Computational Approaches to the Identification and Characterization of Non-Coding RNA Genes

PONTUS LARSSON
Dissertation presented at Uppsala University to be publicly examined in B21, Biomedicinsk Centrum, Husargatan 3, Uppsala, Friday, January 30, 2009 at 10:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Non-coding RNAs (ncRNAs) have emerged as highly diverse and powerful key players in the cell, the range of capabilities spanning from catalyzing essential processes in all living organisms, e.g. protein synthesis, to being highly specific regulators of gene expression. To fully understand the functional significance of ncRNAs, it is of critical importance to identify and characterize the repertoire of ncRNAs in the cell. Practically every genome-wide screen to identify ncRNAs has revealed large numbers of expressed ncRNAs and often identified species-specific ncRNA families of unknown function. Recent years’ advancement in high-throughput sequencing techniques necessitates efficient and reliable methods for computational identification and annotation of genes. A major aim in the work underlying this thesis has been to develop and use computational tools for the identification and characterization of ncRNA genes.

We used computational approaches in combination with experimental methods to study the ncRNA repertoire of the model organism Dictyostelium discoideum. We report ncRNA genes belonging to well-characterized gene families as well as previously unknown and potentially species-specific ncRNA families. The complicated task of de novo ncRNA gene prediction was successfully addressed by developing a method for nucleotide composition-based gene prediction using maximal-scoring partial sums and considering overlapping dinucleotides.

We also report a substantial heterogeneity among human spliceosomal snRNAs. Northern blot analysis and cDNA cloning, as well as bioinformatical analysis of publicly available microarray data, revealed a large number of expressed snRNAs. In particular, U1 snRNA variants with several nucleotide substitutions that could potentially have dramatic effects on splice site recognition were identified.

In conclusion, we have by using computational approaches combined with experimental analysis identified a rich and diverse ncRNA repertoire in the eukaryotes D. discoideum and Homo sapiens. The surprising diversity among the snRNAs in H. sapiens suggests a functional involvement in recognition of non-canonical introns and regulation of messenger RNA splicing.

Keywords: ncRNA, snRNA, U1, splice site, alternative splicing, Dictyostelium, nucleotide composition, partial sums

Pontus Larsson, Department of Cell and Molecular Biology, Box 596, Uppsala University, SE-75124 Uppsala, Sweden

© Pontus Larsson 2009

ISSN 1651-6214
ISBN 978-91-554-7386-0
urn:nbn:se:uu:diva-9518 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-9518)
Till mina föräldrar
List of Publications

This thesis is based on the following papers which will be referred to in the text by their Roman numerals:


† These authors contributed equally

Reprints were made with permission from the publishers.
## Contents

- Introduction .................................................................................................... 9
- Chemical properties of RNA ....................................................................... 9
- Non-coding RNA genes ............................................................................. 11
  - History ................................................................................................. 11
  - The ncRNA families .............................................................................. 12
  - Biogenesis .............................................................................................. 15
  - Structure ............................................................................................... 15
  - Function .............................................................................................. 16
- Protein-coding genes ................................................................................. 16
- Splicing ........................................................................................................ 17
  - Alternative splicing ................................................................................ 19
  - Regulation of alternative splicing .......................................................... 19
  - Alternative splicing and the proteome .................................................... 20
  - Splicing and disease .............................................................................. 21
- Post-genomic era ....................................................................................... 21
  - Genome evolution .................................................................................. 21
  - Comparative genomics ......................................................................... 22
- Identification of functional units ............................................................... 23
  - Experimentally based methods ............................................................... 23
  - Computationally based methods ............................................................ 25
- Present Investigations ................................................................................... 29
  - Non-coding RNAs in *Dictyostelium discoideum* .................................. 29
    - Experimentally based methods ............................................................ 30
    - Computationally based methods ......................................................... 30
  - Analysis of D. discoideum ncRNAs ......................................................... 34
  - Characterization of human spliceosomal snRNA variants .................. 37
    - Biochemical identification and characterization of U1 snRNAs .......... 37
    - Identification of expressed snRNAs using exon microarrays .......... 38
    - Characterization of expressed snRNA variants ................................ 39
- Conclusions and future perspectives ........................................................ 43
- Acknowledgements ...................................................................................... 45
- Swedish summary ....................................................................................... 47
- References .................................................................................................... 50
Abbreviations

A  adenine
AS alternative splicing
bp base pair
BP branch point
C  cytosine
cDNA complementary DNA
CM covariance model
DNA deoxyribonucleic acid
DUSE Dictyostelium upstream sequence element
EST expressed sequence tag
FDR false discovery rate
G  guanine
HMM hidden Markov model
miRNA micro RNA
mRNA messenger RNA
ncRNA non-coding RNA
nt nucleotide
ORF open reading frame
piRNA piwi-interacting RNA
PSE proximal sequence element
RISC RNA-induced silencing complex
RNA ribonucleic acid
RNase P ribonuclease P
RNP ribonucleoprotein
rRNA ribosomal RNA
scaRNA small cajal body-specific RNA
siRNA small interfering RNA
snoRNA small nucleolar RNA
snRNA small nuclear RNA
SRE splicing regulatory element
SRP signal recognition particle
SS splice site
T  thymine
TMG tri-methylated guanosine
tRNA transfer RNA
U  uracil
USE upstream sequence element
Non-protein coding RNA molecules (ncRNA), as the name implies, do not encode proteins and are never translated. Instead, they are functional at the level of RNA as e.g. regulators, structural scaffolds or catalytic factors. Although a number of abundant ncRNA families have been well studied and found to be crucial to core functions in cells, research advancements over the last ten years have made it clear that the number and diversity of ncRNAs have been massively underestimated. The structure, transcription and processing of ncRNA genes are fundamentally different from protein-coding genes which in part explain why ncRNA genes have been largely overlooked. Other methods must often be employed to identify ncRNA genes than can be used to identify protein-coding genes, in particular when it comes to computational prediction of genes in sequence data. Computational methods that are able to accurately and reliably identify ncRNA and protein-coding genes in sequence data are becoming increasingly important as sequencing techniques are becoming cheaper, faster and capable of producing enormous amounts of high-throughput sequence data.

In this thesis I will discuss mainly computational strategies to identify ncRNA genes. I will present our results on ncRNA genes identified in the model organism Dictyostelium discoideum using a combination of experimental and computational strategies as well as an in-depth study of the heterogeneity among human spliceosomal RNA genes.

Chemical properties of RNA

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are highly similar molecules. They are linear polymers built up by nucleotides (nt) which are linked together. A nucleotide is a five-carbon sugar with an organic base and a phosphate group attached at the 1' and 5' carbons, respectively (Figure 1A). In the polymer, the 3' carbon of one nucleotide is linked to the 5' carbon of the next nucleotide via a phosphodiester bond, giving the polymer a directionality. When writing out the sequence of a nucleic acid, a 5' to 3' directionality is implied unless specified otherwise. In DNA nucleotides, the sugar is deoxyribose while in RNA it is ribose. The difference is that in ribose, a hydrogen on the 2' carbon is replaced by a hydroxyl group. This subtle
change has important consequences as it provides RNA with a chemically reactive hydroxyl group that can take part in RNA-mediated catalysis. The hydroxyl group can also attack and break the phosphodiester bond between nucleotides. This makes the RNA molecule unstable compared to DNA.

The bases attached to the sugar are adenine (A), guanine (G), cytosine (C), thymine (T) or uracil (U). While A, C and G are common to DNA and RNA, T is exclusive to DNA and is replaced by U in RNA. Through hydrogen bond formation between the organic bases, inverted nucleotides are able to pair with each other, forming so-called base pairs (bp). In the classical Watson-Crick bp's, G pairs with C and A pairs with T or U. The Watson-Crick pairs are said to be complementary. In RNA, G can also form a bp with U, the so-called G-U wobble. The base pairing capabilities mean that complementary nucleic acid strands can pair with each other with high specificity to form a double-stranded structure (Figure 1B). Because of the steric properties of the linked nucleotides, the double-stranded nucleic acids wind around each other to form a double helix with the paired bases meeting in the center and the sugars forming a backbone on the outside (Figure 1C). It is worthwhile to mention that besides the classical two Watson-Crick bp's there are at least 26 additional possible bp's that involve at least two hydrogen bonds.

An unpaired (single-stranded) nucleic acid can also fold back onto itself and form base pairs between distant complementary nucleotides within the same molecule. This gives the molecule a distinct structure which in the case of RNA is often important for its function. It is often the case that a nucleotide substitution that destroys a base pair is compensated by a second nucleotide substitution that restores the base pairing between nucleotides again, a so-called compensatory base pair change (Figure 1D). While the primary structure of a nucleic acid describes the internal order of the nucleotides, the secondary structure describes the internal order but also which nucleotides are single-stranded and which nucleotides base-pair with each other.
Non-coding RNA genes

History

A few discoveries of high significance mark the history of ncRNA research. The ribosome was known to mainly consist of RNA already in the 40's (Claude, 1943) and transfer RNA (tRNA) molecules functioning as the link between the genetic nucleic acids and proteins were discovered in the late 50's (Hoagland et al., 1958). In the late seventies the concept of split genes and RNA splicing was independently discovered by Phillip A. Sharp and Richard Roberts (Berget et al., 1977; Chow et al., 1977). Subsequently, in the early eighties, ribonucleoprotein (RNP) particles containing a set of small nuclear RNAs (snRNA) were linked to splicing of messenger RNA (mRNA) (Lerner et al., 1980; Yang et al., 1981; Padgett et al., 1983). At about the same time, RNA was shown to have the ability to catalyze chemi-
cal reactions, a property previously only credited to proteins. Thomas Cech and co-workers found that an intron in the endogenous ribosomal RNA (rRNA) from *Tetrahymena thermophila* was able to catalyze splicing of itself *in cis*, in absence of proteins (Kruger et al., 1982; Cech, 1990). Shortly thereafter, Sidney Altman and colleagues were able to show that the cleavage of precursor tRNA molecules into mature tRNA is catalyzed by the RNA component of ribonuclease P (RNase P) *in trans* (Guerrier-Takada et al., 1983; Altman, 1990). Cech and Altman shared the Nobel prize in chemistry in 1989 for the discovery of catalytic RNA and Sharp and Roberts were awarded the Nobel prize in medicine 1993 for their discovery of split genes in eukaryotes. The discovery that exogenous double-stranded RNA injected into the nematode Caernohabditis *elegans* could specifically target and silence gene expression through highly specific anti-sense interactions with the target mRNA (Fire et al., 1998) together with the discovery that an abundant family of endogenous small RNAs, depicted micro RNAs (miRNA), could silence gene expression in a similar manner (Lee et al., 1993; Wightman et al., 1993; Reinhart et al., 2000; Lagos-Quintana et al., 2001; Lee and Ambros, 2001; Lau et al., 2001), clearly demonstrated the potential and involvement of ncRNAs in gene regulation. Consequently, the interest in ncRNAs exploded. For the discovery of gene silencing through small interfering RNAs (siRNA), Fire and Mello received the 2006 Nobel prize in medicine. In 2000, the large subunit of the ribosome from the bacterium *Haloarcula marismortui* was published at such high resolution that the long suspected fact that the peptidyl transfer reaction is catalyzed by the rRNAs could be concluded (Ban et al., 2000; Cech et al., 2006). In addition to these ground breaking discoveries, many other functions have been linked to RNA over the years.

The ncRNA families

The wealth and diversity of ncRNA families presently known makes a complete listing practically impossible. Here I will just focus on a few of the most abundant and widely occurring ncRNA families.

**Ribosomal RNA (rRNA)**

The rRNAs are present in all kingdoms of life. Together with ~50-80 proteins, they make up the ribosome where protein synthesis in the cell takes place. It had long been suspected that the ribosome might actually be a ribozyme, *i.e.* the peptidyltransferase reaction that forms the peptide bonds on the growing peptide would be catalyzed by the rRNAs (Crick, 1968; Moore and Steitz, 2006), but it wasn't until the solution of a high-resolution crystal structure of the large subunit of the ribosome from the bacterium *H. marismortui* in 2000 that this suspicion could be confirmed (Ban, 2000).
Transfer RNA (tRNA)

tRNAs are present in all kingdoms of life and provide the link between the nucleic acids of the genetic material and the amino acids of proteins. Each tRNA folds into a highly conserved secondary structure which exposes a three-nucleotide anticodon sequence that base pair with a complementary sequence on the mRNA known as a codon during protein synthesis (Figure 2A). An amino acid corresponding to the anticodon sequence, dictated by the genetic code, is attached to the tRNA. Upon base pairing between the anticodon and the correct codon on the mRNA at the A site of the ribosome, the peptidyltransferase reaction links the amino acid to the growing peptide (Moore and Steitz 2006).

![Figure 2](image-url)

Figure 2: Schematic cartoon depicting conserved secondary structures and sequence motifs of representative ncRNA families. (A) tRNA. Anticodon sequence is boxed in gray and base paired to the corresponding mRNA codon. (B) U1 snRNA. 5' SS interaction motif is boxed in gray. (C) C/D-box snoRNA. Characteristic boxes C and D are shown as well as optional boxes C' and D. A schematic interaction with a target RNA is indicated. The methylated nucleotide is located at the 5th base pair from the D box and is indicated by a star.

Spliceosomal Small Nuclear RNA (snRNA)

The spliceosomal snRNAs are only present in eukaryotes and are involved in the process of removing introns from pre-mRNA (splicing). Together with protein factors they assemble into RNP (snRNPs) and constitutes the core of the spliceosome. During the splicing process, multiple interactions involving base pairing between the snRNAs and sequence motifs on the pre-mRNA confer specificity to the reaction (Figure 2B) (Staley and Guthrie, 1998; Burge et al., 1999; Tycowski et al., 2006). It is frequently speculated that the catalytic steps involved in splicing are catalyzed by the snRNAs (Cech, 1986; Valadkhan, 2007).

Small nucleolar RNA (snoRNA)

SnoRNAs associate with proteins to form RNP complexes (snoRNPs). SnoRNPs are mainly involved in chemical modifications of a target RNA.
The RNA component interacts with the target RNA through base pairing and thereby brings the target nucleotide into the active site of the associated proteins which catalyze the modification (Figure 2C). Based on sequence motifs within the RNAs, the snoRNPs are divided into two classes. C/D-box snoRNPs perform 2'-O-methylation of nucleotides while H/ACA-box snoRNPs perform pseudouridylation of uridine nucleotides (Bachellerie et al., 1995; Kiss-Lašzló et al., 1996; Smith and Steitz, 1997; Tollervey and Kiss, 1997; Decatur and Fournier, 2003). The exact function of the modifications is unclear but they are likely to increase stability within the molecule. The main targets for snoRNPs are rRNA molecules. An snRNA molecule closely related to snoRNA is the small cajal body-specific RNA (scaRNA). ScaRNAs are structurally and functionally related to snoRNAs and assemble into scaRNPs. Many scaRNAs are a hybrid between the two classes and contain both C/D-box and H/ACA-box domains. ScaRNPs reside within cajal bodies in the nucleus where they modify primarily snRNAs and snoRNAs (Tycowski et al., 2006).

**Micro RNA (miRNA)**

MiRNAs were first observed in the nematode *Caernohabditis elegans* where they were found to regulate switching between developmental stages by specifically down-regulating expression of key proteins by a mechanism dependent on base pairing between miRNA and mRNA (Lee et al., 1993; Wightman et al., 1993; Reinhart et al., 2000; Lagos-Quintana et al., 2001; Lee and Ambros, 2001; Lau et al., 2001). MiRNAs are processed from longer hairpin precursors into ~21-24 nt single stranded RNAs that are incorporated into a larger protein complex known as RNA-induced silencing complex (RISC). The miRNAs are thought to have appeared somewhere close to the base of the metazoan lineage (Grimson et al., 2008). Interestingly, miRNAs are also present in plants although the structure, biogenesis and target interaction are substantially different from the animal counterparts. These differences together with the fact that no miRNAs have been detected in fungi or other intervening lineages have led to the hypothesis that the miRNA machineries in animals and plants have appeared independently (Jones-Rhoades et al., 2006).

**Other RNAs**

Aside from the few ncRNA families mentioned above, numerous other examples, many with regulatory functions, have been characterized (for review, see e.g. Costa, 2005; Hüttenhofer et al., 2005; Mattick and Makunin, 2006). These include the already mentioned RNase P RNA, involved in tRNA processing (Altman, 1990). The signal recognition particle (SRP) is a ribonucleoprotein which include the SRP RNA. It is involved in directing secreted or membrane bound proteins to the cellular membranes (Walter and Blobel, 1982; Egea et al., 2005). Xist RNA is a large mRNA-like transcript with no pro-
tein-coding potential that has been shown to initiate silencing of gene expression from the extra X chromosome in females (dosage compensation) by associating with the chromosome (Clemson et al., 1996). The piwi-interacting RNAs (piRNAs) are germ-line specific ~25-30 nucleotide RNAs that are involved in defense against transposons (Aravin et al., 2007). In addition, a large number of apparently species-specific (or limited to closely related species) ncRNA families have been identified (see e.g. Ruby et al., 2006; Pollard et al., 2006; Missal et al., 2006; Parrott and Mathews, 2007; Mourier et al., 2008 for a few examples).

Biogenesis

The ncRNA genes do not have a general biogenesis pathway. Which RNA polymerase that transcribes a ncRNA depends on the family, e.g. most rRNAs are transcribed by RNA polymerase I, most snRNAs are transcribed by RNA polymerase II and tRNAs are transcribed by RNA polymerase III. While e.g. spliceosomal snRNAs are transcribed as independent units, this is not true for all families of ncRNAs. Most rRNAs are transcribed as a continuous precursor transcript that is cleaved and processed post-transcriptionally (Fatica and Tollervey, 2002). Many snoRNAs and some miRNAs are encoded within introns of protein-coding genes and are co-transcribed together with the host gene and processed out of the spliced introns by exonucleases (Kiss and Filipowicz, 1995; Kim, 2005). The Xist RNA is transcribed, spliced and polyadenylated just like a normal protein-coding mRNA, yet it lacks an open reading frame (ORF) and is never translated (Clemson et al., 1996).

Structure

Many ncRNAs adopt an evolutionarily conserved secondary structure, indicating that the structure is of critical importance for the function of the ncRNA. Most ncRNAs associate with RNA-binding proteins to form active RNP particles. The recognition of the RNA by proteins is often dependent on the structure or a combination of structure and short sequence motifs (Tycowski et al. 2006, Cech et al., 2006). For this reason, unlike protein-coding genes which are often under negative selection to preserve the reading frame of ORFs, amino-acid sequence and three-dimensional fold of the protein product, ncRNAs are free to diverge as long as relevant sequence motifs and secondary structures are preserved, possibly through compensatory base pair changes (Figure 1D). Thus, the nucleotide sequence of ncRNA genes can differ substantially even between relatively closely related organisms (Eddy, 2002a).
Function

The high specificity with which an RNA molecule can interact with other RNA molecules via base-pairing make them ideally suited for functions requiring specific recognition of a particular target sequence. This opens up the possibility of having a generic protein machinery with some function that associates with an interchangeable RNA molecule which guide the machinery to a target RNA with high specificity. Most ncRNA genes, e.g. snoRNAs, miRNAs and piRNAs incorporate into RNP's and function exactly like this (Tycowski et al., 2006; Petersen et al., 2006; Aravin et al., 2007). This brings an enormous flexibility to an organism as new specificities can rapidly evolve without having to alter the protein machinery. A genomic duplication of the ncRNA gene followed by divergence of the sequence bringing specificity would be sufficient (Eddy, 2001).

Protein-coding genes

The DNA in the nucleus contains the information needed for making all the proteins an organism requires. When a specific protein product is required by the cell, the piece of DNA encoding the instructions for making it is transcribed into an RNA copy, known as messenger RNA (mRNA). In eukaryotes, DNA encoding protein-coding genes are transcribed by RNA polymerase II (reviewed in e.g. Roeder, 2005; Kornberg, 2007). Beside the splicing process, reviewed below, the mRNA is modified at the 5' and 3' ends before export into the cytoplasm and translation. The unprocessed mRNA molecule is commonly referred to as pre-mRNA. The mRNA is “capped” at the 5' end by the addition of a reversed guanosine nucleotide by a special capping enzyme. The cap protects the mRNA from degradation from the 5' end but it is also important for splicing, transport and translation. At the 3' end, the pre-mRNA is cleaved at the poly(A)-site by a protein complex that recognizes a number of cis-encoded elements within the mRNA. Subsequently, another protein complex, the poly(A)-polymerase, adds a number of adenosine residues, typically 50-250, to the end of the pre-mRNA. This is known as polyadenylation and the added poly(A)-tail is important for the transport, translation and stability of the mRNA.

The amino-acid sequence of the protein is encoded in the open reading frame of the mRNA in nucleotide triplets known as codons. The genetic code dictates which codon corresponds to which amino acid. Since there are $4^3 = 64$ possible codons but only 20 amino acids, several different codons can encode the same amino acid. Different codons that encode the same amino acid are said to be synonymous and conversely, codons that encode different amino acids are said to be non-synonymous. With very few exceptions, the genetic code is universal and shared between all living organisms. A number
of codons have special meanings: AUG encodes the amino acid methionine as well as being the start codon that indicates where translation should start. Less frequently, CUG, GUG or UUG are used as start codons. UAA, UAG and UGA do not encode amino acids but are stop codons that will cause translation to terminate when encountered. Thus, these special codons impose restrictions on the ORF, i.e. the ORF begins with a start codon and no stop codons are allowed within the ORF since that would cause a premature termination of translation to occur. In addition, because the codons are read in triplets in a fixed so-called reading frame, the length of the ORF must be a multiple of three and the reading frame must not be shifted. If e.g. a mutation in the DNA or a mistake in the transcription into RNA that inserts or deletes a single nucleotide would occur, a frameshift could result which would cause the ORF to be misread and the wrong amino acids to be incorporated into the polypeptide.

Unlike prokaryotes where the ORF is encoded as a continuous nucleotide sequence in the DNA, eukaryotic ORFs are often interrupted by stretches of nucleotides that should not be translated. These non-protein coding sequences are known as introns. The introns have to be excised from the transcript and the coding parts, the exons, have to be put together with high precision as the ORF must not be altered or frameshifted. This process is known as splicing.

**Splicing**

In the splicing process, the introns are removed from the pre-mRNA and the coding pieces, the exons, are joined. This occurs through a two-step transesterification reaction (Figure 3). In the first step, the 2'-hydroxyl of the adenosine residue at the branch point (BP), located close to the 3' end of the intron, attacks and breaks the phosphodiester bond at the 5' splice site (5' SS), between the last nucleotide of the exon and the first nucleotide of the intron. This results in the transesterification of the 5' end of the intron to the BP adenosine, creating a lariat intermediate and a free 3'-hydroxyl at the 5' exon. In the second step, this 3'-hydroxyl attacks and breaks the phosphodiester bond at the 3' splice site (3' SS), between the last nucleotide of the intron and the first nucleotide of the downstream exon, followed by ligation of the 5' exon to the 3' exon through a transesterification reaction and release of the lariat intron (Burge et al., 1999). The nucleotide sequences around the 5' SS, 3' SS usually matches a well-defined consensus sequence motif (Mount, 1982). Although the exact sequence is somewhat variable, the two first nucleotides of the intron are always GU, with very few exceptions. Similarly, the last two nucleotides of the intron are almost always AG.
The splicing reaction is catalyzed and coordinated by the spliceosome, a large RNP complex composed of the five snRNAs U1, U2, U4, U5 and U6 together with more than 200 protein factors (Jurica and Moore, 2003; Will and Lührmann, 2006). The snRNAs are assembled into ribonucleoprotein particles (snRNPs) that are essential for the splicing reaction (Yang et al., 1981; Padgett et al., 1983). The formation of the spliceosome on the pre-mRNA is directed by the interaction of the U1 snRNP with the 5' SS through base pairing to the U1 snRNA and U2 snRNP binding to the branch point through interaction with protein factors and base pairing between the U2 snRNA and the BP sequence (Burge et al., 1999). The two-step transesterification reaction of the splicing process is identical to the reaction by which the group II self-splicing introns splice themselves. In addition, RNA-RNA interactions within the spliceosome closely mimics base pairing interactions within group II introns (Tycowski et al., 2006) and it has been speculated that the spliceosome is in fact a ribozyme (Cech, 1986; Valadkhan et al., 2007; Valadkhan, 2007).

In the shadow of this dominant spliceosome, known as the major spliceosome or U2-dependent spliceosome, there also exists a second spliceosome, known as the minor spliceosome or U12-dependent spliceosome, which catalyzes the removal of a distinct set of introns (Jackson, 1991; Hall and Padgett, 1996; Tarn and Steitz, 1996). The minor spliceosome is present in most
but not all eukaryotic organisms. Although the mechanisms of the splicing reaction and a great number of protein factors are similar between the two spliceosomes, the minor spliceosome contain a different set of snRNAs known as U11, U12, U4atac and U6atac. The U5 snRNA is shared between the spliceosomes (Patel and Steitz, 2003; Will and Lührmann, 2005). The introns spliced by the major and minor spliceosomes are referred to as U2-type and U12-type introns, respectively. It is estimated that approximately one in 300 introns is spliced by the minor spliceosome (Levine and Durbin, 2001).

### Alternative splicing

Through alternative splicing (AS), the same pre-mRNA can be spliced differently in a modular fashion, e.g. introns can be left in the pre-mRNA (intron retention), exons can be spliced out (exon skipping) or alternative 5' and 3' splice sites can be utilized (Figure 4). This way, one transcription unit in the genome can give rise to a large number of different protein products. The most extreme example known is the Dscam gene in *Drosophila melanogaster* (fruit fly) which is involved in neuronal wiring in the nervous system. AS of four of the exons in the gene potentially gives rise to more than 36,000 different protein products (Schmucker et al., 2000, Schmucker, 2007).

![Alternative splicing patterns](image)

**Figure 4: Alternative splicing patterns.** (A) Exon skipping. (B) Alternative 5' splice site or 3' splice site. (C) Mutually exclusive exons. (D) Intron retention.

### Regulation of alternative splicing

Observations of tissue-specific and developmentally regulated alternative splicing events indicated that *trans*-acting factors were involved in regulating AS. It was hypothesized that the snRNPs could be such *trans*-acting factors and that variant snRNPs could produce different splicing products (Las-ki et al., 1986; Padgett, 1986). This hypothesis was supported by reports of developmentally regulated and heterogeneous snRNPs in *Xenopus* oocytes, mouse, sea urchin and drosophila (Forbes et al., 1984; Lund et al., 1985; Santiago and Marzluff, 1989; Lo and Mount, 1990). Later discoveries of *cis*-acting elements, so-called splicing regulatory elements (SRE), which have
been shown to be major regulators of splicing by recruiting various members of the SR protein family, have provided a robust and comprehensive mechanism for AS regulation (Graveley, 2000; Black, 2003; Matlin et al., 2005; Wang and Burge, 2008). However, it is also known that the interaction between the U1 snRNA and the 5' SS influence splicing efficiency and that changes in the strength of the interaction can shift splicing of exons towards constitutive or alternative (Zhuang and Weiner, 1986; Zhuang et al., 1987; Stamm et al., 1994; Sorek et al., 2004). A study of introns with non-canonical 5' SS having a GC-dinucleotide instead of a GU-dinucleotide at the first positions of the intron, found that a large fraction of introns with a non-canonical 5' SS are alternatively spliced and that non-canonical 5'SS are overrepresented among alternatively spliced introns (Thanaraj and Clark, 2001). Thus, regulation of AS seems to depend on a delicate interplay between cis-acting elements in the pre-mRNA and their associated protein factors and snRNP complexes interacting with sequence motifs on the pre-mRNA.

Alternative splicing and the proteome

It has been clear for a long time that the morphological complexity of an organism, measured e.g. as the number of distinct cell types in the body, is not correlated with the size of its genome. For example, the unicellular protist Amoeba dubia has a genome of approximately 670 Gbp which is more than 200 times larger than the human genome (Gregory and Hebert, 1999). Neither does the number of protein-coding genes seem to correlate with the organismal complexity. For example, the nematode Caernohabditis elegans consists of roughly 1000 cells distributed over 24 distinct cell types (Altun, Z. F. and Hall, D. H. 2005. Handbook of C. elegans Anatomy. In WormAtlas. http://www.wormatlas.org/handbook/contents.htm). In contrast, the number of cells in the human body is in the order of \(10^{14}\) and the number of distinct cell types in mammals are around 100 (Bell and Mooers, 1997). In the July 2008 release of Ensembl (Release 50), the number of known and novel human protein-coding genes annotated is 21,528 (Flicek et al., 2008; http://www.ensembl.org). In the same release the number of protein coding genes in the nematode is 20,176. Thus, despite the vast differences in number of cells and cell types between nematodes and humans, we have about the same number of protein-coding genes.

Alternative splicing has been suggested as an explanation for this paradox since AS is a powerful mechanism by which the protein repertoire of an organism can be expanded beyond the number of transcription units encoded within the genome (Black, 2003; Blencowe, 2006). A study using exon junction microarrays estimate that some 74% of human multi-exon genes are alternatively spliced (Johnson et al., 2003). Even more astonishing, a recent study using massive parallel sequencing conclude that 98-100% of human
multi-exon genes undergo AS events, most of which seem to be regulated between tissues, lending support to the idea that AS is an important contributor to phenotypic complexity (Wang et al., 2008a).

Splicing and disease
Among known disease-causing mutations, it is estimated that 20-30% affect pre-mRNA splicing, causing either aberrant splicing or shifting ratios between transcript isoforms. The severity of the disease phenotype depends on the level of correctly spliced transcripts or the ratio between isoforms (see Nissim-Rafinia and Kerem, 2002; Faustino and Cooper, 2003; Wang and Cooper, 2007 for review). Approximately 10-15% of these mutations disrupt intronic splicing motifs e.g. splice sites, SREs or creating decoy splice sites (Krawczak et al., 1992; Faustino and Cooper, 2003).

Post-genomic era
In 1995 the first complete genome sequence of a free-living organism, Haemophilus influenzae, was published (Fleischmann et al., 1995). Since then, the number of completely sequenced genomes has increased exponentially while the cost and time required to sequence have decreased steadily. In October 2008, the number of genome sequences in NCBI's GenBank that are completed or with an available draft assembly was close to 1500 and roughly another 1000 genome sequencing projects are in progress. Approximately 85% represent prokaryotic organisms (http://www.ncbi.nlm.nih.gov/). This calls for large-scale methods capable of more or less automatically analyzing and handling the exponentially growing amount of genome data.

Genome evolution
A genome is not static but is constantly changing due to mutations. Mutations provide the raw material for the evolutionary forces of natural selection and genetic drift to work on. Mutations can arise through e.g. mistakes in the cellular processes of DNA replication or exposure to mutagenic substances or radiation. Mutations can be e.g. nucleotide substitutions where one nucleotide is changed into another or nucleotides can be inserted or deleted. Other types of mutations include larger structural rearrangement events that can be due to e.g. the action of transposons, non-homologous recombination during crossing-over or chromosomal breakage during replication. While the vast majority of mutations are neutral, i.e. they do not have any significant effect on the phenotype of an individual, a small fraction is likely to be
harmful and a small fraction can be beneficial. For example, insertions or deletions in an ORF most likely cause frameshifts and lead to an erroneous protein being synthesized, which is probably harmful to the organism. Conversely, a nucleotide substitution that changes a codon into another synonymous codon would most likely have no effect on the organism and therefore be completely neutral.

If a genomic mutation occurs in the germ-line, the mutation can be inherited by the offspring and passed on to subsequent generations. While severely harmful mutations are unlikely to spread through the population since offspring carrying them are likely to die or otherwise be prevented from breeding, neutral mutations can spread and give rise to genetic variation within a population. Recent genome sequencing projects have shown that the genomes of different human individuals differ from each other by roughly 0.5% (Levy et al., 2007; Wheeler et al., 2008; Wang et al., 2008b).

Comparative genomics

Since all living organisms are related to each other through common descent and since their genomes are shaped by evolutionary forces; comparing the footprints these forces leave in the genomes reveal crucial information. If the corresponding regions in two genomes are significantly well conserved over large evolutionary distances, that suggests selective constraints to keep the regions unchanged which would indicate functionality. Such comparative approaches can also identify elements potentially involved in making a species unique. As an example, comparing the human and mouse genomes might help in identifying genes and regulatory motifs that are of such importance to the organism that they haven't changed substantially in the approximately 80 million years since humans and mice last shared a common ancestor. On the other hand, human and chimpanzee are so closely related that we expect very little to have changed between the genomes (roughly 98.8% of aligned nucleotides are identical) (Mikkelsen et al., 2005). Therefore, it is in the parts of the genomes that are different we would expect to find the genetic variations that set these species apart.

On an even finer scale, having the complete genomes from several individuals of the same species allows for even more detailed analyses such as mapping the genetic cause for various heritable diseases. Recently, a large-scale sequencing initiative, The 1000 genomes project, has been undertaken and the goal is to compile a catalog of human genetic variation that is present at a frequency of one percent or more. The complete genome sequences from at least 1000 individuals will be determined and made publicly available (http://www.1000genomes.org).
Identification of functional units

Having knowledge of the complete genome sequence is extremely valuable when trying to understand how an organism functions since the genome essentially contains all information that defines the organism physically. However, just knowing the genome sequence, *i.e.* the internal order of DNA nucleotides, is not very useful by itself. The next great challenge is to interpret and decode the genome to create a map of all functional units, *i.e.* the genes and the regulatory elements that control their expression. This can be achieved through a wide array of complementary methods which can be separated into experimentally based and computationally based methods. I will mainly focus on identification of RNA genes and to some extent on protein-coding genes.

In general, experimental methods have the advantage of being capable of *de novo*-identification while not suffering so much from false positives (with the exception of microarray technologies). On the other hand, experimental approaches are often expensive, work intensive and may suffer from large amounts of false negatives depending on growth conditions, expression levels etc. Bioinformatic approaches generally have the advantage of being cheap, often scales well with increasing data amounts and are independent of expression levels. However, bioinformatic methods always have to struggle with a trade-off between false positives and false negatives and they are mostly limited to discovering already known ncRNA families. Typically, genes that have been identified computationally have to be verified experimentally in order to be fully trusted.

Experimentally based methods

**Complementary DNA library preparation**

Preparation of complementary DNA (cDNA) libraries by reverse transcription of mature mRNAs extracted from cells have been extensively used to map protein-coding genes. In this process, an oligo-dT primer complementary to the poly(A)-tail of the mature mRNAs, or alternatively a random hexanucleotide primer, is used as a primer for reverse transcriptase enzyme which transcribes the mRNA into cDNA. The cDNA can then be sequenced from the 5' or 3' end, generating transcripts, so-called expressed sequence tags (ESTs), 100-800 nucleotides in length. Since the cDNA is constructed from mature mRNAs, introns will be absent from the resulting sequence and a subsequent mapping back to the genome (if available) will reveal valuable information on the exon-intron structure of the gene. Since this method discriminates against less abundant and non-polyadenylated RNAs, ncRNAs will only very rarely be incorporated into these cDNA libraries.
In 2001, Hüttenhofer and colleagues focused on large scale identification and mapping of ncRNAs by constructing cDNA libraries specifically enriched for ncRNAs (Hüttenhofer et al., 2001). In their approach, total RNA was first size fractionated by electrophoresis and only RNA in the size range 50-500 nucleotides, where the fraction of mRNAs are expected to be very low, was extracted. Subsequently, a poly(C)-tail was added and a cDNA library could then be constructed by reverse transcription using an oligo-dG primer. Subsequent refinements of the technique entails ligating known adapter sequences to the 5’ and 3’ ends of the RNAs in order to generate full-length cDNAs. RNA can also be isolated by performing immunoprecipitation of RNP complexes followed by RNA extraction and adapter ligation.

A drawback with this method is that absence or underrepresentation of certain ncRNA families in the constructed cDNA library may result from low abundance or interference with adapter ligation and reverse transcription into cDNA due to secondary structures or chemical modifications of nucleotides at the ends of RNAs. In addition, unless specifically eliminated, the majority of cDNAs in the library are expected to represent highly abundant RNAs (Hüttenhofer and Vogel, 2006).

**Microarray analysis**

Microarray techniques usually involve spotting 25-70 nucleotide long DNA probes on the silicon or glass surface of microarrays. RNA is then extracted from the cell and either hybridized to the microarray probes directly or chemically labeled and possibly reverse transcribed into cDNA before hybridization. The hybridization patterns can then be visualized through the labeling or if unlabeled RNA was used, through antibodies recognizing the DNA-RNA interaction. A microarray can contain a large number of probes and the design varies depending on the application. Arrays with probes interrogating known or predicted exon junctions have been used to assess alternative splicing events (Johnson et al., 2003). Arrays with probes interrogating known or predicted exons also allow AS events or variations in expression levels to be assessed (Kapur et al., 2007; Clark et al., 2007). Yet another type of array is known as tiling array and consists of overlapping nucleotide probes that span entire chromosomes or even genomes. Here, no a priori knowledge of exons or transcripts is required. Tiling microarrays have revealed a surprising amount of transcription from regions of the human genome assumed to be transcriptionally silent (Bertone et al., 2004; Kampa et al., 2004; Birney et al., 2007).

In the study of ncRNAs, microarrays are usually targeted at verifying or comparing expression levels of transcripts from known or computationally predicted ncRNAs (see e.g. Babak et al., 2004; Hu et al., 2006). However, there are many factors influencing and complicating analysis of the microarray data such as probe specific and array specific hybridization efficiency and varying levels of background noise. Subsequent Northern blot verifica-
tion of transcripts detected on microarray have also failed in many instances so ncRNAs identified on microarray must in many cases be further verified by downstream analyses (Hüttenhofer and Vogel, 2006).

**Direct sequencing**

Recently, powerful high-throughput sequencing platforms such as Roche 454, Illumina Genome Analyzer and Applied Biosystems SOLiD sequencer, capable of producing millions of sequence reads in a short amount of time and at a low cost have emerged (for a recent review, see Mardis, 2008). These techniques are extremely promising and offers the possibility to create a snapshot view of the transcriptome content of a sample. For example, Wang and colleagues performed deep sequencing of the transcriptome of a wide array of human tissues and cell lines and reported that approximately 92-94% of all human mRNAs are alternatively spliced (Wang et al., 2008a). In another study, Lu et al. performed deep sequencing of small RNAs extracted from the plant *Arabidopsis thaliana* and were able to identify a vast repertoire of small RNAs including previously known and unknown miRNAs and siRNAs (Lu et al., 2005). Numerous studies have followed their approach to identify small RNAs (see e.g. Ruby et al., 2006; Landgraf et al., 2007; Fahlgren et al., 2007; Grimson et al., 2008). Similar to the construction of full-length cDNA libraries, these methods rely on an initial preparation step where adapter sequences are ligated to the ends of the RNA molecules. Hence the same bias in representation may arise.

**Computationally based methods**

When describing computationally based methods for gene discovery and annotation, it is convenient to distinguish between situations where we are searching for a particular family of genes where known examples exist and situations where we are searching for genes *de novo*, *i.e.* where we are trying to discover new gene families.

**Searching for known gene families**

The most commonly used method for annotation of functional units in a sequenced genome is probably using an already known sequence from another species to do a sequence similarity search. This is primarily useful for protein-coding genes which are generally under negative selection to preserve amino acid sequence and hence change relatively slowly during evolution. In particular, the amino acid sequence of a known protein can be used to search a genome by translating the genome sequence into amino acids in all six potential ORFs, thereby increasing the sensitivity compared to just doing nucleotide comparisons. The Basic Local Alignment Search Tool (**BLAST**) program suite is widely used for this purpose (Altschul et al., 1990; Altschul
et al., 1997). Other useful tools include the BLAST derivative Washington University BLAST (WU-BLAST) (Gish, W. (1996-2004) http://blast.wustl.edu), The BLAST-like Alignment Tool (BLAT) (Kent, 2002) and FASTA (Pearson and Lipman, 1988; Pearson et al., 1997). For ncRNA genes, searching using nucleotide sequence similarity is not as reliable since many ncRNAs are very much dependent on their structure rather than sequence. Therefore, unless the species with which the comparison is made is closely related, there is a great risk that the nucleotide sequences of homologous genes have diverged substantially (Eddy, 2002a).

A more promising approach to search for homologous ncRNA genes would be to search a database for a similar secondary structure. RScan is one program that implements such an approach using only structural information to search a database (Xue and Liu, 2007). Klein and Eddy have developed the RSEARCH software that is capable of searching a database using both sequence and structure information (Klein and Eddy, 2003). Another option is to manually supply a sequence and structure pattern which can be searched for using e.g. PatSearch (Grillo et al., 2003), HyPa (Gräf et al., 2001; http://www.biophys.uni-duesseldorf.de/hypa) or RNAMotif (Macke et al., 2001).

In cases where homologous or paralogous sequences are known and a reliable multiple alignment is available or can be constructed, the Infernal (Eddy, 2002b) software package can be used to create a so-called covariance model (CM) from the multiple alignment. A CM is a probabilistic model that represents both structural information, inferred from correlation between columns in the multiple alignment, and sequence information (Eddy and Durbin, 1994). The constructed CM can then be used by Infernal to search an input sequence for occurrences of subsequences that receives significant scores according to the CM. The database Rfam is a large-scale effort to collect all available knowledge on a large number of ncRNA families in a community-driven approach. This includes manually curated multiple alignments and pre-computed CMs of ncRNA families (Griffiths-Jones et al., 2005; http://rfam.sanger.ac.uk/). Both Infernal and RSEARCH have been evaluated and shown to perform very well on well-defined ncRNA families but suffer from being computationally expensive (Freyhult et al., 2007).

For many known families of ncRNAs, specially tailored software capable of predicting genes belonging to the family exist. A small selection of examples include tRNAscan-SE (Lowe and Eddy, 1997) and ARAGORN (Laslett and Canbäck, 2004) for tRNAs, snoscan (Lowe and Eddy, 1999), snoGPS (Schattner et al., 2004) and Fisher (Edvardsson et al., 2003) for snoRNAs, MiRscan (Lim et al., 2003), miR-abela (Sewer et al., 2005) and RNAmicro (Hertel and Stadler, 2006) for miRNAs and SRPscan (Regalia et al., 2002) for SRP RNA. Each of these programs is dedicated to the identification of a single family of ncRNAs, often using a combination of sequence and structure motifs known to be crucial for the function of the RNA.
De novo gene prediction

In contrast to the situation above where a well-known family of genes or where homologs to a specific gene are searched for, de novo gene prediction means that we want to identify genes without any specific prior knowledge on sequence or structure features. For protein-coding genes, a number of features, such as start codon, open reading frame, splice signals etc., must be present in order for the gene to function properly in a cellular context. Taken together, these features can be combined to yield de novo protein-coding gene predictions of high accuracy, even though things like e.g. alternative splicing and short ORFs complicate the exact prediction of protein products. A corresponding set of general features for ncRNA genes is as far as we know not available. The approach will also depend on whether the genome sequences of any closely related species are available. In that case it might be possible to do a comparative study.

Single genome search

Without genome sequences from closely related species, the problem of de novo ncRNA gene prediction becomes even harder. We have already seen that secondary structure is important for the function of many ncRNAs so one approach could be to search for locally stable secondary structures within a genomic sequence. RNALfold (Hofacker et al., 2004), RNAplfold (Bernhart et al., 2006) and Rfold (Kiryu et al., 2008) are examples of programs with this capability. However, a number of studies have demonstrated that although secondary structure stability of ncRNAs is statistically significant compared to random nucleotide sequences for certain highly structured ncRNA families, it is not generally true for ncRNA genes (Workman and Krogh, 1999; Rivas and Eddy, 2000; Clote et al., 2005). Hence, this approach is biased towards ncRNAs with stable secondary structures and would probably generate many false positives so further analyses of the predictions would be necessary.

Another possible approach would be to focus on the genomic signals important for transcription and processing of a ncRNA gene. Such approaches have had success in prokaryotes where the promoter and transcription terminator signals are well characterized (Argaman et al., 2001; Chen et al., 2002; Yachie et al., 2006). However, in eukaryotes this approach is complicated by the fact that transcription of ncRNAs can be carried out by any of the polymerases and their promoter structures are poorly understood. In addition, ncRNA genes come in a variety of genomic structures, they are not necessarily independently transcribed. An attempt to correlate human structured ncRNA gene predictions made with the RNAz software (see below) with transcription factor binding sites could not detect any difference between structured ncRNA gene predictions and shuffled background (Backofen et al., 2007).
Yet another option suggested is if there exists a bias in nucleotide composition between the genes encoding ncRNAs and the rest of the genome, regions with a nucleotide composition similar to typical ncRNA genes can be identified (Rivas and Eddy, 2000). This approach is obviously extremely limited to the special cases where such a bias exists. Klein and co-workers as well as Schattner both successfully conducted studies identifying ncRNA genes in hyperthermophiles with extremely AT-rich genomes using this approach independently (Schattner, 2002; Klein et al., 2002).

Comparative study
In case the genome sequence from one or more closely related species is available, the conditions for de novo ncRNA gene prediction are somewhat improved. The basic idea is then that orthologous regions between the genomes are aligned and the substitution pattern can be evaluated to identify regions where nucleotides are substituted in a way that is expected for a ncRNA gene. Since secondary structure is often important to the function of ncRNAs whereas nucleotide sequence is not, it would be expected that evolution preserves the secondary structure through compensatory nucleotide substitutions in base paired regions (Figure 1D). In contrast, for proteins which are very much dependent on their amino acid sequence, the protein-coding genes are expected to have a substitution pattern with mostly synonymous nucleotide changes and no insertions or deletions that would cause frameshift mutations. In order for such an analysis to be successful, the aligned genomes can not be too closely related since there might be too few nucleotide substitutions. On the other hand, they can not be too distantly related, they must be sufficiently close so that a reliable alignment where each column represents an orthologous position can be constructed. Rivas and colleagues pioneered this approach in the software QRNA and demonstrated that they could detect bacterial ncRNAs with high sensitivity and specificity (Rivas and Eddy, 2001; Rivas et al., 2001). Evofold is a similar software capable of considering multiple sequence alignments and taking the underlying phylogenetic relationships into account (Pedersen et al., 2006). Like QRNA, it is based on secondary structure predictions using a probabilistic framework. RNAz also considers multiple sequence alignments and uses thermodynamic parameters for secondary structure prediction (Washietl et al., 2005a). Both Evofold and RNAz have been used to generate a large amount of predicted elements with evolutionarily conserved stable secondary structures, many of which correspond to ncRNA genes (Pedersen et al., 2006; Washietl et al., 2005b).
Present Investigations

To understand the processes and mechanisms governing biological systems, it is crucial to have a comprehensive knowledge about the key elements involved. Non-coding RNAs (ncRNAs) have emerged as highly abundant and diverse factors in e.g. cellular processes and regulation of gene expression. Thus, to determine the ncRNA repertoire is of critical importance to our understanding of biological systems. The aim of my thesis project has been to develop and use computational tools for the purpose of identifying and characterizing ncRNA genes.

Non-coding RNAs in *Dictyostelium discoideum* (Papers I-III)

*Dictyostelium discoideum* is a slime mold belonging to the amoebozoa lineage that branched out early in eukaryotic evolution, after the plant-animal split but before the fungi-animal split (Baldauf and Doolittle, 1997; Bapteste et al., 2002). *D. discoideum* lives as single cells feeding on bacteria. Upon starvation, up to $10^5$ cells come together and differentiate into a multicellular-like organism. *D. discoideum* has for over 50 years been a model organism to study different cellular processes, e.g. cell signaling, multicellular development and host-pathogen interactions. The genome sequence has been determined and was published in 2005 (Eichinger et al., 2005). The genome which is extremely rich in A and T nucleotides, 78% on average, is approximately 34 Mbp in size, organized into six chromosomes and contains roughly 12,500 protein-coding genes. Approximately 10% of the genome consists of interspersed complex repeats (Glöckner et al., 2001). At the start of this study, despite being a well-studied model organism, very little was known about the ncRNA repertoire of *D. discoideum* beside the rRNAs, tRNAs and a few more RNAs (for a review, see Hinas and Söderbom, 2007). We have used a number of experimentally and computationally based techniques to identify and characterize many new ncRNAs in *D. discoideum*. 
Experimentally based methods (Paper I)

In order to get a picture of the ncRNA repertoire of *D. discoideum*, we constructed full-length cDNA libraries of small RNAs (50-500 nt) from *D. discoideum*. Total RNA from cells that had been developing for 16h (slug stage) was size fractionated followed by C-tail addition. To ensure that the cDNAs would represent full-length RNAs of both processed and primary transcripts, an RNA oligo was ligated to the 5' end of the RNAs after treatment with TAP. The RNA was subsequently amplified using RT-PCR and cloned. Before sequencing, colonies were screened to minimize the number of clones corresponding to rRNA sequences. Of the remaining clones, 36 unique sequences were found to represent previously unknown *D. discoideum* ncRNAs. Among the sequenced small RNAs, 18 snoRNAs, 1 U2 snRNA and 1 SRP RNA could be identified. In addition and most interestingly, two highly abundant classes (~14% of the cloned RNAs) of a previously unknown small ncRNA family were isolated and named Class I and Class II RNAs. We note the absence of some of the abundant ncRNA families, e.g. tRNAs and the snRNAs except for U2. This can most likely be attributed to the ligation of the RNA oligo to the 5' end of RNAs as similar studies that perform reverse transcription without the ligation step frequently detect these ncRNAs (Hüttenhofer et al., 2001; Yuan et al., 2003). The ligation step is known to be inefficient and higher-order structures in the RNA as well as chemical modifications near the ends can interfere with ligation and reverse transcription (Hüttenhofer and Vogel, 2006). The cDNA library approach enabled a first glimpse of the ncRNA repertoire of *D. discoideum* and allowed us to identify a highly abundant and, to our knowledge, species specific family of ncRNAs. One should bear in mind however, that the method failed to pick up certain known families of ncRNAs where ligation of oligos or reverse transcription into cDNA are problematic due to extensive structure.

Computationally based methods (Papers II-III)

**Known ncRNA families (Paper II)**

We were confident that *D. discoideum* contained the snRNA components of the splicing machinery since early work had demonstrated the presence of spliceosomal snRNAs on polyacrylamide gels (Wise and Weiner, 1981; Takeishi and Kaneda, 1981). In Paper I, we managed to identify the U2 snRNA. However, we failed to detect any of the other snRNAs associated with the spliceosome. We therefore set out to identify the spliceosomal snRNAs in *D. discoideum* by computational means.

The U5 and U6 snRNAs could be readily identified through a sensitive sequence similarity search with the BLAST software using known U5 and U6 snRNAs as query. However, this approach failed to detect the U1 and U4 snRNAs. We therefore refined the search based on existing knowledge con-
cerning the interactions between the snRNA and pre-mRNA as well as between the snRNAs in the spliceosome. This was combined with knowledge about the evolutionarily conserved secondary structures known to be required for the spliceosome function (Madhani and Guthrie, 1994; Tycowski et al., 2006). This led to the identification of five genomic loci corresponding to U1 snRNA and three loci corresponding to U4 snRNA. In addition, we located a total of seven loci encoding U2 snRNA-like genes by sequence similarity searches using the U2 snRNA identified in Paper I. We did not manage to identify any sequences corresponding to snRNAs belonging to the minor spliceosome, indicating that the minor spliceosome machinery may have been lost in *D. discoideum*. This observation is also supported in a recent study (Dávila López et al., 2008).

The computationally predicted snRNA genes (18 in total) were extensively verified experimentally and expression from 17 loci was detected. This approach illustrates a very cost-efficient and powerful method for locating ncRNAs belonging to known and well-defined families where extensive knowledge of sequence and structure features can be utilized.

**De novo ncRNA gene prediction (Paper III)**

The identification of previously unknown ncRNA families potentially specific to *D. discoideum* in Paper I prompted us to attempt a computational effort to identify ncRNAs *de novo*. Presently, successful methods for *de novo* ncRNA gene prediction have been limited to studies based on comparative genomics or transcriptional signals (e.g. Rivas et al., 2001; Pedersen et al., 2006; Washietl et al., 2005b; Argaman et al., 2001; Chen et al., 2002; Yachie et al., 2006). However, at the time of the study, no genome sequence was available for any sufficiently close *D. discoideum* relative. In combination with the variability in ncRNA gene structure and eukaryotic transcription signals being poorly characterized, these approaches seemed inapplicable. Our characterization of the isolated ncRNAs in *D. discoideum* showed that most genes are situated in between protein-coding genes where the A and T content is very high. However, the ncRNA genes were not particularly skewed towards A and T nucleotides. This fact opened up the possibility for *de novo* ncRNA gene prediction using base composition as a criterion. This approach has previously been successfully used to identify ncRNA genes in hyperthermophiles with extremely AT-rich genomes (Klein et al., 2002; Schattner, 2002).

While the studies by Klein and colleagues and Schattner used a hidden Markov model (HMM) framework and a sliding window approach, respectively, we chose to implement a method based on high scoring segments among partial sums. With this approach, each nucleotide in a sequence is given a score value. Typically, nucleotides that frequently occur in the type of sequences we are looking for, in this case ncRNAs, will receive a positive
score while nucleotides that are not frequent are given a negative score. In this particular case, G and C nucleotides that are unusual in the background genome but relatively frequent in ncRNAs will receive a positive score while A and T nucleotides, which are common in the background genome, will receive a negative score. By summing up scores from consecutive nucleotides we can define a high scoring segment so that the cumulative sum between the start and end point of the segment is maximized. Such high scoring segments would then represent a sequence with a base composition more similar to the known ncRNAs than to the background genome. This method has several advantages. Firstly, the only parameters needed are the scoring scheme used to assign a score value to each nucleotide. These scores can be designed arbitrarily to detect specific features of interest or be derived from e.g. a training set of positive examples and a background set representing negative data. Secondly, there is a solid body of theory developed regarding the statistical distribution of scores of maximal scoring segments which enables the statistical significance of maximal scoring segments to be evaluated analytically (Karlin and Altschul, 1990). In addition, when considering secondary structures of ncRNAs, it is well known that interactions between neighboring nucleotides are important, e.g. stacking interactions between base pairs stabilize helices (Mathews et al., 1999; Workman and Krogh, 1999; Clote et al., 2005). For that reason we implemented a scoring scheme assuming a first order Markov dependence between nucleotides which we refer to as the M1-model with the intention to evaluate any added benefit compared to a scoring scheme considering nucleotides to be independent, the M0-model (Figure 5). The theoretical results regarding the statistical distribution of maximal scoring segments have been extended to sequences with a Markov dependence (Karlin and Dembo, 1992).

As practically none of the known ncRNAs overlap annotated exons or interspersed repeats, we masked those sequences from the D. discoideum genome. We refer to the resulting data set as the non-repetitive, intergenic genome. We used a scaled log ratio between the nucleotide frequency from a “target” set and a “background” set as score for each corresponding nucleotide. The target set consisted of all known ncRNAs filtered for redundancy and the background set consisted of the non-repetitive, intergenic genome with known ncRNAs removed. For the M0-model we scored each nucleotide independently using scores derived from the respective nucleotide frequency. For the M1-model, we scored each overlapping dinucleotide using scores derived from the conditional frequency of the second nucleotide given the first nucleotide. Evaluating the benefit of using the M1-model over the M0-model, we found a slight increase in sensitivity and specificity when using the M1-model. However, the largest part of this increase is most likely due to a characteristic dinucleotide content in the background genome of D. discoideum rather than in the ncRNAs.
We searched the non-repetitive, intergenic genome of *D. discoideum* using the M1-model and identified 94% of previously known ncRNAs as well as a large number of high-scoring segments possibly corresponding to previously unknown ncRNA genes. In order to further refine the predictions we used a few filters which we believed were motivated. The first filter is based on the observation we had made during our earlier work (Papers I-II) that genes from the same ncRNA family are frequently present in multiple copies throughout the genome. Therefore, we clustered the predictions into families according to sequence similarity and experimentally verified expression from sequences belonging to four out of eight families tested. Another observation we had made in previous studies (Papers I-II) was the presence of a conserved sequence element at a fixed distance upstream of many of the ncRNAs identified. We had named this element *Dictyostelium Upstream Sequence Element* (DUSE) because of its similarity to the upstream sequence element (USE) commonly found upstream of snRNA genes in other organisms (Hernandez, 2001). The USE acts as a promoter element in other organisms and we assume that it plays a similar role in *D. discoideum*. Thus, we implemented a transcription signal based filter that searches for the presence of a DUSE element upstream of the predictions. Two predictions were experimentally assayed and expression was detected from both of them. One of these was found to correspond to a previously reported selenocysteine tRNA (Shrimali et. al, 2005). In addition, predictions belonging to two of the experimentally verified clusters from the first filter were also found to be preceded by DUSE elements. We conclude that in the special case of *D. discoideum*, we can successfully exploit the bias in nucleotide composition between ncRNAs and the rest of the genome for *de novo* ncRNA gene prediction. Further refinements of the predictions can be made by employing filters based on existing knowledge of the organism. We note however, just like in previous base composition studies done in hyperthermophiles, that the method seems biased towards detecting ncRNAs with extensive secondary structure (Klein et al., 2002; Schattner, 2002). This manifests itself most clearly through the failure to detect most of the less structured C/D-box snoRNAs and Class I RNAs.
Analysis of *D. discoideum* ncRNAs

**An upstream sequence element**

When analyzing genomic sequences upstream of the isolated ncRNAs (Paper I), we identified an eight nucleotide long sequence element present in front of most Class I RNAs, the U2 snRNA, the SRP RNA and some of the identified snoRNAs. The sequence element is similar to the upstream sequence element (USE, or sometimes known as proximal sequence element (PSE)) found in association with spliceosomal snRNA genes in other organisms (Hernandez 2001). We named this sequence motif *Dictyostelium upstream sequence element* (DUSE). The element is also present at a fixed distance of approximately 63 nt upstream of all snRNAs (Paper II) with no distinction between the U6 snRNA, which is transcribed by RNA polymerase III, and the other snRNAs which are transcribed by RNA polymerase II. In addition, a closer inspection of the genomic sequences preceding RNaseP RNA, RNase MRP RNA and D2 RNA (Marquez et al., 2005; Piccinelli et al., 2005;
Wise and Weiner, 1981; Takeishi and Kaneda, 1981), identified the DUSE ~63 nt from the predicted start of transcription. The importance of this element was corroborated in Paper III where we used the presence of a DUSE upstream of a high-scoring segment as a filter to refine predictions. Two of the predictions were experimentally tested as well as sequences from two clusters preceded by a DUSE and they were all found to be expressed. Typically, we find the DUSE in association with ncRNA genes but not in association with protein-coding genes. Hence, the DUSE seems to be a promoter element specific for a wide range of ncRNA gene families in *D. discoideum*.

**Small nucleolar RNAs**

In Paper I, 18 of the sequenced ncRNAs were found to contain sequence motifs characteristic of snoRNAs. 17 of the *D. discoideum* snoRNAs could be classified as the C/D-box type. We subsequently used a bioinformatic screen to search for target sequences, *i.e.* *D. discoideum* rRNA sequences complementary to the C/D-box snoRNA guide sequences (Cavaillé and Bachellerie, 1998). Candidate targets could be assigned for 12 of the snoRNAs and one of these was experimentally tested and verified to be methylated.

**Spliceosomal small nuclear RNAs (snRNAs)**

We identified and characterized 18 loci encoding snRNA like genes (Paper I-II) and we detected expression from 17 loci. All snRNA classes except U6 are present in two or more copies throughout the genome. These copies are frequently arranged in a pairwise manner where two genes coding for the same snRNA class are typically located close to each other and in all possible relative orientations. The same organization has been reported for *e.g.* *D. discoideum* tRNA genes and is believed to reflect a so far poorly understood mechanism of gene duplication (Eichinger et al., 2005).

The U2 gene family contains seven copies which can be further divided into two subfamilies based on sequence similarity. The first subfamily contains three copies which are highly similar to each other and most similar to the conserved U2 snRNA sequence of other species. The second subfamily consists of four copies that are similar to each other but surprisingly divergent from the first subfamily except for the highly conserved sequence motifs known to be involved in intermolecular interactions within the spliceosome. We also observe compensatory base pair substitutions between the subfamilies, maintaining the predicted secondary structure. Upon closer inspection, the U2 snRNAs in the second subfamily also contain a ~40 nucleotide 5' extension that is predicted to fold into a stem loop structure. Experimental data show that the snRNAs of this second subfamily are downregulated during development and are predominately located in the cytoplasm. This is unexpected since splicing is known to take place in the nucleus and the snRNAs normally only go through a brief cytoplasmic phase dur-
ing the maturation process before being re-imported into the nucleus (Will and Lührmann, 2001).

Most snRNAs could be detected as polyadenylated transcripts in the cytoplasm. This is also surprising as polyadenylation in eukaryotes is associated with mRNAs where it is involved in transcript turnover and translation efficiency. A quality control mechanism in yeast involving polyadenylation of ncRNAs followed by degradation by exonucleolytic activity has recently been reported (LaCava et al., 2005; Vanácová et al., 2005). Whether a similar mechanism exists in *D. discoideum* is so far not known.

In accordance with what is known for other eukaryotes, all snRNAs except for U6 can be immunoprecipitated with an antibody specific for trimethylated cap, indicative of transcription by RNA polymerase II followed by further processing (Will and Lührmann, 2001). In agreement with transcription by RNA polymerase III, the U6 snRNA gene ends in a run of T nucleotides which is typically a signal for termination of transcription.

**A new family of ncRNAs**

**Class I and Class II**

Two highly abundant classes of a previously unknown small ncRNA family were isolated and named Class I and Class II RNAs (Paper I). Transcripts from both classes were found to be mainly cytoplasmic and Class I was found to be developmentally regulated. Class I is a large family with 14 isolated unique sequences mapping to 17 loci in the genome. Sequence similarity searches in the genome revealed 24 additional loci. The distinguishing features for the 55-65 nt Class I RNAs are highly conserved 5' and 3' ends, 16 and 8 nucleotides, respectively and a variable sequence in between. The 5' and 3' ends are predicted to form a helix. Class II sequences are present in two highly similar genomic copies and are predicted to form a similar secondary structure as Class I. The Class I and Class II RNAs share an eleven nucleotide sequence motif. The function of these unknown RNA classes is so far not known. However, one of the Class I RNA genes has been knocked-out, generating a strain with a mild phenotype where the developing multicellular aggregates are reduced in size (Avesson and Söderbom, unpublished).

**Computationally identified ncRNAs**

In a computational screen for ncRNA genes we clustered candidates into families according to sequence similarity (Paper III). Out of eight experimentally tested families, expression could be detected from four. One family in particular is interesting. There are 12 candidates in the family, nine of which are preceded by a DUSE. Through, 5'- and 3'-RACE experiments, we could verify expression from at least three and possibly five individual sequences from the family and the mature 5' end is located at the expected 63
nt downstream of the DUSE motifs. Secondary structure prediction using RNAalifold (Hofacker et al., 2002) suggests a conserved, stable secondary structure supported by several compensatory base pair substitutions. The function of this group of ncRNAs is so far unknown and we have not been able to detect similar sequences in other species.

Characterization of human spliceosomal snRNA variants (Papers IV-V)

Heterogeneity among the expressed snRNA repertoire, often in combination with tissue specificity and/or developmental regulation, has been reported in many species, e.g. Xenopus, mouse, sea urchin, D. Melanogaster, T. thermophila, silk worm and D. discoideum (Forbes et al., 1984; Lund et al., 1985; Santiago and Marzluff, 1989; Alonso et al., 1984; Lo and Mount, 1990; Ørum et al., 1992; Gao and Herrera, 1995; Paper II). Variant snRNAs have also been previously reported in human (Sontheimer and Steitz, 1992). The function of this heterogeneity has not been determined but an involvement in tissue-specific or developmentally regulated alternative splicing is often suggested (Mattaj and Hamm, 1989). Most organisms are known to contain a large amount of snRNA-like sequences in their genomes (see e.g. Paper II; Adams et al., 2000; Lander et al., 2001; Waterston et al., 2002; Manser and Gesteland, 1982). The publication of the human genome (Lander et al., 2001) opened up the possibility to get a complete picture of the snRNA heterogeneity on the genomic level. In order to characterize the heterogeneity among human snRNAs, we have in Papers IV and V used both biochemical procedures as well as computational analyses for this purpose, initially focusing mainly on the U1 snRNA.

Biochemical identification and characterization of U1 snRNAs (Paper IV)

We first compiled a list of human genomic locations encoding U1 snRNA-like sequences through sequence similarity searches with BLAT using the well-characterized, dominant U1 snRNA sequence (U1A) as query (Branlant et al., 1980). For completeness, we also extracted positions of interspersed repeats classified as U1 snRNA from RepeatMasker annotations of the human genome (Smit AFA., Hubley R. and Green P. RepeatMasker Open-3.0. 1996-2004; http://www.repeatmasker.org). After manual inspection of upstream and downstream sequences as well as the RNA gene sequence, a subset of 27 genes were selected as promising candidates for further analysis. Northern blot analysis indicated that eight of these were expressed in HeLa cells and 5' and 3' RACE and molecular cloning confirmed the expression of
three snRNAs. We labeled these variants U1A5, U1A6 and U1A7. Further biochemical characterization of these variants revealed that they are ubiquitously expressed among a wide range of tissues and estimation of the relative abundance levels are at 2-7% of the level of U1A which is in the same range as the U11 snRNA of the minor spliceosome (Montzka and Steitz, 1988). We could show that the variants all posses a tri-methylated guanosine (TMG) cap at the 5' end and that they were associated with Sm proteins. This is indicative of transcription by RNA polymerase II followed by export of the transcript into the cytoplasm where Sm proteins assemble on the RNA followed by hypermethylation of the cap (Kiss, 2004). We could also show that the variants are present in nuclear RNP complexes, consistent with a re-import of the processed snRNA variants into the nucleus. Taken together, these observations indicate that the expressed variant snRNAs undergo the same biogenesis pathway as functional snRNAs (Kiss, 2004).

Identification of expressed snRNAs using exon microarrays (Paper V)

In order to take a large-scale approach to identify expressed snRNA genes in human tissues, we analyze publicly available microarray data. The Affymetrix Human Exon 1.0 ST microarray platform (http://www.affymetrix.com) contains probesets interrogating over one million exon clusters of known and predicted exons and transcribed regions in the human genome. Each probeset consists of up to four individual 25 nt probes that are combined to give a single intensity value for the probeset. Of particular interest to us is the fact that the array also includes probesets interrogating known and predicted RNA genes. Hybridization data for RNA extracted from eleven different human tissues has been made publicly available and can be downloaded from the Affymetrix website. From this dataset we extracted probesets corresponding to snRNA gene predictions. After filtering away cross-hybridizing probesets and probesets with low individual probe coverage we extracted Detection Above Background p-values. Briefly, this p-value has been calculated by comparing each probe intensity against the distribution of intensities from roughly 1000 background probes of the same GC-content. Based on these p-values and setting a false discovery rate (FDR) threshold at 0.05 we assigned predicted snRNA genes as expressed or not expressed.

We find expressed variants for all spliceosomal snRNAs except U12 and U6atac. U6 and U1 show the largest number of expressed variants, 28 and 17, respectively. When comparing how the expressed variants are distributed over the eleven tissue types, we find that cerebellum is significantly overrepresented (p << 0.0001) while heart, muscle and lung are slightly underrepresented (0.05 < p < 0.0025; Figure 6). The same general pattern of tissue ex-
pression is observed for U1 and U6 snRNAs. It is particularly interesting that we find by far the highest number of expressed snRNA variants in brain tissue which has been fingered in several studies as being one of the tissues containing the most alternative splicing events (Xu et al., 2002; Yeo et al., 2004). In addition, many families of ncRNAs have been reported to be over-represented in the nervous system and suggested to be involved in the regulation of nervous system development (see Cao et al., 2006; Mehler and Mattick, 2006 and references therein).

Characterization of expressed snRNA variants

**Complementarity to pre-mRNA 5' splice sites**

A striking feature of the U1 snRNA variants we identified in Paper IV is that they all have at least one nucleotide substitution in the 5' end located motif that interacts with the 5' SS of the pre-mRNA (Figure 7). In particular, they all lack perfect complementarity to the highly conserved GU dinucleotide at the first positions of the intron. In vivo data suggests that the U1 snRNA interaction with the 5' SS can tolerate mismatches to the GU dinucleotide provided that the loss of base pairing is compensated for by increasing base pairing throughout the rest of the interaction motif (Aebi et al., 1986; Jackson, 1991). As the identified variants U1A5 and U1A6 deviate from the canonical U1A 5' SS interaction motif sequence at several positions, a putative interaction with the 5' SS consensus would be suboptimal in this respect. We therefore hypothesized that these variants could function by recognizing non-consensus 5' splice sites and be involved in alternative splicing events.

Similarly, of the 17 U1 snRNA variants we found to be expressed in Paper V, all but three have nucleotide substitutions in the 5' SS recognition motif compared to U1A and seven variants (one of which correspond to the U1A6
variant in Paper IV) do not perfectly match the consensus GU dinucleotide interaction motif. Thus, a fairly large number of variant U1 snRNAs are expressed in human tissues, most of which have a different 5' SS recognition motif than U1A. Provided that the variant U1 snRNAs are assembled into snRNPs and take part in the splicing reaction, it seems plausible that they could give specificity for a subset of splice sites which are not optimally recognized by the U1A snRNA. Similar situations where compensatory mutations in the 5' SS recognition motif of U1 snRNA can suppress 5' splice site mutations and shift pre-mRNA splicing patterns in vivo have been reported. (Zhuang and Weiner, 1986; Zhuang et al., 1987; Siliciano and Guthrie, 1988; Cohen et al., 1993; Lo et al., 1994; Hitomi et al., 1998; Zahler et al., 2004).

Evolutionary conservation
In Paper IV, we looked at orthologous positions to the human U1 snRNA variants in other primate species (chimpanzee, orangutan, rhesus macaque, common marmoset) as well as in dog and cow. The variants are well conserved between human and chimpanzee. U1A5 is also present in rhesus and marmoset, although in marmoset there is a large deletion in the middle that would probably inactivate the gene. U1A6 is present in orangutan but has a large deletion within the gene. U1A7 is present in orangutan and rhesus but the rhesus sequence has a deletion. Surprisingly, we find that the dog and cow orthologs to U1A6 and U1A7 correspond to perfect U1A sequences. This is interesting since it suggests that the variants have evolved from U1A while still being expressed and subjected to the snRNA biogenesis pathway in the cell. Since the emergence of snRNA variants with a preference for a different set of splice sites could have dramatic impacts on the global splicing pattern of pre-mRNAs, it is tempting to speculate that the large internal deletions observed in some of the primate species variants represent a positively selected inactivation of the genes.
To investigate whether more similar situations exist, i.e. where some species have snRNA genes that have retained the original sequence while other species have an snRNA gene variant that has substantially diverged in the same locus, we use the 12-way whole-genome multiple alignment available through the Ensembl databases (Flicek et al., 2008; Paten et al., 2008) to locate orthologous snRNA gene loci between human and non-human species (Paper V). Beyond the primate lineage, we are only able to locate orthologous loci for a few genes. However, we identify 2-3 orthologs to human U1 snRNA genes in rodents, horse, dog, cow and platypus, all of which are lo-

Figure 7: Expressed U6 snRNA variants from Paper V (top) and U1 snRNA variants from Papers IV-V (bottom) and their potential interactions with the 5’ SS consensus. Variant nucleotides mismatching the canonical snRNA are indicated by black boxes. Redundant sequences have been omitted. The GU dinucleotide indicates first positions of the intron. Splice site logo is constructed with Weblogo from 15,263 human EST-supported and corrected canonical introns downloaded from SpliceDB (Crooks et al., 2004; Burset et al., 2001).
cated within the first intron of an evolutionarily conserved testis expressed gene, TEX14, in human located on chromosome 17. Two of these correspond to the U1A6 and U1A7 variants we identified in Paper IV. With the exception of one cow sequence, all correspond to the original U1A sequence in non-primate species while the primate orthologs have diverged sequences in the corresponding positions. A homolog of TEX14 is present in the chicken and *Xenopus* genomes but no U1-like sequences can be located within its introns in those species (Figure 8). Taken together, these results indicate that the first intron of TEX14 harbor a cluster of U1 snRNA genes that were active in an ancestral species close to the root of the mammalian lineage. Somewhere along the primate lineage it seems like there has been a relaxation of negative selection, allowing the primate genes to diverge while the cluster is still under negative selection in non-primate mammals.

![Figure 8: Schematic phylogenetic tree illustrating emergence of U1 snRNA variants at TEX14 locus (square). A cluster of U1 snRNA genes appear close to the base of the mammalian lineage (circle). U1 snRNA variants appear along the primate lineage (star). Branching order from NCBI Taxonomy browser, tree is not drawn to scale.](image-url)
Conclusions and future perspectives

My Ph. D. studies have been aimed at the identification and characterization of ncRNAs using computational methods.

Through a combination of computational and experimental approaches we have identified a large number of ncRNAs belonging to many known as well as several previously unknown gene families in the model organism *Dictyostelium discoideum*. In addition, we developed a new approach to use nucleotide composition as a means to identify ncRNA genes *de novo* and we demonstrated the successful use of this approach. We have also identified a large heterogeneity among human spliceosomal snRNAs and we speculate about a possible link to regulation of alternative splicing. But more knowledge spawns more questions. A few lines of inquiry that might shed light on issues of abundance, function and relevance of the ncRNAs we've encountered and which we should be able to follow up in the not too distant future are sketched below.

- Building CM-models of the ncRNA families that seem to be specific to *D. discoideum* would enable more distantly related genomes to be searched for homologs that might have been missed by sequence homology searches.
- Genome sequencing of *D. purpureum*, a close relative to *D. discoideum* is underway. This opens up the possibility for doing *de novo* ncRNA gene discovery using comparative approaches such as RNAz as well as study homologs to already identified ncRNA families.
- Introns that could potentially be spliced by variant U1 snRNAs remain to be found. Available databases that collect and extract information on splice sites could be searched for this purpose. In particular introns with poor complementarity to U1A snRNA could constitute promising candidates.
- High-throughput sequencing of primate RNA enriched for snRNAs are underway and is likely to yield large amounts of valuable information on the primate snRNA transcriptome.
- Another level of heterogeneity is offered by allelic variations (polymorphisms). Large amounts of data will soon become available from the 1000 genomes project which will shed light on this issue.
In order to test different models of evolution, reliable orthology data is needed. This is notoriously hard to obtain between non-protein coding genomic regions of distantly related organisms because this genomic sequence is mainly evolving neutrally. One approach potentially worth exploring could be to align “genome structure” rather than sequence. For example, annotations of ancestral repeats could be used to find orthologous regions sharing a repeat pattern that is consistent with divergence times and the age of repeats.

As a final point, this work highlights how experimental and computational methods may efficiently complement each other. Experimental data generate knowledge which can guide the design of computational approaches that in turn can be used to search vast amounts of sequence data and generate new predictions. In contrast to experimental methods, computational approaches are generally cheap and fast. In addition, they are always reproducible and are not sensitive to expression levels or growth conditions.
Acknowledgements

I would like to deeply thank my supervisor Anders Virtanen for accepting me as a Ph. D. student, for your support and for everything you have taught me. I have really appreciated how you are always full of ideas and how you can always find the time to discuss projects. Your optimism and your passion for science have been true sources of inspiration!

I would also like to thank my second supervisor Leif Kirsebom for all your valuable help, advice and encouragement throughout my Ph. D. studies and for giving me many opportunities to exciting collaborations.

Many special thanks to Fredrik Söderbom for exciting and fruitful collaborations in the slime mold business! Thanks for enduring those insane working hours prior to submission, looking out for me and being a great friend.

My deepest gratitude also to David Ardell for all the support. You have taught me a lot about science and bioinformatics and your encouragement has meant a lot to me. Also, thanks for great collaborations, your ideas and guidance on Paper III were invaluable!

My fellow members of the Virtanen group: Lei, the snRNA expert. It's been really nice working together with you. Thanks for all the discussions about snRNAs and for teaching me about experimental procedures. Good luck with your projects! Per, for always taking the time to answer my questions, the countless parties in your garden and your endless generosity. Good luck in Japan! Niklas, your turn next! Thanks for all the good times, on conferences, in the (well ventilated) office, in Uppsala and on the west coast. See you in Gravarne!

Thanks to collaborators in the Dicty group at SLU: Andrea, for rewarding collaborations and for making sense out of my predictions. Welcome back to Sweden! Lotta, thanks for collaborations, nice company in Chicago and for making the coffee breaks with me and Johan slightly less geeky (but only slightly).
Collaborators at ICM: Santanu, Jaydip, Nurul M. Islam, Ulrika, Bhupi, Hava and Fredrik P.

Thanks to Gerhart Wagner for many interesting discussions concerning manuscripts and RNA in general.

Thanks also to collaborators and co-authors outside ICM: Christina Kyriakopoulou, Jens Schuster, Anders Aspegren and Vincent Moulton.

I gratefully acknowledge the Lennanders Foundation at Uppsala University for financial support during the conclusion of my Ph. D. studies.

Thanks to all present and former colleagues at ICM! The administrators: Sigrid, Sofia, Anders, and Åsa for all their help and assistance over the years. Akiko for computer support. Christer for technical support. Micro: Johan, I've really enjoyed all our discussions, coffee breaks, going to conferences, playing games, working out, etc. etc. (the list reaches from here to Fyrisfjädern). Thanks man! Helene for tough badminton games. I'm dreading a squash revenge. Karin T and Emma R for helpful discussions on microarrays (and the great parties of course!). Hava for sharing nuggets of pop culture, teaching me english language and showing what a pool party is supposed to be like. Cia – thanks for great travel company in Chicago. Maaike, Erik H, Aurelie, Ulrika, Fredrik P, Shiying, Henrik T, Sonchita, Bhupi, Pernilla, Jon, Marie W, Disa, Linda M (and Mattias). Molbio: Magnus L, Harri, Doreen, David F, Ikue, Magnus J, Chandu, Prune, Brian, Paul, Petter. The Kilimann group: Greta, Nicole and Christoph. Immunology: Kajsa and Sofia.

Tack till oslagbara vänner, som håller det mikrokulära på avstånd:

Min familj:
Ert stöd har varit ovärderligt!
Pia, Daniel, Moa och Ebba. Pernilla, Henrik, Olivia, Agnes och Alice.

Och naturligtvis, allra störst tack till Mamma och Pappa som alltid ställer upp och som alltid har sett till att jag haft möjlighet att göra det jag velat.

Sist men också minst: Last.FM som stått för soundtracket till det här kalaset!


Eftersom människan är en högt utvecklad organism med bl.a. en stor och komplex hjärna, har man trott att det skulle behövas en stor mängd proteiner för att bygga upp och hålla igång en individ. När den mänskliga arvsmassan var färdigsekvserad i början av 2000-talet, blev man därför väldigt förvånad över att upptäcka, att människan har ungefär lika många proteinkodande gener som den betydligt mindre utvecklade rundmasken. Man tror nu, att det snarare är små skillnader i hur de olika proteinerna uttrycks, som bidrar till den ökade komplexiteten hos människan, jämfört med t.ex. rundmasken. Eftersom olika ncRNA är mycket effektiva regulatorer av genuuttryck, tror man att dessa spelar en stor roll i detta sammanhang. Ytterligare en mekanism som kan öka komplexiteten hos uppsättningen proteiner en organ-
ism har tillgång till, är den ovan nämnda alternativa splitsningen. Också den processen är i hög grad reglerad av ncRNA.

Att sekvensera arvsmassan (eller genomet), dvs att bestämma ordningsföljden hos DNA-nukleotiderna, hos en organism är av stor betydelse när man vill förstå hur organismen är uppbyggd och fungerar. Att ta reda på nukleotidsekvensen i arvsmassan är dock bara ett första steg. För att ha någon nyttja av den kunskapen måste man veta vad sekvensen betyder, bl.a. måste man kartlägga bitarna som kodar för gener: både gener som kodar för proteiner och gener som kodar för ncRNA. Att göra detta för ett helt genom är ett stort projekt och till stor del måste man därför förlita sig på datorbaserade metoder, som kan analysera hela genomet och märka ut generna.

I mitt doktorandarbete har jag med hjälp av datorbaserade metoder identifierat och karakteriserat ncRNA utifrån sekvensdata. Vi har också studerat den mångfald som finns bland mänskliga splitsnings-RNA och hur detta kan påverka uppsättningen proteiner hos människan.

I min avhandling presenterar jag strategier vi använt oss av för att identifiera och karakterisera ncRNA gener i amöban Dictyostelium discoideum, på svenska kallad slemsvamp. Slemsvamp är en viktig modellorganism som har studerats under de senaste 50 åren. Trots att den studerats under så lång tid, har man känt till mycket lite om dess ncRNA repertoar. Vi använde oss av flera metoder, både experimentellt baserade och datorbaserade, för att identifiera ncRNAs i slemsvamp.

Vi kunde identifiera många ncRNA tillhörande dels väldokumenterade ncRNA-klasser, som dock inte varit kartlagda i slemsvamp tidigare, men också klasser av tidigare okända ncRNA, som eventuellt är specifika för den här organismen. Funktionen hos dessa okända ncRNA är inte kartlagd. Genom att vi hittade många ncRNA kunde vi upptäcka speciella egenskaper som var ganska utmärkande för ncRNA-gener i slemsvamp. Vi kunde därför, med hjälp av dessa egenskaper, skapa en helt datorbaserad sökmetod som vi använde för att hitta ännu fler okända ncRNA.

Vi har också studerat den naturligt förekommande variationen i splitsnings-RNA (snRNA) hos människan. Eftersom en del av dessa snRNAs är inomvovrerade i lokaliseringen av splitsningsställen i mRNA har det förekommit spekulationer om att varianter av snRNA skulle kunna påverka t.ex. alternativ splitsning. Vi har visat, både genom datorbaserade och experimentella metoder, att ett flertal varianter av snRNA, som skulle kunna påverka lokaliseringen av splitsningsställen, förekommer i mänskliga celler. Vi tror därför att snRNA-varianter skulle kunna påverka alternativ splitsning.
Sammantaget har vi genom att kombinera datorbaserade och experimentella metoder identifierat en rik och varierad ncRNA-repertoar i slemsvamp och i människa. Denna kartläggning kan tjäna som ett första steg för att mer ingående studera funktionen hos de olika ncRNA-generna. Variationen vi ser i mänskliga snRNA-gener skulle kunna ha betydelse för splitsningen av mRNA och därigenom bidraga till att öka antalet olika proteinprodukter, som cellen är kapabel att tillverka.
References


Orum, H., Nielsen, H. & Engberg, J., 1992. Structural organization of the genes encoding the small nuclear RNAs U1 to U6 of Tetrahymena thermophila is very similar to that of plant small nuclear RNA genes. *Journal of Molecular Biology*, 227(1), 114-21.


Zhuang, Y., Leung, H. & Weiner, A.M., 1987. The natural 5'splice site of simian virus 40 large T antigen can be improved by increasing the base complementarity to U1 RNA. Molecular and Cellular Biology, 7(8), 3018-3020.
Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 589*

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)