universität wien

DIPLOMARBEIT

Titel der Diplomarbeit

Investigation and prediction of interactions between AU-rich binding proteins and AU rich elements and the generation of the 'AREsite' webserver

Verfasser

Jörg Fallmann

angestrebter akademischer Grad

Magister der Naturwissenschaften (Mag. rer. nat.)

Wien, am November 25, 2011

Studienkennzahl lt. Studienblatt:	**A 490**
Studienrichtung lt. Studienblatt:	**Diplomstudium Molekulare Biologie**
Betreuer:	**Univ.-Prof. Dipl.-Phys. Dr. Ivo L. Hofacker**

DANKSAGUNG

An dieser Stelle moechte ich meinen Dank an Alle aussprechen die zum Gelingen dieser Arbeit beigetragen haben.

Ivo Hofacker für Betreuung und Unterstützung meiner Diplomarbeit.
Für Ratschläge, Verbesserungen und viele Stunden des Korrekturlesens bedanke ich mich herzlichst bei Berni, Sven, Christian, Ronny und bei Andreas, der schon im Vorfeld immer wieder mit Rat und Tat zur Stelle war.
Des weiteren geht mein Dank an meinen Zimmerkollegen Florian für seinen ungebremsten Optimismus auch wärend längerer Durststrecken, sowie den Rest des TBI, für geteiltes Leid und gemeinsame Freude an der Sache.

Mein besonderer Dank gilt meinen Eltern Klaus und Christa und meiner Schwester Anja, deren Unterstützung ich mir immer sicher sein konnte und die es sich nie nehmen liesen mir in jeder nur erdenklichen Art und Weise unter die Arme zu greifen!
Sowie meiner Gefährtin Andrea, für Unterstützung und Liebe in zehn gemeinsamen Jahren.

Abstract

This thesis is focused on AU-rich elements (ARE) and AU-rich binding proteins (AUBPs). AU-rich elements can be found in the 3' untranslated region of messenger RNAs (mRNAs). They are RNA sequence motifs and their interacting proteins play a a major role for regulation of mRNA stability.
Various layers of regulation exist in an organism that allow it to regulate the level of gene expression according to it's needs. One of those layers is the regulation of mRNA stability, which allows an organism to regulate the amount of protein levels very fast and effective. A well known example for this type of control is the interaction of AU-rich binding proteins and their target mRNA. This proteins can bind to ARE sequence motifs in the mRNAs 3' untranslated region and stabilize or destabilize their target as a function of the type and amount of AUBPs bound and agonistic or antagonistic effects by other AUBPs or RNA binding factors.
As regulation of gene expression via mRNA stability is a very fast and precise method, these proteins and their targets are attracting the attention of researchers in various fields, from cancer research to synthetic biology. The main goal of this thesis is a better understanding of ARE motifs that mark mRNAs as AUBP targets and in the following the prediction of novel AUBP targets with bioinformatical analysis. To analyze these RNA binding proteins, a database was generated, that contains human and mouse transcripts which where annotated for the presence of ARE motifs. Information on the accessibility and fold-enrichment, as well as on phylogenetic conservation of this motifs in known one-to-one orthologs was added by *in silico* folding of the mRNA 3' UTRs, the generation of multiple alignments and analyzation of the nucleotide composition in the 3' UTR with an order-0 and an order-1 markov model . A second database containing literature about the effects of AUBPs on experimentally validated AUBP targets was produced as well. The webserver 'AREsite' has been built and published (1), combining both databases as backends and making the generated information freely accessible, thereby presenting a tool that can be used to examine known, or to predicted novel AUBP targets.
Part of this thesis was the analyzation of information provided by this webservice for the prediction of novel AUBP targets.
The results of this analysis are presented in the section 5. The generated webserver 'AREsite' is avaliable at `http://rna.tbi.univie.ac.at/AREsite` and has already been used as tool for the analysis of AUBP targets (2).

Zusammenfassung

Diese Diplomarbeit beschäftigt sich mit der Rolle von AU-reichen Elementen (ARE) und an sie bindende Proteine (AUBPs). AU-reiche Elemente finden sich in der 3' untranslatierten Region von messenger RNAs(mRNAs). Sie sind RNA Sequenz Motife und die mit ihnen interagierenden AUBPs spielen eine wichtige Rolle in der Regulation der Stabilität von mRNA. In einem Organismus finden sich viele Möglichkeiten die Expression von Genen in Abhaängigkeit des Bedarfs zu regulieren. Eine sehr schnelle und effektive Möglichkeiten stellt die genannte Regulation der mRNA Stabilität dar. Unter den bekannten Mechanismen für diese Art der Regulation finden sich die Interaktionen von AUBPs mit der zu regulierenden mRNA. Diese Proteine binden an AU-reiche Sequenzabschnitte in der 3' untranslatierten Region von mRNAs und können ihre Stabilität in Abhängigkeit ihrer Art und Anzahl, sowie agonistischer oder antagonistischer Effekte von weiteren AUBPs oder RNA bindenden Faktoren, regulieren.
Die Tatsache der schnellen und effektiven Regulation der Genexpression durch mRNA De-/Stabilisierung lässt diese Proteine und durch sie regulierte mRNAs vermehrt in den Fokus wissenschaftlicher Arbeiten in Bereichen von der Krebsforschung bis hin zur synthetischen Biologie rücken. Das entwickeln eines besseren Verständnisses von ARE Motifen die tatsächlich von AUBPs gebunden werden sowie im Folgenden die Vorhersage neuer durch AUBP regulierter mRNAs mit Hilfe bioinformatischer Methoden waren Ziele dieser Diplomarbeit.

Zu diesem Zwecke wurde eine Datenbank generiert, die Transkripte von Maus und Mensch mit von uns annotierten ARE Motifen beinhaltet. Des weiteren enthält die Datenbank Informationen über die Zugänglichkeit und Überrepresentation dieser Motife, sowie phylogenetische Information, welche Allesamt durch das Anwenden von *in silico* Methoden wie RNA Faltung, der Erstellung multipler Alignments und der Analyse der Nukleotidanreicherung mittels Markov Modellen der Ordnung 0 und 1, erzeugt wurden. Eine zweite Datenbank die Literatur bezueglich der Effekte von AUPBs auf experimentell untersuchte Zielen dieser Proteine enthält wurde ebenfalls aufgebaut. Beide Datenbanken wurden anschliessend vereint um einen Webserver namens 'ARE-site' herzustellen.
Dieser Webserver wurde bereits publiziert (1) und ist ein frei zugängliches Werkzeug zur Untersuchung von bekannten sowie der Vorhersage bisher unbekannter mRNAs unter AUPB Regulation.

Die Analyse der über den Webserver verfügbaren Information zur Vorhersage neuer AUPB Ziele war Teil dieser Diplomarbeit und die Resultate dieser Analyse sind im Bereich 5 beschrieben. Der Webserver 'AREsite' ist unter der Adresse `http://rna.tbi.univie.ac.at/AREsite` erreichbar und wurde bereits erfolgreich zur Analyse von AUPB Zielen herangezogen (2).

Contents

List of Figures

1 Introduction

This thesis is focused on AU-rich elements (ARE) and their interaction probabilities with AU-rich binding proteins (AUBP). AU-rich elements are RNA sequence motifs rich in adenine and uracile. They were identified during the 1980's as key players for the stability of certain mRNAs. Located in the 3' untranslated region of these mRNAs, they allow various AUBPs to bind and interact with them. Depending on the kind of AUBP and where in the cell these interactions occur, they can have stabilizing or destabilizing influence on the mRNA. This provides the cell with a fast response mechanism to changes in it's environment, by influencing the gene expression levels via direct interaction between AUBPs and their target mRNAs.

1.1 Subjects of this thesis

The aim of this thesis was the annotation of ARE motifs in the available transcriptome of human and mouse, and the analysis of their role for mRNA stability on transcript and genomic level, producing information about conservation, accessibility and over-representation of ARE motifs in the annotated transcripts.

Evolutionary conservation is always a hint for functionality, as something of functional importance is often being reused and conserved during evolution. Accessibility of a motif is a measure for functionality, as only accessible RNA motifs allow interaction with proteins, although accessibility can be changed by interaction of RNA molecules with each other or proteins for example, and a functional motif does not necessarily have to be accessible in its native state. Over-representation can act as a measure of motif function and importance for regulation, as the ARE core motif 'AUUUA' can sometimes be found in a 3'UTR by pure chance. On the other hand functional motifs should occur in a higher number as expected by chance, given the according 3' UTR sequence.

To analyze above mentioned criteria, a database has been created, which contains all annotated transcripts of human and mouse as well as those of their known one to one orthologs. Latter transcripts and further data were used to analyze phylogenetic conservation of annotated motifs in human and mouse

on transcript and genomic level via the generation of multiple alignments. An analysis of the accessibility in terms of opening energy of annotated ARE motifs was conducted by the application of RNA folding algorithms. To get information about over-represented ARE motifs, annotated 3' UTRs were analyzed using an order-0 and an order-1 Markov model, that returns information on the fold-enrichment of analyzed motifs. Results of those analysis steps were used to filter for transcripts that are predicted to be under strong regulation of ARE motifs. These results were compared to the results of an extensive expert literature search for known AUBP targets and are discussed as part of this thesis. Finally a combination of accessibility and over-representation was used as method for the prediction of novel AUBP targets.

A detailed overview on DNA and RNA, RNA structure and sequence motifs and RNA binding proteins and their role inside cells as well as a section on RNA bioinformatics and the tools that have been used during this thesis is following this short introduction.

2 Background

2.1 RNA synthesis and regulation and the role of RNA binding proteins

2.1.1 Nucleic acids, Transcription and Translation

DNA and RNA Live is based on the information stored at the level of nucleic acids. There are two kinds of these acids available in known live forms, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). For many years now, it has been known that nucleic acids are used to store information in form of genes, which are passed on from generation to generation via reproduction. During evolution a lot of this information has been and will be altered via insertions, multiplications, random or directed mutations as well as transpositions. Some information has been lost due to deletion events. These alterations happen by pure chance or can be seen as response of organisms to their environment. Changes may have no effect at all but changing the wrong part of information may turn out to be lethal. Very complex regulating steps are required to get things under control and to make it possible for a live form to evolve and to survive.
This regulating processes can occur on different stages and layers during the process of gene expression. One of these processes is the regulation of mRNA stability through the interaction of AU-rich binding proteins with AU-rich elements in the 3' UTR of these mRNAs, which is the main topic of this thesis. However, to get a deeper understanding of the regulatory influence of this interactions, this section begins with background information on nucleic acids and the events during gene expression and ends with a detailed description of AU-rich elements, their binding proteins and their regulatory role.
For almost all live forms, DNA provides the genetic information, but in some viruses RNA is used as genetic material. Even though both nucleic acids work as storage for genetic information and have a lot of similarities, there are also a lot of differences between them. One difference can be found in the bases, contained in the nucleic acids. For DNA these are the purines adenine and guanine as well as the pyrimidines cytosine and thymine. In RNA we can find adenine, guanine and cytosine as well, but instead of thymine we find the base uracil which is chemically almost identical to thymine, but lacks a methyl-group at the C5 position, see figure 1.

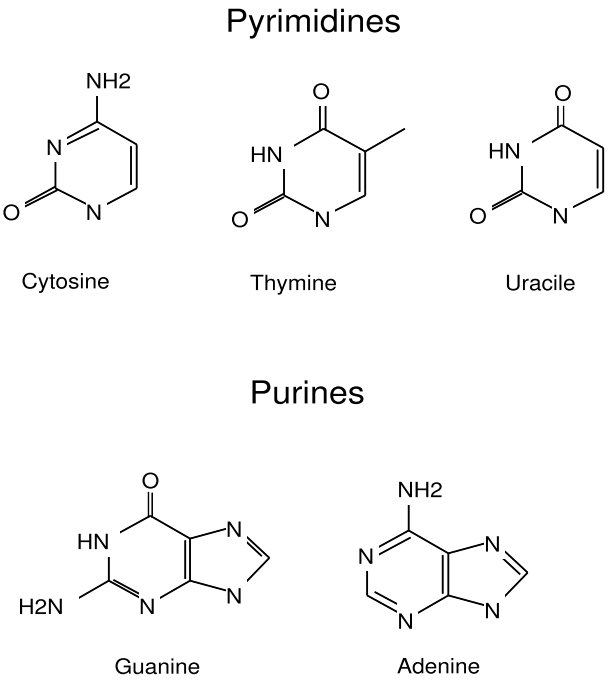


Fig. 1. The nucleobases of RNA and DNA. Whereas DNA has thymine, one can find uracile in RNA.

The use of thymine in DNA is part of a mechanism protecting it from uncontrolled deamination (3; 4). Via deamination, cytosine is metabolized to uracile. As Uracile is usually not found in DNA, it can be used as marker for DNA repair mechanisms. Once uracile is found in DNA, it is immediately exchanged with thymine by cell intern DNA repair mechanisms, restoring the DNA chain to its prior form. This is not possible in RNA as uracile can be found all over the nucleic acid sequence. However, DNA is mostly used as genetic storage material and has a much longer half-life than RNA. Whereas DNA is reproduced and proliferated via semi-conservative replication, a process where one DNA strands acts as template for DNA polymerases to produce the according anti-sense strand, RNA is produced via the process of transcription, which will be explained in more detailed later on. DNA acts as a template during this process, so mutations in DNA can have a more extensive effect on the cell than mutations in RNA.

Similarities between RNA and DNA can be found by taking a closer look at their building blocks and the structures they form inside the cell (3; 4). The building blocks of both nucleic acids are nucleotides. Purines or a Pyrimidines connected to a sugar are termed nucleosides and become nucleotides when bound together via phosphate ester bonds, see figure 2. In a more detailed view the N9 atom

of a purine or the N1 atom of a pyrimidine are connected via a β -glycosidic bond to the C1 atom of the corresponding ribose or deoxyribose to build up a nucleoside. If connected to one or more phosphate groups, the molecule is called nucleotide and acts as building block for a nucleic acid. To produce a nucleic acid, these blocks are connected to each other, whereby the 5' hydroxy end is bound to a phosphate group, and the 3' OH-end is bound to the phosphate group of the next nucleotide via phosphodiesterbonds. This leads to a polarity along the nucleic acid, from the 5' phosphate start to the 3' hydroxy end. The phosphate groups between the nucleosides are negatively charged and provide the whole DNA or RNA backbone with a negative charge which plays an important role in the interaction of binding proteins with their nucleic acid target, as discussed later. Fitting their names, DNA has a deoxyribose - phosphate backbone and RNA a ribose - phosphate backbone, see figure 3.
Due to the lack of an oxy-group on the 2' carbon atom in deoxyribose, DNA is better protected against nucleophilic attacks than RNA.

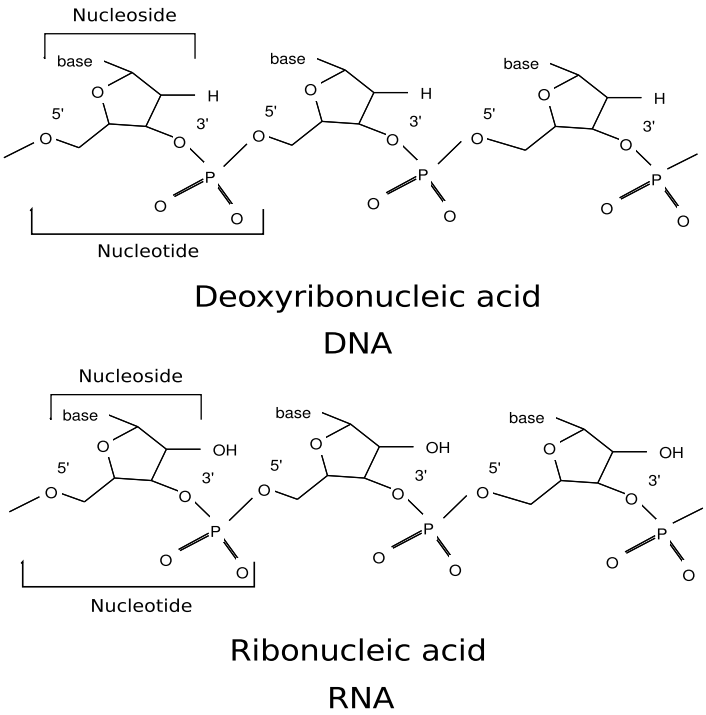


Fig. 2. A nucleoside consists of purine or pyrimidine base, connected via a β -glycosidic bond to the corresponding ribose or deoxyribose.
A nucleotide is a nucleoside bond to one or more phosphate groups via ester bonds.
A nucleic acid is build up by nucleotides connected via 3' – 5' – phosphodiesterbonds.

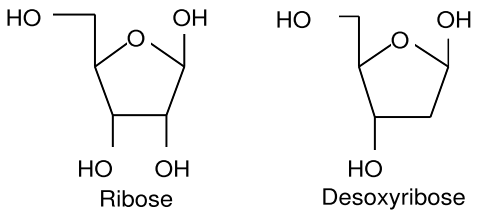


Fig. 3. Ribose is used to build up RNA nucleotides, Deoxyribose is used to build up DNA nucleotides. One can see the missing -OH group in Deoxyribose, giving it its name.

Ribose and Deoxyribose have a slight structural twist and are therefore not planar and called sugar - pucker. Depending on the kind of pucker structure (C2-endo, or C3-endo, see figure 4), a nucleic acid can take different helical shapes (5).

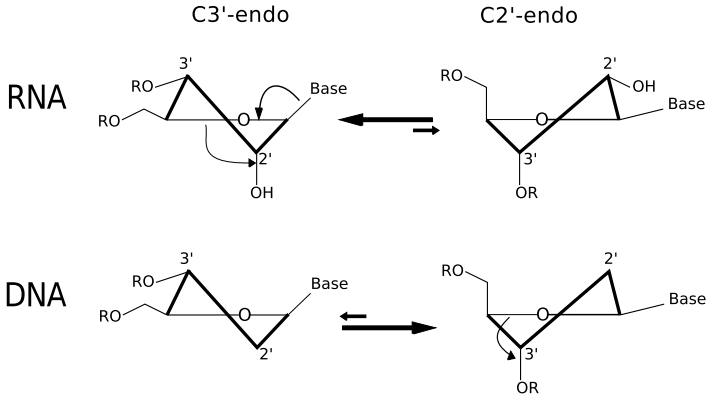


Fig. 4. A nucleic acid can take different helical shapes depending on the sugar-pucker structure (C2-endo, or C3-endo) of Ribose and Deoxyribose.

DNA as well as RNA can be found to take up diverse structures inside cells, as nucleic acids in aqueous solution are under a pressure to form energetically favorable structures.

Thanks to Watson and Crick and based on the work of Pauling and Corey we know that DNA forms a double helical structure, the B-form helix (6), see figure 5. Bases are arranged to each other with a distance of 0.34nm, to perform a full 360° turn after 10 bases or 3.4nm. Two grooves, one called minor and the bigger one called major groove exist in this B-form helix, allowing base-specific interactions with proteins.

RNA forms a more dense structured A-form helix with a deep and narrow groove which is not as easily accessible by proteins (7), see figure 5.

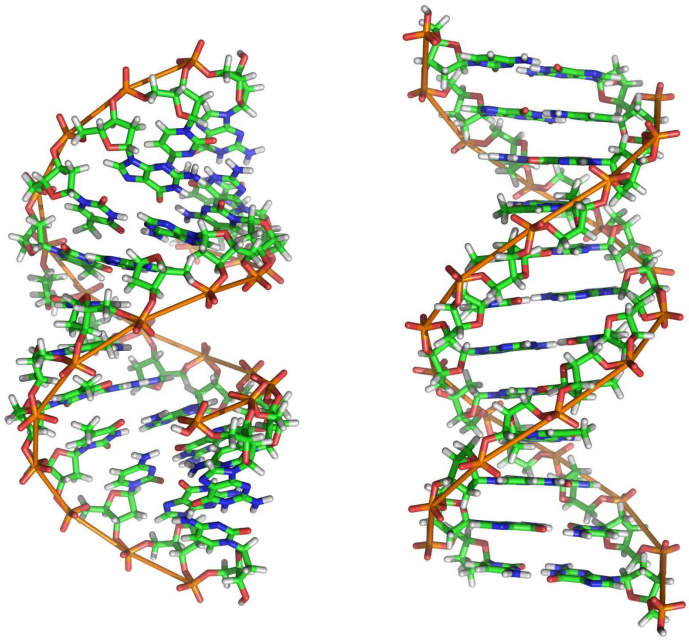


Fig. 5. From left to right, this figure shows the typical RNA A-form helix and the typical DNA B-from helix. Figure adopted from `http://upload.wikimedia.org/wikipedia/commons/b/b1/A-DNA\%2C_B-DNA_and_Z-DNA.png`

The double helical form and the almost perfect stacking of bases inside the DNA or RNA helix make it possible for the weak Van-der-Waals forces to act between bases and additively contribute with a stabilizing effect. Stacking allows electrons in the aromatic heterocyclic rings of the bases to interact and stabilize the structure. The arrangement of hydrophobic pyrimidine or purine bases on the inside of the helix and the hydrophilic groups on the outside add a hydrophobic effect. It occurs inside the helix and is further stabilizing the structure. Hydrogen bonds between the Watson-Crick nucleotide pairs A:T as well as G:C (see figure 6) stabilize the double helix too (8). This bonding effects are true for RNA helices as well, but the C2 - endo sugar - pucker makes the RNA helical structure more dense. In difference to DNA helices which usually form along the whole DNA strand length of two DNA molecules, RNA helices often occur along short complementary regions of RNA molecules, and are mostly generated intramolecular. In eukaryotes double stranded RNA is very unusual and recognized as pathogen RNA and degraded, an effect that is extensively used in molecular biology and known as knockdown or RNA-interference (9) and its discoverers have even been awarded with a Nobel prize.

RNA is able to form versatile other structural motifs besides helices. These motifs and their function will be addressed later on.

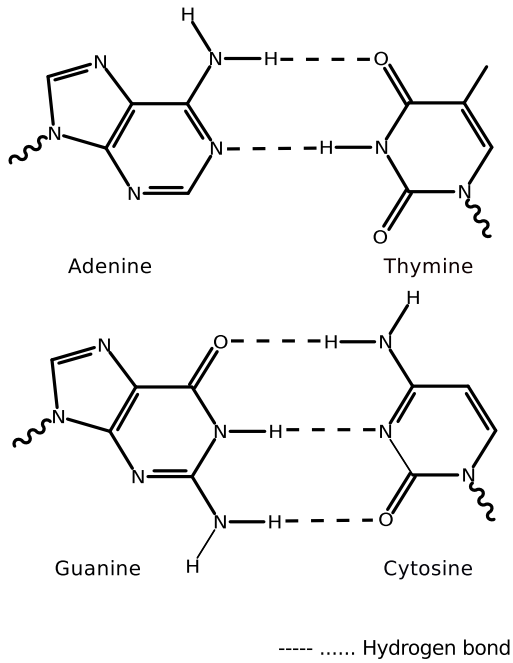


Fig. 6. Hydrogen bonds between adenine and thymine or guanine and cytosine stabilize the double helical structure of DNA. As a G-C basepair is bond via three hydrogen bonds instead of two for A-T basepairs, G-C contributes with a stronger stabilizing effect.

Beside Watson-Crick hydrogen bonds, DNA but mostly RNA structures can be further stabilized by non - Watson-Crick base pairs. These non-canonical base pairs form RNA structure motifs and play a role in RNA-RNA interaction and the generation of 3D structures. Among the non - Watson-Crick base pairs, we can find wobble base pairs, the most prominent one is G:U, see figure 7.

The so called wobble position plays an important role in codon - anticodon recognition during the process of translation. A codon is defined as a triplet of DNA nucleotides that codes for a certain amino acid, and the anticodon is the region of tRNA (transfer RNA) that recognizes this triplet. The role of tRNA for translation will be explained later on, but to understand the necessity of wobble bases one has to know about codons. It is not necessary for all three nucleotides to exactly match the nucleotides of the anticodon to allow bonding. The third position of the codon can alter and was termed wobble position.

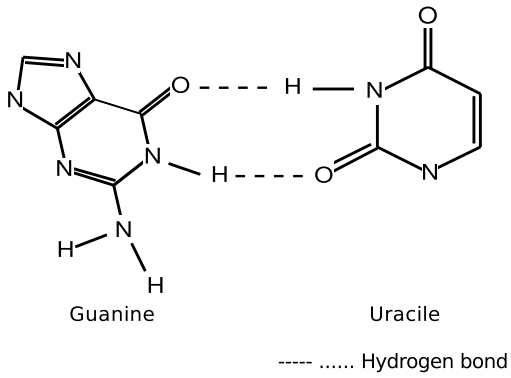


Fig. 7. The G:U Wobble base pair.

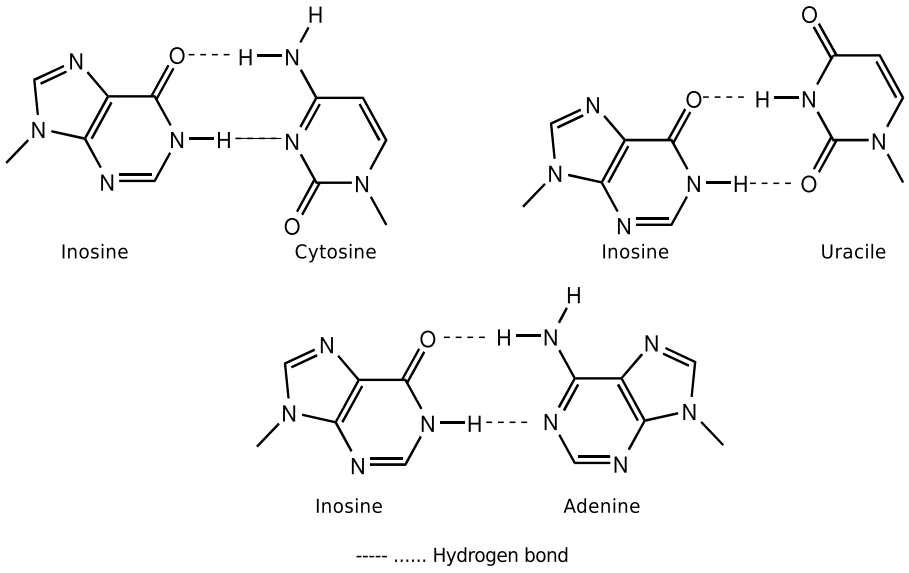


Fig. 8. The Inosine Wobble base pairs are of special interest in tRNA.

As the same counts for all other forms of base pairing, the wobble base pairs enable a molecule of DNA or RNA to introduce bonds even if the corresponding partner nucleotide does not match exactly.
Of special interest in RNA are the wobble pairs with Inosine, see figure 8. Inosine is a nucleobase derived from adenine via the post-transcriptional enzymatic process A-to-I-editing (for more information on this topic refer for example to (10)), and has a series of possible pairing partners. It can pair with adenine, uracile and cytosine and is an essential part of tRNA anticodons, enabling tRNA to create wobble pairs (11).
Furthermore we can find Hoogsteen and sugar-edge base pairs, allowing even more than two nucleotides to interact with each other via non-canonical base pairing, see figure 9 . Hoogsteen as well as sugar edge base pairs form hydrogen bonds between their according edges of purines and pyrimidines and take part in helix stabilization as well as tertiary and quaternary structures of RNA (12).
Changes in sequence can occur without influencing the 3D structure of a nucleic acid if affected bases are replaced via compensatory mutations. This effect is based on the isostericity of Watson-Crick and non-canonical base pairs. Single base mutations as well as mutations concerning a base pair can be compensated if the affected base pair is replaced by an isosteric base pair, which has to display similar distances between the bases and a similar orientation of its glycosidic bond.

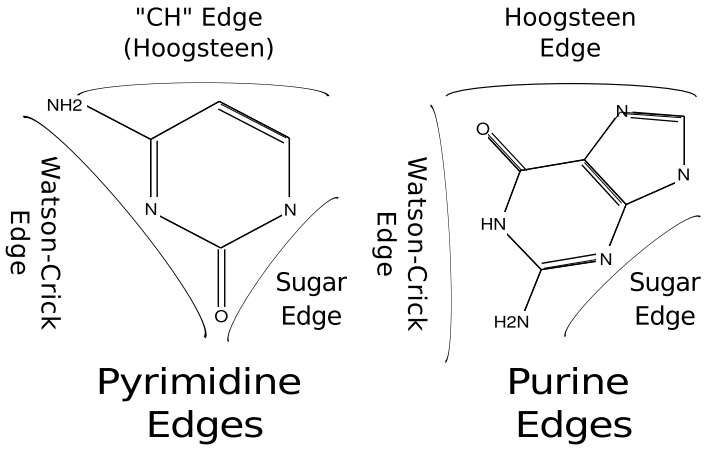


Fig. 9. The interacting edges of purines and pyrimidines. The Watson - Crick edge, the Sugar edge and the CH - edge of pyrimidines which is similar to the Hoogsteen edge in purines.

An interplay between some proteins and DNA and/or RNA has been shown. For example, DNA is packed by interaction with histones, structural proteins that can bind to DNA and allow it to wrap around them, forming chromatin.

This leads to the formation of two types of chromatin. First, euchromatin, which consists of active DNA, meaning it is a region of active transcription. And second, heterochromatin, which is densely packed and not accessible for the transcription machinery. A lot of modifications have been discovered that regulate the active state of DNA, for example methylation of chromatin, making DNA accessible or condensing its structure. By introducing modifications on chromatin, the cell can actively regulate the level of gene expression. The term "gene expression" contains all processes that lead from DNA to the final gene product in defined amount and activity. The interaction with histones allows DNA to take a very densely packed structure inside the nucleus of a cell, the chromosome. This dense packing of DNA allows the simultaneous regulation of gene expression for a group of genes at once by packing or unpacking the strand of DNA where they are located, thereby restricting or relieving the access of the transcription machinery to the regulated genes as a function of the metabolic state of an organism for example.
Summing up, DNA is due to its double helical structure and the lack of an oxy - group, a stable and chemically well suited molecule to store genetic information. Through histones, it can become highly condensed and modifications of histones provide the cell with a regulatory mechanism for gene expression. Therefore it may have been established as genetic storage material instead of RNA in most organisms (13). The current hypothesis is that DNA has replaced RNA as information storage over time, as some hints to a former RNA world exist, where RNA molecules built the base of life without the help of DNA or proteins, see (14).
However, information is not found everywhere along a DNA strand. DNA is divided in different sections with varying function. The current consensus in the community is that a section of DNA coding for a functional product can be seen as gene (15). Eukaryotic genes are divided in intronic and exonic regions. Exons are the sections containing information for the final product. Introns define the boundaries of exonic regions. They can not be found in the mature RNA product, as they become spliced out during RNA processing, a topic that will be discussed later, but can lead to different isoforms of RNA, highlighting its importance. The region of a gene that codes for one amino acid is called codon, as mentioned before.
As amino acids are the building blocks of proteins and influence protein - RNA interaction, a short insertion introducing amino acids follows. Today 22 amino acids are known to act as building blocks for proteins, see figure 10. The generic formula for an amino acid is $H_2NCHRCOOH$, where R is replaced by

an organic substituent, depending on the type of the amino acid.

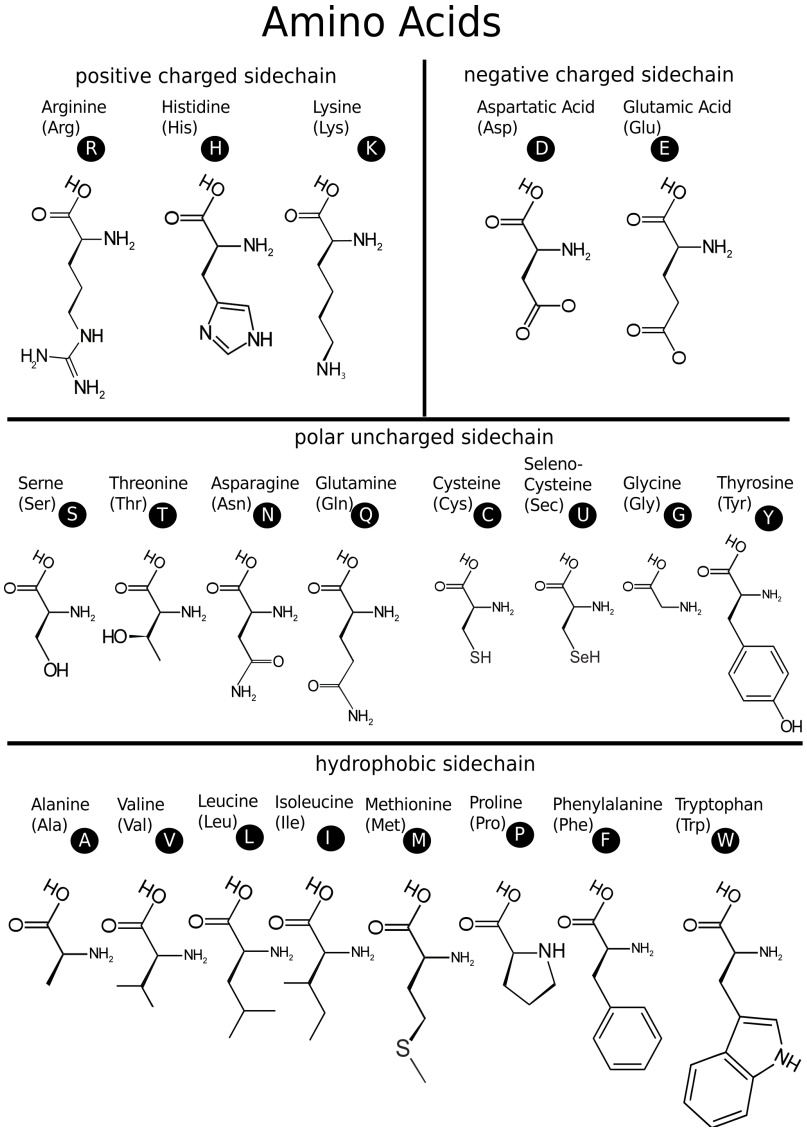


Fig. 10. The 21 common amino acids, figure adopted from:"`http://en.wikipedia.org/wiki/File:Amino_Acids.svg`.

Amino acids play a crucial role in the metabolism and as building blocks for proteins and the amino acid composition of a protein directly affects the interplay with nucleic acids. As mentioned, nucleic acids have a negatively charged phosphate backbone, which allows proteins to interact according to their polarity and charge. Jones et al. (16) presents a full list of amino acid propensities in RNA, ssDNA and dsDNA binding proteins. This bias is strong enough to allow prediction of RNA binding probabilities from amino-acid sequence alone (17). RNA binding proteins will be discussed in detail later. For a more complete picture of amino acids and their function see for example (18), this thesis now

returns to the discussion of the sections in DNA.
In addition to introns and exons, an eukaryotic gene (see figure 11) can contain regulatory elements. Latter can for example interact with the transcription machinery and play a role in gene expression regulation.
A region, shared by almost all genes in all kingdoms, is called promoter and can contain elements like the TATA box, which can be found in about 24% of human genes, or GC-rich and other elements (19). These elements are recognized by RNA - polymerases or other interacting molecules like transcription factors and are therefore essential for transcription. Especially in eukaryotes one can find a lot of other regulatory regions, like upstream or downstream activators or repressors of gene expression. The whole machinery of gene expression regulation is far from being fully understood, but a lot of information is already accessible, providing us with an idea of its complexity.

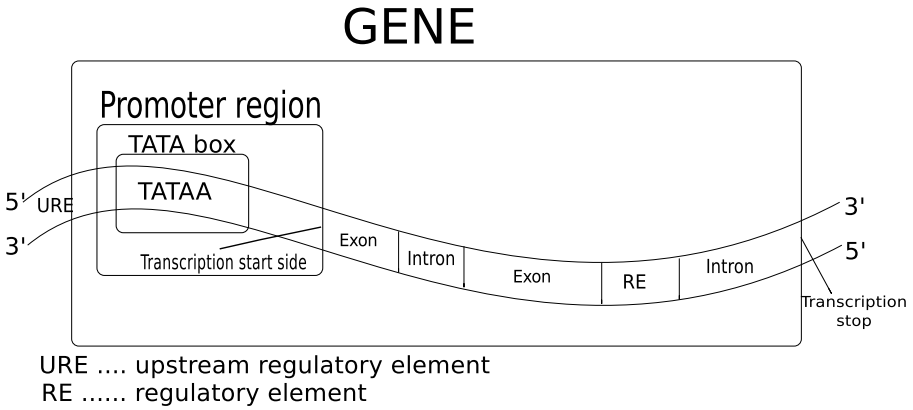


Fig. 11. Selected sections of a typically eucaryotic gene. Upstream regulatory elements are found upstream of the promoter region. The promoter region can contain a TATA - box. The transcription start site marks the first nucleotide which is transcribed into RNA, followed by introns, and exons which later form the coding sequence. Regulatory elements or enhancers can be found upstream of the promoter region or downstream, which means inside the gene sequence.

Decades ago the central dogma of molecular biology: "DNA makes RNA makes protein" was proclaimed by F. Crick (8) and published in 1970 (20). While the core of this dogma, the unidirectional flow of information from DNA to protein, remains almost undisputed (21), the role of RNA as simple intermediate has changed completely. Today it is known that there is much more behind nucleic acids, proteins and transcriptional and translational processes than what was thought in the beginning. More information is going to be revealed using new techniques like high throughput sequencing and it can be expected that additional layers of gene expression will be uncovered and existing models are going to be verified or new models will arise.
The topic of this thesis is the interaction of RNA binding proteins (with focus on AU-rich binding proteins) and RNA (here mRNA). The next sections will discuss the cellular processes that lead to the generation of RNA and proteins in eukaryotic cells, as it is necessary to understand how these molecules are synthesized to discuss the role of their interplay for an organism.

Transcription A cell depends on proteins to maintain its own function. To produce these proteins, a cell has to perform several, highly regulated steps. The first among them is transcription of DNA into RNA via enzymes like RNA - Polymerase. This process occurs in eukaryotes as well as prokaryotes. Topoisomerases, several initiation factors, elongation factors and more are required to initialize and control it. Even subgroups of the mentioned kingdoms do differ in type and amount of factors involved. However, the product of transcription is in all cases RNA.
The focus of this section is on RNA polymerase II transcription in the nucleus of eukaryotes, which produces mRNA that is both, a template for proteins and a target of the RNA binding proteins discussed here.
Whereas one can only find one polymerase for all transcriptional products in prokaryotes, three RNA polymerases exist in eukaryotes with RNA polymerase II as the key player in mRNA transcription, see table 1.

Name	Found in	Product
RNA Polymerase I (Pol I, Pol A)	nucleolus	47S precursor of rRNA subunits 28S, 18S, 5.8S
RNA Polymerase II (Pol II, Pol B)	nucleus	mRNA, miRNA, snRNA and some ncRNAs
RNA Polymerase III (Pol III, Pol C)	nucleus	tRNA and miRNAs snRNAs, 5S rRNA subunit and several repeated sequences (e.g. Alu elements)

Tab. 1. The three RNA polymerases as found in eukaryotes.

The discovery of RNA polymerase in the 1950's provided insights into the first steps of gene expression, a topic which is still not fully understood (22).
Transcription in eucaryotes can be divided into five steps according to (23; 24; 25).

- Pre - initiation: Binding of transcription factor subunits, e.g. TBF(TATA-Binding Protein) to a core promoter region of DNA (30, 75 and 90 base pairs upstream from the transcription start site TSS), in the case of TBF the TATA box, which is found 25 to 30 base pairs upstream from the TSS. Recruitment of a series of other transcription factors and RNA polymerase, forming the pre-initiation complex.
- Initiation: After the pre-initiation complex has attached to the promoter, the carboxy-terminal region of RNA polymerase II becomes phosphorylated. Together with recruitment of activating and repressing factors as well as DNA helicases (that unwind the DNA double helix) transcription starts.
- Promoter clearance: After the first RNA nucleotides are produced and due to phosphorylation of serine 5 in the carboxy-terminal region of RNA polymerase II by transcription factor IIH, resulting in conformational changes of this highly flexible part of the polymerase, the promoter is cleared, giving other molecules of polymerase II the possibility to bind. Release of the RNA transcript during clearance results in truncated transcripts and is called abortive initiation, known in eukaryotes and prokaryotes.
- Elongation: RNA polymerase slides along the template strand of DNA in 3' end to 5' end direction, producing more and more RNA product from 5' end to 3' end, which results in a RNA copy of the DNA coding strand. This step can be done by more than one polymerase simultaneously, giving the cell the possibility to produce vast amounts of RNA in a short

time. RNA polymerase II has proofreading activity, which prevents the synthesizing of RNA strands that do not match the DNA coding strand.

- Termination: After polyadenylation and several other mRNA processing steps occurred, RNA polymerase is released from the DNA template. Thereby releasing a product called pre-mRNA, which undergoes further processing steps to become a so called mature mRNA.

Figure 12 shows a schematic view on transcription in eukaryotes.

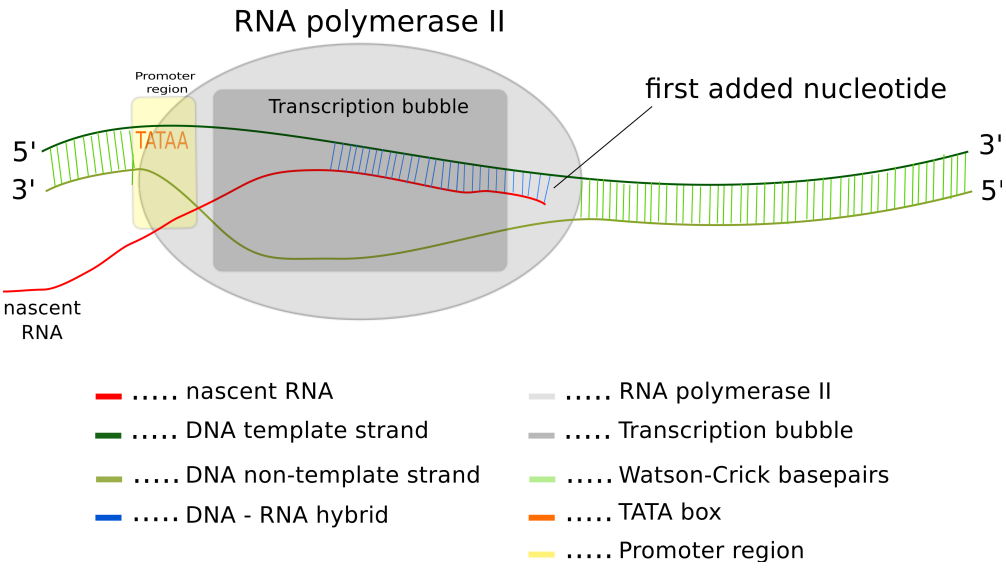


Fig. 12. A simplified model of transcription, showing RNA - polymerase II docking at the promoter region and synthesizing a strand of RNA, with a view inside the transcription bubble.

Several mRNA processing steps happen in parallel to transcription. These so called co-transcriptional processing steps directly influence the speed and effectivity of RNA synthesis (26; 27), as can be seen in figure 13. Among the most important steps are the 5' capping, splicing and the 3' poly-adenylation, as these modifications are necessary for mRNA stabilization and export of the mRNA from the nucleus in the cytosol where translation of the mRNA takes place. In the following, this processing steps will be explained in more detail, as described by (26; 28; 29).

- 5' capping: After RNA polymerase II has transcribed the first 25-30 nucleotides, RNA-triphosphatase removes the γ-phosphate from the 5' end of the nascent RNA strand. Guanylyl-transferase then transfers a guanosine-monophosphate derived from guanosine-triphosphate to form a GpppN end, where N stands for the former 5'-end nucleotide of the RNA. 7-methyl-transferase then methylates the guanine at position 7 of its purine ring. A capping enzyme fulfills the roles of the RNA-triphosphatase and the guanylyl-transferase in mammals. The 5' cap is important for the stability and translation of the mRNA.
- Splicing: Removing the introns and joining exons in a pre-mRNA is termed splicing. In human and yeast a protein complex consisting of the U1, U2, U4, U5 and U6 small nuclear Ribonucleoproteins (snRNPs), the spliceosome, together with a large number of additional proteins catalyzes this step. Two trans-esterification reactions are necessary to join two exons. The first forms a lariat intermediate resulting from a nucleolytic attack of the 2'-OH of a branch point nucleotide on the first nucleotide of the adjacent intron. The second reaction is conducted by a nucleolytic attack between the 3'-OH from the free exon on the last nucleotide of the adjacent intron. Thereby joining the exons and splicing out the intron.
  This process can occur on different sites and provides the cell with a mechanism termed alternative splicing. Through alternative splicing, different isoforms can be generated out of a single transcript, increasing the bandwidth of possible transcription products and adding another layer of gene expression regulation to the repertoire of a cell. As this step has to be under strict regulation, a vast number of enhancers, silencers and other regulatory factors exist.
- 3' poly-adenylation: Endonucleolytic cleavage is the last step of transcription. It occurs 1030 nucleotides downstream of the poly-A signal sequence, a conserved AAUAAA sequence motif in mammals. In mam-

mals cleavage/polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), and two cleavage factors (CFIm and CFIIm) are necessary to conduct this step. This step is followed by poly(A) addition via Poly(A)polymerase (PAP) at the cleaved the 3' end. The poly(A) tail is important for the stability and export, as well as translation of the mRNA.

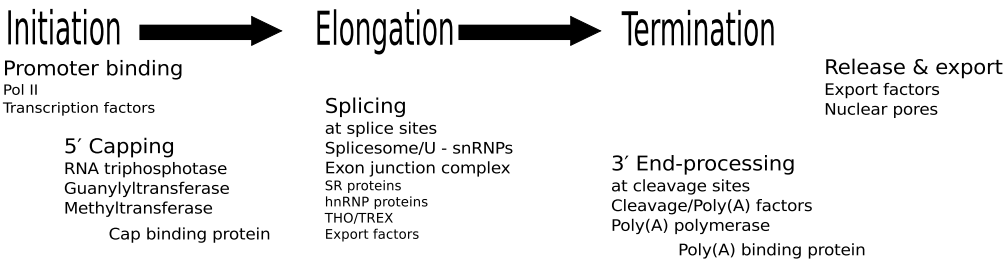


Fig. 13. This figure shows which co-transcriptional steps occur during transcription, at which step they happen and some of the involved factors according to (26).

The product of transcription and processing of a protein coding gene is a mature mRNA molecule, which can be functionally grouped into three regions, see figure 14. From the 5' to 3' end this are the 5' untranslated region (5'UTR), followed by the coding sequence (CDS) which is the sequence where the actual protein coding information is stored, and then the 3' untranslated region (3'UTR).
The mature CDS of a mRNA is also known as open reading frame (ORF). In a mature eukaryotic mRNA, ORF and CDS can be used as synonyms, as they both contain the whole protein coding information. An ORF begins with the start codon (in eukaryotes, AUG, ACG or CUG, which are coding for the amino acid methionine are always the start codons for translation) and ends at the first stop codon (UAA, UAG or UGA are codons where the no tRNA provides fitting anti-codons and therefore stops translation) found downstream.
Untranslated regions, as their name implies, are not translated. They play a major role for mRNA stability and influence the machinery for transcription and translation. 3' UTR's, their sequence motifs and consequences for mRNAs containing these motifs are the central topic of this thesis and will be discussed in more detail later on.

However, not every product of transcription leads to an RNA that codes for a protein. Those that do not are therefore termed non-coding RNAs (ncRNAs). These ncRNAs were first found in bakers yeast in the 1950's, where alanine

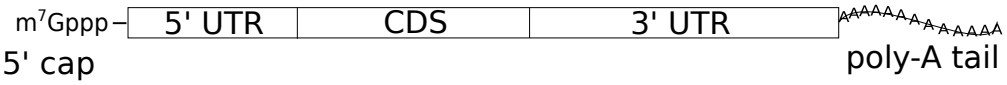


Fig. 14. A typical mature mRNA, with the 5'-cap, the 5'-UTR, the CDS, the 3'-UTR and the poly-A tail.

tRNA was found (30). Since then, more and more types of ncRNAs and their functions have been revealed.

Among the largest projects addressing functional elements in the human genome is the Encyclopedia of DNA Elements (ENCODE) pilot project (31) and the GENCODE project (32). Funded by the National Human Genome Research Institute (NHGRI), its goal is to build a full list of functional elements in the genome of human, and recently, mouse. This list includes elements that act at the protein and RNA levels, as well as elements that regulate the active state of genes. First results of the ENCODE pilot-project, addressing 1% of the human genome, showed that the majority of bases can be associated with transcripts (31). Other projects following the same goals have been funded, an example are the Functional Annotation of the Mammalian Genome (FANTOM) projects, projects I to IV already completed and project V in planning, for more information see (33).

Together with the rising number of identified ncRNAs we begin to get more and more insights in the complexity of regulation of gene expression. As mentioned before very complex regulatory steps are a must for cells and organisms to survive in the highly competitive fight for survival and evolution. Each cell has unique requirements, depending on its environment like neighboring cells or other organisms in close proximity, as well as selective pressure like nutrient shortage and more.

Thus it is no surprise that numerous layers of regulating steps evolved over time, making it possible, to get fast response to changes in the need of proteins, enzymes and other factors. For a long time it was thought that regulation of these steps mainly depends on functional protein products. Today it is known that a lot of ncRNAs play a functional role as well, if not even the major role. RNAs that form highly structured ribonucleic acid chains and have catalytic functions are called ribozymes. Ribosomes are one example for ribozymes, consisting of RNA - protein complexes, where the RNA part is catalyzing the peptidyl transferase activity which forms the amino acid chains (34).

Messenger or mRNA, transcript or tRNA, and ribosomal or rRNA and the right enzymes and factors are required to start the next step on the way to proteins, translation.

Translation During this process mRNA works as a template for the generation of amino acid chains that can then fold to fully functional proteins either by themselves or with the help of other proteins and enzymes. A tRNA is used as a carrier to transport the respective amino acid to the place of its need and rRNA is coding for ribosomal protein subunits which form ribosomes, the actual protein factories of the cell.
The ribosome's role in translation can be compared to the role of RNA-polymerase in transcription. Simplified, reading the information on the nucleic acid template and producing a chain of product, ribosomes bind to mRNA, recruit tRNAs loaded with amino acids and connect the latter, releasing an amino acid chain which can then fold into a functional protein product.
According to (25; 35; 36) the process of translation can be divided into four steps:

- Activation: In this step a tRNA covalently binds to the correct amino acid via an ester bond between the carboxyl group of the amino acid and the 3'OH end of the tRNA.
- Initiation: The start site for translation is always the codon for methionine, AUG or alternatives like ACG or CUG. So methionine is the first amino acid of the newly synthesized poly-peptide product. The 5' cap of eukaryotic mRNA plays a major role in connection of the 40S ribosomal subunit to the mRNA template in vivo. How exactly the 40S subunit finds a start codon and initiates translation is still unclear. Several initiation- and other- factors are needed to get translation up and running.
- Elongation: As the ribosome slides along the mRNA template, amino acids are added to the growing peptide chain, until the ribosome faces a stop codon like UAA, UAG or UGA in the nucleic acid chain.
- Termination: As no tRNA is able to bind to one of these stop codons, the ribosome pauses. This is followed by binding of a release factor that helps disassembling the ribosome - RNA complex.

Again a lot of regulatory steps occur during translation and give the cell a layer of regulation for gene expression. This starts with trans-acting factors, that can interact with the ribosome or the mRNA, to cis-acting regulatory elements or structures on the mRNA itself and, of course, the number of ribosomes that have access to a certain mRNA. Figure 15 shows a schematic view on the process of translation in eukaryotes.

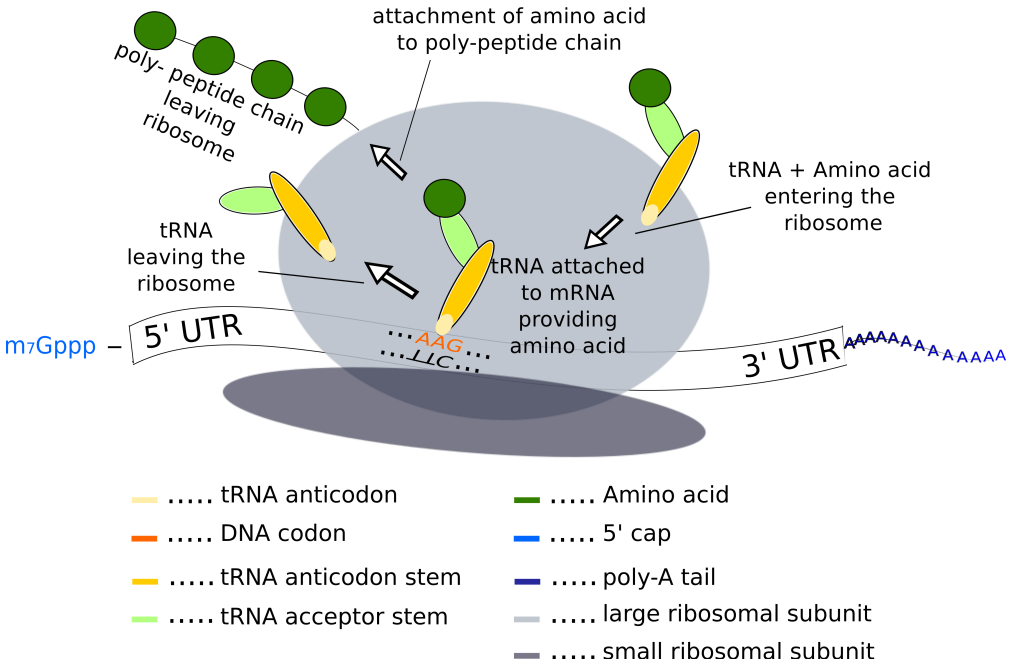


Fig. 15. A simplified view on a charged tRNA, entering the ribosome and binding to a specific position on the mRNA. The ribosome attaches the amino acid to the poly-peptide chain and releases the unloaded tRNA.

After finishing translation, the poly-peptide product can undergo a lot of modifications (phosphorylation, glycosylation, acetylation, see for example (37)) to become the protein of need, be it a membrane protein or a part of a multi-protein enzyme complex. An important modification for RNA binding proteins is for example the methylation of Arginine, which effects the ability of a protein to recognize other proteins or RNA (38)

2.1.2 RNA types and structures

Throughout the cell a lot of different types of RNA exist, each with different functions and often with a well-defined structure. The following section will show what types of RNA are known in eukaryotic cells and to what purpose they are produced. So far three types of RNA: mRNA, tRNA and rRNA have been mentioned. A lot of other non-coding RNAs have been discovered during the last decades, more to be revealed, see for example reference (39) and table 2.

As versatile as the types of RNA which can be found in a cell may be, the structural possibilities for a single molecule are immense. Nevertheless, many RNAs, like for example tRNA have a typical, characteristic structure, see figure 16.

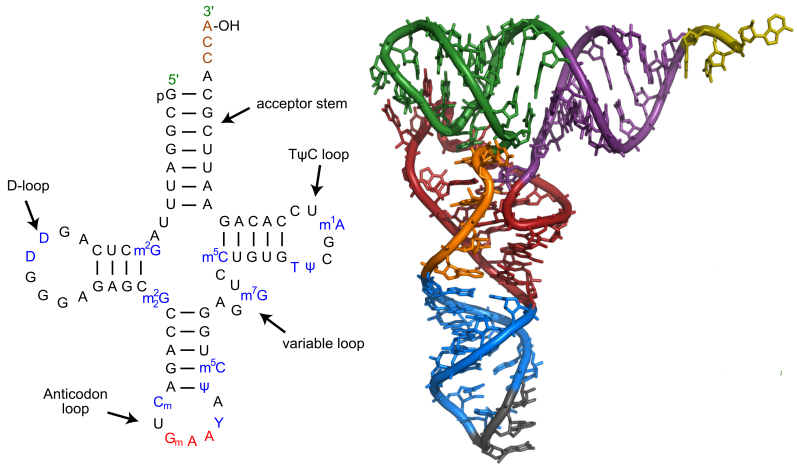


Fig. 16. The typical secondary and tertiary structure of tRNA, figures adopted from:`http://en.wikipedia.org/wiki/Transfer_RNA`.

To go further into detail, we have to distinguish between primary, secondary, tertiary and quaternary structures of RNA.
The **primary structure** is no real structure, instead it is defined by the nucleotide sequence of an RNA.

Secondary structures are formed by introducing intramolecular base pairs (described previously) into an RNA strand, following the laws of thermodynamics. Functional RNA molecules like tRNA usually have a characteristic structure, which is conserved among different species. Structure thus seems to be more important for function than the sequence of an RNA. To fold into

Type of RNA	**Found in**	**Length**	**Function**
tRNA	Eukaryotes, Prokaryotes, Archaea	∼ 70-110	Protein synthesis
rRNA	Eukaryotes, Prokaryotes, Archaea	∼ 120-4500	Protein synthesis
miRNAs	Eukaryotes	∼ 21-23	Regulation of gene expression
snRNAs (U1-U6)	Eukaryotes	∼ 110-120	Splicing of pre-mRNA Processing of rRNAs (U3)
snoRNA	Eukaryotes, Archaea	∼ 50-250	Processing and modification of rRNAs; Regulation of gene expression
Telomerase RNA	Eukaryotes	∼ 400	DNA synthesis at chromosomal ends
RNaseP	Eukaryotes, Prokaryotes, Archaea	∼ 400	Processing of tRNA
7SL RNA	Eukaryotes	∼ 300	Protein secretion
Xist RNA	Eukaryotes	17kb	X - chromosome inactivation
BC200 RNA	Eukaryotes (Primates)	∼ 200	Regulation of translation in dentrites ?
BC1 RNA	Eukaryotes (Rodents)	∼ 152	Regulation of translation in dentrites ?
MRP RNA	Eukaryotes	∼ 270	RNA processing
Lin-4 RNA	Eukaryotes	22	Regulation of translation of mRNAs
OxyS RNA	Prokaryotes	110	Regulation of translation of mRNAs
DsrA RNA	Prokaryotes	86	Regulation of translation of mRNAs
tmRNA	Prokaryotes	∼ 350	Degradation of shortened proteins
6S RNA	Prokaryotes	∼ 180	Regulation of transcription

Tab. 2. Types of RNA, their estimated length and their functions

its characteristic structure a molecule of RNA undergoes diverse steps of folding and unfolding until a structure with a sufficiently deep energy minimum is found. This can correlate with a maximization of base pairs, introducing a number of structural motifs into the RNA sequence. Secondary structure formation does interfere with RNA binding probabilities. If a secondary structure is found in a target region for RNA binding proteins, this structures may prevent the protein from binding, or render it possible. To predict sequence motifs like the ARE motifs as binding sites for RNA binding proteins, this has to be taken into account. Therefore it is necessary to calculate the accessibility of a sequence motif, as has been done for all ARE motifs (see section 2.1.4) that have been annotated during this thesis. Section 2.2 presents the approach that was used for accessibility prediction and section 3 includes more information on the motif annotation process. The influence of secondary structures on the action of AU-rich binding proteins is discussed in section 2.1.4 and has also been shown for miRNAs or snoRNAs that bind their targets in a sequence specific manner (40).
This paragraph describes common secondary structures that can be observed for RNA. As mentioned, base pairs stacking onto each other form the A-type double helical structure. Common RNA secondary structures motifs that are formed if some bases do not take part on Watson-Crick base pairs or if helices fork (see figure 17) include:

- Hairpin stemloops: They are of special interest for RNA secondary structures as they allow an RNA strand to bend back on itself and form an intramolecular duplex.
- Internal loops: Present motifs, that form where both RNA strands have unpaired bases followed by a region of complementarity.
- Bulges: Can form in regions of non-complementarity on one RNA strand.
- Junctions: Can act as intersections to link helices together.
- Multiloops: A section of nucleotides that contains multiple loop structures with an closing base pair.

RNA structure prediction is of importance as it often is the structure that defines the function of a molecule and not the sequence alone, e.g. codon-anticodon recognition by tRNA only works if tRNA is properly folded. This has in fact led to the design of various experimental approaches, for example a genome wide scan for RNA secondary structures in yeast, as done by Kertesz et al. (41). The bioinformatical prediction of RNA secondary structures is discussed in the

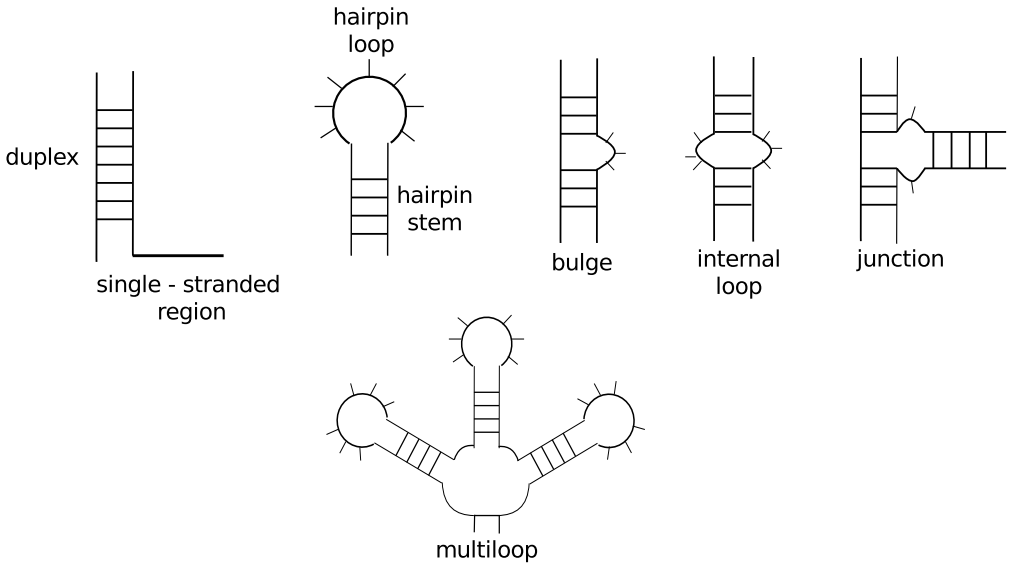


Fig. 17. Common RNA secondary structure motifs.

chapter RNA Bioinformatics.

Positively charged Metal ions like Mg^{2+} influence RNA secondary and tertiary structure formation. They help to minimize the distance between strands by antagonizing the negative charge along the phosphate backbone or even interact directly with the backbone (42).
Even more diverse is the field of RNA **tertiary structures**. Three main strategies are used to assemble these (43):

- Coaxial stacking: Helices are stacked on each other to give more stable helical structures.
- Hydrogen bonding: Whereas Watson-Crick bonds are usually found within RNA double helices, the previously discussed non-Watson-Crick base pairs are used to build up tertiary structures by creating base triples.
  Ribose zippers on the other hand, are interactions where the 2'-OH group of ribose is projected outside of RNA helices and can form hydrogen bonds with adjacent RNA strands. They can join together neighboring strands of RNA.
- Metal ions: As described earlier positively charged metal ions can help to stick RNA secondary structures together or bring them into close proximity, to form tertiary structures.

Among tertiary structural motifs one can find pseudoknots like kissing hairpins (44), see figure 18.

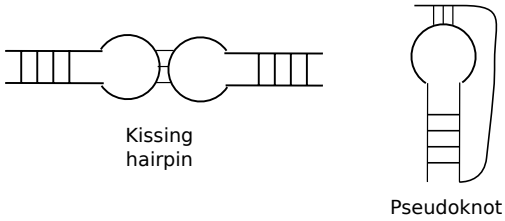


Fig. 18. Common RNA tertiary motifs.

- Pseudoknots form between loops and unpaired regions outside of this loops, again either intra- or intermolecular.
- Kissing loops are formed by hydrogen bonding between two loop regions and can be either intra- or intermolecular.

Correct formation of tertiary structures is necessary to provide an RNA molecule with e.g. catalytical function or for tRNA to function in translation.
Higher order **quaternary structures** can be introduced by interactions between the mentioned secondary and/or tertiary structures. They are of importance for intermolecular interactions of RNA molecules or interactions between RNA and proteins.
Folding of RNA can for example lead to the generation of aptamers, which are shaped RNAs that can bind selectively to ligands like charged molecules, amino acids or nucleotides. They are applicable as biosensors or for future therapeutic strategies by specifically binding a target protein and e.g. alter or block its active- or binding-site. One prominent example for RNA aptamers are riboswitches, which are structured RNA molecules that play an important role in gene expression of prokaryotes (45; 46). Only very few examples for riboswitches in eukaryotes are known and their role in eukaryotic gene expression still remains to be investigated. A review on RNA structures, methods to find and analyze them *in vivo, in vitro* and *in silico* and their role for organisms is available from Wan et al. (47).

miRNAs

Although this thesis is focused on the interactions between AREs and AUBPs, the interplay of AREs and miRNAs is a crucial point in gene expression regulation. Therefore this short section will introduce miRNAs and their influence on ARE containing transcripts. This paragraph is inspired by (48; 49; 50; 51; 52)
Short, single stranded RNA molecules of 20-23nt length are termed miRNAs. They are endogenously expressed in three steps. First a pri-miRNA is transcribed by RNA polymerase II at a miRNA gene locus. This pri-miRNA undergo the same processing steps as protein coding mRNAs, namely 5' capping and 3' poly-adenylation.
Stem loops inside the pri-miRNA mark it as target for the microprocessor complex. Part of this multi-protein complex is the endonuclease DROSHA, that cleaves the pri-mRNA. An alternative is the splicing of mRNAs, which can lead to miRNAs that are part of introns, as proposed by (53).
However, the resulting 70nt long precursor (pre-) mRNA forms secondary structures that are recognized by export proteins. In a Ran-GTP dependent manner, the pre-miRNA is exported into the cytoplasm, where an other dsRNA specific RNAse, DICER, cuts the pre-miRNA into the final mature 22nt long miRNA product.
By loading this miRNA into a member of the Argonaute (AGO) protein family, the latter can specifically bind to targets of the loaded miRNA and degrade them as part of the RNA-induced silencing complex (RISC).
MiRNAs play an important role for the regulation of gene expression. They can have (partially) complementarity to the 3'UTRs of mammal mRNAs, and cause degradation of their targets if latter are accessible for binding. This can interfere with the binding of AUBPs to their targets, increasing or decreasing their effects. A discussion of this topic follows in section 7.3 and more information can be found for example in (51).

2.1.3 RNA binding proteins

Regulation of gene expression often occurs co- and post-transcriptionally and allows the cell to quickly respond to changes in needs and environment.
This is in fact faster than changing the current set of transcription factors, and has a broad bandwidth of possible targets for regulation . The cell has different layers where regulation can take place. Alternative splicing is one example that makes it possible to produce more than one product out of a single mRNA. Polyadenylation, 5'-capping, splicing, RNA editing and RNA degradation are

known examples of internal regulation mechanisms (54; 55). RNA-binding proteins (RBP) take on a central role in this processes, see figure 19. The interaction or competition with miRNAs is another way of mRNA processing regulation (56). In addition, different RBPs interact with short RNAs to form ribonucleoprotein (RNP) complexes. RNPs have a functional role in various cellular processes like DNA replication, histone gene expression regulation and transcription and translation control processes (57).

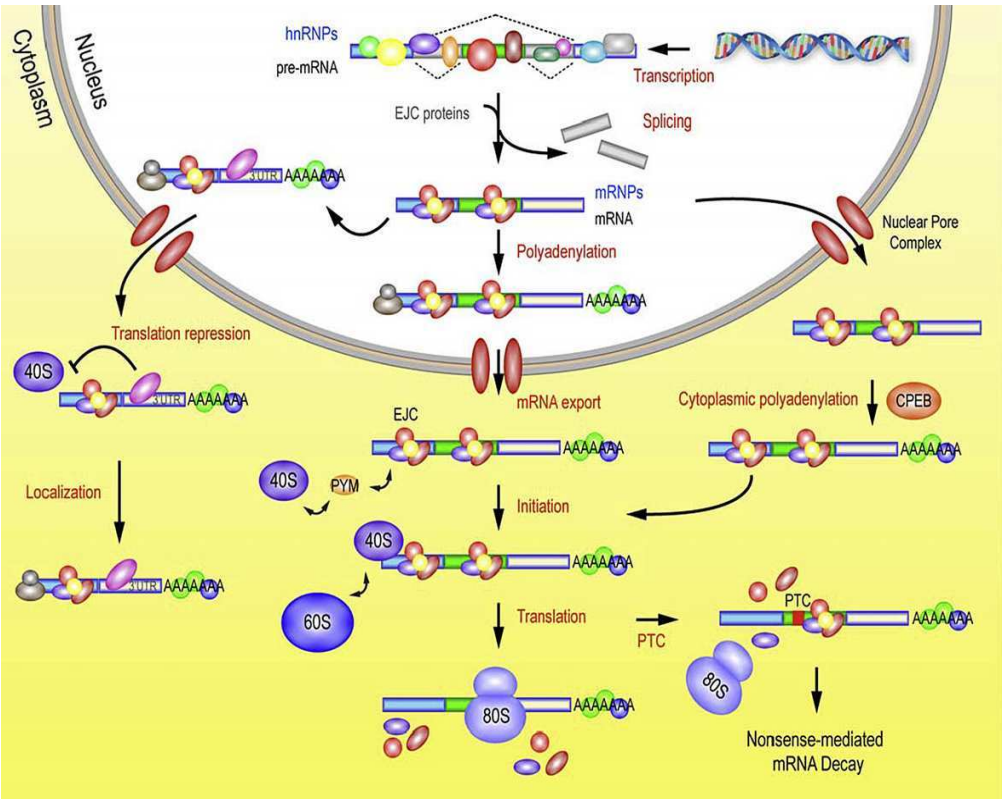


Fig. 19. This figure shows post-transcriptional gene expression regulation as a crucial role of mRNA-binding proteins. Regulation of mRNA splicing, editing, poly-adenylation, export, translation and decay is under control of RBPs. This figure was reprinted from (57) with permission from Elsevier.

RBP can bind to RNA in a sequence specific manner and with varying affinity. The development of versatile RNA binding domains allows these proteins to act on a broad variety of RNA motifs and structures. These domains can interact with single or double stranded RNA and recognize its primary, secondary as well as tertiary structures. Table 2.1.3 shows eleven RNA-binding domains and their properties as excerpt. This table and the following section are inspired by the review from Lunde (58).

One can see that a variety of domains for protein-RNA interactions have been found, each with specific motifs and function. Eukaryotic genomes where scanned for RBP motifs with bioinformatical methods. 5-8% of yeast and 2% of C. elegans and Drosophila protein coding genes where predicted to code for RBPs (59; 60; 61).
Among the best characterized RNA binding domains is the RRM, the RNA recognition motif. Referring to Maris (62), a total of 6056 RRM motifs have been identified in 3541 different proteins in the year 2005, mostly functioning in post-transcriptional processes regulating gene expression. Composed of 80-90 amino acids (63), often multiple copies of these motifs can be found along a single protein.
This modular architecture is common for RNA-binding proteins. Sometimes more of these proteins have to act together to bind a certain target specifically. Interaction between RNA and RRM containing proteins occurs via three conserved arginine or lysine residues that form salt bridges to the phosphodiester backbone and two aromatic residues, that introduce stacking interactions to the nucleobases of the RNA target, establish the sequence specificity in most cases.

Domain	Topology	RNA-recognition surface	Protein RNA interactions
RRM	$\alpha\beta$	Surface of beta-sheet	Interacts with about four nucleotides of ssRNA through stacking, electrostatics and hydrogen bonding
KH (type I and type II)	$\alpha\beta$ between β2, β3 and GXXG loop. Type II: same as type I, except variable loop is between α2 and β2	Hydrophobic cleft formed by variable residues and the bases from the GXXG loop and hydrogen bonding to bases	Recognizes about four nucleotides of ssRNA loop through hydrophobic interactions between nonaromatic sugar-phosphate backbone contacts
dsRBD	$\alpha\beta$	Helix α1, N-terminal portion of helix alpha2 and loop between β1 and β2	Shape-specific recognition of the minormajor minor groove pattern of dsRNA through contacts to the sugar-phosphate backbone; specific contacts from the N-terminal alpha-helix to RNA in some proteins
ZnF-CCHH	$\alpha\beta$	Primarily residues in alpha-helices	Protein side chain contacts to bulged bases in loops and through electrostatic interactions between side chains and the RNA backbone
ZnF-CCCH	Little regular secondary structures	Aromatic side chains form hydrophobic binding pockets for bases that make direct hydrogen bonds to protein backbone	Stacking interactions between aromatic residues and bases create a kink in RNA that allows for the direct recognition of WatsonCrick edges of the bases by the protein backbone
S1	β	Core formed by two beta-strands with contributions from surrounding loops	Stacking interactions between bases and aromatic residues and hydrogen bonding to the bases
PAZ	$\alpha\beta$	Hydrophobic pocket formed by OB-like beta-barrel and small $\alpha\beta$ motif	Recognizes single-stranded 3' overhangs of siRNA through stacking interactions and hydrogen bonds
PIWI	$\alpha\beta$	Highly conserved pocket, including a metal ion that is bound to the exposed C-terminal carboxylate	Recognizes the defining 5' phosphate group in the siRNA guide strand with a highly conserved binding pocket that includes a metal ion
TRAP	β	Edges of beta-sheets between each of the 11 subunits that form the entire protein structure	Recognizes the GAG triplet through stacking interactions and hydrogen bonding to bases; limited contacts to the backbone
Pumilio	α	Two repeats combine to form binding pocket for individual bases; helix α2 provides specificity-determining residues	Binding pockets for bases provided by stacking interactions; specificity dictated by hydrogen bonds to the WatsonCrick face of a base by two amino acids in helix α2
SAM	α	Hydrophobic cavity between three helices surrounded by an electropositive region	Shape-dependent recognition of RNA stemloop, mainly through interactions with the sugar-phosphate backbone and a single base in the loop

Tab. 3. Legend:

dsRBD	double-stranded RNA-binding domain	KH	K-homology
OB-like	oligonucleotide/oligosaccharide binding-like	RRM	RNA-recognition motif
ssRNA	single-stranded RNA	ZnF	zinc finger

Table adopted by permission from Macmillan Publishers Ltd: Nature reviews. Molecular cell biology (58), copyright (2007).

One RNA-binding protein containing an RRM is the poly-A binding protein, one of the key proteins for post-transcriptional processing. It has the ability to bind the poly-A tail of nascent mRNA, thereby protecting the RNA from degradation and tagging it for further steps like transport to the cytoplasm or initiation of translation (64).
Of special interest to this thesis are zinc finger proteins, as they are able to bind ARE motifs. Classically DNA binding, they have been identified to bind RNA as well. Multiple copies of zinc finger domains, classified by the residues that are used to coordinate the core zinc ion, are usually present in multiple repeats in a protein. A family of RNA-binding zinc fingers containing the zinc coordinating motif CCCH is known to bind to AU-rich elements, see figure 20. Sequence specificity is thereby established through hydrogen bonding between the protein sidechains and the Watson-Crick edges of the bases (65).

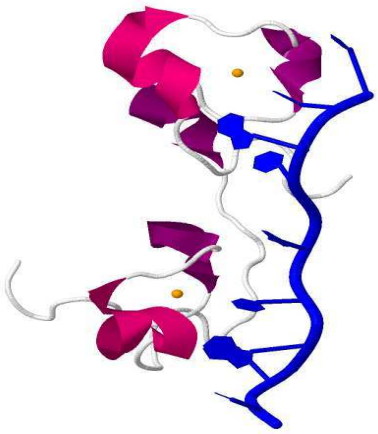


Fig. 20. Sketch of a tandem CCCH-zinc finger domain of TTP in contact with a class II AU-rich element (PDB: 1RGO). The nucleic acid is shown in blue and the two zinc ligands are shown in orange.

2.1.4 UTRs, AREs and ARE binding proteins

As described earlier, mRNAs do not only contain the region that codes for a protein. Beside this CDS, a typical eukaryotic mRNA contains a 5' as well as a 3' untranslated region (UTR).
These regions play a major role in translation and mRNA stability and decay (66; 67; 41).
The 5'UTR is defined as the region between the mRNA-cap and the start-codon. This region is known to play a role in translational control. In prokaryotes one can find the Shine-Dalgarno sequence inside the 5'UTR, in eukaryotes sequence elements like ribosome binding sites (RBS) and internal ribosomal entry sites (IRES) can be found here. The more regulation is needed, the longer the 5'UTR and the more stable secondary structures can be found here (66).
The 3'UTR is defined as the region between the end of the CDS and the poly-A tail. It is known to contain more regulatory elements and is longer as an usual 5'UTR. The 3' UTR plays a role in transcription and translation control. Elements like AU-rich elements, the iron response elements (IRE) and the previously mentioned poly-A signal can be found here (66).
Both regions, the 3'UTR as well as the 5'UTR are processed during or after transcription as mentioned previously and these processing steps are very important for the fate of the mRNA, see for example (28; 67).
The following list presents an overview of the regulatory elements mentioned above as found in (68; 69; 70), not including AU-rich elements, as they will be discussed in more detail in the next section. While SINEs, LINEs and Alu elements are per definition transposable elements, they can contain regulatory motifs and elements, and are common targets of editing effects (69) and have therefore been included in this list.

- Riboswitches: Mostly found in the 5' UTR of prokaryotic mRNAs this regulatory elements effect transcription attenuation as well as translation initiation. Regulation of riboswitches can occur through metal ions, catabolites or trans-acting partially complementary RNAs. Binding to a region of the riboswitch leads to conformational changes and opens or closes the regulated region, making it accessible for, or preventing binding of, interacting molecules.

- IRE: Iron response elements contain a 23- to 27-bp stem with a mismatched C and a 6-nucleotide loop with C at its 5' end. They function by

binding an iron-regulatory protein (IRP or IRE-BP for IRE-binding protein) and effect the mRNA of the transferrin receptor and ferritin, both regulating iron homeostasis in mammalian cells.

- RBS: The eucaryotic ribosome binding site usually includes the 5'-cap. Regulation of translation can occur here for example by masking the binding site via secondary structures or factors that bind this region and block the RBS for the ribosome.

- IRES: Internal ribosome entry sites are usually found in viral RNAs but can also be found in eukaryotes. These binding sites are 5'-cap independent influence the expression of highly regulated genes like genes coding for stress response proteins.

- poly-A signal: The consensus sequence AAUAAA is found in the 3' UTR of eukaryotic mRNAs. This sequence acts as signal for cleavage of the nascent mRNA product 10 to 30 nucleotides downstream of this signal. Cleavage is followed by addition of a poly-A tail by poly-A-polymerase. Often more than one signal exist along the 3' UTR, making it possible for the cell to produce various transcript isoforms. The poly-A tail plays an important role in mRNA stability and processing, as well as export.

- SINE: Short interspersed nuclear elements, are repetitive elements like mobile elements or pseudogenes. With a length of less than 500bp, their main representative are Alu elements.

- LINE: Like SINEs, long interspersed nuclear elements are repetitive elements. Their length is more than 500bp, and they encode their own reverse transcriptase, essential for their amplification. They also contain a poly-A tail, which has, together with the transcriptase, influence on their mobility.

- Alu element: Repeated sequences containing a recognition site for the restriction enzyme AluI. With a full length of about 300bp they belong to the family of SINEs. They can usually be found in gene rich regions, and comprise up to 10% mass of the human genome. Easy to identify by their highly conserved sequence, this mobile elements act on gene expression

regulation in various ways. They can act as regions for modifications like methylation and through their mobility, be inserted into genomic regions that are to be regulated.

There are many more regulatory elements involved in regulation of gene expression, see for example (71). However, the focus here lies on regulation of gene expression via AU-rich elements, which will be discussed in the next chapter followed by a section concerning proteins that bind to these regions and their functions.

AU-rich elements

Regulation of gene expression can occur on different stages. One of them is by control of mRNA stability (72). This post-transcriptional regulation step is often controlled by regulatory elements that can be found in the untranslated regions of mRNAs as previously described. Among these elements the most important for mRNA stability regulation are the cis-acting AU-rich elements, short AREs. They where first discovered in the 1980's, when it was seen that *c-fos* could transform cultured cells after removing a specific sequence from its 3'-UTR, changing its stability (73).

A destabilizing effect of these elements has been confirmed by inserting a 51-nucleotide long AU-rich sequence, from the 3'- UTR of the granulocyte-macrophage colony stimulating factor (GM-CSF) mRNA, into the 3'-UTR of β-globin mRNA, resulting in a profoundly shortened half-life of the target mRNA (74).

AU-rich elements are found in the 3' UTR of approximately 8% of human protein coding genes. Their products play a role in various important processes like: "immune responses, cell cycle/proliferation, inflammation and coagulation, angiogenesis, metabolism, energy, DNA binding and transcription, nutrient transportation and ionic homeostasis, protein synthesis, cellular biogenesis, signal transduction, and apoptosis" (75), highlighting their importance.

A first arbitrary definition of AREs into three classes was proposed by (76).

- Class I ARE contain several dispersed copies of the AUUUA motif within U-rich regions.
- Class II ARE contain at least 2 overlapping UUAUUUA(U/A)(U/A) nonamers.
- Class III ARE are U-rich regions that do not contain the AUUUA pentamere.

Though this characterization by sequence motif only may not be perfect and a characterization of AREs according to their biological function would be more interesting, it has been established and will be used throughout this thesis. Down to the present day, no real consensus sequence has been discovered and ARE motifs reach from the pentamer AUUUA to multiple adjacent copies of the WWWUAUUUAUWWW tridecamer, where W stands for A or U.
As already mentioned, AREs are cis-acting regulatory elements, which means that they affect the molecule of mRNA where they are present. They enable so called AU-rich binding proteins (AUBP) to interact with the mRNA, causing stabilization or decay of their target. These effects do not always have to be actively executed by the AUBP. One can imagine that binding to an ARE alone can effectively prevent other factors like the poly-A polymerase to bind to the mRNA, if both sequence motifs can be found in close proximity, thereby influencing the mRNAs stability and processing, depending on the cell compartment in which the ARE targeting molecule can be found (77; 78).
Degradation of an AU-rich binding protein targeted mRNA is thought to begin with deadenylation. Deadenylation can be induced for example by proteins of the TIS11 family, like TTP (79). A missing poly-A tail would prevent translation of a mRNA, as the poly-A binding protein can no longer bind to the mRNA, which leaves the RNA exposed to exonucleolytic attacks. This step followed by decapping and the attack of a 3'-5' exonuclease-complex, the exosome, or 5'-3' exonucleases via processing-bodies (p-bodies) would degrade the AUBP target.
In the case of stabilization, a blocking of binding sites for degrading proteins may be sufficient to prolong the mRNA lifetime.
The exact mechanism of mRNA degradation initialized by AUBPs remains unclear, but it is fact that regulation of mRNA stability via ARE motifs is an important mechanism for gene expression regulation (80; 81; 82; 55; 79).

AU-rich binding proteins

Active regulation of mRNA stability is possible through interaction between an AU-rich element and an AUBP. According to (65), all zinc finger containing AUBPs have in common that they bind single stranded RNA, the same is thought of other AUBPs. Previously mentioned was the influence of RNA secondary structures on the action of AU-rich binding proteins. One can imagine that an ARE motif that is embedded in a sequence rich in secondary structures can be less accessible for AUBPs than a motif that is found in a longer single stranded sequence or even exposed to the outside of an RNA molecule when found in the loop region of a stem loop for example. Previous work has highlighted the influence of secondary structures on the ARE-AUBP interplay. Meisner et.al presented a molecular switch that allows to regulate the binding of HuR to its target mRNA by introducing a secondary structure containing the HuR target ARE motif which prevents HuR from interaction or breaking up this structure, leaving the ARE motif in a single stranded RNA sequence which was bound by HuR (83).

A lot of AUBP exist, among them, three are of outstanding interest for this thesis, due to the huge amount of data available on their targets and regulatory functions. The focus of this thesis lies on three AUBPs, namely:

TTP: Tristetraprolin is known to have a destabilizing effect on its mRNA target and is predominantly found in the cytoplasm (81). Its tandem zinc finger domain can bind to class II AREs and promote deadenylation and degradation of its target mRNA. TTP can induce deadenylation and is in contact with the mRNA decapping and degradation machinery, after binding the core UUAUUUAUU ARE motif of its target (79). Two CCCH-zinc fingers can bind in a symmetrical fashion to adjacent 5-UAUU-3 subsites on the single-stranded RNA. They combine electrostatic and hydrogen-bonding interactions with stacking between conserved aromatic side chains and the RNA bases (65). A maximum turnover of target mRNA could be observed when two TTP molecules are attached simultaneously (84). TTP was found to regulate its own expression via a negative feedback loop, binding an ARE present in its own mRNA and promoting decay (85). The most prominent target of TTP is the tumor necrosis factor alpha, short TNF-α, which, if not down regulated by TTP, can accumulate inside the cell and lead to cancer (86). Proteins of the TIS11 family (e.g. TTP, BRF1) seem to be able to cross regulate their mRNA targets, a hint on their importance in regulation of inflammatory processes (87) and leukemogenesis (88). Interaction between TTP and a class II ARE can be seen at figure 21C.

AUF1: AUF1 (A+U rich RNA-binding factor1, heterogeneous nuclear ribonucleoprotein D) occurs in four proteins isoforms, p37, p40, p42 and p45. Designated by their molecular masses, these isoforms are spliced from a single transcript and play a role in mRNA decay and translation control (89). AUF1 can bind to class I and class II AREs, and can shuttle between the cytoplasm and the nucleus (81). It has been found to have a destabilizing effect, which can change if influenced by heat-shock processes and (de-)phopshorylation (90). As four isoforms of AUF1 have been discovered, it remains hard to link destabilizing or stabilizing effects to one of them (91). Like TTP and HuR, AUF1 is known to interact with other RBPs to increase its effect on a target or to compete for target sites. Among AUF1 targets one can find a lot of proto-oncogenes like c-myc or c-fos (92) and it is known as key player in the regulation of hematopoesis (88). A scetch of the c-terminal RNA binding motif of Auf1 can be seen at figure 21A, for a more detailed analysis of this region refer to (93).

HuR: HuR is an ELAV (embryonic lethal, abnormal vision) like protein. ELAVL1/HuR is known to act as post-transcriptional gene expression regulator for transcripts containing AU-rich elements (94), in particular the nine nucleotide long sequence 'NNUUNNUUU'. Ubiquitously expressed and predominantly nuclear localized, this protein has the ability to shuttle to the cytoplasm and back to the nucleus, which is different to the other members of the Hu-protein family, as they are restricted to neurons (95). HuR contains three RRM motifs, two N-terminal binding to ARE motifs and the remaining C-terminal motif binding to poly-A tails and stabilizing the RNA-protein complex (96; 97; 98). Although an affinity for U-rich regions exists, a complex between HuR and ARE motifs remains more stable if all RRM motifs and the hinge region are involved (99). The three RRM motifs, as well as the linking hinge region help HuR to bind ARE motifs in a length-dependent manner by forming multimers (94). As HuR and AuF1 have been found in the same tissues, these two AUBPs seem to have some kind of interplay in regulating mRNA turnover (100). TTP seems to play a role in the autoregulation of HuR expression levels. High levels of HuR in the nucleus influence the usage of alternative polyadenylation sites of HuR mRNA, which leads to the expression of a longer, ARE containing version of this mRNA, resulting in a high degree of degradation of HuR mRNA via TTP (78). This would mean that the localization of HuR has influence on its role in mRNA stabilization/decay and displays an other example of autoregulation of RBP expression levels. Targets of HuR seem to overlap with TTP or AuF1, with the difference that HuR has a stabilizing effect on its mRNA targets in most cases. A scetch of the RRM1

of HuR can be seen at figure 21B, for a more detailed analysis of this region refer to (101).

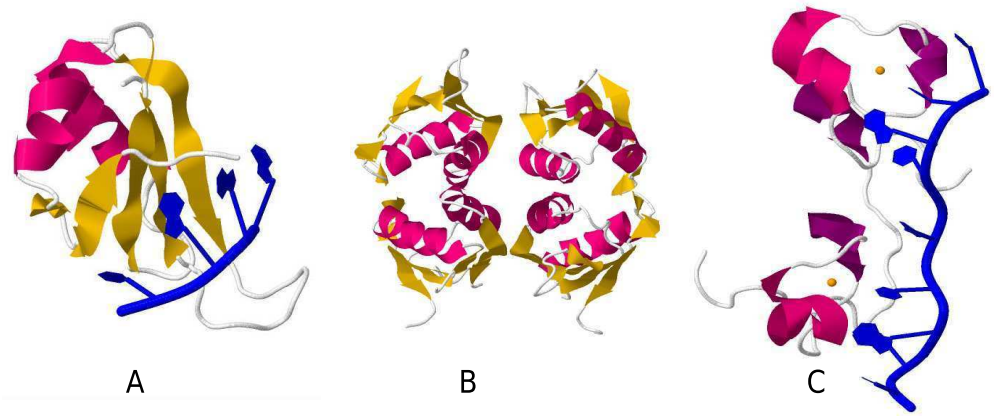


Fig. 21. A) Shows a sketch of the c-terminal RNA binding motif of AUF1 in interaction with a nucleic acid strand (in blue) (PDB-ID: 1WTB)
B) Shows a sketch of the first two tandem RRMs (RRM1) of HuR that is known to bind ARE motifs(PDB-ID: 3HI9)
C) Shows a sketch of a tandem CCCH-zinc finger domain of TTP in contact with a class II AU-rich element (PDB: 1RGO). The nucleic acid is shown in blue and the two zinc ligands are shown in orange.

All three AU-rich binding proteins can be found to compete with each other for single stranded target sites, sometimes having an agonistic and sometimes an antagonistic effect on the stability of their targets. They can have a great influence on the half-life of mRNAs, providing the cell with a fast response mechanism to environmental or developmental conditions (81; 102). The complex interplay of AUBPs with each other or other RBPs as well as their broad spectrum of targets leaves room for a lot of investigations.
Interactions or cross reactions of AUBPs and miRNAs are also known and present further roles of AUBPs in translational control and mRNA decay (51). Summing up all actions of AUBPS is an almost impossible task, but this overview presents main findings of their role in the cell and highlights the large area of functional properties that can be targeted by future scientific analysis.

2.2 Bioinformatics

The combined application of computer science and information technology to the field of (molecular) biology is termed bioinformatics.
Creating algorithms and databases as well as using statistical and computational approaches, the primary goal of bioinformatics is to identify fundamental principals of biological processes. Resulting knowledge may be used to predict data instead of just observing data in the further course.
Bioinformatics is a fast growing field, further stimulated by the explosion in available computing power. Beginning with genomics and proteomics followed by drug design as well as synthetic biology, computational methods are of growing importance. The possibility to search for new drugs in a virtual lab can reduce costs and time compared to wet lab work, a point very important in all fields of research. Some of these fields would not even be possible without computational help. Deep sequencing or all sorts of -omics produce data on large scale, making it absolutely necessary to use computational approaches for analyzing these datasets.
The generation of a database, containing the annotated protein coding transcripts of human and mouse as well as their analysis where part of this thesis. Alignments of ARE containing transcripts where generated to get conservation information for ARE motifs among different species. Furthermore the analysis includes prediction of secondary structures inside the 3' UTRs, leading to information about the accessibility of present ARE motifs. The database and the results of above mentioned analysis steps were then used to generate a webserver that allows to screen for ARE motifs in the 3' UTR of human and mouse transcripts. This tool was named 'AREsite' and can be used to analyze known ARE motifs or screen for novel targets of AUBPs, as will be discussed in section 5.
The next section is inspired by the textbook "Biological sequence analysis" (103) and will focus on alignment algorithms and the programs used to analyze the dataset in the database, followed by a section about folding algorithms and the implementation RNAplfold that was used to calculate ARE motif accessibility.

2.2.1 Methods to generate alignments of ARE containing transcripts

ARE motifs have a known functional role. Their conservation among different species helps do distinguish between ARE motifs that can be found by chance in an A+U rich 3' UTR and motifs that actually have a regulatory function. Comparison of biological sequences is one of the most important operations in computational biology, as sequence (see definition 1) similarity implies similarity in structure and function.
A sequence x is defined as:

$$x = x_1...x_n x \in \mathcal{A} \text{ where } \begin{cases} DNA : \mathcal{A} = \{A, T, C, G\} \\ RNA : \mathcal{A} = \{A, U, C, G\} \\ Protein : \mathcal{A} = \{\text{see figure 10}\} \end{cases} \quad (1)$$

What biologists are searching for is a measure of sequence homology to gain information about sequence conservation.
Whereas sequence similarity means that compared sequences are similar in matters of their nucleotide composition, homology implicates similarity in function. A homologue is a gene or protein found in different species with a common ancestor and often similar structure and function. There are special types of homologues, orthologues and paralogues. An orthologue is a gene or protein with same function but found in an other organism due to a speciation event in the common ancestor, whereas a paralogue is derived by a gene duplication event and results in two copies of a gene on different loci of the same genome with not necessarily the same function. No matter what kind of homology a researcher is looking for, sequence comparison will give a clue about the conservation of an appointed sequence or sequence motif.
Naive pairwise sequence comparison is done by placing two sequences one above the other, thereby placing identical or similar bases in the same column and inserting gaps in form of dashes opposite a non-complementary base, a process called pairwise alignment of two sequences.
To get information about the analogy of these two sequences, one has three possible options, comparison by Hamming distance d_H, by edit distance d_edit or by similarity S.
Hamming distance d_H (see definition 2) simply counts the number of different characters δ, among two sequences placed one above the other. This is very intuitive, but of limited use if sequences differ by more than just single characters (as introduced by point mutations). Hamming distance is thus not related to

real live processes.
Edit distance d_{edit} (see definition 3) is a measure for the minimal amount of operations necessary to transform sequence x into sequence y. This leads to more complex calculations to extract edit distance, but gives results that are more similar to real biological processes.
d_H: The Hamming distance for sequence x and sequence y

$$d_H(x,y) = \sum_{i=1}^{n} \delta(x_i, y_i) \qquad \delta(x_i, y_i) = \begin{cases} 1 \text{ if } x_i \neq y_i \\ 0 \text{ (else)} \end{cases} \tag{2}$$

d_edit: The edit distance for sequence x and sequence y,
given edit operations Replace (R), Insert (I) or Delete (D) with different operation costs:

$$d_{edit(x;y)} = \sum_{i=1}^{n} \delta(x_i, y_i)\delta(x_i, y_i) = \min \begin{cases} R(x_i; y_i) \\ I(x_i; y_i) \\ D(x_i; y_i) \end{cases} \tag{3}$$

While Edit and Hamming distance are a measure for relatedness of two sequences, a measure for similarity can be retrieved, by assigning a similarity score to each pair of characters which allows to fine tune an alignment.
Below the recursion for calculating similarity S of sequence x and sequence y with linear gap costs can be seen, where $\delta_{(x_i,y_j)}$ is the similarity score for the given pair of characters and g is the gapcost:

$$S_{i,j} = \max \begin{cases} 0 \\ S_{(i-1,j-1)} + \delta_{(x_i,y_j)} \\ S_{(i-1,j)} + g \\ S_{(i,j-1)} + g \end{cases} \tag{4}$$

The arising question is, given two sequences can be aligned in various ways, which alignment is the best one to retrieve as much conservation information as possible.
In order to evaluate this, a scoring scheme is needed. Such a scheme has to incorporate various mutational processes, for example insertions or deletions, which will now be referred to as InDels. Other mutations include substitutions (which are point mutations), as well as duplications, inversions and combinations thereof.

The introduction of gaps into an alignment was already mentioned, wherever a InDel or other mutation leads to missing character in the opposite sequence. Gap scores or gap penalties, as they usually contribute with a negative score, can be linear or affine. Linear means that opening of gaps and elongation of gaps add the same score, which can be problematic if one thinks of long insertions or deletions that are the case in many biological processes, as the gap costs would grow very fast. Affine gap costs consist of two scores, a larger one for opening gaps, and one for elongating them, the latter always smaller than the first. As InDels often concern more than one nucleotide, affine gap costs can simulate real live processes more effective than linear gap costs, with rising costs for computational speed and memory usage. One can see that it is very important to use or create a scoring scheme that fits the problem one wants to investigate, or it may lead to suboptimal or in the worst case wrong results.
Substitutions are a special case of mutations and require a special scoring scheme. Depending on the substitution that occurred and the position in the alignment where it was found, the scoring scheme has to differ between mutations that have influence on structure/function and mutations that do not. Substitution matrices like the BLOSUM (BLOcks of amino acid SUbstitution Matrix) matrix for amino acids or the PAM (Point Accepted Mutation) matrix exist for this case and are used in the alignment programs utilized for the analysis of ARE motif conservation. These matrices are built up by taking into account which amino acid or nucleotide can act as substituent for an other, depending on their properties like charge or hydropathy and the probabilities for observing such pairs in real alignments. If amino acids or nucleotides are very similar, they get a positive score, and otherwise a negative score.
Once a scoring scheme is established, an algorithm for prediction of possible and/or optimal alignments has to be found.
To get an optimal alignment, an algorithm has to maximize the number of positive scored pairs while minimizing costs for gaps and mismatches. Due to the huge amount of possible alignments even for short sequences, dynamic programming algorithms are used to find optimal alignments or subsets thereof in an acceptable time frame. The idea behind dynamic programming algorithms is to break down problems into smaller subproblems which can be solved faster, thereby storing results for this subproblems and simply load them if they reoccur during processing instead of recalculating them. As subproblems usually occur more than once during alignment operations, a dynamic programming approach is efficiently saving time and computational power.
Even faster methods use Heuristics, making additional assumptions and sac-

rificing accuracy for time. The big disadvantage of heuristical methods is the possibility of not finding a real optimal alignment for certain sequence pairs. Extensively used and modified algorithms for a variety of bioinformatical implementations concerning alignment operations, including the programs that have been used during this thesis, follows this short introduction.

Global Alignment Algorithm In 1970 Needleman and Wunsch introduced an algorithm for finding optimal global alignments between two sequences (104). It fills a matrix with the best score S, depending on the positions i and j in the alignment of sequence x and sequence y where $\delta_{(x_i,y_j)}$ is the similarity score for the given pair of characters and g is the gapcost (see equation 5).

$$S_{i,j} = \max \begin{cases} S_{(i-1,j-1)} + \delta_{(x_i,y_j)} \\ S_{(i-1,j)} + g \\ S_{(i,j-1)} + g \end{cases} \tag{5}$$

For each position of sequence x aligned with each position of sequence y, a score and a pointer to the cell where the score was derived from, is stored in the matrix. The first row and column of this matrix are filled with values that are derived by aligning each position of the sequences against gaps. The upper left position in the matrix is filled with a zero, marking the start position. Moving to the first position of sequence x, the best score for all possible actions (match, mismatch or gap) and a pointer to the cell where the value was derived from, is stored for alignments with all positions of sequence y by simply calculating the maximum value for each action. This procedure is repeated until the matrix has been filled, see figure 22.
Once the matrix is filled row by row from top left to bottom right, a process termed traceback is conducted, starting at the last position for sequence x and y. This position holds, by definition, the best score for the alignment of the two sequences as the scoring scheme is additive. Moving back to the start of the matrix, always following the pointers to the cells where the last maximum value was derived from, gives the resulting optimal alignment. This algorithm was modified by Gotoh, making it faster and more efficient (105) and with slight modifications, it can give a whole set of optimal or even suboptimal alignments as result.

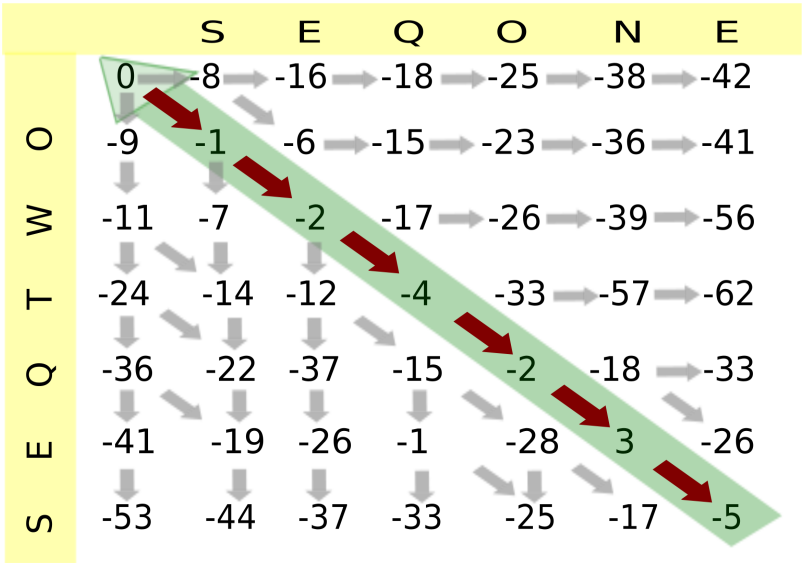


Fig. 22. An implementation of the scoring matrix created by the Needleman - Wunsch algorithm. For each position of the two sequences, the maximum value, and a pointer (grey arrows) to the cell where it was derived from, is stored in the matrix. The backtracing algorithms leads to the best global alignment (green arrow) by following the optimal pointers (red arrows) from the lower right cell where the best score for the alignment is stored, to the beginning of the alignment in the upper left cell.

Local Alignment Algorithm When comparing protein domains, or sections of DNA sequences like promoter regions or in our case ARE motifs, it is of limited use to get optimal global alignments. Analyzing the conservation of a subsequence among other sequences is done by using a local alignment algorithm. The most prominent algorithm was implemented by and named after Smith and Waterman (106).

The big difference to the Needleman-Wunsch algorithm is the possibility of inserting zeros anywhere in the matrix, where the otherwise calculated maximum value would be lower than the value in the source cell or zero. This means, that instead of introducing bad scores into an existing alignment for gaps at the end of a local alignment, one simply starts a new alignment, see equation 6.

Like using the algorithm for global alignments, the best scored operation O for position i,j in the alignment of sequence x and sequence y, but in any case a higher score than the one before is retrieved, or a new alignment is started by inserting 0. Again $\delta_{(x_i,y_j)}$ is the similarity score for the given pair of characters and g is the gapcost:

$$O_{i,j} = \max \begin{cases} 0 \\ O_{(i-1,j-1)} + \delta_{(x_i,y_j)} \\ O_{(i-1,j)} + g \\ O_{(i,j-1)} + g \end{cases} \tag{6}$$

Using edit distance is not productive in this case, therefore similarity is used to evaluate the best alignments, see equation 4. The traceback process can not simply begin at the lower right corner, but has to start at the highest value inside the matrix and ends at the next zero on the way back, indicating the best local alignment inside the matrix. This process can be altered, so that all local alignments can be found instead of just the optimal one, which can be used to screen for conservation in organisms that are no closely related or divided by a large timespan.
To do this, all values contained in the first local alignment are set to zero, and necessary recalculation processes are done, a process known as deblocking. To guarantee that local alignments are not masked by longer non-optimal or global alignments with higher scores, the scoring scheme has to be adapted, so that only very good matches give positive scores. This requires advanced scoring schemes as well as techniques to develop them. Both will not be addressed in this thesis, but can be found among more advanced alignment methods or advanced methods to construct a scoring scheme by analyzing statistical datasets and more in (103).

Heuristic Alignment Methods As mentioned previously, a possibility to speed up the process of finding optimal alignments is to use heuristics. Using this techniques can be a powerful method where exhaustive search would be too expensive, as can be the case when screening large 3' UTRs. This section will introduce two algorithms that use heuristics to speed up the alignment process and are extensively used throughout the community.

FASTA

The FASTA (107) algorithm searches for matches between a given string and strings in a database, by finding the most similar local regions in a dynamic programming matrix. To speed up this search, FASTA is splitting the search string in smaller substrings and looks for exact matches (hot-spots) between

this substring and the database substrings. Thereby FASTA simply tries to find the searched substring inside its database. Once hot-spots are found, they are scored according to an adequate scoring matrix. The ten best hot-spots are re-scored with shorter subsequences using a match score matrix like PAM and the best scoring sub-alignments are extracted. These sub-alignments are then combined into one larger alignment, introducing gaps if necessary. Finally a Smith-Waterman algorithm is used to produce an optimal local alignment, even though this is only a highest scoring region and may not be the best possible alignment.
The corresponding FASTA format is among the most used sequence formats one will encounter if working with bioinformatic tools and databases. A typical FASTA file begins with a '>' followed by one or more identifiers for the protein or nucleic acid sequence separated by '|'. The sequence itself is written in single-letter code below the header and can be easily parsed by scripting languages like Perl or text processing tools or converted to other file formats. Because of its ease of use the FASTA format became a standard in bioinformatics very soon.

BLAST The Basic Local Alignment Search Tool (108) is the most widely used bioinformatical program implemented until now. Again, a heuristic is looking for local alignments between a search string and strings in a database. With the parameters, word length, a given word similarity threshold, and the minimum match score, BLAST computes high-scoring segment pairs (HSPs) between two sequences. A HSP describes a locally maximal segment alignment, which means its score is above the minimum score threshold and can not be improved by shortening or extending the segment pair. To find these HSPs, BLAST uses a heuristic similar to the Smith-Waterman algorithm, see equation 6. Step by step BLAST generates substrings of a given length and calculates scores with all possible sequences of same length. Those having a score greater than the threshold are now compared to sequences in a database for exact matches. If they can be found, a scoring matrix like BLOSUM50 is used to score the surrounding character-pairs and again those with a score greater than the given threshold are saved. Neighboring perfect matches are clustered together and when the scoring is completed, the HSPs found are compared to scores that would be expected if random sequences would be compared. This returns E- and P- values for the resulting alignment. The P-value is the probability to get a score greater or equal to the score of the random sequences, and the E-value or expectation value provides information about how often such a P-value can be expected in a random query of the database. With this information the user

can rate whether or not the resulting alignment is satisfying his needs.

Multiple Alignments For phylogenetic predictions of a certain subsequence like the ARE motifs, pairwise alignments are simply not enough. It requires more than two sequences to determine the level of conservation. The next section will introduce methods to generate multiple sequence alignments and will end with a brief overview of the programs that were used to generate alignments for a phylogenetic analysis of ARE motifs, which has been included in the webserver 'AREsite'.
Whenever more than two sequences are compared to each other by alignment methods, we speak of a multiple alignment. The challenge in producing multiple alignments, is again to find and score the optimal alignment(s). As more than two sequences have to be compared, the usual way of using scoring matrices is no longer feasible. One way to handle this problem is to use the **sum of pairs score**. Here the alignment is produced by dividing a multiple alignment into all pairwise alignments and scoring each one of them, building the optimal multiple alignment with the best scored pairwise alignments as anchors.
An elaborated version hereof is to produce a weighted sum of pair score, where sequences get a weight according to their appearance in the alignment. This helps to circumvent a scoring problem if alignments contain more sequences from species A than B, as otherwise sequences derived from species A would get a better score than appropriate.
As no algorithm can solve the multiple alignment problem for growing numbers and length of sequences in an acceptable time-span, the established programs use heuristics to circumvent limitations in CPU power and memory. An example is the **progressive alignments heuristic**, were smaller multiple alignments are produced and combined to the final large multiple alignment. Beginning with pairwise alignments, sequences with close relationship are combined to a multiple alignment and later again combined to related multiple alignments and so on. To do so, first of all a guide tree, containing the degree of relationship between sequences is produced which is later on it is used to combine the alignments according to this relationship. Afterwards a new guide tree can be calculated and if better than the first, it is used for further alignment steps.
As the calculation of sum of pair scores for high numbers of sequences becomes very complex, predefined profiles containing the probability or frequency of all possible characters are used for easier scoring. To further improve this technique, iterative methods can be used. Thereby an existing alignment is divided

and realigned as long as better alignments are possible, to get the best possible multiple alignment for given sequences.
The following section focuses on programs used during this thesis to produce alignments of ARE containing transcripts in human or mouse with transcripts in other species, leading to information about the conservation of the ARE motifs.

ClustalW This section is inspired by (109), where detailed information can be found. ClustalW's ancestor was written in the 1980's and rewritten and modified over the years to become today's ClustalW, which was first introduced 1994. ClustalW incorporates a position-specific scoring scheme and the name giving **W**eighting scheme for down weighting over-represented sequence groups. Alignments can be produced by using a faster heuristical approach for pairwise alignments and the calculation of a distance matrix, or taking the slower but more precise route of calculating pairwise alignments with a dynamic programming approach and affine gap costs. In both cases a guide tree is build up with the distance matrix calculated, followed by progressive multiple alignment. To get the resulting multiple alignment, improvement steps are introduced, modifying sequence weighting, gap penalties, similarity, divergence and length dependence of the sequences as well as the weight matrices (PAM or BLOSUM) used.
The final alignment does not have to be the optimal alignment, but improvements in sensitivity and accuracy are routine implementations. ClustalW was used during this thesis for its fast and easy handling.

DIALIGN The program DIALIGN (110) is using an internal database of local fragment alignments to search for statistical significant similarity with fragments of the query sequence or the whole sequence if possible. This approach is similar to FASTA or BLAST, but here multiple sequences are compared from begin on. An interesting feature is the possibility to use alignment anchors, forcing the program to align anchored positions to each other. This approach was used when locally aligning ARE motifs found in different species. Providing the options to use various heuristics or combinations with other tools to speed up the alignment process, DIALIGN can be used to align multiple sequences fast and with high precision.

MAFFT In this program, nucleic acid sequences are converted to sequences of four-dimensional vectors, containing the frequencies of the occurring bases.

Fast Fourier transformation (FFT) is used to calculate the correlation of sequences. Homologous regions show a peak in their correlation, and these peaks are used to align the sequences. Using a sliding window approach with window size 30, the sequences are scanned for peaks in correlation, and if segments are identified as homologous segments, they are combined. A modified PAM similarity matrix with a special gap cost and other modifications is used to produce the optimal alignment, segment by segment. MAFFT is a tool that provides various possibilities to calculate and score optimal alignments, as was done during this thesis. Further information can be found at (111).

TBA and MULTIZ The Threaded Blockset Aligner (TBA) (112) is an approach to circumvent exclusion problems in the typical genomic sequence alignment approach. Typically, multiple genomic alignments use one genomic sequence as reference, which means that the other alignments are projected on this sequence, and sequence blocks that are contained by the reference are lost. TBA produces a set of blocks (for example, local alignments of some or all of the given sequences), which contain each position in the given set of sequences exactly once. Assuming that the matching regions occur in the same order and orientation in the whole set, detected matches are represented among this blocks. MULTIZ is the dynamic- programming alignment program inside the TBA program collection. It can be used for sequences containing inversions and duplications, or sequences that are fragmented. MULTIZ was used to build whole-genome alignments for the UCSC Genome Browser (113), which where used during this thesis to extract ARE motif conservation.

2.2.2 RNA folding algorithms as methods for motif accessibility prediction

In section 2.1.2 the types of secondary structures found in RNA molecules were introduced. These structures are very important in the prediction of possible protein binding sites along a RNA molecule. If a very stable structure is found in a protein binding region, this structure may prevent interactions between protein and RNA. The more stable a structure, the more energy has to be provided to open this structure and allow interaction. This strategy can be used for regulation of translation and transcription by e.g. RBPs, miRNAs, or riboswitches, as mentioned previously.
As the interaction of AUBPs with AREs is the main topic of this thesis and AUBPs are single stranded RNA binding proteins, it was a logical step to have a look at the secondary structuredness of ARE containing transcripts.
Folding RNA *In silico* does not come without problems. The first restriction is that tertiary structures are almost impossible to predict except for selected cases. Secondary structures on the other hand are predicted routinely in bioinformatical analysis, where a secondary structure has to fulfill certain constraints. The constraints for a secondary structure (defined as a set of base pairs Ω on a sequence S), cited from (114) are:

- A base cannot participate in more than one base pair, i.e., Ω is a matching on the set of sequence positions (excludes non-canonical base pairs).
- Bases that are paired with each other must be separated by at least 3 (unpaired) bases (RNA backbone can not bend too sharply).
- No two base pairs (i;j) and (k;l) $\in \Omega$ "cross" in the sense that $i < k < j < l$. Matchings that contain no crossing edges are known as loop matchings or circular matchings (excludes pseudoknots).

The first efficient algorithm to predict single stranded RNA secondary structures was introduced by Nussinov in 1980 (115). This algorithm produces a secondary structure with a maximum of introduced base pairs. Similar to alignment algorithms a matrix is filled with all possible base pairs and via a backtracking routine the structure with the maximum number of base pairs is received. A:U and G:C base pair energy contributions are seen as equal and contributions of stacking or destabilizing effects are not regarded in this algorithm.
In vivo the thermodynamic stability of RNA secondary structures is a prerequisite. Based on the work of Nussinov (116) and Waterman&Smith (117), whose model incorporates stacking and destabilizing effects, Zuker introduced

the loop-based energy model (118), where total free energy contributions come from loop regions only, depending on their size and type (interior loops, bulges and multi-branched loops) in 1981 (nearest neighbour model). This is the first approach to divide an RNA sequence in loop- and non-loop regions (loop decomposition). The total free energies $F(S)$ of a secondary structure S is seen as sum of the energy contributions F_L of all loop regions $L \in S$ along an RNA molecule, see equation 7.

$$F(S) = \sum_{L \in S} F_L \tag{7}$$

Zuker presented a dynamic programming approach based on this model to solve the RNA folding problem in an acceptable time scale of $\mathcal{O}(n^3)$. The energy contributions for some loop types have been measured experimentally by the group of Douglas Turner and are used in programs for secondary structure prediction, e.g. (119).
Based on this loop model is the minimum free energy (MFE) approach. Here the RNA sequence is again divided in loop- and non-loop regions. In a first step, the MFE of all loop regions is calculated in the forward recursion, and via a backtracing routine, the actual base pair pattern is retrieved in step two. The outcome of this pattern is a MFE structure, but it does not have to be the only, or the best one for the given sequence.
Coupling this method to the partition function of an RNA sequence leads to the probability of a given MFE structure among the ensemble of all possible structures. The partition function encodes the statistical properties of a system in thermodynamic equilibrium. A dynamic programming algorithm to compute the equilibrium partition function followed by the base pairing probabilities of a given RNA molecule was introduced by McCaskill in 1990 (120).
Calculation of the accessibility of certain RNA sequence intervals is done by calculating the probability that a sequence of nucleotides is found unpaired. Mathematically this is equivalent to the amount of energy needed to make this sequence single stranded, thus to open all structural motifs (121). As ARE binding proteins can only bind to an accessible ARE motif, calculations of motif accessibility have been conducted for all ARE motifs present in the database, using RNAplfold (122).

RNAplfold RNAplfold (122) (part of the Vienna RNA package (123)) is a thermodynamic RNA folding program that calculates local base-pairing probabilities, as well as the probability that stretches of nucleotides are unpaired, which is directly related to the energy needed to open all secondary structures in the respective stretch of nucleotides (opening energy) in cubic time (124). It is based on McCaskills dynamic programming algorithm for calculation of the partition function.
The RNAplfold approach allows to set a fixed window size to derive the average equilibrium base pairing probability over all possible sequence windows of this size. As part of the analysis of annotated ARE motifs was to calculate their accessibility to mark them as possible targets for AUBPs, RNAplfold was used to calculate the accessibility in form of probabilities of being unpaired or opening energy of these motifs.
This output has been integrated in the 'AREsite' webserver and presents a substantial contribution to the screening for novel AUBP targets. Accessibility extracted from RNAplfold has been used to generate a P-value that ranks annotated transcripts for their importance as AUBP binding sites, according to the P-value estimation described in section 4. The runtime parameters and the implementation of RNAplfold results in the web service are discussed in section 3.

2.2.3 Markov chains

An order-0 (for mononucleotides) and an order-1 (for dinucleotides) Markov chain were used during this thesis to calculate the fold enrichment of the ARE motifs of all transcripts in the database.
The fold-enrichment was calculated by extracting the frequency of mono- or dinucleotides in the annotated motifs, divided by their theoretically frequency which was derived from the Markov chains.
A Markov chain is a sequence X of variables x whose present state x_k is independent of the future x_{k+m} and past x_{k-m} states , see definition 8. Markov chains can be of order-0 to order-m, where m has to be finite.

$$P(x_k|x_1 \dots x_{k-m}) = P(x_k) \tag{8}$$

An order-0 Markov chain represents a sequence of states where the probability of the current state x_k is independent of all past states $x_1 \dots x_{k-1}$, see definition 9.

$$P(x_k|x_1 \dots x_{k-1}) = P(x_k) \tag{9}$$

For DNA or RNA sequences an order-0 Markov chain can be used to describe the probability of the sequence X of length L as product of the probabilities of the mononucleotides x found along this sequence, see equation 10.

$$P(X) = \prod_{k=1}^{L} Px_k \tag{10}$$

An example for an order-0 Markov chain for a sequence consisting of the characters A and B can be found in figure 23. The Markov chain defines probabilities p for a change from A to B, p_{AB} and vice versa p_{BA} and the probability for staying in the current state, defined as $1 - p_{AB}$ and $1 - p_{BA}$ respectively for a two state Markov model.

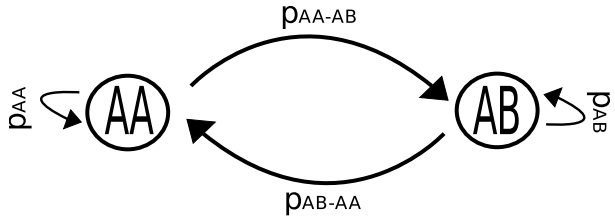


Fig. 23. An order-0 Markov model for a hypothetical sequence consisting of characters A's and B's, where p_{AB} describes the probability for a change from state A to state B and p_{BA} the probability for a change from state B to state A. p_A and p_B respectively is the probability to stay in the current state.

In an order-1 Markov chain (for definition see equation 11) the probability for a given state depends on the current state and the one state before, see 24.

$$P(x_k|x_1 \dots x_{k-2}) = P(x_k|x_{k-1} \times P(x_{k-1})) \tag{11}$$

This can be used to screen for statistically significant dependence of dinucleotide enrichment, by comparing the frequencies of state A ($P(A)$) and B ($P(B)$) derived from an order-0 and the frequencies of A following B and vice versa derived from an order-1 Markov model.

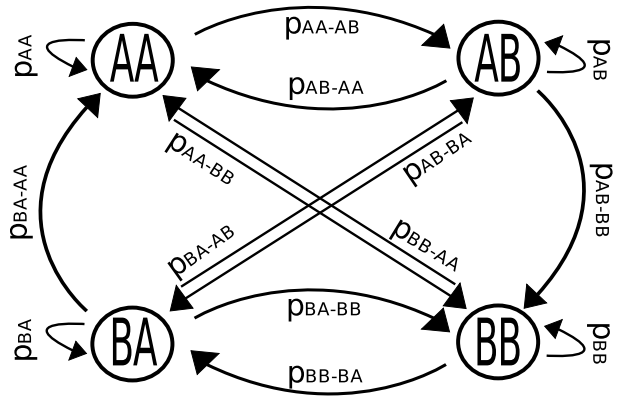


Fig. 24. An order-1 Markov model for a hypothetical sequence consisting of characters A's and B's, where p_{AA-AB} describes the probability to change from state AA to state AB and p_{AB-BB} the probability to change from state AB to state BB and so on. p_{AA}, p_{BB} and so on are the probabilities to stay in the current state.

Equation 12 shows an example where the probabilities for state A and B derived from an order-0 Markov model and the probabilities for A followed by B derived from an order-1 Markov Model are roughly the same and can be seen as independent, thus no statistical significant discrepancy between both models exists.

$$P(A) \times P(B) \cong P(A|B) \times P(B) \quad \forall A, B \tag{12}$$

If the probabilities are not independent, the occurrence of A followed by B shows significance. This was used to compare the fold-enrichment of ARE motifs in the 3' UTR of transcripts in our database based on mono- and dinucleotide probabilities.

2.2.4 Databases

The term database refers to an digital collection of data, managed by a database management system (DBMS) software. This chapter is adopted from the textbooks (125) and (126).
Databases evolved since the 1960s together with Database management systems (DBMSs). The first generation of databases were navigational, meaning that the user has to follow a given path to get to the database entry of interest. Today the most used database system follow the relational model. First proposed in the 1970s, a relational database allows to search for data by content rather

than paths.
The relational database management system (RDBMS) used in this thesis are MySQL and SQlite.MySQL is a free relational database system, running as a server, therefore providing multiple users access to the generated databases. As it is well documented and easy to handle, once the SQL(Structured Query Language) syntax is understood, this database system was chosen to store and manage the data retrieved from the Ensembl database. This data was analyzed and a flat-file database containing annotated transcripts amongst other things was built up as backend of the web server discussed later.
SQlite is an embedded relational database management system, which means that it does not provide a stand alone server version of its database. It uses a simplified SQL syntax and was chosen for the literature database backend of the web server, that is discussed in the next chapter.

3 The AREsite webserver

This section is based on the article (1) and passages taken from this article are used throughout this section without further notice.

Aim of this thesis is the analysis of human and mouse transcript 3'UTRs, for the presence or absence of ARE motifs that are potentially bound by AUBPs. Besides annotation of ARE motifs in these transcripts, analysis of their accessibility seemed a logical steps, due to the fact that AUBPs bind to single stranded RNA. As already mentioned, a hint for significance of AREs is their conservation among different species. Therefore analysis of ARE conservation was a second crucial step during this thesis.

Databases containing ARE motifs in human and other organisms are already available, see for example ARED (75) and its successors. However, these databases contain no additional information like the results of the above-mentioned analysis steps. To investigate as many types of ARE motifs as possible and analyze their potential function on mRNA stability and make the results of our analysis available to other researchers, the decision was to create a new database instead of using the present ones.

This has led to the generation of a websuite named AREsite. As first ARE-focused database, AREsite combines sequence annotation of AREs with the prediction of the accessibility and evolutionary conservation of the motif site. To circumvent restrictions in motif composition, a total of eight different consensus motifs, starting with the plain AUUUA pentamer up to the WWWWAUUUAWWWW 13-mer, which resembles the core motif embedded in a stretch of A/U residues, can be screened using this websuite. Information from extensive expert literature search has been incorporated into a second database that runs as backend for the wenservice and experimentally validated targets of the ARE-binding proteins TTP, HuR and Auf1 are listed.

The database has been published in Nucleic Acids Research (NAR) (1), and is accessible online at `http://rna.tbi.univie.ac.at/cgi-bin/AREsite.cgi` as part of the Vienna RNA websuite (127).

In cooperation with a research team working on tristetraprolin, a lot of data concerning the influcence of TTP on mRNA stability has been included into the database. HuR and AuF1 have been chosen as AUBPs due to the large amount of data, concerning their function and targets, that is freely accessible.

3.1 The developement of the 'AREsite' webserver

3.1.1 Generation of the database

Ensembl release 56 (`ftp://ftp.ensembl.org/pub/release-56/`) was used as data basis for AREsite in its current version. Following the description at `http://www.ensembl.org/info/docs/webcode/install/ensembl-data.html`, the Ensembl release 56 was embedded in the MySQL backhand of the websuite. Any protein coding gene with a transcript that contains at least one ARE motif in its 3'UTR has been added to the internal database.

Furthermore, the database contains all transcripts, one to one orthologous, to the human and mouse transcripts found in following species:

- Anolis carolinensis
- Bos taurus
- Callithrix jacchus
- Callithrix jacchus
- Canis familiaris
- Cavia porcellus
- Danio rerio
- Equus caballus
- Gasterosteus aculeatus
- Macaca mulatta
- Monodelphis domestica
- Ornithorhynchus anatinus
- Oryzias latipes
- Pan troglodytes
- Rattus norvegicus
- Sus scrofa
- Taeniopygia guttata
- Takifugu rubripes
- Tetraodon nigroviridis
- Xenopus tropicalis

Results of an extensive expert literature search where incorporated into a separate database. This database contains all publications concerning the effects of AUBPs on mRNA targets, classified by five criteria, including the results of our cooperating research team working on TTP:

- Direct binding of the protein to the mRNA or its 3' UTR has been shown
- An independent reporter assay confirmed the functionality of the putative binding site
- The loss or overexpression of the ARE-binding protein affects the level of the target mRNA
- The loss or overexpression of the ARE-binding protein affects the protein level of the target mRNA
- The stability of the target mRNA is affected by the lack or excess of the ARE-binding protein

This database can easily be updated as new (scientific) findings emerge, whereas updates of the main database require additional effort and computation time.

3.1.2 Generation of alignments from transcripts

Using data from the Ensembl gene orthology pipeline, alignments of orthologous transcripts were generated. For each gene database entry, all orthologous genes from other species that have a strict one to one relation where collected, followed by a screen for transcripts that have an annotated 3' UTR. Among those, the one that showed the best coverage (at least 75%) of the reference species 3' UTR where selected. Using CLUSTALW multiple species, whole transcript alignments were then generated.

Finally, the region containing the motif site plus five flanking nucleotides on each side from the alignments was extracted to investigate the sequence conservation of the motif site. Each alignment sequence was then searched with the corresponding consensus ARE motif. Using DIALIGN with the detected motifs as sequence anchors, the sequences were realigned. The same procedure was applied to the processed and filtered genomic alignments.

3.1.3 Generation of genomic alignments

Comparative data at the level of transcripts is still limited. Therefore data from genome-wide alignments was incorporated, to get a more refined picture of the conservation pattern of located motifs. Since there is no guarantee that the aligned sequences from other species really belong to the gene of interest, the interpretation of this data has to be done with caution. To circumvent this situation, filtering strategies have been introduced, that ensure that aligned sequences are homologous over a longer stretch of nucleotides than just the motif site.

The UCSC genome browser (113) provides genomic alignments in MAF format, generated using MULTIZ (112), that were obtained for each UTR sequence. Alignments, corresponding to human, were extracted from 46 species multiple alignments based on the human genome assembly hg19. Those corresponding to mouse were extracted from 30 species multiple alignments based on the assembly mm9. A MAF processing and filtering pipeline was developed, as the obtained alignment blocks were often too short for any practical use. The pipeline merges adjacent MAFblocks to longer ones and returns alignment windows of 120 nt and a step size of 30 nt. Using CLUSTALW, these alignments were realigned and filtered for sequences that have at least 50% of the sequence length of the reference species.

3.1.4 Quantifying motif site accessibility

RNAplfold (see section 2.2.2 for a description) was used for the calculation of the motif site accessibility in terms of opening energies and probabilities of being unpaired, as mentioned before. To model the effects of cotranscriptional folding, the parameter set W=80, L=40 m has been used. These parameters have previously been used to predict siRNA binding to mRNAs (128). Results with a different parameter setting (W=240, L=120), provided by the author of (128) are also included in AREsite. These settings consider longer base pair spans and show improved results on siRNA binding as well as on RNA - RNA interaction, and as AUBPs are single stranded RNA binding proteins and an improvement for the prediction of ARE binding protein targets is possible, these results were incoporated into the webservice, but are not included in this thesis.

3.2 AREsite output

The AREsite welcome page provides several choices for the user, see figure 25. First of all, a unique identifier of the gene or transcript of interest has to be inserted (here IL6 for the interleukin-6 precursor). Then the user can choose in which organism and to which type of ARE motif the analysis pipeline should be applied to (here homo sapiens and the ARE core motif AUUUA). By default, the analysis will only be conducted for the representing transcript, which is the one with the most ARE motifs in the 3' UTR.

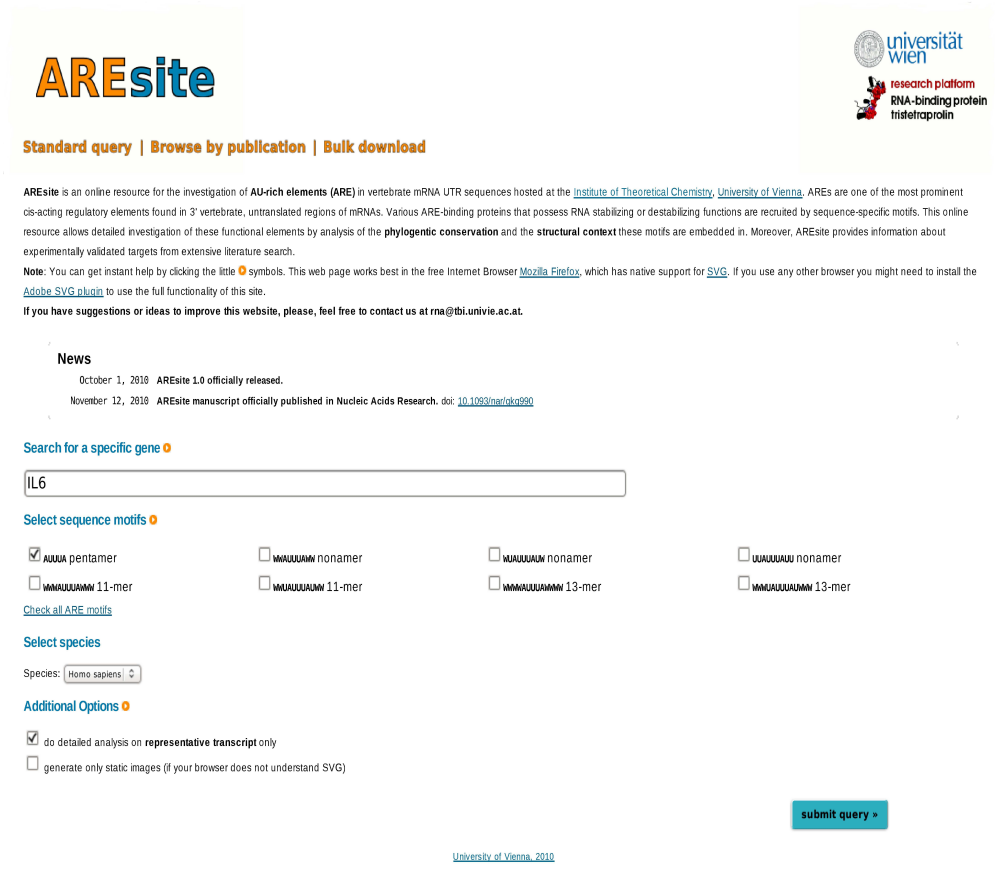


Fig. 25. At the AREsite welcome page, the user can input a unique identifier for the gene of interest and choose the ARE motif he wants to analyze and the organism to look at.

Alternatively the user can choose to conduct a literature search by clicking on "Browse for publication". This opens a window, see figure 26, where the user can choose a publication of interest, sorted by year, and will get a result page for the gene or transcript mentioned in this publication.

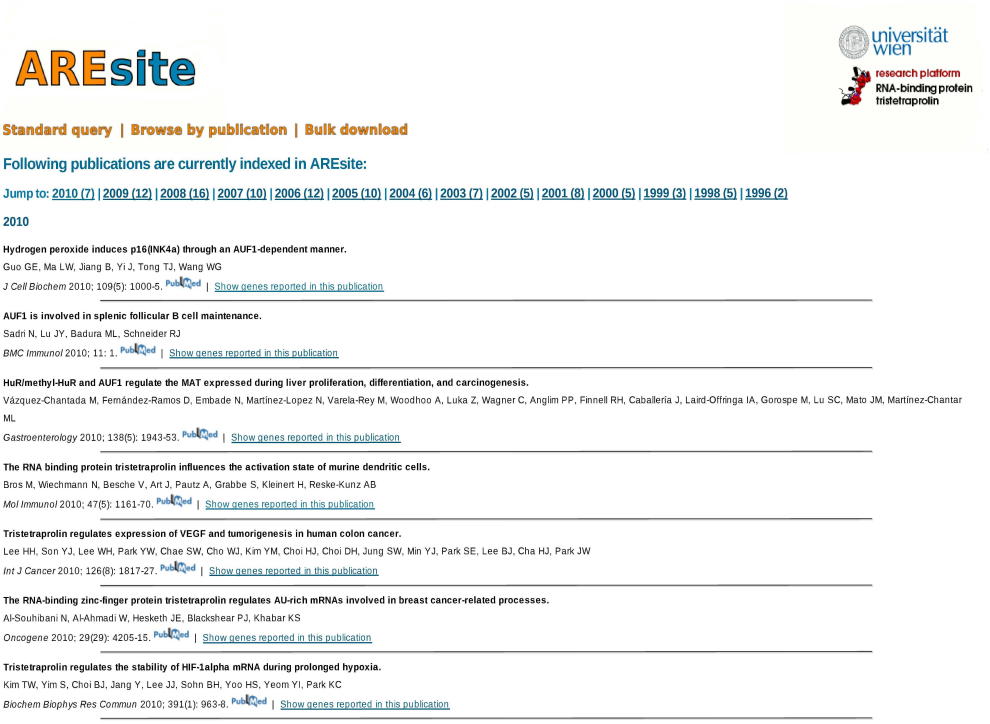


Fig. 26. At the "Browse by publication" page where, the user can conduct a literature search by choosing a publication of interest and will get a result page for the gene or transcript mentioned in this publication.

For users that are interested in all results concerning a certain ARE motif, a click on "Bulk download" opens a window, where all results for this motiv can be downloaded as .zip archive containing annotated Genbank files, see figure 27.

Fig. 27. At the "Bulk download" page, the user can download a .zip archive containing annotated Genbank files for all transcripts containing the ARE motif of interest.

Except for the bulk download, the user will be provided with a result page (in the given example for Interleukin-6) after finishing the input and submitting the query.

The first part of the result page, see figure 28, provides the user with information concerning the number of protein-coding transcripts corresponding to the submitted identifier, the representative transcript, the total number of distinct positions containing the selected motif and literature corresponding the effects of the motif.

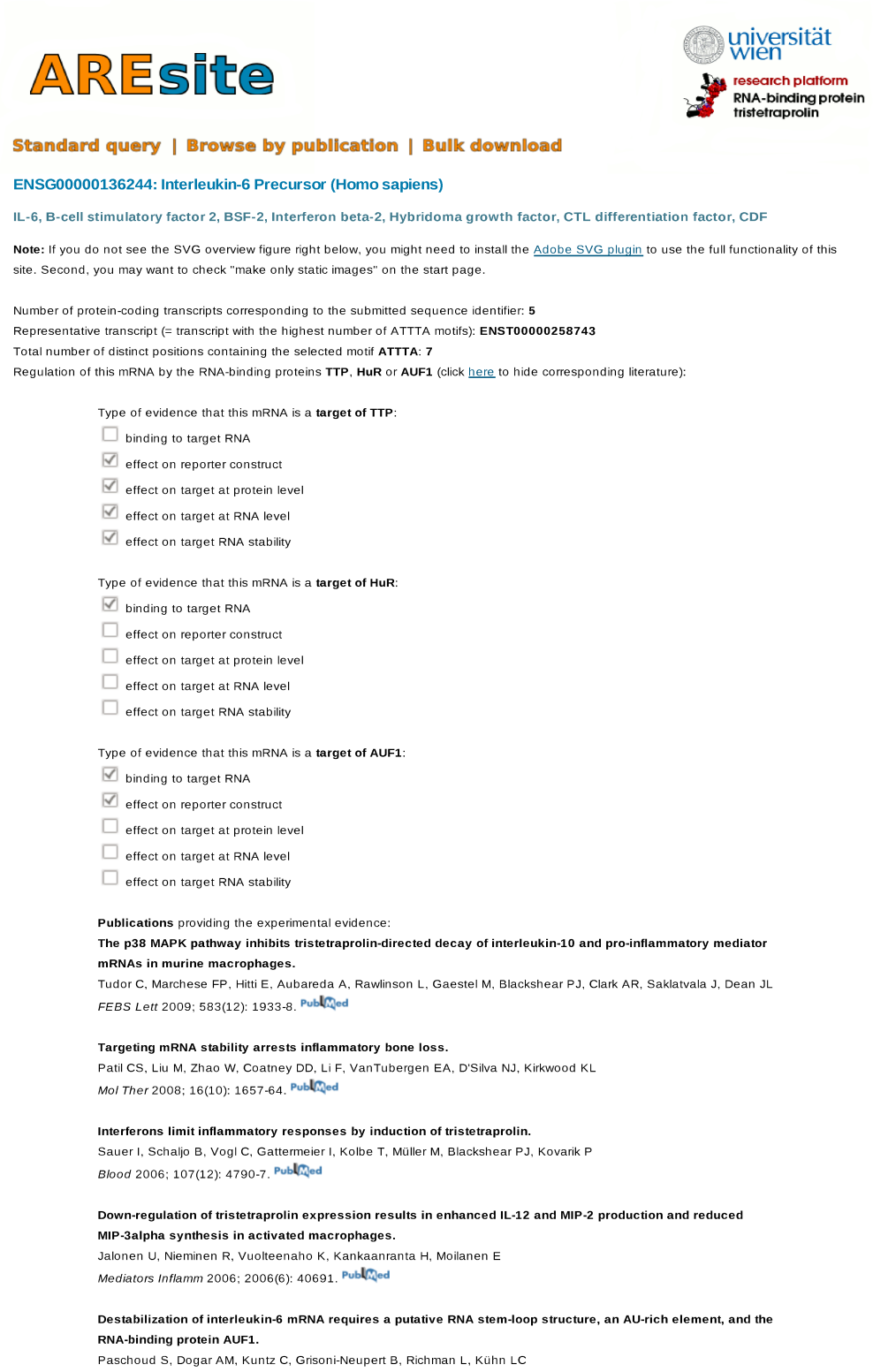
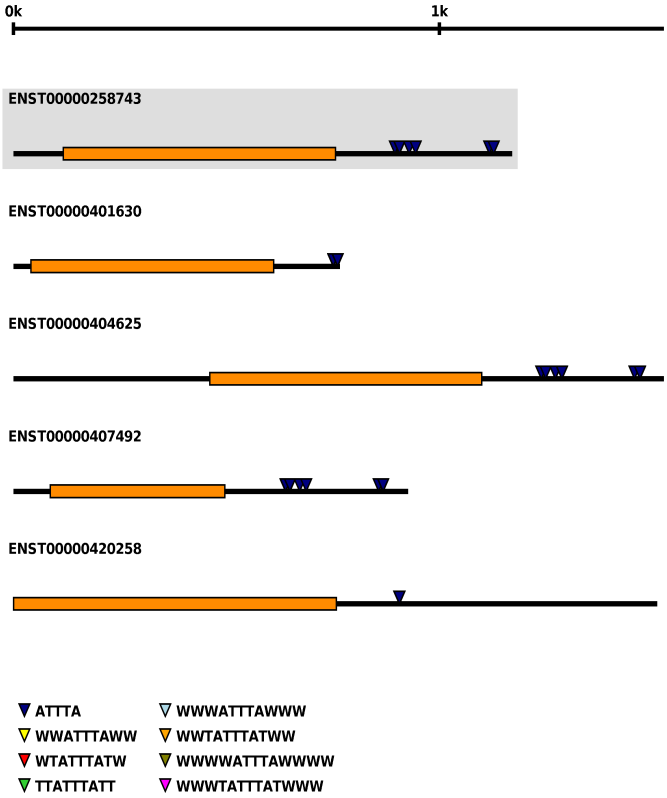


Fig. 28. The first part of the result page for Interleukin-6 shows the number of protein-coding transcripts corresponding to the submitted sequence identifier, five in this example. The representative transcript (= transcript with the highest number of ATTTA motifs), in this example ENST00000258743. The total number of distinct positions containing the selected motif ATTTA, seven in this example. Published evidence for regulation of this mRNA by the RNA-binding proteins TTP, HuR or AUF1, classified by the five criteria mentioned above and the corresponding literature.

The second part of the result page, see figure 29, begins with a plot. It shows the representing transcript in grey and all other transcripts related to the gene of interest and marks the positions of ARE motifs along the sequence. Below this plot is a simple statistics output, showing the length of the 3' UTR, the A/T content and the fold enrichment derived from the Markov models. Furthermore the user can have a look at the RNAplfold output for all transcripts and download the annotated Genbank file for the query. The last part of the figure shows detailed results for each ARE motif and the 3' UTR sequence, with the possibility to highlight ARE motifs and poly-A signals.

For each Motif a sequence logo representing the conservation of the motif, the RNAplfold output as accessibility plot and the alignments can be examined. The accessibility values (u=5) for the core AUUUA pentamer for both parameter settings (short range, mid range) are presented here too.



Transcript ENST00000258743 (representative transcript)

Length 3' UTR:	415 nt
A+T content in 3' UTR:	0.71
RNAplfold output:	[opening energies | probabilities of being unpaired] (whole transcript)
Download/Linkout:	[download as annotated Genbank file | Linkout to Ensembl]
ATTTA:	**2.32** (mononucelotide) / **112.14** (dinucleotide) fold-enrichment
Site 894-898:	ATTTA (ATTTA) Opening energy for the core AUUUA pentamer: **0.60** kcal/mol (short range) / **0.94** kcal/mol (mid range) Probability of being unpaired for the core AUUUA pentamer: **0.38** (short range) / **0.22** (mid range) [Highlight | show accessibility plot | show sequence logo | show alignment]
Site 904-908:	ATTTA (ATTTA) Opening energy for the core AUUUA pentamer: **0.79** kcal/mol (short range) / **1.74** kcal/mol (mid range) Probability of being unpaired for the core AUUUA pentamer: **0.28** (short range) / **0.06** (mid range) [Highlight | show accessibility plot | show sequence logo | show alignment]

3' UTR sequence

Highlight motifs in 3' UTR sequence: all ATTTA | polyA sites (AWTAAA) | none

```
756   CATGGGCACCTCAGATTGTTGTTGTTAATGGGCATTCCTTCTTCTGGTCAGAAACCTGTCCACTGGGCACAGAACTTATG
836   TTGTTCTCTATGGAGAACTAAAAGTATGAGCGTTAGGACACTATTTTAATTATTTTTAATTTATTAATATTTAAATATGT
916   GAAGCTGAGTTAATTTATGTAAGTCATATTTATATTTTTAAGAAGTACCACTTGAAACATTTTATGTATTAGTTTTGAAA
996   TAATAATGGAAAGTGGCTATGCAGTTTGAATATCCTTTGTTTCAGAGCCAGATCATTTCTTGGAAAGTGTAGGCTTACCT
1076  CAAATAAATGGCTAACTTATACATATTTTTAAAGAAATATTTATATTGTATTTATATAATGTATAAATGGTTTTTATACC
1156  AATAAATGGCATTTT
```

Fig. 29. The second part of the result page provides the user with a figure showing the positions of ARE motifs along the transcripts sequences, with the representing transcript highlighted in grey. Furthermore a simple statistic output is provided, presenting information on the length, the A/T content and the fold enrichment of the 3' UTR. The user can have a look at the RNAplfold output and download the annotated Genbank file for this query. A sequence logo representing the conservation of the motif, the RNAplfold output as accessibility plot and the alignments can be examined here.

Furthermore, results are visualized in an interactive SVG plot, see figure 30 that allows the user to explore different parameter settings (u = 1 to 13) for the RNAplfold output.

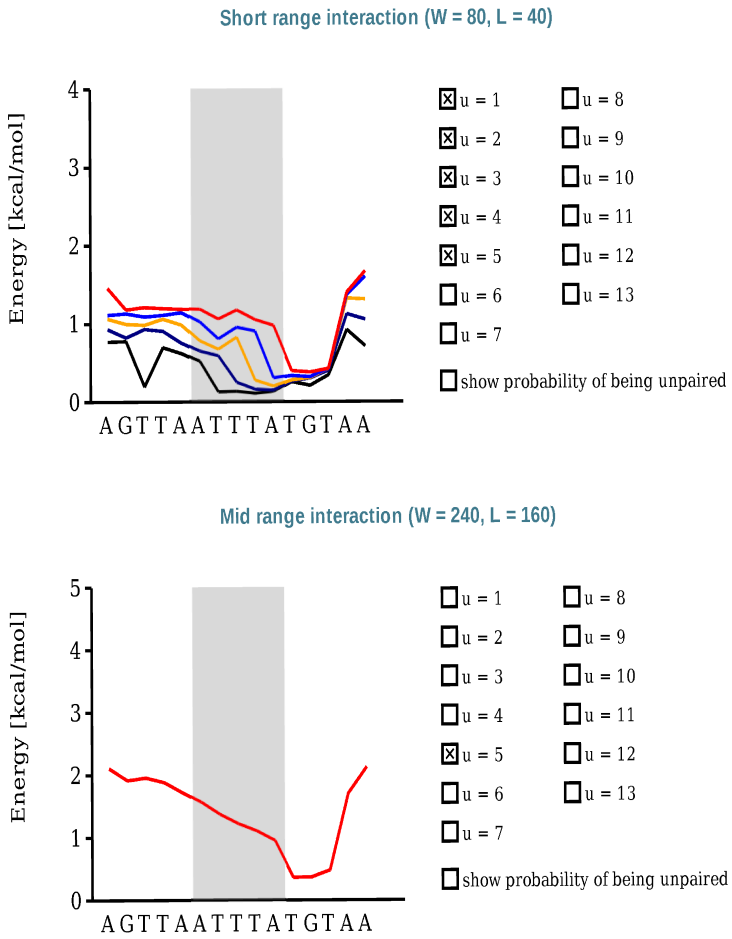


Fig. 30. Results are visualized in an interactive SVG plot, that allows the user to explore different parameter settings (u = 1 to 13) for the RNAplfold output

The last part of the result page shows a phylogenetic tree, see figure 31, where the conservation patterns of the detected ARE motifs are summarized.

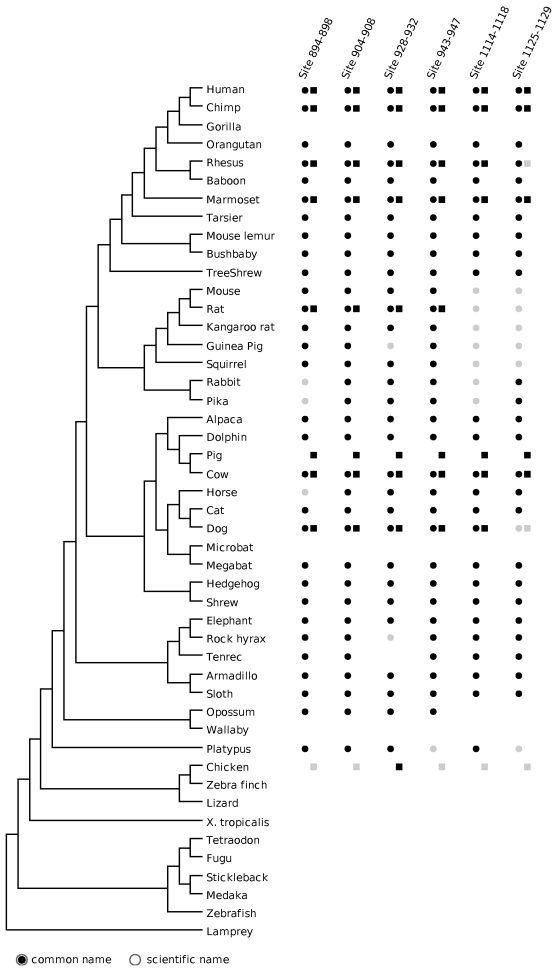


Fig. 31. This phylogenetic tree shows summarized the conservation pattern of the detected ARE motifs. Circles indicate that the motif has been found in genomic MAF alignments, while boxes indicate that the motif was found in transcript alignments. Signs in grey indicate that the sequence is present in the alignment, but the corresponding ARE pattern was not detected.

4 Methods

4.1 Analyzation of ARE motifs as AUBP target sites

In August 2010 Marín and Vanícek introduced a new method for the efficient use of accessibility for miRNA target prediction (129). MiRNAs bind single stranded, partially complementary, accessible regions of RNAs (see section 2.1.2). As the AUBPs investigated in this thesis are also known to bind single stranded RNA molecules, it appeared consequentially to use this new ranking method on the dataset that was generated during this thesis.
In nature two strategies guarantee that a regulatory active sequence motif is bound by the corresponding binding protein. On the one hand such a motif can be made accessible for a protein by embedding it into a structural context that presents the motif at the outside of an RNA structure where a protein can easily interact with it. On the other hand, if this is not possible (e.g. due to the high structuredness of a sequence), or more than one protein has to interact with a sequence motif at the same time in order to regulate the target mRNA or one protein has to bind to more than one target sites for full functionality, a strategy is to make multiple copies of this sequence motif available. This results in an over-representation of the target site in comparison to other motifs in the same sequence and is where the here discussed method can be used to rank predicted target sites. Instead of ranking target predictions according to hybridization energies or total free energies, like earlier approaches, the here presented algorithm ranks target sites according to their over-representation (fold-enrichment) and accessibility in terms of a P-value, which can be in the range from one (not over-represented) to zero (extremely over-represented). This method was called 'Prediction of Accessible MicroRNA Targets (PACMIT)' (129).
The algorithm computes a single-hypothesis P-value (P_{SH}) for each potential miRNA target site c, based on its over-representation, according to equation 13, where l is the length of the 3 'UTR, n is the number of nucleotides in the seed and P is the probability to find a given n-mer by chance at any particular position in the 3' UTR which is computed using an order-1 Markov model (introduced in section 2.2.3). The lower the P_{SH} value derived for a target site, the higher its over-representation. A low P-value means that the motif can be found more often than expected by chance, given the analyzed sequence, which indicates a functional role as mentioned previously.

$$P_{SH} = \sum_{i=c}^{l-n+1} \binom{l-n+1}{i} P^i (1-P)^{l-n+1-i} \tag{13}$$

This approach can be adopted to calculate a single-hypothesis p-value for accessible motifs only (P_{SHacc}), again the lower the P_{SHacc}, the higher is the chance of having found a functional target. Instead of counting all potential target sites c, only (partially) accessible target sites c_{access} are counted, and instead of taking the length l of the whole 3' UTR into account, the total number of accessible sites in the 3' UTR t_{access} is used (see equation 14). This leads to a list of accessible sequence motifs that are ranked according to their over-representation in the analyzed 3' UTR sequences.

$$P_{SHacc} = \sum_{i=c_{access}}^{t_{access}} \binom{t_{access}}{i} P^i (1-P)^{t_{access}-i} \tag{14}$$

To adopt this equation for the use with ARE motifs and AUBPs, certain changes were introduced into this approach. c_{access} is calculated by counting accessible sequence motifs in the 3' UTR, for example the "AUUUA" ARE core motif, with cutoffs of 0.5, 1, 2 and 3kcal/mol opening energy and without any cutoff. These cutoffs where estimations based on the mean accessibility of all motifs of length five in all transcripts (0.94013691 kcal/mol) and the accessibility distribution in known AUBP targets which can be seen in section 5.
t_{access} is derived by counting the total number of accessible nucleotides in the 3' UTR of a transcript, with and without the above mentioned cutoffs. Section 5 presents the PSHacc value distribution for known AUBP targets as well as a list of handpicked AUBP targets, that where ranked using this method. A discussion of the chosen cutoffs and this method as tool for the prediction of novel AUBP binding sides follows in section 6.2.

5 Results

5.1 General

During this thesis the webserver 'AREsite' was created, as described in section 3. This webserver contains annotated ARE motifs in transcripts extracted from the Ensembl database. However, the reason for the generation of the database that works as backend for the webserver was not just to present the annotated motifs, but to act as tool for the analysis of experimentally validated and in the following the prediction of novel AUBP targets that have not yet been examined experimentally.
The webserver provides information on the accessibility and fold-enrichment of annotated ARE motifs. As the fold-enrichment gives clues about whether a certain motif is expected by chance in a 3' UTR or not, it can be used as indicator of a functional role. The higher the enrichment the higher the chance of having found a regulative active motif. This approach is of course limited in its applicability if a 3' UTR is very AU rich and even a high fold-enrichment is no guarantee for functionality, thus it can only be used as additional information for a screen.
Accessibility prediction has been used in combination with wet-lab data to successfully predict RNA binding protein target sites (see e.g. (130; 131)) . This information is included for all ARE motifs in all annotated transcripts and has been used as filtering method for the prediction of regulatory motifs. However, the applicability of accessibility as screening method is limited, as a low accessibility does not exclude a sequence motif from being a target for binding proteins, and high accessibility does not necessarily imply regulative function. This section presents the finding, that accessibility alone can not be used as filtering technique during a screen for regulative active ARE core motifs.
Section 4 introduced a method that uses information on over-representation of accessible target sites in form of the PSHacc-value for miRNA target prediction. This method has been applied, with already discussed changes, to the generated dataset, as the AUBPs covered in this thesis are like miRNAs known to bind single stranded RNA motifs. Its application did indeed lead to better predictions of regulative active ARE motifs compared to the usage of accessibility as only filtering criteria.

The main goal of this thesis was to find methods that allow to extract ARE motifs with regulatory function from the set of annotated motifs. This was accomplished by comparison of the distributions of motif accessibility in terms of opening energies in the transcripts coupled with an analysis of their PSHacc value distributions, which acts as indicator for the over-representation of the analyzed motifs. To see at which cutoffs the PSHacc value or the accessibility of a motif give clues about the importance of an ARE motif as regulatory element, known AUBP targets where extracted from the dataset by screening the publication database, that runs as second backend of the 'AREsite' webserver, for known targets of the AUBPs AuF1, HuR and TTP. This dataset of known targets and "random" pentamers (which were produced for each 3' UTR specific by using a sliding window approach that extracted all pentamers of the given 3' UTR sequence) was used to compare the accessibility and PSHacc value distribution for known targets, new annotated transcripts and other ("random") pentamers in transcripts containing the ARE core motif, to extract cutoffs that allow to distinguish regulatory active core motifs from the rest.
The following sections presents the results of the ARE motif analysis, conducted according to the P-value estimation introduced by Marín and Vanícek (see section 4), together with the results of the motif annotation, accessibility and motif enrichment calculations included in the webserver 'AREsite'. Although many analysis steps have been accomplished with the DNA sequences that code for the ARE containing transcripts, the ARE core motif is presented here as 'AUUUA' and not 'ATTTA'.

5.2 Results with annotated human transcripts

The 'AREsite' webserver contains annotated transcripts for human and mouse. This thesis discusses only results for the human genome, as the principle workflow remains unchanged for an analysis of the mouse genome and the amount of known targets in mouse is far below the amount of known targets in human. Thus it is very hard to extract parameters that allow a distinction between regulatory functional and non-functional ARE motifs in the mouse dataset.

5.2.1 Whole human transcript set analysis

The Ensembl database release 56 contains 20,469 known protein coding genes for the human genome. This section presents an overview of the total number of genes and transcripts, containing ARE core motifs at certain opening energy cutoffs to get a first impression on the available dataset.
12,298 protein coding genes have been annotated to have a 3' UTR containing the 'AUUUA' ARE core motif with an accessibility in terms of opening energy of up to 3kcal/mol. This corresponds to $\sim$ 60% of the protein coding genes in the Ensembl database.
8,941 genes have been annotated to have a 3' UTR containing the 'AUUUA' ARE core motif with an accessibility in terms of opening energy of up to 0.5kcal/mol, which still corresponds to $\sim$ 44% of the protein coding genes in the Ensembl database and 1514 of those genes are known targets of AUBP according to literature included in the webserver 'ARESite'.

The generated flat-file database that acts as backend for the 'ARESite' webserver contains the following entries for the human genome:

- 20,469 genes coding for proteins (number of annotated genes in the Ensembl database)
- Genes coding for proteins with an annotated 3' UTR
  - 19,692 in total
  - Genes coding for proteins with an annotated 3' UTR containing the 'AUUUA' core motif
    - 12,298 with an opening energy of up to 3kcal/mol
    - 12,237 with an opening energy of up to 2kcal/mol
    - 11,570 with an opening energy of up to 1kcal/mol
    - 8,941 with an opening energy of up to 0.5kcal/mol
- Transcripts with an annotated 3' UTR
  - 66,969 in total
  - Transcripts with an annotated 3' UTR containing the 'AUUUA' core motif
    - 28,769 with an opening energy of up to 3kcal/mol
    - 28,528 with an opening energy of up to 2kcal/mol
    - 26,803 with an opening energy of up to 1kcal/mol
    - 19,855 with an opening energy of up to 0.5kcal/mol

However, the presence of an accessible ARE motif does not necessarily mean that a given gene is under expression control of this motif or the according AUBP respectively.
As can be seen the total numbers of transcripts with annotated motifs is comparatively large, so very restrictive cutoffs are necessary to filter for significant hits without loosing too many possible targets.

The next sections presents methods that were used to get cutoffs that fulfill both requirements. It includes a distribution analysis of opening energies and over-representation in terms of the calculated PSHacc values for annotated transcripts and a detailed analysis of known AUBP targets, including some AUBPs themselves, as they have been shown to act autoregulatory and ends with predicted AUPB targets according to their PSHacc values at given cutoffs and a summary of the presented results.

Comparison of opening energy distribution of the 'AUUUA' ARE core motif in known AUBP targets and in all annotated transcripts

To get a hint on a useful cutoff for the accessibility of motifs, an analysis of the opening energy distribution of the ARE core motif 'AUUUA' in known AUBP targets and in transcripts annotated for the 'ARE site' webserver was conducted.
Figure 32 shows almost no differences in the opening energy distribution of ARE motifs in known targets and newly annotated ARE motifs.

Although the total number of 'AUUUA' motifs in known targets is with 8.729 only a small part of the 124.393 annotated motifs in all transcripts, their opening energy distribution differs only marginally. If the opening energy would be taken as the only measure for the regulatory function of an ARE motif, this would mean that every annotated motif must be classified as an active site for AUBP binding. As this is very unlikely, opening energy alone has to be considered as a poor method for AUBP target prediction.
Nonetheless the comparison of opening energy distributions of all pentamers in an ARE containing transcript was compared to the opening energy of the 'AUUUA' ARE core motif, to see whether or not the core motif is more accessible than other pentamers in the same 3' UTR of a transcript. The next paragraph presents the results of this analysis.

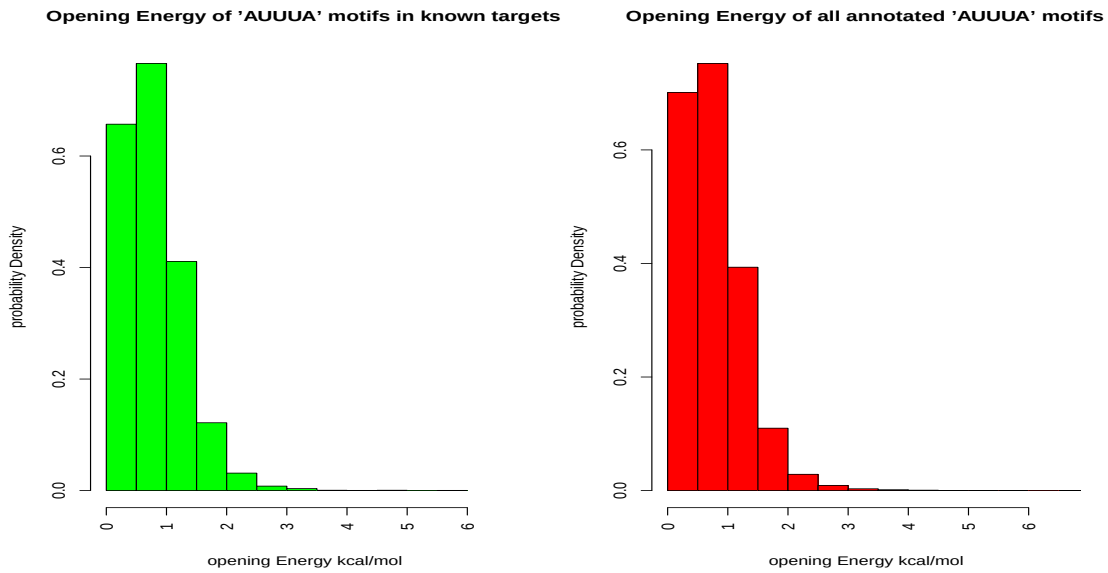


Fig. 32. A plot of the opening energy distribution in transcripts containing the 'AUUUA' ARE core motif that are known targets of AUBPs (green) and of newly annotated ARE core motifs (red).

Analysis of the opening energy distribution of 'AUUUA' compared to "random" pentamers in the same 3' UTR

Figure 33 shows a histogram of the opening energy distribution for all pentamers and for the 'AUUUA' ARE core motif of annotated transcripts in the database. As the opening energy cutoff directly influences the count of motifs that are characterized as accessible, this analysis step was conducted to condense the dataset for a detailed analysis without discriminating ARE motifs that were found in the 3' UTR of annotated transcripts more than necessary.

It can be seen that the opening energy of the 'AUUUA' ARE core motif is lower or equal to one in most cases. The opening energy of other pentamers is in most cases above one. As the relative frequency of 'AUUUA's with an opening energy of ~ 0.5kcal/mol is higher than for the other pentamers, this value has been chosen as a cutoff for the analysis of the PSHacc value distribution in ARE motifs from length seven to thirteen (shown later), as without this very restrictive cutoff the amount of data would have been too large to get results in an acceptable time scale.
For the other analysis steps an opening energy cutoff of 3kcal/mol was taken,

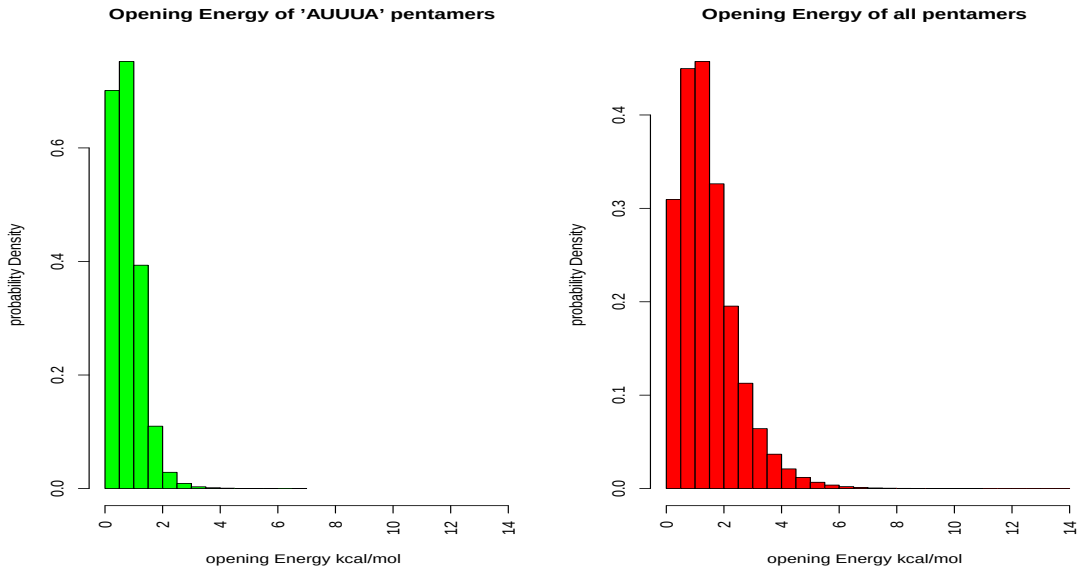


Fig. 33. This histogram shows the opening energy distribution for other pentamers (green) and for the 'AUUUA' ARE core motif (red) of transcripts in the database.

which excludes almost no annotated ARE motifs and reduces the size of the "random" motif dataset at the same time to a quantity that allows a detailed analysis according to the P-value estimation, which is later on used to filter for novel AUBP targets.

Comparison of the PSHacc value distribution of the 'AUUUA' ARE core motif in all annotated transcripts and in known AUBP targets to the distribution of "random" pentamers

The first step was again to compare the 'AUUUA' PSHacc value distribution in known AUBP targets with the distribution in all annotated transcripts, to screen for a possible cutoff that excludes ARE motifs that are not targeted by AUBPS. Figure 34 shows the PSHacc value distribution in known AUBP targets and the ARE core motifs 'AUUUA'.

As can be seen in figure 34, the PSHacc value for the ARE core motif in known targets shows a similar distribution pattern than the PSHacc value of 'AUUUA' in all annotated transcripts. In both cases, we can find a peak at low PSHacc values which shows that both, known targets as well as annotated transcripts, contain over-represented 'AUUUA' ARE core motifs in their 3' UTR. This finding indicates that over-representation is a possible filtering technique

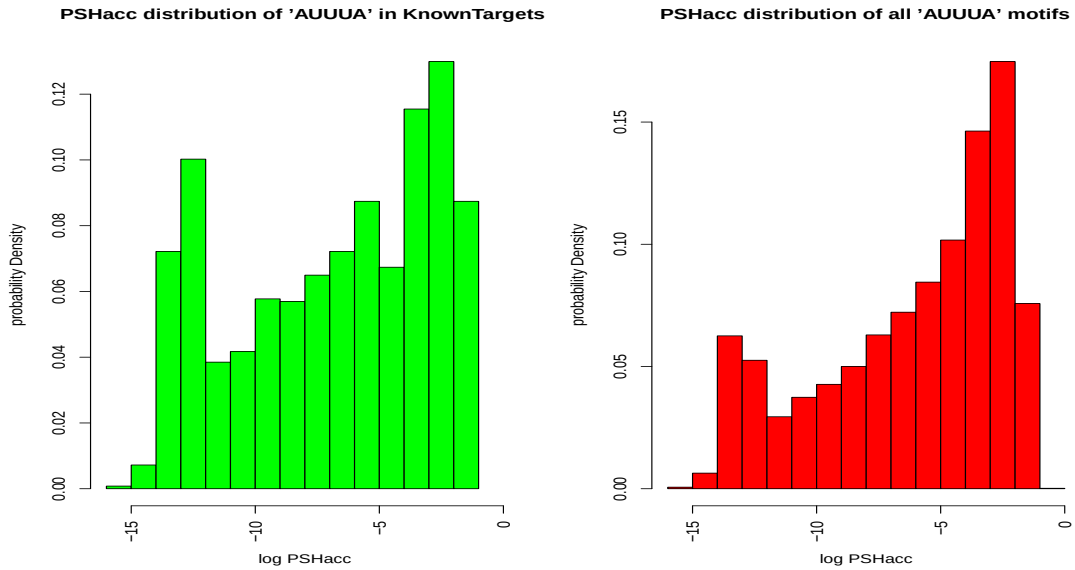


Fig. 34. The $PSHacc$ distribution of the ARE core motif 'AUUUA' in known targets compared to the distribution in all annotated transcripts.

when screening for regulatory active ARE motifs.
Although known targets contain 'AUUUA' motifs that are not over-represented in their 3' UTR as well, which means that the ARE core motif does not necessarily have to be over-represented to act as binding site of AUBPs and further analysis steps are necessary to characterize annotated ARE motifs of this type. Over all, known targets contain more motifs with log PSHacc values in a range between -5 and -10, and a peak at low PSHacc values shows that regulative active ARE motifs that are over-represented are enriched. To see whether low PSHacc values are descriptive for the ARE core motif or if "random" pentamers show a comparable distribution, a comparison with the PSHacc value distribution of other pentamers in the annotated transcripts has been conducted, as can be seen in figure 35.

Comparing both distributions it can be seen that the ARE core motif shows a higher tendency to be over-represented in a 3' UTR than the "random" pentamers. When looking at the PSHacc value distribution for other pentamers in annotated transcripts (see figure 35), it can be seen that some "random" pentamers with low PSHacc values exist, which means that over-represented pentamers exist throughout the annotated transcripts. An analysis of this region shows that the AU content in motifs found here is about 0.65, which

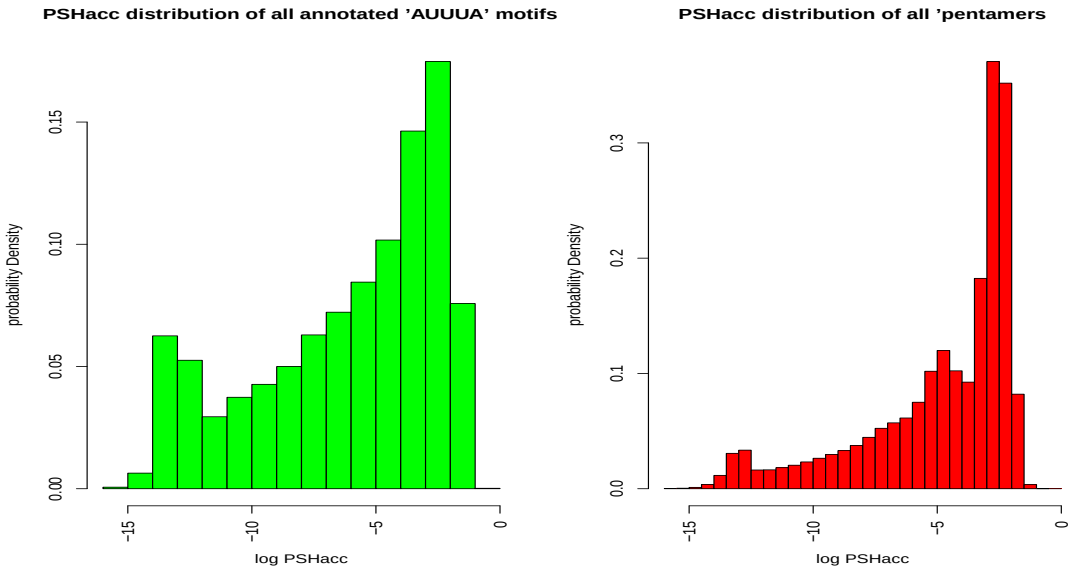


Fig. 35. The $PSHacc$ distribution of the ARE core motif 'AUUUA' compared to the distribution of "random" pentamers in annotated transcripts.

indicates that the peak at about 10^{-13} exists due to a high number of AU-rich elements with PSHacc values between 10^{-13} and 10^{-14}. Visual interpretation of this region indeed shows the presence of many AU-rich elements, which can partially be identified as motifs that usually occur in the flanking regions of the ARE core motif. This leads to the conclusion, that the presence of ARE core motifs with low PSHacc values, indicating their over-representation given the analyzed 3' UTR, can indeed be used as a ranking method for the categorization of regulatory active ARE core motifs.
A detailed analysis of ARE core motif flanking regions has not yet been done, but would be very interesting for the analysis of the previously described peak in the P-value distribution. However, the next section presents PSHacc value distributions in ARE motifs from length seven to thirteen compared to the ARE core motif of same length, but with different flanking regions, in the annotated 3' UTR with an opening energy cutoff of 0.5kcal/mol, due to the previously mentioned high amount of data.

Analysis of the PSHacc value distribution in transcripts containing ARE motifs of length seven to thirteen with an opening energy cutoff of 0.5kcal/mol

The following figures (36, 37, 38, 39) show the PSHacc value distribution for all seven to thirteen nucleotide long ARE motif containing transcripts in the

database with an opening energy cutoff of 0.5kcal/mol. The distribution shown in the figures compares the ARE motifs with one to four flanking 'W' nucleotides (W stands for T or A) and all other motifs containing the ARE core motif 'AUUUA' flanked by one to four 'N' nucleotides (N stands for A,T,C or G).

Figure 36 shows the PSHacc value distribution for motifs containing the 'WAUUUAW' ARE motif compared to 'NAUUUAN' heptamers.
The relative peak for the ARE motif can be found above a log PSHAcc value of -4, whereas the peak for the heptamers is slightly below this value. For a majority of heptamers, the PSHacc value estimation predicts a higher over-representation than for the ARE motifs, but with decreasing PSHacc, the distribution shows a higher relative frequency of ARE heptamers, which means that heptamers with a low PSHacc value are preferentially 'WAUUUAW' ARE motifs.

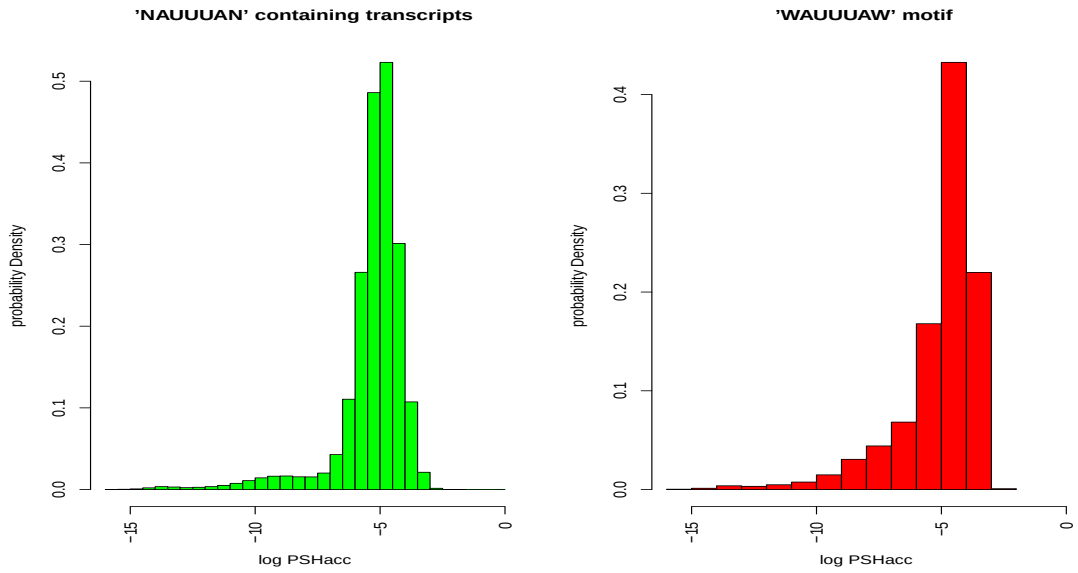


Fig. 36. A plot of the mean PSHacc value of all heptamers in transcripts containing the 'NAUUUAN' ARE motif (green) and the PSHacc values of the 'WAUUUAW' motifs (red) with a cutoff of 0.5kcal/mol opening energy.

Figure 37 shows the PSHacc value distribution for motifs containing the 'WWAUUUAWW' ARE motif compared to the 'NNAUUUANN' nonamers.
Particularly striking is that almost no motifs with a log PSHacc value above -5 can be seen in the distribution. Again the majority of nonamers has its peak at a lower PSHacc value than the ARE nonamer motifs, but with decreasing PSHacc, the distribution shows again a higher relative frequency of ARE nonamers, which means that similar to heptamers, nonamers with a low PSHacc value are preferentially 'WWAUUUAWW' ARE motifs.

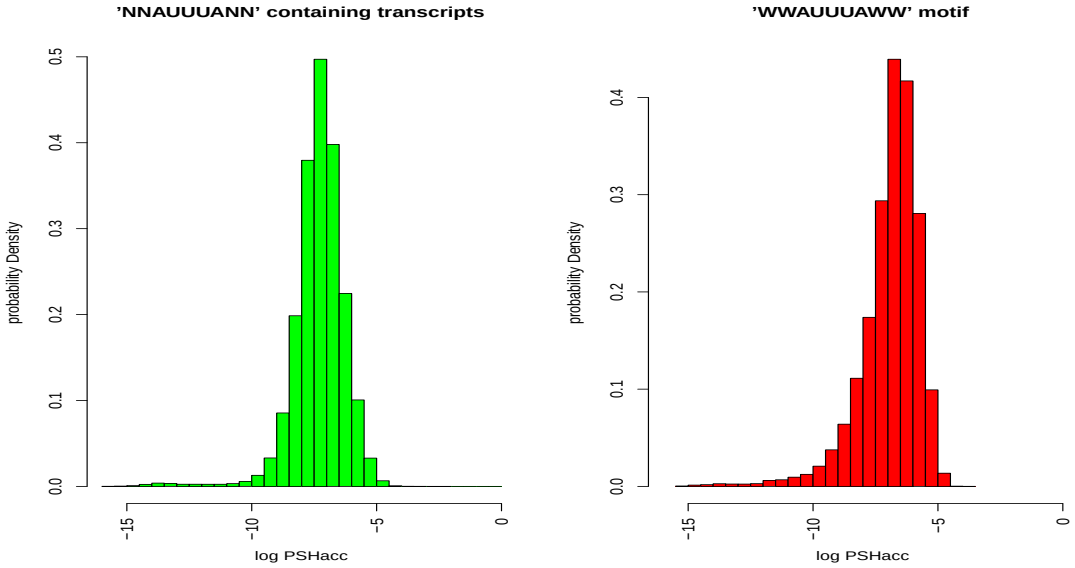


Fig. 37. A plot of the mean PSHacc value of all nonamers in transcripts containing a 'NNAUUUANN' ARE motif (green) and the PSHacc values of the 'WWAUUUAWW' motifs (red) with a cutoff of 0.5kcal/mol opening energy.

Figure 38 shows the log PSHacc value distribution for motifs containing the 'WWWAUUUAWWW' ARE motif compared to the 'NNNAUUUANNN' undecamers.
No motifs with a log PSHacc value above -6 can be seen and the striking finding in this figure is that a higher number of ARE undecamer motifs with a low PSHacc value compared to other undecamers can be seen, whereas the peak is in both cases at a log PSHacc value of below -8, indicating a high overrepresentation for both kinds of motifs.

Figure 39 shows the log PSHacc value distribution for motifs containing the 'WWWWAUUUAWWWW' ARE motif compared to the 'NNNNAUUUANNNN'

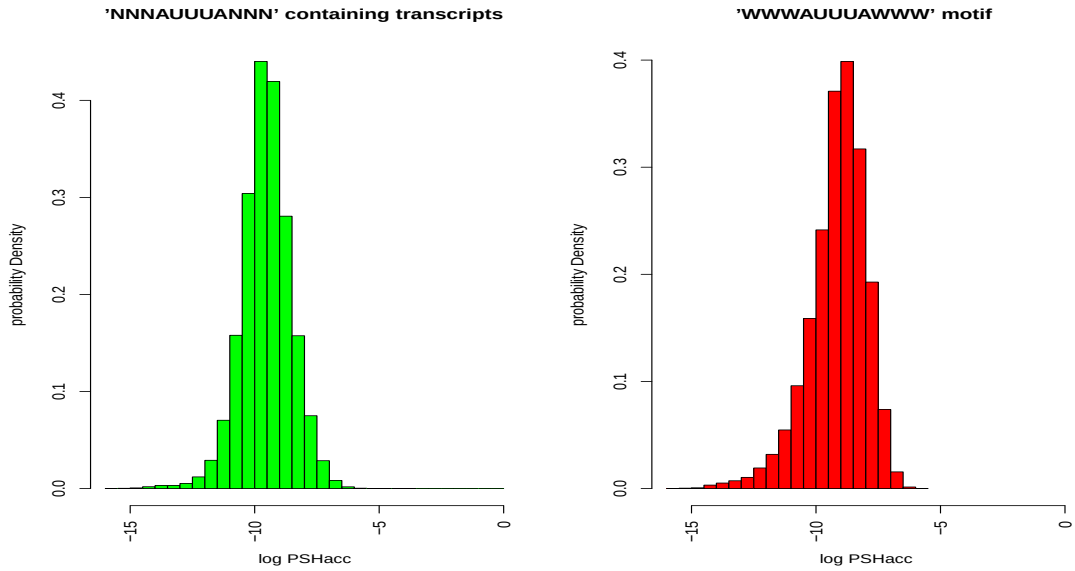


Fig. 38. A plot of the mean PSHacc value of all undecamers in transcripts containing a 'NNNAUUUANNN' ARE motif (green) and the PSHacc values of the 'WWWAUUUAWWW' motifs (red) with a cutoff of 0.5kcal/mol opening energy.

tridecamers.
It is particularly noticeable that motifs of length thirteen are in both cases almost equally distributed around their peak. However, the peak for ARE motifs is again at a higher log PSHacc value than that for the other tridecamers and the relative frequency of ARE motifs at decreasing PSHacc is for the first time higher for the other tridecamers, showing a higher over-representation of "random" tridecamers than ARE motifs.

Summing up the observations of these distributions, ARE motifs, up to the length of eleven, are more over-represented than the other motifs, although the distribution peak is usually found at a lower PSHacc for the other motifs. This means that strong over-representation can be seen as a feature of ARE motifs, although this seems to change at a length of thirteen nucleotides. As ARE motifs are usually characterized as multiple copies of the ARE core motif 'AUUUA' in close proximity, if seems more promising to focus on the over-representation of shorter ARE motifs than extending one ARE sequence above a length of eleven.

To cautiously set this in relation to the biological role of ARE motifs, it could indicate that ARE motifs, if present in a transcript but not used for AUBP binding, have to be less over-represented than other motifs, or that preceding

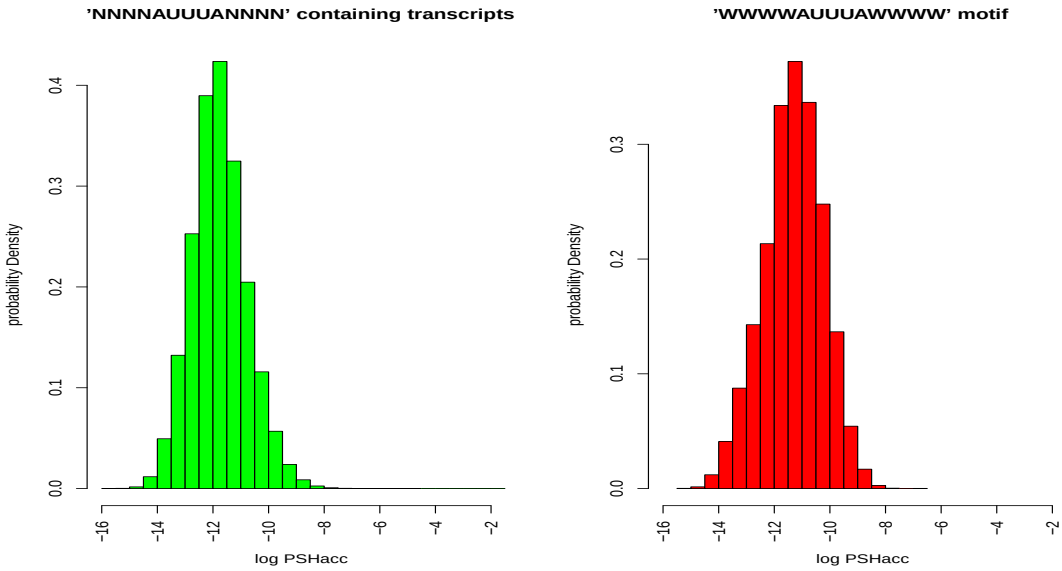


Fig. 39. A plot of the mean PSHacc value of all tridecamers in transcripts containing a 'NNNNAUUUANNNN' ARE motif (green) and the PSHacc values of the 'WWWWAUUUAWWWW' motifs (red) with a cutoff of 0.5kcal/mol opening energy.

regulatory steps are necessary to make these ARE motifs accessible for their binding proteins, which would decrease their PSHacc value. However, a detailed analysis of ARE motif flanking regions in known targets and annotated transcripts remains to be done and is an inevitable task to get more contextual information about the motifs and possible influences on their regulative roles, as the analysis presented here shows no differences that are strong enough to be used as filtering techniques.

5.2.2 Analysis of the PSHacc value distribution in handpicked AUBP targets

Some known targets of the AUBPs HuR, TTP and AuF1 were analyzed in detail to test if these targets would have been found when over-representation in form of the PSHacc value is used as filter method.

The following section presents the PSHacc value distribution of pentamers including the 'AUUUA' ARE core motif in the 3' UTR of the representing transcripts (= transcript with the highest number of AUUUA motifs) in annotated targets, together with information that is part of the 'AREsite' webserver.

TNF-α

Tumor necrosis factor alpha (TNF, cachexin, cachectin, TNF-α) is a cytokine involved in systemic inflammation. Its role as a key mediator of inflammation is well known and TNF-α plays a crucial role in the early phase of the host response to bacterial, viral and parasitic infections (132). High levels of TNF-α due to a systemic release can lead to vascular decompensation and death(133). As described in (134) the TNF-α ARE motif is a very strong motif, meaning that it triggers a dramatic decrease of protein level and has therefore been chosen for this detailed analysis.
According to the webserver 'AREsite', the TNF-α representing transcript ENST00000449264 contains a 3' UTR of 799nt length with an AU content of 0.53 and following ARE motifs:

Motif	**Count**	**Mono- and Dinucleotide fold-enrichment**	
AUUUA	9	8.35	213.77
WWAUUUAWW	5	1,780.03	5,338,794.22
WUAUUUAUW	5	59.63	765.52
UUAUUUAUU	5	401,086.32	104,233,968.85
WWWAUUUAWWW	5	213.54	613.06
WWUAUUUAUWW	5	7,830,760.57	56,851,050.94
WWWWAUUUAWWWW	4	764.72	2,195.41
WWWUAUUUAUWWW	4	34,732,964.59	1,109,953,002.63

It can be seen that the TNf-α mRNA contains multiple AUBP binding sites in forms of 'AUUUA' to 'WWWUAUUUAUWWW'. The dinucleotide based fold enrichment for this motifs shows that they are not to be expected by pure chance in this 3' UTR, which indicates a functional role.
Furthermore 'AREsite' lists experimental evidence that this mRNA is a target of two AUBPs according to eight hits in its publication database backend (86; 135; 136; 137; 138; 139; 140; 87).
Type of evidence that this mRNA is a target of TTP:

- Direct binding of the protein to the mRNA or its 3' UTR has been shown
- An independent reporter assay confirmed the functionality of the putative binding site

- The loss or overexpression of the ARE-binding protein affects the level of the target mRNA
- The loss or overexpression of the ARE-binding protein affects the protein level of the target mRNA
- The stability of the target mRNA is affected by the lack or excess of the ARE-binding protein

Type of evidence that this mRNA is a target of HuR:

- Direct binding of the protein to the mRNA or its 3' UTR has been shown
- An independent reporter assay confirmed the functionality of the putative binding site

To see if the PSHacc value can be used to identify the ARE core motif 'AUUUA' in this AUBP target as functional, the distribution of PSHacc values of all pentamers in the 3' UTR has been analyzed. Following Graph representation (figure 40) shows this distribution in the TNF-α representing transcript ENST00000449264 with a cutoff for the PSHacc value at 10^{-12} for better visualization of motifs with low PSHacc's.

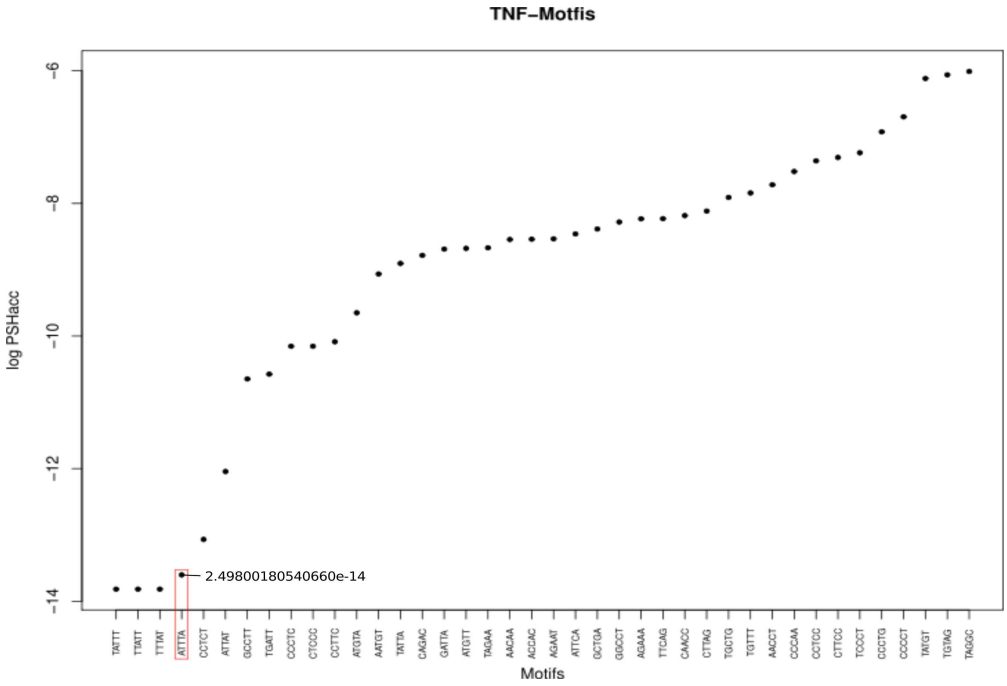


Fig. 40. The $PSHacc$ distribution of all pentamers in the TNF-α representing transcript ENST00000449264 with a PSHacc cutoff of 10^{-12} for better visualization. The 'AUUUA' ARE core motif has been highlighted with a red frame, its PSHacc value has been added to the plot.

As can be seen in figure 40, the 'AUUUA' core motif is among the pentamers with a very low PSHacc value, even though it has not the best PSHacc of all motifs. But motifs with comparative PSHacc value are 'AU' rich as well ('UAUUU','UUAUU' and 'UUUAU') and may present flanking regions of the ARE core motif. In this example, the PSHacc value estimation presents four 'AU' rich motifs in the top five pentamers, indicating a regulatory role for these motifs, which is consistent to the experimental evidence listed for this AUBP target. This can be seen as positive example for target prediction via PSHacc values.

HuR

HuR is one of the three AUBPs discussed in detail in this thesis, please refer to section 2.1.4 for a detailed description. It is known to act on ARE motifs in its own 3' UTR, thereby autoregulating its expression and has therefore been chosen as target for a more detailed analysis.
The HuR representing transcript ENST00000351593 contains a 3' UTR of 4.909nt length with an AU content of 0.56 and following ARE motifs as presented by 'AREsite':

Motif	**Count**	**Mono- and Dinucleotide fold-enrichment**	
AUUUA	13	1.40	11.42
WWAUUUAWW	3	243.12	342,254.09
WUAUUUAUW	2	3.27	3.46
UUAUUUAUU	1	14,141.80	47,837.04
WWWAUUUAWWW	1	7.05	10.98
WWWWAUUUAWWWW	1	191,891.60	485,450.69

The HuR mRNA contains thirteen AUBP binding sites in forms of 'AUUUA' which is an extraordinary high number. Although not as high as in TNF-α, the dinucleotide based fold enrichment for these motifs indicates clearly that they are not to be expected by pure chance in this 3' UTR and indicate a functional role.
Furthermore 'AREsite' lists experimental evidence that this mRNA is a target of two AUBPs according to two hits in its publication database backend (141; 142).

Type of evidence that this mRNA is a target of HuR:

- Direct binding of the protein to the mRNA or its 3' UTR has been shown

Type of evidence that this mRNA is a target of AUF1:

- Direct binding of the protein to the mRNA or its 3' UTR has been shown
- The loss or overexpression of the ARE-binding protein affects the protein level of the target mRNA

To see if the PSHacc value can be used to rank the ARE core motif 'AUUUA' in this AUBP target, the distribution of PSHacc values of all pentamers in the 3' UTR has been analyzed. The Graph representation (figure 41) shows this distribution for all pentamers in the HuR representing transcript ENST00000351593, with a cutoff for the PSHacc value at 10^{-12} for better visualization of motifs with low PSHacc's.

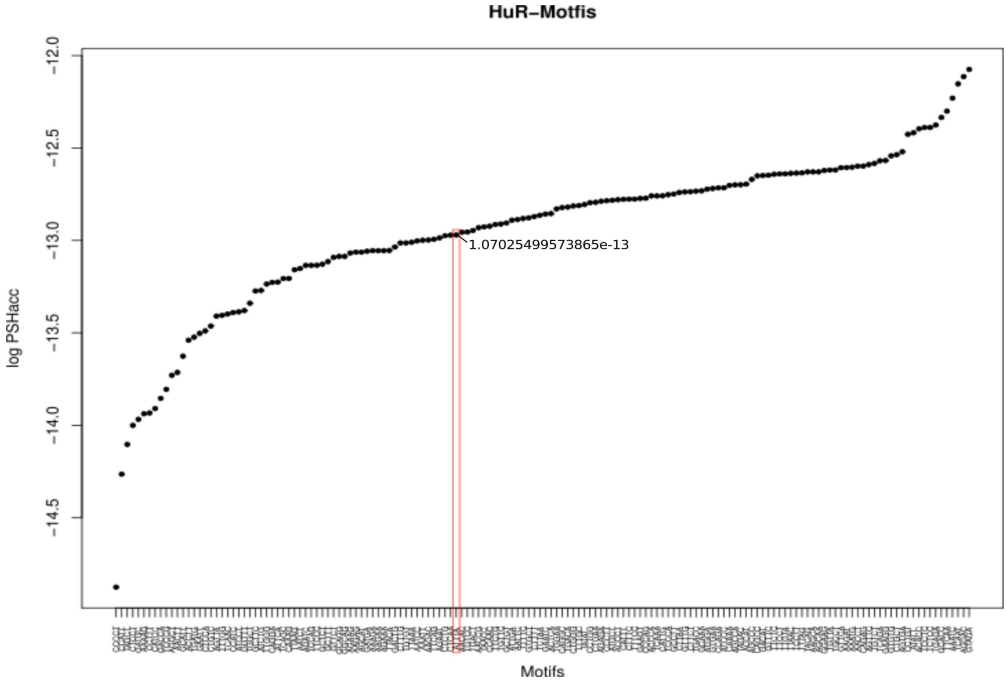


Fig. 41. The $PSHacc$ distribution of all pentamers in the HuR representing transcript ENST00000351593 with a PSHacc cutoff of 10^{-12} for better visualization. The 'AUUUA' ARE core motif has been highlighted with a red frame, its PSHacc value has been added to the plot. Although the ARE core motif if not among the top ten over-represented motifs, it is among the 8% of motifs with a PSHacc value lower or equal to 10^{-13}.

Figure 41 shows that the 'AUUUA' core motif is among the pentamers with a very low PSHacc value, but it is not in the top ten of motifs with the lowest

PSHacc. GC containing motifs can be found here, some of them with better PSHacc values than the ARE core motif, which is not what would be expected considering the high number of ARE core motifs in the 3' UTR. However, HuR is known to autoregulate its own expression in combination with Auf1, thus preceding steps may be necessary to make the ARE motif accessible for HuR, as strong regulation is expected for autoregulatory processes. Competitive binding of more than one AUBP at once would also explain the high number of ARE core motifs found in this 3' UTR. In fact it has been shown that HuR requires single stranded ARE motifs to stabilize its target mRNA, and that HuR can bind to more than one ARE motif in a length dependent manner simultaneously (83; 94). It would also be interesting to have a look at the effects of Auf1 binding on the mRNA structure to see whether or not preceding binding of Auf1 is amplifying or weakening interactions of HuR on its own mRNA, as will be discussed in section refsec:accessibility. This example shows that the PSHacc value alone is not enough to classify ARE motifs in AUBP targets under strong regulation as functionally active.

IL6

IL-6 is a cytokine that regulates the development of the nervous and hematopoietic systems, acute-phase responses, inflammation, immune responses and other biological processes (143). According to (144) ARE motifs present in the 3' UTR of IL6 can be found in proximity of an AU-rich regulatory stem-loop region and are both required for the decay of the mRNA. Some, but not all of the present ARE motifs seem to influence the regulatory function of the stem-loop structure which was a reason to analyze this AUBP target in more detail.
According to the webserver 'AREsite', the IL6 representing transcript ENST00000258743 contains a 3' UTR of 415nt length with an AU content of 0.71 and following ARE motifs:

Motif	Count	Mono- and Dinucleotide fold-enrichment	
AUUUA	7	2.32	8.79
WWAUUUAWW	4	112.14	6,749.40
WUAUUUAUW	2	5.96	9.55
UUAUUUAUU	1	2,392.96	25,254.22
WWWAUUUAWWW	3	9.72	17.29
WWUAUUUAUWW	1	13,430.60	25,382.47
WWWWAUUUAWWWW	3	11.42	18.77
WWWUAUUUAUWWW	1	342,254.09	94,973.53

The IL6 mRNA contains multiple AUBP binding sites in forms of 'AUUUA' to 'WWWUAUUUAUWWW'. The dinucleotide based fold enrichment for these motifs it by far not as high as in TNF-α, but shows they are not to be expected by pure chance in this 3' UTR, which again indicates a functional role.
Furthermore 'AREsite' lists experimental evidence that this mRNA is a target of all three AUBPs discussed in this thesis, according to eight hits in its publication database backend (145; 146; 144; 147; 148; 138; 140; 149).
Type of evidence that this mRNA is a target of TTP:

- An independent reporter assay confirmed the functionality of the putative binding site
- The loss or overexpression of the ARE-binding protein affects the level of the target mRNA
- The loss or overexpression of the ARE-binding protein affects the protein level of the target mRNA
- The stability of the target mRNA is affected by the lack or excess of the ARE-binding protein

Type of evidence that this mRNA is a target of HuR:

- Direct binding of the protein to the mRNA or its 3' UTR has been shown

Type of evidence that this mRNA is a target of AUF1:

- Direct binding of the protein to the mRNA or its 3' UTR has been shown
- An independent reporter assay confirmed the functionality of the putative binding site

To see if the PSHacc value can be used to characterize the ARE core motif 'AUUUA' in this AUBP target as regulatory relevant, the distribution of PSHacc values of all pentamers in the 3' UTR has been analyzed. Following Graph representation (figure 42) shows this distribution for all pentamers in the IL6 representing transcript ENST00000258743, with a cutoff for the PSHacc value at 10^{-6} for better visualization of motifs with low PSHacc's.

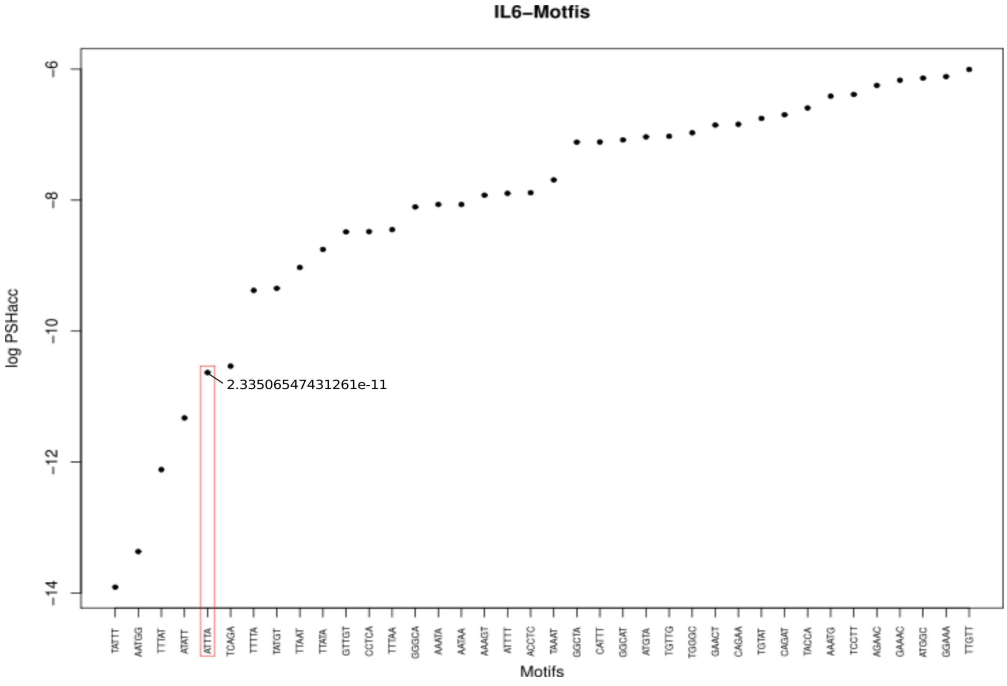


Fig. 42. The $PSHacc$ distribution of all pentamers in the IL6 representing transcript ENST00000258743 with a PSHacc cutoff of 10^{-6} for better visualization. The 'AUUUA' ARE core motif has been highlighted with a red frame, its PSHacc value has been added to the plot.

The ARE core motif 'AUUUA' shown in figure 42 is among the pentamers with outstanding PSHacc values. Among the top ten ranked motifs, seven AU-rich elements (UAUUU, UUUAU, AUAUU, AUUUA, UUUUA, UUAAU, UUAUA) can be found. As the previously mentioned regulatory stem-loop that has been identified in this transcript is AU-rich as well, ranking by PSHacc value may have led to the identification of the ARE core motif, flanking regions and AU-

rich motifs in the stem-loop structure as regulatory elements, but this needs further investigation. This is a nice example for a successful filtering of regulatory motifs in a known AUBP target.

Bcl-2

Bcl-2 is an apoptosis regulating protein. The overexpression of Bcl-2 plays a role in multiple cancers and is associated with resistance to chemotherapy. For the Bcl-2 ARE motifs an antagonistic effect of AuF1 (destabilizing) and Nucleolin (stabilizing) bonding has been reported (150) and it was therefore analyzed in detail.
The Bcl-2 representing transcript ENST00000333681 contains a 3' UTR of 5.279nt length with an AU content of 0.59 and following ARE motifs as presented by 'AREsite':

Motif	**Count**	**Mono- and Dinucleotide fold-enrichment**	
AUUUA	13	1.05	1.83
WWAUUUAWW	2	159.77	14,333.45
WUAUUUAUW	1	1.29	7.14
UUAUUUAUU	1	3,441.16	355,334.10
WWWAUUUAWWW	1	2.52	5.18
WWUAUUUAUWW	1	42,654.12	119406.12
WWWWAUUUAWWWW	1	9.86	20.22
WWWUAUUUAUWWW	1	228,502.30	2,960,143.71

It can be seen that the Bcl2 mRNA contains 13 AUBP binding sites in forms of 'AUUUA', which is similar to what was shown for the HuR mRNA previously. The dinucleotide based fold enrichment for this motif may not be as high as for other AUBP targets, but it is sufficient to show that they are not to be expected by pure chance in this 3' UTR, which is a hint on a functional role of this motifs.
Furthermore 'AREsite' lists experimental evidence that this mRNA is a target of AUF1, according to two hits in its publication database backend (151; 152).
Type of evidence that this mRNA is a target of AUF1:

- Direct binding of the protein to the mRNA or its 3' UTR has been shown
- An independent reporter assay confirmed the functionality of the putative binding site
- The loss or overexpression of the ARE-binding protein affects the protein level of the target mRNA

The distribution of PSHacc values of all pentamers in the 3' UTR has been analyzed to see if the PSHacc value can be used to rank the ARE core motif 'AUUUA' in this AUBP target. The Graph representation in figure 43 shows this distribution for all pentamers in the Bcl2 representing transcript ENST00000333681, with a cutoff for the PSHacc value at 10^{-12} for better visualization of motifs with low PSHacc's.

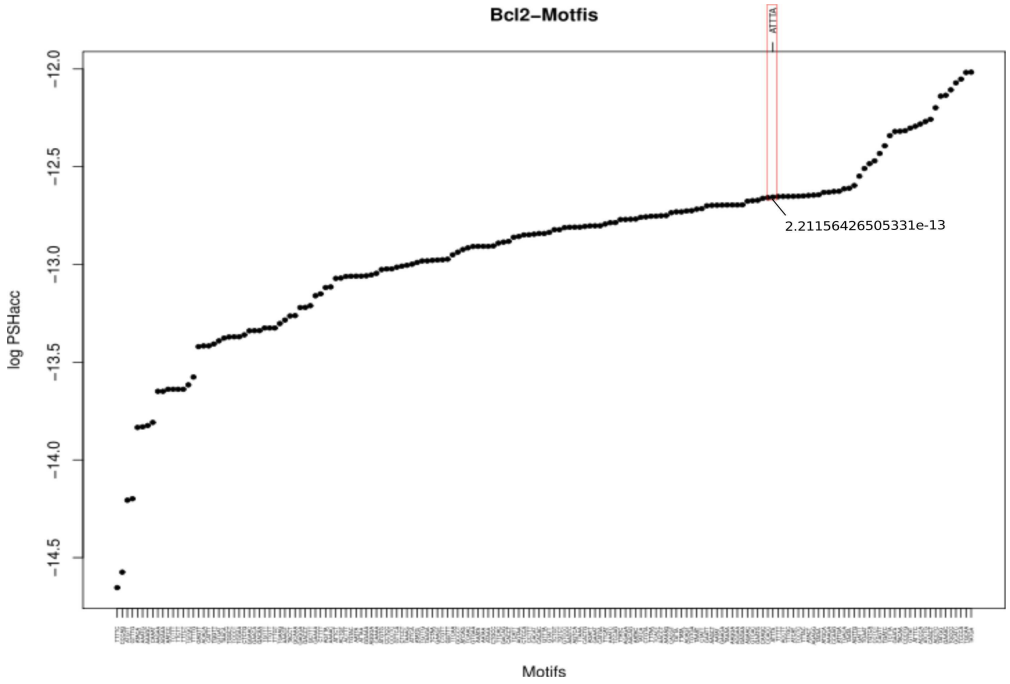


Fig. 43. The $PSHacc$ distribution of all pentamers in the Bcl2 representing transcript ENST00000333681 with a PSHacc cutoff of 10^{-12} for better visualization. The 'AUUUA' ARE core motif has been highlighted with a red frame, its PSHacc value has been added to the plot. Although the ARE core motif can be found in the top ten over-represented motifs, it is among the 17% of motifs with a PSHacc value lower or equal to $10^{-12.5}$

Figure 43 shows that the 'AUUUA' core motif is not among the pentamers with a very low PSHacc value, although some other A+U rich motifs can be found there. Bcl2 is known to be regulated by AuF1 and Nucleoline, so preceding steps may be necessary to make the ARE motif site accessible for binding pro-

teins, as competitive regulation includes binding of more than one protein to the target mRNA. As for HuR, competitive binding of more than one RNA binding protein at once would explain the high number of ARE core motifs found in this 3' UTR and may even be used as indicator for transcripts that are under control of more than one AUBP.
Although in the two shown cases, especially where high competitive binding to an mRNA target occurs, not only ARE motifs are over-represented, the PSHacc value calculation has been successfully applied as filtering method for the prediction of known AUBP targets. To see if the prediction of novel AUBP targets using this method is possible, the next section presents an analysis of transcripts that were chosen, with the exception of TTP, because they show to an extraordinary low PSHacc value for the ARE core motif and no literature concerning their regulation by AUBPs is present in the 'AREsite' literature database.

5.2.3 Analysis of handpicked AUBP targets according to their PSHacc values

TTP

TTP is one of the three AUBPs already described in this thesis, please refer to section 2.1.4 for a detailed description. It is known to act as tumor repressor with targets like the TNF-α mRNA (79) and shows self regulatory funtion by binding to an ARE motif in its own 3' UTR (85) and was therefore chosen for a more detailed analysis. This AUBP is listed in the section of novel AUBP targets, as currently no reference for experimental evidence for its function on its own 3' UTR is listed by 'AREsite', but ARE core motifs with very low PSHacc have been found.
According to the webserver 'AREsite', the TTP representing transcript ENST00000248673 contains a 3' UTR of 674nt length with an AU content of 0.56 and following ARE motifs:

Motif	**Count**	**Mono- and Dinucleotide fold-enrichment**	
AUUUA	5	3.95	25.86
WWAUUUAWW	1	522.08	234,137.40
WUAUUUAUW	1	7.98	83.78
UUAUUUAUU	1	16,837.29	1,465,349.70

The TTP mRNA contains only 5 AUBP binding sites in forms of the 'AUUUA' core motif, but the dinucleotide based fold enrichment for this motif highlights that it is not to be expected by pure chance in this 3' UTR, which is a hint on a functional role.
In its current state, AREsite lists no entries for publications that provide experimental evidence for effects of an AUBP on this target. However, literature exists (see e.g. (85)) and will be added to a future release of the webserver.
The distribution of PSHacc values of all pentamers in the 3' UTR of this transcript have been analyzed to see if the ARE core motif is the only pentamer with a low PSHacc value or if this can be seen as common for pentamers in this transcript. The Graph representation in figure 44 shows this distribution for all pentamers in the TTP representing transcript ENST00000248673, with a cutoff for the PSHacc value at 10^{-12} for better visualization of motifs with low PSHacc's.

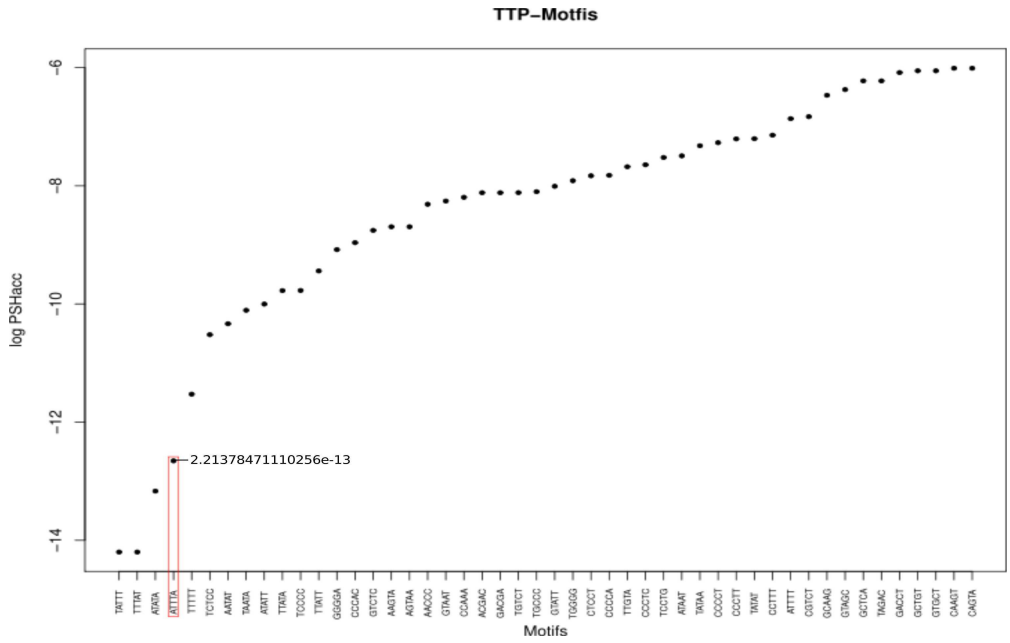


Fig. 44. The $PSHacc$ distribution of all pentamers in the TTP representing transcript ENST00000248673 with a PSHacc value cutoff at 10^{-12} for better visualization. The 'AUUUA' ARE core motif has been highlighted with a red frame, its PSHacc value has been added to the plot.

Figure 44 shows that the 'AUUUA' core motif is among the top four pentamers with a very low PSHacc value. Although the A+U content in the 3' UTR is only 0.56, which is similar to the A+U content in HuR, almost only A+U rich pentamers can be found found among the top ten motifs with very low PSHacc values (UAUUU, UUUAU, AUAUA, AUUUA, UUUUU, UCUCC, AAUAU, UAAUA, AUAUU, UUAUA).

TTP is like HuR known to autoregulate its own expression and contains two zinc finger motifs that can bind their target simultaneously, which means that more than one ARE motif has to be accessible at the same time. Five 'AUUUA' ARE motifs have been annotated for this transcript, so if two or three of them are not well accessible per se this would influence the PSHacc value only marginal, but could in fact be enough to control the autoregulatory effect.

SUMO-specific protease SENP1

The SUMO deconjugation enzymes (SENPs) play an important role in regulation of protein activity by actively regulation their state of SUMOylation. Accoring to (153), SUMO (small ubiquitin-related modifier) proteins are ap-

proximately 10-kD polypeptides that function as reversible post-translational protein modifiers. Sumoylation alters the molecular interactions of modified target proteins by masking or adding interaction surfaces, that can lead to changes in localization, activity and protein stability. Xu et al. (154) shows that SENP1 influences colon cancer cell growth.
Due to it's extraordinary low PSHacc value and a comparably high number of ARE motifs, SENP1 has been chosen for a more detailed analysis.
The webserver 'AREsite', the SENP1 representing transcript ENST00000004980 contains a 3' UTR of 2.424nt length with an AU content of 0.60 and following ARE motifs as presented by 'AREsite':

Motif	**Count**	**Mono- and Dinucleotide fold-enrichment**	
AUUUA	12	1.54	178.08

It can be seen that the SENP1 mRNA contains twelve AUBP binding site in form of the 'AUUUA' ARE core motif which is almost as much as in HuR or Bcl-2. The dinucleotide based fold enrichment for this motif is pretty low compared to the known AUBP targets discussed so far, indicating that the motif may be expected by pure chance in this 3' UTR, which relativizes the occurence of twelve core motifs in reference to their functional role.
In its current state, AREsite lists no entries for publications that provide experimental evidence for effects of an AUBP on this target, and so far no literature concerning these effects has been found.
The distribution of PSHacc values of all pentamers in the 3' UTR of this transcript have been analyzed to see if the ARE core motif is the only pentamer with a low PSHacc value or if this can be seen as common for pentamers in this transcript. Following Graph representation (figure 45) shows this distribution for all pentamers in the SENP1 representing transcript ENST00000004980, with a cutoff for the PSHacc value at 10^{-9} for better visualization of motifs with low PSHacc's.

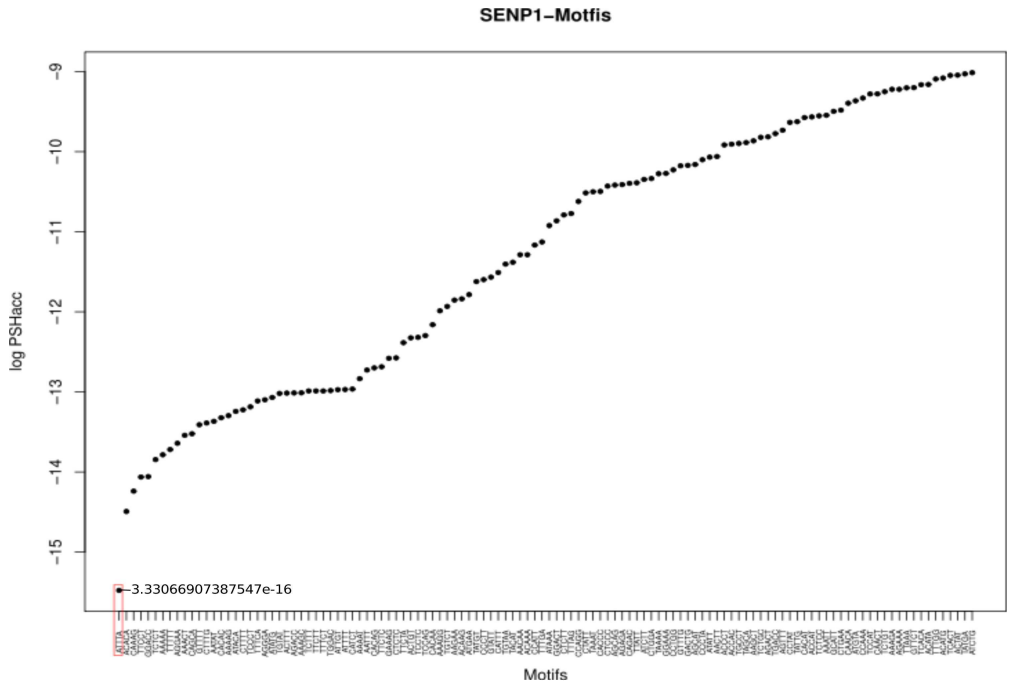


Fig. 45. The $PSHacc$ distribution of all pentamers in the SUMO1 representing transcript ENST00000004980 with a PSHacc value cutoff at 10^{-9} for better visualization. The 'AUUUA' ARE core motif has been highlighted with a red frame, its PSHacc value has been added to the plot.

Figure 45 shows that the 'AUUUA' core motif is the pentamer with the lowest PSHacc value found in this transcript, which was not to be expected according to the relatively low fold-enrichment. Although the A+U content in the 3' UTR is relatively high with 0.69, C+G rich motifs can be found among the top ten motifs with very low PSHacc values (AUUUA, ACACA, CAAAG, UUCCU, GGACC, UCUCU, AAAAA, UUUUU, AGGAA, AAACU).
As so far no experimental evidence for the regulation of SENP1 mRNA via AUBPs exists, this mRNA would represent a good target for further investigation in this direction.

MAGUK p55 subfamily member 7

MAGUK p55 subfamily member 7 is involved in the assembly of protein complexes at sites of cell-cell contact and acts as an important adapter that promotes epithelial cell polarity and tight junction formation via its interaction with DLG1 (155). Due to it's extraordinary low PSHacc value and a high number of ARE motifs, MAGUK has been chosen for a more detailed analysis.

According to the webserver 'AREsite', the MAGUK p55 subfamily member 7 representing transcript ENST00000337532 contains a 3' UTR of 3.072nt length with an AU content of 0.66 and following ARE motifs:

Motif	**Count**	**Mono- and Dinucleotide fold-enrichment**	
AUUUA	13	1.03	2.79
WWAUUUAWW	4	94.19	6,687.58
WWWAUUUAWWW	3	1.64	6.36
WWWWAUUUAWWWW	3	1,671.35	35,678.60

The MAGUK p55 subfamily member 7 mRNA contains thirteen AUBP binding sites in form of the 'AUUUA' ARE core motif, again a very high number, similar to HuR and Bcl-2. The dinucleotide based fold enrichment for this motif indicates that the motif is not to be expected by pure chance in this 3' UTR, which gives a direct hint on a functional role.
In its current state, AREsite lists no entries for publications that provide experimental evidence for effects of an AUBP on this target, and so far no literature concerning these effects has been found.
The distribution of PSHacc values of all pentamers in the 3' UTR of this transcript have been analyzed to see if the ARE core motif is the only pentamer with a low PSHacc value or if this can be seen as common for pentamers in this transcript. The Graph representation in figure 46 shows this distribution for all pentamers in the MAGUK p55 subfamily member 7 transcript ENST00000337532, with a cutoff for the PSHacc value at 10^{-12} for better visualization of motifs with low PSHacc's.

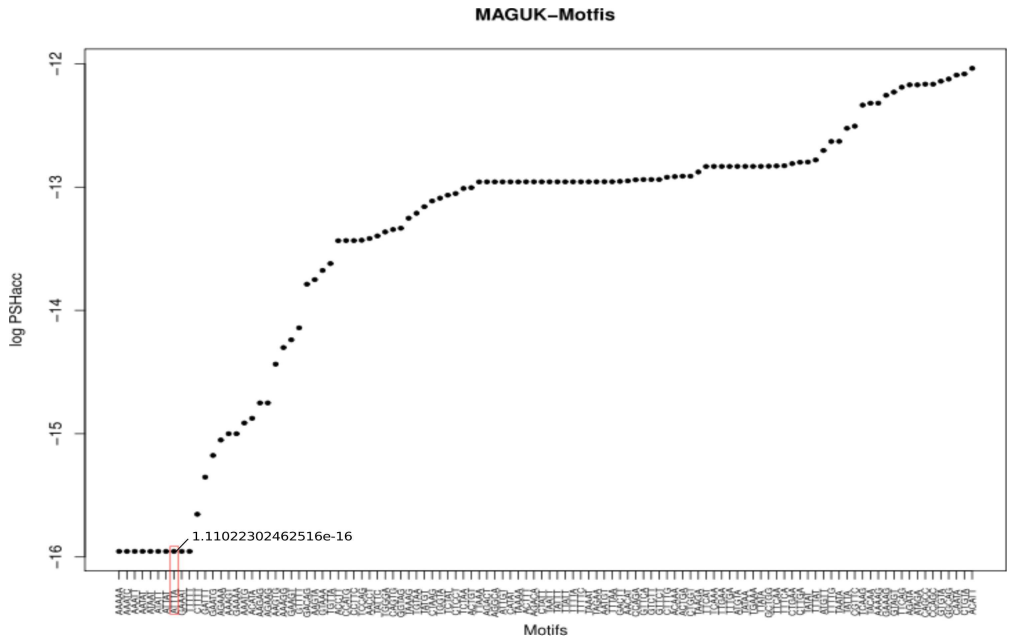


Fig. 46. The $PSHacc$ distribution of all pentamers in the MAGUK p55 subfamily member 7 representing transcript ENST00000337532 with a PSHacc value cutoff at 10^{-12} for better visualization. The 'AUUUA' ARE core motif has been highlighted with a red frame, its PSHacc value has been added to the plot.

Figure 46 shows that the 'AUUUA' core motif is not the only pentamer with the lowest PSHacc value found in this transcript. The A+U content in the 3' UTR is relatively high with 0.66 and almost only AU-rich motifs can be found among the top ten motifs with very low PSHacc values (AAAAA, AAAUC, AAAUU, AAUAU, AUAAU, AUAUU, AUUAU, AUUUA, GAAAU, UUUUU), which may be flanking regions of the ARE core motif.
As so far no experimental evidence for the regulation of MAGUK p55 subfamily member 7 mRNA via AUBPs exists, this mRNA would represent a good target for further investigation in this direction.

6 Conclusion and Discussion

6.1 Conclusion

"In its current state ARElite reports 3275 human protein coding genes which have at least one occurrence of the consensus motif WUAUUUAUW in their 3' UTR sequences. This corresponds to $\sim$ 16% of the human protein coding genes." (1). Summarizing the results of section 5, $\sim$ 60% of the protein coding genes in the Ensembl database have a 3' UTR containing the 'AUUUA' ARE core motif with an accessibility in terms of opening energy of up to 3kcal/mol. This does of course not mean that all of those genes are under control of AUBP's, as the chance of finding the 'AUUUA' pentamer in a random sequence of the length of a common human 3' UTR is comparatively high but in its current state 'ARElite' lists 1514 genes that are validated as targets of AU-rich binding proteins, which leaves us with 1761 genes representing potential targets of AUBPs.

As discussed previously the 'ARElite' webserver and the databases that work in background have not been created to present annotated transcripts, but more to act as tools for the analysis of AU-rich binding protein targets. Some filtering methods were applied to screen the database for regulatory active ARE core motifs. The first approach was to filter annotated transcripts by their accessibility in terms of opening energy. Summing up the results of this analysis, it has been shown that the opening energy alone is an insufficient measure for the regulative role of an ARE motif. A sequence like 'AUUUA' alone can not fold into any kind of RNA secondary structure and is therefore expected to be accessible, which makes accessibility a poor filtering method for this analysis from begin on.

The calculation of over-representation of the annotated ARE motifs in form of the PSHacc value, as presented in (129) and described in section 4 has led to better results. This P-value provides information on how often a sequence motif can be found compared to the expected occurrence of this method in a given sequence. Comparison of the PSHacc value distribution of the ARE core motif in known targets and annotated transcripts showed that the 'AUUUA' motif is over-represented in both cases with a peak at a PSHacc value of 10^{-13}. The PSHacc value distribution of annotated 'AUUUA's has been compared with the

distribution of other pentamers and displayed a stronger over-representation of 'AUUUA's. A peak in the distribution of the other pentamers at PSHacc values between 10^{-13} and 10^{-14} has been analyzed in detail and showed $\sim 65\%$ AU content. By visual inspection of this region many AU-rich elements have been found and could partially be identified as motifs that usually occur in the flanking regions of the ARE core motif. Analysis by PSHacc value have so far shown that the ARE core motif 'AUUUA' as well as ARE flanking regions are over-represented in the analyzed transcripts.

To further test this method, a list of handpicked known AUBP targets (HuR, TNF-α, IL6, Bcl2, TTP, SENP1, MAGUK) has been analyzed in detail. HuR, TNF-α, IL6 and Bcl2 are targets that have been chosen due to a relatively large amount of available literature on their role as ARE regulated proteins for the human organism. The 'AUUUA' motif in their representing transcripts is in all cases significantly enriched, although the ARE core motif in HuR and Bcl2 was not among the top ten of motifs with the lowest PSHacc values.
Analysis of TTP, which is also a known target but was not listed in the reference database of the 'AREsite' webserver, presented the 'AUUUA' motif among the top four over-represented motifs and nine AU-rich motifs in the top ten. With a cutoff for the PSHacc value of 10^{-12} all known ARE targets analyzed here would have been found during a screening, which fits the peak at the distribution of 'AUUUA's in known targets at 10^{-13}.

To see if the very restrictive approach of taking the lowest found PSHacc value with 10^{-16} as cutoff would lead to the identification of other ARE targets, two transcripts with such a low PSHacc value were extracted and analyzed in detail. The SUMO-specific protease SENP1 (SENP1) and the MAGUK p55 subfamily member 7 (MAGUK) both show an over-representation of the 'AUUUA' core motif and AU-rich regions in the top ten motifs with lowest PSHacc values. Unfortunately no literature describing their regulation by AU-rich binding proteins was found, but the phylogenetic analysis provided by 'AREsite' shows a conservation of the over-represented 'AUUUA' motif and is another strong indicator for a regulation of their mRNA stability by AUBPs. Both proteins should be analyzed experimentally for effects of AUBPs on their expression to validate this analysis.

Recapitulating the screening methods used for the database analysis it can be said that the analysis of over-representation of ARE core motifs produced some promising results. Further analysis is required to filter real targets from the rest, as a cutoff for the PSHacc value at 10^{-12} would have included a lot of false positives.
Combining this method with the information provided by the webserver 'ARE-site', which includes phylogenetic analysis, motif fold-enrichment, detailed information on the location of the annotated motifs and literature if available, can result in a very effective method to predict novel AU-rich binding protein targets.
To further improve the screening, more information on the structural context of the ARE motifs is necessary, as is a detailed analysis of flanking regions and ARE motifs that do not exactly match the discussed 'AUUUA' ARE core motif.

6.2 Discussion

Regulation of gene expression is a complex system. The same is true for the regulation of mRNA stability, where many factors are involved that often show a combined set of targets. This is a fact that has to be taken into account when predicting novel targets of AU-rich binding proteins.
'AREsite' is a webserver that does not only present the results of sequence annotation, but contains additional information that can be used to analyze AUBP targets. Ghosh et al. have already published work where this has been done (2). Approaches to filter the presented dataset by the application of filtering methods like the PSHacc value distribution show some promising results although the analysis is not yet complete. So far only the 'AUUUA' ARE core motif has been analyzed. However, the mentioned techniques and ARE motifs as well as their binding proteins are usually not restricted to this core region alone. In fact one 'AUUUA' core motif alone is usually not sufficient for the interaction of known AU-rich binding proteins with their targets. HuR has been shown to require the nine nucleotide long U-rich region 'NNUUNNUUU' more than AU-rich regions for its actions (83) for example. Nonetheless the analysis of the core motif is a first step in a successful prediction of novel AUBP targets. Together with further analysis approaches more information on the requirements for regulative active ARE-AUBP interactions will be revealed and used for more precise predictions. This section discusses the already used approaches for their potential as techniques for the prediction of novel AU-rich binding protein targets and presents methods to enhance their prediction capability. The

section 7 presents methods that will be used for further investigations.

6.2.1 Prediction of target sites using information on their accessibility and over-representation

The analysis of ARE core motif accessibility did not lead to measures that can be used to distinguish regulatory active from inactive motifs so far. As this approach was constricted to the very short region of the 'AUUUA' ARE core motif, this finding was no surprise. However, accessibility of a sequence plays a role in the interaction of AU-rich binding proteins and their targets, which is even increased by the fact that the so far known AUBPs are thought to bind to single stranded RNA sequences which was validated for zinc finger containing AU-rich binding proteins,see (65). The first enhancement using this approach can be done, by taking ARE flanking regions into account. This regions have also been found to be over-represented in ARE containing transcripts (see 5) and information on their accessibility gives clues about the structural context that embeds the ARE motifs. Preceding work on the influence of RNA secondary structures on AUBP binding has shown that their regulatory actions can in the case of HuR be switched off and on by embedding of the according ARE motif into a secondary structure or breaking down this structure (83). Experimental evidence for the existence of AU-rich elements in loop regions of regulatory RNA structures is already available (144) and points out the importance of the structural context for regulation. An analysis of the structural context in which ARE motifs can be found could lead to a more precise categorization of motif and action. Depending on whether the motif can be found in interior loops or the stem or the loop region of a stem-loop can influence the binding probabilities of an AU-rich binding protein. To model this, a lot of information on the AUBP itself, according the region that interacts with the RNA, the footprint of the whole AUBP, the type of RNA binding motif, the number of proteins that interact and distances between binding motifs and proteins, if more than one bind at once, is required. This information has to be collected from literature search and can then be combined with information from secondary structure prediction to get a detailed view on interaction characteristics and mechanisms. Such an approach could even lead to characteristics that can be used to distinguish between certain AU-rich binding proteins in their choice of target, or reveal motifs that are required for up- or down-regulation of mRNA stability respectively.

A lot of AU-rich binding protein targets are known to be regulated by more than one AUBP or the latter have to compete with other RNA binding factors for their target sites (refer to section 2.1.4 for detailed information). Influence of the changes in motif accessibility before and after binding of a factor up- or downstream of an ARE motif can be calculated. Given the information where an AUBP or other RBP binds, restrictions can be introduced when folding the target 3' UTR *in silico* that can give information on accessibility changes at the motif site. The program RNAup (121) can be used to fold RNA sequences with constraints and can be used to model these changes in accessibility. It can even give information on the kind of secondary structure that embeds the ARE motif and will be used for further investigations.
As has been shown the combination of motif accessibility and over-representation to the PSHacc value can be used to filter regulative active ARE motifs. The PSHacc distribution can be applied to the database to screen for transcripts where accessible 'AUUUA' ARE core motifs are over-represented. Although over-representation is only a hint for regulatory function and known targets do contain multiple copies of 'AUUUA's, previously mentioned additional information can be used to improve screening results. Information on the numbers of ARE motifs that are bound by a single AU-rich binding protein can be used to calculate cutoffs which allows the prediction of targets for each AUBP specifically. TTP, for example, is known to contain two zinc finger motifs that can simultaneously bind to RNA. Extracting information on the number of TTPs that have to bind a target at the same time for regulatory function would give a minimal number of accessible ARE motifs required for action. This information can be used to extract mRNAs that contain at least this minimum number of motifs and then be analyzed in detail. Information on the footprint of an AU-rich binding protein can be used to extract motifs that are masked if an ARE motif in proximity is bound by this AUBP. The masked ARE motif is not available for other AU-rich binding proteins or RBPs, so a minimal over-representation of this motif has to be guaranteed to allow concurrent binding of other factors. This can again be used to extract mRNAs that contain a certain amount of ARE motifs if they are regulated by multiple AU-rich binding proteins or RBPs at once. Summing up, this method has been validated with some known AUBP targets, but for a prediction of novel targets or research on the mechanisms of mRNA regulation by ARE motifs, further information has to be combined with this method.

7 Outlook

7.1 Further investigations using the existing ARE motif database

As discussed, the results of the first analysis of annotated transcripts leave room for further investigation. So far, the analysis of ARE motif containing transcripts was limited to accessibility and over-representation of the ARE core motif 'AUUUA'.
AUBPs are thought to bind single stranded regions containing their target motifs, which are often longer than just the core 'AUUUA'. Following this, it would give a hint on the regulative role of a motif to look at its flanking regions. This includes prediction of accessibility of these regions, as well as information on the structural context in which the motif is embedded. For a comprehensive analysis, ARE flanking regions, the structural context and the conservation of ARE motifs have to be included, as will be discussed in more detail in this section.

7.1.1 Analysis of ARE flanking regions and further approaches for the identification of AUBP targets and the underlying mechanisms

ARE flanking regions

As mentioned previously, ARE flanking regions influence the interaction of AUBPs with their targets. The most important task for future analysis is to analyze these regions and extract the structural context that embeds an ARE (core) motif. Rabani et al. (156) present results for the secondary structure prediction of two ARE containing mRNAs that are targeted by RBPs but decay with different rates. The ARE motifs are predicted to be embedded in structurally different loop regions which may be the reason for the decay speed differences. Following this example, the analysis of the structural context of the annotated motifs in the database can be conducted, for example with the program RNAup (121). RNAup was discussed in section 6.2.1 as tool to calculate changes in motif accessibility given prior binding of proteins or factors and predict different types of RNA secondary structures which can embed the ARE motifs. The program allows to distinguish between being in a hairpin-loop, interior-loop, exterior-loop or multi-loop, which can influence the interaction probability of an AUBP with the target mRNA. While an ARE

motif presented in a hairpin loop is easily accessible for an AUBP, interior- or multi-loops as well as exterior-loop regions in close proximity of an other loop structure can hinder a protein from interaction with the sequence motif. This does depend on the type of AUBP and the size of the RNA region that is required for interaction and information on the type of secondary structure that allows or restricts an AUBP from binding to the ARE motif can be used to screen for other sequences with similar structures and analyze their probability of being a target of this AUBP. Concluding this discussion, the next step in analyzing AU-rich elements as gene expression regulator would be to analyze their proximity to secondary structures and find a way to combine the available information provided by the webserver 'AREsite' and the analysis already conducted to a measure that can be used as de novo filtering method for novel AUBP targets. Together with results from the experimental analysis of already predicted targets and literature search, it may be possible to fine tune cutoffs for certain AUBP characteristics. A detailed analysis of known targets for each AUBP may reveal properties that allow to predict targets specifically for each AUBP.

Analysis by distance

Previously mentioned was the influence of the distance between ARE motifs on their functionality. AUBPs that contain more than one ARE binding motif have the capability to interact on multiple sites at once. Extracted information on the exact binding properties of AUBPs from literature can be used to screen for motifs that are found in distances equally to the distances of the protein binding motifs. This information is hard to gather as only little is known about the exact mechanism underlying AUBP binding. Comparing the distances between ARE motifs in known targets of certain AUBPs could lead to the identification of characteristics that are typical for an AUBP or even for actions of AUBPs on their targets.

TTP is one example where an AUBP can bind to more than one ARE motif at once. Using information about the distance of the binding sites in the protein and comparing it with the distance of ARE motifs on annotated transcripts could lead to successful predictions of novel TTP targets. This task is not easily done by counting the number of nucleotides which separate two ARE motifs, but has to be conducted together with information on the structural context of these motifs, as distances in a folded RNA molecule can of course differ from the distances along the (unfolded) RNA sequence.

7.2 Experimental validation of novel AUBP targets

Results retrieved from the ARE motif analysis have to be validated experimentally. On the one hand, only experimental approaches can show if a predicted target really is influenced by the presence of AUBPs, and on the other hand these experiments allow to extract new parameters that help to improve the target prediction.

A common, well established approach is the reporter gene assay with firefly luciferase.
By cloning luciferase with the whole 3' UTR, or segments containing the annotated ARE motifs, of an predicted target (see figure 47) into a vector and transfecting it into an established cell line, it can easily be seen whether this 3' UTR marks the luciferase as target for AUBPs. If this is the case, luciferase will be expressed in higher or lower amounts, depending on the AUBP that binds, than a comparable luciferase without this 3' UTR.

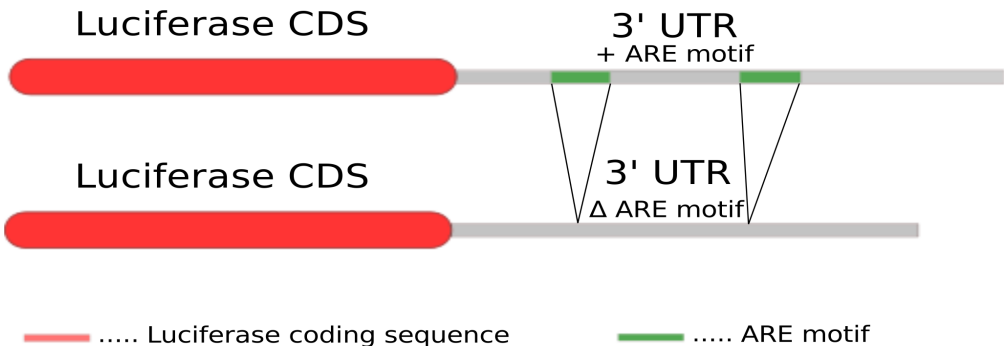


Fig. 47. The firefly luciferase as used for reporter gene assays, with the 3' UTR of a predicted AUPB target. The upper luciferase 3'UTR contains the ARE sequence motifs, that tags it as target for AUBPs, and the lower version is used as control and contains a 3' UTR without the ARE motifs.

This approach has been used many times to validate the actions of AUPBs on a mRNA target, see for example (157; 158; 159; 160) and many more. It can be used comparatively easy to verify or falsify the regulative role of newly annotated ARE motifs.

7.3 Interplay of AUBPs and miRNA

MiRNAs are known to play a major role in gene expression regulation as shown for example in (56). Therapeutic approach of mRNA stability regulation by miRNAs are already being discussed (161). More and more information on the interplay of RNA binding proteins and miRNAs is being revealed and as mentioned in the section 2 this can interfere with the binding of AUBPs to their targets, increasing or decreasing their regulatory effects. A first approach would be to analyze known miRNA binding sites in the annotated transcripts, to see whether or not these binding sites overlap with ARE motifs. If this is the case the binding of a miRNA would sterically prevent interaction of an AUBP with its target site and vice versa. Influence on the accessibility of ARE motifs could also be influences by a prior interaction of a miRNA with the AUBP target and could be analyzed as discussed in section 6.2.1 Extracted information of the distance between RNA binding motifs in AUBPs can be used in combination with the annotation of miRNA binding sites to analyze if such a miRNA target site can be found between the ARE motifs required for the AUBP function, as this would again interfere with both interactions. MiRNAs are in the focus of a growing number of research groups which leads to a lot of available information that can be used to model the influence of miRNAs on AUBP binding and vice versa with the presented methods.

References

[1] Gruber AR, Fallmann J, Kratochvill F, Kovarik P, Hofacker IL: **AREsite: a database for the comprehensive investigation of AU-rich elements.** *Nucleic acids research*(November 2010):1–4.

[2] Ghosh SK, Gupta S, Jiang B, Weinberg A: **Fusobacterium nucleatum and Human Beta-Defensins Modulate the Release of Antimicrobial Chemokine CCL20/Macrophage Inflammatory Protein 3{alpha}.** *Infection and immunity*(11):4578–87.

[3] Berg JM, Tymoczko JL, Stryer L: *Stryer Biochemie, Volume 7th edition.* Palgrave Macmillan 2007.

[4] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: *Molecular Biology of the Cell, Volume 5th edition.* Garland Science 2007.

[5] Williams Aa, Darwanto A, Theruvathu Ja, Burdzy A, Neidigh JW, Sowers LC: **Impact of sugar pucker on base pair and mispair stability.** *Biochemistry*(50):11994–2004.

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

[7] Strobel S, Doudna J: **RNA seeing double: close-packing of helices in RNA tertiary structure**. *Trends in biochemical sciences*(7):262–266.

[8] Crick F: **The Biological Replication of Macromolecules**. *in Symp. Soc. Exp. Biol.* 1958, **XII**(138).

[9] Fire A, Xu S, Montgomery M, Kostas S, Driver S, Mello C: **Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans**. *Nature*(6669):806–811.

[10] Amariglio N, Rechavi G: **A-to-I RNA editing: a new regulatory mechanism of global gene expression.** *Blood cells, molecules & diseases*(2):151–5.

[11] Agris PF, Vendeix FaP, Graham WD: **tRNA's wobble decoding of the genome: 40 years of modification.** *Journal of molecular biology*:1–13.

[12] Leontis NB, Stombaugh J, Westhof E: **The non-Watson-Crick base pairs and their associated isostericity matrices.** *Nucleic acids research* 2002, **30**(16):3497–531.

[13] Leu K, Obermayer B, Rajamani S, Gerland U, Chen Ia: **The prebiotic evolutionary advantage of transferring genetic information from RNA to DNA.** *Nucleic acids research*(18):8135–8147.

[14] Cech T: **The RNA Worlds in Context.** *Cold Spring Harbor perspectives in biology*.

[15] Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M: **What is a gene, post-ENCODE? History and updated definition.** *Genome research*(6):669–81.

[16] Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM: **Protein-RNA interactions: a structural analysis.** *Nucleic acids research*(4):943–54.

[17] Terribilini M, Lee JH, Yan C, Jernigan RL, Honavar V, Dobbs D: **Prediction of RNA binding sites in proteins from amino acid sequence.** *RNA (New York, N.Y.)*(8):1450–62.

[18] Xie J, Schultz PG: **Adding amino acids to the genetic repertoire.** *Current opinion in chemical biology*(6):548–54.

[19] Yang C, Bolotin E, Jiang T, Sladek F, Martinez E: **Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters**. *Gene*:52–65.

[20] Crick F: **Central dogma of molecular biology**. *Nature*(5258):561–563.

[21] Thieffry D, Sarkar S: **Forty years under the central dogma.** *Trends in biochemical sciences*(8):312.

[22] Hurwitz J: **The discovery of RNA polymerase**. *Journal of Biological Chemistry*(52):42477–42485.

[23] Chakalova L, Fraser P: **Organization of transcription.** *Cold Spring Harbor perspectives in biology*(9).

[24] Kornberg R: **The molecular basis of eukaryotic transcription**. *Proceedings of the National Academy of Sciences*(32):12955.

[25] Lewin B: *Genes VIII*. Pearson Prentice Hall 2004.

[26] Hocine S, Singer R, Grünwald D: **RNA Processing and Export**. *Cold Spring Harbor perspectives in biology*(12).

[27] Proudfoot NJN, Furger A, Dye MJM: **Integrating mRNA processing with transcription.** *Cell*(4):501–12.

[28] Millevoi S, Vagner S: **Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation.** *Nucleic acids research*(9):2757–74.

[29] Han J, Xiong J, Wang D, Fu XD: **Pre-mRNA splicing: where and when in the nucleus.** *Trends in cell biology*:1–8.

[30] RajBhandary U, Köhrer C: **Early days of tRNA research: Discovery, function, purification and sequence analysis**. *Journal of Biosciences*(4):439–451.

[31] Consortium TEP: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**(7146):799–816.

[32] Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JGR, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R: **GENCODE: producing a reference annotation for ENCODE.** *Genome biology* 2006, **7 Suppl 1**(Suppl 1):S4.1–9.

[33] Ravasi Tea: **An atlas of combinatorial transcriptional regulation in mouse and man.** *Cell* 2010, **140**(5):744–52.

[34] Smith TF, Lee JC, Gutell RR, Hartman H: **The origin and evolution of the ribosome.** *Biology direct* 2008, **3**:16.

[35] Kapp LD, Lorsch JR: **The molecular mechanics of eukaryotic translation.** *Annual review of biochemistry*:657–704.

[36] Malys N, McCarthy JEG: **Translation initiation: variations in the mechanism can be anticipated.** *Cellular and molecular life sciences : CMLS*:991–1003.

[37] Khoury Ga, Baliban RC, Floudas Ca: **Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database**. *Scientific Reports*:1–5.

[38] Yu MC: **The Role of Protein Arginine Methylation in mRNP Dynamics**. *Molecular Biology International*:1–10.

[39] Caetano-Anollés G: *Evolutionary Genomics and Systems Biology*.

[40] Kedde M, van Kouwenhove M, Zwart W, Oude Vrielink JaF, Elkon R, Agami R: **A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility.** *Nature cell biology*(10):1014–20.

[41] Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E: **Genome-wide measurement of RNA secondary structure in yeast.** *Nature*(7311):103–7.

[42] Pyle AM: **Metal ions in the structure and function of RNA.** *Journal of biological inorganic chemistry : JBIC : a publication of the Society of Biological Inorganic Chemistry*(7-8):679–90.

[43] Batey R, Rambo R, Doudna J: **Tertiary Motifs in RNA Structure and Folding.** *Angewandte Chemie (International ed. in English)*(16):2326–2343.

[44] Staple D, Butcher S: **Pseudoknots: RNA structures with diverse functions**. *PLoS biology* 2005, **3**(6):e213.

[45] Mandal M, Breaker RR: **Gene regulation by riboswitches.** *Nature reviews. Molecular cell biology*(6):451–63.

[46] Cheah MT, Wachter A, Sudarsan N, Breaker RR: **Control of alternative RNA splicing and gene expression by eukaryotic riboswitches.** *Nature*(7143):497–500.

[47] Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY: **Understanding the transcriptome through RNA structure**. *Nature Reviews Genetics*(9):641–655.

[48] Ladomery MR, Maddocks DG, Wilson ID: **MicroRNAs: their discovery, biogenesis, function and potential use as biomarkers in non-invasive prenatal diagnostics.** *International journal of molecular epidemiology and genetics*(3):253–60.

[49] Volk N, Shomron N: **Versatility of MicroRNA Biogenesis**. *PLoS ONE*(5):e19391.

[50] Beezhold KJ, Castranova V, Chen F: **Microprocessor of microRNAs: regulation and potential for therapeutic intervention.** *Molecular cancer*:134.

[51] von Roretz C, Gallouzi IE: **Decoding ARE-mediated decay: is microRNA part of the equation?** *The Journal of cell biology*(2):189–94.

[52] Thomas M, Lieberman J, Lal A: **Desperately seeking microRNA targets.** *Nature structural & molecular biology*(10):1169–74.

[53] Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A: **Identification of mammalian microRNA host genes and transcription units.** *Genome research*(10A):1902–10.

[54] Neugebauer KM: **On the importance of being co-transcriptional**. *Journal of Cell Science* 2002, **115**(20):3865.

[55] Idler R, Yan W: **Control of messenger RNA fate by RNA binding proteins: an emphasis on mammalian spermatogenesis**. *Journal of Andrology*:jandrol–111.

[56] van Kouwenhove M, Kedde M, Agami R: **MicroRNA regulation by RNA-binding proteins and its implications for cancer.** *Nature reviews. Cancer*(9):644–656.

[57] Glisovic T, Bachorik JL, Yong J, Dreyfuss G: **RNA-binding proteins and post-transcriptional gene regulation.** *FEBS letters*(14):1977–86.

[58] Lunde BM: **RNA-binding proteins: modular design for efficient function.** *Nature reviews. Molecular cell biology*(6):479–90.

[59] Keene JD: **Ribonucleoprotein infrastructure regulating the flow of genetic information between the genome and the proteome.** *Proceedings of the National Academy of Sciences of the United States of America*(13):7018–7024.

[60] Lasko P: **The drosophila melanogaster genome: translation factors and RNA binding proteins.** *The Journal of Cell Biology*(2):F51–6.

[61] Lee MH, Schedl T: **RNA-binding proteins**. *WormBook*:1–13.

[62] Maris C, Dominguez C, Allain FHT: **The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression.** *The FEBS journal*(9):2118–31.

[63] Nagai K, Oubridge C, Ito N, Avis J, Evans P: **The RNP domain: a sequence-specific RNA-binding domain involved in processing and transport of RNA**. *Trends in biochemical sciences*(6):235–240.

[64] Mangus Da, Evans MC, Jacobson A: **Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression.** *Genome biology*(7):223.

[65] Hudson BP, Martinez-Yamout Ma, Dyson HJ, Wright PE: **Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d.** *Nature structural & molecular biology* 2004, **11**(3):257–64.

[66] Mignone F, Gissi C, Liuni S, Pesole G: **Untranslated regions of mRNAs.** *Genome biology*(3):REVIEWS0004.

[67] Chatterjee S, Pal JK: **Role of 5'- and 3'-untranslated regions of mRNAs in human diseases.** *Biology of the cell / under the auspices of the European Cell Biology Organization*(5):251–62.

[68] Serganov A: **The long and the short of riboswitches.** *Current opinion in structural biology* 2009, **19**(3):251–9.

[69] Batzer MA, Deininger PL: **Alu repeats and human genomic diversity.** *Nature reviews. Genetics* 2002, **3**(5):370–9.

[70] Ross J: **mRNA stability in mammalian cells.** *Microbiological reviews* 1995, **59**(3):423–50.

[71] Huang HY, Chien CH, Jen KH, Huang HD: **RegRNA: an integrated web server for identifying regulatory RNA motifs and elements.** *Nucleic acids research*(Web Server issue):W429–34.

[72] Blackshear P: **mRNA degradation: an important process in controlling gene expression**. *Biochemical Society Transactions*.

[73] Miller AD, Curran T, Verma IM: **c-fos protein can induce cellular transformation: a novel mechanism of activation of a cellular oncogene.** *Cell* 1984, **36**:51–60.

[74] Shaw G, Kamen R: **A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation.** *Cell* 1986, **46**(5):659–667.

[75] Bakheet T, Frevel M, Williams BRG, Greer W, Khabar KSA: **ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins**. *Nucleic Acids Research* 2001, **29**:246–254.

[76] Chen CYA, Shyu AB: **Au-Rich Elements - Characterization and Importance in Messenger-Rna Degradation**. *Trends in Biochemical Sciences* 1995, **20**(11):465–470.

[77] Guhaniyogi J, Brewer G: **Regulation of mRNA stability in mammalian cells.** *Gene* 2001, **265**(1-2):11–23.

[78] Dai W, Zhang G, Makeyev EV: **RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage**. *Nucleic Acids Research*:1–14.

[79] Sanduja S, Blanco F, Dixon D: **The roles of TTP and BRF proteins in regulated mRNA decay**. *Wiley Interdisciplinary Reviews-RNA*:42–57.

[80] Gingerich TJ, Feige JJ, LaMarre J: **AU-rich elements and the control of gene expression through regulated mRNA stability.** *Animal health research reviews Conference of Research Workers in Animal Diseases* 2004, **5**:49–63.

[81] Barreau C, Paillard L, Osborne HB: **AU-rich elements and associated factors: are there unifying principles?** *Nucleic acids research* 2005, **33**(22):7138–50.

[82] Stoecklin G, Mayo T, Anderson P: **ARE-mRNA degradation requires the 5'-3' decay pathway.** *EMBO reports*:72–7.

[83] Meisner NC, Hackermüller J, Uhl V, Aszódi A, Jaritz M, Auer M: **mRNA openers and closers: modulating AU-rich element-controlled mRNA stability by a molecular switch in mRNA secondary structure.** *Chembiochem : a European journal of chemical biology*(10):1432–47.

[84] Lai WS, Carrick DM, Blackshear PJ: **Influence of nonameric AU-rich tristetraprolin-binding sites on mRNA deadenylation and turnover.** *The Journal of biological chemistry*(40):34365–77.

[85] Tchen CR, Brook M, Saklatvala J, Clark AR: **The stability of tristetraprolin mRNA is regulated by mitogen-activated protein kinase p38 and by tristetraprolin itself.** *The Journal of biological chemistry*(31):32393–400.

[86] Carballo E: **Feedback Inhibition of Macrophage Tumor Necrosis Factor- Production by Tristetraprolin**. *Science*(5379):1001–1005.

[87] Stoecklin G, Tenenbaum Sa, Mayo T, Chittur SV, George AD, Baroni TE, Blackshear PJ, Anderson P: **Genome-wide analysis identifies interleukin-10 mRNA as target of tristetraprolin.** *The Journal of biological chemistry*(17):11689–99.

[88] Baou M, Norton JD, Murphy JJ: **AU-rich RNA binding proteins in hematopoiesis and leukemogenesis.** *Blood*.

[89] Wagner BJ, DeMaria CT, Sun Y, Wilson GM, Brewer G: **Structure and genomic organization of the human AUF1 gene: alternative pre-mRNA splicing generates four protein isoforms.** *Genomics*(2):195–202.

[90] Blum JL, Samarel AM, Mestril R: **Phosphorylation and binding of AUF1 to the 3'-untranslated region of cardiomyocyte SERCA2a mRNA.** *American journal of physiology. Heart and circulatory physiology*(6):H2543–50.

[91] Raineri I, Wegmueller D, Gross B, Certa U, Moroni C: **Roles of AUF1 isoforms, HuR and BRF1 in ARE-dependent mRNA turnover studied by RNA interference.** *Nucleic acids research*(4):1279–88.

[92] Zhang W, Wagner BJ, Ehrenman K, Schaefer aW, DeMaria CT, Crater D, DeHaven K, Long L, Brewer G: **Purification, characterization, and cDNA cloning of an AU-rich element RNA-binding protein, AUF1.** *Molecular and cellular biology*(12):7652–65.

[93] Katahira M, Miyanoiri Y, Enokizono Y, Matsuda G, Nagata T, Ishikawa F, Uesugi S: **Structure of the C-terminal RNA-binding domain of hnRNP D0 (AUF1), its interactions with RNA and DNA,**

and change in backbone dynamics upon complex formation with DNA. *Journal of molecular biology*(5):973–88.

[94] Fialcowitz-White EJ, Brewer BY, Ballin JD, Willis CD, Toth Ea, Wilson GM: **Specific protein domains mediate cooperative assembly of HuR oligomers on AU-rich mRNA-destabilizing sequences.** *The Journal of biological chemistry*(29):20948–59.

[95] Brennan CM, Steitz JA: **HuR and mRNA stability.** *Cellular and molecular life sciences CMLS*(2):266–277.

[96] Deschênes-Furry J, Angus LM, Bélanger G, Mwanjewe J, Jasmin BJ: **Role of ELAV-like RNA-binding proteins HuD and HuR in the post-transcriptional regulation of acetylcholinesterase in neurons and skeletal muscle cells.** *Chemico-biological interactions*:43–9.

[97] Wang X, Tanaka Hall TM: **Structural basis for recognition of AU-rich element RNA by the HuD protein.** *Nature Structural Biology*(2):141–145.

[98] Ma WJ, Chung S, Furneaux H: **The Elav-like proteins bind to AU-rich elements and to the poly(A) tail of mRNA.** *Nucleic acids research*(18):3564–9.

[99] Kim HSH, Wilce MMCJ, Yoga YYMK, Pendini NRN, Gunzburg MMJ, Cowieson NPN, Wilson GMG, Williams BBRG, Gorospe M, Wilce JJa: **Different modes of interaction by TIAR and HuR with target RNA and DNA**. *Nucleic acids research*(3):1117.

[100] Lu JY, Schneider RJ: **Tissue distribution of AU-rich mRNA-binding proteins involved in regulation of mRNA decay.** *The Journal of biological chemistry*(13):12974–9.

[101] Benoit RMR, Meisner NC, Kallen J, Graff P, Hemmig R, Cèbe R, Ostermeier C, Widmer H, Auer M, Cčbe R: **The x-ray crystal structure of the first RNA recognition motif and site-directed mutagenesis suggest a possible HuR redox sensing mechanism.** *Journal of molecular biology*(5):1231–44.

[102] Pascale A, Govoni S: **The complex world of post-transcriptional mechanisms: is their deregulation a common link for diseases? Focus on ELAV-like RNA-binding proteins**. *Cellular and Molecular Life Sciences* 2011, :1–17.

[103] Durbin R, Eddy S, Krogh A, Mitchison G: *Biological sequence analysis: Probabalistic models of proteins and nucleic acids*. Cambridge University Press 1998.

[104] Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *Journal of Molecular Biology* 1970, **48**(3):443–453.

[105] Gotoh O: **An improved algorithm for matching biological sequences.** *Journal of Molecular Biology*(3):705–708.

[106] Smith TF, Waterman MS: **Identification of common molecular subsequences**. *Journal of Molecular Biology*:195–197.

[107] Lipman DJ, Pearson WR: **Rapid and Sensitive Protein Similarity Searches**. *Science*(4693):1435–1441.

[108] Altschul SF: **BLAST: Basic Local Alignment Search Tool**. *Distribution*(3):403–410.

[109] Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Research*(22):4673–80.

[110] Morgenstern B: **DIALIGN: multiple DNA and protein sequence alignment at BiBiServ.** *Nucleic acids research*(Web Server issue):W33–6.

[111] Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic acids research*(14):3059.

[112] Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFa, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W: **Aligning multiple genomic sequences with the threaded blockset aligner.** *Genome research*(4):708–15.

[113] Fujita Pa, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte Ra, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2011.** *Nucleic acids research*(Database issue):D876–82.

[114] Lengauer T: *Bioinformatics: From Genomes to Therapies, Volume 1.* Weinheim, Germany: Wiley-VCH 2007.

[115] Nussinov R, Jacobson aB: **Fast algorithm for predicting the secondary structure of single-stranded RNA.** *Proceedings of the National Academy of Sciences of the United States of America*(11):6309–13.

[116] Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ: **Algorithms for Loop Matchings**. *SIAM Journal on Applied Mathematics* 1978, **35**:68–82.

[117] Waterman M: **RNA secondary structure: a complete mathematical analysis**. *Mathematical Biosciences*(3-4):257–266.

[118] Zuker M, Stiegler P: **Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.** *Nucleic Acids Research*:133–148.

[119] Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.** *Journal of molecular biology*(5):911–40.

[120] McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure**. *Biopolymers* 1990, **29**(6-7):1105–1119.

[121] Mueckstein U, Tafer H, Bernhart SH, Hernandez-Rosales M, Vogel J, Stadler PF, Hofacker IL: **Translational Control by RNA-RNA Interaction : Improved Computation of RNA-RNA Binding Thermodynamics**. *Communications in Computer and Information Science* 2008, **13**:114–127.

[122] Bernhart SH, Hofacker IL, Stadler PF: **Local RNA base pairing probabilities in large sequences.** *Bioinformatics (Oxford, England)*(5):614–5.

[123] Hofacker I: **RNA secondary structure analysis using the Vienna RNA package**. *Current Protocols in Bioinformatics*:Unit12.2.

[124] Bernhart SH, Mückstein U, Hofacker IL: **RNA Accessibility in cubic time.** *Algorithms for molecular biology : AMB*:3.

[125] Date C: *An introduction to database systems*. Introduction to Database Systems, 7th Ed, Addison-Wesley 2000.

[126] Kemper A, Eickler A: *Datenbanksysteme*. Oldenbourg 2006.

[127] Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL: **The Vienna RNA websuite.** *Nucleic acids research*(Web Server issue):W70–4.

[128] Tafer H, Ameres SL, Obernosterer G, Gebeshuber Ca, Schroeder R, Martinez J, Hofacker IL: **The impact of target site accessibility on the design of effective siRNAs.** *Nature biotechnology*(5):578–83.

[129] Marín RM, Vanícek J: **Efficient use of accessibility in microRNA target prediction.** *Nucleic acids research*:19–29.

[130] Li X, Quon G, Lipshitz HD, Morris Q: **Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure.** *RNA (New York, N.Y.)*(6):1096–107.

[131] Kazan H, Ray D, Chan ET, Hughes TR, Morris Q: **RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins.** *PLoS computational biology*(7):e1000832.

[132] Khera T, Dick A, Nicholson L: **Mechanisms of TNF [alpha] regulation in uveitis: Focus on RNA-binding proteins**. *Progress in Retinal and Eye Research*(6):610–21.

[133] Beutler B, Milsark I, Cerami A: **Passive immunization against cachectin/tumor necrosis factor protects mice from lethal effect of endotoxin**. *Science*(4716):869.

[134] Lautz T, Stahl U, Lang C: **The human c-fos and TNF-*alpha* AU-rich elements show different effects on mRNA abundance and protein expression depending on the reporter in the yeast Pichia pastoris**. *Yeast*:1–9.

[135] Lai WS, Carballo E, Strum JR, Kennington EA, Phillips RS, Blackshear PJ: **Evidence that Tristetraprolin Binds to AU-Rich Elements and Promotes the Deadenylation and Destabilization of Tumor Necrosis Factor Alpha mRNA**. *Molecular and Cellular Biology*(6):4311–4323.

[136] Raghavan A, Robison RL, McNabb J, Miller CR, Williams DA, Bohjanen PR: **HuA and tristetraprolin are induced following T cell activation and display distinct but overlapping RNA binding specificities.** *The Journal of Biological Chemistry*(51):47958–47965.

[137] Dean JLE, Wait R, Mahtani KR, Sully G, Clark AR, Saklatvala J: **The 3 Untranslated Region of Tumor Necrosis Factor Alpha mRNA Is a Target of the mRNA-Stabilizing Factor HuR**. *Molecular and Cellular Biology*(3):721–730.

[138] Sauer I, Schaljo B, Vogl C, Gattermeier I, Kolbe T, Müller M, Blackshear PJ, Kovarik P: **Interferons limit inflammatory responses by induction of tristetraprolin.** *Blood*(12):4790–4797.

[139] Deleault KM, Skinner SJ, Brooks SA: **Tristetraprolin regulates TNF TNF-alpha mRNA stability via a proteasome dependent mechanism involving the combined action of the ERK and p38 pathways.** *Molecular Immunology*:13–24.

[140] Patil CS, Liu M, Zhao W, Coatney DD, Li F, VanTubergen EA, D'Silva NJ, Kirkwood KL: **Targeting mRNA stability arrests inflammatory bone loss.** *Molecular therapy the journal of the American Society of Gene Therapy*(10):1657–1664.

[141] Pullmann R, Kim HH, Abdelmohsen K, Lal A, Martindale JL, Yang X, Gorospe M: **Analysis of turnover and translation regulatory RNA-binding protein expression through binding to cognate mRNAs.** *Molecular and Cellular Biology*(18):6265–6278.

[142] Mukherjee N, Lager PJ, Friedersdorf MB, Thompson MA, Keene JD: **Coordinated posttranscriptional mRNA population dynamics during T-cell activation**. *Molecular Systems Biology*(288):288.

[143] Hirano T: **Interleukin 6 in autoimmune and inflammatory diseases: a personal memoir**. *Proceedings of the Japan Academy, Series B*(7):717–730.

[144] Paschoud S, Dogar AM, Kuntz C, Grisoni-Neupert B, Richman L, Kühn LC: **Destabilization of interleukin-6 mRNA requires a putative RNA stem-loop structure, an AU-rich element, and the RNA-binding protein AUF1.** *Molecular and cellular biology*(22):8228–41.

[145] Nabors LB, Gillespie GY, Harkins L, King PH: **HuR , a RNA stability factor , is expressed in malignant brain tumors and binds to adenine- and uridine-rich elements within the 3 untranslated regions of cytokine and angiogenic factor mRNAs**. *BioScience* 2001.

[146] De Silanes IL, Zhan M, Lal A, Yang X, Gorospe M: **Identification of a target RNA motif for RNA-binding protein HuR.** *Proceedings of the National Academy of Sciences of the United States of America*(9):2987–2992.

[147] Jalonen U, Nieminen R, Vuolteenaho K, Kankaanranta H, Moilanen E: **Down-Regulation of Tristetraprolin Expression Results in Enhanced IL-12 and MIP-2 Production and Reduced MIP-3α Synthesis in Activated Macrophages**. *Mediators of Inflammation*(6):40691.

[148] Tudor C, Marchese FP, Hitti E, Aubareda A, Rawlinson L, Gaestel M, Blackshear PJ, Clark AR, Saklatvala J, Dean JLE: **The p38 MAPK pathway inhibits tristetraprolin-directed decay of interleukin-10 and pro-inflammatory mediator mRNAs in murine macrophages.** *FEBS Letters*(12):1933–1938.

[149] Stoecklin G, Stoeckle P, Lu M, Muehlemann O, Moroni C: **Cellular mutants define a common mRNA degradation pathway targeting cytokine AU-rich elements.** *Rna New York Ny*(11):1578–1588.

[150] Ishimaru D, Zuraw L, Ramalingam S, Sengupta TK, Bandyopadhyay S, Reuben A, Fernandes DJ, Spicer EK: **Mechanism of regulation of bcl-2 mRNA by nucleolin and A+U-rich element-binding factor 1 (AUF1).** *The Journal of biological chemistry*(35):27182–91.

[151] Lapucci A, Donnini M, Papucci L, Witort E, Tempestini A, Bevilacqua A, Nicolin A, Brewer G, Schiavone N, Capaccioli S: **AUF1 Is a bcl-2 A + U-rich element-binding protein involved in bcl-2 mRNA destabilization during apoptosis.** *The Journal of Biological Chemistry*(18):16139–16146.

[152] Sadri N, Lu JY, Badura ML, Schneider RJ: **AUF1 is involved in splenic follicular B cell maintenance**. *BMC Immunology*:1.

[153] Geiss-Friedlander R, Melchior F: **Concepts in sumoylation: a decade on.** *Nature reviews. Molecular cell biology*(12):947–56.

[154] Xu Y, Li J, Zuo Y, Deng J, Wang LS, Chen GQ: **SUMO-specific protease 1 regulates the in vitro and in vivo growth of colon cancer cells with the upregulated expression of CDK inhibitors.** *Cancer letters*:78–84.

[155] Stucke VM, Timmerman E, Vandekerckhove J, Gevaert K, Hall A: **The MAGUK Protein MPP7 Binds to the Polarity Protein hDlg1 and Facilitates Epithelial Tight Junction Formation** . *Molecular Biology of the Cell* 2007, **18**(May):1744 –1755.

[156] Rabani M, Kertesz M, Segal E: **Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes.**(39):14885–90.

[157] Arvatz G, Barash U, Nativ O, Ilan N, Vlodavsky I: **Post-transcriptional regulation of heparanase gene expression by a 3' AU-rich element**. *The FASEB Journal*(12):4969.

[158] Accepts MCB, Society A, Authors L, Reserved AR: **Polyamines regulate the stability of JunD mRNA by modulating the competitive binding of its 3 -untranslated region to HuR and AUF1**. *Medicine* 2010, (August).

[159] Li H, Chen W, Zhou Y, Abidi P, Sharpe O, Robinson WHWWH, Kraemer FBF, Liu J: **Identification of mRNA binding proteins that regulate the stability of LDL receptor mRNA through AU-rich elements.** *Journal of lipid research*(5):820–31.

[160] Liao B, Hu Y, Brewer G: **Competitive binding of AUF1 and TIAR to MYC mRNA controls its translation.** *Nature structural & molecular biology*(6):511–8.

[161] Eberhardt W, Doller A, Akool ES, Pfeilschifter J: **Modulation of mRNA stability as a novel therapeutic approach.** *Pharmacology & therapeutics*:56–73.

A Predicted AUBP targets in human and mouse

This section contains human and mouse transcripts (see tables 4, 5) with a PSHacc of e^{-16}, which marks them as possible targets for AUBPs according to the over-representation and accessibility of the ARE core motif 'AUUUA' in their 3' UTR. Targets mentioned in this section are suggested for lab experiments (see section 7.2 for an example) that could validate their role as AUBP targets.

A.1 Predicted AUBP targets in human

Table 4 presents information extracted from the webserver 'AREsite' according to the predicted AUPB targets in human, that were not part of an detailed analysis so far.
As the extraordinary low PSHacc values shows that the ARE core motif 'AUUUA' in this transcripts is highly over-represented as well as accessible, this transcripts would make good targets for experimental validation of their regulation by AUBPS.

Gene	**Transcript**	**AKA**	**PSHacc**
ENSG00000120215	ENST00000381477	MLANA-001: melan-A	1.11022302462516e-16
ENSG00000124172	ENST00000243997	ATP5E: ATP synthase	2.22044604925031e-16
ENSG00000129625	ENST00000379638	REEP5-001: receptor accessory protein 5	8.88178419700125e-16
ENSG00000130513	ENST00000252809	GDF15-201: growth differentiation factor 15	1.11022302462516e-16
ENSG00000136541	ENST00000410096	ERMN-002: ermin, ERM-like protein	6.66133814775094e-16
ENSG00000136709	ENST00000322313	WDR33-001: WD repeat domain 33	1.11022302462516e-16
ENSG00000156030	ENST00000428145	C14orf43-005:chromosome 14 open reading frame 43	2.22044604925031e-16
ENSG00000167005	ENST00000300291	NUDT21-001:nudix (nucleoside diphosphate linked moiety X)-type motif 21	6.66133814775094e-16
ENSG00000178425	ENST00000319550	NT5DC1-002: 5'-nucleotidase domain containing 1	5.55111512312578e-16
ENSG00000183475	ENST00000332783	ASB7-001: ankyrin repeat and SOCS box-containing 7	8.88178419700125e-16
ENSG00000183695	ENST00000329773	MRGPRX2-001: MAS-related GPR, member X2	6.66133814775094e-16
ENSG00000188906	ENST00000430804	LRRK2-005: leucine-rich repeat kinase 2	5.55111512312578e-16
ENSG00000203685	ENST00000366788	C1orf95-201: chromosome 1 open reading frame 95	4.44089209850063e-16

Tab. 4. This table shows predicted AUBP targets in human according to an extraordinary low PSHacc value.

A.2 Predicted AUBP targets in mouse

Table 5 presents information extracted from the webserver 'AREsite' according to the predicted AUPB targets in mouse, that were not part of an detailed analysis so far.
The extraordinary low PSHacc values shows that the ARE core motif 'AUUUA' in this transcripts is highly over-represented as well as accessible, this transcripts would make good targets for experimental validation of their regulation by AUBPS.

Gene	**Transcript**	**AKA**	**PSHacc**
ENSMUSG00000029050	ENSMUST00000030917	Ski-001 ski sarcoma viral oncogene homolog (avian)	5.55111512312578e-16
ENSMUSG00000030148	ENSMUST00000041779	Clec4a2-001 C-type lectin domain family 4, member a2	2.22044604925031e-16
ENSMUSG00000031644	ENSMUST00000034065	Nek1-001 NIMA (never in mitosis gene a)-related expressed kinase 1	3.33066907387547e-16
ENSMUSG00000032727	ENSMUST00000109272	Mier3-001 mesoderm induction early response 1, family member 3	2.22044604925031e-16
ENSMUSG00000049764	ENSMUST00000061617	Zfp280b-201	3.33066907387547e-16
ENSMUSG00000058589	ENSMUST00000078569	Anks1b-201 ankyrin repeat and sterile alpha motif domain containing 1B Gene	3.33066907387547e-16

Tab. 5. This table shows predicted AUBP targets in mouse according to an extraordinary low PSHacc value.

B Usage statistics

AREsite visitors have been logged after publication of the websuite and table 6 shows traffic from all over the world, indicating the importance of web browsable tools for the modern scientist.

IP	Queries	Organization
131.111.189.78	5028	University of Cambridge
202.127.20.33	2257	China Science & Technology Network
149.155.96.6	202	Biotechnology And Biological Science Research Council
149.155.96.5	191	Biotechnology And Biological Science Research Council
138.26.45.133	178	University of Alabama at Birmingham - University Computer Center
133.45.137.205	148	National University Corporation, Nagasaki University
193.51.157.40	147	Universite Montpellier 1
133.1.239.242	117	Osaka University
140.251.50.169	113	Joan and Sanford I. Weill Medical College and Graduate School of Medical Sciences of Cornell University
129.96.252.195	111	Flinders University
193.174.111.250	110	Medizinische Hochschule Hannover
18.4.1.146	102	Massachusetts Institute of Technology
132.239.77.249	102	University of California, San Diego
192.55.208.10	100	St. Jude Children's Research Hospital
115.156.249.82	99	East-Zone for Huazhong University of Science and Technology
204.187.34.100	82	Ottawa General Hospital
129.25.15.197	78	The Drexel University Campus
142.150.92.120	75	University of Toronto
130.223.210.101	75	University of Lausanne
129.125.135.99	71	Rijks Universiteit Groningen

Top ten genes

Gene identifier	Number of queries
ENSMUSG00000028492	667
ENSG00000136244	283
ENSMUSG00000020691	130
ENSG00000232810	68
ENSMUSG00000025746	52
ENSG00000141510	52
ENSG00000171791	41
ENSG00000073756	39
ENSMUSG00000024401	39
ENSG00000169429	36

Top ten visits at one day

Day	Number of queries
Wed Mar 9 2011	2095
Fri Feb 18 2011	1603
Tue Mar 8 2011	1386
Tue Dec 21 2010	1104
Mon Dec 20 2010	793
Mon Jan 10 2011	279
Sun Nov 14 2010	192
Tue Jul 26 2011	171
Thu Dec 23 2010	170
Tue Jan 18 2011	166

Tab. 6. This table shows an excerpt of the top 20 visitors of the webserver 'AREsite' from all over the world, listed with IP, number of queries and the name of the organization filtered via its IP address. If the IP address could not be matched to an organization, the entry was excluded from the list. Following are shown the top 10 genes that were analyzed using the webserver and the top 10 days of usage.

Jörg Fallmann

Education

since 2004 **Diploma study**, *Molecular Biology, University of Vienna*.

since 02.2010 **Diploma Thesis**, *Insitute for theoretical chemistry, University of Vienna, Group Prof. Ivo Hofacker*, 1090 Vienna, *Investigation and prediction of interactions between AU-rich binding proteins and AU rich elements and the generation of the 'AREsite' web-server*.

1999–2004 **Federal Training and Research Institute for Industrial Chemistry HBLVA Rosensteingasse**, *1170 Vienna*.

Professional Experience

14.04.2008-31.10.2008 **System Administrator**, *Insitute for theoretical chemistry, University of Vienna*, 1090 Vienna, Administration of workstations, cluster machines and network.

01.03.2010-31.07.2010 **Tutor**, *Max F Perutz Laboratories*, 1030 Vienna, Tutor in laboratory course: Molecular Biology Techniques.

01.10.2009-28.02.2010 **Tutor**, *Max F Perutz Laboratories*, 1030 Vienna, Tutor in laboratory course: Molecular Biology Techniques.

01.03.2009-31.07.2009 **Tutor**, *Max F Perutz Laboratories*, 1030 Vienna, Tutor in laboratory course: Molecular Biology Techniques.

Languages

German **native**

English **advanced**

Computer skills

Scripting languages: Perl, Bash, Java

OS: Linux, Windows

Misc: R, HTML, Latex, MySQL, SQlite

Talks

- *TBI Winterseminar* in Bled, Slovenia, Feb 13-20,2011
  Title: AREsite proceedings
- *Herbstseminar Bioinformatik* in Vysoká Lípa, Czech Republic, Oct. 5-10, 2010
  Title: Proteins that bind ARE-motifs

Währingerstrasse 17 – 1090 Vienna
☎ *+43-1-4277-52732* • ✉ *fall@tbi.univie.ac.at*