384	9	1609405	384	S	S	1482649	1611112
Locus_18849_Transcript_13_Confidence_0.571_Length_665	2	57728506	665	S	S	57664087	57846539
Locus_18849_Transcript_23_Confidence_0.714_Length_666	2	57728506	666	S	S	57664087	57846539
Locus_18849_Transcript_33_Confidence_0.714_Length_661	2	57728506	661	S	S	57664087	57846539
Locus_21131_Transcript_12_Confidence_0.667_Length_329	1	14190661	329	S	S	14013853	14250582
Locus_28504_Transcript_16_Confidence_0.474_Length_749	7	10090786	749	S	S	9657440	10181006
Locus_28504_Transcript_26_Confidence_0.316_Length_505	7	10091018	505	S	S	9657440	10181006
Locus_28504_Transcript_46_Confidence_0.421_Length_565	7	10090963	565	S	S	9657440	10181006
Locus_36788_Transcript_11_Confidence_1.000_Length_344	2	65799814	344	S	S	65762431	65815815
Locus_37761_Transcript_11_Confidence_1.000_Length_301	8	44858123	301	S	S	44796827	44863499
Locus_39389_Transcript_11_Confidence_1.000_Length_238	3	60394555	238	S	S	60346653	60416947
Locus_40443_Transcript_11_Confidence_1.000_Length_321	6	60038634	321	S	S	60017927	60138003
Locus_43699_Transcript_12_Confidence_0.667_Length_287	6	59494226	287	S	S	59372757	59528894
Locus_43699_Transcript_22_Confidence_0.889_Length_287	6	59494226	287	S	S	59372757	59528894
Locus_44244_Transcript_11_Confidence_1.000_Length_232	6	60669538	232	S	S	60582988	60671012
Locus_50255_Transcript_11_Confidence_1.000_Length_339	6	49172865	339	S	S	49139928	49249627
Locus_52242_Transcript_11_Confidence_1.000_Length_255	1	57370131	255	S	S	57333036	57443445
Locus_52859_Transcript_22_Confidence_0.500_Length_551	6	60045586	551	S	S	60017927	60138003
Locus_53220_Transcript_26_Confidence_0.500_Length_330	9	58112509	330	S	S	58108544	58181612
Locus_54955_Transcript_12_Confidence_0.833_Length_312	4	3323773	312	S	S	3321824	3383740
Locus_54955_Transcript_22_Confidence_0.667_Length_305	4	3323773	305	S	S	3321824	3383740
Locus_55826_Transcript_11_Confidence_1.000_Length_200	4	880309	200	S	S	835652	981672
Locus_58190_Transcript_11_Confidence_1.000_Length_508	1	71295162	508	S	S	71259385	71295765
Locus_59686_Transcript_11_Confidence_1.000_Length_468	3	57810837	468	S	S	57810616	57860000
Locus_59777_Transcript_11_Confidence_1.000_Length_462	2	306228	462	S	S	196763	373026
Locus_61934_Transcript_11_Confidence_1.000_Length_370	4	64747583	370	S	S	64516494	64775108
Locus_62723_Transcript_11_Confidence_1.000_Length_1455	3	12299657	1455	S	S	12296148	12301246
Locus_64490_Transcript_11_Confidence_1.000_Length_262	10	43958550	262	S	S	43505954	43970672
Locus_66397_Transcript_11_Confidence_1.000_Length_739	1	47668315	739	S	S	47446334	48233734
Locus_67674_Transcript_11_Confidence_1.000_Length_552	4	49644603	552	S	S	49596847	49788988
Locus_69038_Transcript_11_Confidence_1.000_Length_204	1	64054655	204	S	S	64001171	64056769
Locus_70267_Transcript_11_Confidence_1.000_Length_375	3	2254615	375	S	S	2152142	2464782
Locus_70463_Transcript_11_Confidence_1.000_Length_470	2	71947099	470	S	S	71945345	72010217
Locus_72924_Transcript_11_Confidence_1.000_Length_449	1	53808301	449	S	S	53807837	53833141
Locus_72937_Transcript_11_Confidence_1.000_Length_250	6	61633986	250	S	S	61631391	61654729
Locus_73116_Transcript_11_Confidence_1.000_Length_267	10	53521026	267	S	S	53353901	53546143
Locus_73167_Transcript_11_Confidence_1.000_Length_507	3	9656122	507	S	S	9582282	9693517
Locus_73509_Transcript_11_Confidence_1.000_Length_242	1	15296022	242	S	S	15203719	15380983
Locus_73538_Transcript_23_Confidence_0.750_Length_613	10	52909362	613	S	S	52625103	52910095
Locus_73538_Transcript_33_Confidence_0.750_Length_618	10	52909362	618	S	S	52625103	52910095
Locus_73840_Transcript_11_Confidence_1.000_Length_286	1	66198176	286	S	S	66194085	66198451
Locus_74602_Transcript_11_Confidence_1.000_Length_223	1	47668148	223	S	S	47446334	48233734
Locus_74807_Transcript_11_Confidence_1.000_Length_221	10	4567937	221	S	S	4537736	4644285
Locus_76244_Transcript_11_Confidence_1.000_Length_338	8	5555841	338	S	S	5468748	5900453
Locus_77183_Transcript_11_Confidence_1.000_Length_258	7	4784699	258	S	S	4747404	4842766
Locus_81085_Transcript_11_Confidence_1.000_Length_283	1	60484588	283	S	S	60429378	60491111
Locus_82615_Transcript_11_Confidence_1.000_Length_476	4	49644168	476	S	S	49596847	49788988
Locus_82741_Transcript_11_Confidence_1.000_Length_258	1	51852863	258	S	S	51749606	51889185
Locus_84316_Transcript_11_Confidence_1.000_Length_318	5	49763392	318	S	S	49761228	49764878