Ribosome profiling, a useful tool in the search for micropeptides

Steven Verbruggen
Promotors: Prof. Dr. ir. Wim Van Criekinge, Dr. ir. Gerben Menschaert
Tutors: Dr. ir. Gerben Menschaert, Dr. ir. Jeroen Crappé

Master’s dissertation submitted in partial fulfillment of the requirements for the degree of Master in Bioscience Engineering: Cell and Gene Biotechnology
Ribosome profiling, a useful tool in the search for micropeptides

Steven Verbruggen
Promotors: Prof. Dr. ir. Wim Van Criekinge, Dr. ir. Gerben Menschaert
Tutors: Dr. ir. Gerben Menschaert, Dr. ir. Jeroen Crappé

Master’s dissertation submitted in partial fulfillment of the requirements for the degree of Master in Bioscience Engineering: Cell and Gene Biotechnology
“The author and the promotors give the permission to make this dissertation available for consultation and to copy parts of it for personal use. Every other use is covered by the limitations of copyright, in particular concerning the obligation to mention the source when citing results from this thesis.”

The promotors: Prof. Dr. ir. Wim Van Criekinge, Dr. ir. Gerben Menschaert

The author: Steven Verbruggen


Eerst en vooral wil ik jou bedanken, beste lezer. Blijk het finale doel van een tekst om hem te lezen, met welke reden en achtergrond dan ook. In ieder geval bedankt dat je de ogen op dit boekje werpt en hopelijk vind je wat je zoekt. De hoofdmoot van dit boekje is net iets serieuzer dan de paragraaf die je net hebt gelezen, maar hoe nostalgisch die paragraaf ook is, hoe zeer deze ook deint op zondagavondmuziek, het kan goed zijn om stil te staan bij de afgelegde reis om dan volop te gaan voor de toch die nog komen zal. Wat je van die paragraaf moet onthouden? Niet veel, tenzij dat een reis onoverkomelijk kan lijken in het begin maar uiteindelijk wel in een plooi valt en dat die plooi net iets anders kan komen te liggen dan je eerst had gedacht. In ieder geval, beste lezer, ik wens je geluk met jouw reis, naar waar die ook gaat.

Een thesis is tenslotte ook een heuse ondertekingsreis, des te meer al een bewijs dat kleine micropeptiden in staat zijn tot grootse dingen. Die reis maakte ik niet alleen, gelukkig. Er
zijn een heleboel mensen die ik daarom nadrukkelijk en oprecht wil bedanken. Mensen die bijgedragen hebben aan mijn thesisreis van één jaar, mensen die meegewerkt hebben aan mijn bio-ingenieursreis van vijf jaar en ja, zelfs ook mensen die bijgedragen hebben aan mijn levensreis van bijna drieëntwintig jaar.

Graag start ik bij mijn promotor, Prof. Wim Van Criekinge. Bedankt voor de enthousiaste manier waarmee de lessen bio-informatica gegeven werden. Hoewel ik het initieel volgde als een plichtvak, kon net dat enthousiasme mijn interesse voor het vakgebied enorm aanwakkeren. Op zo’n manier dat ik me inschreef voor een keuzevak. Nadien kwam ik zelfs informeren voor een thesis. Ik ben er zeker van dat ik zonder die boeiende manier van lesgeven niet in deze richting was afgeslagen. Bedankt dat ik mijn thesis aan dit labo kon doen en dat ik de kans kreeg om er bio-informatica in praktijk om te zetten.

Iemand zonder wie ik deze thesis nooit voor mekaar had gekregen, is mijn copromotor Gerben Menschaert. Gerben, bedankt om mij te katapulteren in een uitermate boeiende onderzoekswereld. Bedankt voor alle tips, voor het nalezen, voor de tonnen gedeelde ervaring en voor het nodige geduld om het waar nodig nog eens uit te leggen. Ik heb veel geleerd van die manier van werken: steeds vergezeld van een lach en grap, maar ook steeds met de nodige passie voor wat het onderzoek zal opleveren. Het was me een interessant jaar waar ik met veel plezier naar terugkijk.

Verder wil ik ook mijn tutor Jeroen Crappé bedanken. Bedankt om mij te introduceren in een onderwerp waar ik nog niets vanaf wist en waar jij al een hele tijd ervaring mee had. Het lijkt mij een lastige taak om telkens te antwoorden op de vragen van een beginner die de schijnbaar meest banale zaken nog niet weet, en dat in de laatste drukke fase van jouw doctoraat. Desalniettemin legde je hulpvaardig alles uit en zette je me die eerste maanden uitstekend op weg. Bedankt!


Een thesisreis van een jaar is een hele uitdaging, een andere opdracht is een tocht van vijf jaar bio-ingenieur. Daarin slagen zou niet evident geweest zijn zonder het clubje bio-ingenieursvrienden dat er de heerlijkste studententijd ooit van maakte. Middagpauzes, blaarmeersen, cafébezoeken, kaas en wijn, gezellige avonden op kot, weekends, festivals, zelfs een eigen muziekgroep. Daarenboven waren er steeds een paar heel speciale mensen die voor elk verhaal een luisterend oor hadden en altijd klaar stonden voor een babbel. Wat begon als een allegaartje mensen die toevallig samen in de les zaten, evolueerde na twee jaar tot een hechte vriendengroep, na vijf jaar tot vriendschappen die enkel nog maar voor meer geniale momenten zorgden. Samen trotseerden we die eerste jaren en nadien koos elk meer zijn eigen richting. Allen stevig bouwend aan een droom, maar op tijd en

Een nog langere reis wordt ondersteund door een zeer warme thuis. Een kleine hechte familie, een eindje weg van de studentenstad, maar een bron van levenswijsheid en steun voor alles wat ik wilde doen. Een gelukzak ben ik om vanuit zo een familie te kunnen bouwen aan mijn weg.

In het bijzonder wil ik mijn ouders bedanken. Altijd klaar om mij bij te staan, te helpen met raad en daad. Hard werken maar toch proberen van elk mogelijk moment te genieten, zeker als familie. Mama en papa, als er nu twee mensen zijn die mij in de meest fundamentele zin op weg hebben gezet, dan zijn jullie dat. Gaande van letterlijk op de wereld zetten tot mij steunen in alles wat ik doe, in echt alle facetten van het leven. Om jullie te bedanken voor alles wat jullie gedaan hebben, is de beschikbare hoeveelheid briefpapier op deze wereld helaas te klein, maar hoewel ik het niet altijd zeg, ben ik jullie enorm dankbaar voor alles wat jullie al voor mij gedaan hebben en dat is echt een hele hoop.

Tot slot, een gemeende dankuwel aan een reeks geweldige vrienden. Of ik jullie nu ken via muziekscholen, zomerkampen, orkesten, bands, middelbare school of nog iets totaal anders. Overal ken ik wel speciale mensen met wie ik geniale tijden heb meegemaakt. Bedankt vrienden en op naar nog meer van dat!

Steven Verbruggen
Gent, mei 2015
Contents

1 Introduction and outline ................................................................. 1

2 Overview of relevant literature ..................................................... 3
   2.1 Micropeptides, what is it about? .............................................. 3
   2.2 Identification of sORFs and micropeptides ............................... 4
      2.2.1 Computational prediction .............................................. 5
      2.2.2 Mass spectrometry ................................................... 6
      2.2.3 Ribosome profiling ..................................................... 7
      2.2.4 Conclusions on sORF identification ............................... 12
   2.3 Categories of sORFs .......................................................... 12
      2.3.1 Inside non-coding RNAs (ncRNAs) .................................. 12
      2.3.2 Intergenic sORFs .................................................... 13
      2.3.3 Upstream sORFs (uORFs) ........................................... 13
      2.3.4 Downstream sORFs .................................................. 14
      2.3.5 Overlapping a coding sequence .................................... 15
      2.3.6 Intronic sORFs ....................................................... 15
   2.4 Functional characterization .................................................. 15
   2.5 Overview of performed studies ............................................. 16
      2.5.1 Bacteriophages ....................................................... 17
      2.5.2 Bacteria ............................................................ 17
      2.5.3 Funghi ............................................................ 17
      2.5.4 Plants ............................................................. 17
      2.5.5 Lower animals ...................................................... 18
2.5.6 Mammals ....................................................... 19
2.5.7 Human cells .................................................... 19

3 Material and methods ............................................. 21
   3.1 Hardware ....................................................... 21
   3.2 Software ....................................................... 22
       3.2.1 Perl and Bioperl ........................................ 22
       3.2.2 R .......................................................... 22
       3.2.3 SQLite ..................................................... 23
       3.2.4 Perl DBI .................................................... 23
       3.2.5 TopHat and STAR ........................................ 24
       3.2.6 Galaxy ...................................................... 25
       3.2.7 SearchGUI and PeptideShaker .......................... 25
   3.3 Single-exon RIBOsORF pipeline .............................. 25
       3.3.1 Mapping .................................................... 26
       3.3.2 TIScalling ................................................ 27
       3.3.3 Single-exon assembly ................................... 28
       3.3.4 PhyloCSF .................................................. 28
       3.3.5 FLOSS ..................................................... 28
       3.3.6 Translation product database construction ............. 30
       3.3.7 Script-based implementation and Galaxy version ........ 30
   3.4 Splice-aware assembly ........................................ 31
   3.5 Validation with proteomics data .............................. 32

4 Results ............................................................. 35
   4.1 Single-exon RIBOsORF pipeline .............................. 35
       4.1.1 Manual check of mapped reads .......................... 35
       4.1.2 Influence of adaptors ................................... 37
       4.1.3 Filtering steps ............................................ 38
       4.1.4 Identification of sORFs in different annotation classes 42
### List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>adenine</td>
</tr>
<tr>
<td>AA</td>
<td>amino acid</td>
</tr>
<tr>
<td>aTIS</td>
<td>annotated translation initiation site</td>
</tr>
<tr>
<td>BLAST</td>
<td>basic local alignment search tool</td>
</tr>
<tr>
<td>BLAT</td>
<td>BLAST-like alignment tool</td>
</tr>
<tr>
<td>bp</td>
<td>base pairs</td>
</tr>
<tr>
<td>C</td>
<td>cytosine</td>
</tr>
<tr>
<td>CAI</td>
<td>codon adaptation index</td>
</tr>
<tr>
<td>CDS</td>
<td>coding sequence</td>
</tr>
<tr>
<td>CFTR</td>
<td>cystic fibrosis transmembrane conductance regulator</td>
</tr>
<tr>
<td>CHX</td>
<td>cycloheximide</td>
</tr>
<tr>
<td>CPAN</td>
<td>comprehensive Perl archive network</td>
</tr>
<tr>
<td>CRAN</td>
<td>comprehensive R archive network</td>
</tr>
<tr>
<td>CRITICA</td>
<td>coding region identification tool invoking comparative analysis</td>
</tr>
<tr>
<td>CHX</td>
<td>cycloheximide</td>
</tr>
<tr>
<td>CSF</td>
<td>codon substitution frequencies</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DBD</td>
<td>database driver</td>
</tr>
<tr>
<td>DBI</td>
<td>database interface</td>
</tr>
<tr>
<td>EMBL</td>
<td>European molecular biology laboratory</td>
</tr>
<tr>
<td>EMT</td>
<td>emetine</td>
</tr>
<tr>
<td>ER</td>
<td>endoplasmatic reticulum</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tag</td>
</tr>
<tr>
<td>FLOSS</td>
<td>fragment length organisation similarity score</td>
</tr>
<tr>
<td>FPKM</td>
<td>fragments per kilobase sequence length per million reads</td>
</tr>
<tr>
<td>GEO</td>
<td>gene expression omnibus</td>
</tr>
<tr>
<td>G</td>
<td>guanine</td>
</tr>
<tr>
<td>GTI-seq</td>
<td>global translation initiation sequencing</td>
</tr>
<tr>
<td>HARR</td>
<td>harringtonine</td>
</tr>
<tr>
<td>HCT116</td>
<td>human colon tumor 116 cells</td>
</tr>
<tr>
<td>HMGCR</td>
<td>3-hydroxy-3-methylglutaryl-coenzym A reductase</td>
</tr>
<tr>
<td>HSPC300</td>
<td>hematopoietic stem progenitor cell 300</td>
</tr>
<tr>
<td>LC</td>
<td>liquid chromatography</td>
</tr>
</tbody>
</table>
LC-MS/MS  liquid chromatography followed by tandem mass spectrometry
lncRNA  long non-coding RNA
LTM  lactimidomicyn
MAF  multiple alignment files
mESC  mouse embryonic stem cells
miscRNA  miscellaneous RNA
mgf  mascot generic format
Mkks  McKusick-Kaufman syndrome
mRNA  messenger RNA
MS  mass spectrometry
MS/MS  tandem mass spectrometry
NCBI  national center for biotechnology information
ncRNA  non-coding RNA
Nelfb  negative elongation factor B
nt  nucleotides
OMSSA  open mass spectrometry search algorithm
ORF  open reading frame
Perl  practical extraction and report language
PRIDE  proteomics identifications
PMY  puromycin
PSM  peptide-to-spectrum match
QTI-seq  quantitative translation initiation sequencing
RDBMS  relational database management system
RIBOseq  ribosome profiling
RP-LC  reverse phase liquid chromatography
RPF  ribosome-protected RNA fragment
RPKM  reads per kilobase of coding region per million aligning reads
RNA  ribonucleic acid
RRS  ribosomal release score
SAM  sequence alignment map
SEC  size exclusion chromatography
Smim20  small integral membrane protein 20
snoRNA  small nucleolar RNA
sORF  small open reading frame
sprcRNAs  short polycistronic ribosome-associated coding RNAs
SQL  standard query language
STAR  spliced transcripts alignment to a reference
T  thymine
TIS  translation initiation site
TOC  translated open reading frame classifier
TrEMBL  translated EMBL
tRNA  transfer RNA
UCSC  university of California, Santa Cruz
<table>
<thead>
<tr>
<th>abbr</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>uORF</td>
<td>upstream open reading frame</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
<tr>
<td>XML</td>
<td>extensible markup language</td>
</tr>
</tbody>
</table>
Abstract

Recently, a new promising sequencing technique was described: ribosome profiling. With this technology, it is possible to analyse the mRNA fragments that are captured in the ribosomes. Therefore, translation can be mapped very precisely and with an extreme sensitivity. As a result, the proteome can be studied on the mRNA level.

Micropeptides on the other hand, are a recently described class of bio-active peptides, which are translated from small open reading frames (sORFs, <100 codons). Up to now, this class has been missed completely in gene prediction as standard gene prediction tools use 100 codons as a lower border. Small coding sequences can occur more easily by chance, so adapted strategies are needed to select for the true coding sORFs.

The identification of functional micropeptides out of the new ribosome profiling data flow, which comes in handy to single out genuine sORFs, requests the development of a custom pipeline. Therefore, RIBOsORF was developed. It is a pipeline which identifies all possible sORFs out of ribosome profiling data and analyses their peptide coding potential based on some score statistics: ribosomal coverage score, PhyloCSF conservation score and FLOSS.

In this thesis, two aspects of this pipeline are emphasised: the implementation of FLOSS, a filter to deprive the RIBO-reads of possible noise based on fragment length similarities, and the adaptation of the pipeline to make it aware of splicing.

Afterwards, the aim is to find validation for the detected sORFs. The most desired validation would be obtained out of mass spectrometry-based proteomics data. Some prudent proofs could be found but extraction and enrichment protocols in proteomics will need further improvements in order to obtain better validation results.
Korte samenvatting


Micropeptiden, anderzijds, zijn een recent beschreven klasse van bio-actieve peptiden die getranslateerd worden vanuit kleine open leesramen (sORFs, <100 codons). Tot nu toe werd deze klasse volledig gemist bij genpredictie omdat standaard predictieprogramma’s werken met een ondergrens van 100 codons. Omdat kleine coderende sequenties veeleerder per toeval kunnen bestaan, zijn aangepaste strategieën vereist om de sORFs die echt coderen, te selecteren.

De identificatie van functionele micropeptiden vanuit deze nieuwe ribosoomprofieleringsgebaseerde datastroom, die zich uitermate leent om de werkelijk coderende sORFs te onderscheiden, vergt de ontwikkeling van een aangepaste pipeline. Daarom werd de RIBOsORF pipeline ontwikkeld. Deze pipeline identificeert alle mogelijke sORFs vertrekende van ribosoomprofieleringsdata en analyseert hun codeerpotentiaal gebaseerd op enkele statistieken: ribosomale coverage, PhyloCSF conservatiescore en FLOSS. In deze masterthesis ligt de nadruk op twee aspecten van deze pipeline: de implementatie van FLOSS, een filter om RIBO-reads te onderscheiden van mogelijke ruis gebaseerd op similariteit tussen fragmentlengtes, en de aanpassing van de pipeline om rekening te houden met splicing.

Nadien wordt getracht om de gedetecteerde sORFs te valideren. De meest gewenste validatie kan verkregen worden met massaspectrometriedata. Enkele voorzichtige bevestigingen konden gevonden worden maar verbeteringen in extractie- en aanrijkingsprotocols zullen nodig zijn om beter validatie te kunnen doen aan de hand van proteomics.
Chapter 1

Introduction and outline

Next generation sequencing methods broadened our knowledge about the genomical structure and its translation. As a rule, a protein is obtained if an open reading frame (ORF) is translated. An ORF consists of a sequence of in-frame codons beginning at a start codon and ending with a stop codon. Additional mechanisms exist to introduce extra variation in protein expression like alternative translation initiation sites, alternative splicing, transcript editing, frame shifting and post-translational modifications. More recently, an additional feature of the proteome complexity was identified: the translation of short open reading frames (sORFs, <100 amino acids (AA)) putatively leading to the existence biologically active micropeptides.

The main objective is to develop a pipeline, RIBOsORF, which identifies all possible sORFs in Eukaryotes out of the results of a new next gen sequencing technique called ribosome profiling (RIBOseq). Subsequently, sORF targets need to be analysed for their coding potential in silico. Out of these results, a search space for mass spectrometry (MS)-based identifications can be constructed, ending up with a proteogenomic approach to validate the outcome.

In chapter 2, an overview of the most important literature concerning this topic is given, including information about sORFs, micropeptides, identification strategies, ribosome profiling and performed searches. Afterwards, the practical work of this research is described as chapter 3 explains the hardware, the software, all the building blocks of the RIBOsORF pipeline and the tools and strategies used to validate the results with MS-based proteomics. RIBOsORF was further tested on available RIBOseq data and these results are given in chapter 4. In chapter 5 the results and the pipeline itself are discussed. Future perspectives are considered in chapter 6. Finally, chapter 7 closes this master thesis by formulating a general conclusion.
Chapter 2

Overview of relevant literature

2.1 Micropeptides, what is it about?

Despite their size, small peptides are known to occupy important functions in living organisms [1]. Known classical small peptides include neuropeptides and peptide hormones [2]. These peptides are generated out of larger precursor proteins carrying an N-terminal signal sequence [2, 3, 4]. When these peptides accumulate in sufficient levels outside the cell, they may function in cell-cell signaling, for example by interacting with a G-protein coupled receptor [4].

More recently, a second class of small peptides was discovered [5, 6]. These non-classical peptides are encoded by sORFs and they never carry an N-terminal signal sequence. In that way, they are released directly in the cytosol [7] (figure 2.1) and may therefore execute totally different functions than the first class of small peptides. However, it could be shown that micropeptides are not bound to the cytoplasm of just one cell. For example, brick1 mutant cells, which are in direct contact with wild type cells, appear to still have a wild type phenotype because of a non-cell-autonomous behaviour of the micropeptide [8]. The molecular mechanism of this non-cell-autonomous effect is unknown, but their size makes these small peptides a good candidate as an intercellular close-distance communicator. Some explanations include the fact that they can be transported by microvesicles, exosomes or transferred towards the secretory machinery by the TAP transporter of the endoplasmatic reticulum (ER) [3]. First known examples of these peptides, also known as micropeptides, have a function in morphogenesis [3, 9].

Brick1 also shows that conservation can be a very particular characteristic of micropeptides. It was first identified in maize [6], but orthologues (HSPC300; hematopoietic stem progenitor cell 300) have been found in Drosophilla [10].

Traditionally in gene prediction tools, a limit of 100 amino acids (equivalent with 300 base pairs (bp)) was assumed for identifying an ORF [12], reasoning that ORFs with lower numbers of codons would have a much higher chance of existing just by chance,
CHAPTER 2. OVERVIEW OF RELEVANT LITERATURE

Figure 2.1: The location of classical peptides and micropeptides in a cell. Classic bioactive peptides contain a signal sequence at their N-terminus. Therefore, they will be sent towards the secretory pathway so that they can be secreted and execute their function extra-cellular. Micropeptides however, lack this N-terminal signal and will be released in the cytoplasm [11].

elevating the frequency of picking up meaningless sequences tremendously. Identifying possible sORFs is rather easy but the hardest job is to determine whether this sORF is encoding a functional micropeptide or is just present by chance.
The effect of this 100 amino acids limit throughout the years is very clear when doing a manual search in the SwissProt protein database. From the 20 193 reviewed human proteins, only 651 were smaller than 100 amino acids at the time of this writing. The smallest translated peptide documented so far has six amino acids [13].

Micropeptides encoded by sORFs are sometimes divided in subclasses. Aspden et al. [14] categorised sORFs in two distinct types after studying the translation in Drosophilla. The ‘longer’ type contains sORFs of around 80 amino acids. These resemble strongly to canonical proteins from a translation, conservation, splicing and protein motifs perspective. The ’dwarf’ sORFs however are much shorter (around 20 amino acids), are less conserved and are mostly situated in 5’ untranslated regions (5’UTR) or non-coding RNAs (ncRNAs). The majority of them will definitely be encoded from a single exon.

2.2 Identification of sORFs and micropeptides

Conventional methods can miss micropeptides [15, 16, 17]. For example, when performing gel-based methods which are not optimized for detecting small molecules, micropeptides
can simply run off without being detected [18, 19]. Also, the staining with Coomassie Blue is less efficient for small proteins [19]. Therefore, particular identification strategies must be developed to search for functional micropeptides.

2.2.1 Computational prediction

Formerly, most gene annotation programs discarded ORFs smaller than 100 codons, stating that these cannot hold coding potential [20]. In a search for micropeptides, this limit is intentionally used in the opposite direction, looking for ORFs under this border. However, in that case, constructing sORFs will result in lots of false positives, forming an ORF just by chance [21]. So, extra criteria are required to select the true coding sORFs. The introduction of multiple criteria lowers the chance that a false ORF is predicted as valid, but on the other hand also increases the chance that a true ORF is not predicted as coding [22]. Another problem is the unavailability of a large experimental validation set of sORFs to train computational prediction models [20].

Sequence-based prediction

Because sORFs mostly lack a lot of features that longer canonical ORFs do possess (e.g. termination signals, splice sites, polyadenylation signals...), one has to rely on other measures for coding potential on sequence level [23]. Nevertheless, it is assumed that expressed ORFs have a biased codon usage, a useful hallmark for building ab initio approaches. Different indices could be proposed to study this characteristic, e.g. the codon adaptation index (CAI) is an index that measure the adaptation of the sequence towards a usage of the optimal codons amongst different synonymous codons [20, 24]. Another method, sORFfinder, is based on the occurrence of hexamer frequencies [23, 25]. A third approach, CRITICA (Coding Region Identification Tool Invoking Comparative Analysis), uses hexamer frequencies as well, however, the hexamer usage information is derived iteratively from the data. Besides that, an additional comparative analysis is executed. Regions of DNA will be aligned to related sequences in a DNA database. If these aligned database sequences have a greater than expected amino acid identity, than this can be interpreted as an evidence for coding potential [26]. CRITICA is originally designed to work in bacteria, however, it can also be used for Eukaryotes [27].

Conservation-based prediction

It is generally accepted that conservation is a hallmark of functional sORFs. By performing the Basic Local Alignment Search Tool (BLAST) against a close species or by doing a multiple alignment against close and more divergent species, a conservation score can be calculated for each theoretical sORF. Based on these scores, false positives can be eliminated. Other controls include examination of syntenic genomic (homologous) locations,
although it can result in false negatives due to gene translocation. Also, protein-coding sequences (CDS) have a difference in the ratio of synonymous (Ks) against non-synonymous (Ka) substitutions rates. This ratio can be used as an additional filter for functionality of sORFs [28].

PhyloCSF uses this methodology as well. First, a multiple alignment with different other genomes is performed along a fixed evolutionary tree, accounting for the different branch distances. Afterwards, the substitutions in the multiple alignment are considered. A higher number of synonymous substitutions results in a higher score whereas non-synonymous substitutions have a decreasing effect on the score [29].

Another conservation score that is frequently used is PhastCons, based on a two-state phylogenetic hidden Markov model [30].

Conservation-based approaches need sufficient alignment data from other organisms to give trustworthy results. This makes it a less suitable option for classification in novel (unannotated) transcripts, so it is better to combine a conservation-based approach with other computational techniques [31].

One should be very cautious when pipelines solely use conservation metrics for searching sORFs, as the focus will be on highly conserved sORFs then. Downstream in vivo testing of the resulting best candidates might show a biased outcome towards essential biological processes like development, morphogenesis, etc. as only the conserved ones were further studied. Non-conserved sORFs should not be totally neglected either, because species-specific sORFs can play a biological role too [16].

One can also search for homology (somehow alike to conservation) of some domains in the sequence by performing BLAST. The expected value (E-value) threshold can be adjusted to set the false positives and negatives rate but still, BLAST loses a lot of its strength for sORFS compared to searches for sequences longer than 100 codons [32, 20]. Also, the BLAST-like Alignment Tool (BLAT) [33] could be used. It allows to search a set of short sequences against a full indexed genome.

### 2.2.2 Mass spectrometry

Mass spectrometry (MS) is the main technique for direct detection of proteins and peptides, being the golden standard in proteomics research. In proteomic studies, a lot of tandem mass spectrometry (MS/MS) fragmentation spectra remain unidentified because there is no peptide-to-spectrum match (PSM) above the threshold. One of the reasons is a possible absence of the peptide in the search space [28]. Parts of these absent peptides can probably be included by taking micropeptides in consideration. Other reasons for failing in finding a valid PSM can be attributed to post-translational modifications or single amino acid variants.

Micropeptides can be included in the search space but the screening for micropeptides with MS remains difficult as standard extraction protocols focus on proteins of at least 10 kDA. Also, micropeptides can be subject of a rapid degradation and are easily lost during
2.2. IDENTIFICATION OF SORFS AND MICROPEPTIDES

purification steps of the sample preparation protocol. Furthermore, the reference database information for identification of small translation products remains too sparse, although information from computational approaches can be used [16].

There are a few studies where MS is used for identification of small translation products, all in K562 (human leukemia) cells. Slavoff et al. [7] found 86 previously unannotated micropeptides in human cells with liquid chromatography followed by tandem mass spectrometry (LC-MS/MS) in an adapted design to enrich for small translation products. They matched their MS/MS data against a custom database of the human transcriptome (RefSeq) combined with assembled transcripts of a comprehensive RNAseq experiment. Taken together with some other studies in K562 cells, several hundreds of micropeptides have been identified with MS [7, 34, 35].

2.2.3 Ribosome profiling

Preliminary techniques

RNAseq and measuring mRNA abundance with microarrays are very important techniques in expression studies. However, due to post-transcriptional regulation, mRNA abundance is an imperfect approximation of expression at protein level. Moreover, prediction of the translation product from the transcript is not always possible due to several effects like internal ribosome entry sites, initiation at non-AUG codons and nonsense read-through [36].

Often, genome-wide translation studies were done by polysome profiling. Herein, polysomes are purified by sucrose gradient separation. Then, the RNA abundance in this polysomal fraction can be analysed by RNAseq or microarray analysis [37, 38]. However, this technique has a limited resolution and accuracy and moreover, it is not very sure that the bound transcript will be translated nor which of the reading frames will eventually be translated [36].

General aspects of ribosome profiling

Ribosome profiling (RIBOseq), the sequencing of ribosome-protected RNA fragments (RPFs), has been proven to be a very useful technique to find sORFs [39]. Ribosomes have the ability to protect a translated RNA segment (around 30 nucleotides (nt) in length) from nuclease activity [40], whereas unprotected mRNA parts will be degraded by the nuclease. With the deep sequencing of the residual ribosome-protected mRNA fragments, it is possible to get a genome-wide view of translation with subcodon resolution, because the position of the active ribosome can be determined at nucleotide level [36]. In that way, also the reading frame can be deduced using additional ribosome-halting treatments prior to measurement (described below). The general procedure is given in figure 2.2.
Figure 2.2: General mechanism of ribosome profiling with cycloheximide (CHX) treatment. First, ribosomes are halted by treating the cells with CHX. Then, cells are lysed so that the nuclear material will be freed. After nuclease digestion, only ribosome-protected fragments will remain. The ribosomes (and thus also the RPFs captured within) are purified using size exclusion chromatography (SEC) or sucrose cushion ultracentrifugation. Finally, the RNA fragments are released from the ribosomes and can be sequenced by means of next generation sequencing [41].
Translation can be halted by using flash freezing [42], although the true strength of RIBOseq arises from the usage of different antibiotics. Cycloheximide (CHX) is often used in ribosome profiling due to its stabilising effect on ribosomes and its inhibition of translation. CHX binds close to the exit (E-)site of the large ribosomal subunit. Therefore, the deacetylated transfer RNA (tRNA) cannot be released and ribosomal translocation is no longer possible. The antibiotic emetine (EMT) has a comparable effect on cells, however, emetine-treated cells show longer footprints, suggesting that a different ribosome conformation is stabilised by EMT, protecting a longer part of the mRNA. It could be proven that a brief treatment of cells with elongation inhibitors did not significantly changed the outcome [42].

Lactimidomycin (LTM) has a very similar structure compared to CHX but it is larger. As a result, LTM acts preferentially on initiating ribosomes where the E-site is still empty. The initiating ribosomes will stall and the elongating ones will run-off [43]. In that way, two RIBOseq experiments are usually performed in parallel, one with CHX for measuring the overall translation profile and one with LTM to determine the (alternative) translation initiation sites (TISs). Harringtonine (HARR) is another translation initiation inhibitor besides LTM but binds free 60S ribosomal subunits whereas LTM binds the total 80S complex assembled at the start site. The full inhibitory mechanism of HARR is however still unknown. Lee et al. [43] performed a comparative study between HARR and LTM and found that HARR shows also some RPFs downstream of the start codon so that LTM gives a higher resolution for mapping TISs.

Generally, coverage is used as a general measure of translation, mostly expressed as the amount of ribosome-protected fragments per kilobase sequence length per million reads (FPKM). Other statistics can also be calculated to introduce more controls. For example, the coverage uniformity can be used to check if the ribosome-protected fragments are equally positioned across the sequence and is defined as:

\[
\text{Coverage uniformity} = \frac{\# \text{Reads in the first half of the sequence}}{\# \text{Reads in the second half of the sequence}}
\]  

(2.1)

Another statistic is the R-value [44]:

\[
R = R_{\text{LTM}} - R_{\text{CHX}}
\]  

(2.2)

\[
R_k = \frac{X_k}{N_k} \times 10
\]  

(2.3)

with \(X_k\) = number of reads on position \(x\) for data \(k\)

\(N_k\) = total number of reads on the transcript for data \(k\)

and \(k = \) LTM/HARR or CHX

RIBOseq is very useful as a tool to experimentally confirm the in silico sORF predictions. In that view, it is probably more efficient to first identify the candidate sORFs from the RIBOseq data (which will result in a smaller amount of targets than with a virtual scan).
Afterwards, the bioinformatic measures can be calculated as a control for all the targets. In that way, the overall process will consume a smaller computational load in comparison to a genome-wide in silico screen followed by RIBOseq validation [17].

RIBOseq can also be applied to investigate other translational features beside sORFs. Ribosomal pausing is supposed to have a role in cotranslational regulation, cotranslational folding and secretion. Ribosome profiling helps in measuring this phenomenon [36]. Also the identification of near-cognate start sites is possible [44].

Advantages over other techniques

In fact, RIBOseq measures the production of proteins and not the abundance. So MS will still be needed to include the steady state and degradation aspects of proteins [39]. Protein abundance is a consequence of synthesis and degradation, but synthesis seems to have the major influence [45]. Therefore, protein synthesis gives an estimation of the protein abundance.

Besides that, RIBOseq has several advantages over MS. RIBOseq exceeds MS in the dynamic range in which proteins can be measured [17, 39], what makes this technique very useful as sORFs are possibly transcribed at low levels [46]. Also, MS always needs a reference database for identification of translation products and is therefore limited to the identification of already known proteins. RIBOseq does not have this limitation and can discover new peptides. This is extremely useful as lots of micropeptides still remain uncharacterized. RIBOseq can also be used to expand the reference database for MS with newly derived peptides [44], so that the combination of RIBOseq and MS makes a good tandem.

RIBOseq makes it possible to determine exact ribosome positions and to perform more quantitative expression studies. This cannot be obtained using polysome profiling techniques. However, polysome profiling monitors the translational status of entire transcripts, which cannot be deduced from footprint sequencing data. In that way, polysome profiling can distinguish a uniform decrease of ribosomes on all transcript copies and a repression of the translation of only a subgroup of the transcripts, for example a certain transcript isoform. Therefore, the combination of ribosome and polysome profiling can be used to get a more complete insight in translational regulation [39].

Disadvantages

However, RIBOseq does not come without limitations. First, like every other experimental approach, RIBOseq needs high quality data and the putative peptide needs to be translated at the moment of sampling to be detected [31]. Also, some known truly non-coding sequences show nevertheless ribosomal coverage. For example, H19, which is a non-coding RNA sequence in mice [47], shows ribosomal occupancy [48]. So, RIBOseq data cannot be used as the only proof for translation and must be combined with other approaches [49].
2.2. IDENTIFICATION OF SORFS AND MICROPEPTIDES

Ribosome profiling data also contains lots of contaminants which have to be filtered out. The most dominant contaminant is ribosomal RNA but also small nucleolar RNAs (snoRNAs), tRNAs and other ncRNAs can result in false sequenced reads. Therefore, mostly additional computational analyses are used to distinguish the truly translated sequences from noise [31].

The translated ORF classifier (TOC) is a classifier that uses different features of sequences to distinguish coding sequences from header (5’) and trailer (3’) sequences, after the program was trained with known RefSeq information. Used features are the general expression level, the ratio of coverage inside versus outside the sequence, the percentage of transcript covered by the ORF and the absence of RPFs downstream of the ORF [50].

The ORFscore measures the bias of the RPFs to the first frame. In that way, bias to the first frame results in a positive value, a uniform frame distribution results in a value close to zero and a bias to the second or third frame results in a negative value [51].

The ribosomal release score (RRS) detects protein-coding transcripts by scoring the termination of the translation at the end of the ORF [52]. The Fragment Length Organisation Similarity Score (FLOSS) [53] deduces the coding probability from its ribosome footprint length distribution but this is explained in detail in part 3.3.5.

Variants of ribosome profiling

Adaptions can be applied to exclude sequences that are just sporadically bound by non-functional monosomes or single subunits. Aspden et al. [14] first performed a polysomal fractionation (by sucrose gradient separation) to select RNA transcripts which were bound by 2 to 6 ribosomes. On this subset, RIBOseq was done afterwards. The total procedure was called PolyRIBOseq.

In global translation initiation sequencing (GTI-seq), one performs a ribosome profiling with CHX and LTM in parallel. In that way, the LTM profile can be corrected with the CHX profile to distinguish elongating ribosomes from ribosomes stalled at a start codon. In that way, a better identification of the possible TISs is achieved [43].

Puromycin (PMY), another antibiotic, leads to a release of all elongating ribosomes. Followed by a general arrest of all ribosomes, this can be used to determine the TISs [54].

GTI-seq lacks however some quantitative features because LTM treatment requires an incubation period during which the elongating ribosomes run off and the initiating ones stay on the transcript. This incubation period can result in some artefacts, for example new ribosomes that start initiation during incubation, which hampers the quantitative relation. Therefore, PMY can be used instead of an incubation period to get rid of the elongating ribosomes, leading to quantitative translation initiation sequencing (QTI-seq). In that way, a sequential treatment, first with LTM and later with PMY, leads to a stronger quantitative measurement. Moreover, because the ribosomes are directly captured from lysates, QTI-seq can be performed outside of cell culture, making a transition to studies in tissues possible [55].
CHAPTER 2. OVERVIEW OF RELEVANT LITERATURE

2.2.4 Conclusions on sORF identification

To conclude this section about identification strategies, it can be stated that a combination of virtual and experimental techniques is probably the best way to have a comprehensive look on the identification of micropeptides [17]. Because of their size, sORFs contain a limited amount of information, so that it is important to validate a putative identification with multiple independent data [18].

2.3 Categories of sORFs

2.3.1 Inside non-coding RNAs (ncRNAs)

An excessive amount of possible sORFs is found in earlier supposed non-coding genes, especially on long non-coding RNAs (lncRNAs) [42]. This is important as there are different ideas about the functions of ncRNAs. First, there could be an intrinsic functional role of the ncRNA itself. Second, the ncRNA could regulate the transcription of flanking genes. Third, it could encode micropeptides [3]. Possibly, also a dual role (both coding and non-coding) is possible [14].

Ribosome profiling questioned the general non-coding character of lncRNAs. Although some transcripts show a clear function as RNA (e.g. Xist, HotAir [56, 57]), highly ribosome-occupied regions were found [42, 43] in a substantial part of the list of candidate lncRNAs [58, 59]. Ingolia et al. [42] assumed that these lncRNAs contain small putatively coding sequences and classified them as short polycistronic ribosome-associated coding RNAs (sprcRNAs), clearly separated from 3’UTR-sequences which are believed to be poorly translated. However, as stated in chapter 2.2.3, ribosome profiling is not a fool-proof method and one must be aware of possible translational noise [60]. In reaction to the ribosome-covered regions in lncRNAs, Guttman et al. [52] investigated this data deeper by developing the ribosomal release score (RRS; see chapter 2.2.3). With this score, most of the lncRNAs were classified in the same group as the known truly non-coding RNAs (like e.g. snoRNAs). Recently, Ingolia et al. [53] redressed this reaction by presenting a different approach. They looked at the size distribution of the ribosome-protected fragments for each sORF separately (FLOSS, see chapter 3.3.5). In that way, they were able to identify the true 80S ribosomal footprints by discriminating coding from truly non-coding sequences. However, the debate about the existence and the function of small translation products in different lncRNAs remains ongoing [14].

Another viewpoint describes the collection of ncRNAs as a translational landscape (see figure 2.3), ranging from functional proteins over translation itself as a process of regulation to truly non-coding RNAs [31]. In this light, leader-like transcripts were defined as transcripts that resembled uORF-containing 5’ leaders of classical genes in translation patterns (see also chapter 2.3.3). Tests with ribosome profiling and MS indicated that
2.3. CATEGORIES OF SORFS

There is a continuum (also called a translational landscape) ranging from protein-coding to truly non-coding RNA transcripts. In between these extrema, there are cases where the translation process itself, as a way of regulation, is more important than the resulting peptide [31].

these transcripts can generate micropeptides, however, these peptides are rarely conserved and believed to be rather non-functional [50].

2.3.2 Intergenic sORFs

It is stated that some intergenic regions in Eukaryotes can be transcribed [61, 62], so possible sORFs can also be located in regions in between the currently considered genic locations. Hanada et al. [23] found a much higher coding potential in some intergenic sORFs of *A. thaliana* than originally expected. Around 5-7% of the possible sORFs in intergenic regions are very likely producing novel peptides [23, 27]. With overexpression studies of sORFs, approximately 10% of the investigated intergenic sORFs resulted in visible phenotypic effects in *A. thaliana* [63].

2.3.3 Upstream sORFs (uORFs)

The existence of ORFs in leader sequences was discovered by doing the first systematic searches in mRNA sequences [64]. However, the unravelling of their main function, controlling the translation of a longer ORF downstream, happened much later [3, 14]. There is a high occurrence of uORFs: approximately 35-50% of the genes in RefSeq contains RPF-covered uORFs [16, 50]. They can overlap with the main protein-coding ORF, sharing regions with dual coding activity [65].

Upstream ORFs can have an *in cis* regulatory function in several ways. They can control translation of a downstream CDS by regulating downstream reinitiation [3, 42]. Otherwise, the uORF can capture a fraction of the scanning complexes and decrease (attenuate) the downstream translation. This was proven in human and mouse cells for the *Atf4* transcript
Finally, uORFs can encode a micropeptide [35, 43]. In that view, the term ‘peptoswitch’ is used [16]. Small regulatory peptides can influence the expression of the downstream main CDS upon interaction with a small molecule, eventually with some help of an intermediary. For example, the CPA1 transcript in S. cerevisiae contains an uORF coding for such a small regulatory peptide. When, during translation of the uORF, there is an interaction with arginine, the ribosome can stall, still bound to the transcript and the micropeptide that is being synthesised. In that way, the whole transcript can be degraded by nonsense-mediated decay which of course decreases the possible translation of the main CDS downstream [67]. However, any positive or negative correlation between the translation of an upstream ORF and its downstream main ORF, that would be excepted under in cis regulation, could not been 100% clearly observed yet [14]. Mostly, more factors than solely the uORF will influence the translation. For example, Lukowski et al. [68] discovered that the combination of an uORF and the secondary mRNA structure of the 5'UTR were determining in cystic fibrosis transmembrane conductance regulator (CFTR) expression control.

A small regulatory peptide could also have a function in trans by binding and regulating other cellular compounds and sequences than those lying on its transcript [3]. In addition, the translation product of an uORF would be possible to have more functions than mere regulation and it can also have its own separate transcript rather than being included in the longer transcript of the canonical downstream ORF [28].

Upstream ORFs start mostly from a non-canonical TIS because, following the leaky scanning model, it would be more difficult for a preinitiation complex to bypass an upstream canonical initiator and start downstream as a canonical TIS is generally stronger than a non-canonical one. The percentage of ribosomes initiating at the upstream start codon, is determined by the environment around that upstream start codon, called the sequence context. Ribosomes that pass this upstream start codon without starting translation, can scan through and initiate at a stronger start codon further downstream. Leaky scanning occurs mostly in uORFs overlapping with the main ORF. Besides leaky scanning, ribosomes can also translate a short uORF and reinitiate at a downstream TIS. This is because some translational factors remain associated with the scanning ribosomal subunit so that reinitiation will be easier. Reinitiation is however assumed to be rather inefficient [16, 43].

### 2.3.4 Downstream sORFs

It is assumed that ORFs in the trailer sequence (3'UTR) are poorly translated [42], so identifying sORFs in these regions attracted little attention so far. However, in general, 3' trailer sequences are longer than 5' leader sequences, so sORFs are more abundant in trailer regions [69, 70]. In 5' leader sequences there are thus less ORFs, even less than expected by chance. This could be explained by assuming that there is selection against ORFs by chance, so that only the ORFs with a real function will be preserved. In 3' trailer sequences, however, the amount of ORFs lies close to the amount expected by chance. This
2.4. FUNCTIONAL CHARACTERIZATION

could explain why most of the ORFs, that are present in the trailer regions, are present by chance and with no real meaning. This also supports the fact that the 3’UTR region can be used as a valid negative control [70]. In addition, ribosome profiling studies show that 3’ trailer sequences are relatively devoid of ribosomes [42, 52]. The expression of micropeptides from these 3’UTR trailer sequences is thus likely to be rare.

2.3.5 Overlapping a coding sequence

Translation from more than one reading frame was thought to be impossible before the first functional peptides, originating from an alternative reading frame, were found [71, 72]. For generating two different translation products from the same genetic material, two different mechanisms are described. First, the ribosome can be triggered to execute a reading frame change by ribosomal frame shifting [73]. Besides that, alternative reading frames can be used through alternative splicing [74].

Recently, Baranov et al. [65] defined a score for identifying dually coding regions. They estimated that minimal 1% of the sequences in the human genome are dually coding. This is an underestimation mainly due to the limited depth of the applied sequencing and the filters used for eliminating false positives. Michel et al. [75] also obtained such an estimation but by using ribosome profiling and triplet periodicity.

Vanderperre et al. [76] constructed an online database of human alternative open reading frames (hAltORF1) wherein the majority of the listed ORFs have a sORF maximum length limit of 300bp.

2.3.6 Intronic sORFs

When performing computational searches in intronic regions, less sORFs are found than expected under an even distribution. This confirms the general non-coding character of the intronic regions. The sORFs that were predicted in the intronic regions of well-known protein-coding genes are frequently used as negative controls [17, 25].

Furthermore, the identification of sORFs, based on ribosome profiling, could also point to mis-annotation of the canonical protein product (e.g. mis-annotated exons and introns, N-terminal extensions, etc.). In that way, searching for small translation products can help the re-annotation of the genome.

2.4 Functional characterization

Recent innovations in computational prediction and proteomics accelerated the identification enormously. The contribution of these new translation products to the cellular

1http://haltorf.roucoulab.com/
processes however remains unclear in many cases [16]. A high number of loci, putatively encoding micropeptides, are still located in intergenic regions and the final goal is to discover the true function of these loci [28]. Micropeptides can probably be involved in key cellular processes like transport, intermediary metabolism, chromosome segregation, genome stability,... [21]. In bacteria, where strategies for characterization are much easier, there are some sORFs demonstrating strong phenotypes, but none is found with an essential function for viability [18]. For Eukaryotes however, this does not hold since sORFs are described that result in lethal phenotypes upon inactivation [46, 77].

So, to retrieve the function of the candidates, a gene-by-gene workflow should be carried out to discover the function accurately. Hanada et al. [63] were able to prove the functionality of some intergenic sORFs with overexpression studies in *A. thaliana*. Results showed that many micropeptides are involved in morphogenesis. Knock-out mutant studies were also successfully applied aiding the functional annotation. Kastenmayer et al. [21] used deletion mutants to do a functional analysis of some theoretical sORFs in *S. cerevisiae*. They constructed a collection of 247 mutants, each with a deletion in a sORF. Phenotypic analysis identified 22 sORFs that were essential for growth or for protection against stressful conditions like higher temperatures, non-fermentable carbon sources and DNA damage. Such a routine analysis is well feasible in yeast but is impossible to achieve on the same scale in higher organisms.

Thus, functional characterisation cannot be performed with the same high-throughput as identification. Also, combining different pieces of evidence should be the optimal way to finally conclude the function of a locus [31]. Furthermore, there are some extra pitfalls. First, it is not always the actual translation product that carries the function. The translation process itself, rather than the translated peptide, can play a regulatory role [31]. Second, sORFs are small, so it is difficult to insert mutations, which makes it hard to do a functional analysis [3, 16]. Also, their small size limits the production of antibodies against them. Third, mostly, mutations in genes for small proteins do not result in a clearly different phenotype on their own, as could been observed for bacteria [18].

Other functional aspects could be interesting to examine as well. The study of interactions can be a very useful way to characterise putative small proteins, although study through tagging of the micropeptide can be misleading as the tag can be much larger than the micropeptide itself [18]. Caution must be taken in order to evade unwanted effects of the tag size. Moreover, degradation and post-translational modification of small proteins could be considered [18], because a large fraction of the translation products seems to be highly unstable and is degraded very soon after production [31].

### 2.5 Overview of performed studies

Below, a short overview of the most important micropeptide studies is listed.
2.5. OVERVIEW OF PERFORMED STUDIES

2.5.1 Bacteriophages

Kobiler et al. [78] found a small peptide of 22 amino acids, translated from the CIII gene, in the bacteriophage λ. This micropeptide inhibits the function of the membrane-bound bacterial protease FtsH, preventing the protease from degrading the transcription factor λ cII, which is essential for the lysogenic cycle of the virus. This example could suggest the existence of small peptides in bacteriophages and other viruses.

2.5.2 Bacteria

Storz et al. [18] list different examples of small proteins in bacteria and draw some conclusions from these findings. Most of the bacterial small peptides are not conserved over a greater group of bacteria, but they are rich in hydrophobic amino acids like leucine, isoleucine and valine and are thus frequently embedded in membranes. An α-helical conformation is the most common structure found in small bacterial transmembrane micropeptides. Other functions of micropeptides found in bacteria are: facilitation of interactions between protein domains, induction of conformational changes and action as a membrane anchor. Micropeptides are less resistant to thermal noise than large proteins. Therefore, it is possible that small proteins need an interacting partner (e.g. a lipid bilayer or an interacting larger protein) to ensure a stable conformation [18].

2.5.3 Funghi

Kessler et al. [79] identified functional sORFs in S. cerevisiae based on homology between fungal genomes. Another functional research, by means of deletions in sORFs, was performed in S. cerevisiae by Kastenmayer et al. [21]. They first used comparative genomics to identify sORFs. They even found conservation over all eukaryotes, sometimes even ranging from yeast to human. Secondly, deletion mutants were constructed in S. cerevisiae to study the functions of these sORFs. Identified functions were growth and protection against stressful conditions.

2.5.4 Plants

Small peptides encoded by sORFs are expected to be especially important in plant morphogenesis. Hashimoto et al. [3] gave several examples of small peptides with putative regulatory functions found in plants.

The POLARIS gene (Arabidopsis thaliana) is needed for a correct auxin-cytokinin response, so it is embryonically involved in root growth and leaf vascularisation. In the POLARIS gene lies a sORF coding for a micropeptide of 36 amino acids. This peptide seems essential
for gene function [3, 5].
In *A. thaliana*, 10% of the intergenic sORFs could be functionally confirmed by overexpression studies resulting in visible phenotypes [63]. The sORFs were selected from an *in silico* generated list of sORFs [23, 25].

In soybean and other Leguminosae, ENOD40 encodes 2 micropeptides (originating from a polycistronic eukaryotic mRNA) that control the use of sucrose during nodulation [3, 80]. Furthermore, research concluded that the RNA structure of ENOD40 is conserved among several plant species. Thus, the RNA can possibly function through translation of its sORFs, leading to regulatory micropeptides, and in the meantime it can control plant mechanisms directly through the secondary mRNA structure [3, 81].

A recessive mutation in *Brick1* in maize leads to a lack of actin dependent polarisation in epidermal leaf cells, because BRICK1 is an essential subunit in the SCAR/WAVE complex, needed in the actin formation process [6, 82]. Although no full genome annotation was available, sORFs could also be identified in *Populus deltoides* by three combined computational approaches [83]. The full transcripts were constructed based on expressed sequence tag (EST) reads.

### 2.5.5 Lower animals

HSPC300, produced from an orthologue gene of *Brick1* (see chapter 2.5.4) is an indispensable component of the SCAR/WAVE complex and plays an important role in nervous system development in *Drosophila*. HSPC300 is needed in the development of the neuromuscular junction, which is also actin dependent [10, 3].

Galindo et al. [9] found the *tarsal-less (tal)* gene (or *polished rice (pri)*) in *Drosophila*. This gene (which was first thought to be a IncRNA [84, 85]) expresses a polycistronic RNA that contains multiple sORFs [9]. Translation leads to micropeptides and subsequent functional analysis revealed that the *tarsal-less* gene plays a role in development, more specifically in the formation of the denticles, apical cuticular structures, and of the taenidia, thickenings in the tracheal tube system of insects. The *tarsal-less* gene is involved in tracheal development because it is essential for the formation of F-actin, which is in turn required to expand the tracheal tubes [77]. Expression of the *tarsal-less* gene leads to modification of the Shavenbaby transcription factor by proteolytic release of a repressor domain [46]. Homologues of *tarsal-less* were also found in other arthropods, demonstrating that conservation of micropeptides is possible. For example, the orthologue gene *mile-pattes* can be found in *Tribolium* [9, 46, 86]. Possibly, the function of the gene is similar in other arthropods, demonstrating segmental development as the general function of *pri*-peptides, however more detailed functional characterization as in *Drosophila* is not yet available for other arthropods [46].

Ladoukakis et al. [28] did a computational approach to identify functional sORFs in *Drosophila*. They found 600 000 sORFs of which 400 classified as functional (2.9% of the currently annotated protein-coding genes) after filtering for transcription, conservation
and synonymous versus asynonymous substitutions. They were able to validate 25 sORFs with proteomics data. However, in several cases splicing was involved before the real translated sORF arised. Aspden et al. [14] on the other hand used PolyRIBOseq (see chapter 2.2.3) to assess the translation probability. They found that around 80% of the already annotated sORFs in *Drosophila* showed translational proof.

A few annotation studies in zebrafish (*Danio rerio*) are reviewed by Pauli et al. [31], some of them focusing on sORFs. Chew et al. [50] used their translated ORF classifier (TOC; see also in chapter 2.2.3) to pick up protein-coding contaminants in lncRNAs. Bazzini et al. [51] identified hundreds of sORFs in zebrafish using ribosome profiling and subsequently ORFscore (see also in chapter 2.2.3). Independently, they also used an *in silico* method based on phylogenetic conservation based on multiple alignments.

### 2.5.6 Mammals

Frith et al. [27] estimated the percentage of sORFs in the collection of all mouse (*Mus musculus*) ORFs at 10%. They applied the CRITICA algorithm [26] on the transcriptome to acquire this estimation. Ingolia et al. [42] defined short polycistronic ribosome-associated RNAs (sprcRNAs), formerly classified as being lncRNAs, in mouse embryonic stem cells (mESCs). These RNAs are engaged by the cellular translation apparatus and frequently contain sORFs. Later, they used the RPF-size distribution and affinity purification to further refine their translation studies in ncRNAs [53]. Crappé et al. [17] performed a genome-wide systematic search for putatively functional sORFs in the mouse genome. They also analysed the coding potential based on evolutionary conservation at the amino acid level. Ribosome profiling data were used to validate the putative translation, leading to putatively encoding sORFs.

### 2.5.7 Human cells

Slavoff et al. [7] discovered 86 previously unannotated micropeptides in human cells with MS. Surprisingly, the discovered micropeptides occurred in the same abundance as canonical proteins, although they are the only group having this conclusion so far. One of their identified micropeptides, MRI-2, could be functionally annotated and may be involved in DNA repair. This emphasises that micropeptides can fulfil important biological functions. Moreover, Oyama et al. [35] could also find proteomic proof for 54 uORF products.

Other studies used ribosome profiling instead of MS. Fritsch et al. [54] used ribosome profiling, applying a PMY antibiotic treatment, to search for uORFs in human monocyte cells. Lee et al. [43] applied GTI-seq in parallel on mouse and human cells to prove the conservation of the global TIS profile. Bazzini et al. (see [51] and also mentioned in chapter 2.5.5) applied the ORFscore also on human HeLa cells profiling data [87], indicating translational activity in RNAs previously classified as non-coding.
Chapter 3

Material and methods

The main purpose of this research is to develop a pipeline, given the name RIBOsORF, which analyses all putative sORFs in a given genome. The pipeline starts from the data obtained out of two parallel RIBOseq experiments, one using freezing, CHX or EMT (generally termed the 'untreated' sample) and one using HARR, LTM or PMY (generally termed the 'treated' sample). After mapping the RPFs, TISs will be determined. From these TISs, sORFs can be constructed \textit{in silico}. Afterwards, filtering for the most promising sORFs can be done given some characteristics (e.g. RPF coverage, conservation, fragment length distribution,...). The pipeline can also generate a RIBOseq derived small translation product database to match the results with mass spectrometry based proteomics data.

The RIBOsORF pipeline was developed at the BioBix lab, the lab of bio-informatics and computational genomics at Ghent University, by Jeroen Crappé, Gerben Menschaert and Steven Verbruggen. Most of this thesis will cover the parts of the pipeline the author of this thesis worked on, in particular the FLOSS calculation (3.3.5), the splice-aware assembly (3.4) and the validation with proteomics data (3.5).

3.1 Hardware

The scripts of this pipeline were written and tested locally on a laptop with four 2.8 GHz Intel® Core™ i7 processors, 16 GB DDR3 RAM and OS X 10.10 Yosemite as operating system. For big datasets, the scripts were executed on one of the three servers of the BioBix lab: \textit{athos}, \textit{aramis} and \textit{midas}. These three Linux servers have respectively 128, 162 and 320 GB RAM and respectively 16, 32 and 64 processors.
3.2 Software

3.2.1 Perl and BioPerl

Perl\(^1\) (Practical extraction and report language) [88] is an interpreted, open source, high-level, multi-purpose scripting language, originally developed by Larry Wall in 1987. Due to its flexibility, it is often called ‘the Swiss army knife of programming languages’\(^2\). Strongly influenced by the features of C and UNIX scripting languages, Perl tries to be a combination of these languages, but it still provides a lot of flexibility for the user, writing the catchphrase ‘there is more than one way to do it’ large. Perl is also a modular programming language, providing different modules in an external library called the Comprehensive Perl Archive Network (CPAN)\(^3\). Besides a strong affinity for regular expressions, Perl possesses also good possibilities for setting up connections with external relational databases, making reading out and writing in them a straightforward job. There is also simple access to different types of files from within Perl scripts.

Newer languages as Python and PHP, although influenced by Perl, suppressed the general popularity of Perl. However, due to the existence of BioPerl\(^3\), a collection of handy bioinformatic modules, Perl is still a general used language in bioinformatics. In that way, Perl can handle necessary functions like reading in genomic and proteomic data, accessing public and local databases, parsing of data and accessing and manipulating most used bioinformatic data formats.

All scripts of the RIBOsORF pipeline are written in Perl version 5.20.1 in combination with different modules, including some of BioPerl 1.6.9. Further details on the functions in the different scripts of the pipeline are explained in chapter 3.3.

3.2.2 R

\(^{1}\)Available on \url{http://www.perl.org/get.html}
\(^{2}\)CPAN homepage: \url{http://www.cpan.org}
\(^{3}\)Available on \url{http://www.bioperl.org}
3.2. SOFTWARE

R\(^4\) \[89\] is an open source programming language perfectly suited for statistical analysis. A broad variation of statistical methods (including classification, classical tests, linear and non-linear modelling,...) are foreseen and can be further expanded by downloading modules from the Comprehensive R Archive Network (CRAN)\(^5\). In R, mathematical formulas can be implemented very easily and plots can be constructed with high quality. R is described as an 'environment', enlightening the fact that it works very flexible but still structured as statistical software. R (version 3.1.2, 64bit) is used for calculating and plotting FLOSS scores (chapter 3.3.5). With the Statistics::R module, the R interpreter is controllable through Perl \[90\].

3.2.3  SQLITE

\[
\text{Available on} \ \text{http://cran.rstudio.com}
\]

SQLite\(^6\) \[91\] provides a self-contained, serverless, zero-configuration SQL database engine. SQLite (version 3.8.8.3) is in the public domain so comes free and is included in different applications. Because SQLite does not have separate sever processes, it writes directly to the ordinary disk. Besides, it is also a very compact library and the database is easily manual accessible in command line by using the sqlite3 command line utility that allows the user to enter and execute standard query language (SQL) statements \[92\].

3.2.4  Perl DBI

In a database structure, tables consist of records, each containing information for different fields. SQLite is a relational database structure. This means that tables are linked by referring from within one table to a record in another table. Reading, searching and editing in a relational database is executed by the relational database management system (RDBMS). Although most systems can be operated by SQL, each system differs slightly in its communication. To simplify the access of different databases through Perl, a database interface (DBI) module has been developed. The DBI module provides methods and variables to make the access independent of the used database. The user only needs to specify which database system needs to be addressed and the DBI makes connection with the appropriate database driver (DBD). In that way, a uniform SQL syntax can be used in the Perl script. The DBI makes connection with the chosen database and sends the query to the DBD, where it will be passed to the RDBMS in the language used by

\[\text{Available on} \ \text{http://cran.r-project.org}
\]

\[\text{Available on} \ \text{http://www.sqlite.org/download.html}\]
CHAPTER 3. MATERIAL AND METHODS

Figure 3.1: Connection scheme. The DBI is accessed from within the script. The DBI makes connection with the appropriate DBD that sends the queries in the appropriate language to the RDBMS of the specified database.

the database. When the DBD receives the results, they will be redirected to the script through the DBI. By using this system (given in figure 3.1), there is no need of knowing the specific structure and conventions of the underlying database [93]. Here, we use the Perl DBI module (version 1.633) [94] in combination with the DBD::SQLite module (version 1.46) [95].

3.2.5 TopHat and STAR

For mapping RPFs to a reference genome, a mapper that can efficiently map short reads will be needed. TopHat\(^7\) (version 2.0.9) is used for mapping splice junctions of RNAseq data to a reference genome in a very fast way [96]. First it aligns the RNAseq reads to the genome using the short read aligner Bowtie (version 2.1.0). Bowtie is a fast and memory efficient aligner written in C++ and developed to map short reads efficiently to mammalian-size reference genomes. The speed is obtained by a Burrows-Wheeler indexation of the reference genome [97]. An adaption, Bowtie2, uses the FM index [98]. After mapping, TopHat analyses the mapped reads and identifies splice junctions [96]. Later, several enhancements (like an improved indel-finding algorithm) were incorporated in TopHat, resulting in TopHat2. It can use either Bowtie or Bowtie2 as core alignment tool [99].

Spliced Transcripts Alignment to a Reference (STAR) is another alignment tool based on an algorithm that starts with searching seeds. Afterwards, seeds will be clustered and the highest scoring cluster will be used as the best possible mapping. STAR outperforms other possible aligners for short reads with a speed factor of minimum 50 [100]. In the RIBOsORF pipeline, STAR\(^8\) (version 2.3.0) is used as the default mapper.

---

\(^7\)Available on http://ccb.jhu.edu/software/tophat/index.shtml
\(^8\)Available on https://github.com/alexdobin/STAR/releases
3.2.6 Galaxy

Galaxy\(^9\) is a web-based open workbench that enables scientists, even if they do not have any programming experience, to easily perform computational analyses on biological data. Analysis tools, data, tutorials and publications are freely available on the web and different tools can be put together in one intuitive analysis chain [101]. Especially GalaxyP, an extension platform focusing on multi-omics analysis, is used to make our pipeline more accessible [102].

3.2.7 SearchGUI and PeptideShaker

Search engines, each with their own algorithm, are needed to identify the proteomic fragmentation spectra by matching and scoring the observed peaks to the theoretical peaks from known protein sequences in a database. Best results are achieved if multiple algorithms are combined, however, different input parameters and interfaces make communication between search engines difficult. Therefore, SearchGUI\(^10\) [103] offers a user-friendly and open-source tool that manages several search engines at once (e.g. OMSSA [104], X!Tandem [105], MS-GF+ [106], etc.) using the same input parameters.

For further proteomics data processing, interpretation and re-analysis of public data, PeptideShaker\(^11\) [107] can be used. It unifies the PSM lists of the different search engines that were used in SearchGUI and combines them to increase the confidence and sensitivity. Confidence of PSMs, peptides and proteins is calculated based on a default false discovery rate of 1%.

Both SearchGUI and PeptideShaker come in a graphical user-friendly interface, which is used locally, as well as in a command line interface, which is suited for working on the midas server.

3.3 Single-exon RIBOsORF pipeline

The RIBOsORF pipeline searches for putatively functional sORFs with a computational approach driven by ribosome profiling (overview of the different steps in figure 3.2). The

---

\(^9\)Homepage: https://usegalaxy.org

\(^10\)Available on: https://code.google.com/p/searchgui/

CHAPTER 3. MATERIAL AND METHODS

Figure 3.2: An overview of the different building blocks of the RIBOsORF pipeline. The pipeline starts off from two differently treated samples. After mapping with STAR or TopHat, rule-based classification makes it possible to determine TISs and assemble possible sORFs in genic and intergenic regions. For each sORF, supplementary characteristics about the coding potential (e.g. PhyloCSF conservation score and FLOSS) can be calculated. The results of each step are stored in an SQLite database and species-specific annotation data are available from iGenomes and Ensembl.

The pipeline starts from two fastq files, earlier termed the ‘untreated’ and the ‘treated’ sample, and consists of 5 major steps. First, the RPF reads are mapped to the reference genome of the model organism. Then, the TISs are identified. Next, starting from these TISs, sORFs are assembled in silico. Afterwards, PhyloCSF [29] is used to calculate a conservation score. After that, a sORF specific FLOSS is determined [53].

All information of these steps can be stored in an SQLite database. From this database, a small translation product datafile can be generated so that MS-based studies can be performed against a RIBOseq derived search space. All input parameters of each step are definable by the user so that they can be optimised in function of the research. For mouse, Ensembl [108] release 72 and later on (chapters 3.4 and 3.5) release 75 genome annotation (assembly GRCm38) from the iGenomes repository were used.

3.3.1 Mapping

For each fastq file, RIBOseq derived reads are aligned with either STAR or TopHat (see chapter 3.2.5). Reads are aligned to indexed sequences in following order: first to the PhiX bacteriophage genome, then to the rRNA of the model organism (obtained from Ensembl) and finally to the complete genome of the model organism (obtained from its iGenome repository). Only uniquely mapped reads are accounted for, although this can be adapted as a pipeline option. Moreover, only genomic reads between 26 and 34 nt are considered.

Available on http://support.illumina.com/sequencing/sequencing_software/igenome.ilmn
as only these are relevant as RPFs. The RPF alignments are assigned to the ribosomal P-site, dependent on the length of the fragment: respectively +12, +13 and +14 from the 5’ end of the alignment for RPFs of ≤30nt, 31-33nt or ≥34nt long [42].

For STAR, the internal clipping function is applied to clip off the 3’ adaptor. The STAR option seedSearchStartLmaxOverLread is always set at 0.5. Furthermore, for the alignment against the PhiX genome no introns and up to 2 mismatches are allowed. TopHat2 does not have an internal clipping tool, so the FASTX clipper (version 0.0.13) is used to initially clip off the 3’ adaptor. For PhiX and rRNA mapping, Bowtie2 with ‘sensitive-local’ option is used. For the genomic alignment afterwards, TopHat2 itself is used with default settings except for the segment length which is set to 15.

The alignments and RPF density data are also returned in BedGraph format which is easy to upload and visualise in a genome browser (e.g. the UCSC genome browser) [109].

### 3.3.2 TIScalling

After the RPFs are mapped, the translation start sites (TISs) can be determined. All profiles from initiating ribosomes (in the ’treated’ sample) can be accumulated at the first nucleotide of an AUG or a near-cognate (i.e. a codon differing 1bp from AUG) start codon within a 3nt range (1nt up- and downstream from the peak position). In that way, it is possible to determine the TIS on nucleotide level [44]. Only accumulated RPF peaks that are positioned in this small 3nt range around a start codon are considered in the TIS calling. Peaks do need to match some other criteria in order to be called as a real TIS [43, 44]. First, they should have the maximum number of reads (in the ’treated’ sample) within a broader range of nucleotides (default 7, i.e. one codon up- and downstream). Secondly, the sum of all RPF counts that comprise the TIS peak should exceed a minimal profile count threshold which is default set at 5. This threshold is however highly dependent on the sequencing depth. And thirdly, the TIS position should have a R value (see equation 2.2) equal or higher than a certain threshold, default set at 0.05. However, for calculating the R value, the sORF begin and end positions are required and these are determined later on during the sORF assembly. So, the evaluation of this characteristic is postponed to the assembly step (chapter 3.3.3).

The same parameter settings are applied for all genic as well as intergenic TISs, choosing for an unbiased approach. The genic regions are obtained from a species-specific Ensembl annotation bundle. To determine the intergenic regions, a sliding window is applied which runs over each chromosome and defines an intergenic region as a sequence where none of the known Ensembl genes is overlapping.
3.3.3 Single-exon assembly

Starting from the TIS positions, single-exon sORFs can be constructed. The minimal and maximal length can be set as input arguments. Between these boundaries, the assembly searches for an in-frame stop codon for each TIS position. However, sORFs starting from a known annotated TIS position are neglected. For each found in-frame start-stop combination the mass of the resulting peptide, the DNA and AA sequence and the annotation class of the TIS (5’UTR, exonic, intronic, 3’UTR, ncRNA or intergenic) are stored. For intergenic sORFs, the distance to the nearest up- and downstream known gene is determined. For 5’UTR, exonic and intronic sORFs, the overlap to known Ensembl exonic regions is calculated, as well as the frame of these known exonic sequences, which is compared to the frame of the sORF. Sequences with an exon overlap that exceeds a certain user defined threshold (default 0.5) are discarded of further investigation. Also the RPKM value (a measure of ribosomal density), which is defined by Ingolia et al. [36] as the number of RPF reads per kilobase of sORF coding region and per million of aligning reads in the data.

The coverage describes the extent to which a sORF has RPF reads along its sequence (in the ‘untreated’ sample). RPF reads are assigned to single genomic positions. Coverage is defined as the percentage of the positions in the sORF that have at least one ribosomal footprint. The sORF is skipped from the analysis if its coverage is not greater than zero. Furthermore, coverage can be used as an extra filtering step when requesting the results from the database.

3.3.4 PhyloCSF

PhyloCSF\textsuperscript{13} is a reformulation of CSF (the codon substitution frequencies metric). It is developed to make an accurate classification of small genomic stretches as protein-coding or non-coding sequence. This is done based on an empirical codon substitution matrix, a common ancestral sequence and the structure of a phylogenetic tree relating the supposed descendants [29].

The tool needs species specific multiple alignment files (MAF)\textsuperscript{14} to function [29, 110]. First, a multiple alignment is created for each sORF in a separate Perl script which stores them in the SQLite database. Then, the sORF multiple alignment is transferred to a fasta file, which can be analysed by the PhyloCSF tool [29].

3.3.5 FLOSS

Ingolia et al. [36, 42] found that the size distributions of ribosome-protected fragments (ranging from 26-34bp) differed between true footprints of protein-coding sequences and

\textsuperscript{13}Download and more information on: https://github.com/mlin/PhyloCSF/wiki
\textsuperscript{14}Multiple alignments are downloadable from http://hgdownload.cse.ucsc.edu/downloads.html
false footprints of rRNA and other non-coding contaminants. However, the length distribution can also vary because of differences in digestion conditions [39]. Moreover, the predominant ribosome conformation can differ between samples, leading to significantly varying footprint lengths [111]. And thirdly, CHX yields general ribosomal footprints that are in general about 1nt shorter than those treated with EMT [42, 53]. Nevertheless, this difference only appears in true footprints, making it another possible approach to identify true footprints. In sum, fragment size distributions are optimally compared within a single sample to prevent all confounding effects [53].

To capture size distribution differences in one number, Ingolia et al. defined the FLOSS [53]. They calculated the metric out of histograms of all footprint length counts in a transcript or reading frame in function of the footprint length. A reference histogram was first constructed using the raw length counts in all annotated nuclear protein-coding transcripts, not overlapping an annotated non-coding transcript. Mitochondrial footprints however, are left out of the analysis, as they show a complete different size distribution. Ingolia et al. defined the FLOSS as:

\[
FLOSS = 0.5 \cdot \sum_{l=26}^{34} |f(l) - f_{ref}(l)|
\]  

where \( f(l) \) is the fraction of reads with length \( l \) included in the transcript histogram and \( f_{ref}(l) \) is the analogue fraction but in the reference histogram. Lower scores are result of a better similarity between two distributions. However, due to sampling errors, the FLOSS is less informative for transcripts with fewer reads. A lower amount of reads results in a higher score [53].

Ingolia et al. developed FLOSS to confirm that the footprints in non-coding transcripts were in fact true footprints instead of artefacts [53]. In the RIBOsORF pipeline however, FLOSS is applied to test the protein-coding character of putatively functional sORFs. Nevertheless, the metric is applicable to every sequence, whether it is a whole transcript or a reading frame.

To introduce the FLOSS calculation in the pipeline, the following algorithm, consisting of 4 important subroutines, was imbedded in a Perl script (available online, see addendum chapter A.1).

The refHist subroutine checks initially if there is already a reference histogram available (stored in the results database as reference fractions in function of the fragment length). If not, it selects all Ensembl nuclear protein-coding genes, except for those genes that overlap with non-coding transcripts. Then, these genomic sequences are reduced to the essential exonic regions by using canonical translation info (i.e. eliminating 5’UTR, intronic and 3’UTR parts of the transcript). All the RPF data, falling in these exonic regions, are used to calculate the reference fractions for each length from 26 to 34bp, resulting in the reference histogram. The reference fractions are stored in the database.

Next, the cutoff subroutine checks for smoother data in the results database. If there is no smoother data available, a new smoother will be constructed starting from the count
data of all canonical coding transcripts (available out of the refHist subroutine) and the overall reference distribution. For each of the given transcripts, a FLOSS can be calculated using formula 3.1. Using a sliding window approach (Ingolia et al. [53] suggested a default window of 200 transcripts), the program goes through the canonical transcripts ranked along an increasing amount of reads. For each window, the first (Q1) and third (Q3) quantile of the FLOSSs of the transcripts in the window are calculated. With Tukey’s method, a raw extreme can be calculated afterwards for each window position:

$$Raw\ Extreme = Q_3 + 3 \cdot (Q_3 - Q_1)$$ (3.2)

With these extreme values, a Loess smoother can be fitted, resulting in a cutoff value for each amount of reads, storable in the results database. The cutoff calculation is easily implemented in R, so here the Statistics::R module is used to give Perl control over the R interpreter. Also the construction of plots can be easily done in R.

The makeLD subroutine is used to fetch all RPF data (fragment lengths, start position, count data, strand) and store it as a length distribution in a Perl hash structure grouped per sORF ID. This makes FLOSS calculation for each sORF much more easy.

Finally, the length distribution per sORF and the reference fractions are used to calculate a FLOSS score for each sORF. Based on the amount of reads for the considered sORF, a cutoff value can be determined on the smoother. With this cutoff value and the FLOSS score of the sORF, a classification can be done in the classification subroutine. The sORF is classified as ‘Good’ if he has a good coding probability and ‘Extreme’ if he is probably non-coding. If the amount of reads of the sORF does not fall in the range where the smoother is defined (due to the sliding window approach), the sORF is classified as ‘Not in cutoff range’. All classification results are stored in the results database.

For further optimisation of the code and the FLOSS algorithm, the Perl code profiler Devel::NYTProf\textsuperscript{15} was used to measure the execution time of each statement and subroutine.

### 3.3.6 Translation product database construction

At the end of the pipeline, a non-redundant custom sORF sequence database can be constructed in fasta format, suited for proteomics identification experiments. The description line of the fasta file consists of the sORF ID, the chromosome, the strand, the begin and end position, the annotation class, the start codon, the biotype, the peptide mass, the conservation score and the FLOSS classification. The sequence data line holds the amino acid sequence of the putative small translation product.

### 3.3.7 Script-based implementation and Galaxy version

All steps of the RIBOsORF pipeline are built up of Perl scripts. In the script-based version, each module needs to be initiated from command line. Although each script includes some

\textsuperscript{15}http://search.cpan.org/~timb/Devel-NYTProf-5.07/lib/Devel/NYTProf.pm
3.4 Splice-aware assembly

To include spliced sORFs in the analysis, the original assembly (chapter 3.3.3) was expanded. In the single-exon assembly, a subdivision was made between sORFs in intergenic and genic region. Because there is no splice structure for intergenic regions, there is no adaption needed for intergenic sORFs and they will all be single-exon.

For protein-coding genes, an exon structure can first be determined, dividing the gene in 5’UTR, exonic, intronic and 3’UTR regions. The exon structure was already used to
classify each TIS (except the class of the already annotated TISs (aTIS)) in a certain annotation class, but this splicing information can also be applied to construct spliced sORFs \textit{in silico}. Single-exon sORFs can be assembled for the protein-coding genes, as it has been done in the single-exon assembly. Afterwards, a new part of code was added to expand the collection of sORFs in protein-coding genes with sORFs that use one or multiple splice junctions (as sORFs will only be added to the collection this time if they show to have a splice site). Then, for all sORFs in protein-coding regions, both single-exon and spliced, exon overlap will be calculated and if there is overlap, the script will calculate if the sORF is in frame with the canonical reading frame. Exon overlap is calculated for sORFs with an 5'UTR, exonic and intronic TIS because a sORF starting in an 5'UTR or intronic region can continue in a downstream exonic region. In-frame sORFs with more than 50% exon overlap will be discarded.

After looking in protein-coding genes, sORFs will also be assembled in non-coding genes. Here as well, single-exon sORFs will be constructed first and will then be completed with spliced sORFs. It is easier compared to protein-coding genes as these regions consist only of exonic and intronic parts, no UTRs. For lncRNAs, there is an extra control to check if the sORF starts in an exon of the lncRNA, as the lncRNA can have introns that span big regions in which other genes can be positioned. Identification of sORFs in these intronic regions is in most cases an identification in the enclosed smaller gene, rather than being a sORF in the lncRNA.

Finally, the reads situated in all found genic sORFs will be fetched for calculating the R value (as explained in chapter 3.3.2) and the coverage. The sORF will only be stored in the database if the R value is higher than a chosen threshold and if the coverage is above zero.

3.5 Validation with proteomics data

The final step in the identification process is confirming the micropeptide candidate with mass spectrometry. Therefore, custom datasets (constructed out of the RIBOsORF pipeline with the translation product database construction step) from mouse and human samples were analysed with SearchGUI and the search results were subsequently interpreted in PeptideShaker. Small spectral datasets in mgf format were searched locally with the graphical user interface, whereas bigger datasets were analysed from command line on the midas server. However, the first analysed dataset was checked in both modes to ensure an unbiased outcome. All input parameters were each time available in the paper that came along with the spectra or were specified in personal communication with the group that generated the data. For samples with an unspecified fragmentation method, the enzyme was set as 'unspecific' and the analysis was performed with the OMSSA and MS-GF+ search engines. For samples that were fragmented with trypsin, three search engines (OMSSA, MS-GF+ and X!Tandem) were applied and the enzyme was set as 'Trypsin'.
For each set of spectra, two subsequent analyses were performed. In this two-stage approach, first, a search against a SwissProt database of all mouse or human known proteins and their isoforms was done. The unvalidated spectra of this first round were then used to search for micropetides against a custom database which was obtained out of the RIBOsORF pipeline.
Chapter 4

Results

4.1 Single-exon RIBOsORF pipeline

So far, the RIBOsORF pipeline is developed to work for fruit fly (Drosophila melanogaster), mouse (Mus musculus) and human (Homo sapiens). Expansion of the pipeline to other species is possible and can be done for further projects. Here, the pipeline was tested on mouse embryonic stem cell (mESC) data of Ingolia et al. [42]. This data (dataset GSE30839) is available from the Gene Expression Omnibus (GEO) [112] of the National Center for Biotechnology Information (NCBI). Two fastq files were selected, one of an analysis with CHX as translation inhibitor (sample GSM765292) and one of an analysis performed with HARR (sample GSM765295) to search for initiating ribosomes. The single-exon RIBOsORF pipeline was tested with release 72 of Ensembl and assembly version GRCm38.

Initially, the RIBOsORF pipeline identified lots of sORF sequences located in pseudogenes primarily due to multi-mapped reads (unpublished). This would introduce lots of false positives as true validation of the protein-coding capacity is lacking due to the difficulty of MS-based proteomics analysis in pseudogenes. Thus, to obtain a stringent set of putatively coding sORFs, only uniquely mapping reads were aligned.

Furthermore, in TIS calling, all parameters were mostly set at their default values, except for the R value cut-off, which was set at 0.01 to primarily allow for a larger number of sORFs and later on filter to a more stringent set.

The scripts of the pipeline are given in addendum chapter A.1. The Galaxy version of the pipeline can be found in addendum chapter A.2.

4.1.1 Manual check of mapped reads

After the mapping step, the pipeline has already generated seven files (see figure 3.3): two SAM (sequence alignment map) files (one for the 'untreated' sample and one for the
Figure 4.1: Example of a visualised sORF in the UCSC genome browser. This putative sORF is located on chromosome 18, ranging from genomic position 56 707 937 till 56 708 035. In red, the sORF itself is highlighted in the nucleotide sequence. On the untreated sense track, in the red rectangle, a good coverage of RPF reads is observed. In blue, the peak (461 reads) of the near cognate start codon (CTG) is highlighted on the treated sense track. The other parts of the treated track have rather low or no RPF reads, except around genomic position 56 707 990 where an internal start site (CTG) is highlighted in orange.

'treated' sample), the SQLite results database that will be further expanded in the next steps, and four bedgraph files. These last four files give genome browser track information for either sample, each time a file for the sense track and another for the antisense direction. After running the complete pipeline, putative sORFs can be checked manually by loading the bedgraph files into a genome browser (e.g. UCSC genome browser \(^1\)). The sORFs can then be visualised with the four different RPF tracks, as can be seen in figure 4.1.

\(^1\)Available at http://genome-euro.ucsc.edu/cgi-bin/hgGateway
4.1.2 Influence of adaptors

Ingolia et al. [42] used 5’CTGTAGGCACCATCAATTCGTATGCGTGCTTCTGCTTGAA3’ (termed ‘long’) as a 3’ adaptor sequence in their mouse data of 2011. However, in their mouse data from 2014 [53], the AIR™ adenylated linker A (5’CTGTAGGCACCATA-3’, termed ‘short’) from Bioo Scientific [113] was used. As the short adaptor consists of the 5’ part of the long adaptor, it was interesting to investigate the influence of the adaptor on the mapping results. Therefore, the data of 2011 [42] were analysed with both adaptors and the mapping statistics were compared. The mapping statistics for the long and short adaptor are given in respectively table 4.1 and 4.2.

**Table 4.1:** Mapping statistics using the long adaptor (CTGTAGGCACCATCAATTCGTATGCGTGCTTCTGCTTGAA).

<table>
<thead>
<tr>
<th>sample type</th>
<th>Total reads</th>
<th>Mapped</th>
<th>Unmapped</th>
<th>Frequency mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Untreated</td>
<td>42 412 547</td>
<td>38 709</td>
<td>42 373 838</td>
<td>0.0009127</td>
</tr>
<tr>
<td>rRNA</td>
<td>42 373 838</td>
<td>166 345</td>
<td>42 207 493</td>
<td>0.003926</td>
</tr>
<tr>
<td>genomic</td>
<td>42 207 493</td>
<td>33 283 571</td>
<td>8 923 922</td>
<td>0.7886</td>
</tr>
<tr>
<td>treated</td>
<td>54 733 912</td>
<td>2 811</td>
<td>54 731 101</td>
<td>5.135 · 10⁻⁰⁵</td>
</tr>
<tr>
<td>rRNA</td>
<td>54 731 101</td>
<td>102 686</td>
<td>54 628 415</td>
<td>0.001876</td>
</tr>
<tr>
<td>genomic</td>
<td>54 628 415</td>
<td>44 211 491</td>
<td>10 416 924</td>
<td>0.8093</td>
</tr>
</tbody>
</table>

**Table 4.2:** Mapping statistics using the short adaptor (CTGTAGGCACCATCAAT).

<table>
<thead>
<tr>
<th>sample type</th>
<th>Total reads</th>
<th>Mapped</th>
<th>Unmapped</th>
<th>Frequency mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Untreated</td>
<td>42 412 547</td>
<td>38 656</td>
<td>42 373 891</td>
<td>0.0009114</td>
</tr>
<tr>
<td>rRNA</td>
<td>42 373 891</td>
<td>166 338</td>
<td>42 207 553</td>
<td>0.003925</td>
</tr>
<tr>
<td>genomic</td>
<td>42 207 553</td>
<td>33 279 593</td>
<td>8 927 960</td>
<td>0.7885</td>
</tr>
<tr>
<td>treated</td>
<td>54 733 912</td>
<td>3 028</td>
<td>54 730 884</td>
<td>5.532 · 10⁻⁰⁵</td>
</tr>
<tr>
<td>rRNA</td>
<td>54 730 884</td>
<td>102 843</td>
<td>54 628 041</td>
<td>0.001879</td>
</tr>
<tr>
<td>genomic</td>
<td>54 628 041</td>
<td>44 208 832</td>
<td>10 419 209</td>
<td>0.8093</td>
</tr>
</tbody>
</table>

The statistics in tables 4.1 and 4.2 are very similar and suggest that the short adaptor can be applied instead of the longer one. To assure this assumption, the unmapped reads, after clipping off the short adaptor, should be inspected. If these unmapped reads still contain adaptor sequences and the part of the read upstream of that present adaptor can be mapped manually to the mouse reference genome, then it is not permitted to use the short adaptor instead of the longer one in the clipping algorithm.
The unmapped reads (after mapping with the short adaptor) were searched manually for remained adaptor sequences but in most cases, they were situated in the 5’end of the read. This makes it a meaningless read since an offset with sequence information upstream of the adaptor is absent. In some cases however, the adaptor was at the 3’end. Then, the part of the read upstream of the adaptor was used for manual mapping to the mouse reference genome with UCSC BLAT. Mostly, there was no mapping possible. For a few reads however, it was possible and the sequence was investigated using BLAST. This resulted in matches to miscellaneous RNA (miscRNAs), snoRNAs and transcript variants, but no real genes. Thus, it could be concluded that the short adaptor can be used instead of the longer one.

4.1.3 Filtering steps

Running the single-exon RIBOsORF pipeline on the mESC data of Ingolia et al. [42], 68 343 putative sORFs are obtained. However, these identifications only comply with the rules for a valid TIS identification (see chapter 3.3.2) and have any elongating RPF covered sequence stretches that result in a coverage above zero. To select the sORFs with the highest chance to be translated, some extra filter scores were introduced through different modules of the RIBOsORF pipeline. Filtering can afterwards be manually applied when fetching the results out of the SQLite database by constructing a selective query based on these scores.

Ribosome profile coverage

RPF coverage is the most intuitive measure to describe the evidence found for a sORF in ribosome profiling. In the assembly, sORFs were already filtered to have a coverage greater than zero. However, when looking for true translation, extra filtering on the coverage may be needed using a threshold. This filtering threshold is set to a value where 90 percent of the 3’UTR sORFs in canonical transcripts are excluded. Without any filter on the coverage, 1931 theoretical 3’UTR sORFs could be found. Because the 3’UTR regions are generally assumed to be non-coding, they are suited to be used as a null distribution. 10% of the 3’UTR sORFs have a coverage above 0.25, making this an ideal threshold for coverage filtering.

At first hand, this threshold seems low, but for each footprint mapped to a single genomic position, the corresponding ribosome is protecting 26-34 bases. Thus, assuming an even distribution of RPF reads, a coverage threshold of for example 0.25 means that the sORF has a coverage close to 100% in terms of ribosomal protection. Also, even a short sORF sequence of 10 AA (30nt) has at least 7 (> 7, 5 i.e. > 25%) reads mapped to its sequence.

2http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
4.1. SINGLE-EXON RIBOSORF PIPELINE

PhyloCSF

PhyloCSF [29] filters based on conservation. Although, PhyloCSF offers an outstanding discrimination for very short regions, it probably leads to more false negative results as well. Volders et al. [114] determined an optimal PhyloCSF score threshold to separate coding and non-coding sequences. This threshold was set at 41.

FLOSS

FLOSS is able to distinguish true ribosomal footprints from contaminants [53]. It is a score based on the similarity between the fragment length distribution of the sORF with a reference distribution. Based on this score, a classification can be done which can be used to filter the results. The classification is based on a cut-off which can be read from a smoother in function of the amount of reads. The more reads, the stronger the filter can work.

First results were controlled manually by checking the FLOSS with SQLite queries. For the given dataset, around 90% of the putative sORFs were classified to have a good FLOSS, making FLOSS not a very stringent filter. Ingolia et al. [53] suggested some examples of truly non-coding RNAs. In two of them (Rmrp and Rny1), some sORFs were found: six in Rmrp and four in Rny1. All of these sORFs were classified as contaminants by the FLOSS algorithm, confirming that these RNAs are truly non-coding.

Furthermore, Ingolia et al. [53] also describe two well known examples very extensively: Malat1 and Gas5, which can be used to validate our scripts. Malat1 is a lncRNA for which they found a general non-coding character although the 5’ part had substantially more RIBO-reads than the rest of the transcript. In this 5’ part, they found a ATG reading frame which could be clearly classified as coding following its FLOSS. The exact position and length of this reading frame was however not specified. The overall FLOSS of the transcript on the other hand classified the whole transcript as non-coding [53]. In contrast, RIBOsORF calculates the FLOSS for each sORF separately and not for the transcript as a whole. But it could be clearly seen that sORFs in the 5’ part of the transcript had a much higher coverage. Because of the high amount of reads in this part, FLOSS can be used as a strong filter and much sORFs can be excluded as non-coding.

Gas5 is a ncRNA that contains several snoRNAs, which resulted in a bad FLOSS. However, Ingolia et al. found that after splicing, most of the snoRNAs were gone and the spliced transcript showed a FLOSS indicating a coding nature [53]. RIBOsORF found sORFs with two biotypes in this transcript: ‘processed transcript’ and ‘snoRNA’. Most of the sORFs with biotype ‘snoRNA’ were classified as non-coding based on their FLOSS, whereas the sORFs in the other biotype passed the filter.

With some adaptions, some interesting plots can be constructed from within the script. For each individual sORF as well as for a collection of sORFs, the RPF length distribution
can be drawn, which represents the essence of FLOSS. As can be seen in figure 4.2, for stringent putatively coding sORF candidates, the distribution resembles the distribution seen in canonical nuclear coding regions, whereas the distribution in a supposed truly non-coding region is totally different, resulting in a much higher FLOSS statistic.

Figure 4.2: Fragment length distributions are able to single out true ribosomal footprints. (A) The length distribution (in green) of a set of stringent sORF candidates in lncRNA (see table 4.3) compared to the reference distribution (in grey) based on the sample specific RPF read lengths falling in all nuclear coding regions. (B) The fragment length distribution of the sORFs in Rmrp (in red), supposed to be a truly non-coding transcript [53], compared to the reference distribution (in grey).

The cut-off smoother as well as the FLOSS itself depend on the amount of reads. Therefore, the FLOSS for each sORF is always plotted in function of the amount of reads located in that sORF, as can be seen in figure 4.3. In part C of that figure, it is shown how the cut-off smoother is constructed in function of the FLOSS of all nuclear coding regions. As one
can see in this figure, the cut-off smoother and the FLOSS statistics are high for regions with a sparse coverage but they decrease with an increasing amount of reads. This can be explained by a decreasing variation, as more reads make the variation on the read length frequencies smaller.

The FLOSS and the classification based on the smoother are presented in part A for all identified sORFs and in part B for all sORFs located in lncRNAs.

**Figure 4.3:** FLOSS calculation in the RIBOsORF pipeline in function of the amount of reads. (A) FLOSS plot for all possible sORFs, (B) for the sORFs in lncRNAs and (C) for the nuclear coding regions used to calculate the cut-off smoother (red line).
Where the FLOSS decrease with an increasing amount of reads, this is not true for the highest covered regions. When the window arrives at the last points, although they have the most reads, the FLOSS cut-off goes up. This is because the observation points are less dense in this part of the plot. Therefore, the interquantile distance, as a measure of variation, (see equation 3.2) will be greater and the smoother will rise. Also, in figure 4.3 A, some points at the far right are classified as 'Not in cut-off range' because the cut-off cannot be calculated for that amount of reads due to the sliding window approach.

Filtered sORFs

The results of the three filtering steps on the 68,343 earlier identified sORFs is given in table 4.3.

4.1.4 Identification of sORFs in different annotation classes

In chapter 2.3, it was described that actively transcribed sORFs can exist in all possible annotation classes. Therefore, identified and filtered sORFs were subdivided based on the annotation wherein the TIS is located (see table 4.3). A schematic representation of the different annotations in the exon structure is given in addendum figure A.1. As one can see, sORFs in 5’UTR and intronic regions can run through in exonic regions, showing a certain percentage of exon overlap. sORFs starting in exonic regions on the other hand, can be fully located in exonic regions, but can also continue partly in intronic or 3’UTR regions. If a sORF overlaps for its greater part with exonic sequences, the chances of being a false positive hit are too high. Therefore, sORFs with too much exon overlap are filtered out during the assembly.

The relative frequency of the identified sORFs in each annotation class can be represented in a pie chart (figure 4.4).

4.1.5 Validation of sORFs using LNCipedia

Mass spectrometry is still the golden standard in analysis and validation of proteomics identifications. Therefore, it would be a very valuable proof of the presence of functional micropeptides. However, current protocols still seem inadequate for identifying small translation products with great confidence and results of MS-based proteomics proof for micropeptides are sparse [7, 34]. That is why an extra way of validation was searched. As discussed in chapter 2.3.1, sORFs are especially of interest in questioning the true non-coding character of lncRNAs. So, an alternative way of validation could be found in LNCipedia4, a publicly available catalogue of annotated human lncRNAs [114, 115].

4http://www.lncipedia.org
Figure 4.4: Pie chart showing the share of each annotation class in the amount of identified sORFs (68,343 sORFs in total) in mESC data. A schematic representation of each annotation class can be find in addendum figure A.1. For 5'UTR and intronic sORFs, separate pie charts show the amount of sORFs that overlap with exonic regions. Furthermore, for exonic sORFs, the fraction which is in frame with the main ORF, is represented in a separate pie chart.
Table 4.3: Table with filtering steps performed in a cumulative way for each annotation class. The first filter selects sORFs with a RPF coverage higher than 0.25. Then a filter based on the FLOSS [53] is applied. And finally, a last filter uses PhyloCSF [29] to screen for the most promising sORFs with a conservation value above 41.

<table>
<thead>
<tr>
<th>5’UTR sORFs</th>
<th>23 593</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>7 360</td>
</tr>
<tr>
<td>FLOSS</td>
<td>5 778</td>
</tr>
<tr>
<td>Conservation</td>
<td>296</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exonic sORFs</th>
<th>33 741</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>4 199</td>
</tr>
<tr>
<td>FLOSS</td>
<td>4 093</td>
</tr>
<tr>
<td>Conservation</td>
<td>231</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intronic sORFs</th>
<th>4 554</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>637</td>
</tr>
<tr>
<td>FLOSS</td>
<td>604</td>
</tr>
<tr>
<td>Conservation</td>
<td>95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3’UTR sORFs</th>
<th>1 025</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>200</td>
</tr>
<tr>
<td>FLOSS</td>
<td>189</td>
</tr>
<tr>
<td>Conservation</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ncRNA sORFs</th>
<th>1 412</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>377</td>
</tr>
<tr>
<td>FLOSS</td>
<td>195</td>
</tr>
<tr>
<td>Conservation</td>
<td>14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intergenic sORFs</th>
<th>4 018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>479</td>
</tr>
<tr>
<td>FLOSS</td>
<td>373</td>
</tr>
<tr>
<td>Conservation</td>
<td>34</td>
</tr>
</tbody>
</table>

The database tries to bundle annotation information of different sources including PRIDE database information, PhyloCSF [29] and RIBOseq data [43, 51]. In table 4.3, 14 sORFs in non-coding RNA were found after stringent filtering. Two of these 14 sORFs were located in lncRNA. However, these two sORFs (on chromosome 2, strand -1, spanning from coordinates 127 792 297 until 127 792 467, see figure 4.5) have the same stop location, were located in the same lncRNA (Ensembl ID ENSMUST00000135091) and differed only in length.

To validate these mouse sORFs, the human homologs of the two sORFs located in this lncRNA, were searched using protein BLAST\(^5\). A hit with an E-value of $1 \cdot 10^{-30}$ for the longer and $1 \cdot 10^{-28}$ for the shorter sORF was found, identifying human lncRNA LINC00116.

\(^5\)http://blast.ncbi.nlm.nih.gov/Blast.cgi
4.2 SPLICE-AWARE ASSEMBLY

(LNCipedia ID: lnc-NPHP1-1:5) as a homolog. In LNCipedia, the homologous human IncRNA is described as coding as there is strong proof for conservation (PhyloCSF score of 460,0287). Moreover, the homologous human sORF was earlier identified using ribosome profiling. This human sORF could be almost perfectly (94% identical) aligned to the sORF found in mouse using Clustal Omega\textsuperscript{6}. The alignment to the longer mouse sORF is given below:

\begin{verbatim}
Mouse MADVSLQVSVLVAFASGVVLQWQANRLRRYLDWRKRRLQDKLATTQKKLDA
Human MADVSLQVSVLVAFASGVVLQWQANRLRRYLDWRKRRLQDKLAATQKKLDA
\end{verbatim}

4.2 Splice-aware assembly

The assembly program was extended and revised to allow the construction of possible spliced sORFs, next to the earlier found single-exon sORFs. The script can be found in addendum chapter A.1. It can be interesting to identify spliced sORFs because they could have more chance to be identified by standard proteomics \cite{14, 28}. The splice-aware version of the assembly identified 164 582 possible sORFs in total in mESC using Ensembl version 75 (table 4.4), approximately 100 000 sORFs more than the original assembly (figure 4.4).

As for the original assembly in chapter 4.1.3, ribosomal coverage can be used to filter for a more stringent collection of sORFs. However, the filter thresholds need to be redetermined. Standardly, sORFs with a coverage of zero are filtered out silently. For the threshold calculation, theoretical sORFs with a coverage of zero are also included, resulting in 1906 single-exon 3'UTR sORFs and 25 spliced 3'UTR sORFs which are supposed to serve as a null distribution. To exclude 90% of the null distribution, thresholds were set at 0.23 for single-exon sORFs and 0.04 for spliced sORFs. Filtering with these thresholds resulted in addendum table A.1.

The splice-aware assembly was also applied on human colon tumor (HCT116) cells. Therefore, in-house lactomidomycin and cycloheximide RIBOseq data, accessible in the NCBI’s Gene Expression Omnibus \cite{112} through accession number GSE58207\textsuperscript{7}, were used. The data were obtained by applying the ArtSeq-protocol\textsuperscript{8} (Epicentre).

For analysing this dataset, following parameters were used. Annotation and translation info was based on Ensembl version 70 and assembly version GRCh37. Only unique STAR mappings were saved and AGATCGGAAGAGCACCGTCTGAACTCC was used as adaptor sequence. Sequences between 10 and 100 AA were allowed. The TIS peak had to be the

\textsuperscript{6}http://www.ebi.ac.uk/Tools/msa/clustalo/
\textsuperscript{7}http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58207
\textsuperscript{8}http://www.illumina.com/products/truseq-ribo-profile.html
Figure 4.5: Two interesting sORFs located in lncRNA ENSMUST00000135091. The sORF is positioned on the negative strand and has two possible TISs (in blue), but with the same stop position. Each TIS can be very clearly distinguished based on the HARR data. The longest sORF sequence is indicated in red. The CHX reads show an approximately equal distribution over the sequence. Interestingly, the homologous human lncRNA, containing an homologous sORF, showed coding capacities in LNCipedia.
### Table 4.4: Amount of sORFs identified by the splice-aware assembly in mESC, grouped by annotation class. Within each class, they are also split in single-exon and spliced sORFs. Also, a distinction is made for sORFs in-frame with the exonic main sequence. NA means that the frame is not applicable because there is no overlap with an exonic canonical sequence.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Spliced</th>
<th>In-frame</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>3’UTR</td>
<td>No</td>
<td>NA</td>
<td>924</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>NA</td>
<td>14</td>
</tr>
<tr>
<td>5’UTR</td>
<td>No</td>
<td>NA</td>
<td>18,757</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>No</td>
<td>5,804</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>1,030</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>NA</td>
<td>1,200</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>1,943</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>302</td>
</tr>
<tr>
<td>Exonic</td>
<td>No</td>
<td>No</td>
<td>81,722</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>15,905</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>26,554</td>
</tr>
<tr>
<td>Intergenic</td>
<td>NA</td>
<td>NA</td>
<td>4,206</td>
</tr>
<tr>
<td>Intronic</td>
<td>No</td>
<td>NA</td>
<td>4,094</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>519</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>ncRNA</td>
<td>No</td>
<td>NA</td>
<td>1,418</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>NA</td>
<td>49</td>
</tr>
</tbody>
</table>
maximal value in range of 1 codon (3bp) at each side. The R value cut-off was set to 0.01 and the minimal coverage count for the TIS was set at 5. An exonic overlap threshold of 50% was chosen.

In total, 163,289 possible sORFs were identified for HCT116 cells (addendum table A.2). The distribution was similar to the one found in mESCs.

4.3 Validation with proteomics data

After performing the splice-aware assembly, both for mESC and HCT116 data, a small translation product database (in fasta format) was constructed. Based on these search spaces, holding sORFs derived from RIBOseq data, different existing proteomics data of either mouse or human samples were searched for micropeptides as described in chapter 3.5.

In the first searches, good PSMs were mostly found for sORFs with an in-frame exonic part. This part always covered less than 50% of the sORF though as was implemented as a filter in the assembly. Nevertheless, all validated spectra and RPFs were found in this exonic part and no evidence could be found for translation in the other part of the sORF. An example of such a sORF is given in addendum chapter A.5.1 and can be considered as a false hit.

Another source of noise could be found in spectra that matched the SwissProt search space not good enough to become validated, although a doubtful match could be suspected manually. Therefore, the spectrum was nevertheless included in the follow-up analysis against the custom sORF database where it resulted in a false positive hit. An example of such a case is given in addendum chapter A.5.2. To prevent such false spectrum matches in subsequent analyses, every spectrum that had a promising match against the custom sORF search space, was manually revised in the preceding search against SwissProt to check if the match is much better against the custom sORF search space compared to the possible match with the SwissProt database.

An N-terminal extension was found in mESC shotgun data for the mouse gene coding for the negative elongation factor B (Nelfb gene). In the 5’UTR region of the canonical gene, 4 sORFs (sORF IDs: 84164, 92030, 86518 and 87272) were initially found, all starting at a different position. Those 4 sORFs run through the first exon of the coding sequence and they all stop at the same end position in the first intron, right behind the first exonic coding part of the canonical gene. This is schematically shown in addendum figure A.6. A spectrum match (addendum figure A.7) and good ribosomal coverage (addendum figure A.8) was found in 5’UTR part of the 4 sORFs, suggesting the existence of this micropeptides in the first place. However, another peptide (NH$_2$ - ATLEAAGER - COOH) in this 5’UTR region was found by Crappé et al. [44] during a search for new proteoforms, suggesting the N-terminal extension of the Nelfb gene rather than the existence of these
4.3. VALIDATION WITH PROTEOMICS DATA

**Figure 4.6:** Location of sORFs 82809 and 89244 (red rectangle) in the Mkks transcript in UCSC genome browser. The sORFs are located in one exon of the 5’UTR region (small blue bars) without any overlap with the canonical coding region (thicker blue bars).

Micropeptides translated from uORFs.

Another N-terminal extension could be found for an isoform of human 3-hydroxy-3-methylglutaryl-coenzym A reductase (HMGCR), although the proteomics and RIBOseq evidence was weaker in this case (addendum chapter A.5.3).

Mouse synaptosome peptidomics data were obtained by the research group of functional genomics and proteomics (Kurt Boonen, KU Leuven). Peptidomics aims to identify the naturally occurring peptides in a tissue or organ instead of analysing peptides from tryptic cleavage of proteins. The setup of this study was designed to enrich the brain sample for synaptosomes. These are small vesicles that arise when the synapses of neurons are sheared off from the cell body. The synaptosomal fraction can be purified (REF) and peptides (neuropeptides as well as micropeptides) are enriched by a specialised peptide extraction protocol. This protocol uses acidified methanol (REF 17559849) with additional delipidation steps [116].

A spectrum in this data was matched to the protein sequence of the small integral membrane protein 20 (Smim20). Smim20 is translated from a spliced sORF of 69 codons. For mouse, this protein is included in TrEMBL (UniProt entry D3Z7Q2) but not in SwissProt, meaning that the sequence is not yet manually annotated and reviewed after protein prediction using *in silico* algorithms. Three spectra confirmed the existence of the protein in mouse (addendum figure A.11). Furthermore, ribosome profiling (addendum figure A.12) suggested the existence of an N-terminal extension, resulting in a micropeptide of 97 amino acids, although no peptide was found to confirm this extension. The human homologue of Smim20 (protein BLAST E-value of $2 \cdot 10^{-11}$) is already manually annotated and reviewed in SwissProt (UniProt entry Q8N5G0). In sum, these results can support the validation of Smim20, a micropeptide translated from a spliced sORF in mouse, to transfer it from TrEMBL to SwissProt.

The most promising result was observed in mouse synaptosome data as well. In the 5’UTR region of the McKusick-Kaufman syndrome (Mkks) transcript, two overlapping sORFs (sORF IDs 82809 and 89244) were found. The two sORFs (strand -1) have the same end position but sORF 82809 starts more upstream and is longer: 60AA. The other sORF, 89244, is 45AA long. Both sORFs are located in the 5’UTR region without any overlap with the canonical coding sequence of the Mkks transcript as can be seen in figure 4.6.

Both sORFs have a very uniform ribosomal coverage in the ‘untreated’ lane (figure 4.7), but it can be seen that the TIS of the longest sORF (82809) is much more clear in the ‘treated’ lane. Therefore, sORF 82809 was investigated and validated further.
Near the end of the sORF, a spectrum of the mouse synaptosome proteomics data could be matched with the peptide $NH_2 - DSLGLASPPKS - COOH$ with 89% confidence (figure 4.8). The spectrum scored much worse against SwissProt, minimising the chance of a false hit. Moreover, the resulting micropeptide did not give any significant hit in a protein BLAST against all known mouse proteins. However, there was a homologue found in human cells called PNAS-117 (E-value $2 \cdot 10^{-26}$). An alignment of the micropeptide of mouse sORF 82809 and the human micropeptide PNAS-117 is given below:

Mouse  MSFQNLWREYKVLVMVPL I GFIHLGWHR1KSSPVPQVPK - DDTTELDSLGLASPPKSQT - - * * : : * * * * : * * * * : * * * * * : * : ** * * * : ** * *** * * * : * * * *  
Human  MSLRNAWRDYKVLVVMPVGLHLGWYRIKSSPVPQ I PKNDD1 PEQDSLGLSNLQSKQI QGK  

Akimoto et al. [117] could identify this human homologue (in their paper termed uMkks1) in human HeLa cells. They stated that 6 uORFs in the Mkks transcript are important in the repression of the downstream main ORF, encoding a 570AA protein. Furthermore, they could conclude that two of these uORFs (uMkks1 and uMkks2) encode micropeptides that are localised to the mitochondria. However, in our knowledge, this is the first time the micropeptide was identified in mouse cells.

A clear link between the uORFs and the Mkks pathogenesis is not yet found. McKusick-Kaufman syndrome is an autosomal recessive disorder that causes hydrometrocolpos, postaxial polydactyly and congenital heart disease [117, 118]. Missense mutations in the main ORF are found in patients as a cause of a defect in the protein [117].

Later, Dr. Kurt Boonen generated a second dataset of the mouse synaptosome. This time, a two-dimensional LC separation was used. The first dimension consist of RP-LC (reverse phase LC) at high pH. The resulting fractions (10) were analysed by LC (low pH)-Orbitrap MS. The data is currently being analysed [116].
Figure 4.7: Ribosomal coverage in sORFs 82809 and 89244 as seen for mESC in UCSC genome browser. The two upper lanes are ‘untreated’ RIBOseq data. The two lanes below are ‘treated’ data. In green are RIBOreads in sense. Antisense reads are in red. At the bottom, possible RefSeq transcript structures are given. The TIS of sORF 82809 is much more clear.
Figure 4.8: A PSM of 89% confidence (mouse synaptosome proteomics data) could be found for peptide $\text{NH}_2 - DSLGLASPPKS - \text{COOH}$ at the end of the sORFs 82809 and 89244 in the 5’UTR region of the Mkks transcript.
Chapter 5
Discussion

The single-exon RIBOsORF pipeline was developed as a specialised pipeline for identifying single-exon sORFs. At the moment, not every existing RIBOseq dataset can be analysed because the pipeline needs input from both an 'untreated' and a 'treated' sample, which is not available for every experiment, although more and more experiments consist of both [42, 43, 53, 55]. So, the pipeline will surely yield interesting results from future ribosome profiling data. Moreover, its implementation in the Galaxy framework allows wet lab researchers to use it setting their own parameters in a very easy and intuitive way.

Besides construction of all possible sORFs out of ribosome profiling data, the pipeline also filters them to select the most promising sORFs based on different statistics which can be added as separate building blocks. The benefit of these filters could be proven as very stringent filtering lead to the identification of two sORFs in lncRNA which could be validated with LNCipedia (chapter 4.1.5). Currently, three filters are already implemented (table 4.3).

The first filter, based on coverage, is calculated in the assembly step as it describes the representation of the sORF in the ribosome profiling data. Coverage can be a meaningful filter as validation with proteomics still remains rather difficult. Therefore, coverage is a good measure to capture the results of RIBOseq and it gives the filtering an experimental base.

PhyloCSF is the second filter of the pipeline and is based on conservation. PhyloCSF involves the strongest selection by preserving around 10% of the original sORFs. The main question remains if conservation is a typical hallmark of coding sORFs as this filter might impede the identification of species-specific sORFs.

On the other hand, FLOSS filtering, based on fragment length similarities, has the lowest impact because the majority (90%) of the sORFs can pass this filter. Nevertheless, Ingolia et al. [53] could prove the benefit of this filter in selecting true ribosomal footprints.

sORFs could be found in different annotation classes (table 4.3 and figure 4.4). Lots of possible sORFs were found in the 5’ leader sequence. However, this does not imply that all of them encode a functional peptide as uORFs can have a rather regulatory
role by decreasing the downstream initiation [31]. But, as can be seen in table 4.3, some small uORFs can have high coverage and conservation, suggesting that the peptide itself is functional [119]. One must always pay attention though when identifying a sORF that overlaps in-frame with a known exonic region as this can rather be an N-terminal extension of a known gene product. Anyhow, integration of RIBOseq and MS-based proteomics or peptidomics can help in the re-annotation of the genome.

Due to a high background signal of known protein-coding genes and a higher level of conservation, more false positive sORFs were found in exonic regions (making it the biggest class). To prevent this high number of false positives, a filter on exon overlap is included in the assembly step. However, a recent ribosome profiling study found that at least 1% of the human protein-coding genes show dual coding capacities [65]. To better take this dual coding into account, the assembly should in fact only use this exonic overlap filter if the sORF is in-frame with an annotated ORF. This frame dependent exon overlap filtering was introduced in the splice-aware version of the assembly.

Also intronic regions can be valuable to investigate as small translation products can be overlooked as a result of a wrong annotation. This is especially true for intronic sORFs that are located completely in intronic regions without exonic overlap.

Intergenic regions are defined as the regions which intervene between known annotated genes without any overlap. However, this does not mean that possible functional sORFs cannot be located in these regions as Hanada et al. [23] identified 49 intergenic sORFs that resulted in a visible phenotype in Arabidopsis. Intergenic sORFs in close proximity to known genes can be the result of misannotation. sORFs at distances of more than several kb from known genes on the other hand, are maybe the result of a yet unknown gene or transcript. Therefore, also the distance to the closest gene is calculated for intergenic sORFs.

The 3'UTR regions are mostly considered as being truly non-coding. However, in a recent study using ribosome profiling, it could be shown that stop codon read-through occurs more than previously thought [120]. So, it is recommended to not exclude these theoretical sORFs as eventually, some interesting findings can be at hand.

In chapter 2.3.1, the possible existence of small translation products in former non-coding transcripts was discussed. Although it is one of the smallest classes identified by the pipeline (figure 4.4), it is expected that lots of sORFs found in several ncRNAs have potential to be responsible for certain biological functions [46, 121]. In some cases, proteomics evidence is already found in human samples for sORFs originating from previously supposed ncRNA transcripts [7].

Aspden et al. [14] suggested the existence of two types of sORFs. First, the ‘dwarf’ sORFs of around 20AA long, being less conserved and less efficiently translated. Secondly, the ‘longer’ sORFs of around 80AA long demonstrating translational efficiencies that are more in line with those of canonical proteins and having a biased function towards membranes. It can be argued if the products of these longer sORFs are really micropeptides pur sang rather than a consequence of the arbitrary border of 100AA [27].

Picking up and validating small translational products with MS-based proteomics remains a
difficult task (as stated in chapter 2.2.2). But as the longer sORFs have better translational efficiencies, they should be much more easy to detect, making them interesting targets in a proteogenomics approach [14, 51]. However, because longer sORFs resemble more to canonical proteins, one can expect that longer sORFs more often arise after splicing in comparison to the dwarf types [28]. Also, there is a positive correlation between the length of the coding sequence and the occurrence of introns [122]. Therefore, the assembly was completely redesigned to allow the construction of a broader splice-aware proteomics search space.

The splice-aware version of the assembly identified approximately 100 000 extra sORFs compared to the original assembly. This has several reasons. During the development of the splice-aware assembly, some things were changed influencing the numbers of the identified single-exon sORFs presented in figure 4.4. First, mouse Ensembl version 75 was used instead of version 72, which resulted in some slight variances in amounts of identified sORFs. Furthermore, the exon overlap filtering was adapted. Before, all sORFs with more than 50% exon overlap were discarded. In the splice-aware assembly, exon overlap filtering was only executed for in-frame sORFs, reasoning that sORFs which are out-of-frame would result in totally different translation products than the overlapping canonical sequence. Therefore, the amount of exonic single-exon out-of-frame sORFs rose drastically (approximately +60 000 sORFs) in table 4.4, compared to all other numbers, which stayed around the same value.

Next to changes for single-exon sORFs, the new assembly introduces of course also spliced sORFs (table 4.4). There are much less spliced sORFs identified compared to single-exon sORFs. This can be easily explained. Smaller coding sequences have fewer introns [28]. For sORFs, it can be expected that splicing is less abundant than for canonical large translation products. Nevertheless, splicing can be important as spliced sORFs can resemble classical proteins so that this class is more easily picked up with standard MS-based proteomics. Most spliced sORFs are found in exonic and 5'UTR regions as most of the splice junctions are located there. However, spliced sORFs can be found in every annotation class except intergenic as splicing is possible in each genic region.

Thresholds for subsequent coverage filtering of sORFs are currently based on assuming all theoretical 3'UTR sORFs as a null distribution. However, a threshold for spliced sORFs based on a null distribution of only 25 spliced 3'UTR sORFs is rather unjustified. Therefore, other methods to determine these thresholds can be used in the future. For example, a totally scrambled genome can be used as null distribution. Otherwise, a null distribution could be sampled from a masked genome where genes with GENCODE level\(^1\) 1 and 2 are discarded [123].

After applying the splice-aware assembly, a small translation product search space can be constructed for MS-based proteomics. As ribosome profiling is only an indirect measure of protein abundance, mass spectrometry remains necessary for true protein or peptide identification [11]. There are a few studies that identified micropeptides with proteomics

---
\(^1\)http://www.gencodegenes.org/gencodeformat.html
When searching for MS-based identifications, attention is needed to prevent false hits. PSMs that are positioned in an in-frame exonic part of the sORF as well as spectra that match with a comparable confidence to both SwissProt and the custom sORF search space, should be left out of the reports. PSMs to sORFs, starting in a 5’UTR, that are in-frame and overlapping with the canonical coding ORF, need to be investigated carefully to verify whether these PSMs point to an N-terminal extension of the canonical ORF.

The need for guidelines to prevent false positive identifications is not only emerging for sORF validations but can be observed in general in the proteogenomics research field as well. The identification of novel events against a customised protein sequence database should not be executed in exact the same way as against a protein reference database. Therefore, proteogenomics research requires adapted statistical procedures and more efforts to eliminate the most common false positives [124]. These needs open the debates to define new standards for analysing and reporting in current proteogenomics [125]. In this research, false positive checks were performed manually but it seems a logic evolution to include these controls in the default proteomics analysis suites.

In the performed MS searches, translational evidence for a spliced mouse sORF (Smim20) which was included in TrEMBL but not yet in SwissProt, was found. Although the search space contains far less spliced candidates than single-exon ones, this spliced sORF could nevertheless been confirmed with MS-based proteomics, supporting the benefits of an extension of the assembly to a splice-aware version.

The most promising proteomics identification (the uORF in the Mkks transcript) was found in mouse synaptosome data. This suggests that micropeptides are also expressed in specific tissues for specific functions and not only for more general purposes as development, morphogenesis and stress conditions. Also other studies found micropeptides in tissue specific samples as for example Alba et al. [126] identified two micropeptides in human testis data. Furthermore, the synaptosome data were generated involving specific enrichment, fractionation and separation procedures. Sample preparation steps that enrich the low-molecular weight proteins, were also reported to be beneficial in other studies confirming micropeptides with peptidomics [7, 124]. Therefore, further investment in better micropeptide extraction and enrichment steps in proteomics protocols is needed so that better MS-based identifications are possible in the future.

This work demonstrated that a multi-perspective approach is the way to go when trying to identify functional micropeptides. The RIBOsORF pipeline starts off with mapping and TIS determination based on RIBOseq data, constructs the sORFs in silico, calculates statistics, filters both using in silico metrics as well as based on RIBOseq experimental data and finally prepares a search space for an MS-based proteomics validation. This proteogenomic combination of building blocks can probably unravel the mystery around this long time overlooked class of molecules, similar to the research done on microRNAs, and in that way the best targets for a further functional characterisation can be selected.
Chapter 6

Further research

Over the last years, the interest in sORFs and micropeptides has increased drastically. Although this research field is still taking its first beginning steps, the RIBOsORF pipeline is designed to allow a genome-wide identification of small translational products. As the field will constantly make drastic progressions in the near future, the RIBOsORF pipeline will be evolving as well, both by improvements in the existing building blocks as by the addition of new ones.

A first optimisation can be applied to the FLOSS cutoff smoother. In figure 4.3, it can be seen that the smoother goes up in the last part of the graph, probably due to sparser observations in function of the amount of reads. To assure this assumption, a sampling test could be useful. There can be evenly sampled for each amount of reads based on the reference fractions which can be used as sampling probabilities. The obtained cutoff smoother of this test can be compared to a second test where a less denser sampling is done in the part where the original cutoff smoother went up. If the effect is present in the second but not in the first test, then this effect can indeed be explained by the lower amount of observations for high amounts of reads. In that case, another goodness-of-fit statistic which can be extrapolated, could be chosen. Now, the Loess smoother is used, based on local regression. With extrapolation, a cutoff value for all amount of reads could be calculated so that also sORFs with very high amounts of reads could be classified instead of falling out of range. This is very interesting as high amounts of reads make those sORFs very informative.

Until now, only the assembly step and the translation product database generation were extended to construct a splice-aware search space for MS-based proteomics. However, other tools like FLOSS and PhyloCSF can also be adapted to work for both spliced and single-exon sORFs, expanding their potential and making the whole existing pipeline splice-aware.

Besides improvements, RIBOsORF can also be extended with new tools. A first tool which is being implemented in the RIBOsORF pipeline is the ORFscore, a score based on triplet periodicity [51] (see addendum figure A.14). With this ORFscore, the pipeline can look at...
mapped RPFs in relation to the frame they are in. This feature can be used to pinpoint
dually coding regions. Because of the frame-awareness of this metric, it can come in handy
to determine whether sORFs should be filtered on exonic overlap.

Another possible extension could search for a link between genomic variations in sORFs
(indels, single nucleotide polymorphisms) and phenotypical variations.

Recently, Gao et al. [55] reported a very important new variation of the ribosome profiling
protocol: QTI-seq. With the sequential treatments of first LTM and later PMY, elongating
ribosomes are depleted very quickly and initiation sites can be determined without needing
an incubation period. Therefore, this technique is very suited to make the transition from
conventional ribosome profiling on cell culture to tests on in vivo samples and tissues.

Tarsal-less has never been picked up in MS-based proteomics studies, although it contains
some of the most studied sORFs. A possible explanation can be found in the fact that
known micropeptides are only expressed at specific locations and at specific times [46, 77].
Therefore, future searches are recommended to be done in different specific samples like for
instance the synaptosome. Besides that, better and alternative extraction methods need
to be developed to obtain a more effective isolation of micropeptides [2]. This can be com-
plemented with efficient enrichment steps. Schwaid et al. [127] could for example identify
16 micropeptides after applying an affinity-based enrichment of cysteine containing human
micropeptides. Thus, a continuous improvement of mass spectrometry based techniques
will be one of the key targets in the search for micropeptides.

After identification, the best candidates can be functionally characterised to finally deter-
mine their function and localization. Functional assays can comprise many techniques, for
example knock out studies involving frame shift mutations with TALEN [128], knock out
studies involving deletion mutants [21], overexpression studies [23], in situ hybridisations
[128], GFP fusions [128], knock down studies involving silencing, interaction studies,...

Given their small size and their putative character to interact with larger proteins, mi-
cropeptides are possibly useful as added agents. These molecules could affect specific
proteins and processes [18]. Therefore, unraveling the function of micropeptides can have
serious implications on biology and medicine.

To guide the multi-disciplinary research on micropeptides, it can be useful to construct a
web-based central repository where all research groups can upload their results on sORFs
and micropeptides. Therefore, our lab is designing such a repository, called SORFS.org
(addendum figure A.15). The purpose is to comprise all information known about sORFs
and their possible small translation products. Moreover, it is especially interesting as
interspecies relations can be constructed between homologous entries.
Chapter 7

Conclusion

Classical genome-wide prediction tools use a 100AA bottom limit to evade false positive identifications. Because of this limit, micropeptides encoded by small open reading frames (sORFs) were long time overlooked. Recently, this new class of bio-active molecules is getting attention and is expanding the complexity of the proteome and its regulation.

Identification is possible \textit{in silico}, based on sequence characteristics or conservation. Besides that, ribosome profiling (RIBOseq), the sequencing of ribosome-protected RNA fragments (RPFs) using the advantage of different antibiotics treatments to stall elongating and/or initiating ribosomes, offers a genome-wide identification technique based on next generation sequencing experiments. Although it is a rather new technique, specialised computational analysis tools are developed right now.

The RIBOsORF pipeline starts with mapping two ribosome profiling experiments: one where all ribosomes and one where only the initiating ribosomes were stalled. From these, translation initiation sites (TISs) can be determined and subsequently sORFs can be constructed \textit{in silico}. Afterwards, different characteristics can be calculated for each sORF: ribosomal coverage, the PhyloCSF conservation score and the FLOSS size distribution similarity score. More tools to further expand this pipeline are surely at hand. Furthermore, the pipeline was recently extended to be able to work with spliced sORFs, hoping to better pick up the more canonical-like sORFs.

Out of this pipeline, a search space for MS-based proteomics validation can be generated. However, one needs to check manually for false hits caused by e.g. exonic overlap. Some first tentative proteomics identifications were found in mouse synaptosome data, suggesting that there is also significance for micropeptides in specific temporal and spacial ranges. Proteomics protocols need to be further improved to obtain a better extraction and enrichment of the micropeptide fraction and to permit a more confident validation. Afterwards, most promising targets can be selected for a functional characterization.

Our RIBOsORF pipeline offers a perfect bioinformatics environment to push this research further from a multi-disciplinary viewpoint.
Bibliography

[23] Hanada K, Zhang X, Borevitz JO, Li WH, and Shiu SH. A large number of novel coding small open reading frames in


[116] Boonen K. Personal communication.


Addenda

A.1 Scripts

All scripts can be found on http://porthos.ugent.be/share/sorfs/.
(user name: guestThesis@biobix.be, password: micropetides)

The single-exon RIBOsORF pipeline consists of the sequence of following scripts:

1. Mapping of the reads: 1_mapping.pl
2. TIS determination: TIScalling_sORFs.pl
3. Creation of an overview table with the TIS calling IDs and parameters: TIS_sORFs_overview.pl
4. Single-exon assembly: assembly_sORFs.pl
5. PhyloCSF: first creation of input with PhyloCSF_sORFs_input2.pl, next score calculation with Phylocsf_sORFs.pl
6. FLOSS calculation: Floss_sORF.pl
7. Translation database construction: Translation_assembly.pl

The splice-aware version of the RIBOsORF pipeline consists of:

1. Mapping of the reads: 1_mapping.pl
2. TIS determination: TIScalling_sORFs.pl
3. Splice-aware assembly: assembly_sORFs_spliceAware.pl
4. Translation product database construction: Translation_assembly_spliced.pl
A.2 Galaxy version

The Galaxy version of the RIBOsORF pipeline is available at http://athos.ugent.be. It is possible to login with the user name guestThesis@biobix.be (password: micropeptides). The sORF pipeline is located in the BioBix tools and consists of following steps:

- Step 1: the mapping of RIBO-seq data using two fastq files as input.
- Step 2: TIS calling in all genic and intergenic regions.
- Step 3: TIS overview to generate a tab separated file with an overview of all parameters for each TIS ID.
- Step 4: SNP calling (this step will be implemented in the near future).
- Step 5: sORF assembly to construct all possible single-exon sORFs based on TISs.
- Step 6a: PhyloCSF input creation to construct sORF specific multiple alignments.
- Step 6b: PhyloCSF sORF coding analysis calculates a PhyloCSF conservation score and classifies the sORFs based on this score.
- Step 7: FLOSS calculation and classification based on fragment length distributions.
- Step 8: generate a small translation product dataset in fasta format.

All these steps can be combined in one workflow, as can be seen in figure 3.3 in chapter 3.3.7. The RIBOsORF pipeline workflow is available in the workflow tab of the guestThesis account. The pipeline is also pre-executed for a set of example parameters on mouse mESC data, making the results available in the Galaxy history.

A.3 Single-exon RIBOsORF pipeline

A.3.1 Identification of sORFs in different annotation classes

A schematic representation of all possible annotations is given in figure A.1.

A.4 Splice-aware assembly

A.4.1 mESC

The 164 582 possible sORFs found in the splice-aware version of the assembly could be filtered based on ribosomal coverage with the thresholds determined by the 3’UTR null distribution. This resulted in table A.1.
Figure A.1: sORFs are possible in different annotation classes, sometimes with overlap to another annotation. sORFs in ncRNA are an extra annotation class which is not included in this scheme.
Table A.1: Amount of sORFs identified by the splice-aware assembly in mESC after filtering on coverage. The results are grouped by annotation class, splicing and frame overlap with the eventual exonic main sequence. NA means that the frame is not comparable because there is no overlap with an exonic canonical sequence. The coverage filtering threshold was set at 0.23 for single-exon sORFs and 0.04 for spliced sORFs, as determined by assuming all possible 3’UTR sORFs as a null distribution.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Spliced</th>
<th>In-frame</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>3’UTR</td>
<td>No</td>
<td>NA</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>NA</td>
<td>3</td>
</tr>
<tr>
<td>5’UTR</td>
<td>No</td>
<td>NA</td>
<td>5924</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>450</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>NA</td>
<td>249</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>561</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>65</td>
</tr>
<tr>
<td>Exonic</td>
<td>No</td>
<td>No</td>
<td>42606</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>2760</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>11986</td>
</tr>
<tr>
<td>Intergenic</td>
<td>NA</td>
<td>NA</td>
<td>605</td>
</tr>
<tr>
<td>Intronic</td>
<td>No</td>
<td>NA</td>
<td>603</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>ncRNA</td>
<td>No</td>
<td>NA</td>
<td>405</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>NA</td>
<td>21</td>
</tr>
</tbody>
</table>
A.5. VALIDATION WITH PROTEOMICS DATA

A.4.2 HCT116

The splice-aware assembly was also performed on human HCT116 data (table A.2).

Table A.2: Amount of sORFs identified by the splice-aware assembly in HCT116 cells, grouped by annotation class. Within each class, they are also split in single-exon and spliced sORFs. Also, a distinction is made for sORFs in or out of frame with the exonic main sequence. NA means that the frame is not comparable because there is no overlap with an exonic canonical sequence.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Spliced</th>
<th>In-frame</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>3’UTR</td>
<td>No</td>
<td>NA</td>
<td>1453</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>NA</td>
<td>50</td>
</tr>
<tr>
<td>5’UTR</td>
<td>No</td>
<td>NA</td>
<td>14 055</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>4 668</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>NA</td>
<td>784</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>648</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>1 368</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>187</td>
</tr>
<tr>
<td>Exonic</td>
<td>No</td>
<td>No</td>
<td>87 937</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>16 719</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>23 552</td>
</tr>
<tr>
<td>Intergenic</td>
<td>NA</td>
<td>NA</td>
<td>4 010</td>
</tr>
<tr>
<td>Intronic</td>
<td>No</td>
<td>NA</td>
<td>4 224</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>232</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>ncRNA</td>
<td>No</td>
<td>NA</td>
<td>3 101</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>NA</td>
<td>227</td>
</tr>
</tbody>
</table>

A.5 Validation with proteomics data

A.5.1 Validation of a sORF within its exonic overlap

As an example, sORF 78402 (mESC, chromosome 19, coordinates 5 274 651 - 5 274 926, strand -1) is described. 49% of this sORF has in-frame exonic overlap. In figures A.2 and A.3, it can be seen that all spectrum matches and RIBO-reads are found in this exonic part of the sORF.
Figure A.2: One of the twelve good PSMs in the exonic part of sORF 78402 (mESC shotgun proteomics data). This spectrum matches to the peptide sequence EQGFDPPII-IFVNQK with 100% confidence. However, no spectrum match could be found in the greater intronic part (figure A.3) of the sORF.

Figure A.3: Only the exonic part of sORF 78402 (mESC data) is covered with ribosome profiling data.
A.5. VALIDATION WITH PROTEOMICS DATA

Figure A.4: Peptide spectrum match with sORF 132127 (HCT116 shotgun proteomics data) in the 5’UTR of HIST1H3D. The spectrum matches to the peptide sequence ILELAGNAAR with 61% confidence.

A.5.2 Spectra that match both against SwissProt and sORFs

The spectrum (HCT116 shotgun proteomics data) in figure A.4 matches sORF 132127 (HCT116, chromosome 6, coordinates 26 199 079 - 26 199 297, strand -1), a sORF located in the 5’UTR region of HIST1H3D, with 61% confidence. However, the spectrum matches HIST1H2AD (figure A.5) with a comparable confidence in the preceding SwissProt search.

Figure A.5: Peptide spectrum match (HCT116 shotgun proteomics data) in HIST1H2AD. The spectrum matches to the peptide sequence ILELAGNAAR with 62% confidence during a search against SwissProt.
A.5.3 Identification of N-terminal extensions

*Nelfb* gene

In the 5’UTR region of the *Nelfb* gene, 4 sORFs could be found (mESC, chromosome 2, strand -1):

- sORF 84164, genomic positions 25 211 075 - 25 211 296
- sORF 92030, genomic positions 25 211 075 - 25 211 302
- sORF 86518, genomic positions 25 211 075 - 25 211 347
- sORF 87272, genomic positions 25 211 075 - 25 211 356

Those 4 sORFs start at different positions of the 5’UTR region of the *Nelfb* gene, cover the coding part of the first exon and end downstream of it in the begin of the first intron. This is schematically shown in the alignment in figure A.6.

![Alignment of sORFs in Nelfb gene](image)

**Figure A.6:** An alignment showing the position of 4 found sORFs in the *Nelfb* gene in mouse. The intron of the *Nelfb* gene is represented schematically.

A peptide was found in those 4 sORFs, outside of the overlapping canonical part (figure A.7).

Moreover, a good ribosomal coverage could be observed outside of the overlapping canonical part (figure A.8). The peptide \( NH_2 - ATLEAAGER - COOH \) (positioned at the start of the longest sORF) was already found in the PROTEOFORMER paper [44], suggesting that the sORFs are rather indicating N-terminal extensions than newly identified micropetides.

3-hydroxy-3-methylglutaryl-coenzym A reductase

A weak PSM (HCT116 shotgun proteomics data) was observed in a sORF which is located in the 5’UTR region of human HMGCR (figure A.9). This sORF starts in the 5’UTR
**Figure A.7:** Peptide-to-spectrum match (mESC shotgun proteomics data) with the 4 found sORFs in the 5’UTR region of the *Nelfb* gene in mouse. The spectrum matched to the peptide sequence \( NH_2 - SGAPSGSATAPAER - COOH \) with 100% confidence. This peptide sequence is located in the 5’UTR region and not in the overlapping part with the canonical coding sequence.

**Figure A.8:** The 4 found sORFs (strand -1) in the *Nelfb* gene show a good overall ribosomal coverage.
Figure A.9: A weakly validated spectrum with confidence of 48% could be found in isoform 3 of HMGCR (HCT116 shotgun proteomics data). The spectrum matches with peptide \( \text{NH}_2 - \text{RRRRGTDFQRES} - \text{COOH} \).

region but overlaps (in-frame) with isoform 3 (identifier P04035-3 in SwissProt) of HMGCR (ribosomal coverage in figure A.10). Therefore, this is probably rather new evidence for an N-terminal extension than for a micropeptide.

A.5.4 Mouse small integral membrane protein 20

Three spectra were found in the mouse small integral membrane protein 20 (Smim20) (figure A.11). With ribosome profiling, evidence for N-terminal extension of this micropeptide could be found as well, resulting in a peptide of 97 amino acids (figure A.12).

A.6 PROTEOFORMER

PROTEOFORMER [44] is a proteogenomic pipeline, somewhat related to the RIBOseq pipeline, available in a script-based version as well as embedded in Galaxy. It searches for true in vivo proteoforms and generates a protein sequence search space for MS/MS identification, all starting from RIBOseq data. The pipeline (schematically given in figure A.13) starts from two RIBOseq fastq files and aligns them (using STAR or TopHat2) against the reference genome. Quality control can be executed by determining in which genes the RPFs are situated and by performing a metagenic functional classification. Based on the mapped ribosome profiles, transcripts can be constructed and TISs determined. Furthermore, SNPs can be identified and all translation products can be assembled. All
A.7. FUTURE RESEARCH

Figure A.10: Ribosome profiling coverage of the N-terminal extended isoform 3 of HMGCR.

These translation products can finally be exported to a non-redundant fasta-formatted database, which makes up a search space for MS-based proteomics.

Although not the main goal of this research, the FLOSS calculation script could be adapted to function in the PROTEOFORMER pipeline, calculating the FLOSS for each (splice-aware) reading frame starting from an earlier identified TIS. For each reading frame an unique ID is constructed based on a concatenation of the transcript ID and the TIS genomic position. In that way, every possible reading frame can be searched for true reads.

A.7 Future research

A.7.1 ORFscore

RIBOsORF could be extended with extra tools like the ORFscore [51] (figure A.14).

A.7.2 SORFS.org

A central repository to store all identified sORFs and their status for different species is developed at our lab (figure A.15).
Figure A.11: One of three validated spectra in the mouse Smim20 peptide (mouse synapstosome proteomics data). This spectrum match had 100% confidence to be situated in the peptide $NH_2 – DVQPPGLKVWSDPFGRK – COOH$.

Figure A.12: Ribosome profiling coverage of the N-terminal extended spliced Smim20 micropeptide.
Figure A.13: Overview of the PROTEOFORMER pipeline. Starting from two in parallel RIBOseq experiments, the pipeline searches for true *in vivo* proteoforms. The pipeline generates a custom protein sequence search space which can be used in MS/MS identification analyses.
Figure A.14: Scheme to define the implementation of the ORFscore. First, all possible ORFs (from a distal AUG until the stop codon) are constructed for each frame (+1, +2 and +3). Next, for each frame, all RPFs falling in one of their ORFs are counted. Subsequently, the RPF distribution is compared to an uniform distribution using a modified chi-squared statistic. The obtained ORFscore is defined negative if the distribution of the RPFs does not accord to the frame of the main CDS and positive if they have the same frame [51].
Figure A.15: Screenshot of the central repository for micropeptides, SORFS.ORG, that is being developed at our lab. The purpose is to make a web-available database where new findings from multiple researchers can be stored and where links between different species are possible.