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

Distinguishing long non-coding RNAs from protein coding transcripts based on machine learning techniques

Hugo Wruck Schneider

Tese apresentada como requisito parcial para
conclusão do Doutorado em Informática

Orientadora
Profa. Dra. Maria Emília Machado Telles Walter

Brasília
2017

Ficha Catalográfica de Teses e Dissertações

Está página existe apenas para indicar onde a ficha catalográfica gerada para dissertações de mestrado e teses de doutorado defendidas na UnB. A Biblioteca Central é responsável pela ficha, mais informações nos sítios:

http://www.bce.unb.br
http://www.bce.unb.br/elaboracao-de-fichas-catalograficas-de-teses-e-dissertacoes

Esta página não deve ser inclusa na versão final do texto.

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

Distinguishing long non-coding RNAs from protein coding transcripts based on machine learning techniques

Hugo Wruck Schneider

Tese apresentada como requisito parcial para
conclusão do Doutorado em Informática

Profa. Dra. Maria Emília Machado Telles Walter (Orientadora)
CIC/UnB

Prof. Dr. Li Weigang
CIC/UnB

Prof. Dr. Marcelo de Macedo Brígido
IB/UnB

Prof. Dr. Nalvo Franco de Almeida Junior
FACOM/UFMS

Profa. Dra. Célia Ghedini Ralha
CIC/UnB (Suplente)

Prof. Dr. Bruno Luiggi Machiavello
Coordenador do Programa de Pós-graduação em Informática

Brasília, 28 de September de 2017

Dedicatória

Dedico esse trabalho a minha esposa, Aline, e ao meu filho, Miguel.

Agradecimentos

Agradeço a todos que colaboraram para que esse trabalho fosse concluído. Agradeço a minha orientadora professora Maria Emília pelos ensinamentos e todo apoio nesse trabalho. À Tainá e ao Professor Marcelo pelas ricas discussões biológicas. Ao Professor Peter que me ajudou muito apesar da distância. E por fim à família que soube ter paciência durante esse longo percurso.

Resumo

Dentre as análises que devem ser realizadas nos projetos de sequenciamento, um problema importante é a distinção entre transcritos codificadores de proteinas (PCTs) e RNAs não-codificadores longos (lncRNAs). Esse trabalho investiga potenciais características dos lncRNAs e propõe dois métodos para distinção dessas duas classes de transcritos (PCTs e lncRNAs). O primeiro método foi proposto com base em máquinas de vetores de suporte (SVM), enquanto o segundo utilizou técnicas de aprendizado semi-supervisionado. O método utilizando SVM obteve excelentes resultados, quando comparados a outras propostas existentes na literatura. Esse método foi treinado e testado com dados de humanos, camundongos e peixe-zebra, tendo atingido uma acurácia de $\approx 98\%$ com dados de humanos e camundongos, e de $\approx 96\%$ para os dados do peixe-zebra. Ainda, foram criados modelos utilizando várias espécies, que mostraram classificações melhores para outras espécies diferentes daquelas do treinamento, ou seja, mostraram boa capacidade de generalização. Para validar esse método, foram utilizados dados de ratos, porcos e drosófilas, além de dados de RNA-seq de humanos, gorilas e macacos. Essa validação atingiu uma acurácia de mais de 85%, em todos os casos. Por fim, esse método foi capaz de identificar duas sequências dentro do Swiss-Prot que puderam ser reanotadas. O método baseado em aprendizado semi-supervisionado foi treinado e testado com dados de humanos, camundongos, ornitorrincos, galinhas, gambás, orangotangos e rãs, tendo sido utilizadas cinco técnicas de aprendizado semi-supervisionado. A contribuição desse método foi que ele permitiu a redução do tamanho do conjunto de dados classificados, utilizados no treinamento. No melhor caso, somente 2 sequências bem anotadas foram usadas no treinamento, o que, comparado com outras ferramentas disponíveis na literatura, indica um ganho expressivo. A acurácia obtida pelo método nos melhores casos foram de $\approx 95\%$ para dados de humanos e camundongos, $\approx 90\%$ para dados de galinhas, gambás e orangutangos, e $\approx 80\%$ para dados de ornitorrincos e rãs. Dados de RNA-seq foram utilizados para teste, tendo sido obtida acurácia de mais de 95%. Esses dados foram utilizados para treinamento dos modelos de orangotango e de rã, que também apresentaram acurácias excelentes.

Palavras-chave: RNAs não-codificadores longos (RNAncl), Máquina de Vetores de Suporte (MVS), Análise de Componentes Principais (ACP), Aprendizagem Semi-Supervisionada

Abstract

Among the analyses that have to be performed in sequencing projects, an important problem to be addressed is the distinction of protein coding transcripts (PCTs) and long non-coding RNAs (lncRNA). This work investigates potential characteristics of the lncRNAs and proposes two methods for distinguishing these two classes of transcripts (PCTs and lncRNAs). The first methods was based on Support Vector Machine (SVM), while the second one used semi-supervised learning techniques. The SVM based method obtained excellent results when compared to other methods in the literature. This method was trained and tested with data from human, mouse and zebrafish, and reached accuracy of $\approx 98\%$ for human and mouse data, and $\approx 96\%$ for zebrafish data. Besides, models with multiple species were created, which improved the classification for species different from those used in the training phase, i.e., these models could also be used in the classification of species different from those that were used in the training phase. To validate this method, data from rat, pig and drosophila, and RNA-seq data from humans, gorillas and macaque were used. This validation reached an accuracy of more than 85% for all the species. Finally, this method was able to identify two sequences within the Swiss-Prot database that were reannotated. The semi-supervised based method was trained and tested with data from human, mouse, platypus, chicken, opossum, orangutan and xenopus, in five semi-supervised learning techniques. The contribution of this method was the reduction of the size of the classified training data set. In the best scenario, only two annotated sequences were used in the training phase, which is an expressive gain when compared to other tools available in the literature. Accuracies obtained by the method in the best cases were $\approx 95\%$ for human and mouse datasets, $\approx 90\%$ for chicken, opossum and orangutan datasets, and $\approx 80\%$ for data platypus and xenopus datasets. RNA-seq data were used for testing, having obtained more than 95% of accuracy. This data was used to train the orangutan and xenopus models, also leading to an excellent accuracy.

Keywords: long non-coding RNA (lncRNA), Support Vector Machine (SVM), Principal Component Analysis (PCA), Semi-supervised learning

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Since Watson and Crick [3] proposed the DNA double helix in the 50's, several fields of genetics and molecular biology have been boosted. Later, in the 90's, the human genome project [4, 5] promoted rapid advances in genomic sequences' experiments in molecular biology labs, and also in sequencing techniques [6]. In the beginning of the 21th century, genomic data generated in many other genome projects around the world needed the support of other scientific and technological areas, e.g, mathematics, statistics and computer science, and gave rise to a new research area, bioinformatics [7]. Since then, billions of nucleotide sequences composing the chromosomes of distinct organisms [8, 9] together with biological functions associated to the genes of these organisms have been discovered [10, 11, 12]. In addition, more advanced findings, e.g., information about proteins, mechanisms of genetic regulation, cellular processes and metabolism [13, 14, 15] have been focuses of recent research.

These studies are based on the so-called central dogma of molecular biology [16], proposed by Watson and Crick [17], who advocated that each RNA assemblied from a region of DNA produced one protein, schematically DNA $\rightarrow$ RNA $\rightarrow$ protein. Later, the dogma was revised, due to findings of RNAs not only participating of the protein synthesis, but in other important regulatory processes and other cellular functions.

From the beginning of the genomic research in the 90's until the early 2000's, the main focus was the study of the DNA and the identification of its genes [18]. Three methods were used: (i) complementary DNA (cDNA) cloning, and sequencing of messenger RNAs (mRNAs) expressed sequence tags (ESTs) [19, 20]; (ii) identification of conserved coding exons by comparative genome analysis [21]; and (iii) computational methods to predict genes [5], e.g., Smith-Waterman [22] and Blast [23], which are efficient and effective methods to find protein coding genes conserved during the evolution of organisms.

Nowadays, high-throughput sequencing techniques [24, 25] allowed to develop hundreds of genome and transcriptome projects all over the world, which have been creating

enourmous volumes of biological data [26]. In this context, more advanced computational techniques to analyze them became essential [27, 7].

In particular, many researches have been discovering and describing a rapidly increasing number of RNAs in eukaryotic genomes that do not produce proteins [28, 29, 30, 31, 8]. Thus, RNAs can be divided in coding and non-conding RNAs (ncRNAs) [32].

NcRNAs are a highly heterogeneous group. A well-known class of structured ncRNAs are involved in the synthesis of proteins, e.g., mRNAs, tRNAs and rRNAs [32, 33, 34]. Other classes of ncRNA, e.g., snoRNAs, snRNAs and RNase P RNAs form an additional elaborate layer in the regulation of gene expression [28], ranging in length from about 20 bases in microRNAs and siRNAs [35] to "macroRNAs" spanning hundreds of kilobases [36, 37], known as long non-coding RNAs (lncRNAs). Also, while the majority of ncRNAs seems to be spliced and processed similar to coding mRNAs, there is also a large body of unspliced transcripts [38, 39] and a vast number of small processing products [40]. The functions of these distinct ncRNAs are analogously diverse. In fact, they appear to be involved in virtually all the regulatory processes in the cell.

LncRNAs, often pragmatically defined as transcripts with a length of more then 200 nucleotides without apparent coding capacity, are still rather poorly understood [41, 33, 42]. Nevertheless, some classes, such as chromatin-associated long *intergenic* ncRNAs (lincRNAs) [43], as well subgroups that are directly involved in transcriptional and post-transcriptional regulation [44, 45, 46], have been identified in high throughput analyses. An extensive literature links lncRNAs with a wide array of diseases [47, 48, 49, 50], although the molecular mechanisms underlying lncRNA action are still largely unknown.

However, despite the importance of lncRNAs, and the existence of very different *in vivo* and *in silico* methods, there are no sets of well-defined attributes that allow to distinguish mRNAs from lncRNAs [32, 51, 52, 53, 54]. Therefore, applying the same strategies for predicting coding genes do not generate good results in the identification and classification of lncRNAs [55, 42].

On the other hand, in artificial intelligence, machine learning is a research field aiming to enhance knowledge stored in machines using learning algorithms. Learning algorithms can be divided in four distinct types [56, 57, 58]: (i) supervised learning, which uses functions taking input and output examples, in order to learn patterns; (ii) unsupervised learning, which finds patterns in *a priori* non classified data; (iii) reinforcement learning, which uses learning functions based on action rewards; and (iv) semi-supervised learning, which uses supervised and unsupervised learning techniques, in order to outperform both techniques.

Recently, many computational methods using machine learning techniques were applied to distinguish protein coding transcripts (PCTs) from lncRNAs. Among them, we

cite CPAT [59], lncRNApred [60], lncRScan-SVM [61], DeepLNC [62] and FEELnc [63].

1.1 Motivation

Today large amounts of biological sequences generated by genome projects are stored in public databases. Particularly, determining characteristics of ncRNAs, as well their precursors, genomic locations, patterns of conservation and responses to cellular changes that correlate with individual processing steps and protein interactions, are very interesting problems.

Regarding to animals, about 1.5% of the human genes are transcribed into mRNAs, i.e., about 98.5% genes do not code for proteins, and there are thousands of lncRNAs among them [42]. This is similar for mammalian genomes in general, in which only 1.2% genes are transcribed into mRNA [64]. Even being overlooked for some years, recent studies pointed that lncRNAs can promote diseases, like cancer metastasis in humans [65]. Despite the lack of knowledge about lncRNAs, the importance of these genes encourage the development of researches to predict them. An important and useful problem is the distinction of lncRNAs and PCTs [32].

The difficulty of these problems makes them good candidates to use machine learning techniques. A well known approach is to use classification methods, such as SVM, to distinguish lncRNAs from PCTs. Also, semi-supervised learning methods, although not yet used for this purpose, can help to solve this task.

1.2 Problem

Although some computational and experimental methods are known in the literature, there is not a broadly used method to distinguish lncRNAs from PCTs.

1.3 Objectives

This project aims to propose computational methods based on machine learning techniques to distinguish lncRNAs from PCTs. In particular, we devise two methods:

- one based on SVM, using a special procedure based on Principal Component Analysis (PCA) to find features that can improve the distinction;
- another one based on semi-supervised learning methods, which can be appliede to organisms that do not have a large amount of known transcripts.

1.4 Chapters' description

In chapter 2, we review the central dogma of molecular biology, and some characteristics of protein synthesis. Non-coding RNAs are detailed, together with their classification and functions. In particular, computational methods to predict lncRNAs, as well as databases containing lncRNAs are described.

In chapter 3, we discuss machine learning methods. The two methods used in this thesis are explored, SVM and semi-supervised learning techniques. For the last method, some applications in bioinformatics are described.

In the next two chapters, we detail our contributions. In chapter 4, we present a SVM based method to distinguish lncRNAs from PCTs.

In chapter 5, we present a semi-supervised learning method for the same problem, which can be used in organisms that do not have a significant volume of transcripts.

Finally, in chapter 6, we conclude this thesis, and suggest future work.

Chapter 2

Long non-coding RNAs

In this chapter, we discuss long non-coding RNAs (lncRNAs), their biological aspects and also computational methods and databases containing lncRNAs. First, in section 2.1, we briefly present the central dogma of molecular biology, and discuss protein synthesis in more detail. Next, in section 2.2, we describe ncRNAs, their types and functions. Following, in section 2.3, we describe small ncRNAs, and present computational techniques and databases containing small ncRNAs. Finally, in section 2.4, we explore lncRNAs, also presenting computational methods and databases containing lncRNAs.

2.1 Central dogma of molecular biology

The classic central dogma of molecular biology [17] states that a DNA molecule is transcribed into a RNA molecule, which is translated into a protein, i.e., it defines the phases of the protein synthesis, as shown in Figure 2.1. When it was first introduced, in 1970, the researchers claimed that the only purpose of the RNAs was to allow the production of a protein, from the DNA.

The protein synthesis, as stated in the classic central dogma of molecular biology, has two phases: (i) transcription; and (ii) translation. The transcription phase begins when the helicase enzyme disrupts the hydrogen bonds between the DNA strands, leaving the strands opened to be copied by the RNA polymerase enzyme. This enzyme reads the DNA from 3' to 5', while synthesizing the messanger RNA (mRNA) in the direction 5' to 3'. The RNA synthesis begins in the promoter region of the DNA strand, and finishes in the terminator region. The transcription is based on the base pairing $A \mapsto U$, $T \mapsto A$ and $C \leftrightarrow G$, taking the gene of the DNA as a template [66, 67].

In eukaryotic cells, the transcription occurs in the nucleous, and the transcribed RNA is called pre-mRNA, which may suffer modifications in some organisms in order to become

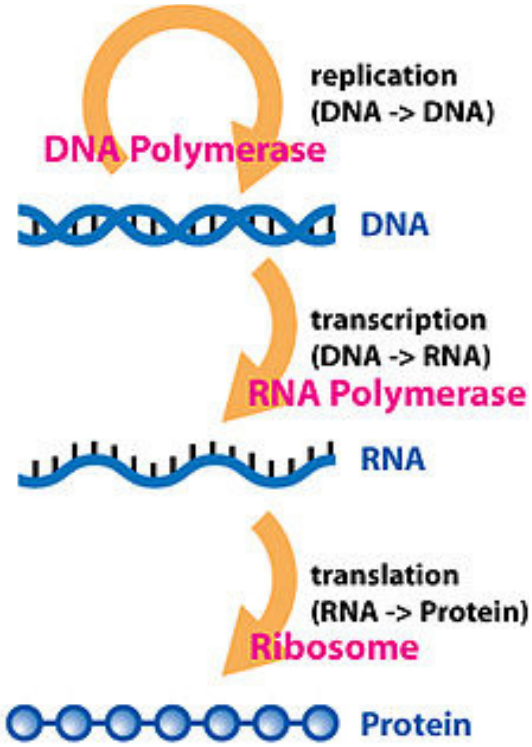

Figure 2.1: Classic central dogma of molecular biology: the protein synthesis.

a mature mRNA. This process, called splicing, removes introns from the pre-mRNA and some times also exons [66, 67].

After the transcription, the translation begins when one mRNA binds itself to a ribosomal RNA (rRNA) in the cytoplasm. The codons of this mRNA are paired with their corresponding anti-codons of a transporter RNA (tRNA) molecule, which transports the amino acids. Each codon is a sequence of three nucleotides that represents an unique amino acid from the genetic code, or a stop signal. Figure 2.2 shows the corresponding amino acid or stop codon, for each codon.

The protein translation always starts with the start codon (AUG), and finishes with one of the stop codons (UAG, UAA or UGA). The outcome of this whole process is a protein molecule.

The protein synthesis also occurs in prokaryote cells. Figure 2.3 shows some diferences between eukaryotes and prokaryotes.

In the 2000's, after finding non-conding genes, previously called "junk DNA" [68], the central dogma was revised [69], to include these novel genes. In other words, the classic central dogma previously considered RNA as a messenger to synthetize proteins, having no independent function besides this process. In recent years, biological discoveries proved that there is a variety of non-coding RNA molecules playing important roles in the cellular

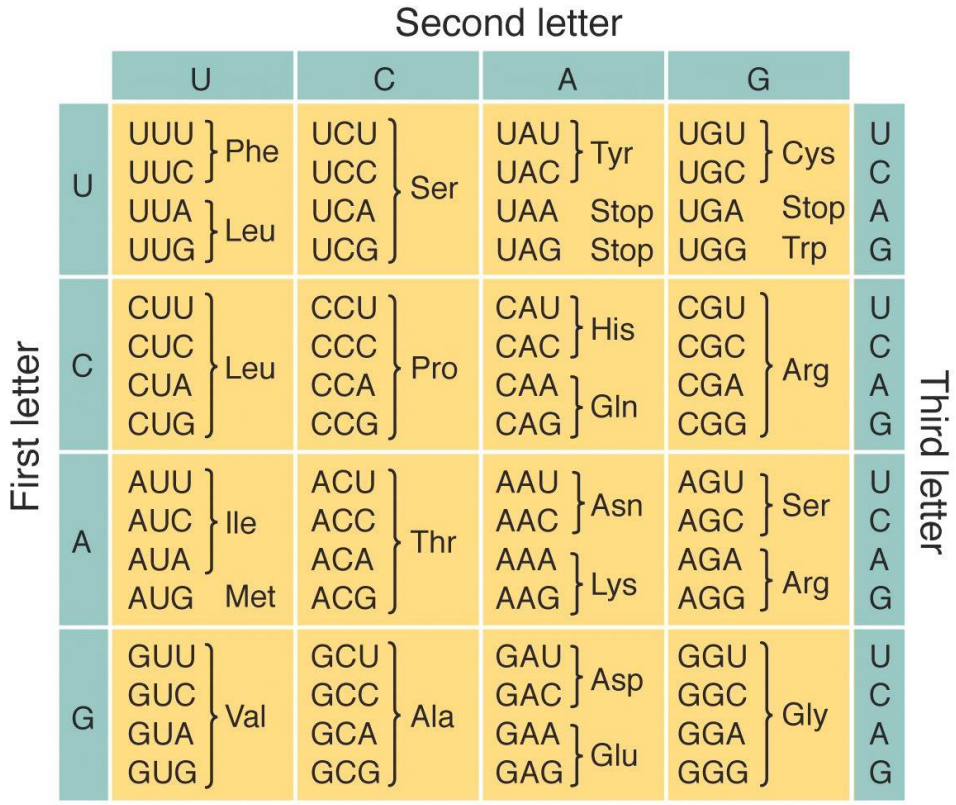

Figure 2.2: The genetic code.

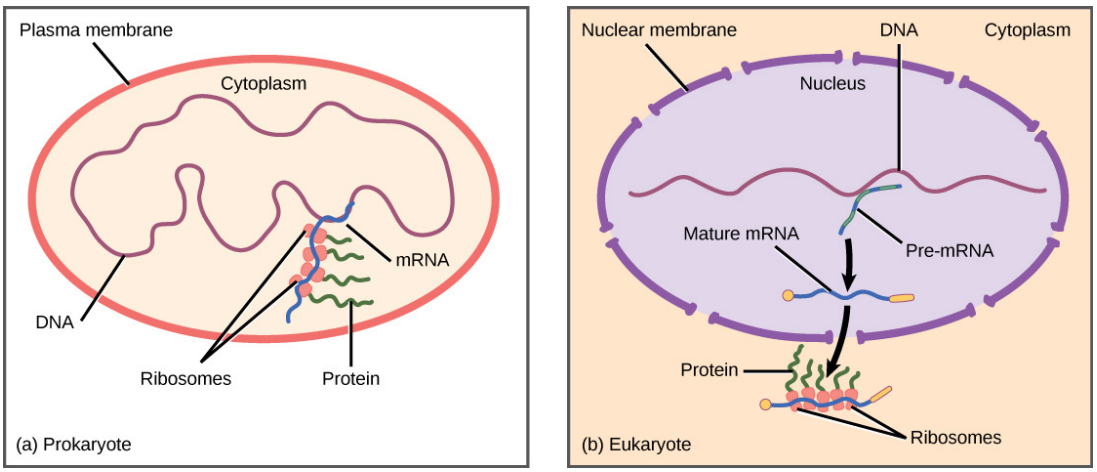

Figure 2.3: The protein synthesis in (a) a prokaryotic cell; and in (b) an eukaryotic cell.

structure as a whole [42, 33, 70, 71, 28, 72]. In the following section, we show some of the functions of these so-called non-coding RNAs.

2.2 Classes of ncRNAs

Non-coding RNAs (ncRNA), also called functional RNAs, are transcribed RNA molecules that are not translated into a protein, but rather perform different functions in the cellular metabolism [32]. These RNAs not coding for proteins represent a large portion of the human genome, having important roles in the cellular structure as well as in catalytic and regulatory processes in the cell [73, 74, 75].

Despite the importance of the ncRNAs, these molecules were identified but not deeply studied in the 80's and 90's, perhaps due to technical difficulties related to identify these non stable and small molecules [32].

From the early 2000's, researches in ncRNAs were resumed, due to the increasing amount of ncRNAs identified by biologists, described in the literature. The most remarkable finding about structural RNAs was related to the development of the nervous system, confirming the observation that the amount of non-coding regions is proportional to the complexity of the organisms [76, 77, 64, 78].

Despite the fact that many characteristics and biological functions have been discovered in recent years to study ncRNAs, computational methods have similar problems as those from experimental methods. Bioinformatics does not have an unique method to predict ncRNAs, although some criteria are used, e.g., ncRNAs in general have no long ORFs, stop codons occur more than expected over a sequence [79], RNAs are conserved regarding to their secondary structures rather than their primary sequences. These characteristics prevent the detection of ncRNAs using traditional tools, like those used to characterize the similarity of DNA in proteins [80, 81, 82]. Studies incorporating the use of codons, synonymous and non-synonymous substitutions, as well as minimum energy folding are also successful to identify ncRNAs [76, 82, 83, 84].

In general, ncRNAs do not have conserved sequences, presenting as their main characteristic the conservation of their spatial structures, including two or three-dimensional, making identification more difficult. The best known ncRNAs have a complex three-dimensional structure, and have catalyst and structural functions [85]. There is still a tendency in bioinformatics to use a combination of several computational methods to characterize ncRNAs using different principles, and then to analyze all the information generated by these methods to decide which RNAs are probably non-coding [86, 87, 88].

Moreover, the absence of a translated protein generated from a transcript is not a sufficient condition to characterize a ncRNA, since this transcript might be translated

once exposed to other conditions, environmental or physiological [86, 87, 33]. This is a drawback faced by computational methods to classify RNAs.

In general, ncRNAs can be divided in two large groups: (i) small ncRNAs, which comprehend a great amount of well known ncRNAs; and (ii) long ncRNAs (lncRNAs), which are the least undestood transcipts today. We explore these two classes in the following sections.

2.3 Small ncRNAs

As said before, non-coding RNAs, such as tRNAs and snRNAs, as well as small bacterial regulatory RNAs, are called small ncRNAs, although they are not related to small eukaryotic RNAs. Eukaryotic small RNAs can be characterized by their sizes, limited from 20 to 30 nucleotides, and their association with the *Argonaute* protein family (Ago). At least three classes of small RNAs are encoded in the human genome, based on their biogenesis mechanisms and their associated type of *Argonaute* proteins: miRNAs, endo-siRNAs or esiRNAs and piRNAs [89]. Figure 2.4 shows three examples of miRNA molecules.

Other small RNAs have been isolated biochemically, such as the small nucleolar RNA (snoRNA), which primarily guide chemical modifications of other RNAs, as shown in Figure 2.5.

Small ncRNAs are classified according to their apparent function within the cell. There is still discussion about the ncRNA amount and division [32]. Table 2.1 shows types of known small ncRNAs and their functions.

2.3.1 Computational methods

Following, we present computational methods to identify and classify small ncRNAs.

Infernal [94] (Inference of RNA alignments) is a secondary structure alignment tool. It builds consensus RNA secondary structure, called covariance models (CM), from multiple sequence alignments in Stockholm format, and it is based on Stochastic Context-Free Grammars (SCFG). This method uses the CMs to search for similarities with the secondary structures of RNAs stored in the Rfam database.

Vienna [84] is a software package used to generate and compare RNA secondary structures. The methods in this package are based on minimum free energy (MFE) and in the probabilities of base pairing.

DARIO [54] is a web application that aims to predict small ncRNAs. For this, an analysis is made for quality control criteria. Then, the analyzed sequence is compared with previously selected ncRNAs of known species. This comparison has two possible outcomes: (i) the sequence overlaps regions of exons, thus disregarding any subsequent analysis;

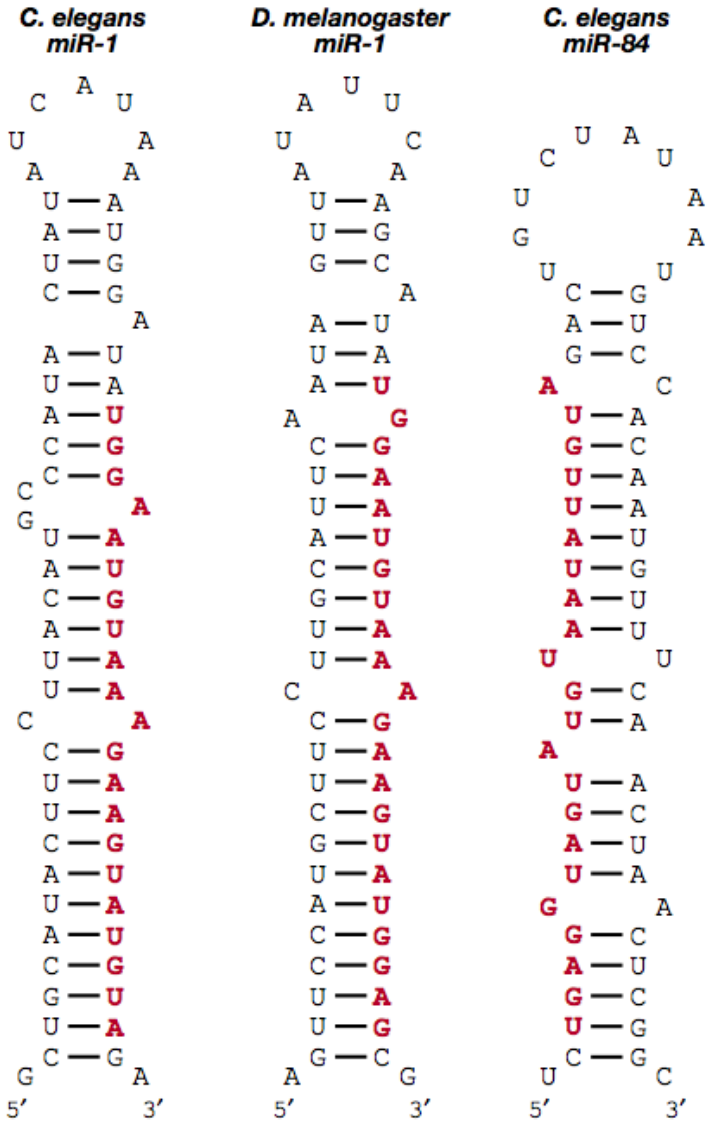

Figure 2.4: Three examples of microRNAs. Proposed structure of the precursor stem is shown, with residues in the mature microRNA (miRNA) shown in red. Comparison of *Caenorhabditis elegans* miR-1 with *Drosophila melanogaster* miR-1 shows perfect conservation of the mature miRNA (except for length variability at the 3' end). Comparison of miR-1 with miR-84 shows an example of how mature miRNAs are asymmetrically produced from either side of the precursor stem [32].

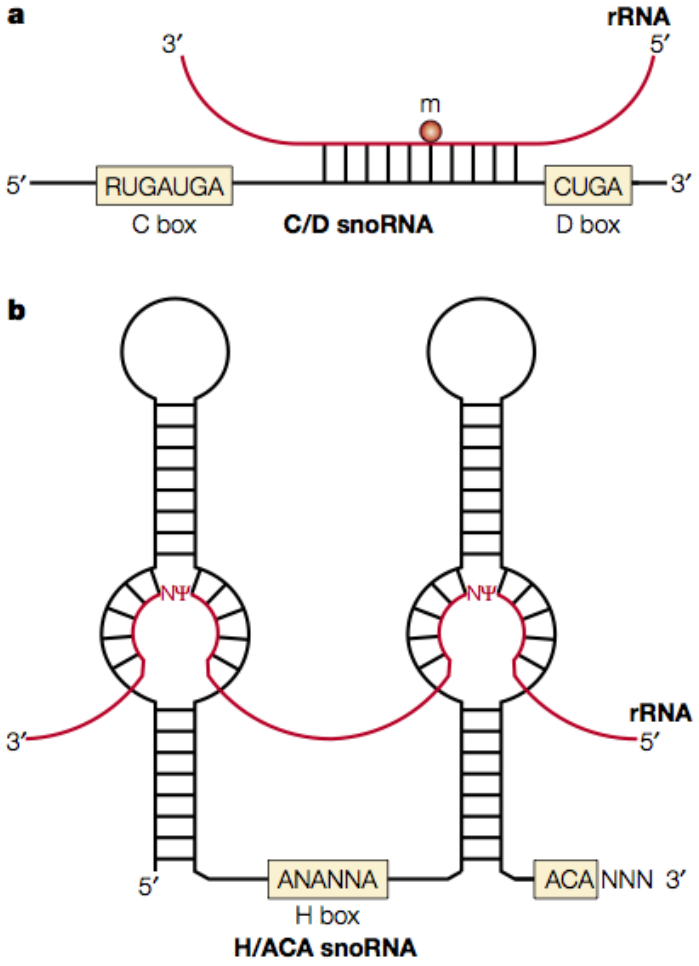

Figure 2.5: Diagrams of snoRNAs guiding modification to target rRNA bases. (a) C/D box small nucleolar RNAs (snoRNAs) use antisense complementarity to target RNA for 2'-O-ribose methylation. (b) H/ACA box snoRNAs use antisense complementarity in an internal loop to target RNA for pseudouridylation [32].

Table 2.1: Small ncRNAs and their functions [32, 90, 89, 91, 92, 93].

Name	Meaning	Function
miRNA	micro RNA	putative translational regulatory gene family
siRNA	small interfering RNA	active molecules in RNA interference
endo-siRNA	endogenous small interfering RNAs	acts as post-transcriptional regulators that target RNAs
snRNA	small nuclear RNA	includes spliceosomal RNAs
snoRNA	small nucleolar RNA	most known snoRNAs are involved in rRNA modification
stRNA	small temporal RNA	interrupt the translation of mRNAs
piRNA	piwi-interacting RNAs	acts in the regulation of translation and mRNA stabilization
rasiRNA	repeat-associated small interfering RNA	acts in the silencing of gene transcription through chromatin remodeling
vtRNA	vault RNA	located at a conserved genomic locus linked to the protocadherin gene cluster
Y RNA	Y RNA	associated with chromosomal DNA replication in a human cell-free system

and (ii) the sequence overlaps regions of intronic or intergenic regions. In this case, the sequence is used for ncRNA prediction. This prediction is based on machine learning techniques to identify characteristic patterns previously identified in several classes of ncRNAs.

PORTRAIT [52] is an algorithm based on a Support Vector Machine (SVM) to predict ncRNAs. For this, features used in the SVM are extracted from each investigated sequence. This tool uses two SVM models: (i) a protein dependent model, in which classical models are used to identify the proteins; and (ii) a protein independent model, when tools for identifying proteins indicate that the sequence is not a protein. The training sets were constructed based on known databases.

PSoL (Positive Sample only Learning) [51] is an algorithm developed to predict small ncRNAs within a set of non classified genes in the genome of *E. Coli*. To solve this problem, genes enconding putative ncRNAs were considered as positive, and the ones that do not code were taken as negative. A SVM was trained with positive data only, i.e., only previously known putative ncRNAs, and it was used to extract positive data within the non classified set. This approach does not use any negative data set to train the SVM, since a non classified gene can be either negative and positive.

SnoStrip [95] is an automatic annotation pipeline, developed specifically for comparative genomics of snoRNAs. It uses sequence conservation, canonical box motifs, and secondary structure to predict putative target sites.

SnoReport [96, 97] is a snoRNA identification software, which relies on the conserved sequence boxes and the secondary structure of snoRNAs. It has a filter based on SVM classifiers, trained to distinguish between microRNA precursors and other types of hairpin-like structures.

RNAsnoop [98] is a target predictor for H/ACA snoRNAs. It computes thermodynamically optimal H/ACA-RNA interactions with dynamic programming (DP), and uses a SVM, trained to distinguish true binding sites, together with a system to evaluate comparative information.

Table 2.2 summarises these methods, and shows their objectives and computational techniques.

2.3.2 Databases

Now, we describe databases containing small ncRNAs.

Rfam [99] is a curated database, containing information about thousands of ncRNA families. It consists of two distinct sets of data: profiles of covariance models (CMs) and seed alignments. CMs are statistical models derived from combinations of information, such as secondary structure and primary sequence represented by multiple sequence align-

Table 2.2: Tools to predict small ncRNAs.

Tool	Purpose	Method
Infernal [94]	secondary structure alignment	SCFG and CM
Vienna [84]	secondary structure prediction and comparison	MFE
DARIO [54]	small ncRNA prediction	Machine learning algorithm
PORTRAIT [52]	ncRNA identification	SVM
PSoL [51]	prediction of small ncRNAs in *E. Coli*	SVM
SnoStrip [95]	Identification of snoRNAs	Comparative Genomics
SnoReport [96]	snoRNA identification	SVM pre-filter
RNAsnoop [98]	prediction of H/ACA snoRNA targets	DP and SVM

ment. Each CM profile corresponds to a family of ncRNA. The seed alignments are stored in a Stockholm format file, and contains representative members of each ncRNA family generated through various structural alignments.

NONCODE [100] is a database containing all classes of ncRNAs, except tRNAs and rRNAs, also including human and mouse lncRNAs.

MiRBase [101] is a database containing miRNAs, in which each entry represents a predicted hairpin portion of a miRNA transcript, with information about its location and mature miRNA sequence.

Table 2.3 summarises these databases and their contents.

Table 2.3: Databases and their contents.

Databases	Contents
Rfam [99]	ncRNA families, mainly for small ncRNAs
NONCODE [100]	ncRNAs except tRNAs and rRNAs
miRBase [101]	miRNAs

2.4 LncRNAs

Currently, despite the lack of knowledge about the roles played by long ncRNAs (lncRNAs) [41, 42], it is known that many of the transcribed sequences, even from non-coding genes, are associated with a lncRNA [102]. The distinction of these ncRNAs is still taken by their sizes, greater than 200 nucleotides, and by the fact that proteins are not syn-

thesized from them. But this is still not enough [41, 42], since lncRNAs sometimes have some protein-coding capability [33, 42].

The amount of lncRNA transcripts in the mouse genome is approximately 30.000 [103] and most of the trancribed genes in the human genome are lncRNAs [104]. LncRNAs can be classified into six major categories [105]: (i) intergenic (lincRNAs), located between two protein-coding genes; (ii) intronic, located within introns of protein-coding genes (iii) bidirectional, transcribed within 1 kb of promoters in the opposite direction from the protein coding transcript; (iv) enhancer, generally <2 kb and transcribed from enhancer regions of the genome; (v) sense, transcribed from the sense strand of protein-coding genes, and can overlap introns and part or all of the exon; (vi) antisense, transcribed from the antisense strand of protein-coding genes, and can overlap an exon of the protein-coding gene in the sense strand, an intron, or both. Figure 2.6 illustrates these categories.

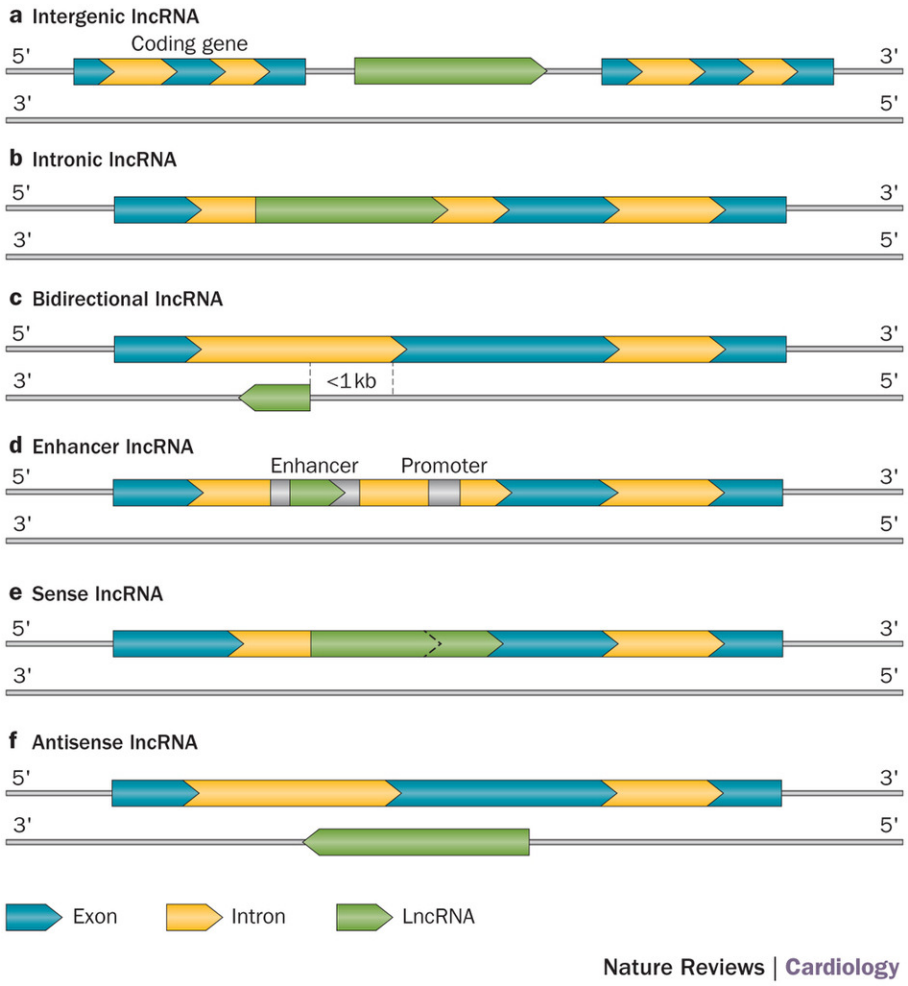

Figure 2.6: Five major lncRNA categories: (a) intergenic; (b) intronic; (c) bidirectional; (d) enhancer; (e) sense; and (f) antisense [105].

Mercer *et al.* [33] discovered some other functions of lncRNAs. They can act in chromatin modifications, mediating epigenetic changes by recruiting chromatin remodelling complexes to specific genomic *loci*. These molecules also play a role in transcriptional regulation. The ability of ncRNAs to recognize complementary sequences also allows highly specific interactions that are responsible to regulating various steps in the post-transcriptional processing of mRNAs, including their splicing, editing, transport, translation and degradation.

An illustrative mechanism by which lncRNAs regulate local protein-coding gene expression at the level of chromatin remodelling, transcriptional control and post-transcriptional processing is shown in Figure 2.7.

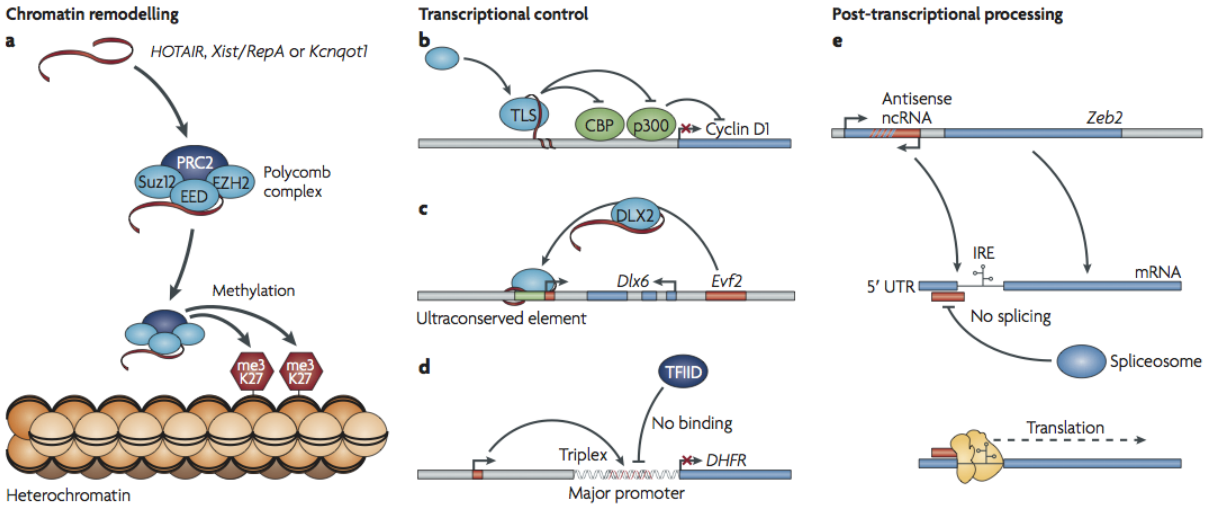

Figure 2.7: Known functions of lncRNAs. (a) ncRNAs can recruit chromatin modifying complexes to specific genomic *loci* to impart their catalytic activity. In this case, the ncRNAs HOTAIR21, Xist and RepA (the small internal non-coding transcript from the Xist *locus*), or Kcnqot1 recruit the Polycomb complex to the HoxD locus, the X chromosome, or the Kcnq1 domain, respectively, where they trimethylate lysine 27 residues (me3K27) of histone H3 to induce heterochromatin formation and repress gene expression; (b) ncRNAs can regulate the transcriptional process through a range of mechanisms. NcRNAs tethered to the cyclin D1 gene recruit the RNA binding protein TLS to modulate the histone acetyltransferase activity of CREB binding protein (CBP) and p300 to repress gene transcription; (c) An ultra conserved enhancer is transcribed as a lncRNA, Evf2, which subsequently acts as a co-activator to the transcription factor DLX2, to regulate the Dlx6 gene transcription; (d) A ncRNA transcribed from the DHFR minor promoter in humans can form a triplex at the major promoter to occlude the binding of the general transcription factor TFIID, and thereby silence DHFR gene expression; (e) An antisense ncRNA can mask the 5' splice site of the zinc finger homeobox mRNA Zeb2 from the spliceosome, resulting in intron retention. The translation machinery can then recognize and bind an internal ribosome entry site (IRE) in the retained intron, resulting in efficient Zeb2 translation and expression [33].

2.4.1 Computational methods

In what follows, we describe computational methods to predict lncRNAs.

ISeeRNA [106] is a tool to identify lincRNAs in transcriptome sequencing data. It is based on a SVM classifier trained to identify these genes, using features as ORF length and k-mers. Human and mouse were used both to train and test the models.

Lnc-GFP [107] (long non-coding RNA global function predictor) is a lncRNA function predictor based on a bi-colored network. It integrates gene expression data with protein interaction data to predict probable lncRNA functions.

For the problem of distinguishing protein coding transcripts (PCTs) from lncRNAs, many computational methods using machine learning techniques were proposed. CPC (Coding Potential Calculator) [108] works well with known PCTs, although it may tend to classify novel PCTs as ncRNAs, if they have not been recorded in protein databases [108]. CPAT tool [59] is based on logistic regression, and it uses four features based on ORFs. Both, CPC and CPAT, are focused on PCTs identification by calculating a coding potential. LncRNApred [60], lncRScan-SVM [61], DeepLNC [62] and FEELnc [63] can predict lncRNAs. IseeRNA [109] was specially designed to predict specifically lincRNAs. LncRScan-SVM and iSeeRNA are methods based on Support Vector Machines (SVM). LncRNApred and FEELnc are methods constructed using Random Forest, having used features extracted from the sequence nucleotides to predict lncRNAs. DeepLNC was built using deep neural networks, having reported high accuracy to predict lncRNAs.

Table 2.4 summarises these methods, indicating their objectives and computational techniques.

Table 2.4: Tools, and their corresponding objectives and techniques, to predict lncRNAs.

Tool	Purpose	Method
iSeeRNA [106]	lincRNA identification	SVM
lnc-GFP [107]	lncRNA function prediction	Bi-colored network
CPC [108]	distinguish PCTs from ncRNAs	SVM
CPAT [59]	distinguish PCTS from ncRNAs	logistic regression
lncRNApred [60]	distinguish lncRNAs from PCTs	random forest
lncRScan-SVM [61]	distinguish lncRNAs from PCTs	SVM
DeepLNC [62]	distinguish lncRNAs from PCTs	deep neural networks
FEELnc [63]	distinguish lncRNAs from PCTs	random forest

2.4.2 Databases

Here, we describe databases containing lncRNAs.

NRED [110] (ncRNA Expression Database) is a database containing lncRNAs in human and mouse genomes. This database also provide ancillary information for featured ncRNAs, including evolutionary conservation, secondary structure evidence, genomic context links and antisense relationships.

NONCODE [100], mentioned before, also covers almost all the published human and mouse lncRNAs.

DIANA-lncBase [111] is a miRNA-lncRNA interactions database. It contains two modules: (i) an experimental module, with detailed information for more than $5,000$ interactions, between $2,958$ lncRNAs and 120 miRNAs; and (ii) a prediction module, with detailed information for more than 10 million interactions, between 56,097 lncRNAs and 3,078 miRNAs, results from the DIANA-microT-CDS algorithm [112].

LncRNADisease [113] is a database containing lncRNAs associated with diseases. It has 600 lncRNA-disease entries and 475 lncRNA interaction entries, including 251 lncRNAs and 217 diseases.

Ensembl [9] is a database containing genome and transcriptome data of various species. Among these data, there are several annotated lncRNAs

lncRNAdb [114] provides comprehensive annotations of eukaryotic lncRNAs.

PLncDB [115] provides information about lncRNAs in plants.

Table 2.5 summarizes theses databases and their contents.

Table 2.5: Databases and their contents.

databases	Contents
NRED [110]	human and mouse lncRNAs
NONCODE [100]	human and mouse lncRNAs
DIANA-lncBase [111]	miRNA-lncRNA interactions
lncRNADisease [113]	lncRNAs associated with diseases
Ensembl [9]	genome and transcript database
lncRNAdb [114]	eukaryotic lncRNAs
PLncDB [115]	lncRNAs in plants

Chapter 3

Machine learning

In this chapter, the techniques of machine learning used in this thesis are detailed. First, in section 3.1, some basic concepts of machine learnnig are presented. The paradigms of supervised learning, unsupervised learning and reinforcement learning are briefly described in section 3.2. In Section 3.3, we explore the method of Principal Component Analysis (PCA), used here to select features. In section 3.4, the Support Vector Machine (SVM) method is explored, while in section 3.5 the semi-supervised learning concepts and techniques are discussed.

3.1 Basic concepts

The machine learning field provided significant progress to several areas, including molecular biology. It is said that a program learns from a set of experience related to a set of tasks with a performance measure if its performance to execute its task improves with increasing experience [56].

Some basic concepts are needed to understand the machine learning techniques [57]:

- **Example** is a pair (x, y), where x is the input and y is the expected output of the function f for the input x. Both x and y do not need to be numbers, instead they can be any kind of value.
- **Feature** or attribute is a characteristic of an object, therefore, we say that a set of attributes defines an object.
- **Class** or **Label** is the classification given to an object.
- **Set of examples** are divided in two sets: (i) training set; and (ii) test set. The training set is meant to train the algorithm, and the test set is meant to validate its training.

- **Overfitting** is a phenomenon that occurs when an algorithm adjusts itself to a very specific data set, thus becaming not effective for a more generic data set.

Machine learning algorithms are classified into four distinct types: (i) supervised learning; (ii) unsupervised learning, (iii) reinforcement learning; and (iv) semi-supervised learning.

3.2 Paradigms

In this section, we generally describe three paradigms of machine learning: supervised learning; unsupervised learning; and reinforcement learning, respectively.

3.2.1 Supervised learning

Supervised learning consists in finding a function from a set of examples of its inputs and outputs. In other words, supervised learning algorithms have to find a function h, also called hypothesis, which approximates the function f using a set of examples. This is called inference. Supervised learning can be easily applied in a fully observable environment, since an agent can observe the outcome of all its actions [57].

Learning is basically finding a function h, from the training set, which performs in the same manner even for new examples beyond the training set. The accuracy of the hypothesis is measured by the test set. The hypothesis generalizes well if it correctly predicts the value of $f(x)$ for novel examples. For stochastic functions, the algorithm have to learn the conditional probability distribution, $P(y|x)$ [57].

The learning problem is called classification when the outputs of a function f are a finite set of discrete values, otherwise it is called regression. A regression problem, technically, is to find a conditional expectation or average value of y. Figure 3.1 represents a regression and illustrates the learning of a mathematical function, where $y = f(x)$.

Examples of supervised learning algorithms are Support Vector Machines (SVM), neural networks, genetic algorithms and decision trees.

3.2.2 Unsupervised learning

The unsupervised learning is the recognition of patterns in the entries, with no information about the desired output. An agent who only uses unsupervised learning can not learn what to do because it has no information on what would be a correct action and a desirable state to be achieved.

Examples of unsupervised learning techniques are Self-Organizing Maps, k-means algorithm and hierarchical clustering algorithms.

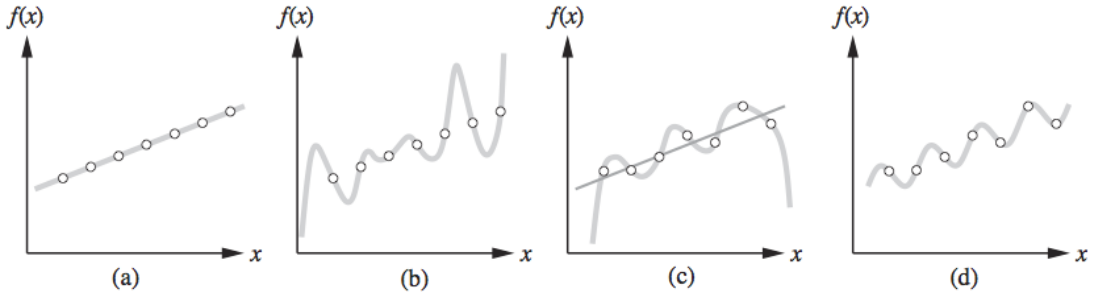

Figure 3.1: (a) Example (x,f(x)) pairs and a consistent linear hypothesis. (b) A consistent degree-7 polynomial hypothesis for the same data set. (c) A different data set, which admits an exact degree-6 polynomial fit or an approximate linear fit. (d) A simple, exact sinusoidal fit to the same data set [57].

3.2.3 Reinforcement learning

Reinforcement learning is associated to learning through iterations to achieve a goal. The agent is the decision maker element, and it is also the entity that learns. The agent perceives and interacts with the environment, which is all but himself, through perceptions and actions. The actions taken by the agent generates an immediate reward, and the buildup of its rewards causes the agent to learn what are the best actions to be taken within the known possible states of the environment [1]. Figure 3.2 illustrates a simple iteration cycle of a reinforcement learning process.

Reinforcement learning is associated to learning through iterations to achieve a goal. The agent is the decision maker element, and it is also the entity that learns. The agent perceives and interacts with the environment, which is all but himself, through perceptions and actions. The actions taken by the agent generates an immediate reward, and the buildup of its rewards causes the agent to learn what are the best actions to be taken within the known possible states of the environment [1]. Figure 3.2 illustrates a simple iteration cycle of a reinforcement learning process.

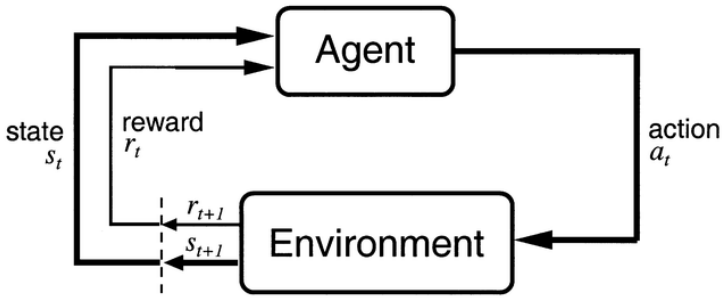

Figure 3.2: Reinforcement learning iteration cycle [1].

The reinforcement learning can be used when there are agents who need greater autonomy, inserted in an environment in which there are no examples of actions that serve as useful parameters to determine the next actions. In this case, the agent, after performing any action, may get a reward. This reward can be good or bad depending on the action taken by the agent. Keeping this in view, the role of reinforcement learning is to use the rewards obtained to learn the optimal actions, or near optimal, within this environment [56].

3.2.4 Semi-supervised learning

The semi-supervised learning, as its name says, is a methodology that seeks to extend the supervised learning methods using techniques of unsupervised learning and vice versa. It aims to solve problems with very few classified examples.

3.3 Feature Selection

Objects can be described by features, which are used in machine learning techniques. If a large number of features are used, this could lead a machine learning technique to the so-called *curse of dimensionality*, but only a small number of features can not lead to the wanted objective. Ideally, the number of selected features should be small enough to avoid the curse of dimensionality [116] and also large enough to achieve the desired classification performance. In our case, transcripts can be characterized by a large number of features, e.g., pattern frequencies (di, tri- and tetra-nucleotides), length and relative length of Open Reading Frames (ORFs).

In this thesis, we used Principal Component Analysis (PCA) to choose the pattern frequencies of a transcript that are most important to distinguish lncRNAs from PCTs. PCA is a mathematical procedure that basically uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables [117]. This procedure rotates the data coordinate system such that variable with the greatest variance lies on the first axis, the second one on the second axis, and so on.

The orthogonal transformation, also called *loading*, takes as input an $n \times n$ matrix with eigenvectors in its columns, where n is the number of dimensions of the coordinate system. Each dimension represents a feature. In order to reduce the number of dimensions, the coordinate system has to be rotated, and the last m dimensions are removed, resulting in a coordinate system with $n - m$ dimensions. In order to find the variables that most contribute to the variability of the data, loadings have to be used, cutting the rows of the least significant dimensions, and calculating the euclidean norm for each vector of the

loading $n \times (n - m)$ matrix. The biggest the norm, the most the dimension contributes to the variability of the data, allowing the method to select the $n - m$ most significant features.

3.4 Support Vector Machine

SVM is a supervised learning model used for classification and regression. In both cases, to achieve its objectives, SVM constructs hyperplanes in a high dimensional space, selecting the ones with the largest computed margin separating distinct classes, related to the training data [118] (Figure 3.3).

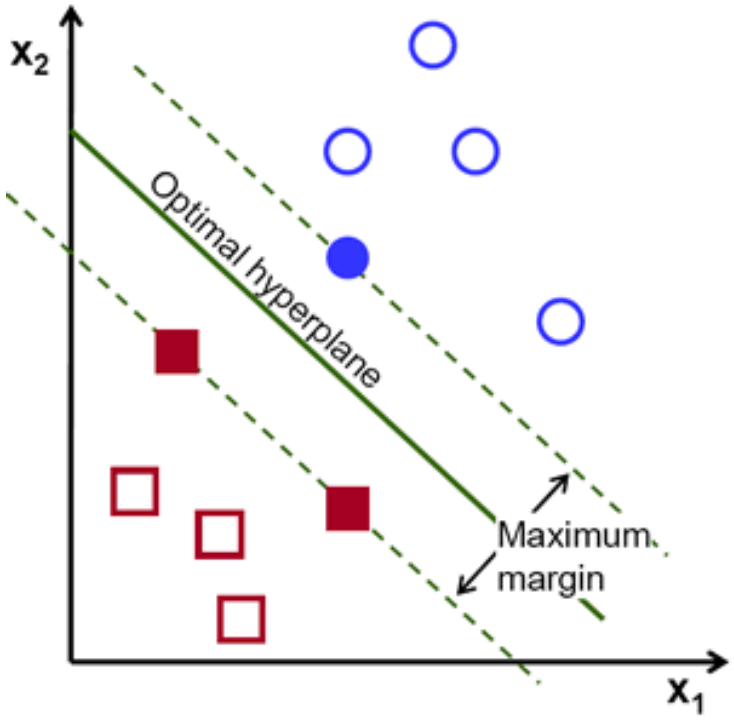

Figure 3.3: SVM hyperplane example [2].

Specifically, for a classification problem, SVM provides a non-probabilistic binary linear classifier that distinguishes one kind of observation from another, classifying them in two classes: positive and negative. For each observation, some features must be extracted, forming an attribute vector that represents the example.

Another relevant aspect to create a non-linear classifier is the choice of an appropriate kernel function $k(x_i, x_j)$, which can be used to transform the attribute vector of each observation in order to use the SVM method [119]. Some kernel functions are non-linear. Some common kernel functions [120], with selected constants γ and $coef0$, are shown:

$$k(x_i, x_j) = x_i \cdot x_j \quad \text{Linear}$$
$$k(x_i, x_j) = (\gamma \cdot x_i \cdot x_j + coef0)^d \quad \text{Polynomial}$$
$$k(x_i, x_j) = exp(-\gamma \cdot ||x_i - x_j||^2) \quad \text{Radial}$$
$$k(x_i, x_j) = tanh(\gamma \cdot x_i \cdot x_j + coef0) \quad \text{Sigmoid}$$

An important parameter for a SVM is the C cost, also known as penalty factor, a value that measures the penalty given for each non-separable points when selecting a hyperplane. If its to large, the SVM may overfit, and if its to small, it may underfit [120].

3.5 Semi-supervised learning

The semi-supervised learning, as said name says, is a methodology that seeks to extend the supervised learning methods using techniques of unsupervised learning and vice versa. It was shown that, for some applications, this algorithm overcomes the performance of the other two methods, used alone. Figure 3.4 illustrates the decision boundaries of each class of algorithm.

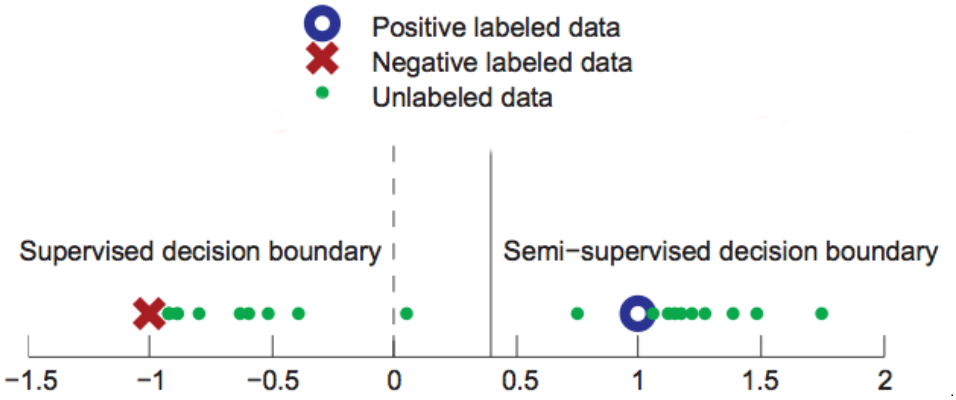

Figure 3.4: Example of supervised and semi-supervised decision boundaries.

The dataset typically applied to a semi-supervised learning algorithm consists of a set $X = \{x_1, ..., x_{i\in[n]}\}$ divided in two subsets: (i) a set $X_l = \{x_1, ..., x_l\}$, for which there is a set $Y_l = \{y_1, ..., y_l\}$ of labels, where y_i is the label of x_i; and (ii) a set $X_u = \{x_{l+1}, ..., x_{l+u}\}$ of unlabeled data [58].

This type of dataset is found at the beginning of every sequencing project. This is also a characteristic of almost all lncRNAs databases, which makes this learning paradigm a strong candidate for the distinction of lncRNAs and PCTs.

For the semi-supervised learning paradigm, Brachman et al. [121] propose two kinds of methods:

Semi-supervised classification

It is an extension of the supervised classification problem. The training set contains a set X_l of labeled data and a set X_u of unlabeled data, and assumes that the amount of unlabeled data is much larger than the amount of labeled data, $l \ll u$. This method aims to train a classifier f with the set of labeled and unlabeled data in order to outperform a supervised classification using only the labeled data.

Constrained clustering

It is an extension of the unsupervised clustering problem. The training set contains a set X_u of unlabeled data. The set X_l of labeled data is used to extract information about the cluster, for example, constraints that determines if two values x_i and x_k can be elements of the same cluster, or not.

There are two methodologies applied to the classes of semi-supervised learning algorithms, transduction and induction. Transduction consists of directly estimating a function $f : X^{l+u} \mapsto Y^{l+u}$ to classify only the unlabeled data, while induction consists of inferring a function $f : X \mapsto Y$ to classify the whole dataset [121].

As other techniques of machine learning, like supervised learning, which have to have some assumptions and prerequisites to rely on, the semi-supervised learning technique also has some prerequisites and assumptions.

A prerequisite of this method is that the distribution of examples has to be relevant to the classification problem, i.e., the knowledge of $p(x)$ that is obtained from the unlabeled data X_u has to carry information useful in the inference of $p(y|x)$. Otherwise, semi-supervised learning will not yield an improvement over supervised and unsupervised learning [58].

In order to generalize from a finite training set to a possibly infinite test case set, four assumptions must be ensured [58]:

Semi-Supervised Smoothness Assumption

If two points x_i and x_j in a high-density region are close, so should be the corresponding outputs y_i and y_j.

This implies that if these two points are in a high-density region, i.e. a cluster, then their outputs are likely to be close, and if they are separated by a low density region, their outputs are far apart.

The Cluster Assumption

If points are in the same cluster, they are likely to be of the same class.

In other words, a cluster represents a class, but a class can be separated in one or many clusters. And if two points are classified in two different classes, there should be a low density region between them, which is called decision boundary.

The Manifold Assumption

The high-dimensional data lie roughly on a low-dimensional manifold.

Transduction

Vapnik's Principle: When trying to solve some problem, one should not solve a more difficult problem as an intermediate step.

For example, if label predictions are only required for a given test set, transduction can be argued to be more direct than induction. Note that transduction is not the same as semi-supervised learning: as said above, some semi-supervised algorithms are transductive, but others are inductive.

The semi-supervised learning methods are divided in four groups: (i) Generative Models; (ii) Low-Density Separation; (iii) Graph-Based Methods; and (iv) Change of Representation.

3.5.1 Generative models

Generative models involve the estimation of the condicional density $p(x|y)$. Any additional information about $p(x)$ for this method can be useful. It can be seen as a clustering problem with additional information about the cluster density, i.e., it implements the cluster assumption [58].

An advantage of this approach is that knowledge of the problem and of the data can be naturally incorporated by modeling. If the modeling is accurate, the unlabeled data will increase the algorithm performance, but if the modeling is not accurate, the unlabeled data will decrease the algorithm performance [58].

3.5.2 Low-density separation

This model is based on a maximum margin algorithm such as SVM. The method for maximazing the margin using labeled and unlabeled data is called transductive SVM (TSVM).

The TSVM problem is nonconvex, and thus difficult to optimize. A possible optimization is done by training the SVM with labeled data. Afterwards, the unlabeled data is classified and used to form a new training set with all data through iterations. After each iteration, the unlabeled data weight is decreased.

3.5.3 Graph-based methods

Today, graph-based methods represent the more active area of research within the semi-supervised learning field. These methods use graph representation for expressing data, where data is associated to a node of the graph, and the edges are labeled with the similarity between neighbouring nodes. Absence of an edge between two nodes is regarded as a lack of similarity between these nodes.

Learning through a graph consists of labeling nodes that have unlabeled data through the edges that connect the unlabeled data with a labeled node. Thus, w_{ij} is the similarity between the nodes x_i and x_j, and y_i and y_j are respectively the labels of this nodes. So the higher the value of w_{ij}, the more likely y_i and y_j are the same.

Two graph based methods are Label Propagation and Label Spreading [58].

3.5.4 Change of representation

This model is not intrinsically semi-supervised learning, but performs a two-step learning. It applies an unsupervised learning method in all data $X = X_l \cup X_u$, ignoring the available labels for a contruction of a new metric or a new kernel, and afterwards, applies a supervised learning method in all the labeled data X_l of this new kernel.

3.5.5 Bioinformatics applications

Next, some semi-supervised learning methods, applied to bioinformatics and computational biology, are briefly discussed.

Chapelle et al. [58] propose two methods. The first one aims to predict a structural class of one protein, given its amino acid sequence. This method extends a SVM to take advantage of the non labeled data, having used two cluster kernels: (i) the neighbourhood mismatch kernel; and (ii) the bagged mismatch kernel. In the same article, a second method aims to predict protein functions using a graph-based semi-supervised learning approach. Multiple graphs were combined and used for function classification of yeast proteins. The use of multiple graphs showed better performance than a single graph approach. When compared to a Semi-Definite Programming based Support Vector Machine (SDP/SVM), it shows comparable results in a shorter time.

Nguyen et al [122] created a semi-supervised learning method using human genes responsible for diseases and a protein-protein interaction database for humans to predict novel disease genes.

Regularized Least Squares for MiRNA-Disease Association (RLSMDA) [123] is also a semi-supervised learning method to identify relationships among diseases and miRNAs. This method does not use negative samples to build the models.

Stanescu and Caragea [124] presented a method that focuses on predicting splice sites in a genome, using self-training and co-training approaches.

Provoost et al [125] proposes a semi-supervised learning method that uses SVM to address gene regulation networks. SVM is used to classify unlabeled data, which creates new samples in the training data sets.

Chapter 4

A Support Vector Machine based method to distinguish long non-coding RNAs from protein coding transcripts

In this chapter, we present the article [126] published in BMC Genomics `https://bmcgenomics.biomedcentral.com`.

4.1 Background

In recent years, thousands of sequencing projects around the world have been creating enormous volumes of RNA data, which has led to the discovery and description of a rapidly increasing number of non-coding RNAs (ncRNAs) in eukaryotic genomes [29, 30, 31, 8]. NcRNAs are a highly heterogeneous group, ranging in length from about 20 bases in microRNAs and siRNAs [35] to "macroRNAs" spanning hundreds of kilobases [36, 37], known as long non-coding RNAs (lncRNAs). While the majority of ncRNAs seem to be spliced and processed similar to coding mRNAs, there is also a large body of unspliced transcripts [38, 39] and a vast number of small processing products [40]. The functions of ncRNAs are analogously diverse. In fact, they appear to be involved in virtually all the regulatory processes in the cell.

Although they are often pragmatically defined as transcripts of a more than 200 nucleotides in length, and without any apparent coding capacity, lncRNAs are still rather poorly understood [41, 33, 42]. Nevertheless, some classes, such as chromatin-associated long *intergenic* ncRNAs (lincRNAs) [43], as well as subgroups that are directly involved in transcriptional and post-transcriptional regulation [44, 45, 46], have been identified in

high throughput analyses. An extensive literature links lncRNAs with a wide array of diseases [47, 48, 49, 50], although the molecular mechanisms underlying lncRNA action are still largely unknown.

Distinguishing between protein coding transcripts (PCTs) and long non-coding transcripts (lncRNAs) is a surprisingly difficult task in practice, and there is still an ongoing controversy whether some or even the majority of the transcripts currently classified as "non-coding" can in fact be translated.

From a computational point of view, distinguishing PCTs from lncRNAs is a paradigmatic machine learning task, and several tools have become available for this purpose. Among these tools, CPC (Coding Potential Calculator) [108] and CPAT [59] have been developed to discriminate PCTs from ncRNAs. While CPC works well with known PCTs, it may tend to classify novel PCTs as ncRNAs, if they have not been recorded in protein databases [108]. The CPAT tool is based on logistic regression, and it uses four features based on ORFs.

Tools such as LncRNApred [60], lncRScan-SVM [61], DeepLNC [62] can predict lncRNAs. IseeRNA [109] was specially designed to predict lincRNAs. LncRScan-SVM and iSeeRNA are methods based on Support Vector Machines (SVM), trained with data from humans and mice, both presenting very good results. To predict lncRNAs, these two methods use GTF as input files, along with conservation data and some nucleotide patterns extracted from the sequences, to predict lncRNAs. LncRNApred is a method that was constructed using Random Forest, and features extracted from the sequence nucleotides to predict lncRNAs. DeepLNC was built using deep neural networks, and reported high accuracy to predict lncRNAs. Unfortunately, it is not clear which features were used, and the DeepLNC site presents an exception when any fasta file is submitted.

Recently, Wucher et al. [63] proposed FEELnc (FlExible Extraction of LncRNAs), a program to annotate lncRNAs based on a Random Forest model, trained with frequency nucleotide patterns and relaxed ORFs. They used FEELnc on a data set of canine RNA-seq samples, having improved the canine genome annotation with 10.374 novel lncRNAs.

Comprehensive reviews of these tools have been provided by Han et al. [61] and Guo et al. [127]. Similarly, Ventola et al. [128] studied features extracted from sequence data, those presented in the literature and some newly proposed features, in order to find signatures (groups of features) that can distinguish lncRNA transcripts from other classes, such as PCTs.

In general, the basic idea of the methods that use information of transcript nucleotides is to create a model to predict ncRNAs from known samples already stored in databases. Despite working well with the species for which they have been trained, these methods do not usually generalize for other organisms. In other words, these approaches are not

capable of reliably predicting lncRNAs in a variety of species.

In addition, there are various databases containing lncRNAs (see Guo et al. [127] and Fritah et al. [129] for detailed reviews). Among them, Ensembl [9], NONCODE v. 4.0 [130], lncRNAdb [114], PLncDB [115], NRED [110] provide information on general and specific lncRNAs, while DIANA-LncBase [131] and lncRNADisease [132] present interactions among lncRNAs and other ncRNAs or proteins.

Moreover, in recent years, experimental and computational models have been developed to predict secondary and tertiary structures of lncRNAs, as explored in Yan et al. [133]. While the prediction of the lncRNAs' secondary structures *in-vitro* has high-experimental costs, *in-silico* methods are low cost, but they exhibit high false-positive rates [127].

Although lncRNAs have very heterogenous characteristics [41, 33, 42], the previous described methods indicate that there are sets of features that allow researchers to distinguish lncRNAs from PCTs.

In this study, we present a SVM based method to distinguish lncRNAs from PCTs, using features extracted from transcript sequences: frequencies of nucleotide patterns selected by Principal Component Analysis (PCA) [117]; open reading frame (ORF) length; and ORF relative length. In addition, in order to analyze the performance of our method, we developed case studies with human, mouse and zebrafish data. We also compared results of our method to other tools found in the literature. To validate our model, we applied it to three different species (human, gorilla and rhesus macaque), as well as to human and mouse pseudogenes. Finally, we re-annotated data from Swiss-Prot, and annotated transcripts derived from RNA-seq data, reported in Necsulea et al [134].

4.2 Methods

4.2.1 Data

Four data sets for training the models were obtained from Ensembl [9]: human (*Homo sapiens*) assemblies GRCh37 patch 13 (hg19, GENCODE 19) and GRCh38 patch 10 (hg38, GENCODE 26), mouse (*Mus musculus*) assembly GRCm38 patch 5 (mm10, GENCODE M13), zebrafish (*Danio rerio*) assembly GRCz10. These transcript FASTA files contain PCT and lncRNA sequences, while the classification was extracted from the transcript biotype, provided by Ensembl.

4.2.2 The SVM based method

We propose a method based on SVM to distinguish lncRNAs from PCTs (see Figure 4.1), using PCA to reduce the number of features calculated from the nucleotides of the transcripts.

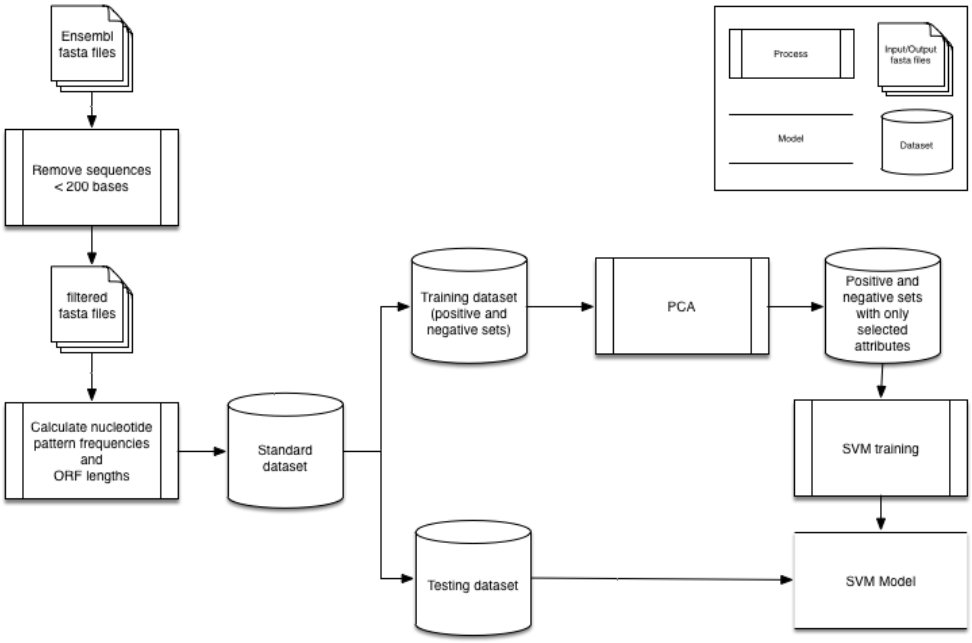

Figure 4.1: The method to distinguish lncRNAs from PCTs. The method is based on SVM and uses as attributes nucleotide pattern frequencies, chosen with the support of PCA, together with the first ORF relative length, a characteristic that informs the coding potential of a transcript.

First, a standard data set was created, removing all the sequences shorter than 200 bases from the original FASTA files. This standard data set contained, besides the transcripts (description and sequence), some calculated features (nucleotide pattern frequencies and ORF lengths) for each transcript, as follows. These features were divided in two sets: the first one contained the average frequency of the di-, tri- and tetra-nucleotide patterns in all the possible frames; and the second set contained the length and relative length of the first and the longest predicted ORFs. The relative length of an ORF is defined by its length divided by its corresponding transcript length.

The standard data set generated two other sets - training and testing, each composed of a positive set (containing lncRNAs) and a negative set (containing PCTs), of equal sizes. The training and the testing data sets were randomly generated, 75% for training and 25% for testing.

First set of features built with PCA

In the standard data set, there was a total of 336 different frequencies of nucleotide patterns in the first set: 16 di-nucleotide pattern frequencies; 64 tri-nucleotide pattern frequencies; and, 256 tetra-nucleotide pattern frequencies. We reduced the number of these possible features, having identified their relative importance, with the PCA method [117]. Thus, PCA was applied to all the nucleotide pattern frequencies of the training data set, to find how many, and which ones, would effectively help to distinguish between lncRNAs and PCTs.

The orthogonal transformation produced by PCA was used to calculate the "contribution" of each nucleotide pattern frequency. This orthogonal transformation is an $n \times n$ matrix with eigenvectors in its columns and features in its rows, where $n = 336$ frequencies of nucleotide patterns. We removed the m least significant columns from this matrix, obtaining a new $n \times (n - m)$ matrix, and calculated the Euclidean norm of the new vectors, also called loadings, represented by its columns. These norms are the contributions of the frequencies after the dimension reduction. This allowed to select sets of nucleotide pattern frequencies in the training data set.

The PCA indicated that a set of 10 features could explain about $\approx 65.0\%$ of data, while a set of 60 features could explain about $\approx 95.0\%$ of data. From this information, we created 6 groups of nucleotide pattern frequencies with sizes 10, 20, 30, 40, 50 and 60. The frequencies of nucleotide patterns that most contributed to the orthogonal transformation were selected to create each group. Each of these groups formed the first potential sets of features.

Second set of features regarding ORFs

In addition, four sets of features were constructed, in order to find the best set of features regarding ORFs: the first ORF length and its relative length; the first ORF relative length; the longest ORF length and its relative length; and the longest ORF relative length.

Implementation

To implement the SVM method [119], a libSVM package [120] was used.

In order to find the best set of features, we combined the 6 sets of features found with the PCA ($10, 20, 30, 40, 50$ and 60 frequencies of nucleotide patterns) and the 4 sets of features described above (the first and the longest ORF lengths and their relative lengths), thereby creating 24 experiments.

In these 24 experiments, the grid search tool[1] with 10-fold cross validation was used in the training data set, to define which experiment performed best. In each experiment, the best C and γ parameters were selected.

Case studies

Four case studies were performed to evaluate the SVM method. We validated all the models created with species different from those used in the training phase, according to the following data sets: rat (*Rattus norvegicus*) assembly Rnor6.0, pig (*Sus Scrofa*) assembly Sscrofa10.2, and fruitfly (*Drosophila melanogaster*) assembly BDGP6. We also applied the models to human and mouse pseudogenes. In addition to this, we re-annotated two sequences from Swiss-Prot database [135], and annotated contigs derived from RNA-seq transcripts of human, gorilla and rhesus macaque, reported in Necsulea et al [134].

4.3 Results and Discussion

4.3.1 Human

In the first case study, only human data from the assemblies GRCh37 (hg19) and the GRCh38 (hg38) were used for training and testing. Our databases included 104,763 PCTs and 24,513 lncRNAs from GRCh37, and 102,915 PCTs and 28,321 lncRNAs from GRCh38. We filtered all the sequences shorter than 200 bases, having obtained 94,830 and 92,716 PCTs from GRCh37 and GRCh38 assemblies, respectively, and 24,266 and 28,024 lncRNAs from GRCh37 and GRCh38 assemblies, respectively.

To train the models, 18,200 PCTs and 18,200 lncRNAs were used from GRCh37, and 21,018 PCTs and 21,018 lncRNAs from GRCh38. GRCh37 testing data set included 6,066 PCTs and 6,066 lncRNAs, while the GRCh38 testing data set contained 7,006 PCTs and 7,006 lncRNAs.

The 6 sets of nucleotide pattern frequencies selected with PCA were used to identify which one produced the best results. To do this, we used two ROC curves (see Figures 4.2 and 4.3). These figures show the results of the models trained with the first ORF relative length and the 6 nucleotide pattern sets. The curve for the model trained with 50 nucleotide frequencies performed slightly better for both assemblies, GRCh37 and GRCh38.

The nucleotide pattern frequencies that achieved the best results for the human data are shown in Table 4.1. The nucleotide pattern frequencies for both GRCh37 and GRCh38

[1] A Python script to find a model with C and γ parameters presenting the best accuracy, which is part of the libSVM package.

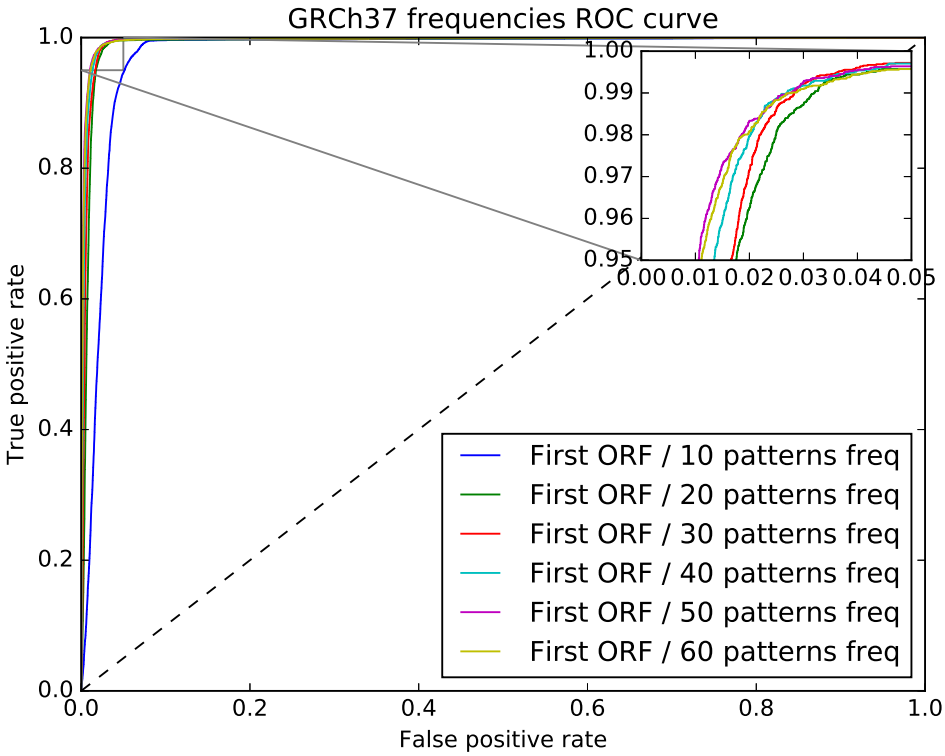

Figure 4.2: GRCh37 ROC curve used to select the set with the best nucleotide pattern frequencies. The model trained with a set composed of 50 nucleotide pattern frequencies performed slightly better than the other models.

data sets were almost equal, the only difference being, "acg" and "gta". We noted that both patterns are among the lowest PCA loadings, compared to all the other patterns.

Using these patterns, together with the first and longest ORF relative lengths as features, we trained 8 models with two kernels, radial and quadratic, having tested them with both data sets, GRCh37 and GRCh38. The results are shown in Table 4.2 and Figures 4.4 and 4.5. The quadratic kernel achieved substantial accuracy in almost all the tests, while the radial kernel achieved very high accuracy in all of them.

In other results, the difference of the ORF relative length was very small when using the first and the longest ORF relative lengths. Although we were able to achieve very close values of accuracy, the first ORF relative length model presented higher sensitivity than the longest one. In addition, finding the first ORF ($O(n)$) has a lower time complexity when compared to finding the longest ORF ($O(n^2)$). From a biological point of view, the canonical model for translation initiation is the scanning model of the ribosome, which is finding the initial "atg" codon [136]. It is worth noting that, in our data sets, in $\approx 94\%$ of the lncRNAs, the first ORF was different from the longest one, while in $\approx 93\%$ of the PCTs, the first and the longest ORFs were the same. Using only this characteristic, we

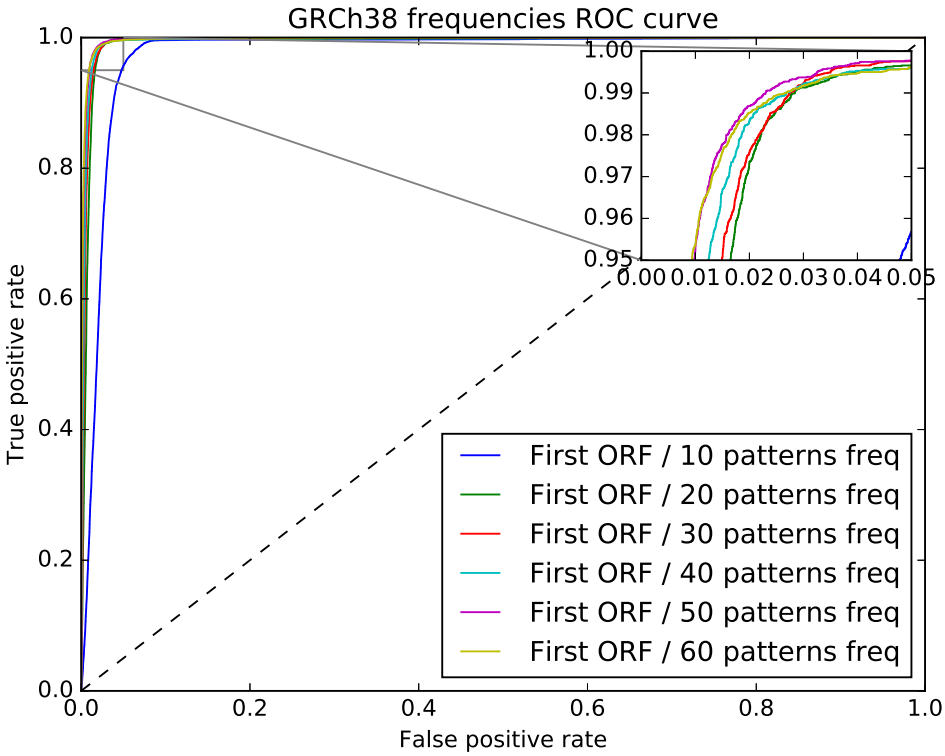

Figure 4.3: GRCh38 ROC curve used to select the set with the best nucleotide pattern frequencies. The model trained with a set composed of 50 nucleotide pattern frequencies performed slightly better than the other models.

Table 4.1: Selected nucleotide pattern frequencies for the human data. GRCh37 and GRCh38 data sets were analyzed to identify 50 pattern frequencies with the highest PCA loadings. The patterns "acg" and "gta", in bold, are the only difference. In the additional files, we listed these nucleotide pattern frequencies, ordered by PCA loadings.

	GRCh37	GRCh38
1	aa, aaa, ac, aca, **acg**	aa, aaa, ac, aca, act
2	act, ag, aga, at, ata	ag, aga, at, ata, atc
3	atc, atg, att, ca, caa	atg, att, ca, caa, cac
4	cac, cag, cat, cc, cca	cag, cat, cc, cca, ccc
5	ccc, cg, cgc, ct, cta	cg, cgc, ct, cta, ctc
6	ctc, ctg, ga, gac, gag	ctg, ga, gac, gag, gc
7	gc, gcg, gg, ggg, gt	gcg, gg, ggg, gt, **gta**
8	gtc, gtg, ta, tac, tag	gtc, gtg, ta, tac, tag
9	tat, tc, tca, tct, tg	tat, tc, tca, tct, tg
10	tga, tgt, tt, ttg, ttt	tga, tgt, tt, ttg, ttt

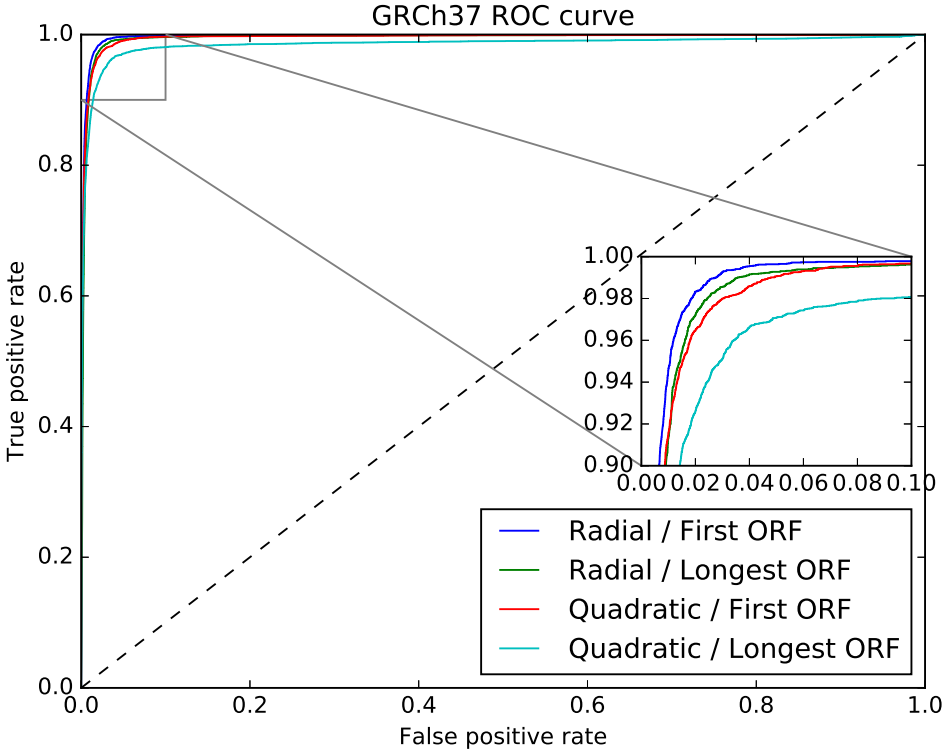

Figure 4.4: GRCh37 ROC curve used to select the best ORF relative length, and the kernel. The model trained with the first ORF relative length and the radial kernel obtained better results.

built a deterministic classifier to distinguish lncRNAs from PCTs, and compared it with the best SVM model (Figure 4.6). This classifier achieved $\approx 93.5\%$ accuracy. Thus, we decided to use the first ORF relative length as a feature in our models.

The contribution of each feature set was also investigated (Figures 4.7 and 4.8). Each feature set can also distinguish lncRNAs from PCTs with high confidence. The model using only the first ORF achieved 92.90% accuracy in GRCh37 data set and 92.95% in GRCh38 data set, while the model using only the 50 frequencies of nucleotide pattern achieved an accuracy of 90.86% and 91.54%, respectively. These results confirm that ORF content is a key characteristic, as reported in the literature, and, also show that other features, such as sets of nucleotide pattern frequencies, can achieve similar performance in distinguishing lncRNAs from PCTs. However, we found that combining all the features in one model presented better results.

We compared our results with the methods and results presented by Sun et al. [109, 137], Han et al. [61], Pian et al. [60] and Wucher et al. [63], as shown in Table 4.3. Note that, in these comparisons, the same human assemblies were used. These results show that, in all the chosen metrics, our method presented better results.

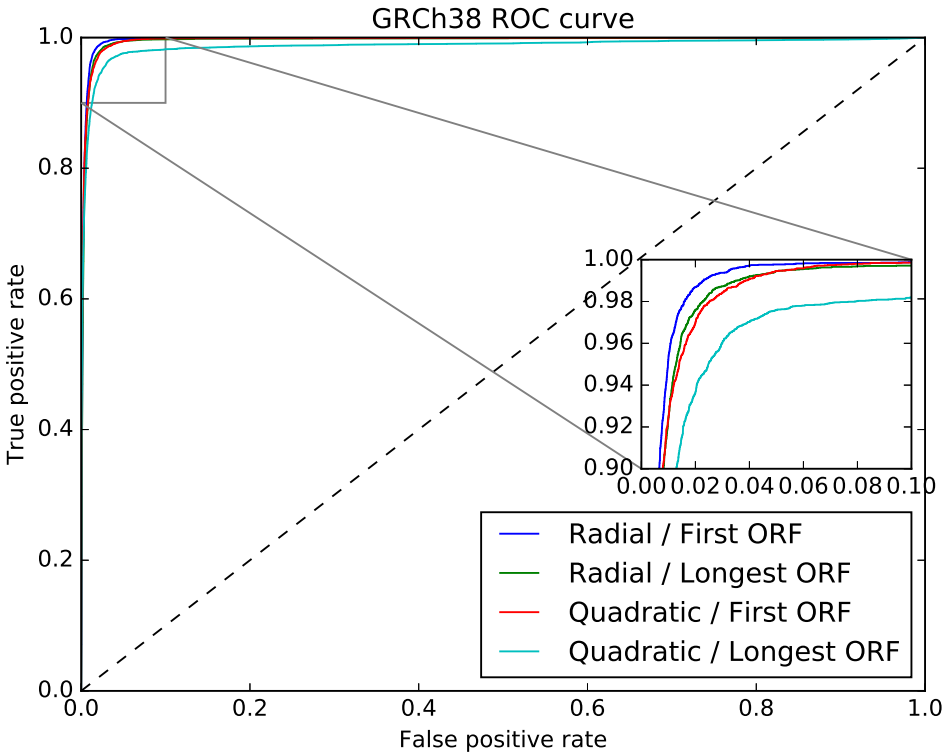

Figure 4.5: GRCh38 ROC curve used to select the best ORF relative lenght, and the kernel. The model trained with the first ORF relative length and the radial kernel obtained best results.

DeepLNC of Tripathi et al. [62] presented almost the same results, when compared to our method. In contrast to the other methods, we did not execute any experiment directly, since DeepLNC uses the lncipedia database [138], and does not clearly indicate the negative data set. We also attempted to use their method with our data set, but the web application (`http://bioserver.iiita.ac.in/deeplnc/`) presented an exception when submitting a fasta file, and failed to report any results. Notably, 98.21% of all the lncRNAs of the lncipedia database were correctly classified by our method.

Moreover, in order to verify the performance of our method in a highly curated set of lncRNAs and PCTs, we selected the best trained model to classify human data, the one trained with data from GRCh37, with 50 PCA selected nucleotide pattern frequencies and the first ORF relative length. This model was used to classify the highly curated data set of 5.322 lncRNAs reported by Nitsche et al. [139] and 5.322 PCTs randomly chosen from the Swiss-Prot reviewed database [135], but not including those annotated as *putative*, *hypothetical*, *unknown* and *predicted*. The model analyzed this data set with 96.15% accuracy, 99.72% sensitivity $(5,307/5,322)$ and 92.58% specificity $(4,927/5,322)$.

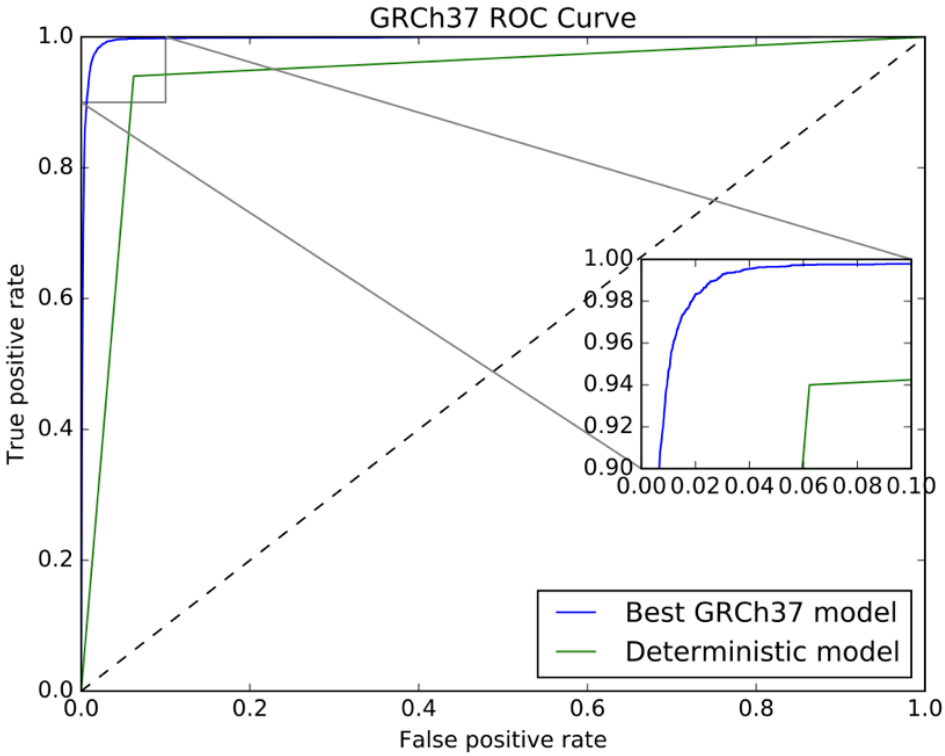

Figure 4.6: GRCh37 ROC curve showing the performance of the deterministic classifier compared to the best SVM model.

4.3.2 Mouse

For the second case study, we used mouse transcript data, from the GRCm38 assembly, with 61,440 PCTs and 11,511 lncRNAs. Again, we removed all the sequences shorter than 200 nucleotides, which resulted in 57,191 PCTs and 11,347 lncRNAs. This data was randomly split in two data sets, a training data set with 8,510 PCTs and 8,510 lncRNAs, and a testing data set with 2,837 PCTs and 2,837 lncRNAs.

Models with the 6 nucleotide pattern sets together with the first ORF relative length were also used to find which set would perform better. The ROC curve in Figure 4.9 shows that the model trained with 50 nucleotide frequencies performed better than the other models. The nucleotide pattern frequencies that achieved the best results for the mouse data are shown in Table 4.4.

Similar to the human case, using these nucleotide pattern frequencies, we also analyzed models trained with radial and quadratic kernels, using the first and the longest ORFs, as well as absolute and relative lengths. Analyzing the results, shown in Table 4.5 and in Figure 4.10, we found that the best model was trained using the radial kernel, with features of the set of 50 frequencies of nucleotide patterns and the first ORF relative

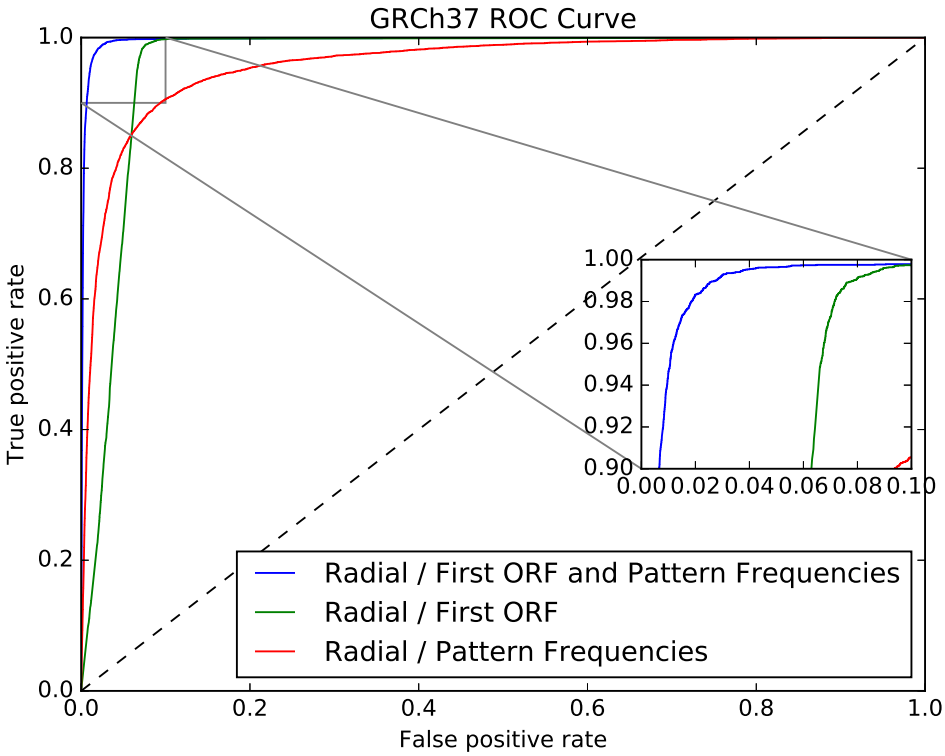

Figure 4.7: GRCh37 ROC curve showing the performance of a model trained with the first ORF relative length only, another model trained with the 50 selected nucleotide patterns frequencies, and a third model using all these features.

length.

Again, the contribution of each feature category was investigated (Figure 4.11). The model using only the first ORF achieved 93.52% accuracy, while the model using only the 50 frequencies of nucleotide patterns achieved an accuracy of 90.68%. Once more, ORF content is confirmed as a determinant characteristic, as well as a set of nucleotide pattern frequencies that achieved similar performance, to distinguish lncRNAs from PCTs. However, we found that combining all the features in one model improved performance.

The comparison of our results with those obtained by Sun et al. [109, 137], Han et al. [61], Pian et al. [60] and Wucher et al. [63] (see Table 4.6), shows that our method achieved better sensitivity and accuracy than the other methods, although the specificity was 1.41% lower than CPC, despite a 23.24% higher sensitivity in this case. Therefore, our method presented better performance in distinguishing lncRNAs from PCTs in mouse transcript data, when compared to the other tools.

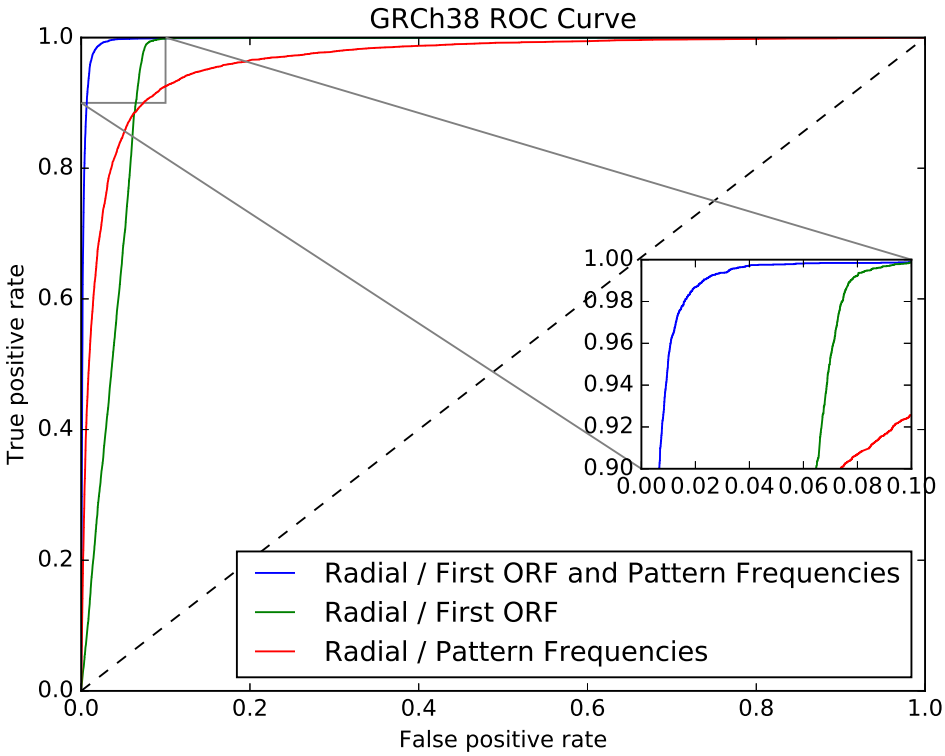

Figure 4.8: GRCh38 ROC curve showing the performance of a model trained with the first ORF relative length only, another model trained with the 50 selected nucleotide patterns frequencies, and a model using all these features.

4.3.3 Human and Mouse

This case study was analyzed to verify if a cross species model would better distinguish lncRNAs from PCTs than the previously tested single species models.

In this case study, we used the same training and testing data from the previous case studies to build the training and testing data sets. We combined data from GRCh37 with GRCm38 and from GRCh38 with GRCm38.

First, we selected the 50 nucleotide pattern frequencies to build the models (see Table 4.7). The least significant patterns (lowest PCA loading), "cca" and "gac", were the only differences in these sets.

Using these patterns, we trained models with the first and the longest ORF relative lengths. The results are shown in Table 4.8.

We noticed a small improvement in accuracy using the bi-species model, when compared to the single species model. These results suggest that a multi-species model can slightly improve distinguishing lncRNAs from PCTs, when compared to a single species model.

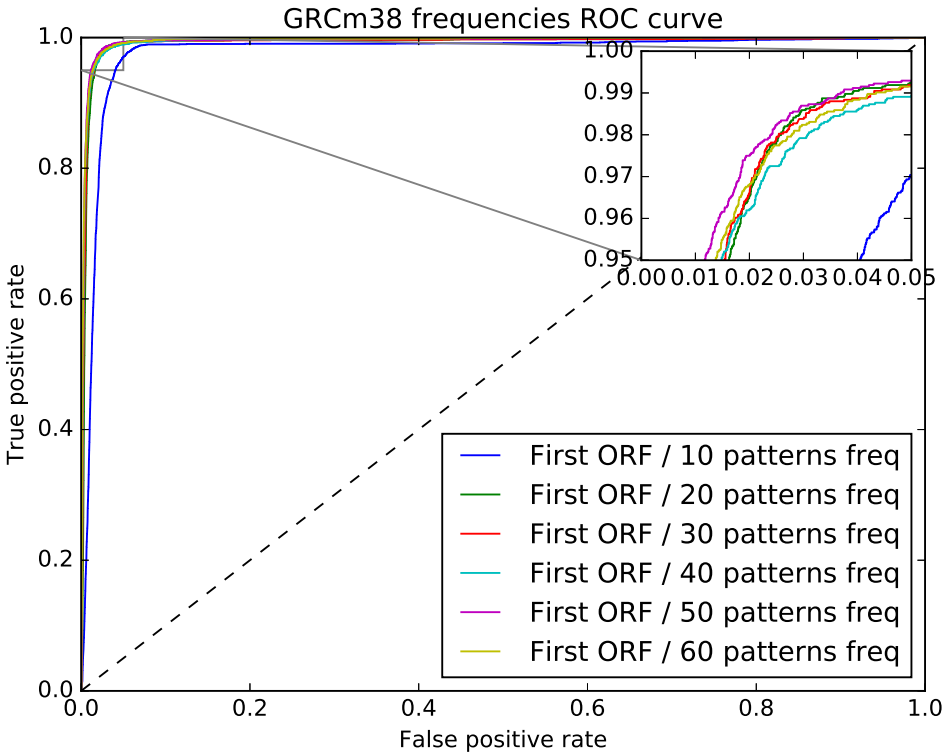

Figure 4.9: GRCm38 ROC curve used to select the best set of nucleotide pattern frequencies. The model trained with the set composed of 50 nucleotide pattern frequencies performed better than the other models.

4.3.4 Mouse and Zebrafish

The last case study was performed to evaluate our method when creating a multi-species model with data from two evolutionary distant species, together with a fewer number of annotated lncRNAs. To do this, we used mouse (GRCm38) and zebrafish (GRCz10).

The same training and testing data sets from the mouse case study were used, together with data from GRCz10, 2,775 PCTs and 2,775 lncRNAs for training, and 926 PCTs and 926 lncRNAs for testing.

The 50 nucleotide pattern frequencies selected by the PCA are shown in Table 4.9. These 50 patterns and the first ORF relative length were used to create the SVM model, which obtained the results presented in Table 4.10.

Once again, the results show that we can use the same method on different sets of species, creating a multi-species model to distinguish lncRNAs from PCTs, with high accuracy.

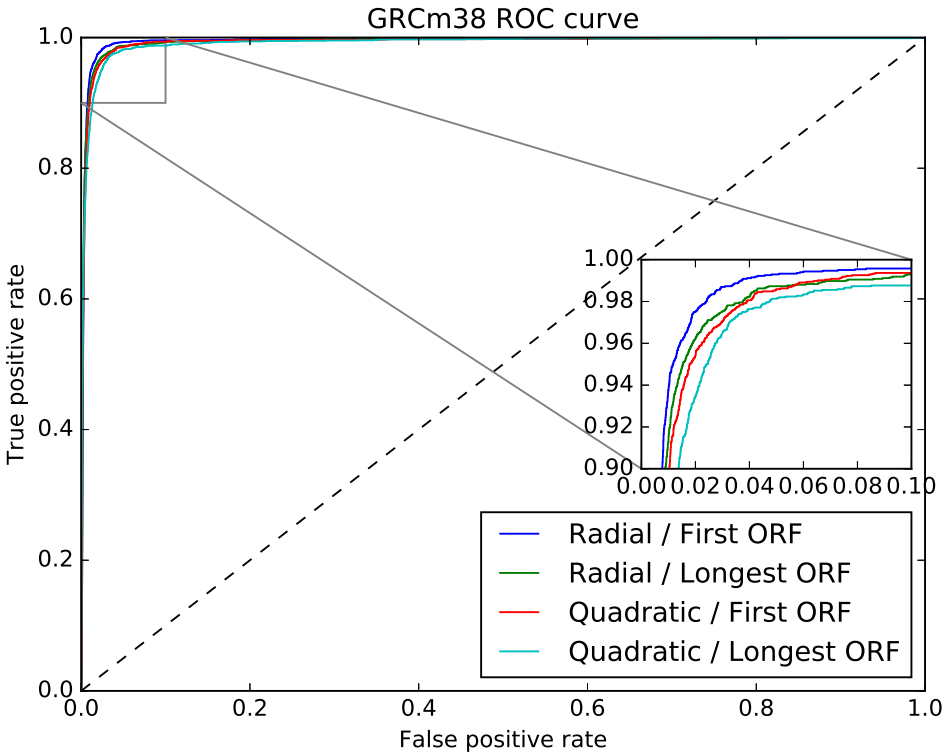

Figure 4.10: GRCm38 ROC curve used to select the best ORF relative lenght and the kernel. The model trained with the first ORF relative length and the radial kernel obtained the better results.

4.3.5 Model validation

To validate our method, we used the best model of each case study to distinguish lncRNAs from PCTs in data sets of species that were not used in the SVM training. The objective was to analyze under- and overfitting, and also whether the models could distinguish lncRNAs from PCTs in data sets of evolutionarily close and distant species.

Besides the data sets used in each case study, we used data from pig (Sscrofa10.2) - 205 lncRNAs and 205 PCTs, rat (Rnor6.0) - 3,537 lncRNAs and 3,537 PCTs, and fruit fly (BDGP6) - 2,776 lncRNAs and 2,776 PCTs. All the results are shown in Table 4.11.

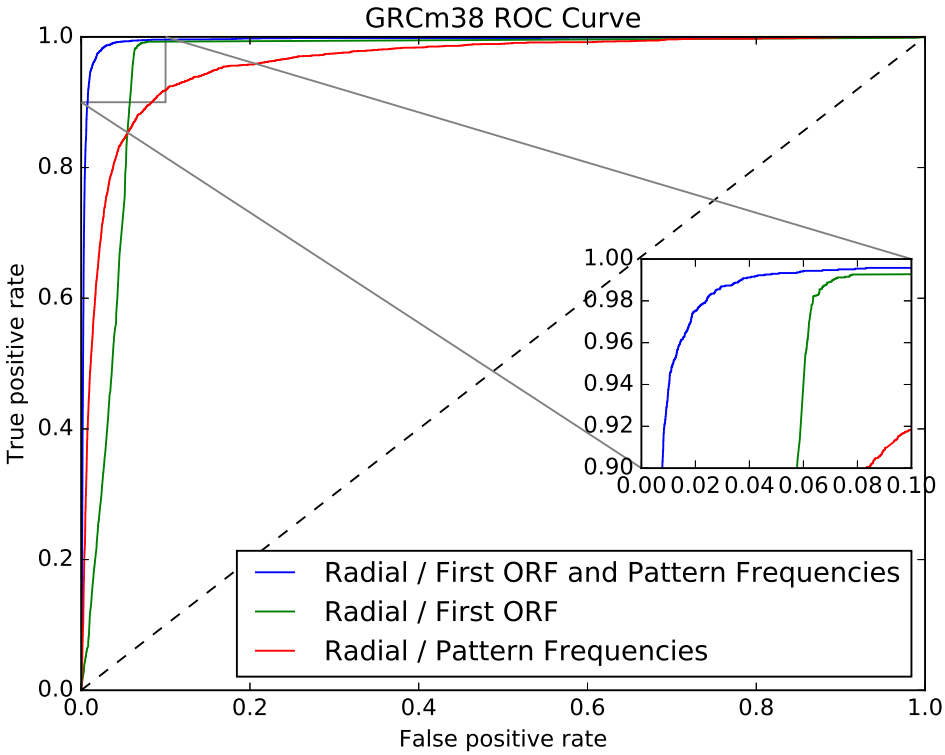

Figure 4.11: GRCm38 ROC curve showing the performance of a model trained with first ORF relative length only, another model trained with the 50 selected nucleotide patterns frequencies, and a third model using all these features

.

Table 4.2: Results of the human case study. We trained 8 models with two data sets, GRCh37 and GRCh38, to select the first, or the longest, ORF relative lengths (the length of the corresponding ORF divided by the length of the transcript). The better results for each data set are in bold.

Model \ Test data set	GRCh37	GRCh38
Radial using GRCh37 and First ORF		
Sensitivity	**98.95%**	**99.43%**
Specificity	97.41%	97.23%
Accuracy	**98.18%**	98.33%
Radial using GRCh37 and Longest ORF		
Sensitivity	98.09%	98.73%
Specificity	97.50%	97.55%
Accuracy	97.80%	98.14%
Quadratic using GRCh37 and First ORF		
Sensitivity	98.15%	98.83%
Specificity	96.60%	96.41%
Accuracy	97.38%	97.62%
Quadratic using GRCh37 and Longest ORF		
Sensitivity	94.79%	95.54%
Specificity	97.23%	97.19%
Accuracy	96, 01%	96.36%
Radial using GRCh38 and First ORF		
Sensitivity	89.86%	97.54%
Specificity	98.64%	**99.26%**
Accuracy	94.25%	**98.40%**
Radial using GRCh38 and Longest ORF		
Sensitivity	98.37%	97.63%
Specificity	97.76%	97.58%
Accuracy	98.06%	97.61%
Quadratic using GRCh38 and First ORF		
Sensitivity	80.43%	98.66%
Specificity	**98.84%**	96.78%
Accuracy	89.63%	97.72%
Quadratic using GRCh38 and Longest ORF		
Sensitivity	94.77%	95.08%
Specificity	97.66%	97.50%
Accuracy	96.21%	96.29%

Table 4.3: Results for models trained with human data. Results in bold are the best for each test data set. Note that our method produced the best results.

Method \ Test data set	GRCh37	GRCh38	NONCODE
Radial using GRCh37 and First ORF			
Sensitivity	**98.95%**	**99.43%**	**96.67%**
Specificity	97.41%	97.23%	-
Accuracy	**98.18%**	98.33%	-
Radial using GRCh38 and First ORF			
Sensitivity	89.86%	97.54%	88.75%
Specificity	**98.64%**	**99.26%**	-
Accuracy	94.25%	**98.40%**	-
CPC [a,e]			
Sensitivity	67.23%	69.90%	-
Specificity	97.62%	73.90%	-
Accuracy	82.43%	71.90%	-
CPAT [a,e]			
Sensitivity	94.60%	89.90%	-
Specificity	85.28%	92.40%	-
Accuracy	89.94%	91.20%	-
lncRScan-SVM [a]			
Sensitivity	93.88%	-	-
Specificity	89.20%	-	-
Accuracy	91.94%	-	-
iSeeRNA [b,c]			
Sensitivity	96.10%	-	-
Specificity	94.70%	-	-
Accuracy	95.40%	-	-
lncRNApred [d,f]			
Sensitivity	-	-	93.40%
Specificity	-	-	-
Accuracy	-	-	-
FEELnc [e]			
Sensitivity	-	92.30%	-
Specificity	-	91.50%	-
Accuracy	-	91.90%	-

[a] Results obtained in Han et al. [61]
[b] Results obtained in Sun et al. [109]
[c] This method was created to classify only lincRNAs
[d] Results obtained in Sun et al. [60]
[e] Results obtained in Wucher et al. [63]
[f] We only considered *sensitivity*, since the negative test data was not clearly specified in the article

Table 4.4: Selected nucleotide pattern frequencies for mouse data. GRCm38 data set was analyzed to identify the 50 pattern frequencies with the higher PCA loadings.

	GRCm38
1	aa, aaa, ac, aca, acg
2	act, ag, aga, at, ata
3	atc, atg, att, ca, caa
4	cac, cag, cat, cc, cca
5	ccc, cg, cgc, ct, cta
6	ctc, ctg, ga, gac, gag
7	gc, gcg, gg, ggg, gt
8	gtc, gtg, ta, tac, tag
9	tat, tc, tca, tct, tg
10	tga, tgt, tt, ttg, ttt

Table 4.5: Results for models trained with mouse data. Results in bold are the best ones for each test data set.

Model \ Test data set	GRCm38
Radial using GRCm38 and First ORF	
Sensitivity	**98.70%**
Specificity	96.96%
Accuracy	**97.83%**
Radial using GRCm38 and Longest ORF	
Sensitivity	97.49%
Specificity	**97.03%**
Accuracy	97.26%
Quadratic using GRCm38 and First ORF	
Sensitivity	98.38%
Specificity	95.80%
Accuracy	97.09%
Quadratic using GRCm38 and Longest ORF	
Sensitivity	96.51%
Specificity	96.99%
Accuracy	96.75%

Table 4.6: Results for models trained and tested with mouse data. Results in bold are the best ones for each test data set.

Method \ Test data set	GRCm38 (mm10)
Radial using GRCm38 and First ORF	
Sensitivity	**98.70%**
Specificity	96.96%
Accuracy	**97.83%**
CPC [a]	
Sensitivity	75.46%
Specificity	**98.37%**
Accuracy	86.91%
CPAT [a]	
Sensitivity	95.34%
Specificity	88.17%
Accuracy	91.76%
lncRScan-SVM [a]	
Sensitivity	95.29%
Specificity	89.14%
Accuracy	92.21%
iSeeRNA [b,c]	
Sensitivity	94.20%
Specificity	92.70%
Accuracy	93.45%
FEELnc [d]	
Sensitivity	94.10%
Specificity	93.80%
Accuracy	93.90%

[a] Results obtained in Han et al. [61]
[b] Results obtained in Sun et al. [109]
[c] This method was created to classify only lincRNAs
[d] Results obtained in Wucher et al. [63]

Table 4.7: Selected nucleotide pattern frequencies for human and mouse data. GRCh37, GRCh38 and GRCm38 data sets were analyzed to identify the 50 pattern frequencies with the highest PCA loadings. The patterns "cca" and "gac", in bold, are the only differences.

	GRCh37 and GRCm38	GRCh38 and GRCm38
1	aa, aaa, ac, aca, acg	aa, aaa, ac, aca, acg
2	act, ag, aga, at, ata	act, ag, aga, at, ata
3	atc, atg, att, ca, caa	atc, atg, att, ca, caa
4	cac, cag, cat, cc, **cca**	cac, cag, cat, cc, ccc
5	ccc, cg, cgc, ct, cta	cg, cgc, ct, cta, ctc
6	ctc, ctg, ga, gag, gc	ctg, ga, **gac**, gag, gc
7	gcg, gg, ggg, gt, gtc	gcg, gg, ggg, gt, gtc
8	gtg, ta, tac, tag, tat	gtg, ta, tac, tag, tat
9	tc, tca, tcg, tct, tg	tc, tca, tcg, tct, tg
10	tga, tgt, tt, ttg, ttt	tga, tgt, tt, ttg, ttt

Table 4.8: Results of the human and mouse case study. We trained four models with two data sets, GRCh37/GRCm38 and GRCh38/GRCm38, and also compared the selection of two attributes, first and longest ORF relative lengths. The best results for each test data set, GRCh37, GRCh38 and GRCm38, are in bold.

Model \ Test data set	GRCh37	GRCh38	GRCm38
Radial using GRCh37, GRCm38 and First ORF			
Sensitivity	**98.86%**	**99.42%**	98.51%
Specificity	97.56%	**97.69%**	97.54%
Accuracy	**98.21%**	**98.55%**	98.02%
Radial using GRCh37, GRCm38 and Longest ORF			
Sensitivity	98.05%	98.67%	97.60%
Specificity	97.53%	97.59%	97.54%
Accuracy	97.79%	98.13%	97.57%
Radial using GRCh38, GRCm38 and First ORF			
Sensitivity	91.22%	99.24%	**98.66%**
Specificity	**98, 65%**	97.46%	97.41%
Accuracy	94.93%	98.35%	**98.03%**
Radial using GRCh38, GRCm38 and Longest ORF			
Sensitivity	98.31%	98.20%	98.23%
Specificity	97.83%	97.63%	**97.74%**
Accuracy	98.07%	97.91%	97.98%

Table 4.9: Selected nucleotide pattern frequencies from mouse and zebrafish. GRCm38 and GRCz10 data sets were analyzed to identify the 50 pattern frequencies with the highest PCA loadings.

	GRCm38 and GRCz10
1	aa, aaa, ac, aca, acg
2	act, ag, aga, at, ata
3	atc, atg, att, ca, caa
4	cac, cag, cat, cc, cca
5	ccc, cg, cgc, ct, cta
6	ctc, ctg, ga, gag, gc
7	gcg, gg, ggg, gt, gtc
8	gtg, ta, tac, tag, tat
9	tc, tca, tcg, tct, tg
10	tga, tgt, tt, ttg, ttt

Table 4.10: Results for the mouse and zebrafish case study. We trained one model with two data sets, GRCm38 and GRCz10.

Test data set / Model	GRCm38	GRCz10
Radial using GRCm38, GRCz10 and First ORF		
Sensitivity	98.56%	97.19%
Specificity	96.86%	95.00%
Accuracy	97.71%	96.09%

Table 4.11: Comparison of all the results for each species, together with their corresponding performances. The best results for each species are in bold. In the columns are the test data set: human GRCh37 and GRCh38; mouse GRCm38; rat Rnor6.0; pig Sscrofa10.2; zebrafish GRCz10; and fruitfly BDGP6.

Model \ Test data set	GRCh37	GRCh38	GRCm38	Rnor6.0	Sscrofa10.2	GRCz10	BDGP6
Radial using GRCh37 and First ORF							
Sensitivity	**98.95%**	**99.43%**	**98.72%**	94.16%	78.89%	95.19%	93.17%
Specificity	97.41%	97.23%	97.04%	94.90%	89.28%	95.23%	99.78%
Accuracy	98.18%	98.33%	97.88%	94.53%	84.08%	95.21%	96.47%
Radial using GRCh38 and First ORF							
Sensitivity	89.86%	97.54%	90.07%	78.13%	55.28%	74.68%	80.87%
Specificity	**98.64%**	**99.26%**	98.51%	**97.89%**	95.93%	98.45%	99.91%
Accuracy	94.25%	98.40%	94.29%	88.01%	75.60%	86.56%	88.67%
Radial using GRCm38 and First ORF							
Sensitivity	98.50%	98.90%	98.70%	93.85%	**79.40%**	95.14%	94.31%
Specificity	97.09%	96.93%	96.96%	94.91%	**89.43%**	94.70%	99.96%
Accuracy	97.79%	97.91%	97.83%	94.38%	**84.41%**	94.92%	96.97%
Radial using GRCh37, GRCm38 and First ORF							
Sensitivity	98.86%	99.42%	98.51%	93.11%	76.38%	94.62%	91.30%
Specificity	97.56%	97.69%	**97.54%**	95.39%	89.94%	95.63%	99.76%
Accuracy	**98.21%**	**98.55%**	98.02%	94.25%	83.16%	95.12%	95.53%
Radial using GRCh38, GRCm38 and First ORF							
Sensitivity	91.22%	99.24%	98.66%	81.00%	55.28%	77.17%	74.95%
Specificity	98.65%	97.46%	97.41%	97.81%	95.85%	**98.74%**	**99.92%**
Accuracy	94.93%	98.35%	**98.03%**	89.40%	75.56%	87.95%	87.43%
Radial using GRCm38, GRCz10 and First ORF							
Sensitivity	98.71%	99.10%	98.56%	**94.64%**	75.89%	**97.19%**	**98.57%**
Specificity	96.89%	96.72%	96.86%	94.69%	89.87%	95.00%	99.65%
Accuracy	97.80%	97.91%	97.71%	**94.67%**	82.88%	**96.09%**	**99.11%**

From these results, we can see that none of the models are overfitted, since they were able to be applied to different species with high accuracy. The models that used GRCh38 data led to worse performance for evolutionarily distant species, especially when compared to models that used data from GRCh37. The newly 3,808 annotated lncRNAs probably contribute to a model more fitted to evolutionarily close species.

The pig data set obtained the worst classification. These results could be explained by the small number of sequences in the data set, and also by the fact that this is not a model organism, so possibly this data is not curated enough. Nonetheless, our method can be used to improve the quality of lncRNA annotation in this species.

On the other hand, it is interesting to note that a multi-species model can improve the accuracy when compared to a single species model, as can be seen in Table 4.11. The accuracy was slightly improved when the GRCh37/GRCm38 model was used to distinguish lncRNAs from PCTs in the human GRCh37 data set. Interestingly, a model created with two evolutionary distant species - mouse and zebrafish - was able to distinguish lncRNAs of the fruit fly, which is an even more distant species.

Finally, we used human and mouse pseudogenes (in GTF files), having predicted 81.2% (12,033 from a total of 15,494) pseudogenes of the human genome, and 91.7% (6,832 from a total of 7,453) pseudogenes of the mouse genome. It is remarkable that there is such a large number of predicted pseudogenes as lncRNA, since pseudogenes are derived from ancient PCTs, and diverge slowly after their generation, losing coding capacity and potential regulatory signal [140]. Nevertheless, our method distinguishes pseudogenes from *bona fide* PCTs.

4.3.6 PCTs re-annotation and RNA-seq annotation

The GRCh38 model was used to search for lncRNAs among *putative*, *hypothetical*, *unknown* and *predicted* human PCTs in the Swiss-Prot reviewed database [135]. We found 1,245 sequences longer than 67 animo-acids (201 bases). To find the corresponding nucleotide sequences, we used the EMBL reference of each entry of the Swiss-Prot database. All these sequences were trimmed, in order to begin with a start codon, because we found sequences that were 5′ UTR long. This avoids introducing bias by the first ORF relative length in the discrimination between lncRNAs and PCTs. Our method found 231 candidates. From these, we focused in 21 candidates - those that had more than a 90% probability of being lncRNA, and shorter than 2,000 bases. After analyzing the EMBL and Swiss-Prot databases and the sequences themselves, we found 2 putative PCTs with multiple "atg" at the 5' UTR, and also with annotation warnings about *dubious prediction*. Thus, both sequences could be re-annotated as lncRNAs with high probability.

In addition, we also used transcripts derived from RNA-seq data to validate our model against annotated lncRNAs, as reported by Necsulea et al. [134]. They presented 11,890, 912 and 12,056 lncRNAs from human, gorilla (*Gorilla gorilla*) and rhesus macaque (*Macaca mulatta*), respectively. Our GRCh37 model correctly classified 11,726 (98.62%) , 737 (80.81%) and 11,086 (91.95%) lncRNAs from human, gorilla and rhesus macaque, respectively.

4.4 Conclusion

In this article, we presented an SVM based method to distinguish long non-coding RNAs (lncRNAs) from protein coding transcripts (PCTs), using features from the nucleotide patterns (frequencies of di-, tri- and tetra-nucleotides) of transcripts, chosen with the support of Principal Component Analysis (PCA), together with ORF length and ORF relative length.

We trained and tested our method with data of human, mouse and zebrafish, obtaining high performance. The best results were an accuracy of 98.18% with human transcripts, 97.83% with mouse transcripts and 96.09% with zebrafish transcripts. We compared our results with other methods in the literature (CPAT, CPC, iSeeRNA, lncRNApred, lncRScan-SVM and FEELnc) and found we had obtained better results.

To validate our model, we first classified the mouse data with the human model, and vice-versa, obtaining accuracy of $\approx 97.8\%$ in both cases, showing that our model is not overfitted, and can be used with evolutionarily close species. We also validated the multi-species models human/mouse and mouse/zebrafish, which also produced excellent results. Next, we tested our models with data from rat, pig and fruit fly, having obtained accuracies from 84% to $\approx 99\%$ in all these organisms. Our method classified 81.2% of human pseudogenes and 91.7% of mouse pseudogenes as non-coding, and also found 2 uncharacterized sequences, among $1,245$, in the Swiss-Prot reviewed database, indicating a high probability of being lncRNAs. Furthermore, the method successfully annotated the majority of the assembled transcripts derived from RNA-seq data from human $(98,62\%)$, gorilla $(80,81\%)$ and rhesus macaque $(91,95\%)$.

We intend to investigate if a semi-supervised learning method could reduce the size of the training data sets, while simultaneously maintaining high accuracy in the testing phase. This could be very useful to train models for organisms with a small amount of known lncRNA transcripts. Lastly, novel features (see Ventola et al. [128]) could be used in machine learning methods, also indicating potential biological characteristics of lncRNAs.

4.5 Availability of data and materials

All data sets analyzed during the current study are available in the Ensembl Database [9]: http://www.ensembl.org.

The trained models and the program to predict long non-coding RNA are available at: https://github.com/hugowschneider/longdist.py

Chapter 5

A semi-supervised learning method to distinguish long non-coding RNAs from protein coding transcripts

In this chapter, we present the article submitted to Journal of Bioinformatics and Computational Biology `http://www.worldscientific.com/worldscinet/jbcb`.

5.1 Introduction

High-throughput sequencing projects[29, 30, 31, 8] have been generating a large amount of genomic data, including protein coding sequences (PCTs), non-coding RNAs (ncRNAs), and unannotated or uncharacterized sequences. PCTs are usually identified by methods based on sequence comparison (or alignment), which methods allow fast annotation that transfer functions already known of similar sequences[141]. NcRNAs are a high heterogenous group of sequences, ranging in length from 20 bases, e.g., snoRNAs that can be identified through C/D and H/ACA boxes in their sequences [96, 97], to hundreds of bases[36, 37], e.g., long non-coding RNAs (lncRNAs), which can be spliced and processed very similar to a messenger RNA (mRNA).

Identification and classification of lncRNAs, defined as transcripts longer than 200 bases and without apparent coding capabilities[41, 33, 42] is still a challenge, although there are many methods to predict lncRNAs. Chromatin-associated long *intergenic* ncRNAs (lincRNAs)[43] and subgroups involved in transcriptional and post-transcriptional regulation[44, 45, 46] have been identified in high throughput analyses and some of them also in wet labs. Besides, an extensive literature shows that lncRNAs are related to a wide array of diseases[47, 48, 49, 50]. Although the molecular mechanisms underlying lncRNA action are still largely unknown.

From a computational point of view, machine learning techniques have been applied to predict lncRNAs, and some tools became available for this purpose. CPC (Coding Potential Calculator)[108] and CPAT[59] discriminate protein coding genes from ncRNAs. LncRNApred[60], lncRScan-SVM[61], DeepLNC[62] and FEELnc[63] predict lncRNAs

among another sequences. A Support Vector Machine (SVM) based method that uses Principal Component Analysis (PCA) to find sets of k-mers that can be used to discriminate lncRNAs from PCTs was presented by Schneider et al.[**?**] being successful to address the problem in a variety of species.

In addition, different databases containing lncRNAs are known, and Guo et al.[142] and Fritah et al.[129] give detailed reviews. Among them, Ensembl[9], NONCODE v. 4.0[130], lncRNAdb[114], PLncDB[115], NRED[110] have information of general and specific lncRNAs, while DIANA-LncBase[131] and lncRNADisease[132] provide interactions among lncRNAs and other ncRNAs or proteins.

In general, these tools and methods use information of annotated transcripts to build models capable of predicting lncRNAs. Although working well, all of them need a considerable amount of annotated sequences to build these models, with 50% to 75% of data for training, and none of them use unannotated sequences to enhance the models.

In this context, we present a method to discriminate lncRNAs from PCTs based on semi-supervised learning techniques, using features extracted from transcripts' sequence, as described by Schneider et al.[**?**]: frequencies of nucleotide patterns selected by PCA[117]; and relative length of open reading frames (ORFs). Also, the method presented in this article uses a small amount of classified data (this occurs in many sequencing projects), together with a large number of non classified data to enhance the distinction of lncRNAs and PCTs. Besides, in order to analyze the performance of our method, we developed case studies with data of three organisms - human, mouse and zebrafish. After, we compared both the results and the amount of required data of our method to other tools found in the literature. To validate our model, we applied it to three different species, as well as to pseudogenes of human and mouse. Finally, we annotated transcripts derived from RNA-seq data, as reported in Necsulea et al.[134].

5.2 Materials and Methods

Six data sets to build the models were created from data obtained from Ensembl[9], and from transcripts derived from RNA-seq provided by Necsulea et al.[134]: human (*Homo sapiens*) assemblies GRCh37 patch 13 (hg19, GENCODE 19) and GRCh38 patch 10 (hg38, GENCODE 26), mouse (*Mus musculus*) assembly GRCm38 patch 5 (mm10, GENCODE M13), platypus (*Ornithorhynchus anatinus*) assembly OANA5, chicken (*Gallus gallus*) assembly Gallus_gallus-5.0, opossum (*Monodelphis domestica*) assembly monDom5, orangutan (*Pongo abelii*) assembly PPYG2, and xenopus (*Xenopus tropicalis*) assembly JGI 4.2. These transcript FASTA files contain PCT and lncRNA sequences, while the classification was extracted from the transcript biotype, provided by Ensembl and from the annotation provided by Necsulea et al.[134].

5.2.1 Semi-supervised learning method

The proposed method is based on the method proposed by Schneider et al. [**?**] where the main difference is in the machine learning technique applied in this work. Different semi-supervised machine learning methods, Transductive Support Vector Machines (TSVM)[143], Label Propagation and Label Spreading[58], were selected to drastically

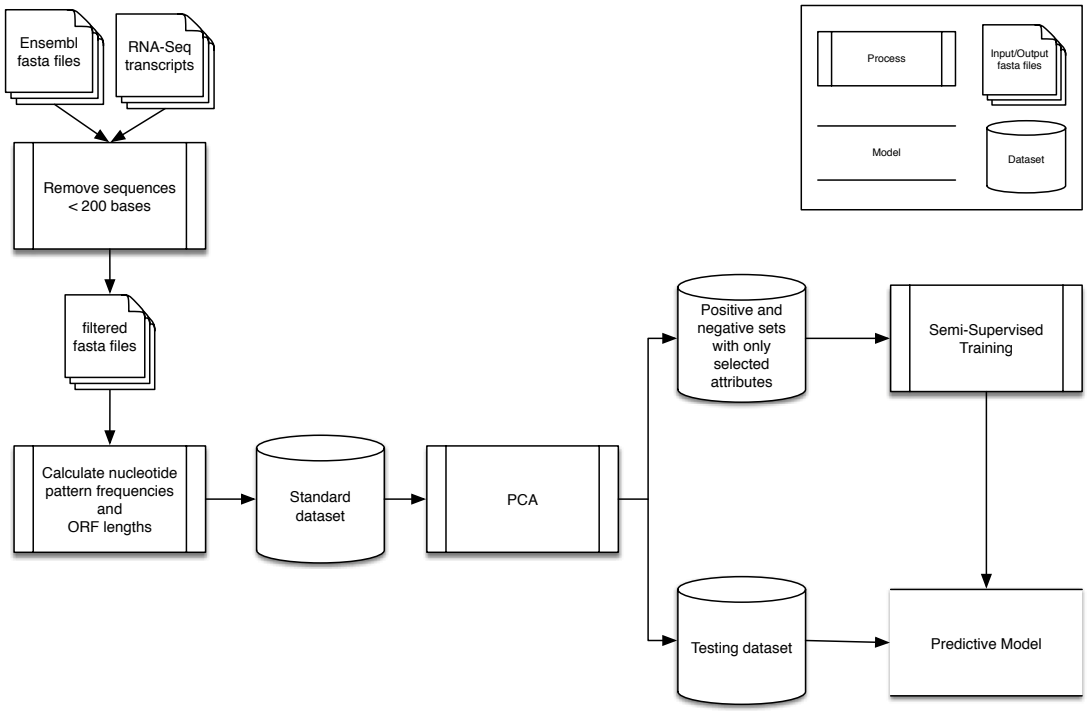

Figure 5.1: Semi-supervised method to distinguish lncRNAs from PCTs

reduce the size of the training dataset and still maintain excellent performance measures. Figure 5.1 presents our method to distinguish lncRNAs from PCTs using PCA[117] to reduce the amount of features (k-mers) calculated from transcripts' nucleotides, as described by Schneider et al.[**?**], and a semi-supervised learning technique to classify the sequences.

Standard data set

Our standard data set contains all the transcripts longer than 200 bases, from those obtained at the original FASTA files. Each entry of our standard dataset is composed of the description and nucleotide sequence of the transcript, the relative length of the first open reading frame (ORF), and the average frequencies of each nucleotide patterns.

As a result, 50 nucleotide pattern frequencies were selected using PCA[**?**] and the data, with only these selected frequencies and with the relative ORF length, were separated in two classes. The positive class is built with the lncRNAs, and the negative with the PCTs. The unclassified data, randomly selected for each experiment, are those transcripts that are neither lncRNAs nor PCTs.

Training data sets

Training data sets were created based on the assumption that there are a few classified examples and lots of unclassified data. The classified and unclassified examples in the

training data sets were selected as follows: one positive (lncRNA) and one negative (PCT) example (two classified examples) for 20 unclassified examples. The classified/unclassified data ratio were selected to fit all datasets. We variated the amount of positive and negative samples, by selecting 1, 10, 50 and 100 samples of each class. These amounts were selected from the smallest classified training dataset, one classified sequence from each class, to small number of classified examples that could represent any improvement in the classification process. With this, we built four different experiments to compare the performance of each corresponding semi-supervised learning method.

Selected features

The selected features for this method are based on Schneider et al.[**?**] method. The first selected feature is the first ORF's relative length of each transcript. The first ORF is defined in this method by the first start codon (ATG) found in the sequence and it does not require a stop codon.

The method also used nucleotide pattern frequencies selected using PCA. PCA was applied to the whole dataset and the frequencies of nucleotide patterns that most contributed to the PCA's orthogonal transformation, after the dimension reduction, were selected to create the group of 50 features used for classification.

Case studies

We performed two case studies to evaluate the semi-supervised method using human and mouse data. These case studies used data from Ensembl database to train and test models, and transcript derived from RNA-Seq data, as reported in Necsulea et al.[134], to validate the created model. After, we validated our method using Ensembl data and transcripts derived from RNA-Seq data of less curated species: chicken, opossum, platypus, orangutan and xenopus.

5.3 Results and Discussion

In this section, we present the results of the semi-supervised method based on two case studies using the different proposed learning algorithms. The main objective is to evaluate the performance of each algorithm and compare them to select the one with the best performance.

5.3.1 Human

In the human case study, data from GRCh37 and GRCh38 were used. The datasets contain $94,830$ PCTs and $24,266$ lncRNAs from GRCh37 and $94,044$ PCTs and $28,165$ lncRNAs from GRCh38 longer than 200 bases.

We trained 20 models for each human dataset based on the four previously described experiments (1, 10, 50 and 100 samples of each class) with five semi-supervised methods: (i) TSVM, (ii) Label Propagation with kNN kernel, (iii) Label Spreading with kNN kernel, (iv) Label Propagation with RBF kernel, (v) Label Spreading with RBF kernel. The

results of each model were compared based on accuracy, sensitivity, specificity and training and testing time.

Time comparison

Time measures were calculated using the GRCh37 models in a machine with an Intel® i7 processor and 8Gb RAM. This comparisons show that graph based methods are faster than the TSVM method, specially with the growth of the training data sets (results shown in Figure 5.2).

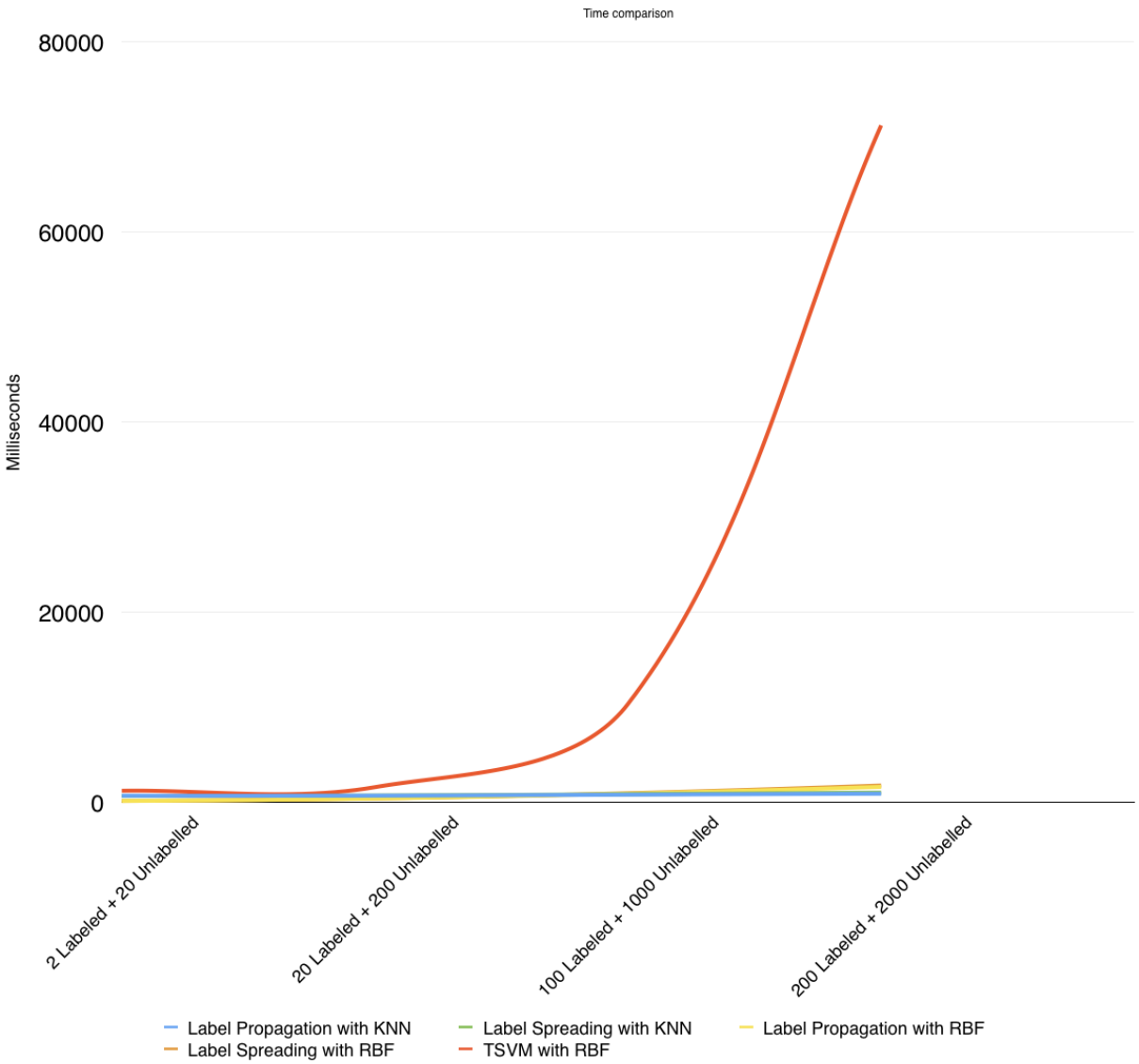

Figure 5.2: Comparison of the training and testing times of the different models

Classification

All models performed similarly with different sized training datasets except for the TSVM with only two labeled examples, which performed worst than other models. Tables 5.1

and 5.2 shows the results and Figures 5.3 and 5.4 the best models performance for the GRCh37 and GRCh38 models, respectively. Also, it is interesting to note that the TSVM is the only method that shows an improvement in the performance with the growth of training data.

Table 5.1: Comparison of the GRCh37 models

	1 Positive 1 Negative 20 Unlabeled	10 Positives 10 Negatives 200 Unlabeled	50 Positives 50 Negatives 1000 Unlabeled	100 Positives 100 Negatives 2000 Unlabeled	Average
RBF Label Propagation					
Sensitivity	**97.19**%	95.22%	97.32%	97.44%	96.79%
Specificity	92.86%	**93.25**%	**92.86**%	92.90%	**92.97**%
Accuracy	**95.03**%	94.24%	95.09%	95.17%	94.88%
kNN Label Propagation					
Sensitivity	97.02%	98.20%	98.14%	**99.23**%	98.15%
Specificity	92.92%	92.55%	92.50%	91.82%	92.45%
Accuracy	94.97%	95.38%	95.32%	95.52%	95.30%
RBF Label Spreading					
Sensitivity	95.55%	96.65%	97.48%	97.44%	96.78%
Specificity	**93.15**%	92.94%	92.80%	92.91%	92.95%
Accuracy	94.35%	94.79%	95.14%	95.18%	94.87%
kNN Label Spreading					
Sensitivity	96.63%	**98.93**%	**98.70**%	99.02%	**98.32**%
Specificity	92.95%	92.08%	92.36%	92.15%	92.39%
Accuracy	94.79%	**95.51**%	**95.53**%	**95.58**%	**95.35**%
TSVM					
Sensitivity	99.74%	87.90%	90.02%	97.09%	93.69%
Specificity	62.23%	89.52%	94.61%	**93.82**%	85.05%
Accuracy	80.99%	88.71%	92.31%	95.46%	89.37%

Methods comparison

Comparing the results of the semi-supervised methods to others methods or tools used to distinguish lncRNAs from PCTs (see Table 5.3), we observe that the semi-supervised methods achieved excellent sensitivity and accuracy, although it was not the best performing method. But, when analyzing the size of the training data sets, none of the other methods used so few annotated sequences compared to the semi-supervised method. In this scenario, we claim that the kNN label spreading method, as the best performing semi-supervised method, can be used in transcriptome projects with few annotated sequences, to improve and speed up lncRNA annotation.

RNA-Seq

We used transcripts derived from RNA-seq data to validate our method against annotated lncRNAs, as reported by Necsulea et al.[134]. The models trained with one positive and 1

Table 5.2: Comparison of the GRCh38 models

	1 Positive 1 Negative 20 Unlabeled	10 Positives 10 Negatives 200 Unlabeled	50 Positives 50 Negatives 1000 Unlabeled	100 Positives 100 Negatives 2000 Unlabeled	Average
RBF Label Propagation					
Sensitivity	**97.79**%	96.13%	97.18%	97.16%	97.07%
Specificity	92.75%	**93.07**%	**92.89**%	92.89%	**92.90**%
Accuracy	**95.27**%	94.60%	95.03%	95.03%	94.98%
kNN Label Propagation					
Sensitivity	96.70%	97.78%	**99.03**%	**99.24**%	98.19%
Specificity	**92.99**%	92.77%	92.27%	92.05%	92.52%
Accuracy	94.85%	95.28%	95.65%	**95.64**%	95.36%
RBF Label Spreading					
Sensitivity	97.51%	97.79%	97.21%	97.20%	97.43%
Specificity	92.80%	92.77%	92.86%	92.83%	92.81%
Accuracy	95.15%	95.28%	95.03%	95.01%	95.12%
kNN Label Spreading					
Sensitivity	97.02%	**98.97**%	98.91%	99.05%	**98.49**%
Specificity	92.90%	92.32%	92.39%	92.19%	92.45%
Accuracy	94.96%	**95.64**%	**95.65**%	95.62%	**95.47**%
TSVM					
Sensitivity	92.46%	94.30%	96.76%	93.76%	94.32%
Specificity	89.55%	90.68%	92.89%	**95.64**%	92.19%
Accuracy	91.01%	92.49%	94.82%	94.70%	93.26%

Table 5.3: Comparison among methods of lncRNA prediction, with the GRCh37 data set. The table shows the performances measures and the number of labeled and unlabeled sequences used for training.

	Sensitivity	Specificity	Accuracy	Training Data set Size
kNN Label Spreading	96.63%	92.95%	94.79%	**2 labeled** **20 unlabeled**
kNN Label Spreading	98.93%	92.08%	95.51%	20 labeled and 200 unlabeled
SVM-Schneider et al [**?**]	**98.95**%	97.41%	**98.18**%	36,400 labeled
CPC	67.23%	**97.62**%	82.43%	8,280 labeled
CPAT	94.60%	85.28%	89.94%	20,000 labeled
lncRScan-SVM	93.88%	89.20%	91.94%	10,000 labeled
iSeeRNA[1]	96.10%	94.70%	95.40%	30,039 labeled

negative samples, ten positive and ten negative samples, 50 positive and 50 negative samples and 100 positive and 100 negative samples classified correctly 95.48% (11352/11889), 97.64% (11609/11889), 98.77% (11743/11889) and 97.78% (11625/11889) of the annotated lncRNAs, respectively.

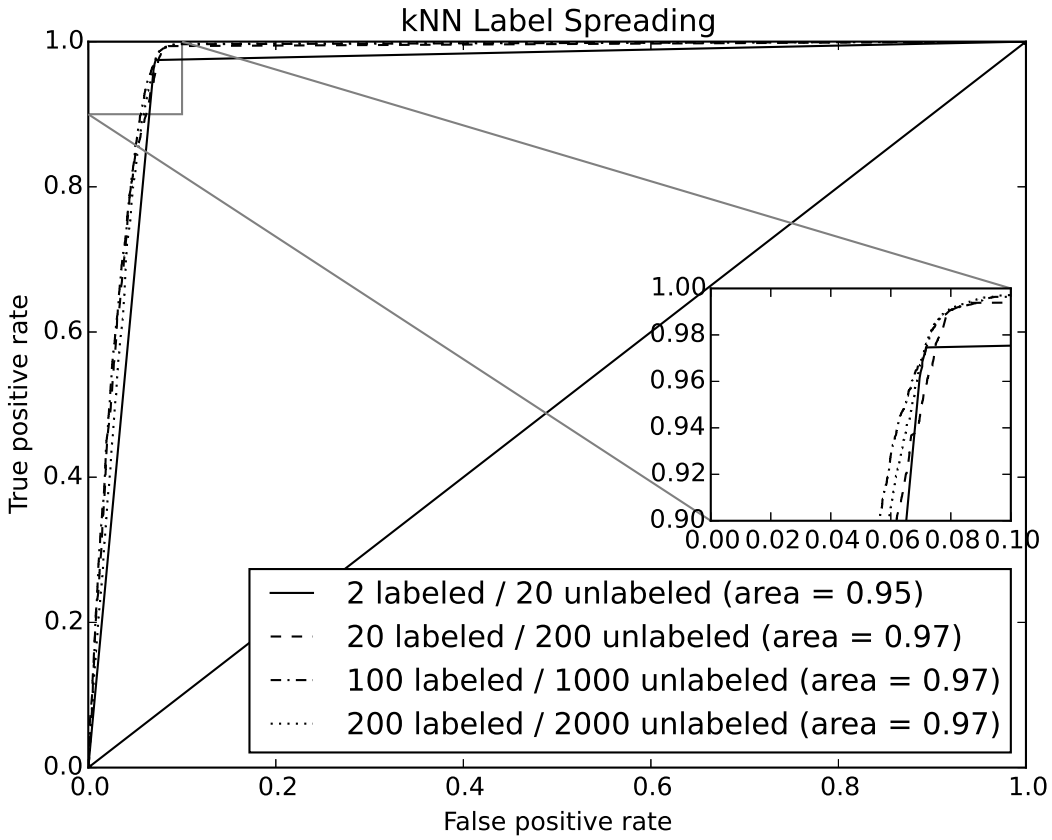

Figure 5.3: kNN label Spreading with GRCh37 data. Best performing algorithm for this dataset.

5.3.2 Mouse

Mouse dataset contains 57, 723 PCTs and 11, 842 lncRNAs longer than 200 bases, and this data were used to train 20 models to compare the five semi-supervised techniques (Label propagation with kNN and RBF kernels, label spreading with kNN and RBF kernels and TSVM).

Classification

In this case study, similar to the human case study results, all semi-supervised methods presented similar performance, except for the TSVM with the smallest training dataset. Also the only method that shows a significant improvement with the growth of the dataset is the TSVM. These result can be seen in table 5.4 and the best model performance is showed in Figure 5.5

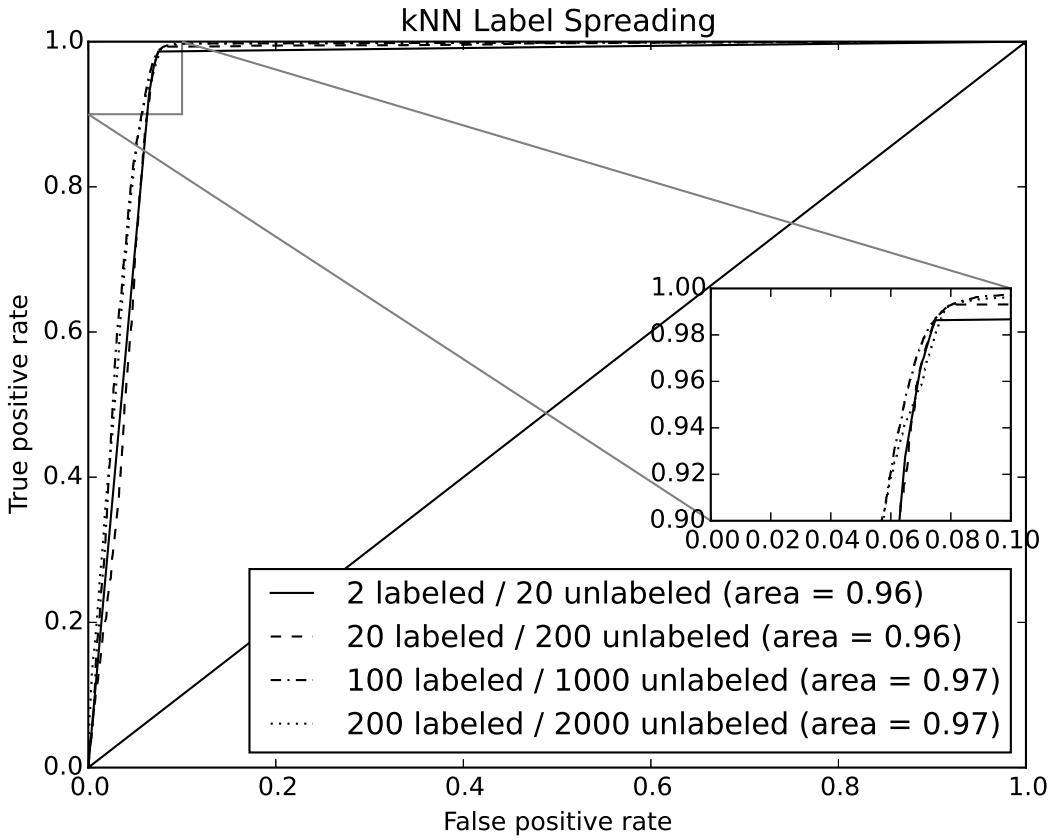

Figure 5.4: kNN label Spreading with GRCh38 data. Best performing algorithm for this dataset.

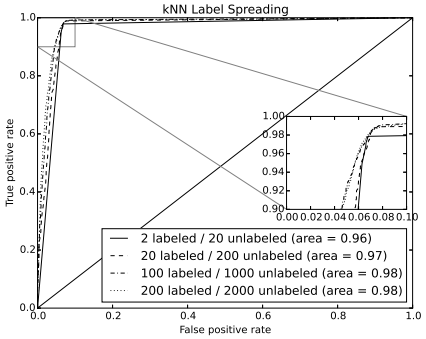

Figure 5.5: kNN label Spreading with GRCm38 data. Best performing algorithm for this dataset.

Methods comparison

Again, we compared the results of the kNN Label Spreading model results to the methods and tools available in the literature. We note that the performance are impressive when considering the size of the training dataset. Theses results are shown in table 5.5.

Table 5.4: Comparison of the GRCm38 models

	1 Positive 1 Negative 20 Unlabeled	10 Positives 10 Negatives 200 Unlabeled	50 Positives 50 Negatives 1000 Unlabeled	100 Positives 100 Negatives 2000 Unlabeled	Average
RBF Label Propagation					
Sensitivity	**97.80**%	91.75%	97.13%	98.44%	96.28%
Specificity	93.23%	94.18%	93.39%	92.83%	93.41%
Accuracy	**95.52**%	92.96%	95.26%	95.64%	94.85%
kNN Label Propagation					
Sensitivity	97.36%	93.79%	98.50%	**98.63**%	97.07%
Specificity	93.33%	93.91%	92.84%	92.80%	93.22%
Accuracy	95.35%	93.85%	95.67%	**95.72**%	95.15%
RBF Label Spreading					
Sensitivity	97.31%	96.27%	97.11%	97.45%	97.04%
Specificity	93.34%	93.52%	**93.46**%	**93.35**%	**93.42**%
Accuracy	95.33%	94.89%	95.29%	95.40%	95.23%
kNN Label Spreading					
Sensitivity	97.13%	**98.65**%	**98.81**%	98.60%	**98.30**%
Specificity	**93.38**%	92.71%	92.69%	92.72%	92.88%
Accuracy	95.26%	**95.68**%	**95.75**%	95.66%	**95.59**%
TSVM					
Sensitivity	86.35%	81.59%	94.86%	93.78%	89.15%
Specificity	90.92%	**94.36**%	90.58%	92.90%	92.19%
Accuracy	88.64%	87.97%	92.72%	93.34%	90.67%

Table 5.5: Comparison among methods of lncRNA prediction using the GRCm38 data set. The table shows the performances measures and the number of labeled and unlabeled sequences used for training.

	Sensitivity	Specificity	Accuracy	Training Data set Size
kNN Label Spreading	97.13%	93.38%	95.26%	**2 labeled** **20 unlabeled**
kNN Label Spreading	98.65%	92.71%	95.68%	20 labeled 200 unlabeled
SVM-Schneider et al [**?**]	**98.70**%	96.96%	**97.83**%	17,020 annotated
CPC	67.23%	**97.62**%	82.43%	8,280 labeled
CPAT	94.60%	85.28%	89.94%	20,000 labeled
lncRScan-SVM	93.88%	89.20%	91.94%	5,000 labeled
iSeeRNA^2	96.10%	94.70%	95.40%	16,010 labeled

RNA-Seq

Looking into the lncRNAs annotated by Necsulea et al [134], the kNN Label Spreding models trained with one positive and one negative samples, ten positive and ten negative samples, 50 positive and 50 negative samples and 100 positive and 100 negative samples classified correctly 96.46% (600/622), 98.39% (612/622), 98.39% (612/622) and 98.39%

(612/622) of the RNA-seq transcripts, respectively.

5.3.3 Validation

We investigated the performance of our method in organisms less curated with a few number of annotated lncRNAs. We built models with chicken, opossum and platypus datasets, containing 5,884 lncRNAs and 29,886 PCTs, 7,476 lncRNAs and 22,052 PCTs, and 4,007 lncRNAs and 22,973 PCTs, respectively.

We built three models using only two annotated sequences, one positive sample (lncRNA) and one negative sample (PCT) and 20 unannotated sequences from each organism to investigate the performance of the method with these organism. The results are shown in Figure 5.6 and Table 5.6.

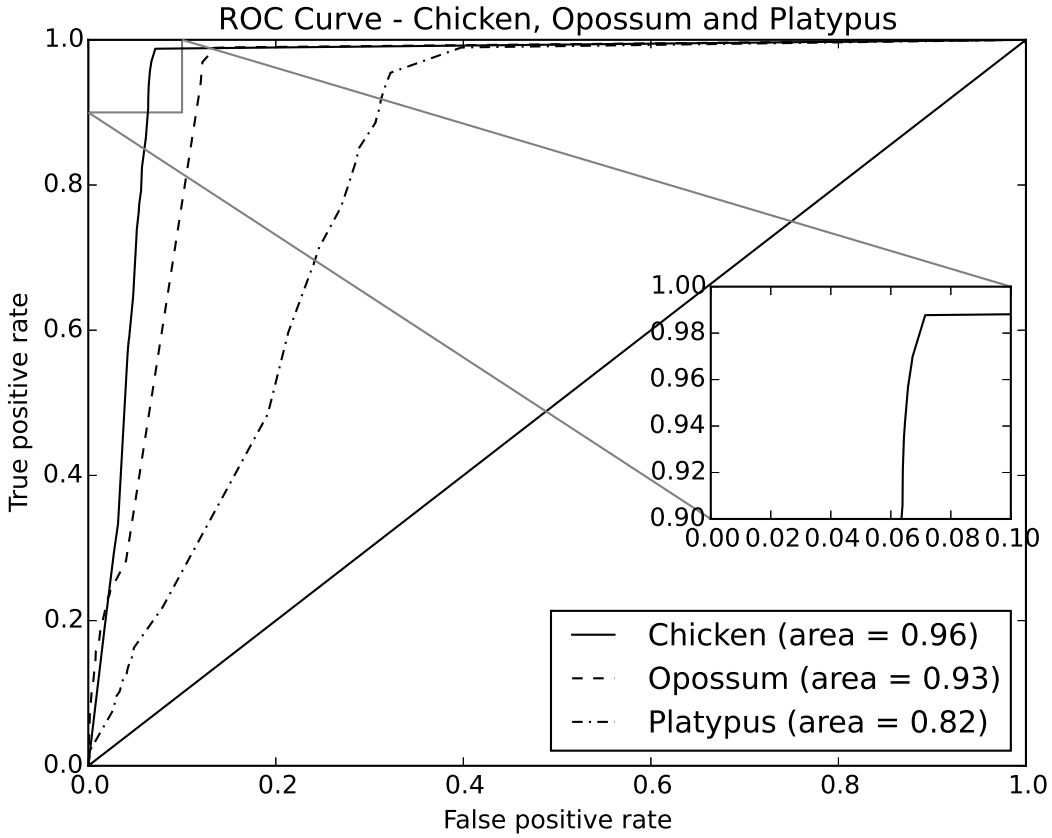

Figure 5.6: ROC curve for chicken, opossum and platypus

Table 5.6: Results for chicken, opossum and platypus

	Sensitivity	Specificity	Accuracy
Chicken	91.38%	93.60%	93.23%
Opossum	94.03%	88.02%	89.54%
Platypus	80.18%	71.27%	72.59%

Annotated lncRNAs derived from RNA-Seq data[134] were also used to test our semi-supervised model. For chicken, opossum and platypus, 97.84% (819/837), 98.31% (873/888) and 95.15% (1236/1299) lncRNAs were correctly annotated.

Besides, we built models using PCTs from the Ensembl database and lncRNAs derived from RNA-Seq data. We used $9,609$ and $3,146$ lncRNAs and $20,749$ and $22,665$ PCTs from orangutan and xenupus, respectively.

Both models were trained with two annotated sequences and 20 unannotated sequences. Results are shown in Figure 5.7 and Table 5.7.

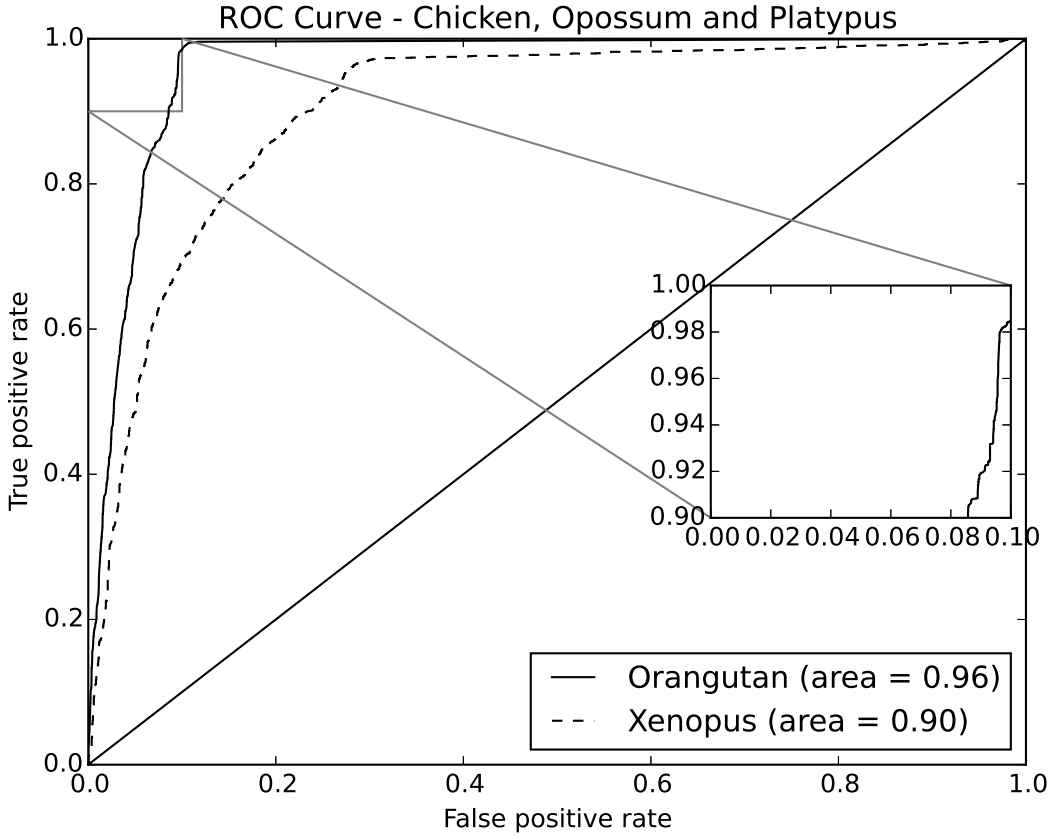

Figure 5.7: ROC curve for orangutan and xenopus

Table 5.7: Results for xenopus and orangutan

	Sensitivity	Specificity	Accuracy
Orangutan	98.35%	90.13%	94.24%
Xenopus	82.62%	82.39%	82.50%

5.4 Conclusion

In this article, we presented a semi-supervised based method to distinguish long non-coding RNAs (lncRNAs) from protein coding transcripts (PCTs), using features from transcripts' nucleotides patterns chosen with the support of Principal Component Analysis (PCA), together with ORF relative length, as described by Schneider et al [**?**].

We trained and tested our method with data of human, mouse, chicken, opossum, platypus, orangutan and xenopus having obtained a high performance with small training datasets. The models were trained with small datasets, compared to other methods and tools, obtaining high performance. The accuracy of the models using only two labelled example ranged from 72.59% with the platypus to 95.26% with the mouse. Also, the performance have not improved significantly with the growth of the dataset.

The method was validated using annotated lncRNAs transcripts derived from RNA-seq data and also showed a good performance. More than 95% of the annotated lncRNAs were correctly annotated.

We intend to investigate the biological characteristics of the selected features and how each feature contributes to the distinction. This could help reduce the amount of features and improve the understanding of lncRNAs.

Chapter 6

Conclusion

In this thesis, we studied the problem of distinguishing lncRNAs from PCTs. We proposed two computational methods based on machine learning techniques, to distinguish lncRNAs from PCTs.

The first one is a method based on SVM, using a PCA to find features that can improve the distinction. This method classified correctly 98.21% and 98.55% of the human and mouse data, respectively. We also build multi-species models which improved the classification. Also, to validate this method, we classified data from species that were not used for the model building, achieving accuracies from $\approx 84\%$ to $\approx 99\%$. LncRNAs derived from human, macaque and gorilla RNA-seq data were also used to validate the method, classifying correctly the majority of the transcripts.

The second method is based on the semi-supervised technique, and uses the procedure to find features defined in the SVM method. We trained a model using a small amount of annotated transcripts (2, 20, 100 and 200) achieving accuracies of 95.65% and 95.75% for human and mouse data, respectively. We also validated this method by classifying and training models using lncRNAs derived from RNA-seq. Chicken, opossum and platypus data were used for training and classifying RNA-seq data, achieving an accuracy of $\approx 93\%$. Orangutan and xenopus models were trained using this RNA-seq data, and $\approx 92\%$ of the data was correctly classified. This method can be used to organisms presenting an small amount of annotated lncRNAs.

6.1 Contributions

During the Doctorate, we published an article in the Brazilian Symposium on Bioinformatics (BSB 2014), focusing on a genome wide identification of long non-coding RNA in *Komagatella pastoris* [144].

We also created two methods to distinguish lncRNAs from PCTs, with excellent performance, even for organisms with a small amount of annotated lncRNAs. One method was published in the BMC Genomics, while the second one was submitted to the Journal of Bioinformatics and Computational Biology.

With this work, we developed a method using PCA for feature selection with excellent results, and two methods for lncRNA and PCT distinction. The first one based on SVM and the second using semi-supervised learning algorithms.

6.2 Future work

Next, we intend to investigate the biological meaning of the features obtained with PCA, and also the applicability of novel features (see Ventola et al. [128]) in machine learning methods. Also, an interesting way of research is the characterization of features, and methods, to discover the six classes of lncRNAs. Finally, the identification of lncRNAs in genomes is another important problem.

Referências

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, March 1998. [Online]. Available: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&amp xi, 21

[2] Itseez, "Open source computer vision library," https://github.com/itseez/opencv, 2015. xi, 23

[3] J. D. Watson and F. H. Crick, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," *Nature*, vol. 171, pp. 737–738, Abril 1953. 1

[4] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural *et al.*, "The Sequence of the Human Genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001. [Online]. Available: http://www.sciencemag.org/cgi/content/abstract/sci 1

[5] E. S. Lander and Others, "Initial Sequencing and Analysis of the Human Genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001. [Online]. Available: http://www.bibsonomy.org/bibtex/23e9396c3be637322748a1bdd44ce2fce/schaul 1

[6] W. J. Ansorge, "Next-generation dna sequencing techniques," *N Biotechnol*, vol. 25, no. 4, pp. 195–203, Apr 2009. 1

[7] R. Gentleman, *Bioinformatics and computational biology solutions using R and Bioconductor*, ser. Statistics for biology and health. New York: Springer Science+Business Media, 2005. [Online]. Available: https://libproxy.usc.edu/login?url=http://link.springer.com/10.1007/0-387-29362-0 1, 2

[8] The ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57–74, 2012. 1, 2, 29, 54

[9] F. Cunningham and co authors, "Ensembl 2015," *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D662–9, Jan 2015. 1, 18, 31, 53, 55

[10] M. S. S. Felipe, R. V. Andrade, F. B. M. Arraes, A. M. Nicola, A. Q. Maranhão *et al.*, "Transcriptional Profiles of the Human Pathogenic Fungus Paracoccidioides brasiliensis in Mycelium and Yeast Cells," *Journal of Biological Chemistry*, vol. 280, no. 26, pp. 24 706–24 714, 2005. [Online]. Available: http://www.jbc.org/content/280/26/24706.abstract 1

[11] P. Ângelo, C. Nunes-Silva, M. Brígido, J. Azevedo, E. Assunção *et al.*, "Guarana (Paullinia cupana var. sorbilis), an anciently consumed stimulant

from the Amazon rain forest: the seeded-fruit transcriptome," *Plant Cell Reports*, vol. 27, pp. 117–124, 2008, 10.1007/s00299-007-0456-y. [Online]. Available: http://dx.doi.org/10.1007/s00299-007-0456-y 1

[12] E. Lau, "Non-coding rna: Zooming in on lncrna functions," *Nat Rev Genet*, vol. 15, no. 9, pp. 574–5, Sep 2014. 1

[13] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, Jan. 2000. [Online]. Available: http://dx.doi.org/10.1093/nar/28.1.27 1

[14] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Research*, 2011. [Online]. Available: http://nar.oxfordjournals.org/content/early/2011/11/10/nar.gkr988.abstract 1

[15] P. D. Karp, C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa *et al.*, "Expansion of the BioCyc collection of pathway/genome databases to 160 genomes," *Nucleic Acids Research*, vol. 33, no. 19, pp. 6083–6089, 2005. [Online]. Available: http://nar.oxfordjournals.org/content/33/19/6083.abstract 1

[16] B. Lewin, J. E. Krebs, E. S. Goldstein, and S. T. Kilpatrick, *Lewin's Genes X*, ser. Jones and Bartlett books in computer science. Jones and Bartlett, 2009. [Online]. Available: http://books.google.com.br/books?id=0pM4KbFIEb0C 1

[17] F. Crick, "Central Dogma of Molecular Biology," *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970. [Online]. Available: http://dx.doi.org/10.1038/227561a0 1, 5

[18] F. S. Collins, A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland *et al.*, "New Goals for the U.S. Human Genome Project: 1998-2003," *Science*, vol. 282, no. 5389, pp. 682–689, 1998. [Online]. Available: http://www.sciencemag.org/content/282/5389/682.abstract 1

[19] F. Liang, I. Holt, G. Pertea, S. Karamycheva, S. L. Salzberg *et al.*, "Gene index analysis of the human genome estimates approximately 120,000 genes," *nature genetics*, vol. 25, no. 2, pp. 239–240, 2000. 1

[20] B. Ewing, P. Green *et al.*, "Analysis of expressed sequence tags indicates 35,000 human genes," *nature genetics*, vol. 25, no. 2, pp. 232–234, 2000. 1

[21] H. R. Crollius, O. Jaillon, A. Bernot, C. Dasilva, L. Bouneau *et al.*, "Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence," *nature genetics*, vol. 25, no. 2, pp. 235–238, 2000. 1

[22] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences." *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, Mar. 1981. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/7265238 1

[23] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990. 1

[24] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature genetics*, vol. 40, no. 12, pp. 1413–1415, 2008. 1

[25] D. Schmidt, M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield *et al.*, "Chip-seq: using high-throughput sequencing to discover protein–dna interactions," *Methods*, vol. 48, no. 3, pp. 240–248, 2009. 1

[26] A. Bernal, U. Ear, and N. Kyrpides, "Genomes OnLine Database (GOLD): a monitor of genome projects world-wide," *Nucleic Acids Research*, vol. 29, no. 1, pp. 126–127, 2001. [Online]. Available: http://www.genomesonline.org/cgi-bin/GOLD/index.cgi 2

[27] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007. 2

[28] L. R. Sabin, M. J. Delás, and G. J. Hannon, "Dogma Derailed: The Many Influences of RNA on the Genome," *Molecular Cell*, vol. 49, no. 5, pp. 783–794, 2013. 2, 8

[29] The ENCODE Project Consortium, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," *Nature*, vol. 447, pp. 799–816, 2007. 2, 29, 54

[30] N. Maeda, T. Kasukawa, R. Oyama, J. Gough, M. Frith *et al.*, "Transcript annotation in FANTOM3: Mouse Gene Catalog based on physical cDNAs," *PLoS Genetics*, vol. 2, p. e62, 2006. 2, 29, 54

[31] M. B. Clark, P. P. Amaral, F. J. Schlesinger, M. E. Dinger, R. J. Taft *et al.*, "The reality of pervasive transcription," *PLoS Biology*, vol. 9, p. e1000625, 2011. 2, 29, 54

[32] S. R. Eddy, "Non-coding RNA genes and the modern RNA world." *Nature Reviews Genetics*, vol. 2, no. 12, pp. 919–929, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11733745 2, 3, 8, 9, 10, 11, 12

[33] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding RNAs: insights into functions," *Nat Rev Genet*, vol. 10, no. 3, pp. 155–9, Mar 2009. 2, 8, 9, 15, 16, 29, 31, 54

[34] J. S. Mattick and I. V. Makunin, "Non-coding rna," *Human molecular genetics*, vol. 15, no. suppl 1, pp. R17–R29, 2006. 2

[35] R. W. Carthew and E. J. Sontheimer, "Origins and mechanisms of miRNAs and siRNAs," *Cell*, vol. 2009, pp. 642–655, 136. 2, 29

[36] P. Kapranov, G. St Laurent, T. Raz, F. Ozsolak, C. P. Reynolds *et al.*, "The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' unannotated RNA," *BMC Biol.*, vol. 8, p. 149, 2010. 2, 29, 54

[37] J. Hackermüller, K. Reiche, C. Otto, N. Hösler, C. Blumert *et al.*, "Cell cycle, oncogenic and tumor suppressor pathways regulate numerous long and macro non-protein coding RNAs," *Genome Biol.*, vol. 15, p. R48, 2014. 2, 29, 54

[38] H. I. Nakaya, P. P. Amaral, R. Louro, A. Lopes, A. A. Fachel *et al.*, "Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription," *Genome Biology*, vol. 8, no. 3, p. R43, 2007. 2, 29

[39] J. Engelhardt and P. F. Stadler, "Evolution of the unspliced transcriptome," *BMC Evol Biol*, vol. 15, p. 166, 2015, doi: 10.1186/s12862-015-0437-7. 2, 29

[40] P. Kapranov, J. Cheng, S. Dike, D. Nix, R. Duttagupta *et al.*, "RNA maps reveal new RNA classes and a possible function for pervasive transcription," *Science*, vol. 316, pp. 1484–1488, 2007. 2, 29

[41] C. P. Ponting, P. L. Oliver, and W. Reik, "Evolution and functions of long noncoding RNAs." *Cell*, vol. 136, no. 4, pp. 629–641, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2009.02.006 2, 14, 15, 29, 31, 54

[42] U. A. Orom and R. Shiekhattar, "Noncoding RNAs and enhancers: complications of a long-distance relationship," *Trends Genet*, vol. 27, no. 10, pp. 433–9, Oct 2011. 2, 3, 8, 14, 15, 29, 31, 54

[43] J. L. Rinn, M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu *et al.*, "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs," *Cell*, vol. 129, no. 7, pp. 1311–23, Jun 2007. 2, 29, 54

[44] Y. He, B. Vogelstein, V. E. Velculescu, N. Papadopoulos, and K. W. Kinzler, "The antisense transcriptomes of human cells," *Science*, vol. 322, no. 5909, pp. 1855–7, Dec 2008. 2, 29, 54

[45] M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young, "A chromatin landmark and transcription initiation at most promoters in human cells," *Cell*, vol. 130, no. 1, pp. 77–88, Jul 2007. 2, 29, 54

[46] H. L. Ashe, J. Monks, M. Wijgerde, P. Fraser, and N. J. Proudfoot, "Intergenic transcription and transinduction of the human beta-globin locus," *Genes Dev*, vol. 11, no. 19, pp. 2494–509, Oct 1997. 2, 29, 54

[47] T. Weirick, D. John, S. Dimmeler, and S. Uchida, "C-It-Loci: a knowledge database for tissue-enriched loci." *IBioinformatics*, p. 10.1093/bioinformatics/btv410, 2015. 2, 30, 54

[48] J. R. Hall, Z. J. Messenger, H. W. Tam, S. L. Phillips, L. Recio *et al.*, "Long noncoding RNA lincRNA-p21 is the major mediator of UVB-induced and p53-dependent apoptosis in keratinocytes," *Cell Death and Disease*, vol. 6, p. e1700, 2015. 2, 30, 54

[49] S.-S. Tang, B.-Y. Zheng, and X.-D. Xiong, "LincRNA-p21: Implications in Human DiseasesLong noncoding RNA lincRNA-p21 is the major mediator of UVB-induced and p53-dependent apoptosis in keratinocytes," *Int. J. Mol. Sci.*, vol. 16, pp. 18 732–18 740, 2015. 2, 30, 54

[50] V. Kumar, H.-J. Westra, J. Karjalainen, D. V. Zhernakova, T. Esko *et al.*, "Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression," *Cell Death and Disease*, vol. 9, no. 1, p. e1003201, 2013. 2, 30, 54

[51] C. Wang, C. Ding, R. F. Meraz, and S. R. Holbrook, "PSoL: a positive sample only learning algorithm for finding non-coding RNA genes," *Bioinformatics*, vol. 22, no. 21, pp. 2590–2596, 2006. 2, 13, 14

[52] R. Arrial, R. Togawa, and M. d. M. Brígido, "Outlining a Strategy for Screening Non-coding RNAs on a Transcriptome Through Support Vector Machines," vol. 4643, pp. 149–152, 2007, 10.1007/978-3-540-73731-5_14. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-73731-5_14 2, 13, 14

[53] T. C. Silva, P. A. Berger, R. T. Arrial, R. C. Togawa, M. M. Brigido *et al.*, "SOM-PORTRAIT: Identifying Non-coding RNAs Using Self-Organizing Maps." in *BSB*, ser. Lecture Notes in Computer Science, K. S. Guimarães, A. Panchenko, and T. M. Przytycka, Eds., vol. 5676. Springer, 2009, pp. 73–85. 2

[54] M. Fasold, D. Langenberger, H. Binder, P. F. Stadler, and S. Hoffmann, "DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments," *Nucleic acids research*, vol. 39, no. suppl 2, pp. W112–W117, 2011. 2, 9, 14

[55] A. Machado-Lima, H. A. Del Portillo, and A. M. Durham, "Computational methods in noncoding RNA research," *Journal of mathematical biology*, vol. 56, no. 1, pp. 15–49, 2008. 2

[56] T. M. Mitchell, *Machine Learning*, 2nd ed., E. M. Munson, Ed. New York: McGraw-Hill, 1997. [Online]. Available: http://www.cs.cmu.edu/~tom/mlbook.html 2, 19, 22

[57] S. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Pearson Education Inc., 2009. [Online]. Available: http://www.bibsonomy.org/bibtex/21f6eedfb10d8c4ad7369e3d994de5234/lantiq 2, 19, 20, 21

[58] O. Chapelle, B. Sch?olkopf, and A. Zien, *Semi-supervised learning*, ser. Adaptive computation and machine learning. Cambridge, Mass.: MIT Press, 2006. [Online]. Available: https://libproxy.usc.edu/login?url=http://site.ebrary.com/lib/uscisd/Doc?id=10173579 2, 24, 25, 26, 27, 55

[59] L. Wang, H. J. Park, S. Dasari, and co authors, "CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model," *Nucl Ac Res*, vol. 41, no. 6, pp. e74–e74, Apr 2013. 3, 17, 30, 54

[60] C. Pian, G. Zhang, Z. Chen, Y. Chen, J. Zhang *et al.*, "LncRNApred: Classification of Long Non-Coding RNAs and Protein-Coding Transcripts by the Ensemble Algorithm with a New Hybrid Feature," *PLoS One*, vol. 11, no. 5, p. e0154567, 2016. 3, 17, 30, 37, 40, 46, 54

[61] S. Han, Y. Liang, Y. Li, and W. Du, "Long Noncoding RNA Identification: Comparing Machine Learning Based Tools for Long Noncoding Transcripts Discrimination," *Biomed Res Int*, vol. 2016, p. 8496165, 2016. 3, 17, 30, 37, 40, 46, 48, 54

[62] R. Tripathi, S. Patel, V. Kumari, P. Chakraborty, and P. K. Varadwaj, "Deeplnc, a long non-coding rna prediction tool using deep neural network," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 5, no. 1, p. 21, 2016. [Online]. Available: http://dx.doi.org/10.1007/s13721-016-0129-2 3, 17, 30, 38, 54

[63] V. Wucher, F. Legeai, B. Hédan, G. Rizk, L. Lagoutte *et al.*, "FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome," *Nucleic Acids Research*, pp. 1–12, 2016. 3, 17, 30, 37, 40, 46, 48, 54

[64] T. R. Mercer, M. E. Dinger, S. M. Sunkin, M. F. Mehler, and J. S. Mattick, "Specific expression of long noncoding rnas in the mouse brain," *Proc Natl Acad Sci U S A*, vol. 105, no. 2, pp. 716–21, Jan 2008. 3, 8

[65] R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings *et al.*, "Long non-coding rna hotair reprograms chromatin state to promote cancer metastasis," *Nature*, vol. 464, no. 7291, pp. 1071–6, Apr 2010. 3

[66] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts *et al.*, *Molecular biology of the cell*, 4th ed. Garland, 2002. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0815332181 5, 6

[67] H. Lodish, A. Berk, C. A. Kaiser, M. Krieger, M. P. Scott *et al.*, *Molecular Cell Biology (Lodish, Molecular Cell Biology)*, 6th ed. W. H. Freeman, Jun. 2007. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0716776014 5, 6

[68] J. Setubal and J. Meidanis, *Introduction to computational molecular biology.* PWS Publishing, 1997. 6

[69] J. A. Shapiro, "Revisiting the central dogma in the 21st century," *Ann N Y Acad Sci*, vol. 1178, pp. 6–28, Oct 2009. 6

[70] I. A. Ilik, J. J. Quinn, P. Georgiev, F. Tavares-Cadete, D. Maticzka *et al.*, "Tandem stem-loops in rox rnas act together to mediate x chromosome dosage compensation in drosophila," *Mol Cell*, vol. 51, no. 2, pp. 156–73, Jul 2013. 8

[71] S. Maenner, M. Müller, J. Fröhlich, D. Langer, and P. B. Becker, "Atp-dependent rox rna remodeling by the helicase maleless enables specific association of msl proteins," *Mol Cell*, vol. 51, no. 2, pp. 174–84, Jul 2013. 8

[72] C. Ernst and C. C. Morton, "Identification and function of long non-coding rna," *Front Cell Neurosci*, vol. 7, p. 168, 2013. 8

[73] S. R. Eddy, "Noncoding RNA genes," *Current opinion in genetics & development*, vol. 9, no. 6, pp. 695–699, 1999. 8

[74] V. A. Erdmann, M. Z. Barciszewska, M. Szymanski, A. Hochberg, N. d. Groot *et al.*, "The non-coding RNAs as riboregulators," *Nucleic acids research*, vol. 29, no. 1, pp. 189–193, 2001. 8

[75] V. Erdmann, M. Barciszewska, A. Hochberg, N. d. Groot, and J. Barciszewski, "Regulatory RNAs," *Cellular and Molecular Life Sciences*, vol. 58, no. 7, pp. 960–977, 2001. 8

[76] J. H. Badger and G. J. Olsen, "Critica: coding region identification tool invoking comparative analysis," *Mol Biol Evol*, vol. 16, no. 4, pp. 512–24, Apr 1999. 8

[77] C. Presutti, J. Rosati, S. Vincenti, and S. Nasi, "Non coding rna and brain," *BMC Neurosci*, vol. 7 Suppl 1, p. S5, 2006. 8

[78] J. S. Mattick and M. J. Gagen, "The evolution of controlled multitasked gene networks: the role of introns and other noncoding rnas in the development of complex organisms," *Mol Biol Evol*, vol. 18, no. 9, pp. 1611–30, Sep 2001. 8

[79] C. Wahlestedt, "Natural antisense and noncoding rna transcripts as potential drug targets," *Drug Discov Today*, vol. 11, no. 11-12, pp. 503–8, Jun 2006. 8

[80] E. Rivas and S. R. Eddy, "Noncoding rna gene detection using comparative sequence analysis," *BMC Bioinformatics*, vol. 2, p. 8, 2001. 8

[81] K. C. Pang, M. C. Frith, and J. S. Mattick, "Rapid evolution of noncoding rnas: lack of conservation does not mean lack of function," *Trends Genet*, vol. 22, no. 1, pp. 1–5, Jan 2006. 8

[82] E. Torarinsson, M. Sawera, J. H. Havgaard, M. Fredholm, and J. Gorodkin, "Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common rna structure," *Genome Res*, vol. 16, no. 7, pp. 885–9, Jul 2006. 8

[83] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li *et al.*, "Classification of real and pseudo microrna precursors using local structure-sequence features and support vector machine," *BMC Bioinformatics*, vol. 6, p. 310, 2005. 8

[84] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm *et al.*, "Viennarna package 2.0," *Algorithms Mol Biol*, vol. 6, p. 26, 2011. 8, 9, 14

[85] S. R. Eddy, "Computational genomics of noncoding rna genes," *Cell*, vol. 109, no. 2, pp. 137–40, Apr 2002. 8

[86] M. C. Frith, T. L. Bailey, T. Kasukawa, F. Mignone, S. K. Kummerfeld *et al.*, "Discrimination of non-protein-coding transcripts from protein-coding mrna," *RNA Biol*, vol. 3, no. 1, pp. 40–8, 2006. 8, 9

[87] C. Liu, B. Bai, G. Skogerbø, L. Cai, W. Deng *et al.*, "Noncode: an integrated knowledge database of non-coding rnas," *Nucleic Acids Res*, vol. 33, no. Database issue, pp. D112–5, Jan 2005. 8, 9

[88] W. Arruda, C. G. Ralha, T. Raiol, M. M. Brígido, M. E. M. T. Walter *et al.*, "ncRNA-Agents: A multiagent system for non-coding RNA annotation," in *Advances in Bioinformatics and Computational Biology, 8th BSB*, ser. Lect. Notes Comp. Sci., J. C. Setubal and N. F. Almeida, Eds., vol. 8213, 2013, pp. 136–147. 8

[89] V. N. Kim, J. Han, and M. C. Siomi, "Biogenesis of small RNAs in animals," *Nature Reviews Molecular Cell Biology*, vol. 10, no. 2, pp. 126–139, 2009. 9, 12

[90] S. S. Lakshmi and S. Agrawal, "piRNABank: a web resource on classified and clustered Piwi-interacting RNAs," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D173–D177, 2008. 12

[91] P. F. Stadler, J. J. L. Chen, J. Hackermüller, S. Hoffmann, F. Horn *et al.*, "Evolution of vault RNAs," *Molecular biology and evolution*, vol. 26, no. 9, pp. 1975–1991, 2009. 12

[92] E. Meiri, A. Levy, H. Benjamin, M. Ben-David, L. Cohen *et al.*, "Discovery of microRNAs and other small RNAs in solid tumors," *Nucleic acids research*, vol. 38, no. 18, pp. 6234–6246, 2010. 12

[93] C. P. Christov, T. J. Gardiner, D. Szüts, and T. Krude, "Functional requirement of noncoding Y RNAs for human chromosomal DNA replication." *Mol Cell Biol*, vol. 26, no. 18, pp. 6993–7004, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1128/mcb.01060-06 12

[94] E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy, "Infernal 1.0: inference of rna alignments," *Bioinformatics*, vol. 25, no. 10, pp. 1335–7, May 2009. 9, 14

[95] S. Bartschat, S. Kehr, H. Tafer, P. F. Stadler, and J. Hertel, "SnoStrip: A snoRNA annotation pipeline," *Bioinformatics*, vol. 30, no. 1, pp. 115–6, Jan 2014. 13, 14

[96] J. Hertel, I. L. Hofacker, and P. F. Stadler, "SnoReport: computational identification of snoRNAs with unknown targets," *Bioinformatics*, vol. 24, no. 2, pp. 158–64, Jan 2008. 13, 14, 54

[97] J. V. de Araujo Oliveira, F. Costa, R. Backofen, P. F. Stadler, M. E. Machado Telles Walter *et al.*, "Snoreport 2.0: new features and a refined support vector machine to improve snorna identification," *BMC Bioinformatics*, vol. 17, no. Suppl 18, p. 464, Dec 2016. 13, 54

[98] H. Tafer, S. Kehr, J. Hertel, I. L. Hofacker, and P. F. Stadler, "RNAsnoop: efficient target prediction for H/ACA snoRNAs," *Bioinformatics*, vol. 26, no. 5, pp. 610–6, Mar 2010. 13, 14

[99] S. W. Burge, J. Daub, R. Eberhardt, J. Tate, L. Barquist *et al.*, "Rfam 11.0: 10 years of rna families," *Nucleic Acids Research*, vol. 41, no. D1, pp. D226–D232, 2013. 13, 14

[100] D. Bu, K. Yu, S. Sun, C. Xie, G. Skogerbø *et al.*, "Noncode v3.0: integrative annotation of long noncoding rnas," *Nucleic Acids Res*, vol. 40, no. Database issue, pp. D210–5, Jan 2012. 14, 18

[101] A. Kozomara and S. Griffiths-Jones, "mirbase: integrating microrna annotation and deep-sequencing data," *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D152–7, Jan 2011. 14

[102] J. S. Mattick and I. V. Makunin, "Non-coding RNA," *Human Molecular Genetics*, vol. 15, no. suppl 1, pp. R17–R29, 15 April 2006. [Online]. Available: http://hmg.oxfordjournals.org/content/15/suppl_1/R17.abstract 14

[103] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith *et al.*, "The Transcriptional Landscape of the Mammalian Genome," *Science*, vol. 309, no. 5740, pp. 1559–1563, Sep. 2005. [Online]. Available: http://dx.doi.org/10.1126/science.1112014 15

[104] P. Kapranov, J. Cheng, S. Dike, D. A. Nix, R. Duttagupta *et al.*, "RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription," *Science*, vol. 316, no. 5830, pp. 1484–1488, Jun. 2007. [Online]. Available: http://dx.doi.org/10.1126/science.1138341 15

[105] Y. Devaux, J. Zangrando, B. Schroen, E. E. Creemers, T. Pedrazzini *et al.*, "Long noncoding rnas in cardiac development and ageing," *Nat Rev Cardiol*, vol. 12, no. 7, pp. 415–25, Jul 2015. 15

[106] K. Sun, X. Chen, P. Jiang, X. Song, H. Wang *et al.*, "iseerna: identification of long intergenic non-coding rna transcripts from transcriptome sequencing data," *BMC Genomics*, vol. 14 Suppl 2, p. S7, 2013. 17

[107] X. Guo, L. Gao, Q. Liao, H. Xiao, X. Ma *et al.*, "Long non-coding rnas function annotation: a global prediction method based on bi-colored networks," *Nucleic Acids Res*, vol. 41, no. 2, p. e35, Jan 2013. 17

[108] L. Kong, Y. Zhang, Z.-Q. Ye, X.-Q. Liu, S.-Q. Zhao *et al.*, "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine," *Nucleic Acids Res*, vol. 35, no. Web Server issue, pp. W345–9, Jul 2007. 17, 30, 54

[109] K. Sun, X. Chen, P. Jiang, X. Song, H. Wang *et al.*, "iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data," *BMC Genomics*, vol. 14 Suppl 2, p. S7, 2013. 17, 30, 37, 40, 46, 48

[110] M. E. Dinger, K. C. Pang, T. R. Mercer, M. L. Crowe, S. M. Grimmond *et al.*, "NRED: a database of long noncoding RNA expression," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D122–D126, 2009. 18, 31, 55

[111] M. D. Paraskevopoulou, G. Georgakilas, N. Kostoulas, M. Reczko, M. Maragkakis *et al.*, "Diana-lncbase: experimentally verified and computationally predicted microrna targets on long non-coding rnas," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D239–45, Jan 2013. 18

[112] M. Reczko, M. Maragkakis, P. Alexiou, I. Grosse, and A. G. Hatzigeorgiou, "Functional microrna targets in protein coding sequences," *Bioinformatics*, vol. 28, no. 6, pp. 771–6, Mar 2012. 18

[113] G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu *et al.*, "Lncrnadisease: a database for long-non-coding rna-associated diseases," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D983–6, Jan 2013. 18

[114] X. C. Quek, D. W. Thomson, J. L. Maag, N. Bartonicek, B. Signal *et al.*, "lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs," *Nucleic Acids Research*, vol. 43, no. Database issue, pp. D168–D173, 2015. 18, 31, 55

[115] J. Jin, J. Liu, H. Wang, L. Wong, and N. H. Chua, "PLncDB: plant long non-coding RNA database," *Bioinformatics*, vol. 29, no. 8, pp. 1068–1071, Apr 2013. 18, 31, 55

[116] R. Bellman, *Dynamic Programming*, ser. Dover Books on Computer Science Series. Dover Publications, 2003. [Online]. Available: https://books.google.com.br/books?id=fyVtp3EMxasC 22

[117] I. T. Jolliffe, *Principal component analysis*, 2nd ed. New York: Springer-Verlag, 2002. 22, 31, 33, 55, 56

[118] S. S. Haykin, *Neural networks: a comprehensive foundation*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall, 1999. 23

[119] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ser. COLT '92. New York, NY, USA: ACM, 1992, pp. 144–152. [Online]. Available: http://doi.acm.org/10.1145/130385.130401 23, 33

[120] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 23, 24, 33

[121] R. Brachman, W. W. Cohen, and P. Stone, Eds., *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2012. 25

[122] T. P. Nguyen and T. B. Ho, "A semi-supervised learning approach to disease gene prediction," in *Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on*. IEEE, 2007, pp. 423–428. 27

[123] X. Chen and G.-Y. Yan, "Semi-supervised learning for potential human microrna-disease associations inference," *Sci Rep*, vol. 4, p. 5501, Jun 2014. 27

[124] A. Stanescu and D. Caragea, "An empirical study of ensemble-based semi-supervised learning approaches for imbalanced splice site datasets," *BMC Syst Biol*, vol. 9 Suppl 5, p. S1, 2015. 28

[125] T. Provoost and M.-F. Moens, "Semi-supervised learning for the bionlp gene regulation network," *BMC Bioinformatics*, vol. 16, no. 10, p. S4, 2015. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-16-S10-S4 28

[126] H. W. Schneider, T. Raiol, M. M. Brigido, M. E. M. T. Walter, and P. F. Stadler, "A support vector machine based method to distinguish long non-coding rnas from protein coding transcripts," *BMC Genomics*, vol. 18, no. 1, p. 804, Oct 2017. [Online]. Available: https://doi.org/10.1186/s12864-017-4178-4 29

[127] X. Guo, L. Gao, Y. Wang, D. K. Y. Chiu, T. Wang *et al.*, "Advances in long non-coding RNAs: identification, structure prediction and function annotation," *Brief Funct Genomics*, vol. 15, no. 1, pp. 38–46, Jan 2016. 30, 31

[128] G. M. Ventola, T. M. R. Noviello, S. D´ Aniello, A. Spagnuolo, M. Ceccarelli *et al.*, "Identification of long non-coding transcripts with feature selection: a comparative study," *BMC Bioinformatics*, vol. 18, no. 187, pp. 1–16, 2017. 30, 53, 68

[129] F. S, S. P. Niclou, and F. Azuaje, "Databases for lncRNAs: a comparative evaluation of emerging tools," *RNA*, vol. 20, no. 11, pp. 1655–1665, Jun 2014. 31, 55

[130] C. Xie, J. Yuan, H. Li, M. Li, G. Zhao *et al.*, "NONCODEv4: exploring the world of long non-coding RNA genes," *Nucleic Acids Research*, vol. 42, no. D1, pp. D98–D103, Nov 2014. 31, 55

[131] M. D. Paraskevopoulou, G. Georgakilas, N. Kostoulas, M. Reczko, M. Maragkakis *et al.*, "DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs," *Nucleic Acids Research*, vol. 41, no. Database issue, pp. D239–45, Jan 2013. 31, 55

[132] G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu *et al.*, "LncRNADisease: a database for long-non-coding RNA-associated diseases," *Nucleic Acids Research*, vol. 41, no. Database issue, pp. D983–6., Jan 2013. 31, 55

[133] K. Yan, Y. Arfat, D. Li, and co authors, "Structure prediction: New insights into decrypting long noncoding RNAs," *Int. J. Mol. Sci.*, vol. 17, no. 1, p. 132, Jun 2016. 31

[134] A. Necsulea, M. Soumillon, M. Warnefors, A. Liechti, T. Daish *et al.*, "The evolution of lncrna repertoires and expression patterns in tetrapods," *Nature*, vol. 505, no. 7485, pp. 635–40, Jan 2014. 31, 34, 52, 55, 57, 59, 63, 65

[135] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal *et al.*, "Uniprotkb/swiss-prot, the manually annotated section of the uniprot knowledgebase: How to use the entry view," *Methods Mol Biol*, vol. 1374, pp. 23–54, 2016. 34, 38, 52

[136] P. Agarwal and V. Bafna, "The ribosome scanning model for translation initiation: implications for gene prediction and full-length cdna detection," *Proc Int Conf Intell Syst Mol Biol*, vol. 6, pp. 2–7, 1998. 35

[137] L. Sun, H. Liu, L. Zhang, and J. Meng, "lncRScan-SVM: A Tool for Predicting Long Non-Coding RNAs Using Support Vector Machine," *PLoS One*, vol. 10, no. 10, p. e0139654, 2015. 37, 40

[138] P. J. Volders, K. Verheggen, G. Menschaert, K. Vandepoele, L. Martens *et al.*, "An update on lncipedia: a database for annotated human lncrna sequences," *Nucleic Acids Res*, vol. 43, no. 8, pp. 4363–4, Apr 2015. 38

[139] A. Nitsche, D. Rose, M. Fasold, K. Reiche, and P. F. Stadler, "Comparison of splice sites reveals that long non-coding RNAs are evolutionarily well conserved," *RNA*, vol. 21, pp. 801–812, 2015, doi: 10.1261/rna.046342.114. 38

[140] A. Frankish and J. Harrow, "Gencode pseudogenes," *Methods Mol Biol*, vol. 1167, pp. 129–55, 2014. 52

[141] S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman *et al.*, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990. 54

[142] X. Guo, L. Gao, Y. Wang, D. K. Chiu, T. Wang *et al.*, "Advances in long noncoding RNAs: identification, structure prediction and function annotation." *Brief Funct Genomics*, vol. 15, no. 1, pp. 38–46, Jun 2016. 55

[143] T. Joachims, "Transductive support vector machines," *Chapelle et al.(2006)*, pp. 105–118, 2006. 55

[144] H. Schneider, S. Bartschat, G. Doose, L. Maciel, E. Pizani *et al.*, *Genome-Wide Identification of Non-coding RNAs in Komagatella pastoris str. GS115.* Cham: Springer International Publishing, 2014, pp. 115–122. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-12418-6_15 67