universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„RNA-RNA long range interactions in mammalian genomes“

verfasst von / submitted by

Dominik Steininger BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2019 / Vienna 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 834
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Molekulare Biologie
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

Abstract

RNA-splicing by the spliceosome is an important characteristic of eukaryotic organisms and the foundation of protein diversity and synthesis in humans. During the splicing process introns are removed from the coding sequences to yield a mature RNA transcript that can be further translated to proteins. The size of these introns can reach from a few to hundreds of thousands basepairs.

For small introns we can argue that the spliceosome or associated auxiliary factors detect both ends of the intron simply by the spacial proximity of the two splice-sites. In large introns, however, this argument loses its weight as the size of the spliceosome is limited. Nevertheless, there has to be a mechanism that brings the two splice-sites together to be recognized and processed.

Considering the fact that RNA is prone to form intra-molecular interactions, we suggest that long introns form secondary structures and bring both splice-sites closer together so that the spliceosome can recognize and remove them. We predicted possible interactions between the 5' and the 3' ends of intronic regions and compared them to a background model which revealed that interactions between the ends of native intron sequences are much more stable than average interactions. Additionally, we found an over-representation of interactions between the outermost parts of the 5' and the 3' regions of the intron. This suggests that at least part of the introns show some kind of structuredness that could affect the splicing regulation.

A closer investigation revealed that inverted Alu-sequences are involved in forming strong connections between intron-ends. Moreover, Alu-elements may influence the transcription process and are in some cases involved in genetic diseases, even if they are not inserted directly into the coding regions. Possible long-ranged structures between intron ends could indicate another mechanism of splicing regulation and closer investigation could provide an approach for better understanding the complex regulation in alternative splicing.

Zusammenfassung

Das Spleißen von RNA durch das Spleißosom ist ein bedeutendes Markenzeichen eukaryotischer Organismen und der Eckpfeiler der Proteindiversität und -synthese beim Menschen. Während des Speißprozesses werden die nicht-kodierenden Teile, sogenannte Introns, aus dem RNA-Transkript entfernt, sodass nur mehr kodierende Sequenzen übrig bleiben und als Vorlage für die Proteinsynthese dienen. Die Größe dieser Introns variiert zwischen einigen wenigen und einigen hunderttausend Basenpaaren. Das Spleißosom muss den Anfang und das Ende der intronischen Regionen erkennen und diese in weiterer Folge entfernen.

Für kurze Introns kann man damit argumentieren, dass das Spleißesom mit Hilfe von zusätzlichen Faktoren die beiden Spleißstellen erkennen und herausschneiden kann. Betrachtet man jedoch lange Introns, ist es schwer vorstellbar, dass die Spleißenden allein durch die Größe des Spleißkomplexes erkannt werden können. Dennoch muss ein Mechanismus existieren, der die beiden Intronenden soweit zusammenführt, dass das Spleißosom diese erkennen und spleißen kann. In Anbetracht der Tatsache, dass RNA intramolekulare Strukturen ausbildet, schlagen wir vor, dass die beiden Spleißenden durch RNA-Faltung in räumliche Nähe gebracht werden und dadurch vom Spleißosom erkannt und weiter prozessiert werden können.

Wir haben die bestmöglichen Wechselwirkungen zwischen den 5’ und 3’ Enden der nativen Introns berechnet, diese mit einem Hintergrundmodell verglichen und konnten feststellen, dass native Sequenzen im Vergleich zum Kontrollmodell deutlich stabilere Interaktionen ausbilden. Zusätzlich fanden wir gehäuft Interaktionen zwischen den äußersten Region der Introns. In weiteren Untersuchungen konnten wir zeigen, dass unter anderem Alu-Elemente an der Bildung starker Wechselwirkungen beteiligt sind. Darüber hinaus ist es möglich, dass Alu-Elemente auf diese Art den Transkriptionsprozess beeinflussen und damit auch etwaige genetische Krankheiten verursachen können, selbst wenn Alus durch ihre transposablen Eigenschaften nicht direkt in kodierende Regionen springen. Hier präsentieren wir Hinweise auf einen weiteren Mechanimus der Spleißregulation

durch Interaktionen zwischen den äußersten Enden von langen Introns, die dazu führen, dass die Enden in räumliche Nähe gelangen und die Erkennung der Spleißstellen erleichtert wird.

Acknowledgements

First of all I want to thank Ivo Hofacker for supervising this work, his support, patience and his knowledge and explanations. Following this, I also want to thank all members of the one and only room 309, which provided a pleasant and cheerful working environment. Especially Stefan, who did not get tired reading parts of this work over and over.

Additionally, I want to express my gratitude to Florian Eggenhofer, who brightened the days with his legendary jokes, to Sven Findeiß for the social gatherings and proof-reading this work, and Christoph Flamm, who is a source of new ideas and interesting (yet distracting) projects.

Thanks to Judith and Richard for all the administrative and mental support, the cooking expertise and for being kind of a social glue.

I would like to thank all members of the institute for providing a constructive and enjoyable place to work, a nice community and of course the soccer-table.

Special thanks goes to my wife Katharina for her endurance and support in writing this thesis, for encouraging and motivating me and standing by my side. On this occasion I also want to thank my family for all support throughout my studies.

Contents

Chapter 1

Motivation

In the last decades the known role of RNAs (ribonucleic acids) in life has transformed from a general intermediate product of protein synthesis to an important signalling and regulatory element with a lot of functionality [Crick, 1970]. This opened a wide field for researchers to study RNA and its complex functions in the cell. In spite of all the ongoing discoveries the involvement of RNA in the transcription process still raises a lot of new questions. Especially the role of intergenic sequences in the eukaryotic transcriptome and their role in evolution is still keeping a lot of mysteries. Compared to other eukaryotic organisms, humans have a relatively small number of genes but produce a lot more proteins than suggested by the "one gene, one enzyme" theory, postulated by Beadle and Tatum [Beadle and Tatum, 1941]. This is due to a process called alternative splicing that allows a cell to combine various exonic regions of a gene to create multiple transcripts. An extreme example is the DSCAM-homologue gene in *Drosophila* with about 38,000 protein-isoforms [Schmucker et al., 2000]. It is not yet fully known how this alternative splicing mechanism is regulated and how exons and introns are processed to yield different transcripts in the mRNA (messenger RNA) maturation process. In most splicing events a large complex called the spliceosome is responsible for stripping out intronic regions and transforming pre-mRNA to mature-mRNA, which then can be translated into proteins.

The length of these introns can reach from short sequence snippets with a few nucleotides up to a length of hundreds of thousands of bases. It has been found that short introns are recognized across exonic markers, but intergenic regions longer than 200 nt are usually only recognized based on intronic features, which

means that the spliceosome has to know somehow where it has to splice out the right sequence [Fox-Walsh et al., 2005]. For smaller introns we can argue that the spliceosome may detect both ends of the introns simply by its size and the spacial proximity of the two splice-sites. But the process of long intron splicing still raises a lot of questions. We can find introns with a length up to 1.5 Mb in the human genome. In such dimensions the argument of spacial proximity loses its weight. Even if there are many additional splicing factors involved in recognizing the splice-sites, it seems very unlikely that the whole complex can span such large regions. Nevertheless, somehow the spliceosome needs to find the appropriate downstream-splice-site to cut out the non-coding sequence. Considering the fact that RNA is prone to form intra-molecular interactions, we hypothesize that introns form secondary structures that bring both the 5' and the 3' ends into vicinity so the spliceosome can recognize the splice-sites and process the non-coding regions.

Chapter 2

Background

2.1 RNA

Ribonucleic acid (RNA) is one of life's most basic and most indispensable molecules, as it is one of three major macromolecules in all forms of known life. While DNA (desoxyribonucleic acid) was discovered to carry all the information a cell needs to function and to replicate, at the beginning of the 20th century RNA was not given a lot of attention. In the 1930s RNA was discovered in plant viruses. At this time it was still believed that only proteins are complex enough to pass on hereditary information. In 1957 Fraenkel-Conrat and Singer could proof that viral RNA carries the genetic information of *tobacco mosaic virus* (TMV), but it was not yet clear how the protein coding functions of RNA were realized [Fraenkel-Conrat and Singer, 1957]. One year later Richard Roberts discovered ribosomes as the site of protein synthesis in *Escherichia coli* (E.coli) and at the end of the 50s Jacob and Monod proposed the idea of a messenger RNA (mRNA) as the template for protein synthesis in bacteriophage T4 [Roberts, 1958] [Jacob and Monod, 1959].

The central dogma of molecular biology, first proposed by Crick in 1958, states that DNA maintains the information of all proteins, that can be synthesised via an intermediate product - the mRNA (Figure 2.1) [Crick, 1970]. In recent decades it has been shown that RNA has a much more complex role in life than proposed in Cricks famous statement. Involvement of so called non-coding RNA (ncRNA) in regulatory processes has led to the assumption that the "RNA world" preceded the evolution of DNA and proteins [Gilbert, 1986]. Since then ncRNAs have

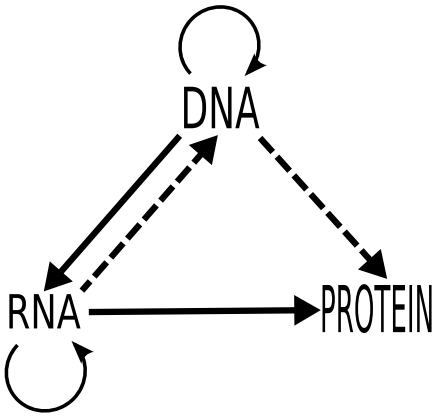

FIGURE 2.1: Depiction of Francis Cricks central dogma of molecular biology. Solid lines indicate the processes proposed by Crick, DNA can be replicated or transcribed into RNA, which acts as template for protein synthesis [Crick, 1970]. Nowadays, it is known that RNA has a more complex regulatory function depicted by the dashed lines. Additionally to the Cricks proposal, RNA can act as a template for reverse transcription or regulate and interfere with DNA. Figure adopted from [Lorenz, 2014].

been grouped into different classes according to their function. Transfer RNAs (tRNA) and ribosomal RNAs (rRNA) are important in the translation process. Micro RNA (miRNA), small interfering RNA (siRNA) and piwi-interacting RNA (piRNA) regulate RNA expression. Furthermore, small nucleolar RNAs (snoRNAs) located in the nucleolus act as guides for chemical modifications on other RNAs and play an important role in rRNA processing. A short overview of some non-coding RNAs can be found in Table 2.1.

tRNA	physical link between amino acid and mRNA, important for translation
rRNA	component of ribosomes, essential for protein translation
miRNA	part of the RNA interference pathway, regulates RNA expression
siRNA	part of the RNA interference pathway, regulates RNA expression, artificial
piRNA	part of RNA interference pathway, regulates RNA expression, especially in spermatogenesis
snoRNA	guide chemical modification of tRNA, snRNA and rRNA
lncRNA	non-coding transcripts longer than 200 nt

TABLE 2.1: General overview of some prominent non-coding RNA classes and their role in eukaryotes.

2.1.1 The structure of RNA

RNA is a polymeric molecule consisting of nucleobases that are interconnected via a ribose-phosphate backbone. The monomers, so called nucleotides, are made of a five carbon ribose sugar with its C atoms numbered 1 to 5, a nitrogenous base that is attached to the 1' carbon atom via a glycosidic bond and a phosphate group on the 5' position, that is important for polymerization (see Figure 2.2). The bases are modified in numerous ways during a complex maturation process that fulfills a wide range of functions. In general an RNA-sequence is a chain of nucleobases interconnected by a 5'-3' phosphodiester bond [IUPAC-IUB Comm. on Biochem. Nomencl, 1970]. The nitrogenous bases are either the pyrimidines cytosine (C), uracil (U) or purine bases adenine (A) and guanine (G) abbreviated by their first letter. In DNA uracil is substituted by an almost identical thymine (T) base that features an additional methyl group on its 5' position.

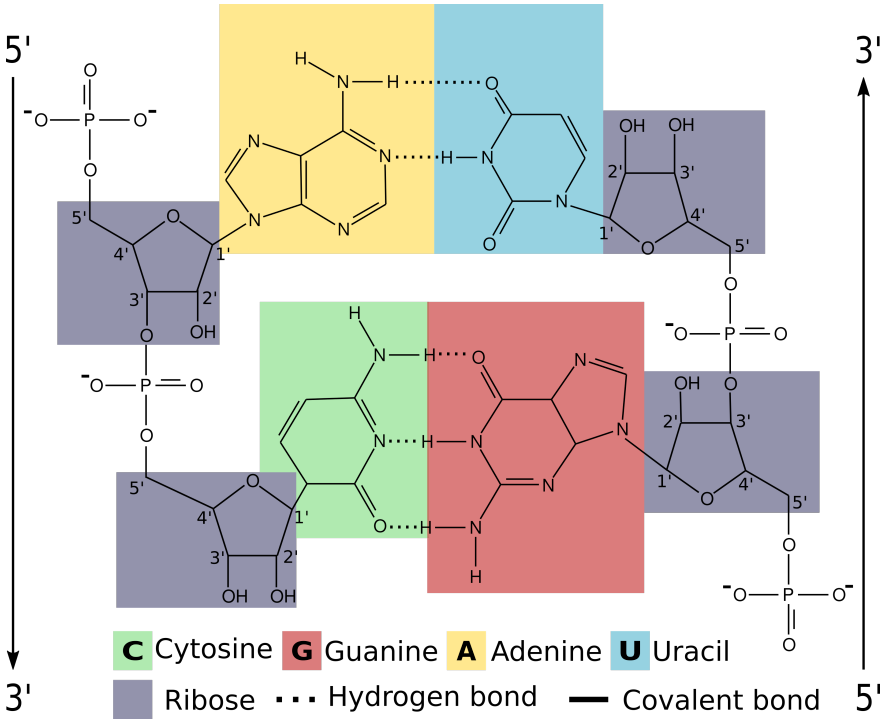

FIGURE 2.2: Two RNA strands oriented in opposite directions. The strand on the left shows an adenine (yellow) paired with an uracil (blue) via two hydrogen bonds (dotted lines), and a cytosine (green) that forms three hydrogen bonds with guanine (red). The nucleotides are connected to each other via alternating ribose (grey) and phosphate residues that indicate the strand orientation. The nucleobase is attached to the 1' position of the ribose, at the 3' position the ribose is chained to the next nucleotide via a phosphate residue. Figure adopted from [Eggenhofer, 2011]

In standard Watson-Crick base-pairing an A pairs with an U and G pairs with a C under the formation of hydrogen bonds [Watson and Crick, 1953]. Besides these canonical base-pairs, RNA can also form energetically weaker non-canonical pairs, for example the Wobble pair between a G and an U [Leontis and Westhof, 2001]. Most computational approaches do not consider non-canonical pairs other than GU, because measuring the energy parameters for such weak interactions is a challenging task and not yet satisfactorily solved.

The sequence of nucleotides in the polymeric chain is called the **primary structure** (Figure 2.3a). It is usually depicted by the initial letter of the base name

$$5'AUGC3'$$

and is generally written from the 5' to 3' end.

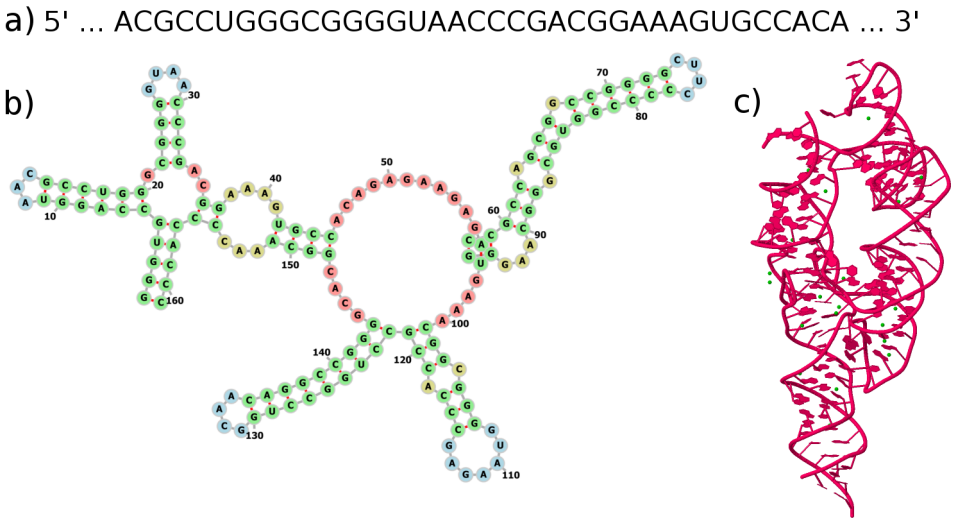

FIGURE 2.3: Examples of different representations of *Thermus aquaticus* RNAseP (GENBANK: Z15006.1, [Brown et al., 1993]). The primary structure (a) is represented by the initial letter of the base, the secondary structure (b), created with *forna* ([Kerpedjiev et al., 2015]), shows additional intramolecular basepairing information and the tertiary structure (c) adds spacial information. The 3D structure is created with pymol [Schrödinger, LLC, 2015].

The **secondary structure** (Figure 2.3b) of RNA refers to basepairing interactions within a molecule. These intramolecular interactions are responsible for the shape of nucleic acids. A single stranded RNA-chain usually folds back on itself to form a thermodynamically more stable structure, for example a stem (Figure 2.4a). Characteristic pairing patterns in RNA molecules are stems, hairpin loops, bulges, multi loops and internal loops (see Figure 2.4). These structures

are the building blocks of energy calculations in RNA bioinformatics, but they do not carry the complete spatial information. A special motif is the so called **pseudoknot** shown in Figure 2.5, where nucleotides within a loop interact with nucleotides that are not part of the loop. Pseudoknots are a special case in RNA secondary structure prediction, because they add another layer of complexity and are computationally very expensive.

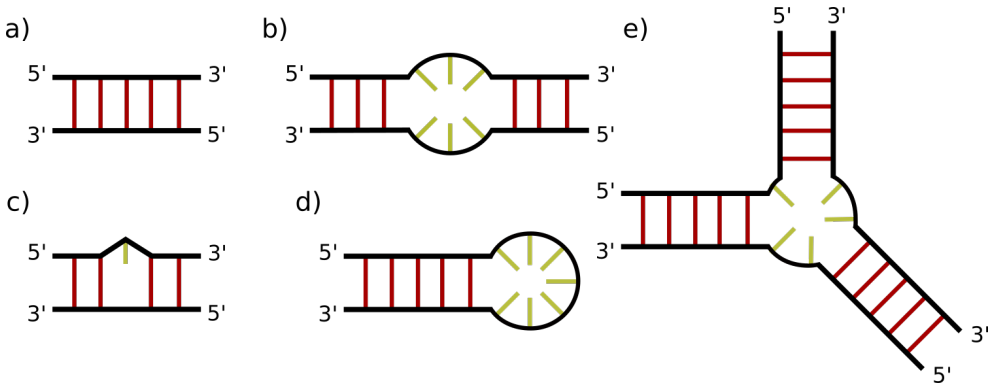

FIGURE 2.4: RNA secondary structure motifs for structure decomposition. Red lines indicate paired nucleotides (Watson-Crick or Wobble pairs), green unpaired nucleotides and the black ones the backbone. a) shows a double helix or stem structure, b) is called an interior loop, with a bubble of unpaired bases in between a double-helical structure, c) depicts a bulge, d) a hairpin or stem loop and e) is named a multi-loop.

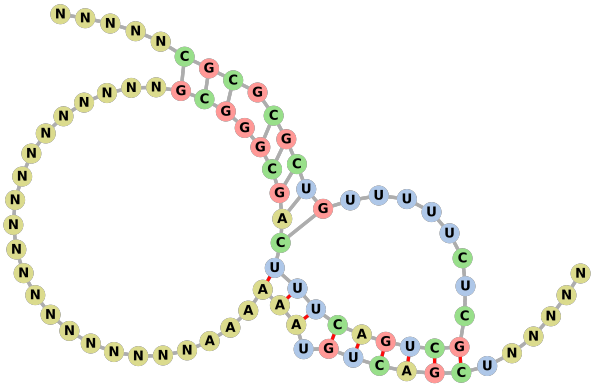

FIGURE 2.5: In a pseudoknot nucleotides from a loop pair with another region located outside of the loop, in this case a G-C rich region. This image depicts the pseudoknot structure of the P3 helix of the human telomerase (hTR) [Chen and Greider, 2005] and was created with forna [Kerpedjiev et al., 2015]. Us are colored blue, A, G, C are colored yellow, red and green, respectively. N is a placeholder for any of the bases AUGC.

The three dimensional information is introduced by the **tertiary structure** (Figure 2.3 c) and describes the spatial shape of the macro-molecule. Therefore

it also appraises intrinsic steric and geometric constraints that are important for biochemical activity.

When considering the spatial arrangement of the tertiary structure in context with other molecules it is called the **quaternary structure**. These interacting molecules can be different RNA strands, DNA or even proteins. A famous example for a quaternary structure complex is the spliceosome with its 5 snRNAs (Figure 2.3.1.2) interacting with a multitude of proteins.

2.1.2 RNA vs DNA structure

From a chemical point of view DNA and RNA are very similar molecules, DNA misses a hydroxy-group on its 2' position, hence called deoxyribonucleic acid. This small difference is the cause for DNA's stability, that makes it a lot more stable than RNA and less prone to hydrolysis under alkaline conditions. But this "instability" is the cause for RNA's most outstanding feature, it can form so called ribozymes. Those are catalytically active conformations of RNA-sequences with important roles in cellular mechanisms e.g. the translation process or the splicing reaction. Because of the active C2 atom, the RNA is able to chemically attack the adjacent phosphodiester-bond and cleave the RNA backbone . These active conformations are a direct result of the tertiary structure of RNA that in turn is encoded on sequence level. Unlike DNA, RNA is mostly found as a single stranded molecule which allows a higher number of structures that can perform catalytic reactions. Generally, RNA is more prone to form intramolecular base pairs and DNA tends to build intermoleculare base pairs resulting in the well known double helix structure. In helix form RNA mostly adopts the A-form geometry that leads to a deep and narrow major groove and a wide and shallow minor grove. DNA is usually found in the more compact B-form.

2.2 RNA Bioinformatics

2.2.1 RNA secondary structure prediction

In biology the structure of a molecule is generally linked to its function. Secondary structure prediction, also known as RNA folding, is the basis for more complex predictions that may help to reveal the function of certain RNAs. Modelling RNA folding still is a complex task, because in nature everything is a fluent system influenced by a phletora of parameters. Mostly, when we talk about the secondary structure of an RNA molecule we refer to the minimum free energy structure (MFE-structure), the structure which is thermodynamically the most stable in a certain environment. We can also use secondary structure prediction to calculate suboptimal structures a sequence can adopt. An efficient and often used way to calculate secondary structures is the dynamic programming (DP) approach.

DP is a divide and conquer strategy where the problem is split into subproblems that are solved individually. It makes use of memoization, a method where the results of already solved functions are stored and don't have to be recalculated every time, instead the solutions of the subproblems are looked up. This saves a lot of computation-time at the expense of increased memory requirements. DP has proved very useful for secondary structure prediction.

In terms of **RNA folding** this means that the molecule is first decomposed into defined substructures (e.g. Figure 2.4), following an algorithm specific recursion scheme. A simple example for DP is the maximization of basepairs calculated by the Nussinov algorithm shown in Figure 2.6 [Nussinov, 1978].

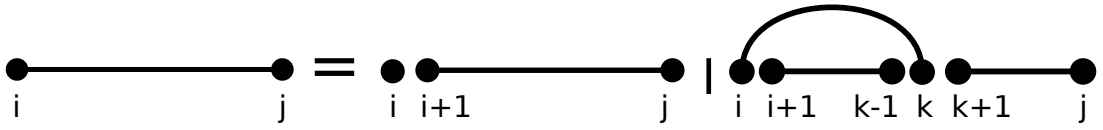

FIGURE 2.6: Example of dynamic programming recursion: Decomposition of a secondary structure into subproblems according to the Nussinov algorithm for maximization of basepairs in RNA sequence folding [Nussinov, 1978]. A nucleotide i can either be unpaired and we get a substructure $s[i+1;j]$ or it is paired with another nucleotide on position k resulting in the substructure $s[i+1;k-1]$ and $s[k+1;j]$.

This algorithm evaluates the maximum number of basepairs (N) a stretch of RNA can form but it neglects actual binding energies. It has a runtime of $O(n^3)$

and needs $O(n^2)$ space. Based on the Nussinov recursion (Equation 2.1) Zuker and Stiegler developed a more sophisticated model by assigning actual binding energies to loops (Figure 2.4), which allowed them to determine the most stable secondary structure a sequence can fold into under certain conditions, the so called minimum free energy (MFE) [Zuker and Stiegler, 1981]. These binding energies were taken from melting experiments of the Turner group [Turner and Mathews, 2010].

$$N_{i,j} = \max \begin{cases} N_{i+1,j-1} + (i,j) \\ N_{i+1,j} \\ N_{i,j-1} \\ \max_{i<k<j}(N_{i,k} + N_{k+1,j}) \end{cases} \tag{2.1}$$

2.2.1.1 Minimum Free Energy calculations

Often one is interested to find the most abundant structure an RNA can fold into under certain conditions, which is the structure with the lowest free energy called the Minimum Free Energy (MFE) structure. It is important to keep in mind that the MFE structure represents only one out of many possible structures or states an RNA can adopt, therefore it still may be poorly populated. One can imagine the set of all possible structures, also called an (thermodynamic) ensemble, as a landscape with energy barriers that have to be overcome to get from one state into another. The MFE structure then is located at the lowest point of this landscape, the global minimum, the other structures are referred to as suboptimal structures.

The Zuker algorithm as implemented in the *Vienna RNA Package* uses a loop based model for energy assignment [Lorenz et al., 2011]. In general, stacked pairs have a negative energy contribution, whereas loops (see Figure 2.4) add destabilizing effects. The complete implementation of the RNA folding algorithm including grammar rules can be found in [Bompfünewerer et al., 2007]. Based on the Zuker algorithm John McCaskill published an algorithm that applied ideas from statistical mechanics to RNA structure prediction to derive statistical properties of RNAs by transforming energies to probabilities [McCaskill, 1990].

2.2.1.2 The partition function

By using the partition function Z it is possible to derive equilibrium properties of RNA sequences from its free energies and determine the abundance of certain secondary structures in an ensemble. The partition function can also be used to compute the likeliness of two bases forming a basepair or the energies needed to separate complementary stretches of RNA, which is very useful in RNA-RNA interaction search, that we will discuss later.

$$Z = \sum_{s \in S} \exp\left(\frac{-E(s)}{RT}\right) \tag{2.2}$$

According to Equation 2.2, Z is the sum over the Boltzmann-factors of all structures S, where $E(s)$ is the energy of a certain structure, T the absolute temperature in degrees Kelvin and R the molar gas constant.

The probability of a structural feature f in the ensemble can then be computed by calculating the partition function of all structures expressing a certain feature and setting it in relation to the overall partition function Z (Equation 2.3).

$$P(f) = \frac{Z_f}{Z} \tag{2.3}$$

2.2.2 RNA-RNA interaction search

With the discovery of miRNAs in 1993 a new era in RNA biology was ushered [Lee et al., 1993]. The group of Rosalind Lee discovered that the lin-4 gene associated with the development of *C. elegans* does not produce a protein-coding RNA, instead it codes for a short sequence that post-transcriptionally inhibits the translation of the lin-14 gene. Nowadays we know that RNA is an important regulator of cellular activity and a key part in understanding the complex networks in cells. Finding such regulators still is not an easy task, although there are a lot of useful computational methods out now.

Pattern matching approaches for DNA, for example BLAST [Altschul et al., 1990], have proved very badly for RNA-RNA interaction prediction, because two interacting RNAs rarely form a perfect duplex and higher level structures are neglected.

The most simple approach for interaction prediction is to concatenate target and query sequence via linker nucleotides and apply a standard folding algorithm such as Zuker's *mfold* [Zuker and Stiegler, 1981]. Beside a runtime of $O(n^3)$ it also has limitations, that render it impractical for most RNA-RNA interaction search, because it allows intra-molecular interactions in the target- and the query-regions as well as interactions with the linker region itself, and therefore the predictions are very inaccurate.

RNAcofold and *pairfold* tackled these problems by restricting the formation of pseudoknots and using adopted energy parameters for the linker regions [Hofacker et al., 1994] [Andronescu et al., 2005]. The exclusion of pseudoknots eliminates a lot of possible interactions from the ensemble with the advantage of reducing the computational complexity. Therefore, the most naive implementation has a time complexity of $O(n+m)^3$ and a space complexity of $O(n+m)^2$. Because of the cubic time-complexity it is not suited for whole genome scans as needed for this study. An improvement of this technique was introduced with *RNAhybrid*, which omits secondary structure formation in monomers and only considers inter-molecular basepairs [Rehmsmeier et al., 2004]. With a time-complexity of $O(mnL^2)$ it is made for finding the most favourable interaction of a small RNA in a large sequence.

A more sophisticated concept for RNA-RNA interaction search was introduced with *RNAup* [Muckstein et al., 2006]. Here the RNA-RNA-interaction is decomposed into two steps, where first the probability that an interaction site is unpaired or accessible is evaluated and then it is combined with the hybridization energy to yield an overall binding energy considering intra-molecular as well as inter-molecular parameters. This leads to a more accurate prediction with the downside of additional computational cost resulting in a time complexity of $O(n^3 + nw^5)$ and space requirements of $O(n^2 + nw^3)$, where w is the maximum length of the interaction. This approach also allows in-loop-binding and the prediction of kissing hairpins, which is not possible in the previously-described tools, and it additionally provides a wide range of interesting applications, e.g. riboswitch-design, but it is too expensive for genome-wide interaction scans.

Another tool, *IntaRNA*, extends the *RNAup* approach with a seed-search algorithm, that makes use of the nearly perfect complementary miRNA seed regions [Busch et al., 2008]. With a time complexity of $O(n^2m^2)$ and memory requirements of $O(nm)$ it is pretty fast in miRNA scans and can handle whole genomes,

but because of its seed recognition it is mainly suited for miRNA searches and therefore not appropriate for this study.

A different approach was introduced by *RNAplex*, which focuses on speed at the expense of accuracy to perform interaction searches on whole genomes [Tafer and Hofacker, 2008]. It uses a simplification of the loop-energy calculation by replacing the logarithmic scoring function with an affine function, similar to the Gotoh-algorithm, resulting in a time-complexity of $O(mn)$ [Gotoh, 1982]. Later on *RNAplex* was extended to also mimic intra-molecular interactions by deriving position-dependent nucleotide-penalties from local pair-probabilities computed with *RNAplfold* [Tafer et al., 2011] [Bernhart et al., 2006]. Initially *RNAplex* was designed to search miRNA targets in the whole genome. Because of the relatively cheap runtime and the ability to approximate intra-molecular interactions by calculating the strand opening energies, *RNAplex* combined with *RNAplfold* were the programs of choice for interaction calculations.

2.2.2.1 Calculating the strand opening energies

As already mentioned, RNA is susceptible to fold back on itself, forming so called intra-molecular base pairs. Based on the partition function as previously described (Equation 2.2), it is possible to compute the probability of a single basepair being unpaired. The probabilities then can be used to derive the energy that is needed to unfold a certain stretch of the secondary structure by converting probabilities to opening energies as shown in Equation 2.4. This opening energy is used in programs like *RNAup*, *IntaRNA* and *RNAplex* to evaluate the binding energies of interactions.

A tool which is used for the efficient calculation of local base-pairing-probabilities is *RNAplfold*. It uses a sliding window approach where it averages base-pairing probabilities over a given window size W and a maximum base separation of L [Bernhart et al., 2011]. *RNAplfold* is a helpful tool for scanning large sequences or even genomes for short RNA-structures by pre-calculating base-pairing features.

$$\Delta G^u_{ij} = \frac{-1}{RT} \ln(\frac{Z_u[i,j]}{Z}) = \frac{-1}{RT} \ln P^u_{ij} \tag{2.4}$$

2.2.2.2 Evaluating RNA-RNA interactions using *RNAplex*

As previously mentioned *RNAplex* from the *Vienna RNA Package* is a fast tool for RNA-RNA interaction search which focuses on speed in expense of accuracy and it is used to find possible interactions between two RNA sequences [Tafer and Hofacker, 2008]. Along with the most stable interaction in terms of energy, suboptimal interactions are returned in kcal/mol. Intra-molecular interaction effects due to RNA refolding can be approximated either by a fixed energy penalty which is usually derived from sequence composition, or by accessibility profiles calculated with *RNAplfold*. These profiles then can be used to derive a position dependent per nucleotide penalty that is more accurate than a fixed cost. As mentioned in the section before, opening energies can be derived from the probability of a region being unpaired. Because *RNAplex* stores only the last four results of the recursion matrices during scan phase, the probability that region $[i, j]$ is unpaired has to be approximated. Considering Equation 2.4 the probability of a region being unpaired can then be transformed to the corresponding opening energies and the penalty cost for the addition of a certain nucleotide can than be written as in Equation 2.5. To obtain the most stable interaction between two molecules, the energy is minimized over the closing pairs of the duplex, and the high scoring interactions then are evaluated using the full energy model of *RNAduplex*.

$$\begin{aligned} \Delta G_{i,j} &= -RT \ln P(j+1|i,j) = -RT \ln \frac{P(i,j+1)}{P(i,j)} \\ &\approx -RT \frac{P(j-\delta, j+1)}{P(j-\delta, j)} \end{aligned} \tag{2.5}$$

2.3 From Genes to Proteins

When we think about DNA and RNA, another major building block of life comes into our minds: proteins. It all starts with a gene, which describes the basic functional and physical unit of the DNA that contains the instructions to build a protein via an mRNA intermediate. A gene consists of different elements, often the protein coding sequences make up only a small fraction of these elements. In eukaryotes the coding regions of genes are termed exons, they are interrupted

by introns that have non-coding characteristics. Generally, the key step of the transcription process is carried out by RNA polymerase II, which first yields an unprocessed transcript that contains a capping structure followed by the 5' untranslated regions (UTR), exons, introns and a 3' UTR with a poly A-tail (see Figure 2.7). This raw transcript is called precursor messenger RNA (pre-mRNA) which then is further processed to the mature messenger RNA (mRNA). This modification or editing process encloses three steps: **5' capping**, **3' polyadenylation** and **RNA splicing**.

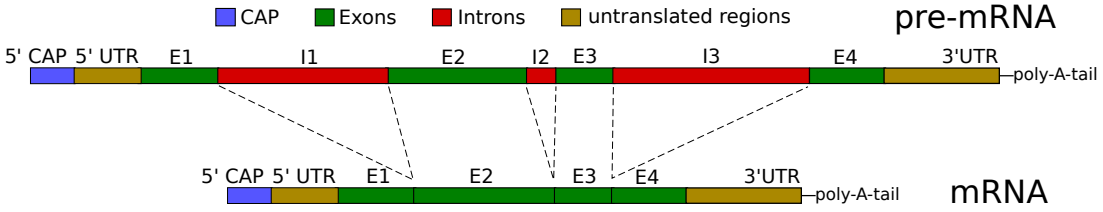

FIGURE 2.7: mRNA before and after the splicing process. The introns (I) get spliced out of the pre-mRNA and the exons (E) are fused together to build the mature mRNA, which is the template for protein synthesis.

2.3.1 The mRNA editing process

2.3.1.1 5' and 3' processing

The 5' capping is a co-transcriptional modification process on the 5' end of the mRNA. Hereby a guanine is added via an unusual 5'-to-5' triphosphate link to stabilize the RNA and prevent it from early degradation. Additionally it promotes translation, intron excision and regulates the nuclear export. Downstream of the capping region follows the 5' UTR containing the famous Kozak sequence, which plays a key role in the initiation of the translation process and also other regulatory elements [Kozak, 1986].

When the transcription of a gene is done, the 3' end of the RNA gets polyadenylated. In contrast to polyadenylation in bacteria the polyadenylation process in eukaryotes facilitates stability of the mRNA by saving the molecule from enzymatic degradation in the cytoplasm. After transcription termination a set of proteins cleaves the outer most part of the molecule, then the poly-A polymerase (PAP) adds approximately 250 adenosins to the 3' end. Poly-A binding proteins (PAB) then bind these nucleotides, prevent degradation and promote export into

the cytoplasm. In eukaryotes a short poly-A tail usually results in degradation and decreased translation efficiency [Guhaniyogi and Brewer, 2001].

2.3.1.2 mRNA splicing

Splicing is the process of removing non-coding sequences (introns) from pre-mRNA and joining the coding sequences (exons) together to obtain a mature mRNA molecule that can then be translated into protein. Because it is a co-transcriptional process it usually takes place in the nucleus of eukaryotes, but it is also found in different forms in mitochondria, archaea, and some viruses [Tilgner et al., 2012]. In higher organisms the most frequent variant is **spliceosomal splicing**, which uses a large complex called the spliceosome to remove intronic sequences from the pre-mRNA. More rare variants of splicing are **self-splicing** introns where RNAs form catalytically active ribozymes, and the **tRNA splicing**.

Introns recognized by the spliceosome are one of the defining characters of eukaryotes and are found in nearly every sequenced eukaryotic genome except in *Hemiselmis andersenii*, a cryptophyte [Lane et al., 2007]. The excision of a spliceosomal intron requires a large molecular complex termed the major spliceosome comprising 5 snRNAs and more than 150 proteins [Will and Lührmann, 2011]. This complex catalyzes the splicing reaction (Figure 2.9) of introns usually containing a GU at the 5', the donor-, and an AG at the 3', the acceptor splice site, which is preceded by a poly-pyrimidin tract (PPT) that acts as a binding site for spliceosomal subunits and recruits factors to the branch point sequence (BPS). The BPS contains a reactive A residue (Figure 2.8) that is important for lariat formation. Introns containing a GU and an AG at the 5' and 3' end respectively are dubbed canonical splice-sites. About 99 % of all known introns are spliced by this machinery mostly co-transcriptionally [Ng et al., 2004]. Besides canonical splice-sites, some eukaryotes also contain U12 introns, that are excised by another complex similar to the major spliceosome, called the minor spliceosome. Unlike the major one it is found outside the nucleus and splices out non-canonical introns [König et al., 2007]. A variant of the spliceosomal splicing process is termed intrasplicing or **nested splicing**, where the intron itself is shortened by consecutive splicing reactions and in a final step the complete intron can be removed from the pre-mRNA [Ott et al., 2003].

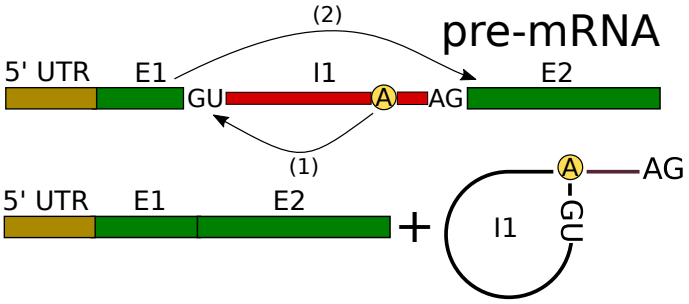

FIGURE 2.8: Spliceosomal splicing reaction: (1) the branch point adenine (yellow) in the poly pyrimidine tract attacks (PPT) the 5' splice site and the intron circularizes. In the next step the free 3' hydroxy group of the 5' exon attacks the 3' splice site (2), the exons fuse and the lariat gets released (bottom).

The splicing reaction is a highly regulated and complex process influenced by a lot of factors. The general thought is, the more complex an organism the higher the rate of introns in the genome. An average human for example has about eight introns per gene. Less convoluted organisms like yeast generally have relatively few and short introns [Sakharkar et al., 2004]. Some claim that spliceosomal introns are a non-negligible burden, because the organism needs to maintain the whole machinery participating in the splicing process. Lane and Martin state in their 2010 paper, that the energetic cost is probably tolerable but with an average elongation rate of 60 bases per second, the RNA polymerase II dependent transcription process is very costly in terms of time [Lane and Martin, 2010]. As a consequence, the splicing process must imply an evolutionary advantage for organisms to invest the resources.

The assembly of the spliceosome usually takes place at the transcription-site. It is a multi-step process mediated by the C-terminal-domain of the polymerase II as shown in Figure 2.9. In the initial phase the U1 subunit binds to the 5' splice-site of the pre-mRNA, U2 then is recruited to the branchpoint near the 3' splice-site where it interacts with U2AF (U2 auxiliary factor) forming the ATP independent commitment complex or early (E) complex. When U1 and U2 start interacting with each other they form the pre-spliceosomal complex, which is also called complex A, that depends on the ATP-driven helicases pre-mrna-processing 5 (prp5) and sub2. Subsequently the tri-snRNP consisting of subunit U4, U5 and U6 are recruited to form the pre-catalytic spliceosome (complex B). After internal rearrangements facilitated by a set of additional helicases, U1 and U4 are released from the complex leaving a catalytically active complex B* [Matera and Wang, 2014]. The splicing reaction requires a large quantity of ATP [Will

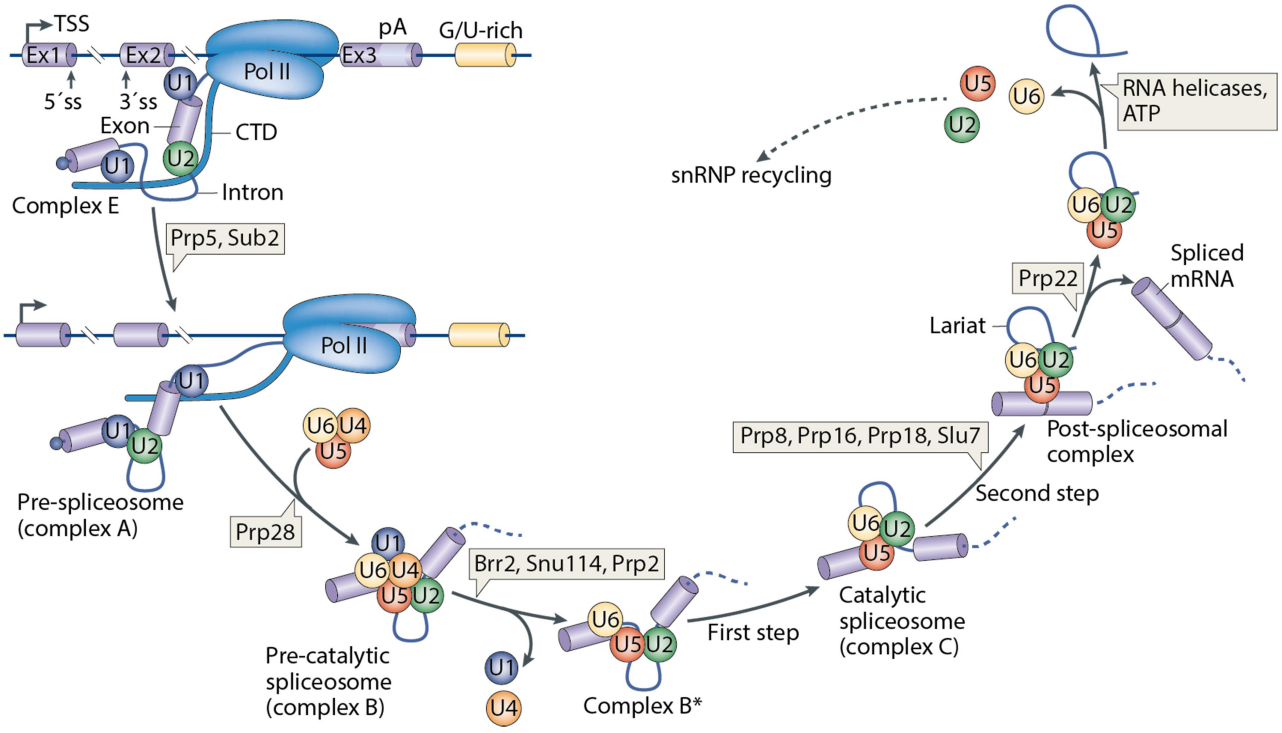

FIGURE 2.9: Figure depicts the spliceosomal cycle of the major spliceosome, starting with the initial splice site recognition by the two subunits U1 and U2 (complex A), followed by the formation of the catalytic complex B. Various proteins and structural rearrangements lead to spliceosome activation via two nucleophilic attacks, that cut out the intron. Afterwards the remaining snRNPs get recycled for further splicing steps. This figure was adapted from [Matera and Wang, 2014].

and Lührmann, 2011]. After the splicing process, the spliceosome is disassembled and the subunits are recycled.

The splicing process is mainly regulated via cis-acting regulatory elements on the pre-mRNA, which are recognized by certain proteins forming either an repressor or an activator. Depending on a motifs' location it can either act as an exonic splicing enhancer (ESE) or an intronic splicing enhancer (ISE). Motifs that inhibit the splicing process are called silencers and can be located in the exon (ESS) or the intron (ISS). These cis-acting regulators lead to alternatively spliced transcripts, which is another important hallmark of the eukaryotic splicing machinery.

Alternative splicing allows the synthesis of various products from only one gene. According to Lander et al. around 60 % of all eukaryotic genes undergo alternative splicing [Lander et al., 2001]. In Humans even 98 % of multi-exonic genes are alternatively spliced [Pan et al., 2008]. The most common variant

of alternative splicing in animals is the exon skipping event (see Figure 2.10), where under certain circumstances a particular exon is included in the mRNA and bypassed in another mRNA [Zhang et al., 2015].

An impressive example for alternative splicing is the DSCAM homologue in *Drosophila*. This gene can generate more than 38,000 isoforms from only one pre-mRNA transcript [Schmucker et al., 2000]. Besides changes in gene expression, alternative splicing also effects elements controlling translation, stability, and the localization of mRNA.

To what is known, not only enhancers and silencers influence the splicing patterns. Kristi Fox-Walsh et al. showed in 2005, that the efficiency of splicing also depends on the length of the introns [Fox-Walsh et al., 2005]. Exons flanked by short introns with a length of about 200 nt maximum are mostly recognized across the exon, if they are flanked by longer introns, splicing factors gain more weight [Sterner et al., 1996]. Splice-sites that are recognized across the intron are significantly more efficient and result in enhanced incorporation of exons with weak splice-sites. The intron also influences the probability that an exon is alternatively or constitutively spliced. For example, in *Drosophila* exons preceded by a long intron have an up to 90 times higher probability of being alternatively spliced than exons flanked by short introns.

Because alternative splicing is highly regulated, mutations in introns or regulatory sequences can result to unpredictable consequences. A modified splicing motif may lead to reduced specificity or activation of cryptic splice-sites, which again could cause insertions, exclusions, deletions, changes in reading-frames and in consequence aberrant mRNAs and proteins. Despite sophisticated quality-assurance processes like nonsense-mediated decay (NMD) around 50 % of all genetic disorders in humans are caused by abnormal splicing [López-Bigas et al., 2005] [Wang and Cooper, 2007].

A well known example for splicing-dependent disorders are beta-thalassemias. These thalassemias are a group of inherited blood-cell disorders, mostly prevalent in Mediterranean people, leading to massive forms of anemia. A defect in the HBB-gene interferes with the synthesis of hemoglobin-β-chains, which leads to an excess of γ- and δ-globins and destabilizes erythrocytes. This defect, originally a selective advantage in malaria-prone regions, is often caused by activation of

cryptic splice-sites and mutations in splicing motifs, that lead to aberrant mRNA-sequences. [Cao and Galanello, 2010].

In the early 80s T. Chech and his team revealed that some RNAs have special chemical features that allow these RNAs to be processed without protein-dependency [Cech and Bass, 1986]. They discovered so called self-splicing introns, which are catalytically active RNA-sequences that form ribozymes. In 1989, Chech was accredited the Nobel-prize for his findings, together with Sidney Altman. Self-splicing RNAs are found in a wide range of organisms and can be split into two major groups.

Group I self-splicing introns are found in the mRNA of chloroplasts and mitochondria as well as in rRNA of lower eukaryotes. In this two step process an active G-group binds to a specific sequence near the 3' end of the RNA and attacks the phosphodiester bonds on the 5' end of the intron. Then, the newly developed 5' OH acts as a nucleophile and attacks the 3' splice site. The exons are interconnected and the intron is set free. In **group II self-splicing introns** on the other hand an active A in the sequence is the attacking group and forms an unusual 5'-2' bond with the 5' end of the intron. Afterwards the 3' end of the upstream exon attacks the 3' splice site of the intron, the exons fuse and the intron is released in a lasso like structure called the lariat. These introns are found in tRNA, rRNA and mRNA of chloroplasts and mitochondria. Because of the similarity of the splicing reactions of group II introns and pre-mRNA splicing, it has been proposed that the spliceosome originated from group II self splicing introns [Rogozin et al., 2012].

Although introns are non-coding regions, they do not consist only of "dead" sequences. Because they are not under the same selective pressure as exon, they harbour various, even active, sequence elements, e.g. SINEs.

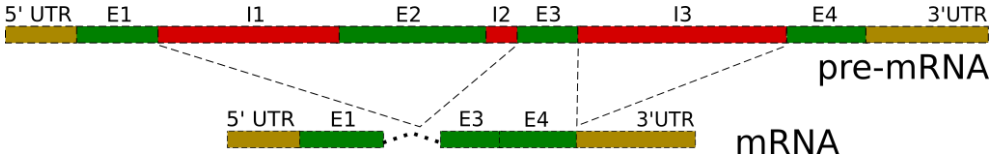

FIGURE 2.10: Example of an exon skipping event. The region starting at intron 1 (I1) until the end of I2 also including exon number 2, is not included in the major RNA. 5' cap and poly-A-tail not shown.

2.3.2 SINEs

SINEs (short interspersed elements) are a class of retro-transposable elements that often amplify themselves through RNA intermediates. The largest subclass of SINEs is the Alu-repeat family [Schmid and Deininger, 1975]. Alus account for about 11 % of the whole human genome and hence, are the most abundant transposable elements found in humans. With an usual length of around 300 bp there are more than one million elements spread throughout the whole genome [Lander et al., 2001]. Alu-sequences usually consist of two homologous parts on their 5' and their 3' sites that are flanked by direct repeats of variable length which probably are artifacts of duplication. They have an internal RNA polymerase III promoter but no transcription terminator and are mainly found in non-coding segments of gene rich R-bands [Versteeg et al., 2003]. L1 elements (a subclass of LINEs, long interspersed elements) on the contrary are commonly enriched in gene poor regions. Originally, Alus are derived from the small cytoplasmic 7SL RNA, probably at the clade of supraprimates, and have a relatively high sequence homology among their subfamilies [Kriegs et al., 2007]. Therefore, Alus are also used to get insights into the evolution of primate and humans.

Around 5-10 % of all mRNAs contain Alu elements in their 3' ends. Because they are transposable elements there is a chance that every time an element inserts near or directly in a gene it may influence the expression of that gene and under specific circumstances become a regulator of that gene [Deininger, 2011].

It has also been found that Alus provide transcription factor binding sites and that sequences located in exons (Alu exonization) influence alternative splicing [Polak and Domany, 2006] [Sorek, 2002]. If located in introns, pairs of inverted repeats can fold into secondary structures leading to ADAR (adenosine deaminases acting on RNA) activity that converts adenin to inosine by deamination (see Figure 2.11) and can result in retention in the nucleus or activation of cryptic splice-sites [Chen et al., 2008] [Nishikura, 2010].

The transposable activity of Alus, about one new Alu per 200 births, leads to genetic diseases in nearly one out of ten thousand persons. Alu elements are providing one of the most common sources of sequence homology for non-allelic recombination events that could lead to instability of the genome, misregulation, and disruption of coding regions or splicing signals [Deininger and Batzer, 1999] [Hedges and Deininger, 2007].

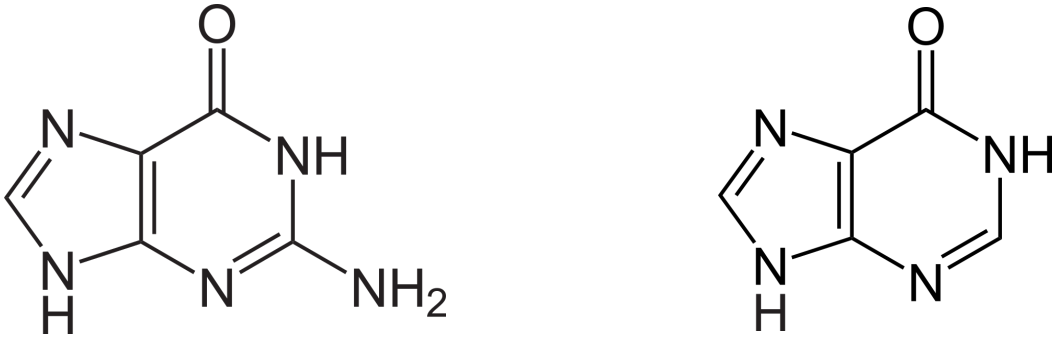

FIGURE 2.11: Structure of guanine on the left and of inosine on the right. The only difference is the missing NH_2 - group on the 2' position. ADARs convert adenin to the guanine related inosine which can lead to changes in the nucleotide sequence.

Chapter 3

Methods

3.1 Overview

This section gives an overview over the methods used in this work, starting with the extraction of the data, followed by the RNA-RNA-interaction search and finally to the cross-referencing of possible long-range interactions with known repeats and disease data. If not mentioned otherwise, all tools were part of the *Vienna RNA package* (version 2.1.9), a toolbox for prediction, comparison, and visualization of RNA secondary structures developed at the TBI Vienna [Lorenz et al., 2011].

3.2 Data preparation and filtering

As a first step the human comprehensive-gene-annotation-file release number 19 (July 2013) (gencode.v19.annotation.gtf.gz) was downloaded from genecodegenes.org [J et al., 2012]. The start- and end-coordinates of annotated exons were extracted and used to infer the introns inbetween. Redundant intron information from alternatively spliced transcripts was filtered out by comparing the coordinates (chromosome, start- and end-points) of the introns.

In the next step the nucleotide sequences of the human genome were downloaded from gencode in fasta-format (GRCh37.p13.genome.fa.gz) and the DNA-sequences corresponding to the intron-coordinates were extracted.

3.3 RNA-RNA interaction search

To prepare the data for interaction search, the introns were split either in half if the sequence length was shorter than 4,000 nt or if longer, fragments of 2,000 nt were taken from each, the 5' and the 3' end, resulting in two sub-sequences with a length between 200 and 2,000 nt (see Figure 3.1) termed the 3' and the 5' fragment. For each of these fragments the accessibility-profiles were computed and transformed to opening-energies using *RNAplfold*. These profiles were then used together with *RNAplex* to find the best interactions between each of the two fragments.

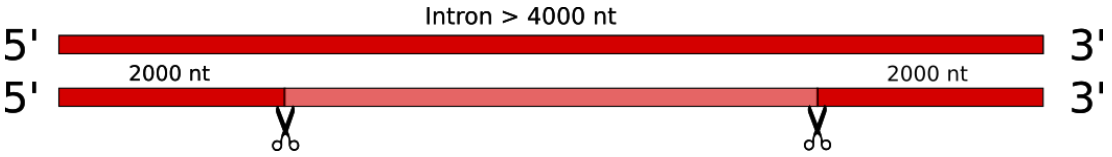

FIGURE 3.1: If the introns are longer than 4,000 nt only fragments of 2,000 nt were taken from the 5' and the 3' end. The light red part in the middle of the intron was neglected. Sequences shorter than 4,000 nt were cut in the middle yielding two fragments with the same length.

3.3.1 Accessibility calculation

The basepair accessibilities were calculated using *RNAplfold* from the ViennaRNA package [Lorenz et al., 2011]. To get rid of so called border effects as described by Lange et al, the window size over which the pair probabilities are averaged (-W) was increased to 50 nt larger than the maximum basepair span (-L) [Lange et al., 2012]. Because *RNAplex* requires opening energies to simulate intra-molecular base-pairing, *RNAplfold* needs to transform the probabilities of region $[i, j]$ being unpaired into opening energies (see Equation 2.4). This is done by invoking *RNAplfold* using option -O. Additionally the parameter -u was set to 5 as suggested by the author of *RNAplex* to calculate the mean probabilities of region $[i, i + 5]$ being unpaired. The output was saved as binary file (-b) to speed up the read search done by *RNAplex* in the next step. The fasta files containing the 5' and 3' sequence fragments were generated with by a perl script that cuts the sequences as described above and marks the data with an unique id.

Following command shows the invocation of *RNAplfold* with the parameters described above:

```
RNAplfold -O -b -u 5 -W 200 -L 150 --plex_output < TARGET.fa
RNAplfold -O -b -u 5 -W 200 -L 150 --plex_output < QUERY.fa
```

3.3.2 Intermolecular interactions

For the interaction evaluation process *RNAplex* from the Vienna RNA Package release 2.2.4 was used. This version of *RNAplex* uses a slightly different backtracking approach than the version shipped with package release 2.1.9, where it first finds the end of the best interactions and then uses this mark to trace back the interaction start. In contrast to the older version no fixed maximum interaction length has to be defined that caused an error in energy calculation when swapping target and query sequences. If swapping the sequences still yielded different interaction energies and/or interaction coordinates, the one with the lower energy was considered.

Here *RNAplex* was used to find the best interactions between 5' and 3' fragments of the introns. Therefore, the pre-computed binary accessibility-files generated with *RNAplfold* were taken as additional input to account for intra-molecular interactions. Options -a and -b set the directory for the pre-calculated binary files. The duplex distance -z was set to 0 and for the backtracking the approximated *RNAplex* energy model (option -f 2) was used. The input consisting of the 5' and the 3' fragments was supplied in two fasta files (option -t TARGET, -q QUERY). The program was called two times, with each sequence once as target and once as query sequence.

```
RNAplex -b -a . -f 2 -t TARGET.fa -q QUERY.fa
RNAplex -b -a . -f 2 -t QUERY.fa -q TARGET.fa
```

In an additional approach, the 2,000 nt regions on the 5' and 3' fragments were extended by one window length of *RNAplfold* into the exonic part to get more exact accessibilities for the first bases of the intronic sequence and avoid possible border-effects [Lange et al., 2012]. After the accessibility calculation these additional nucleotides where substituted by Ns to act as constrains for *RNAplex*.

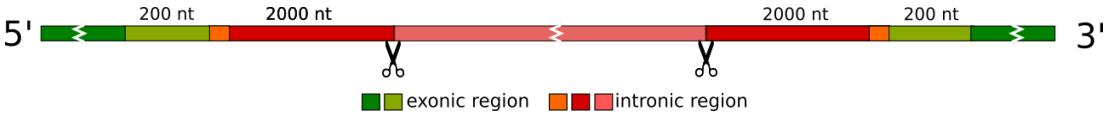

FIGURE 3.2: Depiction of the sequence extension into the exonic part. The 5' and the 3' region of the intron are extended by a 200 nt stretch into each direction, to counteract so called border effects in accessibility calculation. The green section marks the exonic region, the shapes of red the intronic region. Light green indicates a 200 nt long (equals to one window length in RNAplfold) extension of the 2,000 nt long intronic regions (red), to avoid border effects due to accessibility computation [Lange et al., 2012]. Light-red illustrates the part of the intron that is not accounted for in the calculations. The orange part refers to the 50 outermost bases on the 2,000 nt flanking regions.

In another experiment everything except the first 50 nt of the 5' fragment and the last 50 nt of the 3' fragment were shuffled (Figure 3.3) and the best interactions predicted, following the same steps as shown above, to see if these regions influence the accessibility prediction of the whole structure or the probability of favourable interactions in the outermost parts of the intron. In this experiment only introns larger than 3,999 nt were used.

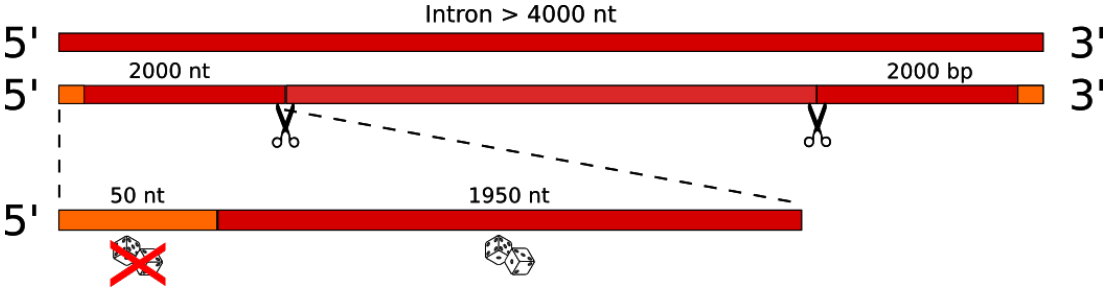

FIGURE 3.3: For introns larger than 4,000 nt a 2,000 nt region on the 5' and the 3' ends were cut off and then everything but the outermost 50 nt were shuffled using k-lets of two to see if and how these regions affect possible long-range interactions.

Interaction binning

When interaction binning is used, the interaction distance is represented by assigning the start on the 5' region and the end of the predicted interaction on the 3' region to a corresponding bin. Therefore, each fragment of 2,000 nt is split into five bins with a width of 400 nt each. If the interaction ranges from [250, 410] on the target strand and [1300, 1460] the ends fall into bin 1 and bin 4, which is represented as 1.4 in according plots.

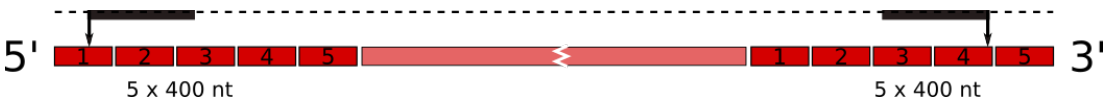

FIGURE 3.4: Each fragment with a length of 2,000 nt gets split up into five bins of 400 nt. Then, the interaction start-point on the 5' and the interaction end on the 3' fragment is allocated to an according bin to visualize the interaction distance. Here, an interaction between coordinates [210, 900] on the 5' and [1180, 1590] on the 3' end is formed.

3.4 Background model

To evaluate the statistical significance of the intron interactions, a background model is needed. Here, a randomized model was used to calculate the z-score of the native interactions and to compare them to the scores of a shuffled sequence, with the idea in mind, that if there are certain long range interactions between the 5' and the 3' ends of the introns, these patterns or affinities get disrupted by randomizing the sequence but keeping the GC-content.

In secondary structure prediction the stability of a structure depends substantially on stacking energies of adjacent basepairs, therefore one should conserve doublets to yield correct energies. To conserve the di-nucleotides, an in-house shuffling module based on the algorithm by Ericson and Altschul was used [Ericson and Altschul, 1985]. The native sequence fragments were taken and shuffled 150 times each (see section 3.4). Subsequently the interactions between the 5' and 3' regions of the sequences were evaluated using *RNAplfold* and *RNAplex* as described above. The randomized dataset was then used for for the z-score calculations.

k-let shuffling

At first a directed multi-graph of k-lets was created and stored in a perl hash-table where the keys consisted of the k-lets, the values contained the edges to neighbours. Then the neighbours where re-picked in a random order, a tree was generated, and the newly shuffled sequences with conserved k-lets were generated by defining an Euler-path, that includes each vertex. The k-let shuffling script was implemented as an easy-to-use perl module and used to create the background model.

z-score calculation

The z- or standard score was used to compare the interaction-energies of the native model and the background model. This score is a dimensionless positive or negative value obtained by subtracting the mean of the population (μ) from a certain score (x) and dividing this through the standard deviation (σ) yielding a normalized score, that shows the deviation of a value from the mean (see Equation 3.1).

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

Here μ is defined as the mean interaction energy over each of the shuffled sequences derived from a specific native one and x being the actually best interaction energy calculated with RNAplex from the native dataset. For z-score calculation of the randomized model x was defined as a randomly picked sequence from the shuffled data. σ is the standard deviation of the interaction energies from the shuffled sequences. Calculations were done with a custom made python script and the resulting data plotted with the R-package ggplot2.

3.5 Extracting repeat coordinates

The information about repeats were retrieved via the UCSC table browser (http://genome-euro.ucsc.edu/cgi-bin/hgTables). Coordinates were downloaded as *.gtf file from the group "repeats" human assembly GRC37/hg19. The "repeatMasker" track was selected and depending on the repeats needed, the filter function was used to select the according repeat class, e.g. "alu". In the next step these repeats were intersected with interaction data.

3.6 Intersection of dataset by coordinates

To find the intersection of the two datasets, namely the calculated interactions between 5' and 3' set and the repeats, a nice collection of tools from Guinlan and Hall called *BEDtools* version 2.25.0 was used [Quinlan and Hall, 2010]. This genomic tool-kit contains a program called *intersectBed*, which allows one to

find the intersections between datasets according to their coordinates with an optional minimal and/or a maximum overlapping-threshold function.

3.7 Visualization of RNA secondary structures

For the visualization of RNA secondary structures *forna* an online visualization toolbox was used [Kerpedjiev et al., 2015]. This application relies on Bruccolerie's NAVIEW algorithm implemented in the ViennaRNA package and additionally administers the d3.js javascript visualization library to apply a force field [Bruccoleri and Heinrich, 1988]. The *forna* webapp can be found at http://nibiru.tbi.univie.ac.at/forna/.

Chapter 4

Results

In this work the human genome was systematically screened for long range interactions in intronic RNA sequences with the assumption, that interactions between the 5' and the 3' end probably bring these regions in spacial proximity and may influence the splicing process in long introns. If not mentioned otherwise, all analyses are based on introns longer than 4,000 nt extracted from the human gene-annotation-file release number 19 (hg19, Juli 2013) from the gencode database [J et al., 2012].

4.1 Intron length distribution

At first, we analyzed the overall length distribution of the extracted introns to get an overview of the available data. Therefore, the introns were extracted from the genome annotation file and redundant sequences were removed from the dataset as described in the methods section (section 3.2). Figure 4.1 gives an overview of the intron-length-distribution, with many introns shorter than 1,000 nucleotides and relatively few long introns. The number of introns is declining the longer the sequences are. A closer analysis revealed that sequences between a length of 100 to 400 nt represent about 16 % or 51,666 out of 331,109 sequences. Sequences with a length between 400 and 3,999 nt cover around 60 % of the whole dataset whereas in total there are 110,930 introns longer than 4,000 nt. Introns longer than 60,000 nt make up only 3.4 %. In total we find 331,109 intronic sequences longer than 100 nt and 279,443 sequences with at least 400 nt. The shortest

intron found was 1 nt in length the longest about 1.55 Mbp according to the gencode annotation file. A more detailed breakdown can be found in Table A.1.

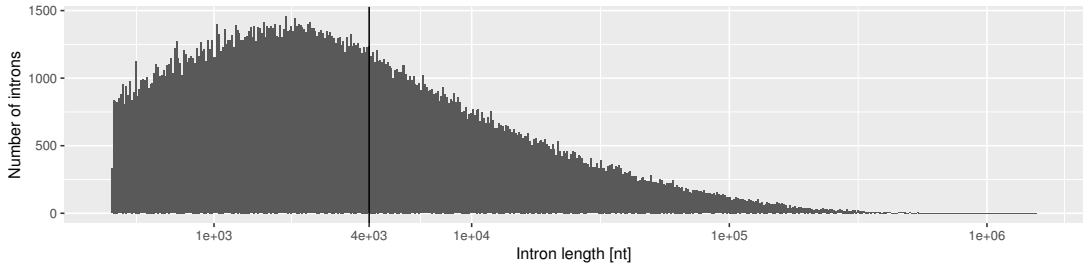

FIGURE 4.1: Intron-length-distribution of extracted introns from the gencode hg19 dataset. The dataset consists of a lot of short sequences and relatively few sequences longer than 10,000 nt. The black vertical line at 4,000 nt marks the lower length-boundary for introns used in interaction-experiments. Note, that only introns with a minimal length of 400 nt are shown.

4.2 Evaluation of interaction energies

After getting acquainted with the general length-distribution of the introns, we checked if there are indications that introns can form energetically stable interactions between their 5' and their 3' end and compared these interactions to a background model (described in section 3.4). If the quality of the background model were similar to the native model it would indicate that these introns formed some kind of randomized interactions. If they differed it could be an indication that there were certain non-random structures formed between the ends. Here, the 5' and the 3' regions of introns longer than 400 nt were evaluated using *RNAplfold* and *RNAplex* as described in section 3.3. These tools were applied to predict the most favourable interactions, which in consequence is the interaction with the lowest free energy ($\min \Delta G$) a sequence can fold into given a certain condition. For the background model, the sequences were shuffled and evaluated following the same protocol as for the native data. Subsequently, the z-scores were computed and compared (see Figure 4.2). The z-score indicates the number of standard deviations σ an interaction is better (lower value) or worse (higher value) than the mean predicted energy for a certain sequence. Figure 4.2 depicts the z-score distribution of the native (black) and the randomized dataset (grey) with a conserved flanking region of 50 nt. The distribution of the unshuffled data indicates a pronounced left skewed distribution with values going down as low as -950.8 and up to $+5.4$ compared to the more narrow distribution of the shuffled

data with values between −7.0 and +3.6. A basic statistical comparison of the analysis can be found in Table 4.1. A Wilcoxon-Mann-Whitney test of the two distributions resulted in a p-value $< 2.2^{-16}$ supporting the claim that the native interactions between the 5' and the 3' regions of introns are more stable than randomized sequences.

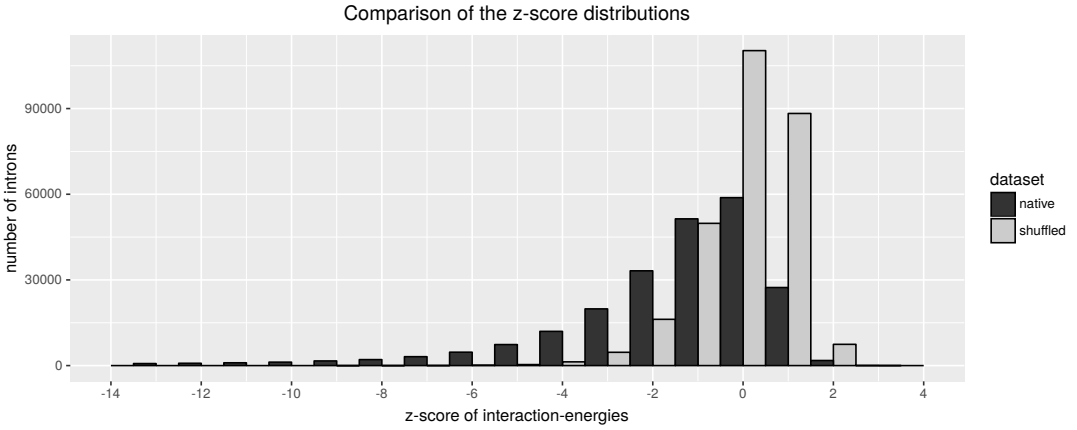

FIGURE 4.2: Snapshot of the interaction-energy z-score distribution of both, the native (black) and the shuffled (grey) dataset. The x-axis shows the z-scores binned in intervals of 1, on the y-axis the number of introns falling into these bins is shown. For better visualization only values with a score between -14 and +4 are shown. A short statistical summary of the datasets is shown in Table 4.1.

	native model	background model
Minimum	-950.7614	-6.9620
1st Quantile	-5.1468	-0.4593
Median	-1.5068	0.0529
Mean	-13.7331	-0.0011
3rd Quantile	-0.1996	0.6558
Maximum	5.4457	3.6220

TABLE 4.1: Basic statistical comparison of the z-score distribution of interaction-energies of both the native and the background model. By definition of the z-score, the median- and mean-value of the reference data, the background model, has to be close to 0.

The median- and mean-value of the randomized dataset by definition is close to 0 whereas the native data hold a median-value of −1.5 and a mean-value of −13.7. It is interesting to see that a fair number of the native sequences form energetically more stable interactions than their shuffled complement (Figure 4.2).

This may be an indication of intron-structuredness, that gets lost in the shuffling process and needs to be investigated further.

4.2.1 Analysis of predicted intron interactions

Figure 4.2 suggests that a substantial amount of introns form more stable interactions between their 5' and their 3' region than shuffled sequences. The position of these interactions may indicate that there are certain regions more prone to interact with each other. These regions may be sequential or structural motifs known to be found in a specific distance from the intron start.

Therefore, introns longer than 4,000 nt are cut into two 2,000 nt long pieces (see section 3.3), a 5' and a 3' fragment. Each fragment has an additional 200 nt long stretch, that extends into the exonic area and is used to get rid of possible bias introduced by the accessibility-calculations with *RNAplfold* as described in Lange et al. [2012]. Subsequently, *RNAplex* is used to find only interactions in the intronic region, while having the additional accessibility-information of the extended regions.

Figure 4.3 shows histograms of interactions start- (A, C), end- (B, D) and mid-points (E, F) on each fragment. The y-axis always marks the interaction count, the x-axis depicts the appropriate start-, mid- or end-point on a nucleotide based scale. Unexpectedly, we find an increased number of introns starting and ending in the most outer parts of the introns. An extreme example can be found in Figure 4.3 B where the number of interactions ending in the last bin (between 1990 and 2,000 nt) is about six times higher than the average, preceded by a small decrease in numbers. In Figure 4.3 A we also find an increased number of interactions starting in the first few bins, though less pronounced than on the 3' fragment (Figure 4.3 C). The number of start-points seems to decline at around 1750 nt, whereas the interaction-mid-points are nearly evenly distributed with small valleys around 200 nt and around 1800 nt (Figure 4.3 E). The mid-points on the 3' fragment (Figure 4.3) F show a general decrease in numbers starting around 1750 nt.

Next, we compared the native dataset from Figure 4.3 to the randomized data (Figure 4.4) to check for differences in the distributions that could suggest a more narrow analysis of certain regions. As already mentioned, only sequences longer

than 4,000 nt were used, cut into two parts and each fragment was extended by 200 nt into the exonic area. The extended 200 nt long regions were shuffled independently (see section 3.4) to avoid a bias from the differing basepair-composition of the coding- and non-coding-areas. The evaluation was done using the same tools as before.

Again, we can see an increased number of interactions starting in the first few bins (Figure 4.4 A, C) and a decrease of start-points after around 1750 nt. The number of interactions ending in the last bins are also above average, though less pronounced than in Figure 4.3. Surprisingly, histogram Figure 4.4 D does not show increased numbers of end-points near the last bin. The mid-points on the 5' fragment (Figure 4.4 E) on the other hand show a slight peak at the beginning and then leveling back to the average. On the 3' end (Figure 4.4 F) we see a less abrupt decline in mid-points near the end compared to the native data.

Although the start- and end-regions of the introns were extended to get the additional opening energies for the regions in the beginning, we observed increased

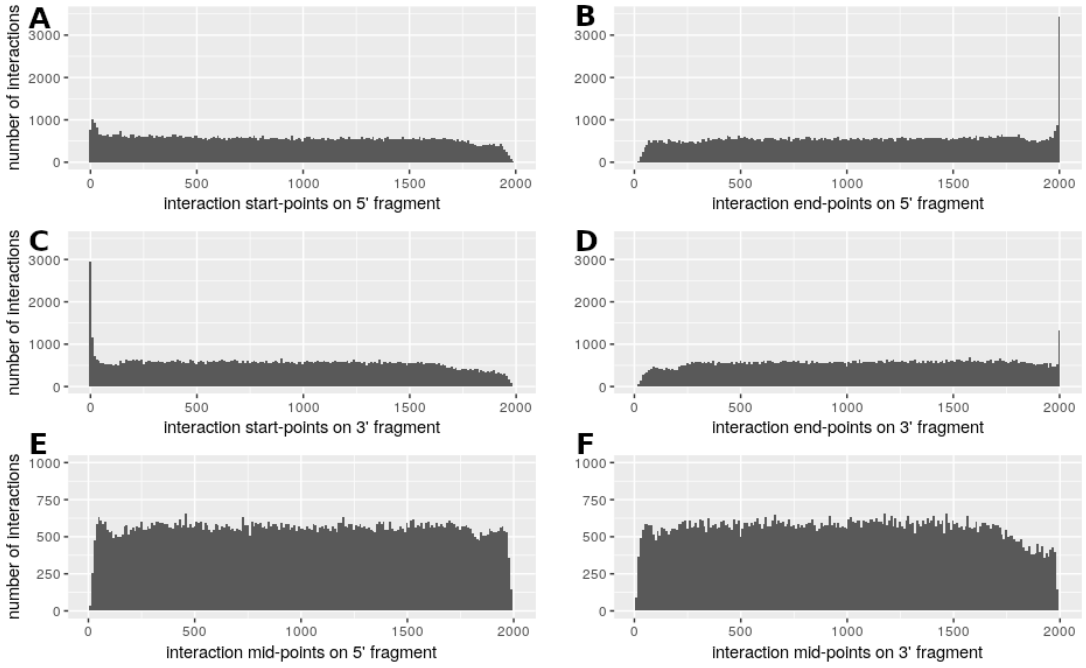

FIGURE 4.3: Overview of start- (A, C) and end-points (B, D) of predicted interactions as well as the mid-points (E, F) in native intronic sequences longer than 4,000 nt, where each fragment is exactly 2,000 nt + 200 nt long. Each sequence was extended by 200 nt into the exonic region for accessibility-calculations with *RNAplfold*. Here, the interactions were evaluated using *RNAplex* with the constraint, that pairing is only possible within the intronic bounds. For representation the binwidth is set to 10 nt.

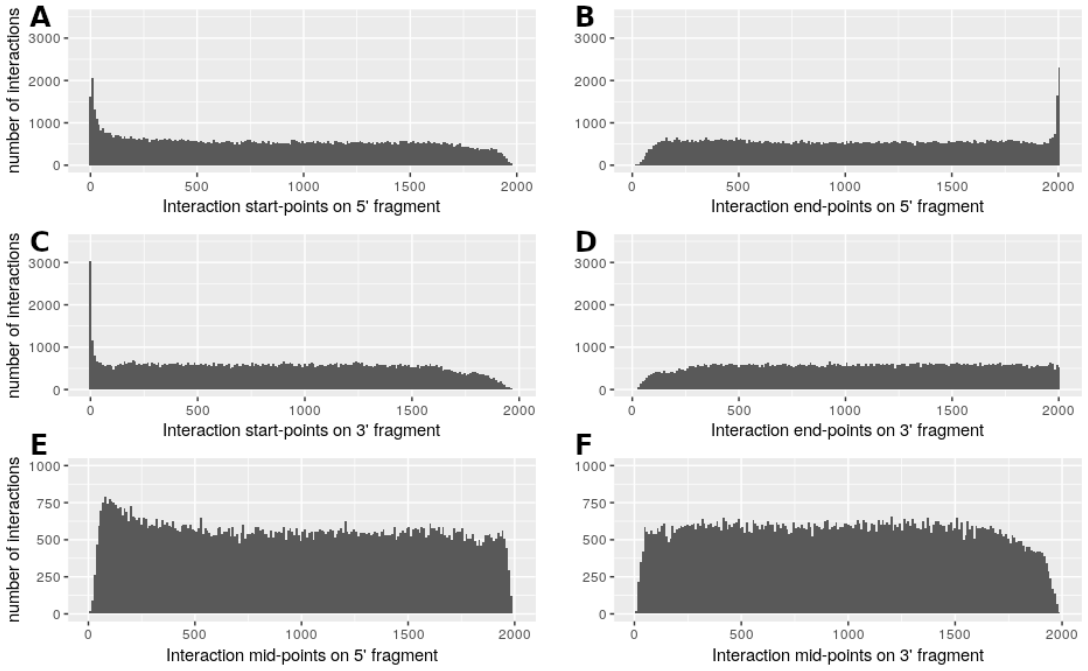

FIGURE 4.4: Overview of start- (A, C), end-points (B, D) an mid-points (E, F) of predicted interactions in di-nucleotid shuffled intron-sequences longer than 4,000 nt, where each fragment is exactly 2,000 nt long. Additionally, each sequence was extended by 200 nt into the exonic region for accessibility-calculations with *RNAplfold*. In this plot the interactions were evaluated using *RNAplex* with the constraint, that pairing is only possible within the intronic bounds. The binwidth is set to 10 nt.

numbers of interactions start- and end-points at the respective end, which still looked like border-effects. To see, if these accumulations are interactions with a span of nearly 2,000 nt or artifacts from the calculation, the start- and end-position were randomly chosen with regard to the interaction-length Figure 4.5. Here, we did not find any indications of an aggregation of start- or end-points near the outermost positions.

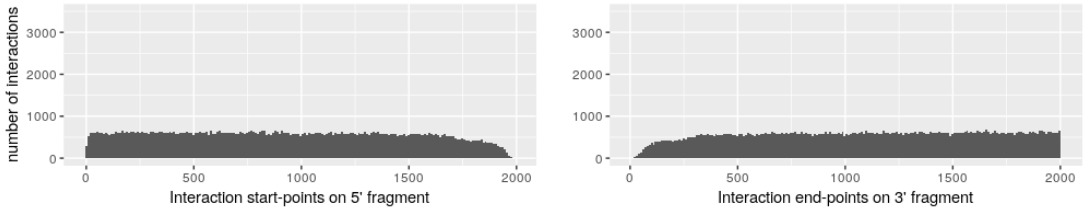

FIGURE 4.5: Randomly chosen interaction start- and end-points of the native data with respect to the interaction length. In contrast to Figure 4.3 and Figure 4.4 we don't find peaks at the outermost parts of the sequence.

Next, we investigated if some regions on the 5' end are more prone to bind to a specific region on the 3' fragment or at least show some preference of paring

within distinct intervals from both of the ends. Therefore, the two 2,000 nt long pieces without the 200 nt extensions where divided into five intervals of 400 nt in length. Then the interaction start coordinates on the 5' sequences were assigned to a bin, e.g. an interaction that starts at position 410 will be assigned to bin number 2. Analogue steps were taken for the interaction-endpoints on the 3' fragment on all interaction sites (see section 3.3.2).

Figure 4.6 indicates the binned interaction start- on the 5' and the end-points on the 3' part which represent the interaction span. When comparing both datasets one can see that the shuffled data is more evenly distributed than the native data. In the heatmap representing the native set (Figure 4.6 left) we see a lot of interactions starting in interval-1 on the 5'- (5'-1) and ending in interval-5 on the 3' (3'-5) end. Interactions starting on 5' interval-1 and ending on 3' interval-5 span the widest, interactions starting in 5'-5 and ending in 3'-1 have the smallest span, the ends are further apart.

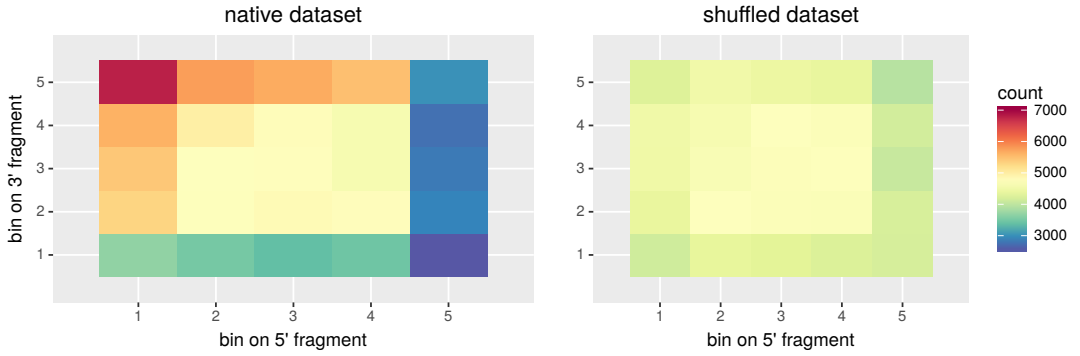

FIGURE 4.6: Comparison of the interacting bins in the native and the shuffled dataset. Each part is 2,000 nt in length and splits into intervals of 400 nt named 1 to 5. The x-axis depicts the interactions that start in a certain interval on the 5' fragment, on the y-axis the intervals on the 3' fragment are shown. The shuffled dataset on the right shows evenly distributed interactions in all bins compared to the native set, which seems to prefer long interaction-spans (5'-1 with 3'-5) over short ones (5'-5 interacting with 3'-1).

The long spanning interactions from the native model seem interesting, but one has to keep in mind that the majority of the native interactions are also longer than the ones in the shuffled set. Long interactions can therefore not start or end in every region of the 2,000 nt long fragments and in consequence are biased by the length of the interactions. By looking at the mid-points of the interactions (Figure 4.7) we can reduce this bias, but cannot get rid of it completely. In Figure 4.7 we find, that in the native sequences the interaction-mid-points are enriched near the middle of the fragments. It also clearly indicates that very few

interactions have their mid-points somewhere in the last 400 nt (5'-5) of the 5' fragment, whereas in the shuffled data the least populated bins are associated with the last 400 nt (3'-5) on the 3' fragment. In general the shuffled data seems more evenly distributed than the interactions from the native set.

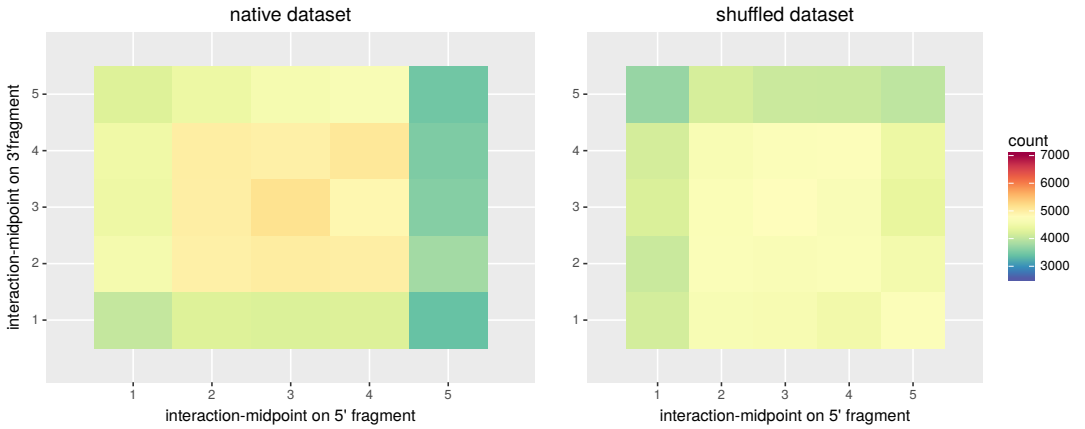

FIGURE 4.7: Comparison of the interacting regions in both, the native and the shuffled dataset based on the interaction mid-points. Every fragment is 2,000 nt in total and split into intervals of 400 nt named 1 to 5. The x-axis shows the midpoints of the interactions in a certain interval on the 5' fragment, the y-axis shows the mid-points on the 3' fragment.

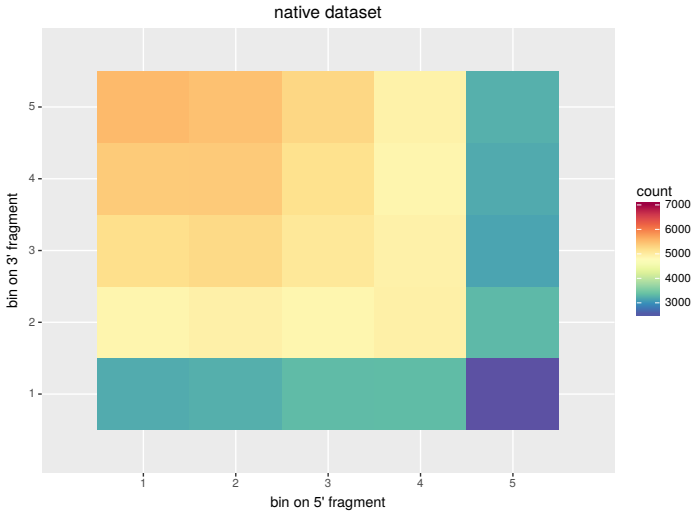

FIGURE 4.8: Distribution of randomly chosen interaction sites on both 2,000 nt long fragments by keeping the interaction length from the native interaction. Each fragment of 2,000 nt is split into 5 intervals á 400 nt and the start of the interaction on the 5' fragment (x-axis) as well as the end of the interaction on the 3' fragment (y-axis) is assigned to a bin.

If the interaction start-coordinates from the native set were randomly chosen and the length of the original interactions is kept, a pattern similar to Figure 4.6 appears. Figure 4.8 shows an over-representation of wide-spanning interactions

in the left upper edge (5'-1, 3'-5), which are enriched by a factor of at least two compared to the shortest-spanning interactions from bin 5'-5, 3'-1. Similar to Figure 4.6, there are relatively few interactions starting in interval 5'-5 and ending in 3'-1.

To investigate the contribution of the interactions from the smaller peak (found in Figure 4.11), with lengths between 270 and 340 nt (termed the 300-peak), in Figure 4.9 we separated the two datasets to view them independently. Here, the heatmap without the interactions of the 300-peak shows that the bigger part of the interactions are located at the beginning of the 5' fragment and end on the outer edge of the 3' fragment. Examining the interactions of the 300-peak separately we see an accumulation of interactions between the start of the 5' and the end of the 3' fragment as before, but also interactions near the end- and the beginning of the 5' and 3' part, respectively.

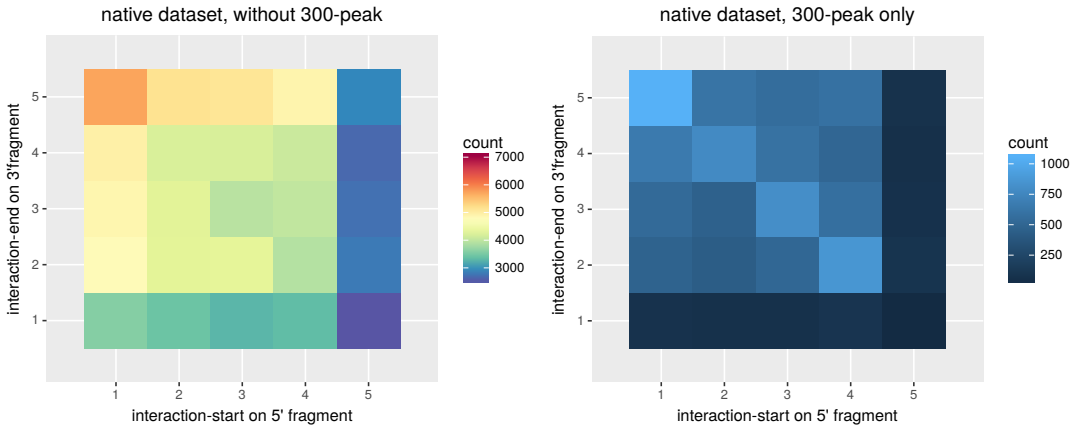

FIGURE 4.9: Interacting regions on the 5' and on the 3' fragment of the native dataset. Each fragment of 2,000 nt is split into 5 parts á 400 nt and the interactions are assigned to the corresponding intervall of their start- and endpoints. The x-axis shows the interaction start-points on the 5' fragment and the y-axis the interaction end-points on the 3' fragment. On the left we find the native dataset, without interactions with a length between 270 and 340 nt (the 300-peak) and on the left the 300-peak interactions exclusively.

The z-score distribution as shown in Figure 4.2 measures the number of standard-deviations of the native data diverging from the background-model. It is not an ideal indicator for quality or stability of the interactions. The minimum free energy (MFE) is much more suited to compare the strength of the interactions. According to Figure 4.10, it seems like the native sequences (red) form longer and more stable interactions than the shuffled ones (blue), which supports the

assumption that, though non-coding, there is some intrinsic structure in the native intron-sequences that gets lost when shuffled.

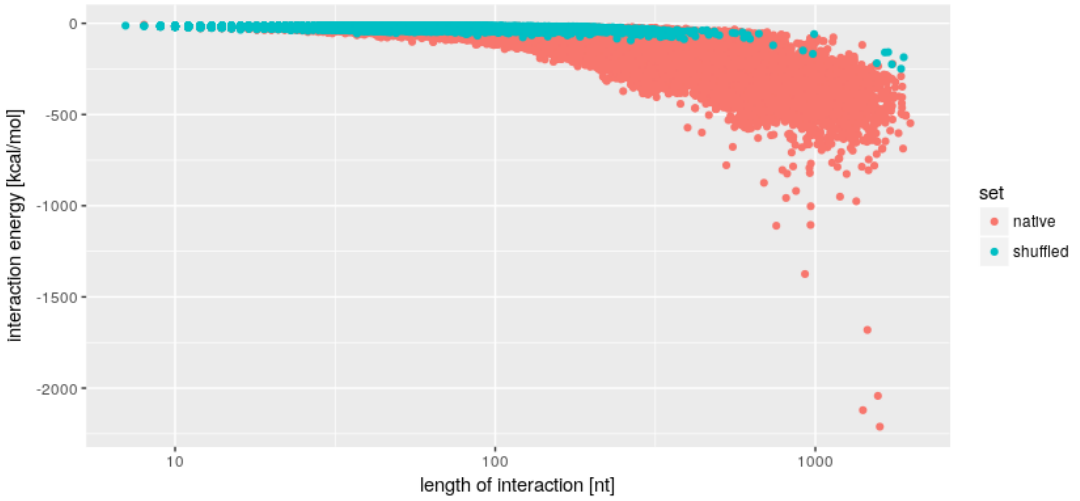

FIGURE 4.10: This figure shows the comparison of the two datasets with regards to the interaction-energy ΔG in kcal/mol on the y-axis and the length of the interaction on the x-axis. The lower the energy, the more stable the interaction.

4.2.2 Analysis of the interaction length-distribution

According to Figure 4.6 there are increased numbers of interactions between the beginning of the 5' fragment and the end of the 3' fragment. For one thing, this effect may be caused by many very long interactions and needs to be examined more closely. Therefore, the length of the predicted interactions is determined by subtracting the interaction-start from the -end position on the 5' segment, the distribution is shown in Figure 4.11. In the native data we observe a right skewed distribution with a lot of short interactions and a relatively rare amount of long interactions. In the random model on the other hand the distribution looks more narrow and less skewed with a mean interaction-length of about 41 nt and a median of 34 nt compared to a 103 nt and 49 nt of the native data for introns longer than 400 nt.

In the native model we find 205,483 or 73.5 % from a total of 279,308 introns longer than 400 bp forming interactions with less than 100 nt in length, whereas only 5,753 or 2.1 % are longer than 500 nt. Compared to the shuffled data where 96.1 % are shorter than 100 nt and only 37 interactions span at least 500 nt. Interestingly, we see increased numbers of interactions with a length between

270 to 340 nt in the native data (Figure 4.11) that spike at roughly 300 nt length. This "peak" from now on also called the "300-peak" covers 24,843 sequences or around 9 % of intronic sequences longer than 400 nt. Interactions longer than 1,000 nt are very rare and account only for 0.5 % of the native interactions. The longest one found is 1,980 nt and spans nearly the whole stretch.

For most analyses, introns longer than 4,000 nt were used and split into two fragments. Figure 4.12 shows the interaction-length-distribution only for sequences with a length greater then 4,000 nt. Again we find, that most interactions are between 10 and 100 nt in length, but now the second peak looks more pronounced than in Figure 4.11 and still covers 15,942 introns or 14.5 % of a total of 110,930 sequences longer than 4,000 nt. The average interaction-length of the native data is 149 nt, in the shuffled data it is only 48 nt.

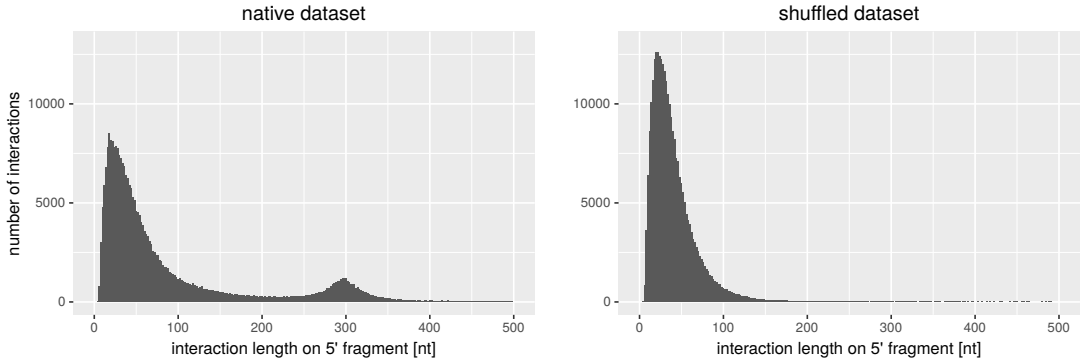

FIGURE 4.11: Comparison of the interaction-length-distribution of both the native (left) and the shuffled dataset (right) for sequences longer than 400 nt. For better visualization, only interactions smaller than 500 nt are shown. The x-axis indicates the interaction length in nucleotides, the y-axis the number of interactions within one bin with a width of 2 nt. Beside the large amount of relatively short interactions, we find an interesting second peak at around 300 nt in the native data, that is absent in the shuffled data. The randomized data shows a more narrow distribution with about 96 % of the interactions shorter than 100 nt.

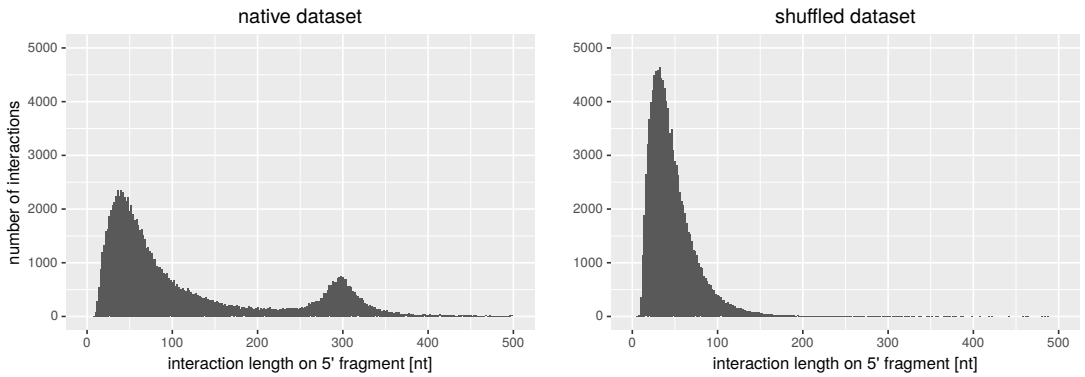

FIGURE 4.12: Comparison of the interaction-length-distribution of both the native (left) and the shuffled dataset (right) for sequences longer than 4,000 nt. For better visualization, only interactions smaller than 500 nt are shown. The x-axis indicates the interaction length in nucleotides, the y-axis the number of interactions within one bin with a binwidth of 2.

4.2.3 Mapping of repeat-elements against predicted interactions

The accumulation of interactions with a length between 270 and 340 nt in Figure 4.11 suggests that there are specific motifs involved in forming these interactions. A known repetitive element similar in length is the Alu-sequence-family (see subsection 2.3.2), which generally are about 300 nt in length and account for around 10 % of the human genome [Deininger and Batzer, 1999]. To pursue this further, we cross-checked against data gained from the UCSC-repeat-masker and used the *intersectBed* tool from the *BEDtools-collection*. It turned out that a lot of the interactions found in this region actually are overlapping with annotated Alu-repeats [Quinlan and Hall, 2010]. From a total of 1,194,735 Alus annotated by the repeat-masker, 690,128 elements overlap with introns by at least 1 nt. An evaluation based on the nucleotide coverage of the entire intron-Alu-overlaps results in a coverage of 12.4 % (354,467,736 nt of 2,848,981,677 nt). Keep in mind, that the extracted introns are based on transcript information. As a consequence some nucleotides may be counted more than once. When only the bases involved in interactions are counted a coverage rate of 24.5 % (7,048,724 nt out of 28,826,825 nt) is reached.

Figure 4.13 shows the number of Alu-sequences overlapping with predicted interaction sites for the full dataset and for interactions with a length between 270 and 340 nt under various minimum-sequence-coverage-restraints. From a total of 24,497 intron sequences with a length between 270 and 340 nt, 10,056 or 40 % could be mapped onto Alu-repeats following the condition that at least 90 % of the predicted interaction overlaps with at least 90 % of an annotated Alu-segment. If this overlapping restriction is set to a minimum of 10 %, 17,443 interactions or 70 % of the subset are overlapping with known Alu sites, whereas for the total dataset these are around 15 %. Table 4.2 shows the number of hits found with at least 90 % overlap of the annotated sequences and predicted interaction-region grouped roughly by their Alu-family as annotated by UCSC.

To analyze, whether this interacting Alus are more stable than interactions of their shuffled counterpart we plotted the mid-to-mid distance of the interactions overlapping by at least 10 % against the interaction energy, and compared the results of both datasets. In Figure 4.14 we observe that the mid-to-mid distance of the shuffled data is much more evenly distributed. Whereas in the native

family	number of hits	%
AluJ	3188	15.85
AluS	14210	70.65
AluY	2714	13.50
$\sum$	20112	100.00

TABLE 4.2: The number of Alu-sequences overlapping with predicted 5'-3' intron-interactions are shown under the constraint that at least 90 % of the Alu-sequences and the interaction-region are covered.

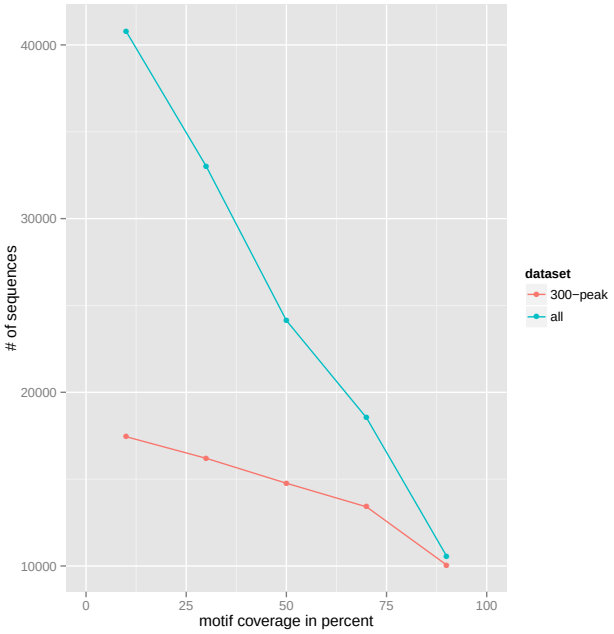

FIGURE 4.13: Number of predicted interactions that overlap with annotated Alus by a certain ratio. The x-axis shows the sequence coverage constraints, with 90 % meaning that the start- and end-coordinates of the interaction and the Alu-motifs are both covered by at least 90 %. On the Y-axis the number of sequences fulfilling these conditions are depicted. The colors indicate the different datasets, where the 300-peak dataset (red) covers only interactions with a length between 270 and 340 nt and is a subset of the whole dataset (blue).

data very short or extremely long distances are rare. Both datasets can also be clearly differentiated in terms of interaction energies, where the native interactions have a much lower energy than the interactions of the corresponding shuffled sequences. To get a maximum interaction mid-to-mid distance the interactions at the outer ends of the 5' and the 3' fragment have to be very short. Such short interactions would probably have a relatively high interaction-energy and therefore be less stable. A short mid-to-mid distance means that the interactions are located at the inner ends of the fragments.

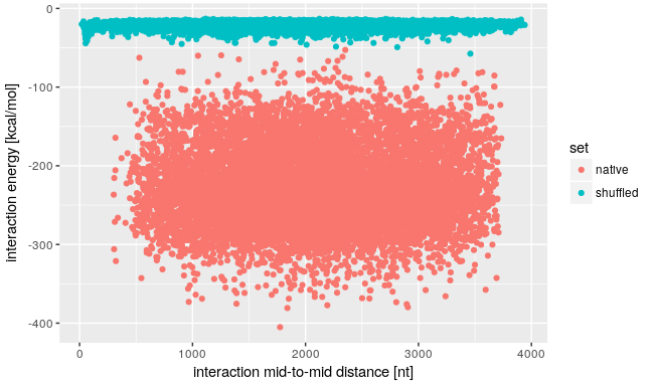

FIGURE 4.14: Comparison of the interaction mid-to-mid distance on the x-axis and the according interaction energy on the y-axis of the native (red) dataset under the constraint that at least 10 % of annotated Alus-sequences overlap with at least 10 % of the interaction region for sequences longer than 4,000 nt. The shuffled dataset (blue) shows the corresponding sequences when shuffled.

On considering Alu-regions interacting with each other, one can argue, that these repeats may be the cause of the "negative shift" of the native data compared to the background model as depicted in Figure 4.2. A closer look at the z-score distribution of the native introns under the condition that interactions are overlapped by at least 10 % with known Alu-sequences (Figure 4.15) showed, that the majority of these introns has a low z-score but the number of Alu-Alu interactions is only a small fraction (17,443 out of 279,443) of the total sequence set analyzed and therefore is not decisive.

Due to their transposable characteristics, Alu-elements can insert in various orientations. Therefore, it is interesting to know their orientation, because one can assume that an Alu prefers an interaction with another Alu facing the opposite direction. By intersecting the introns with the list of 1,194,735 Alus a total of 1,365,979 overlaps were found, whereas 504,607 sequences were non-overlapping Alu-regions. As shown in Table 4.3 overlaps with the opposite strand are more frequent than on the same strand. The number of overlaps exceeds the number of

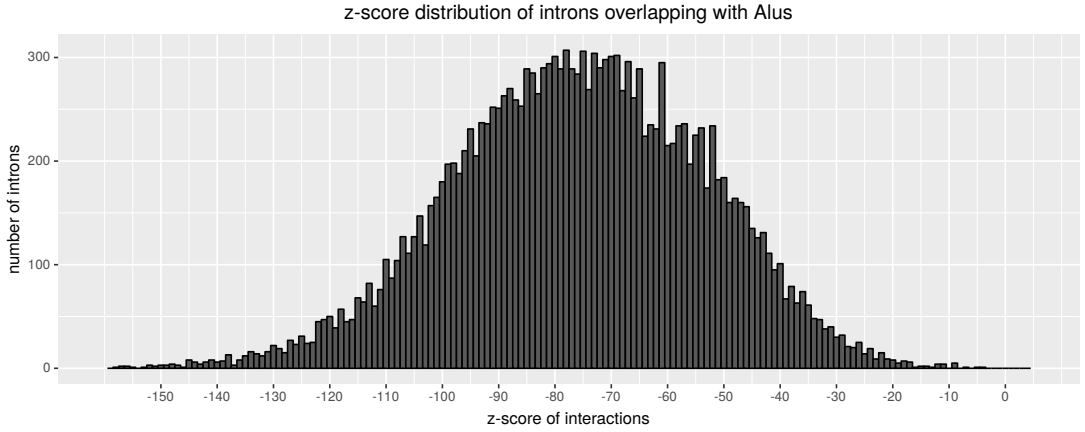

FIGURE 4.15: Z-score distribution of introns forming interactions that are overlapping by at least 10 % with annotated Alu-elements. Most of the Alu-Alu interactions have a very low z-score.

annotated Alu-sequences, because the intron information is based on transcript-information which may hold parts of redundant sequence-stretches.

Introns overlapping with Alus on same or opposite strand		
opposite strand	+/-	370,685
	-/+	388,587
same strand	+/+	308,737
	-/-	297,970
no strandedness		1,365,979

TABLE 4.3: An intron can either overlap with an Alu-element that is on the same or on the opposite strand. Here, an overview of the number of overlaps between introns and Alus is shown.

Next we checked if there are Alu-sequences actually involved in the interactions between the intron ends. As listed in Table 4.4, the majority of Alu-repeats that are overlapping by at least 10 % with an interacting-regions are inverted Alus and only a few elements are oriented head-to-head or tail-to-tail.

Additionally, we checked whether there are interactions overlapping with one of the numerous other repeats annotated in the repeat-masker data (see Table 4.5). From the native data 115,081 interactions overlap by at least 10 % with any of the annotated 5,298,130 repeats. In the shuffled dataset on the other hand, only 22,985 overlapping repeats were found. Note, that overlaps between any repeats are counted, which means a region that partly overlaps with an L1-repeat could

interact with a region overlapping with an Alu or MIR and therefore these results are not very conclusive. With the additional constraint, that an interaction must be between two repeats with an identical "annotation" and has to be on concurrent strands, exactly 5,799 hits remain. The resulting hits are about 85 % Alu sequences, around 6 % MIRs, 4 % L2, and a few L1, FLAM and MLTx repeat element classes.

To analyze the impact of these repeats further and check their influence on interaction stability we looked at the z-score distribution of the interactions between repeats only (Figure 4.16) and compared them to interactions with no repeats involved. Interestingly, we found that interactions with repeats involved often show a lower z-score with values between -160 and -20, when compared to interactions with no repeat-involvement. The number of interactions not having interactions between repeat is a lot higher, which simply happens because there are more interactions between non-repeat-regions.

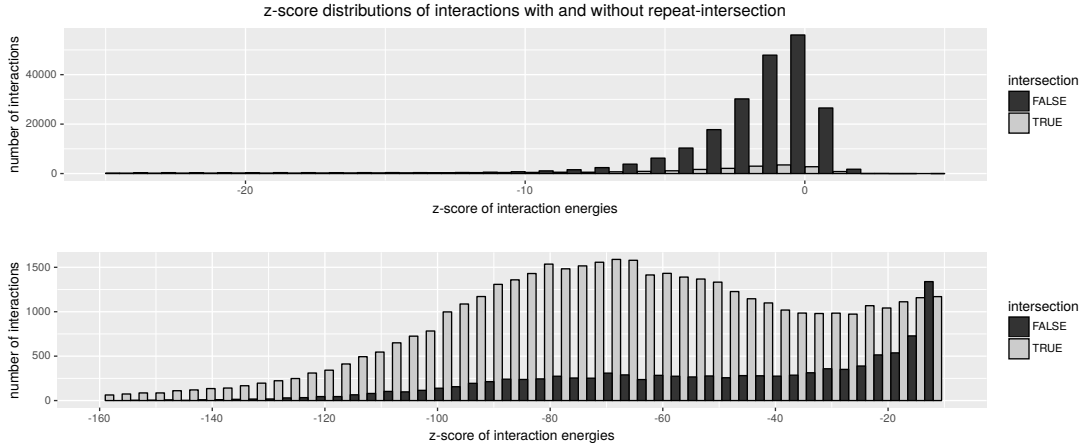

FIGURE 4.16: Section of the z-score distribution of introns overlapping with repeats annotated by the repeat-masker (grey) compared to introns not overlapping with repeats. The upper graph shows the distribution of the z-scores between -25 and +5 and the corresponding intron count, whereas the lower graphs represent a close-in of interactions with a z-score between -160 and +15.

Orientation of Alu-sequences found in interactions in p10 dataset		
+	+	645
+	-	8869
-	-	906
-	+	9024

TABLE 4.4: Here the orientation of interacting Alu-sequences under the constraint, that at least 10 % of the annotated Alu is overlapping with a minimum of 10 % of an interaction is shown. Head-to-head (+/+) and tail-to-tail (−/−) interacting Alus are very scarce, whereas the most frequent interactions are sequences interacting with an Alu complement ($+/-$ *and* $-/+$).

dataset	all repeats (%)	alus only (%)
minimum overlap of 10 %		
300p	20,489 (83.64)	17,443 (71.20)
all	115,081 (41.20)	40,474 (14.49)
minimum overlap of 90 %		
300p	9,971 (40.70)	9,966 (40.68)
all	10,524 (3.77)	10,469 (3.75)

TABLE 4.5: Here, the number of annotated repeats from the UCSC repeatmasker overlapping with by at least 10 or at least 90 % are shown. The numbers are given separately for the whole native dataset with introns longer than 400 nt and for a subset of this dataset with an interaction-length between 270 and 340 nt, termed the 300p data. The values in brackets represent the percentage of hits under the given condition in relation to the total number of sequences (279,443).

We have shown in section 4.2 that interactions between the 5' and the 3' ends of introns are more stable than expected by chance and that this effect still persists even if repeats are excluded. This suggests that interactions between intronic ends are indeed beneficial and that inverted Alu pairs are one way nature achieves this.

Deininger and Batzer revealed that Alus within exonic regions interfere with gene expression and can be the cause of a plethora of diseases [Deininger and Batzer, 1999]. We cross-referenced some genes known to be disrupted by exonic Alu-sequences and found a few that also have interacting Alu-elements in their intronic parts, namely the GK, ATP7A and CLCN5-gene. Two out of those three genes (GK, ATP7A) have introns that are overlapping by at least 90 % with an annotated Alu-repeat. CLCN5 has a coverage of more than 10 %. All of them are located on the X-chromosome.

Chapter 5

Discussion

Introns are a defining part of eukaryotic genomes and their exact function and involvement in the splicing process still raises a lot of questions. Although heavily researched, the process of RNA processing and especially intron splicing, which allows a cell to produce multiple variations of proteins derived from one gene, bears a lot of mysteries. A well known example of alternative splicing is the DSCAM-homologue in *Drosophila*, where around 38,000 isoforms are created from a single gene [Schmucker et al., 2000].

An average exon has a length of about 200 nt, where on the other side the length of an intron varies between a few and several hundred thousand nucleotides (see Figure 4.1) [Sakharkar et al., 2004]. In alternative splicing the spliceosome cuts out the non-coding sequences, but although it is a relatively large complex, it does not span hundreds or even thousands of nucleotides. So how can the spliceosome splice out an intron and ligate the exons when it is not able to recognize both ends of a long intron? Here, we suggest that intra-molecular interactions between the ends of intronic sequences are a possible mechanism that may support the splicing process of long introns.

The intron length distribution (Figure 4.1) showed, that the bigger part of the introns is relatively short. Because we wanted to examine long range interactions, we only used introns with a minimum length of 400 nucleotides for basic statistical analysis. For experiments concerning possible interactions between the 5' and the 3' ends a minimum sequence length of 4,000 nt was chosen to yield two stretches, a 5' and a 3' region, of 2,000 nt each and to ensure that we actually could find long enough interactions. The length of 400 nt was chosen arbitrarily as a lower

limit to exclude short introns. These limitations reduced the dataset and as a consequence the CPU time needed for creating the datasets.

This study required to find the best interaction of many long RNA sequences, which rendered it mandatory to solve these computations in reasonable time. Therefore we chose *RNAplex* over more accurate tools because of its relatively low memory- and time-complexity and combined it with *RNAplfold* to increase the accuracy of the predictions. A few years ago Pervouchine et al tried a different approach using seed-search combined with local alignment scoring (BLASTZ) to find complementary nucleotide stretches near splice-sites of genes [Schwartz et al., 2003] [Pervouchine et al., 2012]. They tried to evaluate the context of long ranged structures in alternative splicing by arbitrary combination of donor and acceptor boxes. Using local alignments for RNA-RNA interaction prediction is very inaccurate, because it relies simply on sequence alignments without making use of an energy model and therefore neglects actual interaction-energies and intra-molecular competition.

We found interactions between the 5' and the 3' end of the introns being significantly more stable than one would expect simply by chance. This is an indication, that long introns actually form stable intra-molecular structures between their outer ends and supports our assumption, that there is a mechanism based on the intrinsic structure or the introns, which allows these ends to come into spacial proximity and facilitates splice site recognition. Because we had to make the results comparable first, we had to build a background model and normalize the data by calculating their z-scores. We also considered keeping possible splicing motifs intact by excluding the first 50 nt from the shuffling process. Additionally, we found that favourable interactions between the outermost ends of the introns were more prevalent than interactions between the inner ends of the introns (Figure 4.6), while in the shuffled dataset we found them to be evenly distributed. The accumulation near the outer ends made sense when we consider that these interactions lead to a better splice-site recognition by bringing them closer together.

By analyzing the interaction start- and endpoints based on their relative distance (Figure 4.3) we found an accumulation of interaction start-points at the beginning of the 5' fragment. Same was true for the end of the 3' fragment. At first we thought that the peaks at the beginning of Figure 4.3 (A) and at the end (B) might have been caused by interactions spanning nearly the whole fragments,

because even a few long interactions with limited positioning-possibilities could lead to that phenomenon. By considering the interaction-length and randomly choosing an interaction start-point as shown in Figure 4.5 we would expect to see a similar effect, but instead it looked uniformly distributed. Additionally, we tried extending the regions for the accessibility-calculations (Figure A.1, Figure A.2) to avert border effects as suggested by Lange et al. [Lange et al., 2012]. We put a lot of effort in this question but in the end the results were inconclusive and probably are an artifact of the interaction calculation.

A more thorough analysis of the interactions showed an accumulation of introns with interactions between the first 400 nt on the 5' fragment and the last 400 nt of the 3' fragment as shown in Figure 4.6. In general we found a lot of interactions reaching from the start of the 5' to the end of the 3' fragment and therefore bridging the 5' to the 3' end, indicating that the ends of these introns were in spacial proximity. Short spanning interactions between the inner parts of the analysed introns are relatively rare. These introns may have used different mechanisms or combinations of mechanisms to get spliced out, for example, it would have been possible that the interactions contributed in bringing the ends closer together such that auxiliary splicing factors could recognize splicing motifs and the combination of both facilitates the splicing reaction.

When we analysed the intron length distribution, we found an interesting second peak around 300 nt in length (Figure 4.1). It turned out, that the these long, stable interactions were often caused by Alu-elements. Additionally, we could show that nucleotides associated with Alu-elements in general were more prevalent in interactions compared to the total introns by a factor of two. Following that, an intersection with annotated repeats showed that repeats interacting with each other seemed to be very stable in general. To make sure that repeats were not the cause for the more stable interactions found in the dataset, we excluded repeat-only interactions and found that although there were interactions between repeats with very low z-scores, the total number of these interactions is relatively low. One could theorize that some of these repeats may not be artifacts but possibly have had regulatory functions regarding alternative splicing or being responsible for an exon to be either included or skipped. Tajaddod and colleagues suggested that iSINEs forming a double-stranded RNA structures may provide binding motifs for certain RNA-binding proteins [Tajaddod et al., 2016]. These Alu-Alu interactions could not only bring the two intron-ends in close

proximity but additionally formed a binding site for regulatory splicing factors and supported the splicing process.

Deininger and Batzer hypothesized in a study that Alu-sequences disrupting exonic regions can be the cause of various different splicing-related diseases [Deininger and Batzer, 1999]. Hence, about 10 % of the whole genome and about 12 % of the introns are Alu-elements, it is not unlikely that we also find Alus interacting in genes or in exons that interact with introns which lead to aberrant splicing diseases. It is interesting that we found interacting intronic Alus in genes which showed a disease-associated phenotype caused by splicing defects. Although currently this is only shown for exonic regions, one can hypothesize that these exonic Alus could interact with intronic Alu-sequences which ultimately result in aberrant splicing or that the aberrant splicing is caused by Alus competing with different Alu-interaction-partners.

An example for a disease caused by Alu-associated splicing errors is the GK-gene which codes for the glycerol kinase and causes Hyperglycerolemia if it is defect. In this region 13 intronic Alu sequences along with an inverted counterpart have been found to form strong interactions. This could be subject to closer investigation examining whether this specific gene is susceptible for splicing defects caused by secondary structure formation in intronic sequences or if the secondary structures are essential and get disrupted by newly inserted Alus as suggested by Zhang et al. [1998].

In summary, the presented results indicate that long range RNA-RNA interactions may be another mechanism influencing the splicing process by bringing the 5' and the 3' intronic ends in spacial proximity to be recognized by the spliceosome. In a wider study it probably would make sense to group all introns by specific features, for example by GC-content, by the frequency of being skipped, by their translation frequency and furthermore, analyze each group individually. In the future, it also would be interesting to check for long range interactions between introns spanning one or more exons that may facilitate exon skipping events via large secondary structures. A closer investigation on the role of Alu-elements in forming intronic-interactions could also be interesting. Especially diseases caused by aberrant splicing should be checked. Because, this study is based on statistical analysis only, it would be interesting to check if our calculations fit to actual data from lab-work. Therefore, we have assembled a list (Table A.2) with possible candidates that could be interesting to check in the

lab. In general, we see this work as a starting point for further investigations of the role of intronic long range interactions.

List of Figures

List of Tables

Appendix A

Appendix

FROM	-	TO	number of introns
100	-	399	51,666
400	-	999	58,037
1,000	-	3,999	11,0475
4,000	-	9,999	53,301
10,000	-	19,999	26,057
20,000	-	59,999	22,468
60,000	-	99,999	4,983
100,000	-	199,999	3,143
		> 200000	979
		$\sum_{\geq 100}$	331,109
		$\sum_{\geq 400}$	279,443

TABLE A.1: This tables gives an overview of the number of introns within a given length in a basepair interval (FROM,TO). Most of the intervals are arbitrarily taken. Sequences with a length between 100 and 399 are not considered in further analysis, because they are not considered long. In total there are 331,109 intronic sequences $\geq$ 100 nt and 279,443 sequences longer than 399 nt. The later ones are used in the next experiment.

We find, that the first bin of the interaction start sites (Figure A.2 A, C) as well as the last bins (Figure A.2 B, D) On both fragments are pronounced. Additionally there are slightly less interactions formed in the coding regions (-200, 0) than in the non-coding regions.

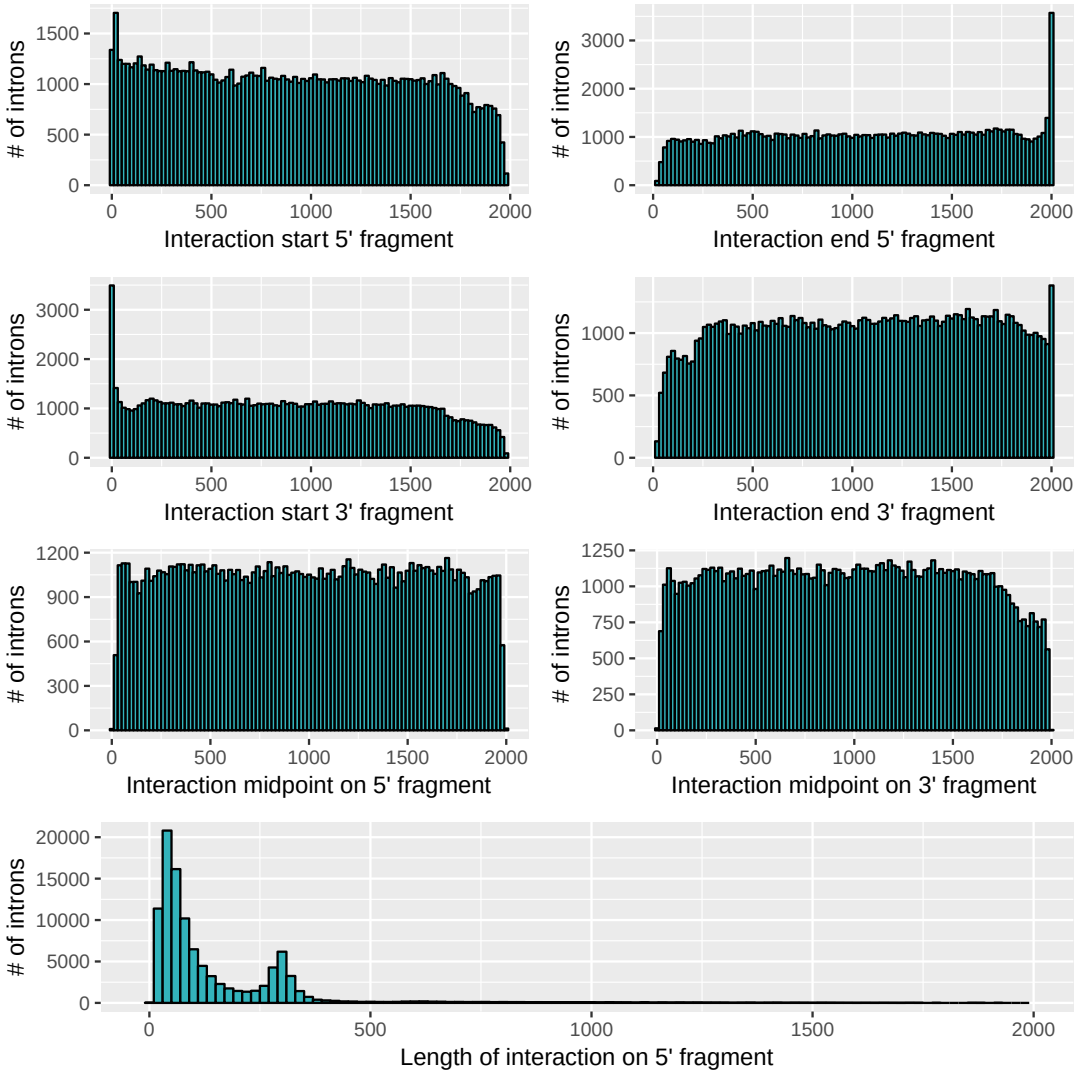

FIGURE A.1: Overview of start- and end-points of predicted interactions in introns longer than 4,000 nt (two fragments á 2,000 nt + 200 nt) extended by one windowlength (200 nt) into the coding region on each fragment. binwidth = 20 nt A) shows the number of interactions starting in a specific interval on the 5' or "target" sequence. B) shows the end of the interactions on the 5' fragment. C) and D) show the number of interactions starting and ending in a specific bin width a width of 20 nt. E) and F) depict the midpoint of the interactions on the 5' and the 3' fragment respectively. G) and H) represent the length distribution of the interactions on both fragments for all introns $> 4,000nt$. I) shows the length distribution of the introns in this dataset.

A similar experiment with extended sequences but with *RNAplex* constraint to find only interaction-sites in intronic regions also shows a higher frequency in the first (Figure A.1 A, C) and last bins (B, D) on both segments suggesting that these are caused by the method and not by actual accumulations of interactions in this segment.

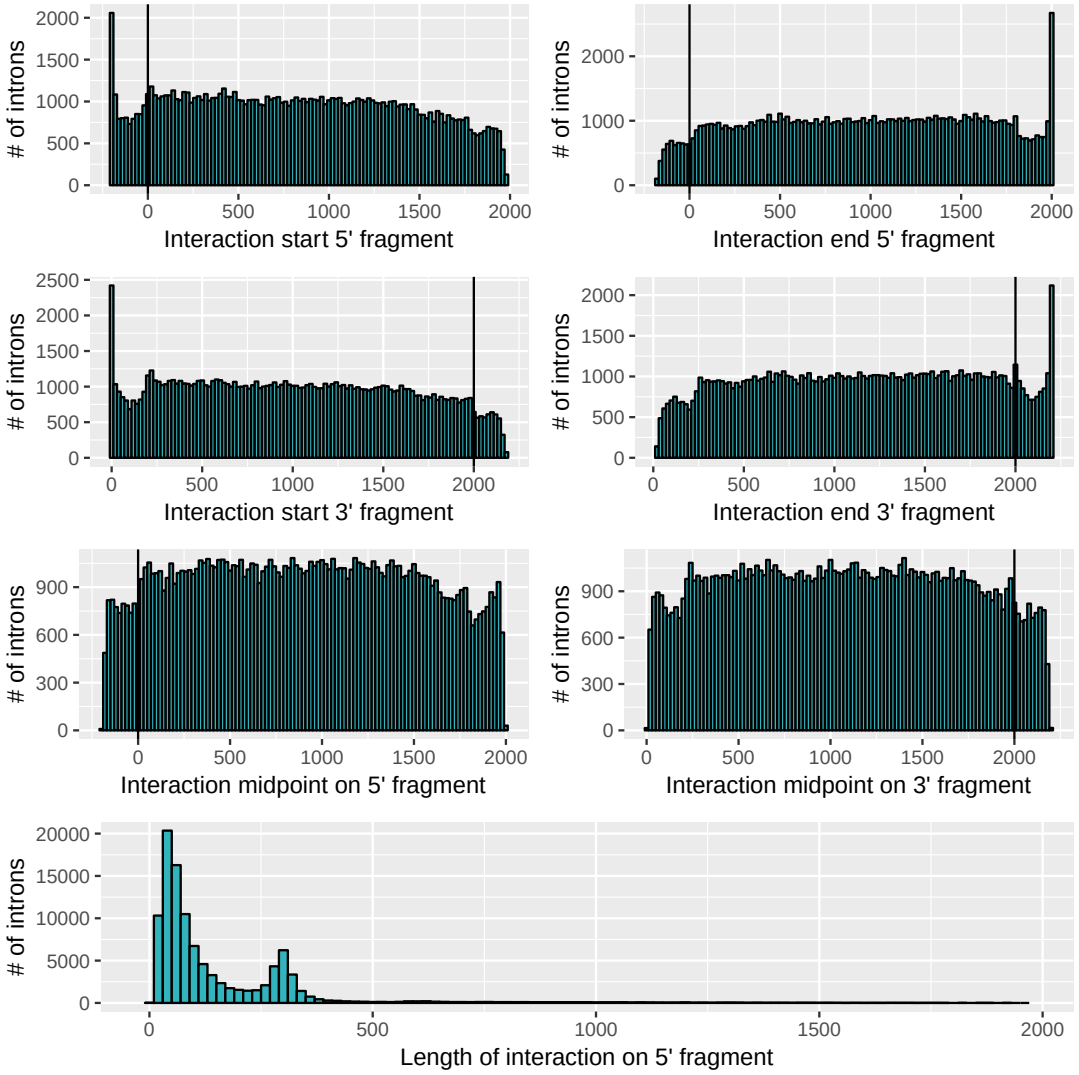

FIGURE A.2: Overview of start and end point of predicted interactions in introns longer than 4,000 nt (two fragments of 2,000 nt in length) with each fragment extended by 200 nt into the exonic region. The black line indicates the actual start of the intron on the 5' piece and the end of the intron on the 3' segment. One bin is equal to 20 nt. A) shows the number of interactions starting in a specific interval on the 5' or "target" sequence. B) shows the end of the interactions on the 5' fragment. C) and D) show the number of interactions starting and ending in a specific bin width a width of 20 nt. E) and F) depict the midpoint of the interactions on the 5' and the 3' fragment respectively.

transcript-id	5'exon	z-score	mfe	intron-length	IA-5'	IA-3'	bins	AluO
gene_name: GK, id: ENSG00000198814.8								
ENST00000427190.1	ENSE00003477733.1	-106.9	-285.09	6183	[233,530]	[1064,1360]	1.4	*
ENST00000427190.1	ENSE00003498504.1	-79.3	-206.11	11895	[687,986]	[846,1136]	2.3	*
ENST00000487652.1	ENSE00003462844.1	-68.0	-211.24	23759	[421,716]	[845,1133]	2.3	*
ENST00000487652.1	ENSE00003486611.1	-0.2	-19.11	13670	[638,663]	[609,633]	2.2	*
ENST00000471362.1	ENSE00003598373.1	-132.6	-289.12	11790	[424,714]	[1292,1578]	2.4	*
ENST00000471362.1	ENSE00003574058.1	-124.9	-300.92	16947	[606,913]	[95,401]	2.2	*
gene_name: ATP7A, id: ENSG00000165240.13								
ENST00000350425.4	ENSE00001760012.1	-116.3	-266.91	59639	[747,1043]	[586,885]	2.3	*
ENST00000341514.6	ENSE00001694839.1	-14.7	-60.59	4362	[178,281]	[1893,1994]	1.5	
ENST00000341514.6	ENSE00001656404.1	-64.9	-170.68	4388	[983,1309]	[937,1263]	3.4	*
ENST00000341514.6	ENSE00001683584.1	-2.4	-54.79	5014	[660,784]	[1227,1356]	2.4	
ENST00000341514.6	ENSE00001633722.1	-93.0	-269.61	5865	[220,513]	[1374,1668]	1.5	
ENST00000341514.6	ENSE00001594340.1	-20.5	-61.09	8170	[778,855]	[1277,1353]	2.4	
ENST00000341514.6	ENSE00001712655.1	-4.3	-26.57	8520	[1797,1844]	[1240,1288]	5.4	
ENST00000341514.6	ENSE00001636041.1	-0.3	-18.46	16479	[1221,1249]	[71,99]	4.1	

TABLE A.2: List of interesting interactions that should be checked in vitro. To identify the correct intron the transcript-id and the id of the 5'-exon is given. Additional to the interacting regions (IA-5', IA-3') the interaction-bin is shown (bins). A mark in the column AluO means, that the interaction is overlapping with an annotated Alu. Here, interactions from two genes (GK, ATP7A) are given as examples with a variety of z-scores and interaction-regions (bins). These genes are also known to have aberrant splicing-variants caused by Alu-inserts that lead to certain diseases [Deininger and Batzer, 1999].

Bibliography

Altschul, S. F., Gish, W., et al. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):pages 403–410, 1990. ISSN 0022-2836. doi:10.1016/S0022-2836(05)80360-2.

Andronescu, M., Zhang, Z. C., and Condon, A. Secondary structure prediction of interacting RNA molecules. *Journal of Molecular Biology*, 345(5):pages 987–1001, 2005. ISSN 0022-2836. doi:10.1016/j.jmb.2004.10.082.

Beadle, G. W. and Tatum, E. L. Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Sciences of the United States of America*, 27(11):pages 499–506, 1941. ISSN 0027-8424.

Bernhart, S. H., Hofacker, I. L., and Stadler, P. F. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):pages 614–615, 2006. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/btk014.

Bernhart, S. H., Mückstein, U., and Hofacker, I. L. RNA accessibility in cubic time. *Algorithms for Molecular Biology*, 6(1), 2011.

Bompfünewerer, A. F., Backofen, R., et al. Variations on RNA folding and alignment: lessons from Benasque. *Journal of Mathematical Biology*, 56(1-2):pages 129–144, 2007. ISSN 0303-6812, 1432-1416. doi:10.1007/s00285-007-0107-5.

Brown, J. W., Haas, E. S., and Pace, N. R. Characterization of ribonuclease P RNAs from thermophilic bacteria. *Nucleic Acids Research*, 21(3):pages 671–679, 1993. ISSN 0305-1048.

Bruccoleri, R. E. and Heinrich, G. An improved algorithm for nucleic acid secondary structure display. *Bioinformatics*, 4(1):pages 167–173, 1988. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/4.1.167.

Busch, A., Richter, A. S., and Backofen, R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics (Oxford, England)*, 24(24):pages 2849–2856, 2008. ISSN 1367-4811. doi:10.1093/bioinformatics/btn544.

Cao, A. and Galanello, R. Beta-thalassemia. *Genet Med*, 12(2):pages 61–76, 2010. ISSN 1098-3600. doi:10.1097/GIM.0b013e3181cd68ed.

Cech, T. R. and Bass, B. L. Biological catalysis by RNA. *Annual review of biochemistry*, 55(1):pages 599–629, 1986.

Chen, J.-L. and Greider, C. W. Functional analysis of the pseudoknot structure in human telomerase RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 102(23):pages 8080–8085, 2005.

Chen, L.-L., DeCerbo, J. N., and Carmichael, G. G. Alu element-mediated gene silencing. *The EMBO journal*, 27(12):pages 1694–1705, 2008. ISSN 1460-2075. doi:10.1038/emboj.2008.94.

Crick, F. Central Dogma of Molecular Biology. *Nature*, 227, 1970.

Deininger, P. Alu elements: know the SINEs. *Genome biology*, 12(12):page 236, 2011.

Deininger, P. L. and Batzer, M. A. Alu repeats and human disease. *Molecular Genetics and Metabolism*, 67(3):pages 183–193, 1999. ISSN 1096-7192. doi: 10.1006/mgme.1999.2864.

Eggenhofer, F. RNApredator: A web-based tool to predict small RNA targets. Ph.D. thesis, University of Vienna, 2011.

Erickson, B. W. and Altschul, S. F. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Molecular Biology and Evolution*, 2(6):pages 526–538, 1985. ISSN 0737-4038. doi:10.1093/oxfordjournals.molbev.a040370.

Fox-Walsh, K. L., Dou, Y., et al. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proceedings of the National Academy of Sciences of the United States of America*, 102(45):pages 16,176–16,181, 2005.

Fraenkel-Conrat, H. and Singer, B. Virus reconstitution. *Biochimica et Biophysica Acta*, 24:pages 540–548, 1957. ISSN 00063002. doi:10.1016/0006-3002(57)90244-5.

Gilbert, W. The RNA world. 1986.

Gotoh, O. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3):pages 705–708, 1982. ISSN 0022-2836. doi:10.1016/0022-2836(82)90398-9.

Guhaniyogi, J. and Brewer, G. Regulation of mRNA stability in mammalian cells. *Gene*, 265(1):pages 11–23, 2001.

Hedges, D. and Deininger, P. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 616(1-2):pages 46–59, 2007. ISSN 00275107. doi:10.1016/j.mrfmmm.2006.11.021.

Hofacker, I. L., Fontana, W., et al. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly*, 125(2):pages 167–188, 1994. ISSN 0026-9247, 1434-4475. doi:10.1007/BF00818163.

IUPAC-IUB Comm. on Biochem. Nomencl. Abbreviations and symbols for nucleic acids, polynucleotides, and their constituents. *Biochemistry*, 9(20):pages 4022–4027, 1970. ISSN 0006-2960, 1520-4995. doi:10.1021/bi00822a023.

J, H., A, F., et al. GENCODE: the reference human genome annotation for The ENCODE Project., GENCODE: The reference human genome annotation for The ENCODE Project. *Genome research, Genome Research*, 22, 22(9, 9):pages 1760, 1760–1774, 2012. ISSN 1088-9051. doi:10.1101/gr.135350.111,10.1101/gr.135350.111.

Jacob, F. and Monod, J. Genes of structure and genes of regulation in the biosynthesis of proteins. *C. R. Hebd. Seances Acad. Sci.*, 249:pages 1282–1284, 1959. ISSN 0001-4036.

Kerpedjiev, P., Hammer, S., and Hofacker, I. L. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, 31(20):pages 3377–3379, 2015. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/btv372.

Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, 44(2):pages 283–292, 1986. ISSN 00928674. doi:10.1016/0092-8674(86)90762-2.

Kriegs, J. O., Churakov, G., et al. Evolutionary history of 7sl RNA-derived SINEs in Supraprimates. *Trends in genetics: TIG*, 23(4):pages 158–161, 2007. ISSN 0168-9525. doi:10.1016/j.tig.2007.02.002.

König, H., Matter, N., et al. Splicing segregation: The minor spliceosome acts outside the nucleus and controls cell proliferation. *Cell*, 131(4):pages 718–729, 2007. ISSN 00928674. doi:10.1016/j.cell.2007.09.043.

Lander, E. S., Linton, L. M., et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):pages 860–921, 2001. ISSN 0028-0836. doi:10.1038/35057062.

Lane, C. E., van den Heuvel, K., et al. Nucleomorph genome of hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function. *Proc. Natl. Acad. Sci. U.S.A.*, 104(50):pages 19,908–19,913, 2007. ISSN 1091-6490. doi:10.1073/pnas.0707419104.

Lane, N. and Martin, W. The energetics of genome complexity. *Nature*, 467(7318):pages 929–934, 2010. ISSN 0028-0836, 1476-4687. doi:10.1038/nature09486.

Lange, S. J., Maticzka, D., et al. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Research*, 40(12):pages 5215–5226, 2012. ISSN 0305-1048, 1362-4962. doi:10.1093/nar/gks181.

Lee, R. C., Feinbaum, R. L., and Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):pages 843–854, 1993. ISSN 0092-8674. doi:10.1016/0092-8674(93)90529-Y.

Leontis, N. B. and Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):pages 499–512, 2001. ISSN 13558382. doi:10.1017/S1355838201002515.

López-Bigas, N., Audit, B., et al. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Letters*, 579(9):pages 1900–1903, 2005. ISSN 00145793. doi:10.1016/j.febslet.2005.02.047.

Lorenz, R. RNA Secondary Structure Thermodynamics and Kinetics. Ph.D. thesis, University of Vienna, 2014.

Lorenz, R., Bernhart, S. H., et al. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):page 1, 2011.

Matera, A. G. and Wang, Z. A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology*, 15(2):page nrm3742, 2014. ISSN 1471-0080.

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

Muckstein, U., Tafer, H., et al. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):pages 1177–1182, 2006. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btl024.

Ng, B., Yang, F., et al. Increased noncanonical splicing of autoantigen transcripts provides the structural basis for expression of untolerized epitopes. *Journal of Allergy and Clinical Immunology*, 114(6):pages 1463–1470, 2004. ISSN 00916749. doi:10.1016/j.jaci.2004.09.006.

Nishikura, K. Functions and Regulation of RNA Editing by ADAR Deaminases. *Annual Review of Biochemistry*, 79(1):pages 321–349, 2010. ISSN 0066-4154, 1545-4509. doi:10.1146/annurev-biochem-060208-105251.

Nussinov, R. Algorithms for Loop Matchings. *Society for Industrial and Applied Mathematics*, 1978. doi:10.1137/0135006.

Ott, S., Tamada, Y., et al. Intrasplicing–analysis of long intron sequences. *Pacific Symposium on Biocomputing*, pages 339–350, 2003. ISSN 2335-6928.

Pan, Q., Shai, O., et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):pages 1413–1415, 2008. ISSN 1546-1718. doi:10.1038/ng.259.

Pervouchine, D. D., Khrameeva, E. E., et al. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. *RNA*, 18(1):pages 1–15, 2012. ISSN 1355-8382. doi:10.1261/rna.029249.111.

Polak, P. and Domany, E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC genomics*, 7(1):page 1, 2006.

Quinlan, A. R. and Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):pages 841–842, 2010. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/btq033.

Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. Fast and effective prediction of microRNA/target duplexes. *RNA (New York, N.Y.)*, 10(10):pages 1507–1517, 2004. ISSN 1355-8382. doi:10.1261/rna.5248604.

Roberts, R. Microsomal particels and protein synthesis. In *Microsomal particels and protein synthesis*. Pergamon Press, 1958.

Rogozin, I. B., Carmel, L., Csuros, M., and Koonin, E. V. Origin and evolution of spliceosomal introns. *Biol Direct*, 7(11):pages 6150–7, 2012.

Sakharkar, M. K., Chow, V. T. K., and Kangueane, P. Distributions of exons and introns in the human genome. *In Silico Biol. (Gedruckt)*, 4(4):pages 387–393, 2004. ISSN 1386-6338.

Schmid, C. W. and Deininger, P. L. Sequence organization of the human genome. *Cell*, 6(3):pages 345–358, 1975. ISSN 0092-8674.

Schmucker, D., Clemens, J. C., et al. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6):pages 671–684, 2000. ISSN 0092-8674 0092-8674.

Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8, 2015.

Schwartz, S., Kent, W. J., et al. Human-mouse alignments with BLASTZ. *Genome Research*, 13(1):pages 103–107, 2003. ISSN 1088-9051. doi:10.1101/gr.809403.

Sorek, R. Alu-containing exons are alternatively spliced. *Genome Research*, 12(7):pages 1060–1067, 2002. ISSN 10889051. doi:10.1101/gr.229302.

Sterner, D. A., Carlo, T., and Berget, S. M. Architectural limits on split genes. *Proc Natl Acad Sci U S A*, 93(26):pages 15,081–15,085, 1996. ISSN 0027-8424.

Tafer, H., Amman, F., et al. Fast accessibility-based prediction of RNA-RNA interactions. *Bioinformatics*, 27(14):pages 1934–1940, 2011. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/btr281.

Tafer, H. and Hofacker, I. L. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, 24(22):pages 2657–2663, 2008. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/btn193.

Tajaddod, M., Tanzer, A., et al. Transcriptome-wide effects of inverted SINEs on gene expression and their impact on RNA polymerase II activity. *Genome Biology*, 17(1):page 220, 2016. ISSN 1474-760X. doi:10.1186/s13059-016-1083-0.

Tilgner, H., Knowles, D. G., et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, 22(9):pages 1616–1625, 2012. ISSN 1088-9051. doi:10.1101/gr.134445.111.

Turner, D. H. and Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 38(Database):pages D280–D282, 2010. ISSN 0305-1048, 1362-4962. doi:10.1093/nar/gkp892.

Versteeg, R., van Schaik, B. D., et al. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome research*, 13(9):pages 1998–2004, 2003.

Wang, G.-S. and Cooper, T. A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, 8(10):pages 749–761, 2007. ISSN 1471-0056, 1471-0064. doi:10.1038/nrg2164.

Watson, J. D. and Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):pages 737–738, 1953. ISSN 0028-0836 0028-0836.

Will, C. L. and Lührmann, R. Spliceosome structure and function. *Cold Spring Harb Perspect Biol*, 3(7):page a003,707, 2011. ISSN , 1943-0264. doi:10.1101/cshperspect.a003707.

Zhang, C., Yang, H., and Yang, H. Evolutionary Character of Alternative Splicing in Plants. *Bioinformatics and Biology Insights*, 9(Suppl 1):pages 47–52, 2015. doi:10.4137/BBI.S33716.

Zhang, Y., Huang, B., et al. Alu sx insertion in a patient with benign glycerol kinase deficiency. *Am J Hum Genet Suppl*, 63:page A395, 1998.

Zuker, M. and Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):pages 133–148, 1981.