View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Unitn-eprints PhD

PhD Dissertation

International Doctorate School in
Information and Communication Technologies

DISI - University of Trento

PROTEIN-DEPENDENT PREDICTION OF MESSENGER RNA BINDING USING SUPPORT VECTOR MACHINES

Carmen Maria Livi

Advisor:

Prof. Enrico Blanzieri

University of Trento

March 2013

Abstract

RNA-binding proteins interact specifically with RNA strands to regulate important cellular processes. Knowing the binding partners of a protein is a crucial issue in biology and it is essential to understand the protein function and its involvement in diseases. The identification of the interactions is currently resolvable only through in vivo and in vitro experiments which may not detect all binding partners. Computational methods which capture the protein-dependent nature of the binding phenomena could help to predict, in silico, the binding and could be resistant against experimental biases.

This thesis addresses the creation of models based on support vector machines and trained on experimental data. The goal is the identification of RNAs which bind specifically to a regulatory protein. Starting from a case study, done with protein CELF1, we extend our approach and propose three methods to predict whether an RNA strand can be bound by a particular RNA-binding protein. The methods use support vector machines and different features based on the sequence (method Oli), the motif score (method OliMo) and the secondary structure (method OliMoSS). We apply them to different experimentally-derived datasets and compare the predictions with two methods: RNAcontext and RPISeq. Oli outperforms OliMoSS and RPISeq affirming our protein specific prediction and suggesting that oligo frequencies are good discriminative features. Oli and RNAcontext are the most competitive methods in terms of AUC. A Precision-Recall analysis reveals a better performance for Oli. On a second experimental dataset, where negative binding information is available, Oli outperforms RNAcontext with a precision of 0.73 vs. 0.59. Our experiments show that features based on primary sequence information are highly discriminative to predict the binding between protein and RNA. Sequence motifs can improve the prediction only for some RNA-binding proteins. Finally, we can conclude that experimental data on RNA-binding can be effectively used to train protein-specific models for in silico predictions.

Keywords
bioinformatics, RNA-protein binding, RNA binding site, support vector machines

A Oma,
e a me :)

Acknowledgements

First of all I would like to thank my advisor Prof. Enrico Blanzieri for his supervision and his guide through my PhD activity. Special thanks go to Yann Audic and Luc Paillard of the Institut Génétique et Développement de Rennes (France) for the internship opportunity and for giving me access to their dataset. Moreover their suggestions regarding CELF1-binding and high-throughput methods gave the right hint for my research. I am grateful to Dr. Michela A. Denti and Dr. Francesca Demichelis for their help when I was lost in biology. My sincere thanks also to Nicola Segata who taught me a lot at the beginning and for his support at the ending phase of my PhD work.

Ich danke meiner ganzen Familie für Ihre Unterstützung, ganz besonders Mami und Tati, meiner Schwester Marion und meiner lieben Oma ... auf das klein Emma vielleicht auch einmal eine These schreibt.
Alla fine vorrei ringraziare una persona che era sempre al mio fianco, che ha sopportato i miei drammi personali quando tutto andava storto e che mi ha spinto quando non volevo più. Purtroppo le nostre strade si sono divise ma senza di te Paolo non sarei la persona che sono oggi, alla fine di questo PhD.

Carmen

Contents

1 Introduction 1
1.1 The context 1
1.2 The problem 3
1.3 Our approach 4
1.4 Structure of the thesis 4

2 Background 7
2.1 RNA-protein binding. A biological point of view 7
2.1.1 RNA 7
2.1.2 Protein 8
2.1.3 The binding 8
2.1.4 Laboratory experiments 11
2.2 RNA-protein binding. A computational point of view 12
2.2.1 Dissection and analysis of RNA-protein complexes 12
2.2.2 Binding residue prediction 13
2.2.3 Motif finding tools 19
2.3 Machine learning 20
2.3.1 Support vector machine (SVM) 20
2.4 Data balancing 21
2.5 Performance measures 21
2.6 Biological databases 22
2.6.1 The University of California, Santa Cruz (UCSC) Genome Browser . . . 22
2.6.2 The Ensembl project 23
2.6.3 The RCSB Protein Data Bank (PDB) 23
2.6.4 The Atlas of UTR regulatory activity (AURA) 23
2.6.5 The National Centre for Biotechnology Information (NCBI) resources . 23
2.6.6 The Gene Expression Omnibus (GEO) 24

3 Preliminary analysis 25
3.1 Abstraction of the binding . . . 25
3.2 The Approach . . . 27
3.3 Material and Methods . . . 28
3.3.1 Dataset . . . 28
3.3.2 Binding Residue Identification . . . 28
3.3.3 Feature Extraction and Representation . . . 30
3.4 Results and Discussion . . . 31
3.5 Conclusions . . . 33

4 RNA-binding prediction for CELF1. A case study 35
4.1 The Approach . . . 36
4.2 Material and Methods . . . 36
4.2.1 Datasets . . . 36
4.2.2 Feature Extraction and Representation . . . 38
4.2.3 Experiments and Results . . . 40
4.3 Discussion and Conclusions . . . 42

5 Predicting mRNA binding with sequence features, motifs and secondary structures 45
5.1 The Approach . . . 45
5.2 Material and Methods . . . 47
5.2.1 Datasets . . . 47
5.2.2 Feature Extraction and Representation . . . 48
5.2.3 Evaluation and Comparison . . . 49
5.3 Results and Discussion . . . 49
5.3.1 Evaluation 1 . . . 49
5.3.2 Evaluation 2 . . . 59
5.3.3 Evaluation 3 . . . 59
5.4 Conclusions . . . 64

6 Conclusions and future work 69

Bibliography 71

List of Tables

2.1 Physical-chemical amino acid properties 9
2.2 Methods for RNA-binding site predictions 17

3.1 Dataset description 29
3.2 Performance of five different feature combinations 32
3.3 Predicted binding and non-binding residues in six complexes using SVM1.4 . . 33
3.4 Constructed versus real RNA-strands 34

4.1 Summary of the CELF1 datasets 38
4.2 Performance of the CELF1 model with different feature combinations 41
4.3 Results of the first experiment 41
4.4 Results of the second experiment 42

5.1 Performance of *Oli*, *OliMo*, *OliMoSS*, *RNAcontext* and *RPISeq* on the *AURA-dataset* 50
5.2 p-values of the Wilcoxon signed-rank test for *Oli*, *OliMo*, *OliMoSS*, *RNAcontext* and *RPISeq*. 51
5.3 Precision values for *Oli*, *OliMo*, *OliMoSS*, *RNAcontext* and *RPISeq*. 52
5.4 Performance of *Oli*, *OliMo*, *OliMoSS*, *RNAcontext* and both *RPISeq* methods on *PUM2+* in combination with two different negative datasets 63
5.5 Additional analysis to test the ability of the model to identify binding sequences among general 3'UTRs. 64

List of Figures

2.1 *BindN+* output example for the protein sequence PUM2_HUMAN (Q8TB72). 15

3.1 RNA-sequence construction based on predicted binding residues. 29

4.1 CELF1-binding motif of length 11. The corresponding PSSM produces scores which are used as features in our model 37
4.2 CELF1-binding motif of length 15 37
4.3 Schematic diagram of oligo feature extraction 39
4.4 Schematic diagram of the motif-score calculation. 39

5.1 Precision-Recall curves for *Oli* and *RNAcontext* on protein SLBP, MSI1, TIAL1, CEPB4, AGO2 and CPEB1. 53
5.2 Precision-Recall curves for *Oli* and *RNAcontext* on protein CUGBP1, TNRC6A, PUM1, TNRC6C, PABP and U2AF65. 54
5.3 Precision-Recall curves for *Oli* and *RNAcontext* on protein AGO4, QKI, TNRC6B, ELAVL1, AUF1 and AGO1. 55
5.4 ROC curves for *Oli* and *RNAcontext* on protein SLBP, MSI1, TIAL1, CEPB4, AGO2 and CPEB1. 56
5.5 ROC curves for *Oli* and *RNAcontext* on RBP CUGBP1, TNRC6A, PUM1, TNRC6C, PABP and U2AF65. 57
5.6 ROC curves for *Oli* and *RNAcontext* on RBP AGO4, QKI, TNRC6B, ELAVL1, AUF1 and AGO1. 58
5.7 ROC curve of the *Oli* performance on the *AURAdataset*. 60
5.8 ROC curves of *Oli*, *OliMo*, *OliMoSS*, *RNAcontext*, *RPISeq-SVM* and *RPISeq-RF* on *3K-* 61
5.9 ROC curves of *Oli*, *OliMo*, *OliMoSS*, *RNAcontext*, *RPISeq-SVM* and *RPISeq-RF* on *PUM2-* 62
5.10 PR curve of *Oli* on *PUM2+* in combination with *3K-* and *PUM2-*. 63
5.11 Precision-Recall curves for *Oli* and *RNAcontext*. 65
5.12 The approach 66

Chapter 1

Introduction

Protein-protein and RNA-protein interactions are crucial mechanisms in cells as they are involved in many processes like the regulation of gene transcription or the regulation of molecular pathways. A deeper understanding of the specificity of the binding is a basic step towards the construction of computational models for simulations and *in silico* binding predictions.

1.1 The context

Inside the eukaryotic nucleus genes are transcribed into ribonucleic acids (RNAs). An RNA is a consecutive sequence of four nucleotides: Adenine (A), Guanine (G), Cytosine (C) and Uracil (U). There exists several types of RNAs: mRNA codifies for a protein and tRNA is involved in the translation of mRNAs into protein sequences. Proteins are organic structures made of amino acids which fold into a globular and compact form. In eukaryotic cells proteins functionally act as enzymes, structural proteins or attend to the cell signalling.

Proteins can interact with nucleic acids, amino acids or organic compounds. In particular when they are able to bind ribonucleotides they are called RNA-binding proteins (RBPs). The human genome encodes a large number of RNA-binding proteins (RBPs) (Glisovic et al., 2008; Uren et al., 2011; Zhang and Darnell, 2011) which carry out different functions and determine a variety of biological processes. Some RBPs are well-studied and their function is partially known as for example ELAVL1, an RBP which has been identified to be involved in diseases and cancer (Uren et al., 2011). Unfortunately this information is not available for all RBPs and investigations about the binding properties are needed (Mukherjee et al., 2011). RBPs are also involved in the post-transcriptional regulation, in splicing and in phenomena like RNA stability and RNA translation. This suggests that RBPs have to interact specifically with their mRNA targets. Each mRNA contains, beside the coding regions, also untranslated regions: the 5' and the 3' untranslated region (5'UTR, 3'UTR). Especially on the 3'UTR several regulatory sequences and RBP binding sites are established (Corà et al., 2007).

The interplay between RBPs and RNA is a fine-tuned system in the cell and its perturbation can cause disorders. Many abnormal splicing proteins are found in tumours. Therefore knowing more about RBP-RNA binding for splicing and transcription factors is of interest in cancer research. Consequently a better understanding of such interactions could help to explore their importance in diseases (Khalil and Rinn, 2011; Zhang and Darnell, 2011). Moreover, the identification of RNA targets is of special interest in biology and essential to understand proteins function (Uren et al., 2011; Auweter et al., 2006; Lichtarge and Sowa, 2002). More precise binding descriptions and accurate binding predictions are therefore needed (Mukherjee et al., 2011).

In the last decade different computational approaches focused on the prediction of RBP-RNA interactions. One category of approaches relies on the use of machine learning techniques like Neural Networks (Jeong et al., 2004), Random Forest (RF) (Liu et al., 2010), Naïve Bayes (Terribilini et al., 2006) and Support Vector Machines (SVM) (Wang and Brown, 2006; Cheng et al., 2008; Wang et al., 2011) to predict single amino acids that on the protein surface potentially interact with ribonucleotides. These methods are based on the binding information extracted from 3-dimensional binding complexes and exhibit high prediction accuracies. Unfortunately they do not consider the RNA-binding partner and give no information about the RNA sequence potentially bound by an RBP. *RPISeq* (Muppirala et al., 2011) instead addresses this binding-partner problem and predicts if a given RNA sequence is bound by a specific RBP, obtaining high positive-prediction rates. Another class of computational approaches is that of motif-finding tools which search for binding sites on RNA strands (Glisovic et al., 2008; Lichtarge and Sowa, 2002). These methods need experimental data to extract significant sequence motifs within the bound sequences (Bailey et al., 2009) or to search for significant sequences and structural motifs by learning from both bound and non-bound data (Kazan et al., 2010). Another category of studies concentrate on the amino acid propensity (Jeong et al., 2003; Pérez-Cano and Fernández-Recio, 2010; Gupta and Gribskov, 2011) and the structural analysis (Bahadur et al., 2008; Jones et al., 2001) of RNA-protein complexes concluding that amino acids and nucleotides have no significant preference for binding.

RNA protein interactions are not sufficiently understood and investigative techniques suffer from different biases. The binding is highly protein specific (Westhof and Fritsch, 2011) and despite many experimental investigations the interaction mechanism between RBPs and target RNAs is not always well described (Änkö et al., 2012; Mukherjee et al., 2011; Khalil and Rinn, 2011). This circumstance may due to different binding preferences: some RBPs bind specific target sequences on the RNA strand (Uren et al., 2011), others recognize their binding site within the RNA secondary structure (Li et al., 2010; Draper, 1999). Even within the same RBP class the binding recognition can be different (Guzman et al., 1998; Änkö et al., 2012).

1.2 The problem

Currently the detection of RNA targets and the identification of real binding sites has to be done through *in vitro* and *in vivo* experiments like for example the *systematic evolution of ligands by exponential enrichment (SELEX)* technique (Tuerk and Gold, 1990) or *cross-linking and immunoprecipitation (CLIP)* techniques (Kishore et al., 2011; Hafner et al., 2010; Jaskiewicz et al., 2012). Unfortunately they are costly and time consuming and each such technique has its assumptions and its limitations due to experimental biases (Kishore et al., 2011; Änkö and Neugebauer, 2012; Puton et al., 2011). Furthermore the assessed RBP-RNA interactions are limited to the deployed species, the deployed cell type and the experimental conditions. Moreover the observed binding sequences are restricted to the currently transcribed genes and not all binding-partners might be detected. The same holds for the non-binding information: the transcriptome in a specific sample does not cover all the possible transcripts even in the same species. Therefore by performing one experiment one may have only a subset of all possible binding and non-binding sequences.

Computational methods, capable to capture the specific and protein-dependent nature of the binding phenomena, could help to detect the interaction partners *in silico* and could be resistant against introduced biases. On the other hand high-throughput datasets contain precious information on detected RNA-protein interactions. Exploiting the information contained in these experimental data to predict *in silico* the other RNA-protein bindings seems therefore a promising strategy.

RNA-binding prediction could be achieved by applying motif finding tools to detect an RBP-binding site on an RNA strand and infer a consequential interaction but it may underestimate the complex binding mechanism (Änkö and Neugebauer, 2012) and it is not reliable (Westhof and Fritsch, 2011). A single RNA can contain binding sites for more than one protein (Jain et al., 2011) and the binding of an RBP can depend on the binding of another protein. Some RBPs may need more than one binding site spread along the folded RNA sequence. These specific binding mechanisms can not be covered by motif finding tools alone but could be rather caught by features describing the general sequence composition (Änkö and Neugebauer, 2012). The most important question remains: “what are the RNA targets of RBPs; and how is their *in vivo* binding specificity achieved?” (Änkö and Neugebauer, 2012). A comparison of several available RBP-binding-site prediction tools concludes that this kind of methods need to be improved. A suggestion is to include more sequences and structural information to develop methods able to predict protein-RNA interactions (Puton et al., 2011).

1.3 Our approach

Knowledge about the binding partners is of great interest in molecular biology. Computational predictions could reduce the number of laboratory experiments, which are costly and time intensive, and the lack of literature in this domain let us conclude that there is a niche which is not yet occupied. Therefore in this thesis we aim to create an *in silico* binding prediction based on SVM. Experimental *in vivo* data constitute a precious source for model constructions because they may contain important information regarding RBP affinities and binding preferences. Consequently we create our models on this kind of data. Since each RBP interacts differently with its target RNAs (Glisovic et al., 2008), it seems reasonable to train one SVM per RBP to model the specific binding phenomena. We represent each RNA sequence initially by its oligonucleotide composition and include significant binding patterns as features. Afterwards we extend the model with secondary structure and accessibility features and apply it to more experimentally derived RBPs. The use of SVMs is motivated by the good classification performance shown in previously published studies (Wang et al., 2010; Tong et al., 2008; Wang et al., 2011).

In this thesis we present a method which predicts RNA-target sequences in a protein specific way. The innovative aspect of our approach is the use of experimental datasets of detected RBP-RNA binding partners to construct prediction models. Another innovative property is the application of secondary structure and accessibility information as features. Even if our contribution to the literature is a first step, the treated argument is of importance in biology and still far away from being resolved (Puton et al., 2011). Therefore in the thesis we will show that:

- negative data, even if not detected in laboratory can be used to train models;
- secondary structure information appears to be not necessary to predict the binding;
- high-throughput data can effectively be used to create models.

1.4 Structure of the thesis

This document is divided into 6 chapters. Chapter 2 explains basics and gives background information important to understand the thesis. The chapter is divided into two sections: in Section 2.1 we describe the RNA-protein binding first from a biological point of view by specifying RNA, protein and their binding interactions. Then we explain shortly some relevant *in vivo* and *in vitro* experiments. Section 2.2 describes the literature concerned with the prediction of binding sites in protein sequences, details their performance values and the applied features. We resume also studies which analyse the binding site in RNA-protein complexes, describe SVMs, data balancing and list diverse important biological databases which are relevant in the thesis. Chapter 3 describes a preliminary analysis in which we address the prediction of the RBP binding partner. The approach exploits 3-dimensional binding complexes and is based on binding residue predictions. Afterwards we report the obtained results and discuss them. Chapter 4 delineates a case study which represents the basic framework of our approach. The case study has been done on RBP CELF1 and exploits a high-throughput dataset to construct the SVM model. Sequence and motif features are used. The result of the case study conducts to the extension of the approach to more RBPs and to different features, all described in Chapter 5. The chapter reports the obtained results and the comparison with other published methods. Finally, in the last Chapter 6 we draw our conclusions and propose future works.

Chapter 2

Background

In this chapter we introduce the RNA-protein binding first from a biological point of view by describing RNA, protein, their binding and the involved physical and chemical forces. Then we explain relevant *in vivo* and *in vitro* experiments used to detect an occurred binding. From the computational point of view this chapter introduces the state-of-the-art literature concerned with RNA-protein binding and presents some databases where biological data is stored.

2.1 RNA-protein binding. A biological point of view

Post-transcriptional regulations like alternative splicing or RNA translation are crucial processes in cells and are mediated by RBPs and transacting RNAs. Furthermore RBPs are part of complex regulatory networks and their interplay with RNA sequences is crucial in cells whereas perturbations can cause disorders (Glisovic et al., 2008; Hogan et al., 2008). Consequently a better understanding of their interactions helps to explore their importance in diseases (Khalil and Rinn, 2011; Zhang and Darnell, 2011).

It is known that RBPs interact specifically with their mRNA targets and understanding the exact molecular recognition (Westhof and Fritsch, 2011) is important to determine the mechanism of these specific binding. The identification of the binding partners is important in biology to determine a protein function (Uren et al., 2011; Auweter et al., 2006) and consequently better descriptions of the binding are required (Mukherjee et al., 2011).

2.1.1 RNA

Inside the nucleus of an eukaryotic cell genes are transcribed into RNA. An RNA is a consecutive sequence of the four ribonucleotides Adenine (A), Guanine (G), Cytosine (C) and Uracil (U). Several types of RNAs are present in the cell: messanger RNAs (mRNAs), non-coding RNAs (ncRNAs) like the transfer RNA (tRNA) or regulatory RNAs like the microRNA (miRNA). They have all different assignments, for example mRNAs codify for proteins. An mRNA contains several regions:

- the coding region
- the 5' untranslated region (5'UTR)
- the 3' untranslated region (3'UTR)

The coding region of the mRNA strand is used as a kind of "blueprint" and translated into a protein sequence. Especially on the 3'UTR several regulatory sequences and RBP binding sites are established (Corà et al., 2007; Hogan et al., 2008). Usually RNA strands appear single stranded but guided by molecular forces, such as hydrogen bonds and stacking interactions, some RNAs can form secondary structures like stems, hairpin-loops and bulges.
There exists *in silico* methods which predict the formed secondary structure of an input sequence. For instance the *Vienna RNA package* (Lorenz et al., 2011) provides software tools to predict RNA secondary structures, to analyse and to compare them. The predictions use different approaches based on the minimum free energy, on the pair probabilities or based on suboptimal structure folding. For instance *RNAfold* calculates the minimum free energy secondary structure of a corresponding RNA (Hofacker et al., 1994).

2.1.2 Protein

An mRNA sequence codifies for a protein sequence. Proteins are organic structures build up by a consecutive sequence of amino acids, also called the primary structure of the protein. 20 amino acids exist in nature. Table 2.1 lists the 20 amino acids and some of their molecular properties. Similarly to RNA also the protein sequence forms secondary structures and folds into a 3-dimensional structure. Small molecular forces like hydrogenbonds, Van der Waals forces, stacking interactions and electrostatic interactions (Auweter et al., 2006) induce the formation of three types of secondary structures called α-helix, β-sheets and coils. The arrangement of these structures in the 3-dimensional space is called the tertiary structure in which the protein folds into a compact and globular form. Each element of the protein chain, namely each amino acid, has individual chemical and physical properties such as charge, molecular mass or hydrophobicity. These properties determine the folding of the protein and are involved in the destined function of the folded protein.

2.1.3 The binding

The RNA binding is protein dependent because each protein has its own binding characteristic. Even similar RBPs and RBPs with the same protein surface bind differently (Guzman et al.,

Table 2.1: **Physical-chemical amino acid properties.** The Table lists each amino acid by name and gives some values regarding its hydrophobicity (Kyte and Doolittle, 1982), its molecular weight (Artimo et al., 2012) and the acid dissociation constant pKa for each side-chain (Parrill, 1997).

Name	Hydrophobicity	Chemical structure	Molecular weight [g/mol]	pKa^{1a}	pKa^{1b}
Alanine	1.8	C3H5ON	71.0788	2.35	9.87
Arginine	-4.5	C6H12ON4	156.1875	2.01	9.04
Asparagine	-3.5	C4H6O2N2	114.1038	2.02	8.80
Aspartic Acid	3.5	C4H5O3N	115.0886	2.10	9.82
Cysteine	2.5	C3H5ONS	103.1388	2.05	10.25
Glutamic Acid	-3.5	C5H7O3N	129.1155	2.10	9.47
Glutamine	-3.5	C5H8O2N2	128.1307	2.17	9.13
Glycine	-0.4	C2H3ON	57.0519	2.35	9.78
Histidine	-3.2	C6H7ON3	137.1411	1.77	9.18
Isoleucine	4.5	C6H11ON	113.1594	2.32	9.76
Leucine	3.8	C6H11ON	113.1594	2.33	9.74
Lysine	-3.9	C6H12ON2	128.1741	2.18	8.95
Methionine	1.9	C5H9ONS	131.1926	2.28	9.21
Phenylalanine	2.8	C9H9ON	147.1766	2.58	9.24
Proline	-1.6	C5H7ON	97.1167	2.00	10.60
Serine	-0.8	C3H5O2N	87.0782	2.21	9.15
Threonine	-0.7	C4H7O2N	101.1051	2.09	9.10
Tryptophan	-0.9	C11H10ON2	186.2132	2.38	9.39
Tyrosine	-1.3	C9H9O2N	163.1760	2.20	9.11
Valine	4.2	C5H9ON	99.1326	2.29	9.72

[1] the pKa is defined as the logarithm of the dissociation constant Ka[mol/L]: pKa $= -\log_{10}(\text{Ka})$

[a] pKa of the carboxylic-acid group

[b] pKa of the ammonium-group

1998; Cléry et al., 2008). After the analysis of several RNA-protein complexes Draper (Draper, 1999) confirms that some proteins recognize a specific sequence of ribonucleotides whereas other proteins "bind to RNA hairpins and loops" (Guzman et al., 1998).

- Some RBPs use so-called binding motifs to dock RNA. Binding motifs are specialised domains arranged on the protein surface and able to bind ribonucleotides. Many scientific works address the analysis of the domain specific binding. The most studied domains are probably the RNA recognition motif (RRM), the K-Homology (KH) domain, the Zinc binding domain, the double stranded RNA-binding domain (dsRBD) and the Pumilio homology domain (PUF or PUM-HD) (Glisovic et al., 2008; Auweter et al., 2006; Guzman et al., 1998; Draper, 1999; Cléry et al., 2008).

- Other RBPs, independently from the presence or from the absence of binding domains on the protein surface, bind specific sequences on the target RNA, called sequence motifs or consensus sequences (Ray et al., 2009). Some proteins can bind also structural motifs of the RNA sequence such as bulges or stem-loops (Jones et al., 2001; Draper, 1999; Kishore et al., 2010).

- Yet other RBPs use for binding neither motifs on the protein surface nor motifs on the target RNA: they dock to the RNA backbone phosphate or ribose group (Guzman et al., 1998; Draper, 1999).

The forces guiding the binding are similar to the forces which lead to the formation of the secondary and the tertiary structure of the protein. The binding is directed by several factors including hydrogen bonds, base-stacking, electrostatic and hydrophobic interactions between amino acids and ribonucleotides (Guzman et al., 1998; Draper, 1999; Pérez-Cano and Fernández-Recio, 2010; Jones et al., 2001; Auweter et al., 2006) . Also the 3-dimensional structure of the protein and its binding-pocket influences the binding: the "mutual accommodation of the protein and RNA-binding surfaces" (Draper, 1999) determines significantly the ligation (Auweter et al., 2006). The RNA can insert itself spatially and "…the side chains of the protein have to access the edges of the RNA bases" (Westhof and Fritsch, 2011).
It is known that RBPs can bind both "sequence specific" and "non-sequence specific" (Gupta and Gribskov, 2011) to their RNA targets and are able to select an RNA out of all the transcriptome (Draper, 1999). Moreover different RBPs can associate with a sequence on the target RNA (Glisovic et al., 2008) and vice-versa RBPs with the same binding surface may be able to associate with different target sequences (Cléry et al., 2008). These versatility makes it difficult to define an exact binding mechanism. Therefore laboratory experiments are necessary to detect an occurred binding between a protein of interest and the transcribed RNA.

2.1.4 Laboratory experiments

Different *in vivo* and *in vitro* experimental techniques are used in biology to detect RBP-binding targets (Glisovic et al., 2008). *In vivo* high-throughput techniques are able to identify binding interactions genome-wide and to observe the binding in the living cell. However each of these techniques has its strength and its weakness regarding experimental errors (Corden, 2010; Änkö and Neugebauer, 2012; Khalil and Rinn, 2011).

Systematic Evolution of Ligands by Exponential Enrichment (SELEX)

SELEX is an *in vitro* method whose goal is to identify the nucleic acid (DNA or RNA) sequences bound by an RBP of interest. Out of a pool of artificially generated RNA sequences the technique detects high-affinity targets. Kishore and colleagues (Kishore et al., 2010) argue that this high-affinity is not always the most specific one and Jeffry L. Corden (Corden, 2010) affirms that "...such short sequences have limited value in predicting *in vivo* binding sites". However a frequent application of SELEX is the detection of a consensus sequence in the experimental outcomes by searching for significant motifs. Several motif based sequence analysis tools implement algorithms which discover motifs in these sequences and calculate the corresponding position-specific scoring matrices (PSSM) (for a more detailed description see Section 2.2.3).

Cross-linking and immunoprecipitation assay (CLIP)

The *in vivo* detection of RBP-RNA interactions can be done by the "fixation", also called cross-linking, of the occurred binding in the living cell using UV light. After the cross-linking the cell is broken and the interacting elements are isolated. This technique is called *Cross-linking and immunoprecipitation assay (CLIP)*. The obtained pieces of transcripts can be analysed via high-throughput sequencing and is than called *HITS-CLIP (high-throughput sequencing CLIP)* (Zhang and Darnell, 2011). An advancement of the CLIP technique is *Photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP)* which facilitates the cross-linking and results in a much higher number of detected binding couples (Hafner et al., 2010). The advantage of these techniques is the detection of "real" binding in living cells and it is genome-wide. On the other hand it has been shown that certain RBPs are active only during specific cellular life-cycles (Hogan et al., 2008). Hence the binding is limited to the current cell stage, the currently transcribed genes and the tissue. Therefore *in vivo* techniques may not be able to catch all possible targets (Änkö and Neugebauer, 2012).

Computational methods have been developed ad hoc to align detected reads to the genome and to analyse the experimental outcomes. For example *PARalyzer* (Corcoran et al., 2011) can be used to analyse CLIP-data and in combination with motif-finding tools it can create

RNA-protein interaction maps. Also *Piranha* (Uren et al., 2012) detects genome-wide binding sites based from CLIP data.

2.2 RNA-protein binding. A computational point of view

"How proteins selectively bind specific sites on nucleic acids has been a challenging and interesting problem since the earliest days of molecular biology" (Draper, 1999). Therefore extensive studies on protein-protein and on DNA-protein interactions have been done in the past and only a decade later RNA-binding sites on proteins have been studied and have been dissected. These investigations revealed the diverse nature of RNA-binding sites compared to the well-studied DNA-binding sites.
In this section we review different computational attempts which study and predict RNA-RBP interactions:

- analysis of the binding site and the binding-interaction (Section 2.2.1): dissection and statistical analysis of the 3-dimensional binding complex to study the interaction mechanism and the binding preferences for each amino acid or the involved forces;
- prediction of single elements on the protein surface able to bind ribonucleotides (Section 2.2.2): machine learning techniques like Random Forest Method, Support Vector Machines or Neural Networks are applied to predict for each amino acid in a protein sequence if it binds to a ribonucleotide or not;
- tools to search significant patterns in the sequences i.e. binding motifs (Section 2.2.3);
- molecular recognition and docking simulations: these simulations are used to predict the RNA-protein binding based on structural information, free energy calculations, RNA plasticity and hydrogen-bonds. These approaches will not be addressed as they are beyond the scope of this thesis.

2.2.1 Dissection and analysis of RNA-protein complexes

To discover the mechanisms with which RBPs bind to their RNA-targets, the structures and the binding sites of 3-dimensional RNA-protein complexes (downloaded from PDB) have been analysed (Bahadur et al., 2008; Ellis et al., 2007; Jones et al., 2001). An accumulation of databases classify and store the results. The *Amino Acid-Nucleotide interaction database AANT* (Hoffman et al., 2004) creates sub-models of residue-nucleotide interactions and divides them in several classes. Similarly to *RsiteDB* (Shulman-Peleg et al., 2009) which, additionally to the interacting nucleotides, analyses also the interacting dinucleotides and predicts the binding pocket based on 3-dimensional consensus binding patterns. The *Biological Interaction* *database for Protein-nucleic Acid (BIPA)* (Lee and Blundell, 2009) just as *NPIDB (database of nucleic acids-protein interactions)* (Spirin et al., 2007) provide information about size, shape, residue propensity, secondary structure composition, domain and intermolecular interactions of the binding sites.

Statistical analysis of the 3-dimensional binding complexes revealed that even if there is no significant tendency of a residue to bind a specific ribonucleotide some combinations are preferred (Terribilini et al., 2006). Exploring amino acid propensities, secondary structure motifs and atom-atom contacts shows that Arginine and Lysine seem to be favoured within protein binding sites and van der Waals interactions seem much more frequent than hydrogen bonds (Jones et al., 2001; Bahadur et al., 2008; Morozova et al., 2006; Gupta and Gribskov, 2011; Pancaldi and Bähler, 2011). Considering the general negative charge of the RNA sequence, a tendency to find positively charged amino acids in the binding pocket was expected. Indeed Arginine and Lysine are positively charged amino acids. Different numbers are reported in literature regarding the interactions of amino acids: Hoffman et. al. (Hoffman et al., 2004) notes that in 23% of the cases residues bind to ribonucleotides ribose, 51% to the phosphate and only 26% to the base. This tendency is confirmed by other studies too (Morozova et al., 2006; Bahadur et al., 2008; Ellis et al., 2007; Pérez-Cano and Fernández-Recio, 2010; Gupta and Gribskov, 2011).

Analysis of hydrogen bonds (Jeong et al., 2003; Kim et al., 2003), secondary structure interactions (Jones et al., 2001), backbone contacts (Bahadur et al., 2008; Gupta and Gribskov, 2011), steric exclusions and binding pocket shapes (Morozova et al., 2006; Shazman et al., 2011) try to determine the mechanism of the binding interaction. And in fact protein atoms are able to form a dense binding pocket around the ribonucleotide to create a complementary shape between base and pocket. Also the stacking interactions seem to be determinant for RNA recognition (Morozova et al., 2006). Other investigations address the specific interaction of ribosomal proteins (Ciriello et al., 2010) and the difference between the binding pockets regarding the bound RNA-type (Ellis et al., 2007; Bahadur et al., 2008).

2.2.2 Binding residue prediction

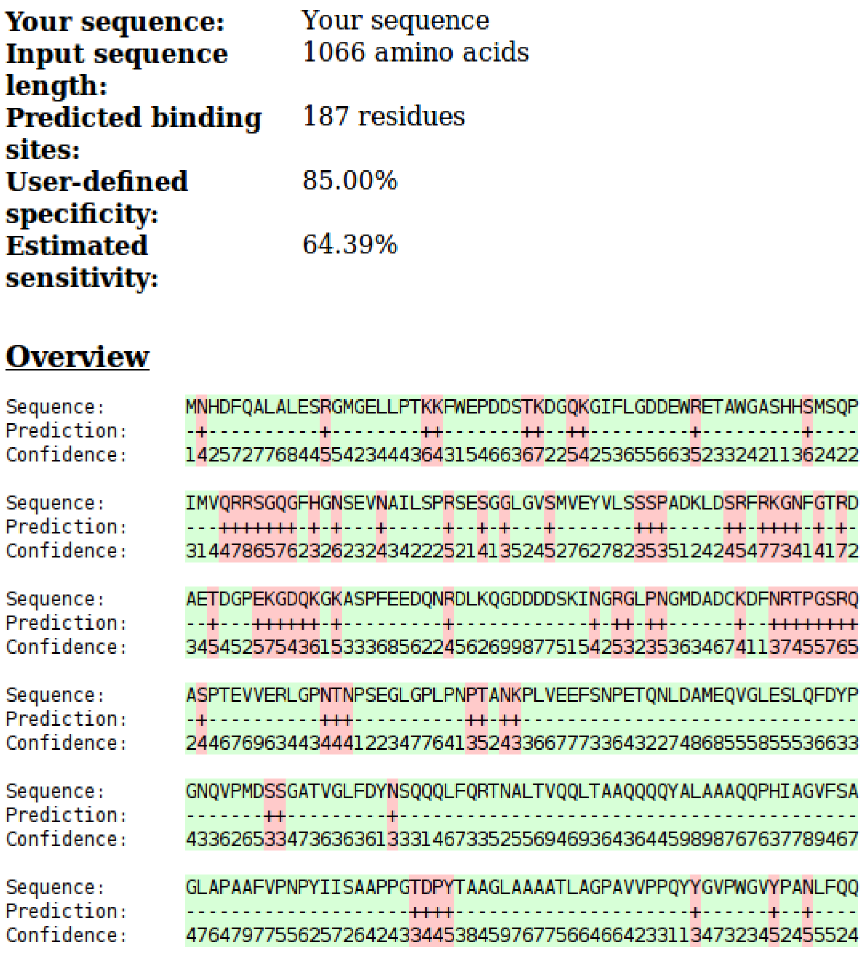
The basic idea common in these approaches is the prediction of the single binding residue in the protein sequence. The binding information is extracted from 3-dimensional RNA-protein complexes which have been downloaded from PDB (see Section 2.6), as it is the only database which provides 3-dimensional binding structures. The classification of each residue in "binding" and "non-binding" is done using tools like *HBPLUS* (McDonald and Thornton, 1994) or *Entangle* (Allers and Shamoo, 2001) which calculate the distances between amino acids and ribonucleotides within each complex. If an amino acid is spatially close to a ribonucleotide, which is defined by a threshold distance, it is defined as binding otherwise as non-binding. In this way a dataset with binding and non-binding residues is constructed and represents the positive and the negative examples to train classifiers.

Table 2.2 shows an overview of previously published works reporting their performance, the classifier and the used features. The aim of the table is to list the works and not to compare them. In particular the reported performance measures have no comparative function as the single classifiers are trained on different datasets and apply different validation procedures, for instance a 5-,10-, or 7-fold cross or leave-one-out validations.

Wang and Brown (Wang and Brown, 2006) propose an SVM-based approach, called *BindN*, which takes as input the protein sequence. Each amino acid is described by three biochemical features: the acid dissociation constant value (pKa) of the side chain, the hydrophobicity (H) index and the molecular mass (MM). The pKa specifies the acidity of an amino acid. Weak acids have a pKa value in range of -2 to 12, whereas a pKa value smaller than -2 is said to be a strong acid. The hydrophobicity is a chemical property and means the rejection against water molecules. Hydrophobic elements tend to avoid the contact with water. Then a model is trained with several sequence instances, where an instance represents a subsequence of a certain length w. From a protein sequence with n amino acid residues a total of $(n-w-1)$ data instances can be extracted using a sliding window technique. The target residue is positioned in the middle of the window and $\frac{(n-w-1)}{2}$ neighbour-residues on each side provide additional information. The target residue is labelled as positive when it is defined as "binding" otherwise it is labelled as negative. This sliding window technique is used in all methods to extract several data instances from the input sequence. *BindN* performs with a specificity of 0.69, a sensitivity of 0.66 and an accuracy of 0.69. An extended version, called *BindN+* (Wang et al., 2010), adds evolutionary information in form of position-specific scoring matrices (PSSMs) and uses the mean and the standard deviation of the three biochemical features introduced before. Combining these features results in an increased performance. Figure 2.1 shows an example output of *BindN+*.

A similar approach but based on Naïve Bayes classifiers is implemented in *RNAbindR* (Terribilini et al., 2006). Each amino acid is codified in a binary way and additionally described by its relative accessible surface area (rASA), sequence entropy, H, secondary structure (SS) and electrostatic potential. Comparisons with a previously published neural network classifier (Jeong et al., 2004) did not show significant improvements. *PPRInt* (Kumar et al., 2007) was among the first methods to introduce the use of PSSM profiles in the prediction of binding sites in RBPs. A PSSM profile is generated by running a PSI-Blast (Altschul et al., 1997) search against a non-redundant database of protein sequences. The PSSM indicates how probable is the appearance of an amino acid at a certain position in the sequence. In *PPRInt* SVMs are trained with different features: one with the amino acid sequence and one with PSSM profiles. The latter performed better with a specificity of 0.89 and a sensitivity of

Figure 2.1: **Output example of BindN+.** *BindN+* output for the protein sequence PUM2_HUMAN (Q8TB72). Predicted binding residues are coloured in red, predicted non-binding residues are coloured in green.



Your sequence:	Your sequence
Input sequence length:	1066 amino acids
Predicted binding sites:	187 residues
User-defined specificity:	85.00%
Estimated sensitivity:	64.39%

Overview

```
Sequence:    MNHDFQALALESRGMGELLPTKKFWEPDDSTKDGQKGIFLGDDEWRETAWGASHHSMSQP
Prediction:  -+----------+--------++-------++--++---------+---------+----
Confidence:  142572776844554234443643154663672254253655663523324211362422

Sequence:    IMVQRRSGQGFHGNSEVNAILSPRSESGGLGVSMVEYVLSSSPADKLDSRFRKGNFGTRD
Prediction:  ---+++++++-+-+---+-----+--+-+---+-------+++-----++-++++-+-+-
Confidence:  314478657623262324342225214135245276278235351242454773414172

Sequence:    AETDGPEKGDQKGKASPFEEDQNRDLKQGDDDDSKINGRGLPNGMDADCKDFNRTPGSRQ
Prediction:  --+---++++++-+---------+------------+-++-++------+--++++++++
Confidence:  345452575436153336856224562699877515425323536346741137455765

Sequence:    ASPTEVVERLGPNTNPSEGLGPLPNPTANKPLVEEFSNPETQNLDAMEQVGLESLQFDYP
Prediction:  -+----------+++----------++-++------------------------------
Confidence:  244676963443444122347764135243366777336432274868555855536633

Sequence:    GNQVPMDSSGATVGLFDYNSQQQLFQRTNALTVQQLTAAQQQQYALAAAQQPHIAGVFSA
Prediction:  -------++---------+-----------------------------------------
Confidence:  433626533473636361333146733525569469364364459898767637789467

Sequence:    GLAPAAFVPNPYIISAAPPGTDPYTAAGLAAAATLAGPAVVPPQYYGVPWGVYPANLFQQ
Prediction:  --------------------++++---------------------+------+--+----
Confidence:  476479775562572642433445384597677566466423311347323452455524
```

0.53. Similar feature combinations are applied in *RISP*(Tong et al., 2008), *RRINTR* (Wang et al., 2008) or *ProteRNA* (Huang et al., 2010). PSSM profiles which incorporate evolutionary information are a common features not only in the prediction of RNA-protein interactions, but also in the prediction of DNA-protein and protein-protein interactions. *RNAProB* (Cheng et al., 2008) went a step further and adopted a "smoothed PSSM", created on a traditional PSSM profile and considering the PSSM neighbourhood of the target residue. The performance of the classifier applied on three different benchmark datasets is rather high with an average specificity of 0.88 and an average sensitivity of 0.74.

Each amino acid has its own physical and chemical characteristics (see Section 2.1.2) represented by properties like hydrophobicity or molecular mass, which may determine its ability to be involved in binding or not. Therefore it can be more likely to find specific residues in the binding site then elsewhere. Following this reasoning *PiRaNhA* (Spriggs et al., 2009) adds the interface propensity (IP) to its features. A similar approach is integrated also in *PRNA* (Liu et al., 2010) which uses a Random Forest (RF) classifier. The residues are encoded, besides other features, as interaction propensities extended to three amino acids. Taking into account also the nearest neighbourhood of the considered residue the IP is calculated on triplets instead on single residues. Subtracting one descriptor after another the feature importance can be evaluated. The triplet IP and structural features like accessible surface area (ASA) and secondary structures turned out to be the most powerful features in *PRNA*.

More sophisticated approaches, instead of using the residue of interest and its sequential neighbours to investigate the binding, use surface patches and clefts which surround the target residue. Patches and clefts consider the 3-dimensionality of the binding pocket and the binding surface. In that respect a surface patch is defined as RNA-interacting if a limited amount of accessible surface residues belong to it. Whereas a binding-cleft, including cavities and grooves, contains at least 10 accessible and interacting residues (Chen and Lim, 2008). *PRIP* (Maetschke and Yuan, 2009) applies sequential, graph-topological and spatial information to individuate the binding patches. The application of both SVM and Naïve Bayes classifier confirms that the former performs better.

The above presented approaches have the final goal to predict RNA-binding residues on the protein surface taking as input only the protein sequence. All the features are based on the sequential, structural and evolutionary identity of the protein and focus on the binding site. A recent study (Pancaldi and Bähler, 2011) attempts a more global approach and analyses different features in correlation with RBPs, their target RNAs and their specific association by using: mRNA properties such as UTR characteristics, RNA structure, translational features, expression level; protein properties such as physical-chemical features and Gene Ontology associations. After the detection of the most important features, SVM and RF applied to predict the interactions.

Table 2.2: **RNA-binding site prediction methods.** The approaches attempt to predict RNA-interacting residues in protein sequences. Listed are name, classifier, applied features and some performance values: accuracy (Acc), sensitivity (Sens), specificity (Spec), area under the ROC curve (AUC), correlation coefficient (CC) and net prediction (NP). Shown is the performance of the best feature combination presented in the paper. This table is not intended for comparison.

Name	Classifier	Features[a]	Performance measures
BindN	SVM (Wang and Brown, 2006)	pKa, H MM	Acc=0.69, AUC=0.73 Sens=0.66, Spec=0.69
RNAbindR	Naïve Bayes (Terribilini et al., 2006)	amino acid, rASA sequence entropy, H SS, electrostatic potential	Acc=0.76,CC=0.30 Sens=0.43, Spec=0.47
PPRInt	SVM (Kumar et al., 2007)	amino acid PSSM	MCC=0.45, AUC=0.81 Sens=0.53, Spec=0.89
RISP	SVM (Tong et al., 2008)	amino acid PSSM	CC=0.35, NP=0.72 Sens=0.61, Spec=0.83
PrintR	SVM (Wang et al., 2008)	amino acid, PSSM SS, ASA	Acc=0.87, AUC=0.83 MCC=0.45, Sens=0.48
RNAproB	SVM (Cheng et al., 2008)	Smoothed PSSM	Acc=0.87, MCC=0.68
PiRaNhA	SVM (Spriggs et al., 2009)	PSSM, H predicted accessibility interface propensity	Acc=0.87, AUC=0.86 Sens=0.56, Spec=0.92 MCC=0.49
ProteRNA	SVM (Huang et al., 2010)	PSSM, SS sequence conservation	Acc=0.89, MCC=0.26 Sens=0.25, Spec=0.96 Prec=0.39

[a] H= hydrophobicity, SS=secondary structure, pKa=acid dissociation constant, ASA=accessible surface area, rASA=relative ASA, PSSM=position specific scoring matrix, MM=molecular mass

Name	Classifier	Features[a]	Performance measures
BindN+	SVM (Wang et al., 2010)	PSSM mean±std of pKa mean±std of H mean±std of MM	Acc=0.77, AUC=0.82 Sens=0.71, Spec=0.78 MCC=0.39
PRNA	Random forest (Liu et al., 2010)	rASA, SS, PSSM pKa, H, triplet IP number of atoms electrostatic charge hydrogen bonds	Acc=0.82, MCC=0.48 Sens=0.81, Spec=0.86
no name	SVM (Wang et al., 2011)	Smoothing PSSM, ASA pKA, H, MM hydrophobic moment net charge index of side chain net charge index moment propensity propensity moment	Acc=0.88, MCC=0.68 Sens=0.78, Spec=0.91
no name	SVM (Choi and Han, 2011)	amino acid frequency, H amino acid position, ASA chain length, MM, pKa triplet IP, partner information	Acc=0.90, CC=0.24 Sens=0.60, Spec=0.91 NP=0.75

[a] H= hydrophobicity, SS=secondary structure, pKa=acid dissociation constant, ASA=accessible surface area, rASA=relative ASA, PSSM=position specific scoring matrix, MM=molecular mass

Many of the previous methods predict binding residues without considering the RNA-binding partner. They give no information about the RNA-sequence potentially bound by an RBP. Only *RPISeq* (Muppirala et al., 2011) addresses this issue and predicts the binding between a given RBP and its RNA-target using both SVM (*RPISeq-SVM*) and RF (*RPISeq-RF*) method. The classifiers are trained with simple sequence features using Conjoint Triad Feature (CTF) of length 3 to encode the protein and oligonucleotides (oligos) of length 4 to encode the RNA sequence. In CTF the amino acids are divided in seven different groups reducing the alphabet from 20 elements to 7 elements. A protein is encoded with 343 features by counting the (normalized) frequency of all occurring triplets. The RNA sequence is encoded similarly by counting and normalizing the frequencies of all occurring oligos of length 4. Finally each RBP-RNA pair is represented by 599 features. Two different datasets are created using interacting RBP-RNA complexes as positive examples, and randomly coupled "non-interacting" RBP-RNA pairs as negative examples. The first dataset contains mainly interactions with ribosomal proteins and ribosomal RNAs and the second dataset only non-ribosomal complexes. In a 10-fold cross validation *RPISeq-RF* performs better on both datasets than *RPISeq-SVM*. Both classifiers achieve high performance values when applied on the ribosomal-dataset with an accuracy of 0.89 and 0.87, for RF and SVM respectively. The RF-model achieves a precision of 0.89 and an AUC of 0.97, whereas the SVM applied on the ribosome dataset achieves a precision of 0.87 and an AUC of 0.92. Much lower values are reached on the second dataset without ribosomal complexes: the accuracies vary between 0.76 (*RPISeq-RF*) and 0.72 (*RPISeq-SVM*). The precision does not exceed the threshold of 0.75 and also the AUC values with 0.85 (RF) and 0.81 (SVM) are lower compared to the ribosomal-dataset.

2.2.3 Motif finding tools

Some RBPs can bind specific sequence patterns on their targets, called motifs. Motifs are patterns in RNA, DNA or protein sequences which can be modelled by position-specific scoring matrices, called PSSMs. PSSMs can be used to describe potential binding sites and are also applied to identify evolutionary similar proteins and to discover evolutionary conserved functional sites (Lichtarge and Sowa, 2002). Experimentally detected binding motifs are not always available, but there exist *in silico* tools which search for significant patterns in a group of RNA sequences known to be bound by an RBP.

The MEME Suite (Bailey et al., 2009) provides several motif-based sequence analysis tools, for example *MEME* (Bailey and Elkan, 1994) which detects motifs in a set of sequences. Based on the provided input *MEME* calculates the position-dependent letter probabilities of the discovered motif. These probabilities inform about the appearance of each letter in each position within the motif.

RNAcontext (Kazan et al., 2010) discovers structural motifs within related sequences. The tool searches for significant sequential and structural binding preferences in a set of target and non-target examples by using a structural context alphabet (i.e. paired, unpaired, hairpin loop) based on secondary structure predictions. Finally the model can be applied to detect the identified motif in a set of unknown sequences.

2.3 Machine learning

New high-throughput technologies and next-generation sequencing produce a huge amount of biological data. These data needs to be stored and to be searchable for biologically relevant questions. Artificial intelligence and machine learning are applied to mine, to explore, to analyse and to extract the knowledge contained in these genomic, proteomic or transcriptomic data. Hence machine learning is an important field in bioinformatics (Inza et al., 2010; Jensen and Bateman, 2011).
SVMs (Vapnik, 1995) are one of the most used supervised machine learning techniques and have been widely applied in bioinformatics (see Section 2.2). The basic idea is that the classifier learns a mathematical function on input examples and classifies than new unknown examples.

2.3.1 Support vector machine (SVM)

The SVM classifier discriminates linearly between input vectors $x_i \in \mathbb{R}^p$, with p being the number of features. The input examples $i = 1, 2, ..., n$ are associated with different classes $y_i \in$ {+1,-1}: an input vector x_i belongs to the positive class with label $+1$ or to the negative class with label -1. The goal of a SVM is to find a discrimination function $f(x)$ which divides the two classes in such a way that the label for new vectors can be predicted: $f(x) = sign(< \mathbf{w}, x > +b)$ where $\mathbf{w}$ is the weight vector, the scalar b the bias and $sign$ returns the sign of the argument. $f(\mathbf{x}) = 1$ assigns the positive class label, otherwise the negative one.
During the training the hyperplane tries to divide the two classes linearly. This can be controlled by parameters that provide a more flexible classification of the input vectors. The "softmargin" approach introduces so-called slack variables allowing an example to be misclassified. The use of these slack variables can be regulated by the constant $C > 0$. The SVM can map the input space into a feature space using kernel-functions. To implement the classifier we use the freely available SVM package *LibSVM* (Chang and Lin, 2011) and apply both the linear kernel and the RBF kernel.

Fast Local Kernel Support Vector Machine (FaLK-SVM) apply a set of SVMs to create an appropriate local model on each training point. We use the freely available software package

FaLKM-lib (Segata, 2009) with the linear kernel.

2.4 Data balancing

A frequent problem with biological datasets is that they are unbalanced. Usually the number of negative data points is much higher than the number of positive ones (Terribilini et al., 2006; Cheng et al., 2008; Wang and Brown, 2006). Machine learning techniques learn on input examples and using highly-unbalanced datasets affects their performance. To overcome the unbalancing different approaches are proposed in literature such as several forms of data re-sampling (over-sampling, under-sampling), one-class learning or by using different class weights (Kotsiantis et al., 2006). One re-sampling method is a synthetic minority over-sampling technique called *SMOTE* (Chawla et al., 2002). The method amplifies the positive dataset by creating new synthetic instances and forces the classifier to become more general.

2.5 Performance measures

A binary classifier like the SVM assigns to predicted binding sequences the positive class label ($+1$) and to sequences predicted as non-binding the negative class label (-1). Correct assignments to the positive or the negative class increase the numbers of the true positives (TP) or the true negatives (TN), respectively. Wrongly attributed elements increase the false negatives (FN) or false positives (FP). Different measures can be calculated to assess the performance of a classifier. In this thesis we use:
Accuracy (Acc) is the rate of correct and false predicted elements over the total dataset

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.1}$$

A weakness of the ACC is that its value can be high even if one class is never or poorly predicted. For instance when all elements of the negative class are predicted as negatives but no positive element was assigned to the positive class.
The Matthews correlation coefficient (MCC) instead is a balanced measure and indicates the correlation between observed and predicted classification

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{2.2}$$

The sensitivity

$$Sens = \frac{TP}{TP + FN} \tag{2.3}$$

measures the ability of the classifier to identify positive elements whereas the specificity

$$Spec = \frac{TN}{TN + FP} \tag{2.4}$$

measures the proportion of correctly classified negative elements.
The precision (Prec), also called positive predictive value, indicates the portion of positive classified examples that are really positive:

$$Prec = \frac{TP}{TP + FP} \tag{2.5}$$

The creation of the receiver operating characteristic (ROC) curve is a common way to visualize a model performance. The x-axis shows the false positive rate and the y-axis displays the true positive rate by varying a parameter, in our case the classification threshold. True positive rate and false positive rate are defined as

$$TPR = \frac{TP}{TP + FN} \tag{2.6}$$

and

$$FPR = \frac{FP}{FP + TN} \tag{2.7}$$

respectively. More the ROC curve advances to the upper-left corner, the better is the classification ability of the model. A curve near the diagonal characterizes a “random” classification (Fawcett, 2006). A similar visualization gives the Precision-Recall (PR) curve showing the precision on the y-axis and the recall on the x-axis:

$$Recall = \frac{TP}{TP + FN} \tag{2.8}$$

More the PR curve advances to the upper-right corner, the better is the prediction of the model (Davis and Goadrich, 2006). By calculating the area under the ROC curve (AUC) the performance of a classifier can be reduced to a single value. A perfect classifier has an AUC of 1.

2.6 Biological databases

In this section we describe some databases which are relevant in this thesis.

2.6.1 The University of California, Santa Cruz (UCSC) Genome Browser

UCSC (Fujita et al., 2011) gives access to different databases which store sequences and annotations for several genomes. Different utilities and tools are provided to search, to visualize and to analyse the data. The collection contains linked data regarding genes, their transcripts and protein products deriving from well-known sources like GenBank (Benson et al., 2013), Ensembl Genes (Flicek et al., 2010), RefSeq (Pruitt et al., 2007) or the UniProt Knowledgebase (UniProtKB) (Magrane et al., 2011).

2.6.2 The Ensembl project

Another collection of databases with genomic material of about 60 species (September 2011), annotations and sequence data is "The Ensembl project" (Flicek et al., 2010). Ensembl provides links to UCSC Genome Browser and integrates DNA data, genes, transcripts and information about sequence variations or regulation. Tools such as BioMart (Smedley et al., 2009) can be used to search across the databases and to perform complex queries.

2.6.3 The RCSB Protein Data Bank (PDB)

PDB (Bernstein et al., 1977) stores 3-dimensional macromolecular crystal structures obtained by nuclear magnetic resonance, X-ray crystal structure determination, cryo-electron microscopy and theoretical modelling (Berman et al., 2000). Each entry is identified by a unique accession number and the file format contains the atomic coordinates for each structure. The 3-dimensional structure can be visualized by molecular viewers. Beside this other information regarding the macromolecule is stored, e.g. the sequence information, molecular functions or annotations, referencing other databases such as UniProtKB.

2.6.4 The Atlas of UTR regulatory activity (AURA)

The manually curated AURA (Dassi et al., 2012) is a database containing information about human UTRs. Currently (February 2013) it contains the binding data for 103 RBPs and more than 127000 UTRs, deriving from 63138 transcripts of about 29000 genes. The online database provides information regarding RBPs, their function and their bound 3'UTR and 5'UTR sequences. The datasets and the binding information derive from laboratory experiments like CLIP, PAR-CLIP or SELEX. AURA can be searched in different ways: by searching directly for UTRs, by searching for UTRs bound by an RBP of interest or by using the BioMart-equivalent called AURA Mart.

2.6.5 The National Centre for Biotechnology Information (NCBI) resources

The NCBI resources are accessible online and offer a variety of tools for sequence analysis, tools for data mining and the access to many literature and molecular databases. Important databases are for example GenBank or RefSeq. The "non-redundant" protein database (nr-database) contains non-redundant sequences from GenBank (Genbank CDS translations) together with protein sequences from Refseq, PDB and UniProtKB. The nr-database can be used to perform a Blast-search.

The *Basic Local Alignment Search Tool (BLAST)* (Altschul et al., 1997) has been developed at the NCBI and searches for genes and proteins in databases. The algorithm tries to find similar gene or protein sequences and queries the sequence in question against large datasets. BLAST can also be used to identify functional and evolutionary sequence information. Several types of BLAST are available: to search for a protein sequence in a protein database or to search for a protein sequence in a nucleotide database. A special way to search in databases is the use of PSI-BLAST (Position-Specific Iterated BLAST) (Altschul et al., 1997) which creates a PSSM by matching the query sequence against the database. Afterwards the PSSM can be used to search for evolutionary similar sequences.

2.6.6 The Gene Expression Omnibus (GEO)

A slightly different kind of database is *The Gene Expression Omnibus (GEO)* (Edgar et al., 2002). It is a repository for high-throughput data on the NCBI. Experimental outcomes can be uploaded and than be searched by dataset (for example RBP name) or accession number.

Chapter 3

Preliminary analysis

Predicting amino acids to be involved in ribonucleotide-binding is challenging and has been addressed conspicuously often (see Section 2.2). Machine learning techniques like SVM, Random Forest or Naïve Bayes have been trained with sequence features, like molecular mass (MM), hydrophobicity (H), the acid dissociation constant value (pKa) or secondary structures (SS); with evolutionary features, like position-specific scoring matrix (PSSM) or with 3-dimensional structure-properties, like the accessible surface area (ASA) (Wang and Brown, 2006; Kumar et al., 2007; Wang et al., 2008). These approaches do not look to the RNA sequence which can be bound by the analysed RBP but focus on the prediction of binding elements in the protein sequence. Only recently published methods involve also the RNA sequence in the binding prediction by using both, protein and RNA sequences (Muppirala et al., 2011) or global features like UTR characteristics or expression levels (Pancaldi and Bähler, 2011). Considering the predicted binding amino acid it should be possible to detect the bound RNA target-sequence. Therefore our preliminary analysis addresses the prediction of the RNA binding partner.
In this Chapter we give a formal description of the binding, describe our approach and present the obtained results. Then we discuss them and conclude by outlining the next steps.

3.1 Abstraction of the binding

In order to give a precise description of the problem we define the main concepts involved in binding. The RNA is a consecutive sequence of n nucleotides $dNTP$

$$dNTP_0 \dots dNTP_i \dots dNTP_{n-1}$$

The i-th nucleotide $dNTP_i$ is composed of a base, a phosphate and a ribose sugar. There are four types of bases: Adenine (A), Cytosine (C), Guanine (G) and Uracil (U). For instance an RNA sequence of length 10 can be: ACGUUCGGAA. The protein P is a consecutive and folded sequence of m amino acids. There are 20 types of amino acids

$$aa_v \in \{\mathit{A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V}\}$$

with $v = 1 \ldots 20$. Every amino acid aa_v can be described by some physical-chemical properties prop(aa_v)$\in$ {*polarity, molecular mass, hydrophobicity, pKa, charge*}. The protein can form several structures: the primary (1str), secondary (2str) and tertiary (3str) structure. To form the primary structure *1str* of the protein the amino acids bind together through the so-called peptide bound. After this connection the single amino acid aa_v is called residue r_j at the sequence position j inside the protein chain

$$r_0 \ldots r_j \ldots r_{m-1}$$

The protein chain can form the secondary structure *2str* by shaping in the secondary structures α-helices, β-sheets or coils. Thus a secondary structure $ss_l \in \{\alpha, \beta, coil\}$ can be assigned to every r_j within the protein P. Suppose that $aa_v \in \{A, R, N, D, C...\}$ and r_j is the amino acid aa_v at the j-th position in the primary structure of a protein P, than:

$$\begin{aligned} 2str_P \colon \{A, R, N, D, C, \ldots\} &\rightarrow \{\alpha, \beta, coil\} \\ 2str_P(r_j) &\mapsto ss_l \end{aligned}$$

For our purpose it is sufficient to define the tertiary structure *3str* as the transformation of the amino acid r_j into a 3-dimensional space with the coordinates x,y,z:

$$\begin{aligned} 3str_P \colon \{A, R, N, D, C \ldots\} &\rightarrow R^3 \\ 3str_P(r_j) &\mapsto (x, y, z) \end{aligned}$$

An interaction is a kind of physical-chemical phenomenon which occurs between two or more objects. The interaction between a nucleotide $dNTP_i$ and a residue r_j is defined as I_{ij} where $I_{ij} \in \{van\ der\ Waals,\ electrostatic\ interaction,\ hydrogen\ bond,\ base\ stacking\}$. One type of interaction can be stronger than another. The binding is the result of an interaction and ends in a stable association and in the formation of a molecular complex. A complex C is the collectivity of all residues r_j of the protein P and all nucleotides $dNTP_i$ of the RNA which are involved in an interaction I_{ij} between r_j and $dNTP_i$. In this context the protein binding is a stable association of a specific piece on the protein, called binding site, to a specific piece on the RNA strand, called RNA-target sequence. The formed structure is called protein-RNA complex where the binding site of the protein P is defined as a subset of the protein chain $r_0 \ldots r_{m-1}$, consisting of all k residues

$$b_0 \ldots b_{k-1}$$

which are involved in an interaction with the nucleotides of the target RNA. The target sequence of an RNA is defined as a subset of the RNA sequence $dNTP_0 \ldots dNTP_{n-1}$, consisting of all w nucleotides

$$t_0 \ldots t_{w-1}$$

which are involved in an interaction with a residue. Given a target and a binding site the interaction I_{ij} between the nucleotide t_i with $0 \leq i \leq w-1$ and the binding residue b_j with $0 \leq j \leq k-1$ is denoted as

$$t_i \xrightarrow{I_{ij}} b_j$$

Given the formed interactions $t_i \xrightarrow{I_{ij}} b_j$ between every $t_i \in t_0 \ldots t_{w-1}$ and every $b_j \in b_0 \ldots b_{k-1}$ the binding is defined as the set of all interactions formed in the complex.

3.2 The Approach

Our approach exploits a binding-residue prediction and uses the predicted elements to construct the bound RNA sequence. The bound RNA sequence is constructed by using a previously created propensity statistic. If a residue is predicted as binding we will use the propensity statistic to detect its favoured nucleotide binding partner and create the targeted RNA sequence.

First we perform a simple sequence analysis on the dataset, extract all amino acid triplets which are present in the protein sequences and calculate for each its interaction propensity with ribonucleotides. In other words we detect triplets (r_{j-1}, r_j, r_{j+1}) and check for each of them if the target residue r_j is involved in binding $dNTP_i \xrightarrow{I_{ij}} r_j$. In such case we save the triplet with its target residue r_j and its nucleotide binding partner $dNTP_i$. Hence we detect the binding preference for each triplet and construct its propensity table.

To implement the binding-residue prediction we use SVMs and apply different features. The goal is to predict the binding elements of the binding site $b_0 \ldots b_{k-1}$ on the RBP. Evolutionary information like PSSM based features seem to be more powerful than features based on the simple amino acid sequence (see Table 2.2). Beside this each amino acid has its own physical and chemical characteristic which strongly determine the availability of binding. Therefore in our prediction we apply different properties to describe each amino acid: the MM, the pKa value of the carboxyl and the amino group, number of atoms, electrostatic charge, number of potential hydrogen-bonds and H. It is known that a protein binds through its binding-pocket or through surface residues which means that the accessibility of an amino acid is an important condition for binding. If an amino acid is buried no interaction with a nucleotide can occur. For this reason we include also ASA into the features. To assess which feature combination fits better with our dataset we test five classifiers:

- *SVM1.1*: MM, number of atoms, electrostatic charge, number of potential hydrogen-bonds and two pKa values;
- *SVM1.1Beta*: MM, electrostatic charge, H and two pKa values;
- *SVM1.2*: MM, number of atoms, electrostatic charge, number of potential hydrogen-bonds, two pKa values and ASA;
- *SVM1.3*: MM, number of atoms, electrostatic charge, number of potential hydrogen-bonds, two pKa values, ASA and SS;
- *SVM1.4*: MM, number of atoms, electrostatic charge, number of potential hydrogen-bonds, two pKa values, ASA, SS and PSSM;

The classifier with the best performance is chosen for our final approach.
The last step is the construction of the target sequence $t_0 \dots t_{w-1}$ on the RNA which is bound by the predicted binding-residues of the binding site $b_0 \dots b_{k-1}$ on the RBP. If r_j of the triplet (r_{j-1}, r_j, r_{j+1}) is predicted as binding we search the triplet in the propensity table and go the way backward: $dNTP_i \xleftarrow{I_{ij}} r_j$. In this way we construct a target sequence. Figure 3.1 illustrates the described idea.

3.3 Material and Methods

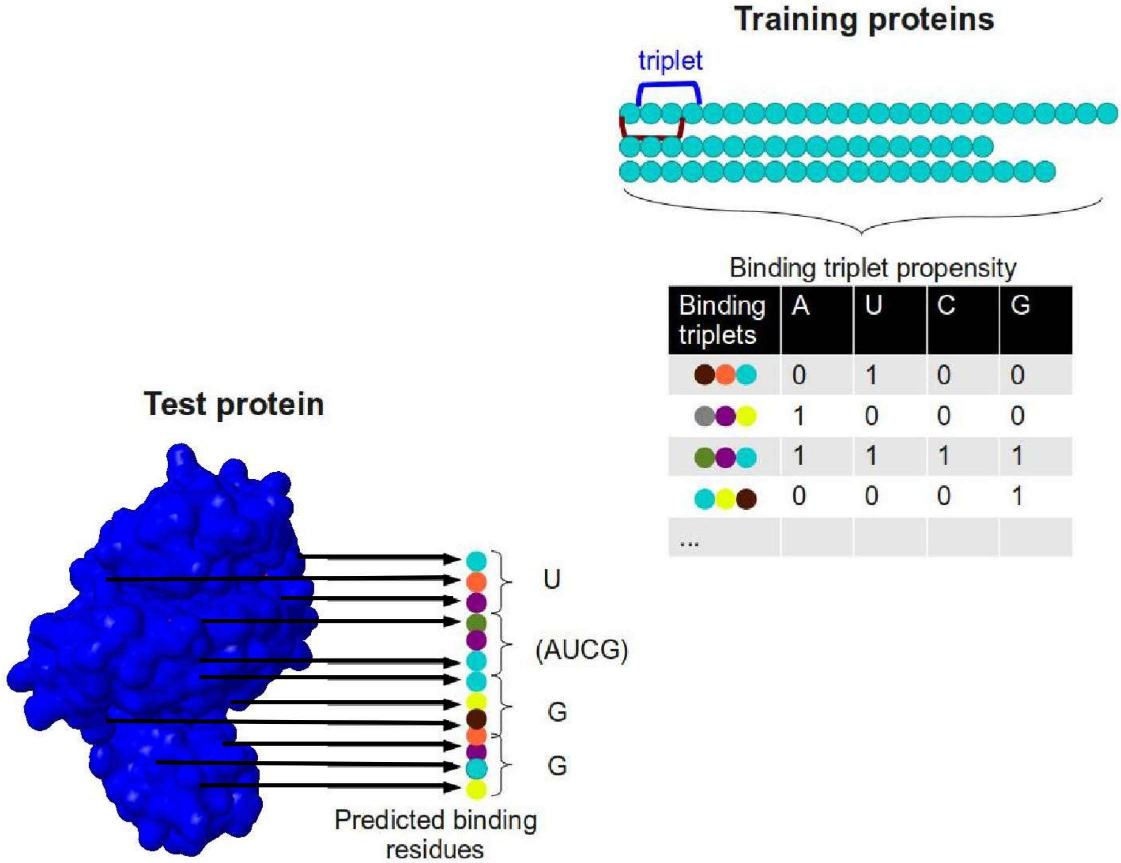
3.3.1 Dataset

Our dataset is composed of 50 protein files (Terribilini et al., 2006; Kumar et al., 2007) and contains 3-dimensional RBP-RNA binding complexes downloaded from PDB (Bernstein et al., 1977), with a resolution higher than 3Å . A pdb file can contain multiple protein and RNA chains. Therefore strands which are not involved in any kind of intermolecular interaction are eliminated. HBPLUS (McDonald and Thornton, 1994) could not resolve the structure for three files (1FJG.pdb, 1H38.pdb, 1JJ2.pdb), so they are cancelled from the dataset. Table 3.1 summarizes the properties of the dataset.
Test proteins: to carry out the test we use six randomly selected protein chains from the dataset which will not make part of the training. These are: 1A9N.pdb, 1ASZ.pdb, 1AV6.pdb, 1B7F.pdb, 1B23.pdb and 1C0A.pdb.

3.3.2 Binding Residue Identification

To identify a binding residue we check if it participates in an interaction I_{ij}. HBPLUS is used to calculate the attractions and atom-atom contacts within the RNA-protein complexes.

Figure 3.1: **RNA-sequence construction based on predicted binding residues.** The main idea of our preliminary analysis: the interaction propensity between amino acid and ribonucleotide is calculated on a set of training proteins. This propensity is finally used to construct a consequential RNA target-sequence based on the predicted binding residues on a test protein.


Training proteins
triplet
Binding triplet propensity
Test protein
U
(AUCG)
G
G
Predicted binding residues

Binding triplets	A	U	C	G
●●●	0	1	0	0
●●●	1	0	0	0
●●●	1	1	1	1
●●●	0	0	0	1
...				

Table 3.1: **Dataset description.**

Number of protein files	50
Number of protein chains	116
Number of amino acids	26276
Number of binding triplets	1452
Number of non-binding triplets	11405
Number of RNA chains	91
Number of nucleotides	3188
Number of RNA binding triplets	605
Number of RNA non-binding triplets	1050

The cut-off distance defines the maximal allowed distance between an interacting amino acid and a ribonucleotide. Different cut-off distances are used and vary from 2.5Å (Kumar et al., 2007; Wang and Brown, 2006) to 5Å (Terribilini et al., 2006). Analyses of the distribution of misclassified non-binding residues recommend a distance between 3.5Å and 8.5Å (Tong et al., 2008), therefore we decide to use a cut-off distance of 3.9Å (Wang et al., 2008; Spriggs et al., 2009; Liu et al., 2010). Ribonucleotides which are spatially near an amino acid, within this cut-off distance, are labelled as binding residues otherwise as non-binding residues. We extract in total 11405 non-binding and 1452 binding residues. Running the classifier on such an unbalanced dataset caused a constant non-binding prediction for each residue. Therefore we choose randomly 1452 negative elements to adapt the number of negatives to the number of positive residues.

3.3.3 Feature Extraction and Representation

PSSM

A PSSM is created by running a PSI-BLAST (Altschul et al., 1997) search against the nr-database downloaded from NCBI. With each sequence in our dataset we perform a PSI-BLAST search to obtain the PSSM of the corresponding protein.

Physical-chemical amino acid properties

Each amino acid has individual physical and chemical properties. We think that these characteristics strongly determine the willingness to bind. Hence we use the following features to describe each amino acid:

1. molecular mass (Table 2.1);
2. pKa of the carboxyl group (Table 2.1);
3. pKa of the amino group (Table 2.1);
4. number of atoms (Liu et al., 2010);
5. electrostatic charge (Liu et al., 2010);
6. number of potential hydrogen-bonds (Liu et al., 2010);
7. hydrophobicity (Table 2.1);

ASA

DSSP (Kabsch and Sander, 1983) is applied to calculate the ASA. The program takes as input the 3-dimensional complex and calculates the accessible area for each amino acid on the protein surface by using the atomic coordinates of the pdb file.

Secondary Structure

To extract the secondary structure of the protein, which is given in the pdb-file, *DSSP* is used as well. The secondary structures are divided into three groups: α-helix, β-sheet and coil. Each of them is encoded binary in a 3-dimensional vector (Wang et al., 2008) where the corresponding vector field is set to 1: for α-helix (100), for β-sheet (010) and for coil (001). If there is no structure assigned the vector-elements are all zero (000).

Classifier Evaluation

LibSVM (Chang and Lin, 2011) package is used to construct the SVM classifier. The best parameter combination of cost C and RBF-kernel parameter γ is optimized with respect to the highest value of

$$\frac{(sensitivity + specificity)}{2}$$

To evaluate our model we perform a 10-fold cross validation and calculate sensitivity, specificity and accuracy. Our dataset is strongly unbalanced with 11405 non-binding and 1452 binding residues. Therefore we balance the dataset by choosing randomly 1452 residues for each class.

3.4 Results and Discussion

Five classifiers with different feature combinations have been applied to our dataset to predict the binding amino acids within a protein sequence. The result for each model is shown in Table 3.2. SVM1.1 and SVM1.1Beta are the simplest classifiers based only on a few features. SVM1.1 performed worst with a sensitivity of 0.57 whereas SVM1.1Beta goes slightly better with a sensitivity of 0.62. SVM1.4 includes evolutionary information in form of PSSM and reaches, as expected, the highest performance with a sensitivity of 0.84 and a specificity of 0.70. Therefore the feature combination of classifier SVM1.4 fits best the dataset and will be used in the following to predict the binding amino acids. To check the performance of SVM1.4 we apply it to 6 test proteins and predict their binding residues. The predicted and the real binding residues, detected by HBPLUS within the complexes, are reported in Table 3.3. For instance when applied on test protein 1A9N SVM1.4 predicts only 40 out of 71 triplets as

Table 3.2: **Performance measures of five different feature combinations.** *SVM1.1*: MM, number of atoms, electrostatic charge, number of potential hydrogen-bond and two pKa values. *SVM1.1Beta*: MM, electrostatic charge, H and two pKa values. *SVM1.2*: MM, number of atoms, electrostatic charge, number of potential hydrogen-bond, two pKa values and ASA. *SVM1.3*: MM, number of atoms, electrostatic charge, number of potential hydrogen-bond, two pKa values, ASA and SS. *SVM1.4*: MM, number of atoms, electrostatic charge, number of potential hydrogen-bond, two pKa values, ASA, SS and PSSM.

Model	Sensitivity	Specificity	ACC
SVM1.1	0.57	0.64	0.73
SVM1.1Beta	0.62	0.58	0.72
SVM1.2	0.66	0.61	0.76
SVM1.3	0.67	0.51	0.71
SVM1.4	0.84	0.70	0.77

binding; for test protein 1AV6 the predicted residues are 68 whereas HBPLUS detected only 19 amino acids as binding.

Starting from the predicted binding-triplets the RNA sequence is reconstructed using the propensity statistic. The reconstructed RNA sequence for test protein 1A9N is:

$$G(AG)AAAG(AG)AAA$$

Nucleotides in brackets can not be assigned uniquely, the binding of the investigated triplet can be with ribonucleotide A or G. The real RNA strand extracted from the corresponding pdb file is

$$CCUGGUAUUGCAGUACCUCCAGGU$$

Table 3.4 contains the reconstructed and the original RNA strands for all 6 test proteins.

The final results are not satisfying which is evidenced by comparing the constructed and the real RNA strands. This insufficiency has several reasons. Without any doubt one reason is the true positive prediction of SVM1.4. Despite a high sensitivity of 0.84 in an independent 10-fold cross validation, the prediction on the test proteins shows that too many triplets are wrongly assigned to the binding class. We base our RNA-construction on these predictions so they should be more precise, which is evidently not the case when looking to the results of Table 3.3 and Table 3.4.

Furthermore also the applied propensity statistic may be too simple. If a triplet binds more than one nucleotide the assignment should rely on more information. Maybe by including also the nucleotide-triplet in the statistic. On the other hand the dataset was strongly unbalanced,

Table 3.3: **Predicted binding residues using SVM1.4.** Predicted binding and non-binding residues in six complexes. The "real" number of binding elements extracted with HBPLUS is shown in brackets.

Complex	# Binding residues	# Non binding residues	Total # of residues
1A9N	40 [71]	485 [454]	525
1ASZ	324 [124]	654 [854]	978
1AV6	68 [19]	226 [275]	294
1B7F	146 [86]	188 [248]	334
1B23	149 [48]	255 [356]	404
1C0A	247 [119]	337 [465]	584

there were much more negative (11405 non-binding) than positive (1452 binding) residues. So we were constrained to create a dataset with an equal number of negative and positive elements. Also this procedure can interfere with the prediction ability of the model.

3.5 Conclusions

The predicted binding-residues can be sparse over the protein surface instead be accumulated within a binding pocket. This makes the ligation to an RNA, from a biological point of view, improbable. Additionally this kind of predictions ignore the 3-dimensional structure of the input protein and they do not care where the binding residue is positioned in the folded protein. Taking into account the 3-dimensionality is important but not easy to implement.

Assuming that the binding information is available on both sides, on the protein and on the RNA, one could think to ignore the RBP because its binding site remains always the same. What changes is the RNA target. Therefore one future direction could be the examination of the binding site on the RNA target without considering the binding site of the RBP. Another problem we faced in our preliminary analysis, is the lack of data: for each RBP only one bound RNA is available. This may be enough for binding-residue predictions but probably not enough to create an RNA sequence on propensity data. The fact is that some complexes of the dataset are incomplete. For example they contain only fragments instead of the entire protein sequence and the RNA is not always a real one but an artificial one. Therefore we think that it can have more sense to exploit high-throughput datasets which detect a vast number of RNA sequences bound by a specific RBP.

Table 3.4: **Constructed versus real RNA-strands.** Real RNA sequences bound by the corresponding RBP are shown in *italic* and constructed RNA sequences based on our binding-triplet prediction are written “normal”.

Structure	Constructed RNA vs. real RNA
1A9N	G(AG)AAAG(AG)AAA *CCUGGUAUUGCAGUACCUCCAGGU*
1ASZ	(AG)CA(AG)(UG)CU(AG)(UC)UA(CG)(CG)(AUC)GG(UC)(AG)CG(AU)(CG)A(AG)AA(AG)(AU)GG(CG)GGGU(AUCG)(ACG)U(CG)(AG)(AU)U(AG)CA(AG)(UG)CU(AG)(UC)UA(CG)(CG)(AUC)GG(UC)(AG)CG(AU)(CG)A(AG)AA(AG)(AU)GG(CG)GGGU(AUCG)(ACG)U(CG)(AG)(AU)U *UCCGUGAUAGUU-AA-GG-CAGAAUGGGCGC-UGUC-CGUGCCAGAU-GGGG–CAAUUCCCCGUCGCGGAGCCA*
1AV6	UUC(CG)G(UG)AAAGC(CG)(AG)AC(ACG)UAC *GAAAAA*
1B7F	(AUG)UUU(AU)(ACG)GA(AC)GGUU(AG)UAU(UCG)(UC)(AUC)C(AG)(AG)(AG)(AC)(CG)AC(AUG)UUU(AU)(ACG)GA(AC)GGUU(AG)UAU(UCG)(UC)(AUC)C(AG)(AG)(AG)(AC)(CG)AC *GUUGUUUUUUUU*
1B23	CUAUGA(CG)(CG)AUU(CG)(UG)UAUC(AG)GGG(CG)(CG)UCCGGAGC(UC)ACGAUG(AU)(AG)AC(AG)UA(CG) *GGCGCGU-AACAAAGCGG–AUGUAGCGGA-UGC-AA-CCGUCUAGUCCGG–CGACUCCGGAACGCGCCUCCA*
1C0A	(UG)CCC(AUC)C(UC)AU(CG)C(UC)AAGA(AG)GA(UG)G(AC)(AC)GC(ACG)(UG)CGGAG(AUC)CCGU(CG)ACG(UG)(UC)(CG)UCG(ACG)U(AU)U(AG)(ACG)(UC)G(UC)CAUAA(AU)G(CG)G(AC) *GGAGCGG-AGUUCAG-CGG–AGAAUACCUGCCU–CACGCAGGGGG-CGCGGG–CGAGUCCCG-CCGUUCCGCCA*

Chapter 4

RNA-binding prediction for CELF1. A case study

The interplay between RBPs and RNAs is highly specific and crucial for cellular physiology. Identifying the RNA targets for a given RBP is important to understand its function and therefore of interest in biology. Several molecular approaches like SELEX or CLIP-seq detect RBP-RNA interactions and consequently the RBP-specificity but computational models and binding predictions would greatly reduce "hands-on" experimental time and experimental costs. On the other hand experimental data, especially from high-throughput techniques, constitute an important source and host precious information regarding the binding of the analysed RBP. Tapping the full potential of such *in vivo* and *in vitro* datasets seems a good strategy, because the data can be used to predict *in silico* RNA-protein bindings.

RBPs have different ways to bind their RNAs: they can bind particular patterns on the sequence or associate secondary structures. For instance CELF1 (also known as CUGBP1 or EDENBP) is a human RNA-binding protein which binds mainly single stranded UGU-rich RNA-segments (Marquis et al., 2006; Kress et al., 2007). Being present in the nucleus and cytoplasm of the cell, CELF1 controls post-transcriptional processes at many levels.

In this chapter we propose the basic framework of our approach and apply it to CELF1. We use SVMs to classify CELF1-binding sites and to discriminate binding from non-binding RNAs. Additionally we perform two experiments which verify the prediction ability of the proposed approach. First we briefly describe the concept, the dataset and the applied features. Then we detail the experiments, the obtained results and discuss them.[1]

---

[1] Part of this work was published in the proceedings of the *6th International Conference on Practical Applications of Computational Biology & Bioinformatics- PACBB12, 2012, Salamanca, Spain*

4.1 The Approach

The goal is to discriminate between CELF1-target RNAs and sequences which are not targeted by CELF1. Therefore we exploit the results of a CLIP-seq dataset and use SVMs for classification. We describe each RNA sequence with 287 features: 256 features are obtained by encoding the RNA sequence in oligonucleotides (oligos), 30 features are motif scores calculated by applying PSSMs and the last single feature is the presence of a UGU-rich motif in the sequence.

RBP CELF1 is known to bind single stranded RNA sequences by targeting a specific sequence pattern. Due to a previously performed SELEX experiment (Marquis et al., 2006) two similar binding motifs have been detected with MEME: Figure 4.1 and Figure 4.2 show their sequence logo. A third and independent binding motif has been searched also within the CLIP-seq dataset (by using MEME). The motif is not shown as it changes slightly for each fold. MEME provides PSSMs which are finally used to calculate the motif-scores along each RNA strand. In our approach the presence of the binding site is described by means of these scores and they are calculated for each motif on each sequence. In other words we encode the binding site by using the ten highest scores for each motif (10×3) as features.

The binding is not only determined by the presence of a specific binding motif but can depend also on the sequence-context of the motif. To incorporate the individual sequence composition we encode each RNA strand by its oligo frequency. The appearance of each oligo in the sequence sets the corresponding feature vector field.

Based on structural information a particular UGU-rich binding motif is known to be bound by CELF1. We describe the UGU-rich binding motif by means of a binary feature which is set to 1 if the motif is present in the sequence, otherwise it is set to 0.

All the described features fit the CELF1-binding specificities.

A 10-fold cross validation and two experiments are performed to validate the model. In the first experiment we apply SVM and localSVM on cluster sequences bound by CELF1 and on non-bound sequences. Additionally we attempt to balance the dataset by applying a balancing algorithm. The second experiment divides the training data in subsets. Each set contains sequences of a limited length l and validates the prediction performance on subsequences of the test data.

4.2 Material and Methods

4.2.1 Datasets

CLIPData This dataset originates from a CLIP-seq experiment realized in Hela-kyoto cells (Olivier le Tonqueze, unpublished data) and represents the positive data for our model. It

Figure 4.1: **CELF1-binding motif of length 11.** Two CELF1-binding motifs have been found in a previously performed SELEX experiment. The corresponding PSSMs produce scores which are used as features in our model.

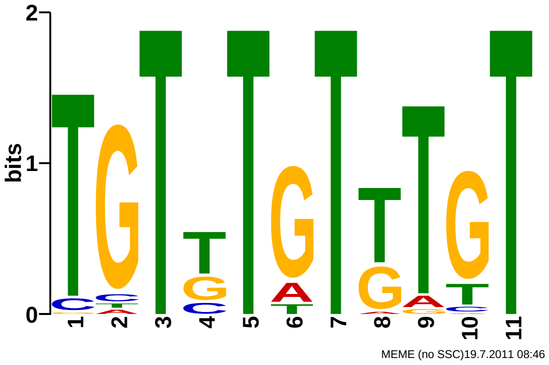


Figure 4.2: **CELF1-binding motif of length 15.**

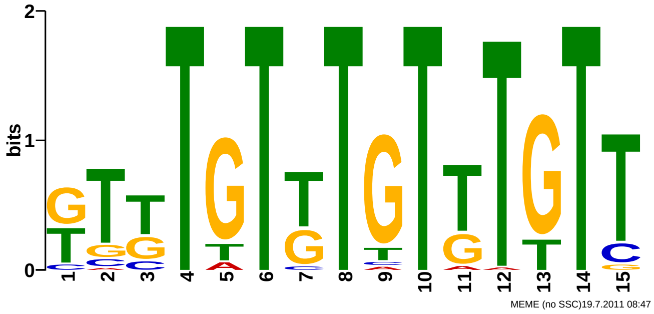


Table 4.1: **Summary of the datasets.** The positive data contains binding clusters from a CLIP-seq experiment with CELF1. The negative examples used to create the model are formed by expressed but not bound transcripts.

Dataset	No. of chains	Shortest/longest	Average length
CLIPData	1932	29/1401	170
original NegData	36701	11/16193	952
NegData	24680	19/999	379

consists of 1932 cluster sequences identified to be bound *in vivo* by CELF1.

NegData The dataset is constituted of 3'UTRs from genes expressed in Hela-kyoto cells and not bound by CELF1. It represents the negative data for the SVM. Originally the dataset consists of 36701 transcripts. We decided to use only a subset of 24680 transcripts in order to provide a negative dataset with a similar sequence length distribution as in the positive one. Table 4.1 shows a summary of the datasets.

4.2.2 Feature Extraction and Representation

Oligos

To incorporate the sequence specificity we codify the RNA strand with oligos of length 4. Oligos are all possible combinations of nucleotides $\Omega = \{A, U, C, G\}$:

$$oligo_x = w_1 w_2 w_3 w_4 \tag{4.1}$$

with $w_i \in \Omega$ and $i = 1, 2, 3, 4$. For example $AAAA, AAAU, AAUC$ and so on. A sliding window of length 4 is scrolled over the sequence and the words frequency is extracted. Figure 4.3 shows the oligo construction. This procedure adds 256 features.

PSSM

The identification of a motif in a sequence $s = s_1, ..., s_n$ of length n is based on the PSSM matrix-value $pssm(g(\hat{s}_k), k)$ which calculates a motif-score

$$score_{\hat{s}} = \sum_{k=1}^{m} pssm(g(\hat{s}_k), k) \tag{4.2}$$

for each subsequence $\hat{s} = s_{1+i} \ldots s_{m+i}$. m indicates the motif length $m \leq n$, k the position in the subsequence $\hat{s}$, $g(\hat{s}_k)$ the symbol emitted at position $\hat{s}_k$ and $pssm(g(\hat{s}_k), k)$ the value

Figure 4.3: **Schematic diagram of the oligo extraction.** The general composition of the sequence is encoded via oligos. A sliding window of length 4 scrolls over the RNA and extracts the oligo frequency in the sequence.

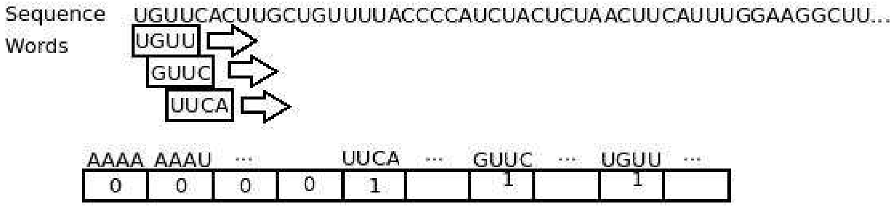
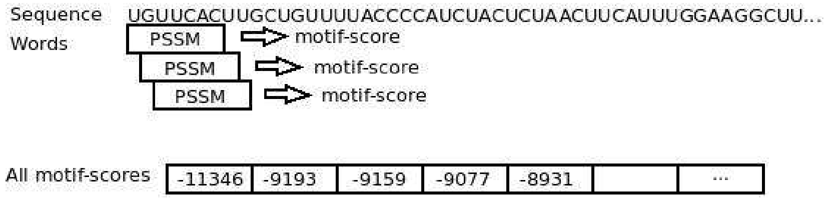

Sequence
UGUUCACUUGCUGUUUUACCCCAUCUACUCUAACUUCAUUUGGAAGGCUU...
Words
UGUU
GUUC
UUCA
AAAA
AAAU
...
UUCA
...
GUUC
...
UGUU
...
0
0
0
0
1
1
1

Figure 4.4: **Schematic diagram of the motif-score calculation.** The PSSM profile scrolls along the RNA and is used to calculate the motif-score for each subsequence. The ten highest scores of the sequence are finally chosen as features to represent the binding-motif.


Sequence
UGUUCACUUGCUGUUUUACCCCAUCUACUCUAACUUCAUUUGGAAGGCUU...
Words
PSSM
motif-score
PSSM
motif-score
PSSM
motif-score
All motif-scores
-11346
-9193
-9159
-9077
-8931
...

of the matrix at column k and row $g(\hat{s}_k)$. Figure 4.4 shows the PSSM score extraction. This procedure adds 30 features (10 motif scores for 3 motifs) to the final feature vector.
Several motif-based sequence analysis tools, such as *MEME* (Bailey and Elkan, 1994), discover motifs in sequences and calculate their PSSMs. Consider the following example sequence $x_{example}$:

```
UGUUCACUUGCUGUUUUACCCCAUCU
```

The sorted motif-scores produced by the binding motif in Figure 4.2 on the example sequence are:

```
[-11346.0, -9193.0, -9159.0, -9077.0, -8931.0, -8841.0, -8779.0,
-8630.0, -8623.0, -8299.0, -8219.0, -8148.0, -8063.0, -8060.0,
-8015.0, -7972.0, -7878.0, -7828.0, -7721.0, -7712.0, -7605.0,
-7568.0, -7560.0, -7412.0, -7330.0, -7193.0]
```

Binding Motif

CELF1 contains 3 RNA recognition motifs. Based on structural information of the CELF1 protein, each RRM may recognizes one UGU-trinucleotide. Therefore a known binding pattern for CELF1 is a UGU-motif with the following structure:

$$[UGU]N_1[UGU]N_2[UGU] \qquad (4.3)$$

where N_i with i=1,2 are two nucleotide sequences $n_{i_0}...n_{i_s}$ with length $s \in [0 \ldots 20]$.

Balancing Dataset

To overcome the unbalanced data, with 1932 binding and 24680 non-binding sequences, we attempt different approaches:

1. The application of a synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) which amplifies the positive dataset and creates new synthetic instances to force the classifier to become more general.
2. The application of local kernel machines (localSVM) which use a set of SVMs to create an appropriate local model on each training point.

Classifier Evaluation

The performance of the model is evaluated in a 10-fold cross validation by calculating Matthews correlation coefficient, sensitivity and specificity. We optimize the RBF kernel parameters C and γ for the SVM and the localSVM on the highest MCC. To avoid a circularity problem we extract the third CLIP-seq motif in each training fold separately.

4.2.3 Experiments and Results

Several combinations of the previously described features have been tested using a 10-fold cross validation on a subset of 1932 binding and 1989 randomly selected non-binding sequences (Table 4.2). The combination with oligos, 10×3 motif-scores and UGU-motif achieves the best performance values with an MCC=0.63, Sens=0.81 and Spec=0.82. Therefore we build the SVM on these features and describe two experiments to test its classification ability.

First Experiment

We first test the performance of the SVM and the localSVM on the total dataset formed by 1932 cluster sequences and 24680 3'UTRs. The results, reported in Table 4.3, provide important information regarding the influence of the unbalanced dataset to the performance. With a high specificity of 0.9 and a low sensitivity of around 0.4 is it evident that both methods are not discriminative. This is likely induced by the unbalanced dataset. Therefore we decide to apply the two classification methods after balancing the dataset with SMOTE. Sensitivity and specificity increase to 0.92 and 0.77 with the SMOTE+localSVM harbouring an MCC of 0.40. Much better performs the SMOTE+SVM approach which increases both, sensitivity and specificity, to 0.86 to reach an MCC of 0.48 (see Table 4.3).

Table 4.2: **Performance of different feature combinations.** The impact of the different features is tested on a dataset with 1932 positives and 1989 randomly selected negatives. The combination which uses all previously described features obtains the best classification values (last row) and is used for the final model.

PSSM11	PSSM15	PSSMClip	Oligo	TGT	Sens	Spec	MCC
3	3	3	0	0	0.63	0.63	0.27
5	5	5	0	0	0.63	0.65	0.28
7	7	7	0	0	0.63	0.66	0.29
10	10	10	0	0	0.62	0.68	0.30
10	10	10	1	1	0.81	0.82	0.63

Table 4.3: **Results of the first experiment.** Applying SVM and localSVM on the total dataset with 1932 binding and 24680 non-binding examples shows that both methods are not discriminative on the unbalanced dataset. After the balancing with SMOTE, SVM and localSVM classify the sequences with high sensitivity and specificity.

Method	Sens	Spec	MCC
SVM	0.39	0.99	0.52
SMOTE+SVM	0.86	0.86	0.48
localSVM	0.41	0.98	0.52
SMOTE+localSVM	0.92	0.77	0.40

The positive and the negative sequences differ in the composition of mRNA regions (clusters) known to bind CELF1 and in the full length 3'UTR sequences, respectively. To overcome this difference we apply the model, in a second experiment, on subsequences of fixed length.

Second Experiment

From an RNA sequence to be tested we extract subsequences of length x, with a sequence overlap of $x/2$, and apply the classifier on the subsequences. As classifier we use SVM with the previously described features. If a subsequence is predicted as binding, the tested sequence is classified as a binding sequence. To bypass the unbalanced data problem described above, the total dataset is divided into several training sets set_x. Every training set contains sequences of

Table 4.4: **Results of the second experiment.** An SVM with the previously described features is trained on different sequence subsets, called set_x. Each set contains sequences of a limited length x. The models are used to classify RNA subsequences of length $l \leq x$ extracted from longer sequences. The Table shows the percentages of true positive (TP) and true negative (TN) predicted RNAs and the MCC.

set_x	Subsequence length x	TP prediction (%)	TN prediction (%)	MCC
set_{100}	100	96.5	14.8	0.20
set_{200}	200	90.5	49.4	0.44
set_{300}	300	87.8	62.6	0.46
set_{400}	400	78.9	75.8	0.45
set_{500}	500	68.6	80.2	0.36

a limited length $l \leq x$, x=100,200,300...500. For instance: $x = 200$ indicates that the model is trained on sequences with a maximum length of 200 (set_{200}). The corresponding model classifies subsequences of the same length, extracted from a test RNA longer than x. Table 4.4 reports the predicted percentages of binding and non-binding sequences for each model.

4.3 Discussion and Conclusions

The identification of RNAs associated with a given regulatory RBP is costly and time consuming when realized experimentally. Moreover it depends highly on the cell type or on the organism used in the experiment. So far there is no other approach to detect specific protein-RNA binding *in vivo*. Machine learning techniques have been successfully applied on biological data and SVMs are known to be a powerful method to solve classification problems with high performance.

We proposed to classify RNA sequences as binding or as non-binding to CELF1 protein using SVMs trained on CLIP-seq data. Two experiments are described to test the prediction ability of the approach. The first experiment applied SVM and localSVM on a CLIP-seq dataset. Due to the unbalanced data the application of a balancing algorithm was necessary. Results in Table 4.3 show a high sensitivity and specificity (on the balanced dataset) for both SVM and localSVM, indicating that binding information from SELEX and CLIP-seq experiments can be efficiently used to train SVMs. To avoid bias, potentially introduced by the data balancing, we designed a second experiment in which we grouped sequences by their length to obtain homogeneous and equalized datasets. The results in Table 4.4 indicate that subsequences of length $x = 400$ allow to train a SVM to discriminate binding from non-binding RNAs more accurately than shorter sequences. Codifying short sequences in oligos causes numerous features to be zero and therefore they probably contain not enough information for a proper classification.

Furthermore our results show that exploiting experimental data to create models seems reasonable. An advantage of the SVM defined on subsequences (see Second Experiment in Section 4.2.3) is the localisation of potential binding sites for CELF1. Therefore a model based on binding-clusters works potentially better. Unfortunately not all experimental outcomes contain such information and the binding site is not always known. The same limitation holds for previously known binding-motifs which are not always available.

Chapter 5

Predicting mRNA binding with sequence features, motifs and secondary structures

In this chapter we extend the previously defined approach (see Section 4.1) to cover a broader area of binding properties like the secondary structure of the RNA sequence and the accessibility of the binding site. Encouraged by the results of our case study with CELF1, and taking advantage of other experimental and publicly available datasets, we address a bigger set of RBPs and their RNA targets. We propose three approaches based on different features which are applied to 19 RBPs with the goal to detect whether an RNA sequence is bound or not by the corresponding RBP. To evaluate the methods we compare our results with the predictions of *RPISeq* (Muppirala et al., 2011) and *RNAcontext* (Kazan et al., 2010).

5.1 The Approach

RBPs bind in different ways to RNA: some of them associate sequence motifs and others structural motifs. Similar to the approach presented in Chapter 4 we embed the binding site information in form of motif scores. Known binding motifs as for CELF1 are not always available. Therefore we use *MEME* to automatically discover a significant binding motif within a set of bound RNA-sequences, without using previously available binding information. Other RBPs do not bind sequence patterns but can associate structural patterns, so-called secondary structures. For that reason we include simple secondary structure information as features. The tertiary structure of the protein and the accessibility of the binding site influences the RNA binding (Auweter et al., 2006; Mukherjee et al., 2011). The accessible surface area is calculable only on 3-dimensional structures but realistically speaking no high-throughput dataset provides such 3-dimensional information. So we decide to use a simple accessibility feature: we define an RNA-subsequence as accessible if at least four consecutive nucleotides are single stranded, i.e. not involved in a stem. We argue that a ribonucleotide which forms a stem with another ribonucleotide should be less accessible to an RBP.

We propose three methods: method *Oli* which uses oligos of length 4 as features, extension *OliMo* which adds protein-specific binding motifs and extension *OliMoSS* which adds secondary structure information. The methods are applied to experimental datasets which are downloaded from *The Atlas of UTR Regulatory Activity (AURA)* (Dassi et al., 2012). Additionally we use the well-studied human RBP Pumilio-2 (PUM2), extracted from *Gene Expression Omnibus (GEO)* (Edgar et al., 2002), to assess the influence of true negative RNA sequences on the prediction ability of the models. RBP PUM2 is involved in the translation and stability of mRNA and binds sequence-specific to 3'UTRs.

Similar to CELF1, also for PUM2 the entire experimental data with bound and non-bound RNA sequences is available. Unfortunately experimental non-binding examples do not exist for the remaining RBPs. Therefore we use 3000 3'UTRs, randomly downloaded from *Ensembl Genome Browser* (Flicek et al., 2010), as negatives. Usually the number of non-bound sequences in high-throughput experiments is much higher than the number of bound sequences. To preserve this proportion we choose 3000 human 3'UTRs as negatives, as it is the double of the highest amount of sequences used in *AURAdataset* (see Section 5.2.1), and small enough to train models in a reasonable time. The dataset is balanced using the over-sampling algorithm *SMOTE*.

To study our methods we perform different evaluations. In Evaluation 1 we assess the prediction of *Oli*, *OliMo* and *OliMoSS* with the linear kernel applied on 18 different RBPs. As said above experimental non-binding data is not available for these RBPs, so we use the randomly selected 3'UTRs as negative examples. In Evaluation 2 and Evaluation 3 the methods are applied on PUM2 protein where experimental detected non-bound sequences are available. In Evaluation 2 the negative data is formed as in Evaluation 1 by the randomly selected 3'UTRs whereas in Evaluation 3 the models are trained on PUM2's real negatives. Comparing the predictions of Evaluation 2 and Evaluation 3 we can assess the importance of real negative training data.

In both, Evaluation 1 and 3, we compare the results with *RPISeq* (Muppirala et al., 2011) and *RNAcontext* (Kazan et al., 2010). *RPISeq* is directly comparable with our methods because it takes as input proteins and RNA sequences and predicts their binding using SVM (*RPISeq-SVM*) and RF (*RPISeq-RF*). Similarly *RPISeq* applies oligos of length 4 to describe the RNA sequence. A totally different approach is applied in *RNAcontext*: the tool is designed to find structural and sequence motifs in a pool of training sequences and searches for them in a set of test sequences assigning a score. In Evaluation 1 we compare our methods with *RPISeq* and

RNAcontext to assess a) the influence of the applied approach (motif based tools vs. machine learning), b) the influence of the used features and in Evaluation 3 to assess the influence of the negative training data on the prediction.

5.2 Material and Methods

5.2.1 Datasets

AURAdataset

The *AURAdataset* is formed by RBPs and related RNA sequences downloaded from AURA (Release 2.4). AURA is an online database which contains experimentally-derived mRNA-RBP pairs in human. For simplicity we call the set of RNA sequences reported in AURA as bound by a specific RBP "*RBP+*" set. For example *CPEB1+* is formed by 256 sequences and *PUM1+* by 668 RNA sequences. Using only proteins with more than 50 associated 3'UTRs (to have enough positive examples to train an SVM) and proteins for which *MEME* was able to detect binding motifs in a reasonable time, we obtain the *AURAdataset* with 18 *RBP+* sets (see first two columns of Table 1). In order to eliminate similar sequences we run CD-HIT (Li and Godzik, 2006) within each *RBP+* set and filter out those with more than 80% identity. By choosing randomly 3000 human 3'UTRs from Ensembl Genome Browser we construct an "artificial" negative dataset, in the following called *3K-*, which is applied for each *RBP+* set. For example consider RBP CPEB1: its *CPEB1+* set contains 256 binding sequences representing the positive data. The negative data, used to train the SVM, is formed by *3K-*.

PUM2dataset

The PUM2 data originates from a PAR-CLIP experiment done on human embryonic kidney cells (HEK293) and has been downloaded from GEO (GSM545210). In the experiment (Hafner et al., 2010) 7523 clusters on about 3000 transcripts were identified and 93% were found within the 3'UTRs. We extracted all 3'UTRs in such way that each cluster appears only once, to avoid multiple copies of the same cluster. In this way we create a dataset called *PUM2+* which contains 2151 positive 3'UTRs known to be bound by PUM2. The non binding data for RBP PUM2 originates from an RNA-seq experiment (Kishore et al., 2011) done on the same HEK293 cells and under the same conditions. Two repetitions, A GSM714678 and B GSM714678, are available on GEO. To avoid a loss of data we merge the results of the two repetitions and download the sequences from Ensembl Genome Browser (NCBI36/hg18 release 54, May 2009). Hence the dataset called *PUM2-* is constituted by 3000 of the 12329 negative 3'UTRs not detected to be bound by PUM2.

5.2.2 Feature Extraction and Representation

Oligos

We codify the individual RNA sequence composition using the frequency of oligos of length 4. Oligos are all possible combinations of nucleotides, for example AAAA, AAAU, AAUC and so on. The corresponding feature contains the number of oligo occurrences in the sequence. See Chapter 4.2.2 for more details.

PSSM

Motifs are patterns in RNA, DNA or protein sequences which can be modelled by position-specific scoring matrices, called PSSMs. *MEME* detects a significant motif in a set of sequences and creates the corresponding PSSM which we finally use to compute the motif score. The motif score $score_{\hat{s}_i}$ is calculated for each m-length subsequence $\hat{s}_i = b_{i+1} \ldots b_{i+m}$, $i \in \{0, ..., n-m+1\}$ along the RNA sequence $b_1 b_2 \ldots b_j \ldots b_n$ where b_j is the ribonucleotide at the j-th position and m the motif length $m \leq n$:

$$score_{\hat{s}_i} = \sum_{k=1}^{m} pssm(b_{i+k}, k) \tag{5.1}$$

where $pssm(b, k)$ returns the value of the matrix for $b \in \{A, U, C, G\}$ and position k. We search for significant motifs in each *RBP+* set using the following *MEME* property settings: mod=zoops, minw=5 and maxw=10.

Simple Secondary Structure Features

Based on the predicted RNA secondary structure, using RNAfold (Hofacker et al., 1994), we calculate the following features:

1. predicted folding energy (calculated by RNAfold);
2. stem density: proportion of paired base pairs (Pancaldi and Bähler, 2011);
3. number of stems (Pancaldi and Bähler, 2011);
4. accessibility: the accessibility is computed on subsequences with at least four consecutive nucleotides in single stranded form, i.e. not involved in a stem secondary structure. We codify these subsequences using oligos. Oligos are all possible combinations of nucleotides of length 4. The corresponding feature is set to 1 if a specific subsequence is single stranded otherwise it is set to 0.

5.2.3 Evaluation and Comparison

The analysis has been done as follows: **Evaluation 1** tests the prediction of methods *Oli*, *OliMo* and *OliMoSS* on the *AURAdataset* and each *RBP+* set is assisted with *3K-*. AUC and precision are calculated. The same has been done for *RNAcontext* and *RPISeq*. Finally we apply the Wilcoxon signed-rank test on the AUCs to compare the performances. **Evaluation 2** tests the performance of *Oli*, *OliMo* and *OliMoSS* on *PUM2+* with *3K-* as negatives. We calculate AUC and precision and compare the predictions with *RNAcontext* and *RPISeq*. **Evaluation 3** tests the performance of *Oli*, *OliMo* and *OliMoSS* on *PUM2+* with the real negatives *PUM2-*. AUC and precision are used to compare the predictions with *RNAcontext* and *RPISeq* and to compare the predictions with the results of Evaluation 2.

In each evaluation we identify the best value for the linear kernel parameter C according to the highest MCC on each fold. The machine learning method we use is the freely available SVM package *LibSVM* (Chang and Lin, 2011). All the scripts in this chapter are implemented in Python. Due to the unbalanced dataset with much more negative than positive examples, we apply the oversampling algorithm *SMOTE* to balance the data. Therefore in our approach we will use only the linear kernel. *SMOTE* is executed only on the training folds. To ensure a fair evaluation and to avoid circularity in the folds we search binding motifs with *MEME* in each of the 10 training folds separately.

5.3 Results and Discussion

5.3.1 Evaluation 1

The performance of *Oli*, *OliMo*, *OliMoSS*, *RNAcontext*, *RPISeq-SVM* and *RPISeq-RF* on each RBP of the *AURAdataset* is assessed by calculating the AUC. All AUC values are shown in Table 5.1. In order to check if two samples, in our case the AUC values of two methods, belong to the same distribution we use the Wilcoxon signed-rank test. The statistical test allows to analyse if the predictions of two distinct methods on the same dataset are significantly different. The p-values, shown in Table 5.2, are calculated by testing each method against the others.

Regarding the AUC *Oli* and *OliMo* achieve the highest mean of 0.76, followed by *RNAcontext* with a mean of 0.72. *OliMoSS*, *RPISeq-SVM* and *RPISeq-RF* obtain the lowest performances with means of 0.69, 0.66 and 0.61, respectively. The Wilcoxon signed-rank test shows a statistically significant difference between the prediction of *Oli* and *OliMoSS* ($p=0.0003$) and between the prediction of *OliMo* and *OliMoSS* ($p=0.0004$). The same can not be claimed between *Oli* and *OliMo* ($p=0.77$). All approaches show a statistically significant difference in their prediction compared to *RPISeq-RF*. *RNAcontext*, compared to *RPISeq-SVM* and *RPISeq-RF*, predicts differently ($p=0.0033$ and $p=0.0003$). In the same way *Oli* and *OliMo*

Table 5.1: **Performance of Oli, OliMo, OliMoSS, RNAcontext and RPISeq on the AURA-dataset.** The performance is evaluated calculating the AUC for each RBP. The first column lists the protein name and the second column shows the number of RNA sequences contained in each *RBP+* set. The last rows of the table show the mean and the standard deviation of the AUC for each method. The negative data is always formed by *3K-* (see Evaluation 1).

Name	#(*RBP+*)	*Oli*	*OliMo*	*OliMoSS*	*RNAcontext*	*RPISeq-SVM*	*RPISeq-RF*
SLBP	54	0.67	0.67	0.58	0.49	0.54	0.35
MSI1	69	0.75	0.84	0.64	0.80	0.84	0.78
TIAL1	73	0.55	0.52	0.51	0.57	0.48	0.50
CPEB4	109	0.59	0.63	0.48	0.50	0.55	0.54
AGO2	213	0.85	0.85	0.72	0.82	0.70	0.61
CPEB1	256	0.66	0.67	0.61	0.64	0.63	0.55
CUGBP1	309	0.78	0.77	0.70	0.75	0.73	0.63
TNRC6A	249	0.87	0.86	0.80	0.83	0.67	0.67
PUM1	668	0.75	0.76	0.71	0.72	0.68	0.64
TNRC6C	157	0.80	0.83	0.69	0.80	0.71	0.61
PABP	403	0.61	0.59	0.55	0.55	0.52	0.50
U2AF65	363	0.79	0.76	0.77	0.75	0.65	0.66
AGO4	279	0.87	0.86	0.76	0.84	0.76	0.62
QKI	725	0.87	0.86	0.84	0.84	0.77	0.76
TNRC6B	760	0.84	0.86	0.83	0.84	0.70	0.68
ELAVL1	1872	0.77	0.77	0.73	0.75	0.62	0.63
AUF1	1987	0.72	0.72	0.68	0.68	0.57	0.62
AGO1	1873	0.86	0.86	0.86	0.84	0.74	0.62
Mean		0.76	0.76	0.69	0.72	0.66	0.61
standard deviation		0.10	0.11	0.11	0.12	0.10	0.10

Table 5.2: **p-values of the Wilcoxon signed-rank test for Oli, OliMo, OliMoSS, RNAcontext and RPISeq.** The statistical test is used to compare the performance of all methods on the *AURAdataset*. A p-value under a predefined significance level indicates a significant difference in the prediction. We use a significance level of 0.05.

	Oli	*OliMo*	*OliMoSS*	*RNAcontext*	*RPISeq-SVM*
OliMo	0.7744				
OliMoSS	0.0003	0.0004			
RNAcontext	0.0055	0.0025	0.0284		
RPISeq-SVM	0.0006	0.0003	0.0650	0.0033	
RPISeq-RF	0.0002	0.0002	0.0040	0.0003	0.0085

are statistically different from *RNAcontext* ($p=0.0055$ and $p=0.0025$) and from both *RPISeq* methods ($p<0.0007$).

Low precision values characterize the six approaches: the mean ranges from 0.14 for *RPISeq-RF* to 0.38 for *OliMo*. Table 5.3 contains the precisions calculated at a threshold of 0.5 for each method on each RBP. *Oli* performs similar to *OliMo*. *RNAcontext* (Prec=0.33) and *OliMoSS* (Prec=0.31) outperform both *RPISeq-SVM* (Prec=0.15) and *RPISeq-RF* (Prec=0.14). The computation of the precision at a threshold of 0.5 does not show the overall potential of the methods, that is best visualized by PR curves. In the following we compare *Oli* with *RNAcontext*, as it seems to be more competitive than *RPISeq* (Wilcoxon signed-rank test $p=0.0055$ vs. $p<0.007$). To visualize the classification ability of the two approaches we plot the PR curve (Figures 5.1-5.3) and the ROC curve (Figures 5.4-5.6) for each RBP. The optimum in a PR curve is the upper-right corner and both methods have similar difficulties to reach it for SLBP, TIAL1, CPEB4 and PABP. *Oli* outperforms *RNAcontext* for nearly each RBP and its curve is visibly shifted over the y-axis. The ROC curves of both approaches instead show a competitive behaviour which essentially reflects the AUCs of Table 5.1.

The performance of our approaches is protein dependent. For several RBPs like AGO1, TNRC6A or QKI, *Oli* and *OliMo* achieve an $AUC \geq 0.80$ whereas for other proteins, like TIAL1, they perform worse with an $AUC \leq 0.6$. This may due to the fact that each RBP binds in a specific way and the adopted features may not always capture the particular binding property. We expected that providing more binding information with motif scores and accessibility, could improve the discrimination between binding and non-binding RNA. But contrary to this expectation *OliMo* and *OliMoSS* do not outperform *Oli*. Moreover, *OliMoSS* shows low AUCs and a statistically significant difference in its prediction, affirming that it is the weakest of our approaches. We conclude that our secondary structure features are not neces-

Table 5.3: **Precision values for Oli, OliMo, OliMoSS, RNAcontext and RPISeq.** For each RBP in the *AURAdataset* the precision is calculated at a threshold of 0.5. The last row shows the mean and the standard deviation for each method.

Name	*Oli*	*OliMo*	*OliMoSS*	*RNAcontext*	*RPISeq-SVM*	*RPISeq-RF*
SLBP	0.13	0.16	0.19	0.01	0.02	0.01
MSI1	0.18	0.17	0.09	0.08	0.03	0.02
TIAL1	0.03	0.04	0.03	0.02	0.02	0.02
CPEB4	0.05	0.06	0.04	0.04	0.03	0.03
AGO2	0.30	0.29	0.23	0.32	0.07	0.07
CPEB1	0.20	0.19	0.13	0.22	0.08	0.08
CUGBP1	0.31	0.32	0.24	0.31	0.11	0.10
TNRC6A	0.30	0.32	0.27	0.33	0.08	0.08
PUM1	0.46	0.44	0.38	0.35	0.22	0.19
TNRC6C	0.23	0.25	0.17	0.21	0.06	0.05
PABP	0.17	0.17	0.13	0.13	0.12	0.12
U2AF65	0.32	0.42	0.27	0.32	0.11	0.11
AGO4	0.38	0.39	0.29	0.36	0.09	0.09
QKI	0.57	0.66	0.49	0.54	0.21	0.21
TNRC6B	0.59	0.58	0.52	0.53	0.22	0.21
ELAVL1	0.78	0.78	0.74	0.73	0.44	0.40
AUF1	0.73	0.71	0.68	0.67	0.43	0.42
AGO1	0.80	0.81	0.76	0.68	0.40	0.39
Mean±sd	0.36±0.24	0.38±0.24	0.31±0.23	0.33±0.23	0.15±0.14	0.14±0.13

Figure 5.1: **Precision-Recall curves for Oli and RNAcontext on the AURAdataset.** The PR curves visualize the performance of *Oli* (red line) and *RNAcontext* (green line) on RBP SLBP, MSI1, TIAL1, CEPB4, AGO2 and CPEB1. *Oli* outperforms *RNAcontext* for nearly each RBP.

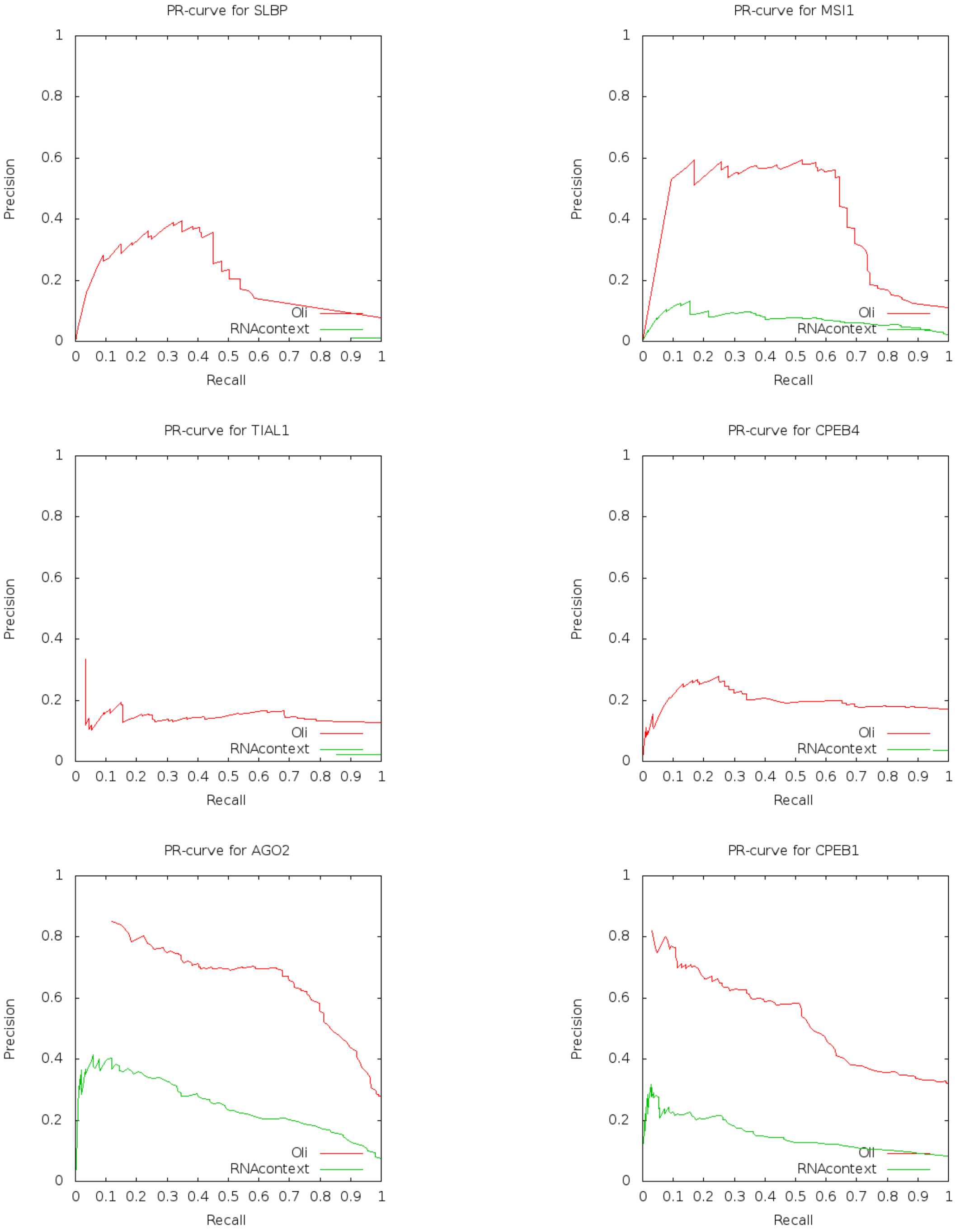


Figure 5.2: **Precision-Recall curves for Oli and RNAcontext on the AURAdataset.** The PR curves visualize the performance of *Oli* (red line) and *RNAcontext* (green line) on protein CUGBP1, TNRC6A, PUM1, TNRC6C, PABP and U2AF65.

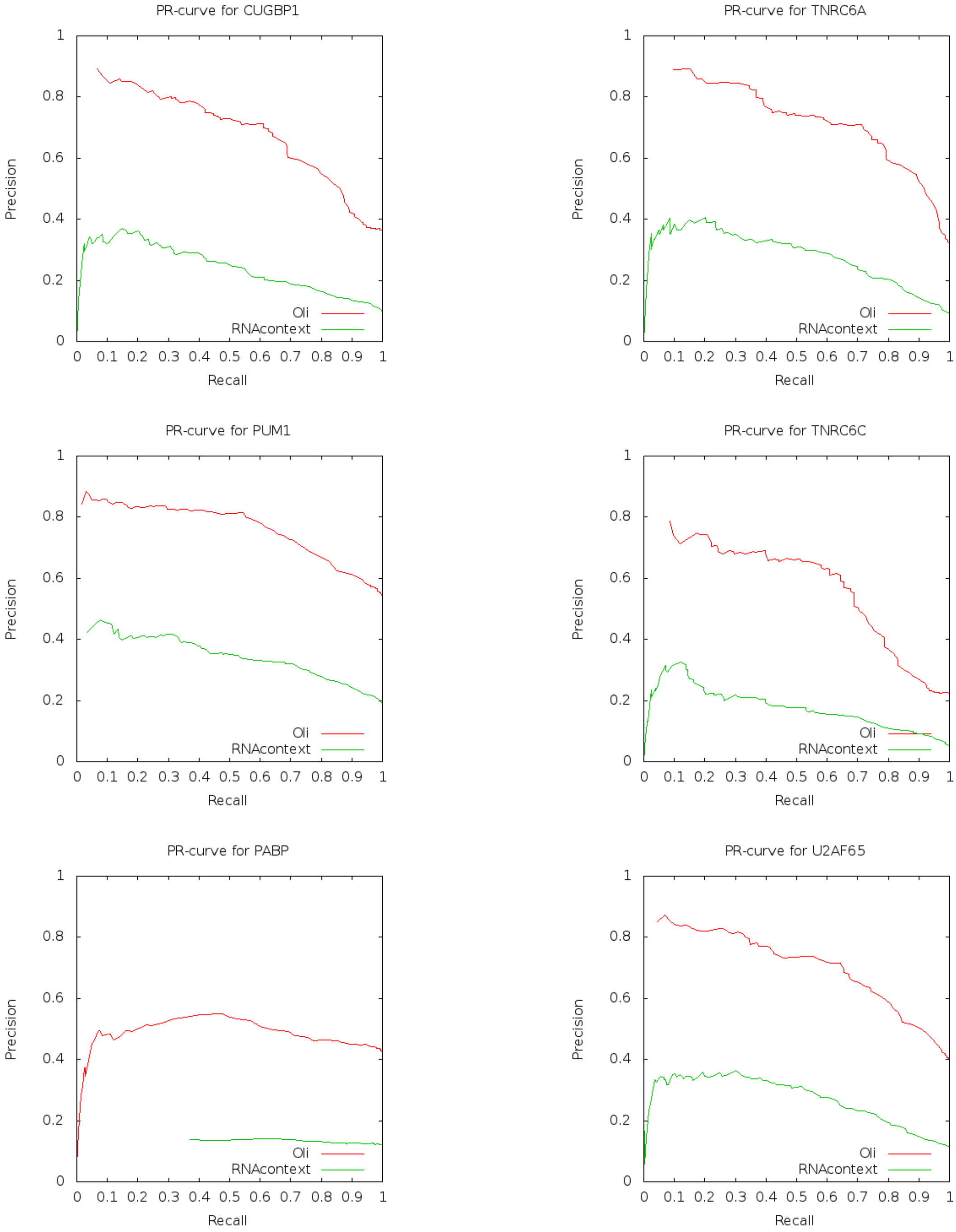


Figure 5.3: **Precision-Recall curves for Oli and RNAcontext on the AURAdataset.** The PR curves visualize the performance of *Oli* (red line) and *RNAcontext* (green line) on RBP AGO4, QKI, TNRC6B, ELAVL1, AUF1 and AGO1.

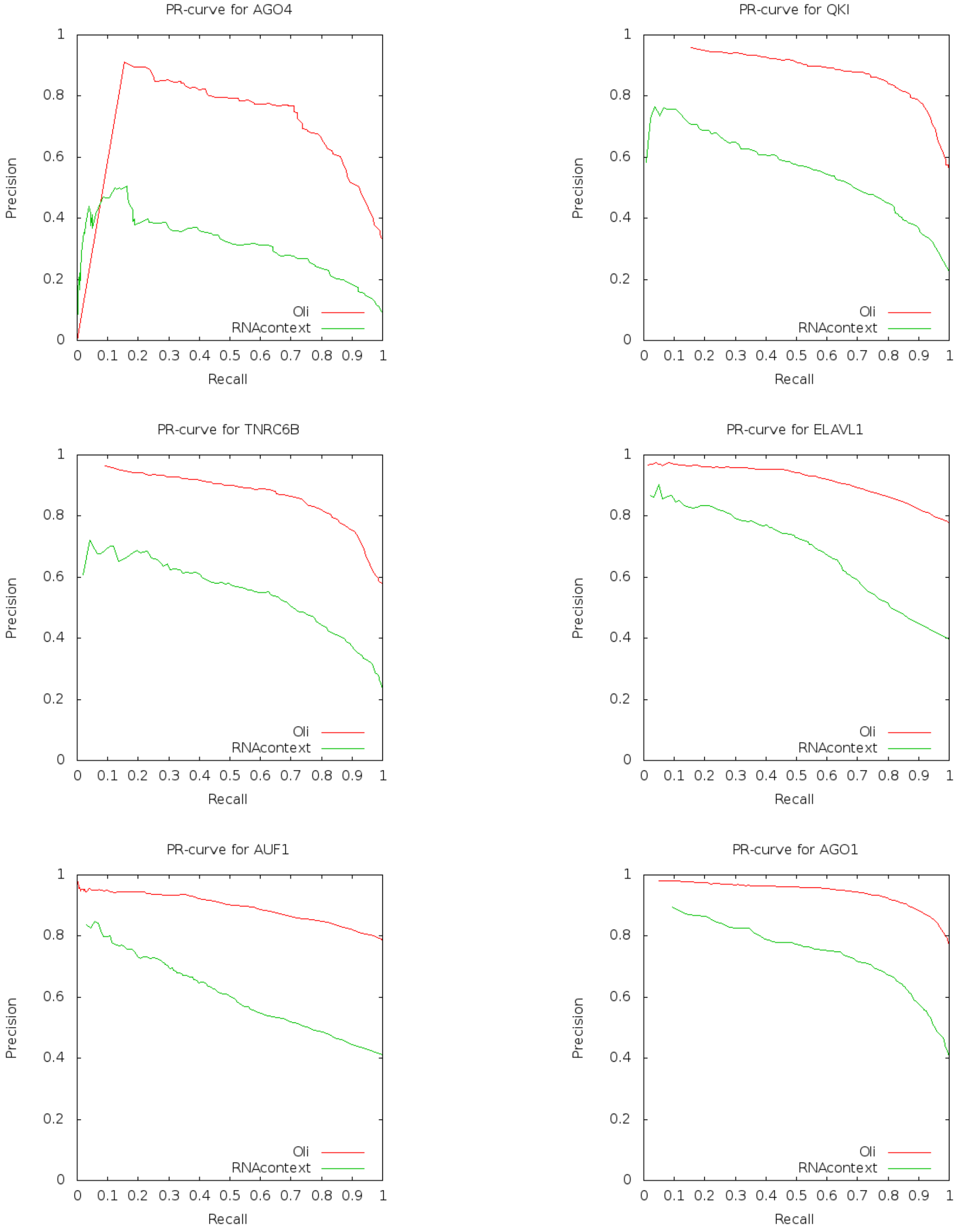


Figure 5.4: **ROC curves for Oli and RNAcontext on the AURAdataset.** The ROC curves for SLBP, MSI1, TIAL1, CEPB4, AGO2 and CPEB1 visualize the performance for *Oli* (red line) and for *RNAcontext* (green line). The curves reflect the AUCs in Table 1 and show the differences in the prediction of the two methods.

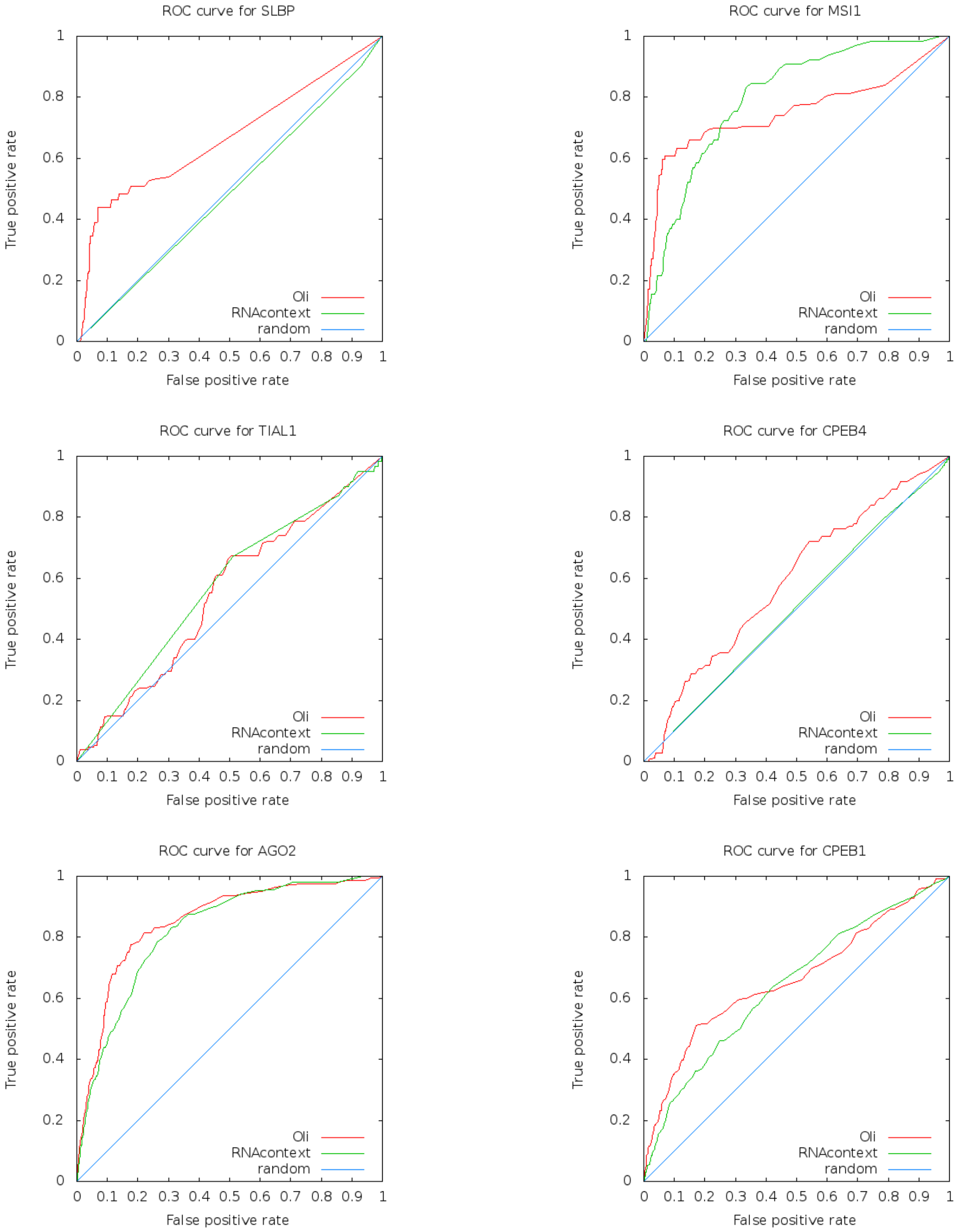
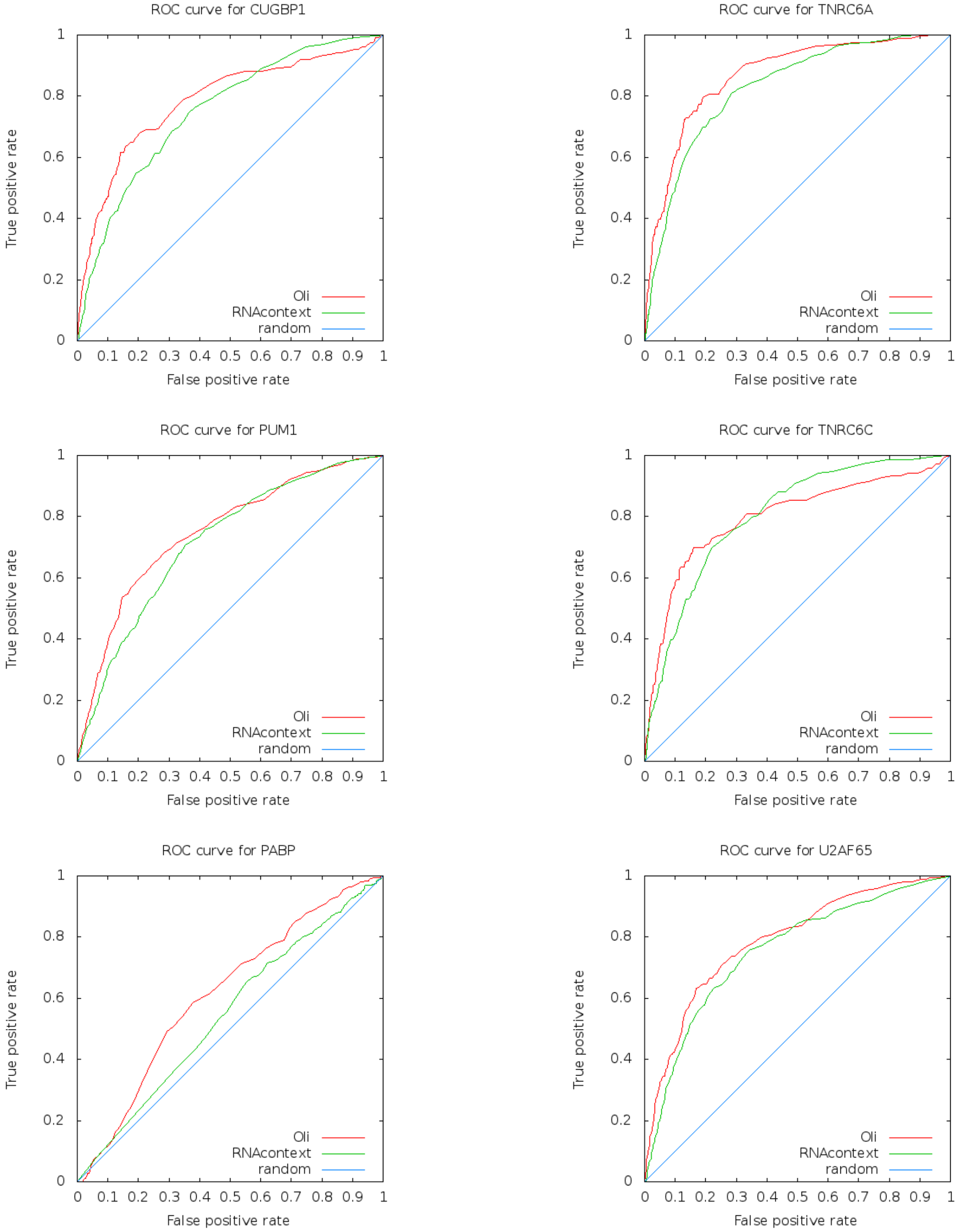


Figure 5.5: **ROC curves for Oli and RNAcontext on the AURAdataset.** The ROC curves for CUGBP1, TNRC6A, PUM1, TNRC6C, PABP and U2AF65 visualize the performance for *Oli* (red line) and for *RNAcontext* (green line).



ROC curve for CUGBP1

ROC curve for TNRC6A

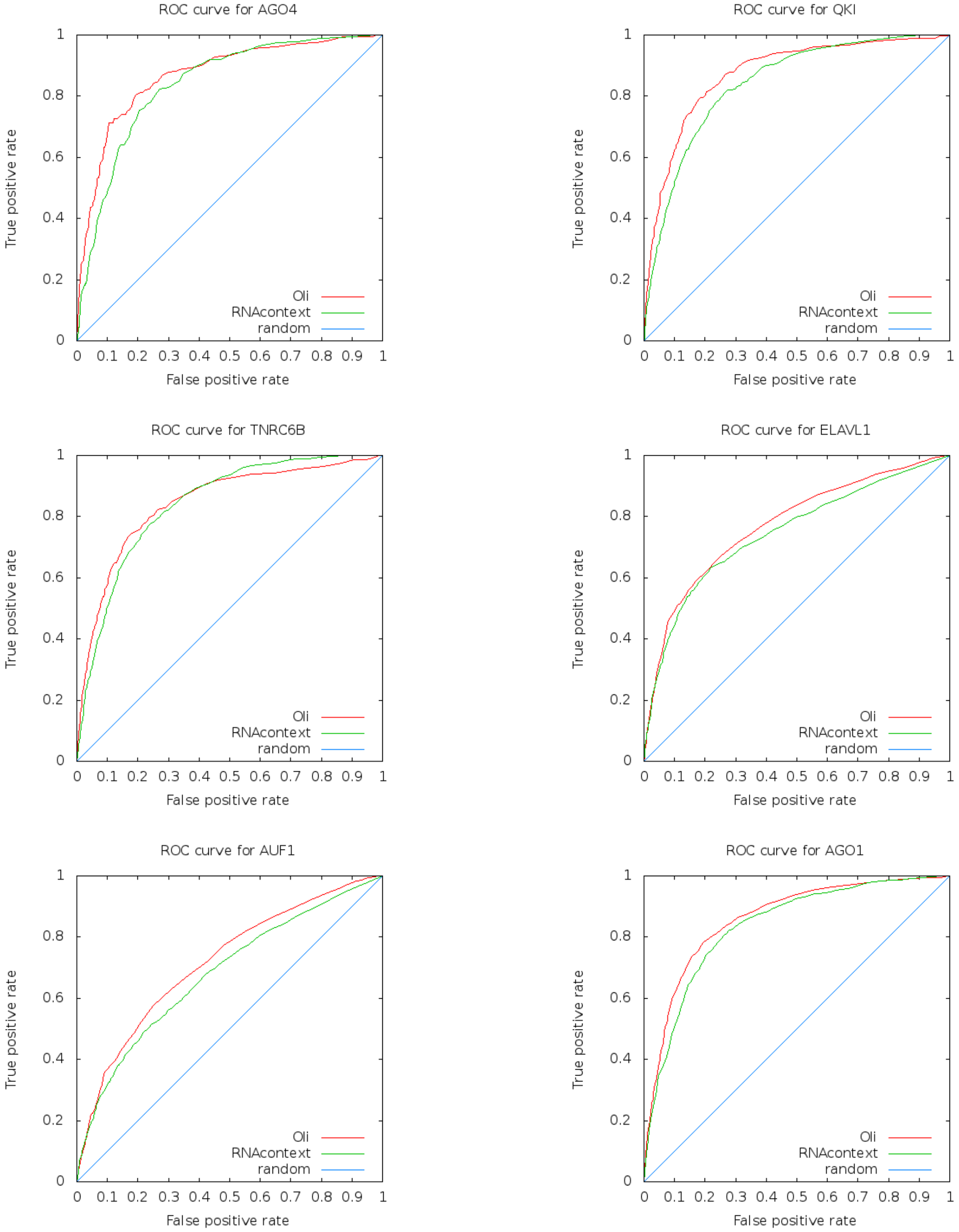
ROC curve for PUM1

ROC curve for TNRC6C

ROC curve for PABP

ROC curve for U2AF65

Figure 5.6: **ROC curves for Oli and RNAcontext on the AURAdataset.** The ROC curves for AGO4, QKI, TNRC6B, ELAVL1, AUF1 and AGO1 visualize the performance for *Oli* (red line) and for *RNAcontext* (green line).



sary, maybe because enough binding information is already included in the oligo representation. Furthermore also the accessibility feature can have a limited impact as some RBPs bind RNA backbones and not accessible ribonucleotides. The motif scores in contrast appear to be more useful. For some proteins, e.g. MSI1, CPEB4 or TNRC6C, *OliMo* performs better than *Oli*. However the oligo-based features are generally better to apprehend the specific binding properties. The relatively good performance of the *Oli* approach on the *AURAdataset* is visualized in Figure 5.7, showing the ROC curve for each RBP.

Although applying the same sequence features, our protein specific discrimination (i.e. one model for each RBP) turns out to be more sensible compared to the general discrimination approach of *RPISeq*. Besides, inferring RBP-RNA binding based only on the presence of specific binding motifs may underestimate the complexity of the binding process, explaining the lower performance of *RNAcontext* in the PR curves. Considering that high-throughput methods produce many data, even a little change in the precision of an *in silico* method results in more correctly predicted binding RNAs.

5.3.2 Evaluation 2

Here we test the prediction of *Oli*, *OliMo*, *OliMoSS*, *RNAcontext*, *RPISeq-SVM* and *RPISeq-RF* on the *PUM2+* data. Just as in Evaluation 1 the negative examples are formed by the randomly selected human 3'UTRs of *3K-*. AUC and precision are calculated to evaluate the models and are reported in Table 5.4. *Oli* and *OliMo* perform equal and obtain identical AUC and precision values. The poorest performance shows *RPISeq-RF*. *RNAcontext* achieves a similar AUC than our three methods which confirm also the ROC curves in Figure 5.8. The curves for *Oli* (red line), *OliMo* (green line), *OliMoSS* (dark blue line) and *RNAcontext* (lila line) are similar, whereas the prediction of *RPISeq-RF* (brown line) is almost "random". However even if the AUC for *Oli*, *OliMo* and *RNAcontext* is similar, their precision is it definitively not: *Oli* and *OliMo* outperform *RNAcontext* clearly with a Prec=0.80 vs. Prec=0.68.

5.3.3 Evaluation 3

In Evaluation 3 we assess the performance of the six methods on *PUM2+* when it is combined with the experimentally derived negatives (*PUM2-*). The results (see Table 5.4) confirm the previously observed behaviour: *Oli* and *OliMo* perform equal on the dataset. Also *RNAcontext* and *RPISeq-SVM* achieve good AUCs but much lower precisions than *Oli* and *OliMo*. The worst performance demonstrates again *RPIseq-RF* with an AUC of 0.52 and a precision of 0.42. Figure 5.9 shows the ROC curves of the methods on the *PUM2+* and *PUM2-* dataset. As before *Oli* (red line) and *OliMo* (green line) perform identically and also *OliMoSS* (dark blue line), *RNAcontext* (lila line) and *RPISeq-SVM* (cyan line) show similar ROC curves. Whereas

Figure 5.7: **Oli performance.** The ROC curves describe the performance of method *Oli* on the *AURAdataset* and on *PUM2+*. The negative data is always formed by *3K-*. More the ROC curve advances to the upper-left corner, the better is the classification ability of the model. A curve near the 45-degree diagonal characterizes a "random" classification.

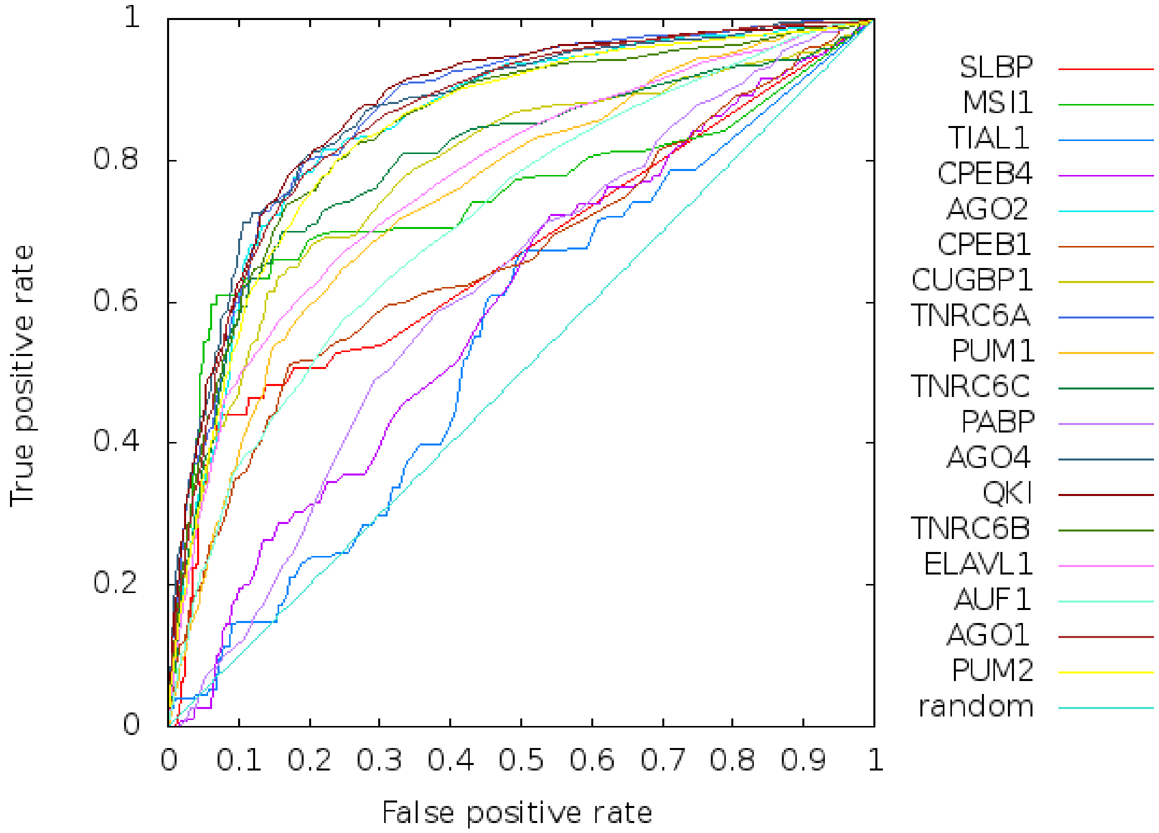


Figure 5.8: **ROC curves of PUM2+ in combination with 3K-.** *Oli* (red line), *OliMo* (green line), *OliMoSS* (dark blue line) and *RNAcontext* (lila line) perform similar. The worst prediction shows *RPISeq-RF* (brown line) which is almost "random".

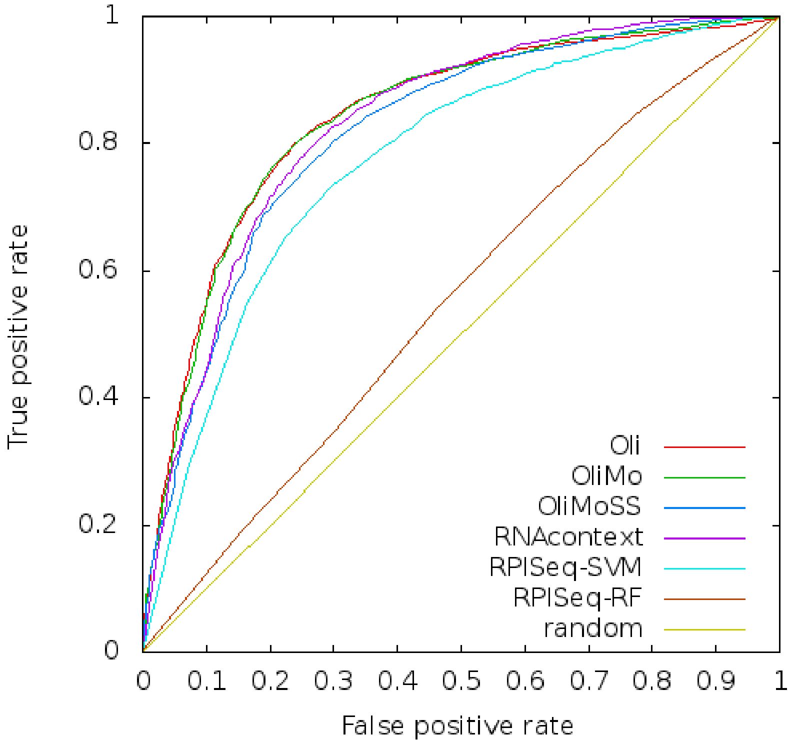
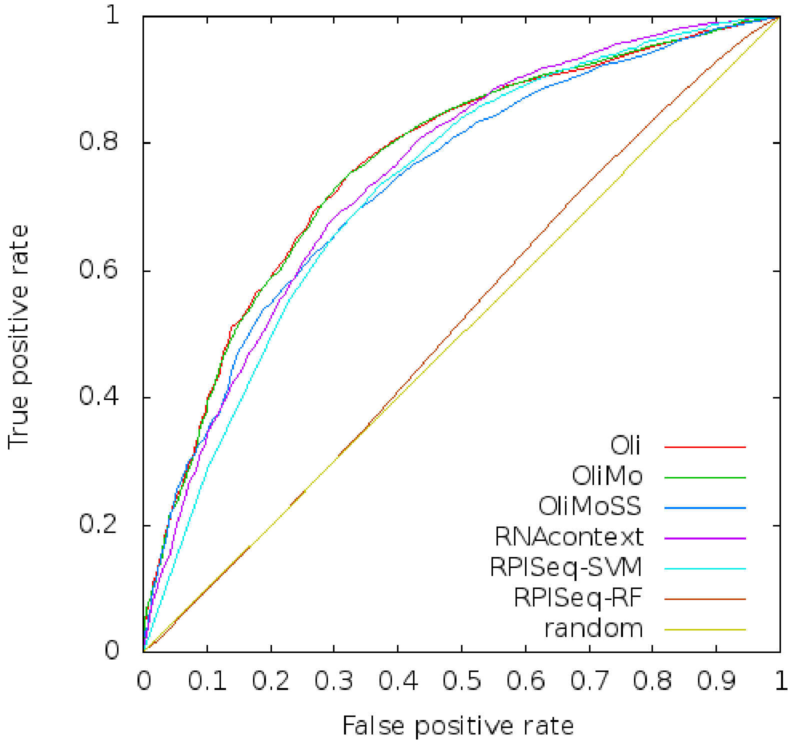


Figure 5.9: **ROC curves of PUM2+ in combination with PUM2-.** The performance of *Oli* (red line) and *OliMo* (green line) is identical. Also the curves for *OliMoSS* (dark blue line), *RNAcontext* (lila line) and *RPISeq-SVM* (cyan line) show a similar behaviour, whereas *RPISeq-RF* (brown line) experiences difficulties to predict the right class. Its ROC curve converges to "random".



RPISeq-RF (brown line) converges to "random".

Secondary structure features do not improve the prediction, confirming the results of Evaluation 1. Above all we expected models trained on real binding data to increase the discrimination, but apparently it decreases the performance. This is evidenced by the different AUCs of the two datasets and by comparing the ROC curves in Figure 5.8 with the ROC curves in Figure 5.9. The curves produced by the dataset with real negatives are flatter then the ones produced by *3K-*.

In order to check the ability of a model based on real negative data to find binding sequences among general 3'UTR sequences, we perform an additional analysis. We consider the models generated by the 10-fold cross validation done with *PUM2+* and *PUM2-* and test them, substituting the negatives of each of the 10 test sets with negatives taken from *3K-*. For this we use *Oli*. The results are shown in Table 5.5. Evidently an approach based on real data is also able to distinguish between real positives and randomly selected sequences. Moreover, the task appears to be easier than distinguishing real negatives because the model improves

Table 5.4: **Performance of Oli, OliMo, OliMoSS, RNAcontext and both RPISeq methods on PUM2+ in combination with two different negative datasets.** The table contains the performance values AUC and precision (Prec) for each method on two different datasets: one with *PUM2+* and randomly selected 3'UTRs *3K-* and one with *PUM2+* and experimental negatives *PUM2-*.

Pos. data	Neg. data	Value	*Oli*	*OliMo*	*OliMoSS*	*RNAcontext*	*RPISeq-SVM*	*RPISeq-RF*
PUM2+	*3K-*	AUC	0.84	0.84	0.82	0.83	0.77	0.56
		Prec	0.80	0.80	0.74	0.68	0.47	0.40
PUM2+	*PUM2-*	AUC	0.77	0.77	0.74	0.75	0.73	0.52
		Prec	0.73	0.73	0.69	0.59	0.48	0.42

Figure 5.10: **PR curve of Oli on PUM2+ in combination with 3K- and PUM2-.** Method *Oli* when trained on real negatives (red line) and trained on the random 3'UTRs (green line) does not show big differences in recall and precision. Therefore random sequences can be a good approximation if no experimental negatives are available.

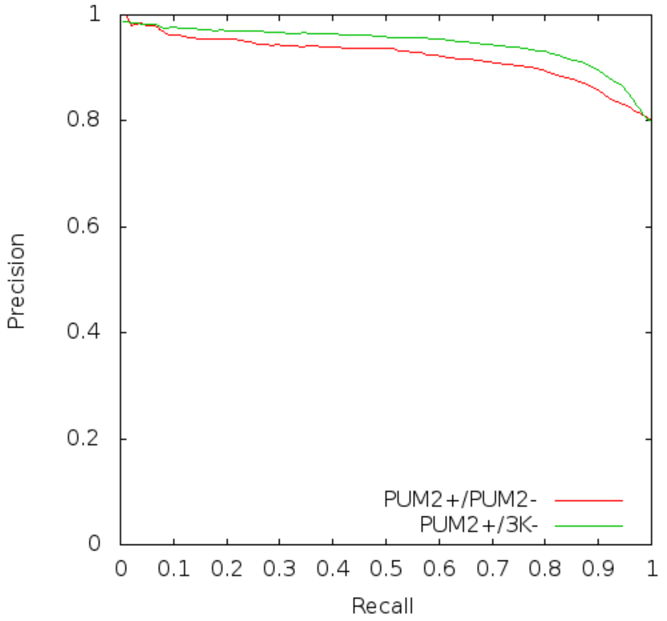


Table 5.5: **Additional analysis to test the ability of a model trained on real negatives to identify binding sequences among general 3'UTRs and vice versa.** First the models are trained on *PUM2+* and *3K-* and *PUM2+* and *PUM2-* are used to test. The results are shown in the first row. Then the models are generated with *PUM2+* and *PUM2-* and the test is performed with *PUM2+* and the negatives from *3K-*. The results are shown in the second row.

Training	Testing	Sens	Spec	Prec
PUM2+/3K-	PUM2+/PUM2-	0.6	0.8	0.69
PUM2+/PUM2-	PUM2+/3K-	0.52	0.9	0.81

in precision (Prec=0.81). This is consistent with the fact that all the methods reach better performances when the negatives are formed by *3K-*. Therefore a complete dataset obtained by *in vivo* experiments can be effectively used to train SVMs with simple sequence features. This all highlights the importance of high-quality negative training data. The problem is that non-binding information is rarely available but necessary to build models.

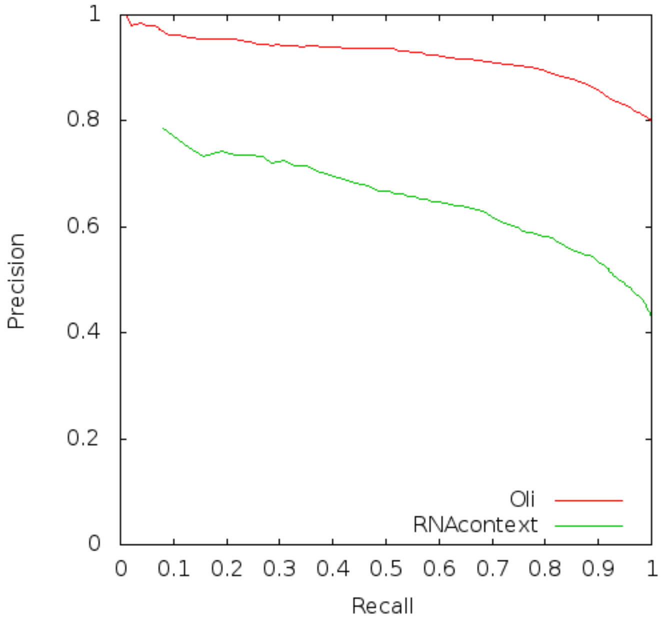
Calculations of the Pearson coefficient on each row of Table 5.4 confirms a correlation between the predictions of *PUM2-* and *3K-* for all methods. In fact a Pearson coefficient of 0.99 let us conclude that random sequences could be a good approximation if no experimental negatives are available. Moreover, they can be used to assess the relative performance of the methods, as has been done in Evaluation 1. This is confirmed also by the precision and recall in the PR curves of Figure 5.10. The curves for *Oli* trained on the real negatives (red line) and on the random 3'UTRs (green line) do not show a big difference in recall and precision.

Finally the difference in precision, regarding the training with *PUM2-*, is of 0.14 between *Oli* and *RNAcontext* and of 0.25 between *Oli* and *RPISeq-SVM*. As discussed also in Evaluation 1, even a little change in the precision is important. Considering 1000 RNAs a 0.14 higher precision results in additional 140 right classified RNA sequences. The impact of the difference between *Oli* and *RNAcontext* gets clearer in Figure 5.11 which draws their PR curve on *PUM2+* and *PUM2-*.

5.4 Conclusions

The knowledge of RBP-RNA interactions is of interest in biology. More specifically the identification of RNA-RBP binding is important to understand the protein function, something that currently can be done only through *in vivo* and *in vitro* laboratory experiments. In this chapter we applied SVMs to experimental datasets and attempted to predict the RNA targets for different RBPs. We proposed to describe RNA sequences in 3 different ways: the first method, called *Oli*, uses oligos as features; the second method, called *OliMo*, where we add

Figure 5.11: **Precision-Recall curves for Oli and RNAcontext.** Regarding the precision and recall *Oli* (red line) outperforms *RNAcontext* (green line) on the experimental dataset composed of *PUM2+* and *PUM2-*. More the curve advances to the upper-right corner, better performs the classifier.

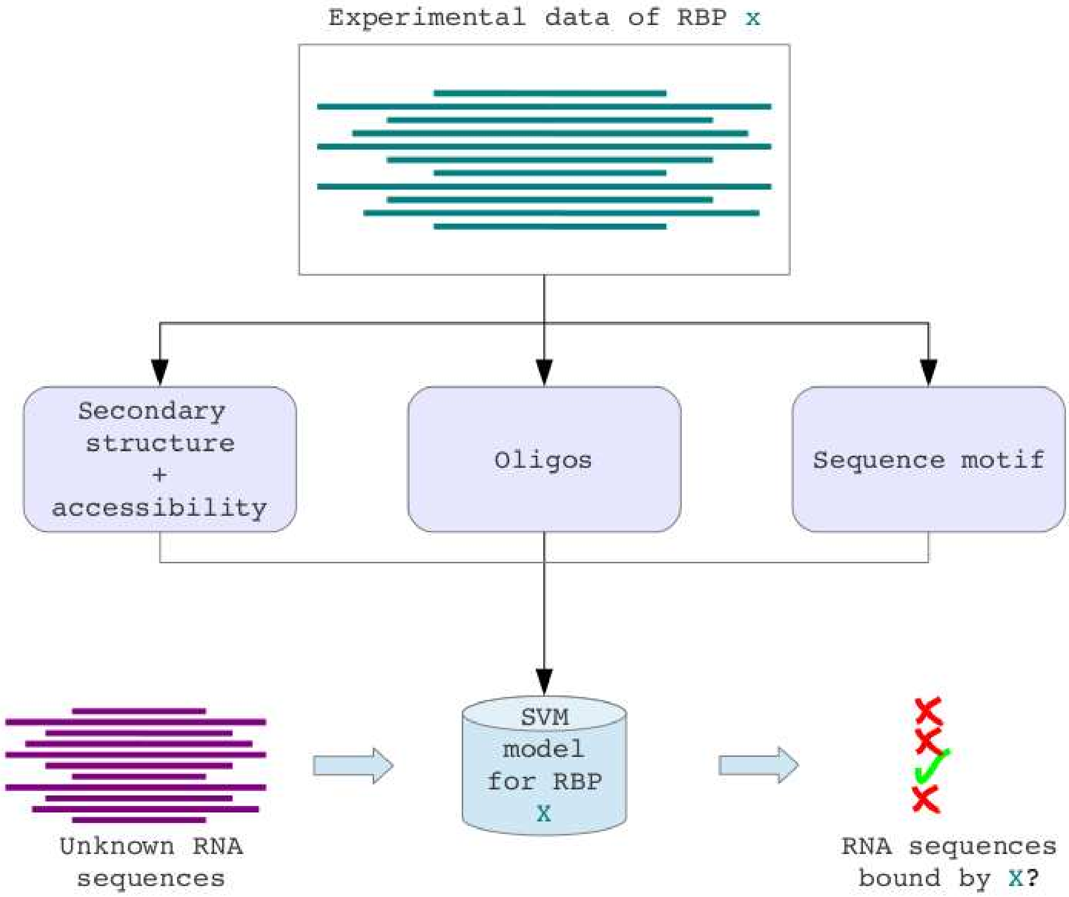


motif scores of an automatically detected binding motif; the third method, called *OliMoSS*, where we extend *OliMo* with secondary structure features. Finally we compared the predictions of our methods with *RNAcontext* and *PRISeq*. *Oli* and *OliMo*, achieved higher performance values than *OliMoSS* and *RPISeq*. Despite applying the same oligo-based features, *Oli* outperforms *RPISeq* which reinforces our decision to train a separate SVM for each RBP. Also binding motifs alone seem not to be discriminative enough on our datasets, observable on the higher precision of *Oli* and *OliMo* compared to *RNAcontext*. Comparing models trained on experimentally detected non-binding sequences with models trained on randomly chosen RNA sequences shows that the latter performs better.

A suitable method is therefore to use RBP-specific experimental data, combine it with negative examples and train an SVM with the proposed features. Such a model can than be applied to predict the binding of the corresponding RBP with other RNA sequences. The methods can be chosen on the basis of the individual binding properties of the RBP, if they are previously known. Figure 5.12 illustrates the idea. In fact our proposed approach is easily extendible to more RBPs and to any type of experimentally derived dataset.

We conclude that simple sequence information, like the oligo representation of the RNA sequence, in combination with experimental binding data can be effectively used to construct predictive models. The right choice of negative training examples is important. They can be

Figure 5.12: **The approach.** Our approach is easily extendible to more experimental datasets which are used to train an SVM with different features. The model is then applied to predict the binding of the corresponding RBP with other RNA sequences. If the individual binding specificities of the RBP are previously known, the most appropriate feature composition can be chosen.


Experimental data of RBP x
Secondary structure + accessibility
Oligos
Sequence motif
Unknown RNA sequences
SVM model for RBP X
RNA sequences bound by X?

approximated with random sequences if real data is not available, but ideally they derive from the same experiment, under the same conditions and from the same cell line.

Chapter 6

Conclusions and future work

Each protein has its own binding characteristics and the different recognition strategies make it difficult to express one general binding rule over all RBPs. Several works focus on the identification and the dissection of protein binding sites, but the obtained information can not be applied to all RBPs. This fact makes the development of computational methods, predicting *in silico* the binding between an RBP and an RNA, tricky. Laboratory experiments instead are able to detect *in vivo* and *in vitro* the bound RNAs. They produce data which contains important information about the specific RBP binding. The advantage of these techniques is the genome-wide detection of real binding couples in living cells.

In this thesis we focused on the *in silico* prediction of RBP-binding partners, used SVMs and exploited experimental datasets. In the preliminary work of Chapter 3 we implemented five classifiers, based on different features, to predict binding amino acids in protein sequences. Thereafter we attempted the construction of the targeted RNA sequence by combining the predicted binding residues with a basic propensity statistic. The poor results let us conclude that the prediction of the binding elements was not sensitive enough and the propensity statistic we used was too simple. Additionally the dataset contained only one RNA-binding partner for each RBP, which provides not enough binding information. Therefore in Chapter 4 we started to exploit a CLIP-seq dataset, performed on RBP CELF1, which assures much more binding sequences. We trained a SVM with different features to discriminate between the detected binding and the non-binding RNAs. The features are based on simple sequence properties like oligo frequencies and sequence motifs. Regarding the results we conclude that high-throughput datasets can successfully be used to create predictive models. Hence in Chapter 5 we applied SVMs to more experimental datasets with different RBPs and extended the idea with more features like the secondary structure of the RNA sequence and the accessibility. In total we proposed three different approaches: the first method is called *Oli* and uses simple oligo frequencies as features; the second method is called *OliMo* and extends *Oli* with motif scores. The third method is called *OliMoSS* and uses, additionally to the previously described

features, secondary structures and accessibility. A comparison with *RNAcontext* and *PRISeq* showed that *Oli* and *OliMo* performed better on our data.
"Simple reliance on sequence motifs or regularities for predicting protein binding to a RNA element is dangerous" (Westhof and Fritsch, 2011). Even if such an affirmation seems reasonable from a biological point of view, we have to consider the fact that an exact and precise overall description of RBP-RNA binding does not exist. Bioinformatics can mine and analyse data but can give good results only and only if the problem is well described and if enough high quality data is available.

Future directions

So far it is unclear whether the use of PSSM scores alone performs better or worse than the described SVM. To determine which approach performs better a thorough analysis between SVMs and PSSMs is in progress. In future we plan also to test our CELF1-model described in Chapter 4 on other species (e.g. CLIP-seq data for mouse available) with the goal to predict CELF1-binding RNAs in other cell lines.
In Chapter 5 we use the same oligos for each RBP. Calculating the information gain for each oligo individually on each RBP , the number of features could be reduced and only the most important oligos could be used to create the RBP-dependent model. In this way we adapt the features to each RBP which maybe improves the sensitivity of the approaches. Beside this we plan to apply one-class SVMs to avoid the unbalance of the datasets and to avoid the necessity of "artificial" negative data.
Parts of this thesis have been published in the proceedings of the *6th International Conference on Practical Applications of Computational Biology & Bioinformatics- PACBB12, 2012, Salamanca, Spain* and have been submitted to *BMC Bioinformatics* (currently under revision).

Bibliography

Allers, J. and Shamoo, Y. (2001). Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *Journal of Molecular Biology*, 311(1):75.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Änkö, M.-L., Müller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J., and Neugebauer, K. M. (2012). The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biology*, 13(3):R17.

Änkö, M.-L. and Neugebauer, K. M. (2012). RNA–protein interactions in vivo: global gets specific. *Trends in Biochemical Sciences*, 37(7):255–262.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I., and Stockinger, H. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research*, 40(W1):W597–W603.

Auweter, S. D., Oberstrass, F. C., and Allain, F. H.-T. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, 34(17):4943–4959.

Bahadur, R. P., Zacharias, M., and Janin, J. (2008). Dissecting protein–RNA recognition sites. *Nucleic Acids Research*, 36(8):2705–2716.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(suppl 2):W202–W208.

Bailey, T. L. and Elkan, C. (1994). *Fitting a mixture model by expectation maximization to discover motifs in bipolymers*. Department of Computer Science and Engineering, University of California, San Diego.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, 41(D1):D36–D42.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer Jr., E. E., Brice, M. D., Brice, J. R., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112(535).

Chang, C.-C. and Lin, C.-J. (2011). *LIBSVM: A library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chen, Y. C. and Lim, C. (2008). Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Research*, 36(5):e29.

Cheng, C.-W., Su, E. C., Hwang, J.-K., Sung, T.-Y., and Hsu, W.-L. (2008). Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinformatics*, 9(Suppl 12):S6.

Choi, S. and Han, K. (2011). Prediction of RNA-binding amino acids from protein and RNA sequences. *BMC Bioinformatics*, 12(Suppl 13):S7.

Ciriello, G., Gallina, C., and Guerra, C. (2010). Analysis of interactions between ribosomal proteins and RNA structural motifs. *BMC Bioinformatics*, 11(Suppl 1):S41+.

Cléry, A., Blatter, M., and Allain, F. H. (2008). RNA recognition motifs: boring? Not quite. *Current Opinion in Structural Biology*, 18(3):290–298.

Corà, D., Di Cunto, F., Caselle, M., and Provero, P. (2007). Identification of candidate regulatory sequences in mammalian 3'UTRs by statistical analysis of oligonucleotide distributions. *BMC Bioinformatics*, 8(1):174.

Corcoran, D., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R., Keene, J., and Ohler, U. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biology*, 12(8):R79.

Corden, J. L. (2010). Shining a new light on RNA-protein interactions. *Chemistry & Biology*, 17(4):316–318.

Dassi, E., Malossini, A., Re, A., Mazza, T., Tebaldi, T., Caputi, L., and Quattrone, A. (2012). AURA: Atlas of UTR Regulatory Activity. *Bioinformatics*, 28(1):142–144.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240. ACM.

Draper, D. E. (1999). Themes in RNA-protein recognition. *Journal of Molecular Biology*, 293(2):255–270.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

Ellis, J. J., Broom, M., and Jones, S. (2007). Protein–RNA interactions: Structural analysis and functional classes. *Proteins: Structure, Function, and Bioinformatics*, 66(4):903–911.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Gräf, S., Haider, S., Hammond, M., Howe, K., Jenkinson, A., Johnson, N., Kähäri, A., Keefe, D., et al. (2010). Ensembl's 10th year. *Nucleic Acids Research*, 38(suppl 1):D557–D562.

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., et al. (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Research*, 39(suppl 1):D876–D882.

Glisovic, T., Bachorik, J. L., Yong, J., and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, 582(14):1977–1986.

Gupta, A. and Gribskov, M. (2011). The role of RNA sequence and structure in RNA–protein interactions. *Journal of Molecular Biology*, 409(4):574–587.

Guzman, R. N. D., Turner, R. B., and Summers, M. F. (1998). Protein-RNA recognition. *Biopolymers (Nucleic Acid Sciences)*, 48:181–195.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188.

Hoffman, M. M., Khrapov, M. A., Cox, J. C., Yao, J., Tong, L., and Ellington, A. D. (2004). AANT: the amino acid-nucleotide interaction database. *Nucleic Acids Research*, 32:D174–D181.

Hogan, D., Riordan, D., Gerber, A., Herschlag, D., and Brown, P. (2008). Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biology*, 6(10):e255.

Huang, Y.-F., Chiu, L.-Y., Huang, C.-C., and Huang, C.-K. (2010). Predicting RNA-binding residues from evolutionary information and sequence conservation. *BMC Genomics*, 11(Suppl 4):S2.

Inza, I., Calvo, B., Armañanzas, R., Bengoetxea, E., Larrañaga, P., and Lozano, J. (2010). Machine learning: an indispensable tool in bioinformatics. *Methods in Molecular Biology*, 593:25–48.

Jain, R., Devine, T., George, A., Chittur, S., Baroni, T., Penalva, L., and Tenenbaum, S. (2011). RIP-Chip analysis: RNA-binding protein immunoprecipitation-microarray (Chip) profiling. In Nielsen, H., editor, *RNA*, volume 703 of *Methods in Molecular Biology*, pages 247–263. Humana Press.

Jaskiewicz, L., Bilen, B., Hausser, J., and Zavolan, M. (2012). Argonaute CLIP – A method to identify in vivo targets of miRNAs. *Methods*, 58(2):106–112.

Jensen, L. J. and Bateman, A. (2011). The rise and fall of supervised machine learning techniques. *Bioinformatics*, 27(24):3331–3332.

Jeong, E., Chung, I.-F., and Miyano, S. (2004). A neural network method for identification of RNA-interacting residues in protein. *Genome Informatics*, 15(1):105–116.

Jeong, E., Kim, H., Lee, S.-W., and Han, K. (2003). Discovering the interaction propensities of amino acids and nucleotides from protein-RNA complexes. *Molecules and Cells*, 16(2):161–167.

Jones, S., Daley, D. T., Luscombe, N. M., and Berman, H. M. (2001). Protein-RNA interactions: a structural analysis. *Nucleic Acids Research*, 29(4):943–954.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637.

Kazan, H., Ray, D., Chan, E. T., Hughes, T. R., and Morris, Q. (2010). RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Computational Biology*, 6(7):e1000832.

Khalil, A. M. and Rinn, J. L. (2011). RNA–protein interactions in human health and disease. *Seminars in Cell & Developmental Biology*, 22(4):359–365.

Kim, H., Jeong, E., Lee, S.-W., and Han, K. (2003). Computational analysis of hydrogen bonds in protein–RNA complexes for interaction patterns. *FEBS Letters*, 552(2–3):231–239.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature*, 8(7):559–564.

Kishore, S., Luber, S., and Zavolan, M. (2010). Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Briefings in Functional Genomics*, 9(5-6):391–404.

Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30:25–36.

Kress, C., Gautier-Courteille, C., Osborne, H., Babinet, C., and Paillard, L. (2007). Inactivation of CUG-BP1/CELF1 causes growth, viability, and spermatogenesis defects in mice. *Molecular and Cellular Biology*, 27(3):1146–1157.

Kumar, M., Gromiha, M. M., and Raghava, G. (2007). Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins: Structure, Function, and Bioinformatics*, 71(1):189–194.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132.

Lee, S. and Blundell, T. L. (2009). BIPA: a database for protein–nucleic acid interaction in 3D structures. *Bioinformatics*, 25(12):1559–1560.

Li, W. and Godzik, A. (2006). CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

Li, X., Quon, G., Lipshitz, H. D., and Morris, Q. (2010). Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, 16(6):1096–1107.

Lichtarge, O. and Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Current Opinion in Structural Biology*, 12(1):21–27.

Liu, Z.-P., Wu, L.-Y., Wang, Y., Zhang, X.-S., and Chen, L. (2010). Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics*, 26(13):1616–1622.

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26.

Maetschke, S. and Yuan, Z. (2009). Exploiting structural and topological information to improve prediction of RNA-protein binding sites. *BMC Bioinformatics*, 10(1):341.

Magrane, M. et al. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database: The Journal of Biological Databases and Curation*, 2011.

Marquis, J., Paillard, L., Audic, Y., Cosson, B., Danos, O., Bec, C. L., and Osborne, H. B. (2006). CUG-BP1/CELF1 requires UGU-rich sequences for high-affinity binding. *Biochemical Journal*, 400(2):291–301.

McDonald, I. K. and Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology*, 238(5):777–793.

Morozova, N., Allers, J., Myers, J., and Shamoo, Y. (2006). Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Structural Bioinformatics*, 22(220):2746–2752.

Mukherjee, N., Corcoran, D. L., Nusbaum, J. D., Reid, D. W., Georgiev, S., Hafner, M., Jr., M. A., Tuschl, T., Ohler, U., and Keene, J. D. (2011). Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Molecular Cell*, 43(3):327–339.

Muppirala, U. M., Honavar, V. G., and Dobbs, D. (2011). Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics*, 12(1).

Pancaldi, V. and Bähler, J. (2011). In silico characterization and prediction of global protein–mRNA interactions in yeast. *Nucleic Acids Research*, 39(14):5826–5836.

Parrill, A. (1997). Educational materials for organic chemistry . http://www.cem.msu.edu/ cem252/.

Pérez-Cano, L. and Fernández-Recio, J. (2010). Optimal protein-RNA area, OPRA: A propensity-based method to identify RNA-binding sites in proteins. *Proteins: Structure, Function, and Bioinformatics*, 78(1):25–35.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(suppl 1):D61–D65.

Puton, T., Kozlowski, L., Tuszynska, I., Rother, K., and Bujnicki, J. (2011). Computational methods for prediction of protein-RNA interactions. *Journal of Structural Biology*.

Ray, D., Kazan, H., Chan, E., Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B., Morris, Q., and Hughes, T. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology*, 27(7):667–670.

Segata, N. (2009). FaLKM-lib v1.0: a Library for Fast Local Kernel Machines. Technical report, DISI, University of Trento, Italy. Software available at `http://disi.unitn.it/~segata/FaLKM-lib`.

Shazman, S., Elber, G., and Mandel-Gutfreund, Y. (2011). From face to interface recognition: a differential geometric approach to distinguish DNA from RNA binding surfaces. *Nucleic Acids Research*, 39(17):7390–7399.

Shulman-Peleg, A., Nussinov, R., and Wolfson, H. J. (2009). RsiteDB: a database of protein binding pockets that interact with RNA nucleotide bases. *Nucleic Acids Research*, 37(suppl 1):D369–D373.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., and Kasprzyk, A. (2009). BioMart–biological queries made easy. *BMC Genomics*, 10(1):22.

Spirin, S., Titov, M., Karyagina, A., and Alexeevski, A. (2007). NPIDB: a Database of Nucleic Acids-Protein Interactions. *Structural Bioinformatics*, 23(23):3247–3248.

Spriggs, R., Murakami, Y., Nakamura, H., and Jones, S. (2009). Protein function annotation from sequence: prediction of residues interacting with RNA. *Bioinformatics*, 25(12):1492–1497.

Terribilini, M., Lee, J., Yan, C., Jernigan, R., Honavar, V., and Dobbs, D. (2006). Prediction of RNA binding sites in proteins from amino acid sequences. *RNA*, 12(8):1450–1462.

Tong, J., Jiang, P., and Lu, Z.-h. (2008). RISP: A web-based server for prediction of RNA-binding sites in proteins. *Computer Methods and Programs in Biomedicine*, 90(2):148–153.

Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968):505–510.

Uren, P. J., Bahrami-Samani, E., Burns, S. C., Qiao, M., Karginov, F. V., Hodges, E., Hannon, G. J., Sanford, J. R., Penalva, L. O., and Smith, A. D. (2012). Site identification in high-throughput RNA–protein interaction data. *Bioinformatics*, 28(23):3013–3020.

Uren, P. J., Burns, S. C., Ruan, J., Singh, K. K., Smith, A. D., and Penalva, L. O. F. (2011). Genomic analyses of the RNA-binding protein Hu Antigen R (HuR) identify a complex network of target genes and novel characteristics of its binding sites. *Journal of Biological Chemistry*, 286(43):37063–37066.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag.

Wang, C.-C., Fang, Y., Xiao, J., and Li, M. (2011). Identification of RNA-binding sites in proteins by integrating various sequence information. *Amino Acids*, 40:239–248.

Wang, L. and Brown, S. J. (2006). BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Research*, 34(suppl 2):W243–W248.

Wang, L., Huang, C., Yang, M., and Yang, J. (2010). BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Systems Biology*, 4(Suppl 1):S3+.

Wang, Y., Xue, Z., Shen, G., and Xu, J. (2008). PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids*, 35(2):295–302.

Westhof, E. and Fritsch, V. (2011). The endless subtleties of RNA-protein complexes. *Structure*, 19(7):902–903.

Zhang, C. and Darnell, R. B. (2011). Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nature Biotechnology*, 29(7):607–614.