Bioinformatics Analyses of Alternative Splicing

Non-EST based Prediction, Influence of Secondary Structures and Tandem Splice Sites

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat
der Fakultät für Angewandte Wissenschaften
der Albert-Ludwigs-Universität Freiburg

von Diplom-Informatiker (Univ.)

Michael Hiller

Gutachter

Prof. Dr. Rolf Backofen

Prof. Dr. Peter Stadler

Tag der Disputation: 4. Dezember 2006

Abstract

After realizing its frequency in the last decade, alternative splicing has attracted considerable attention. Although several biological phenomena can be explained by alternative splicing today, research has just started to uncover all of its aspects.

This thesis investigates three aspects of alternative splicing, mostly by means of computational large-scale analyses. In the first part, we introduce a new approach to predict alternative splicing without using expressed sequence information. Given that our knowledge about the human transcriptome is still incomplete, *ab initio* prediction of alternative splicing is a rather recent but important research area. In contrast to existing methods, our approach is independent of orthologous sequences, thus it is applicable to a single genome. By introducing an efficient dynamic programming algorithm, we reduce the computational complexity during the search for new splice events compared to a naive algorithm. The use of this algorithm is demonstrated in a genome-wide application, where we predict and verify novel human splice events.

In the second part, we investigate the influence of mRNA secondary structures on the regulation of the splicing process. We show that experimentally verified binding sites of splicing regulatory proteins have a higher single-strandedness. As alternative and constitutive splicing often depends on several such binding sites, this indicates a general importance of mRNA secondary structures for splicing. Then, we develop a new motif finding method that benefits from using an informative prior probability distribution, which takes the single-strandedness of putative motif occurrences into account. We perform extensive tests with artificial and biological data sets and demonstrate that the additional information about secondary structures help to discriminate real binding sites from spurious ones.

In the third part, we analyze a group of splice events that have mostly escaped attention in the past. These splice events occur at tandem acceptor splice sites and result in minor changes of the mRNA and the protein. Genome-wide analyses provide evidence for a non-random distribution of these splice events at the genome and protein level, for tissue-specific regulation, and for evolutionary conservation. Moreover, we find that SNPs affecting such acceptors have a highly predictive effect on splicing. Extending our studies to tandem donors, we investigate differences between alternatively and not alternatively spliced tandem donors. We conclude that these donor and acceptor splice events represent one major mechanism to increase the proteome diversity and that some of them have consequences for protein function and human disease. Finally, we develop a relational database, which stores extensive information about tandem splice sites.

In summary, in this thesis, we introduce a new approach for *ab initio* splice event prediction, uncover another detail about the regulation of splicing, develop a new *de novo* motif finding method, perform the first detailed genome-wide analysis of tandem splice sites, and develop a specific database of tandem donors and acceptors.

Zusammenfassung

Nachdem die Häufigkeit von alternativ gespleißten Genen im letzten Jahrzehnt erkannt wurde, hat das alternative Spleißen in der Wissenschaft große Aufmerksamkeit erfahren. Mehrere biologische Phänomene können heute durch alternatives Spleißen erklärt werden. Trotzdem hat die Forschung gerade erst begonnen alle Aspekte aufzudecken.

Diese Dissertation untersucht drei verschiedene Aspekte des alternativen Spleißens, hauptsächlich durch Anwendung von computerbasierten Analysen. Im ersten Teil wird eine neue Methode für die Vorhersage von alternativen Spleißformen ohne Verwendung von exprimierten Sequenzen vorgestellt. Wenn man bedenkt, dass unser Wissen über das humane Transkriptom noch unvollständig ist, stellt die *ab initio* Vorhersage von Spleißformen ein neues, aber wichtiges Forschungsgebiet dar. Im Gegensatz zu anderen Methoden ist unser Ansatz unabhängig von Informationen über orthologe Sequenzen und daher auf einzelne Genome anwendbar. Die Komplexität der Suche nach neuen Spleißformen kann durch die Entwicklung eines effizienten Algorithmus, der auf dem Prinzip der dynamischen Programmierung basiert, deutlich reduziert werden. Wir zeigen den Nutzen dieser Methode durch eine Anwendung auf das humane Genom, bei der wir neue Spleißvarianten vorhersagen und nachweisen.

Der zweite Teil der Arbeit untersucht den Einfluss von mRNA Sekundärstrukturen auf die Regulation des Spleißprozesses. Dabei zeigen wir, dass experimentell bestätigte Bindungsstellen von regulatorischen Spleißfaktoren eine signifikant höhere Einzelsträngigkeit aufweisen. Da alternatives und auch konstitutives Spleißen von mehreren solcher Bindungsstellen abhängt, deutet dieses Ergebnis auf einen generellen Einfluss von mRNA Sekundärstrukturen auf den Spleißprozess hin. Wir nutzen dieses Prinzip bei der Entwicklung eines neuen Algorithmus für die Erkennung von Motiven in biologischen Sequenzen. Dieser Algorithmus berücksichtigt die Einzelsträngigkeit möglicher Bindungsstellen, was durch eine sequenzspezifische a priori Wahrscheinlichkeitsverteilung modelliert wird. Umfassende Tests mit künstlichen und biologischen Datensätzen zeigen, dass diese zusätzliche Information hilfreich ist, um zwischen echten und falsch-positiven Bindungsstellen zu unterscheiden, was genauere Motivbeschreibungen erlaubt.

Im dritten Teil analysieren wir eine Gruppe von alternativen Spleißereignissen, die bisher wenig Beachtung gefunden haben. Diese Ereignisse geschehen an Tandemakzeptor-Spleißstellen und führen zu subtilen Veränderungen der mRNA und des entsprechenden Proteins. In genomweiten Untersuchungen fanden wir Hinweise, dass diese Spleißereignisse nicht zufällig im Genom und im Proteom verteilt sind; dass sie gewebespezifisch reguliert werden können; und dass eine Teilmenge evolutionär konserviert ist. Weiterhin konnten wir zeigen, dass SNPs in solchen Spleißstellen einen vorhersagbaren Effekt auf Veränderungen im Spleißmuster haben. Wir erweitern die Untersuchungen auf Tandemdonor-Spleißstellen und analysieren Unterschiede zwischen alternativen und konstitutiven Tandemdonoren. Wir kommen zu dem Schluss, dass Tandem-Spleißstellen

einen wichtigen Mechanismus zur Vergrößerung der Proteom Vielfalt darstellen. Außerdem haben einige dieser Spleißstellen Auswirkungen auf die Proteinfunktionalität sowie auf menschliche Erkrankungen. Um weitere Forschungen zu erleichtern, erstellen wir eine spezifische Datenbank, die umfassende Informationen über Tandem-Spleißstellen öffentlich zugänglich macht.

Zusammengefasst lässt sich sagen, wir entwickeln in dieser Dissertation einen neuen Ansatz für die *ab initio* Spleißformvorhersage; beschreiben ein weiteres Detail der Regulation des Spleißprozesses; stellen einen neuen Algorithmus für die Erkennung von unbekannten Sequenzmotiven vor; führen die erste umfassende Analyse von Tandem-Spleißereignissen durch und erstellen eine spezifische Datenbank über Tandemdonoren und -akzeptoren.

Danksagung

Zuerst möchte ich mich bei meinem Doktorvater Rolf Backofen ganz herzlich bedanken für die Betreuung dieser Arbeit, für die gute und erfolgreiche Zusammenarbeit, aus der mehrere gemeinsame Publikationen entstanden sind, für viele wertvolle Ideen und Anregungen, und für all das, was ich von ihm lernen konnte. Ihm verdanke ich mein Interesse für diverse Bioinformatik Probleme algorithmischer und biologischer Natur.

Bei Peter Stadler möchte ich mich für das Interesse an dieser Arbeit bedanken und für die Bereitschaft diese zu begutachten.

Weiterhin bedanke ich mich bei meinen Kollegen Anke Busch, Martin Mann, Rainer Pudimat, Sven Siebert, Sebastian Will für interessante und lustige Diskussionen und für gemeinsame Freizeitaktivitäten. Ganz herzlich bedanken möchte ich mich bei meiner Zimmergenossin Anke Busch für das Aushalten meiner Person in 'kommunikativen' und 'unkommunikativen' Phasen, sowie für das Tolerieren der teilweise lauten Tastaturbearbeitung und der Zimmertemperaturen, welche auf offene Fenster und Klimaanlagen zurückzuführen waren.

Bedanken möchte ich mich auch bei der Jenaer Genome Analyse Gruppe (Matthias Platzer, Klaus Huse, Karol Szafranski, Stefanie Schindler, Swetlana Nikolajewa, Rileen Sinha) für die zahlreichen, endlosen und interessanten Diskussionen und 'Autoren Biwacks' sowie die produktive Zusammenarbeit, in der ich viel gelernt habe und die ich sehr genossen habe (und auch weiterhin genießen werde). Ein ganz besonderer Dank geht dabei an Klaus Huse für das Prägen des Begriffes 'Buchstabenrechner', an Matthias Platzer für das ständige Perfektionieren von allen Tabellen, Bildern und Begriffen sowie an Karol Szafranski, der auch in der hektischsten Diskussion stets einen kühlen Kopf behielt.

Weiterhin bedanke ich mich bei Stefan Stamm und Zhaiyi Zhang für die ebenfalls erfolgreiche und interessante Kooperation, welche hoffentlich in Zukunft so weiterläuft. Für das Korrekturlesen dieser Arbeit bedanke ich mich vielmals bei Anke Busch und Klaus Huse.

Außerdem bedanke ich mich bei meinen Eltern, meinem Bruder und allen Freunden für die ständige Unterstützung während meines Studiums und meiner Zeit in Jena und Freiburg. Ganz herzlichen Dank auch an meine Freundin Manu für die Liebe und Unterstützung, die mir trotz knapper Zeit und großer Entfernung zuteil wurde.

List of own publications

[1] Hiller M, Nikolajewa S, Huse K, Szafranski K, Rosenstiel P, Schuster S, Backofen R, and Platzer M. TassDB: a database of alternative tandem splice sites. *Nucleic Acids Res.* database issue, 2007. in press

[2] Szafranski K, Schindler S, Taudien S, Hiller M, Huse K, Jahn N, Schreiber S, Backofen R, and Platzer M. Violating the splicing rules: TG dinucleotides function as alternative splice acceptors in U2–dependent introns. submitted

[3] Hiller M, Szafranski K, Backofen R, and Platzer M. Alternative splicing at NAGNAG acceptors: Simply noise or noise and more? *PLoS Genet.* **2**(11), e207, 2006.

[4] Hiller M, Pudimat R, Busch A, and Backofen R. Using RNA Secondary Structures to Guide Sequence Motif Finding towards Single–Stranded Regions. *Nucleic Acids Res.* **34**(17), e117, 2006.

[5] Platzer M, Hiller M, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, and Huse K. Sequencing errors or SNPs at splice–acceptor guanines in dbSNP? *Nature Biotechnol.* **24**(9), 1068–70, 2006.

[6] Hiller M, Huse K, Szafranski K, Rosenstiel P, Schreiber S, Backofen R, and Platzer M. Phylogenetically widespread alternative splicing at unusual GYNGYN donors. *Genome Biology* **7**(7):R65, 2006

[7] Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, and Platzer M. Single–Nucleotide Polymorphisms in NAGNAG acceptors are highly predictive for variations of alternative splicing. *American J Human Genet.* **78**(2), 291–302, 2006.

[8] Hiller M, Huse K, Platzer M, and Backofen R. Non–EST based prediction of exon skipping and intron retention events using Pfam information. *Nucleic Acids Res.*, **33**(17), 5611–21, 2005.

[9] Hiller M, Huse K, Platzer M, and Backofen R. Creation and disruption of protein features by alternative splicing – a novel mechanism to modulate function. *Genome Biology*, **6**(7):R58, 2005.

[10] Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, and Platzer M. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nature Genet.*, **36**(12), 1255–7, 2004.

[11] Hiller M, Backofen R, Heymann S, Busch A, Glaesser TM, and Freytag JC. Efficient prediction of alternative splice forms using protein domain homology. *In Silico Biology*, **4**(2), 195–208, 2004.

Organization of the thesis

This thesis describes three studies that address open questions concerning alternative splicing. These studies have in common that we mainly use algorithmic work and large-scale computational analyses. However, from a biological viewpoint, these studies address different aspects of alternative splicing. Specifically, these aspects are the *ab initio* prediction of splice events, the regulation of splicing by secondary structures, and the novel field of tandem splice sites. Therefore, I decided to describe the three studies in the chapters 2, 3, and 4.

Each of these chapters starts with a brief summary and then gives a specific introduction, mentions related work, and motivates the following study. Each chapter ends with a specific discussion of the findings. These three chapters are preceded by an introduction chapter 1, which provides general background knowledge about alternative splicing. A broader conclusion and an outlook are given in chapter 5.

Writing style

Current research is mostly team work and consequently rather '*we*' instead of '*I*'. Most of this work was done in collaboration with other researchers who contributed with ideas, wet-lab experiments, and biological expert knowledge. Therefore, I decided to write this thesis in the '*we*' style. At this point, it should be mentioned that all wet-lab experiments are not performed by myself. However, as they are important to verify and extend our computational analysis, the results of these experiments are briefly described.

Contents

Chapter 1

Introduction into constitutive and alternative splicing

"Genes in pieces"

Soon after the discovery of exons and introns, Walter Gilbert 1978 asked "*Why genes in pieces?*" [1]. Gilbert speculated that differential usage of exons can lead to functionally different proteins. Contradicting the fundamental dogma "*one gene, one polypeptide chain*", this would allow evolution to "*seek new solutions without destroying the old*". Furthermore, he speculated that mutations at exon boundaries and at silent codon positions can affect the splicing process. In the last decade, genome projects and large-scale studies have started to uncover the importance of alternative splicing and confirmed many of the hypotheses suggested by Gilbert in 1978.

Alternative splicing is the main topic of this thesis. This chapter gives an introduction to the biology of the splicing mechanism and to important bioinformatics studies. Because of the complexity of this field, we mainly focus on topics that are relevant for the understanding of the work described in this thesis.

1.1 The splicing mechanism

A fundamental difference between the gene structures of prokaryotes and eukaryotes is the existence of introns in the latter. Most genes of higher eukaryotes consist of several exons and introns, with an average of ten exons (and nine introns) per human gene [2]. During the *splicing* process, introns are removed from the pre-mRNA and only exons are retained in the mature transcript. This process is carried out by one of the largest molecular machines, the *spliceosome*, which consists of several small nuclear RNAs (snRNAs) and more than 150 proteins. The snRNAs are bound to multiple proteins forming *small nuclear ribonucleoprotein particles* (snRNPs). Five snRNPs (called U1, U2, U4, U5, and U6) are involved in the splicing process. Noteworthy, despite the involvement of

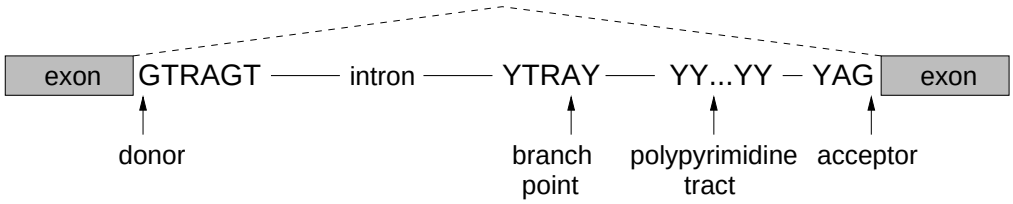

Figure 1.1: Schematic illustration of an intron with its neighboring exons.
The basic splicing signals are depicted. The dashed line indicates the splicing pattern. Y stands for C or T, R stands for A or G.

numerous proteins, the RNA components of the snRNPs seem to be mainly responsible for the catalytic activity of the spliceosome.

The recognition of an intron in the pre-mRNA sequence requires three basic splicing signals (Figure 1.1). The first two signals are the 5' intron end, which is called the *donor* splice site, and the 3' intron end, which is called the *acceptor* splice site. In mammals, the donor site has an extended intronic consensus sequence GTRAGT (R stands for A or G), where the first intronic dinucleotide GT[1] is nearly invariant. Apart from very rare exceptions, the terminal acceptor dinucleotide is always AG. The third signal, the *branch point*, is usually located about 40 nucleotides (nt) upstream of the acceptor site and has the consensus YTRAY (Y stands for C or T, the branch point adenosine is underlined) [3]. The acceptor is preceded by a stretch of pyrimidines, which is called the *polypyrimidine tract*.

Simplified, the splicing of an intron occurs in the following order [4]:

- the U1 snRNP binds to the donor site by specific base pairings between the snRNA and the mRNA,
- the protein heterodimer U2AF binds to the polypyrimidine tract and acceptor site,
- the U2 snRNP binds to the branch site by base pairings,
- the tri-snRNP consisting of U4, U5, and U6 enters the spliceosome,
- the U6 snRNP replaces U1 by binding to the donor site, and U1 and U4 are released from the spliceosome,
- the mRNA is cleaved at the donor site and the 5' intron end is attached to the branch point adenosine forming a lariat structure,
- the mRNA is cleaved at the acceptor site, the upstream exon is ligated to the downstream exon, and the intron is released.

[1]Since we analyze alternative splicing mainly from a genomic viewpoint in the following, we write T instead of U throughout this thesis, also when referring to an RNA sequence.

1.2 Alternative splicing

As anticipated by Walter Gilbert [1], genes can produce more than one mature transcript by allowing alternative splicing decisions. This process is called *alternative splicing*. Different transcripts from one gene are often translated into different proteins, thus violating the classical 'one gene, one polypeptide chain' rule. Although the earliest reports of alternative splicing date back to 1980 [5], the frequency of alternative splicing was revealed in the last decade. Surprisingly, the 'one gene, one polypeptide chain' rule applies to only a minority of eukaryotic genes.

Given the exon-intron structure of a gene, one distinguishes *constitutive* from *alternative* splice sites. A constitutive splice site is always used to produce a mature transcript, while an alternative splice site can be omitted sometimes. Likewise, the terms *constitutive* and *alternative* are applied to exons and to the splicing process in general.
Most alternative splice events can be classified into the following basic types:

- the inclusion or exclusion of one (or more) exons (denoted exon skipping, Figure 1.2A),
- the usage of alternative donor or acceptor sites (Figure 1.2B and C),
- the mutual exclusion of exons (Figure 1.2D),
- the retention of an intron (Figure 1.2E)

Apart from these basic events, genes sometimes produce rather complex splicing patterns (Figure 1.2F). Furthermore, alternative splicing can be coupled to transcriptional variation such as alternative promoter or polyadenylation site usage. Since these events are not addressed in this thesis, we refer the types described in Figure 1.2 as alternative splice events.

1.2.1 Frequency of alternative splicing

A common strategy to detect alternative splice events in a genome-wide manner is based on available mRNAs and expressed sequence tags (ESTs). An EST is a partial sequence of a transcribed DNA sequence. With an average length of about 500 nt, ESTs are often shorter than the entire transcript and error-prone since their sequence is determined in a single sequencing step. Nevertheless, the abundance of automatically sampled ESTs (currently about eight million for human) allows the detection of alternative splice events in a genome-wide manner. To this end, ESTs and mRNAs are often aligned to the human genome sequence (Figure 1.3). Since ESTs represent parts of spliced transcripts, larger gapped regions usually indicate the position of introns, while aligned parts generally correspond to exons. To increase the specificity, one often demands that putative introns have the typical GT-AG splice site dinucleotides. However, alignments at the exon-exon junctions can be ambiguous (Figure 1.3C). Therefore, special computer programs such

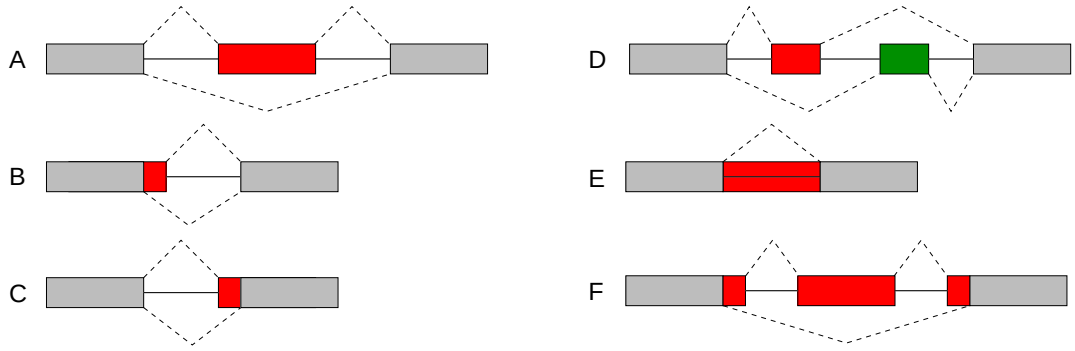

Figure 1.2: Types of alternative splicing.
(A) Exon skipping: the red exon is included in some transcripts and excluded in other transcripts. Such exons are called *alternative exons*. (B) Alternative donor: the upstream exon has an alternative donor site. (C) Alternative acceptor: the downstream exon has an alternative acceptor site. (D) Mutually exclusive exons: either the red or the green exon is included in the transcript. (E) Intron retention: the entire intron can be retained in some transcripts. (F) Complex events: An example of a complex event is the simultaneous skipping of an exon and usage of alternative donor and acceptor sites. Exons are shown as boxes, introns as horizontal lines. Dashed lines indicate the splicing pattern.

as sim4 [6] have been developed to yield a spliced alignment with correct splice site dinucleotides. Alternative splice events are then detected by searching exons (or parts thereof) that are contained in some ESTs and excluded in others (Figure 1.3A). Likewise, intron retention and other splice events can be identified from these alignments.

EST based studies found that between 35% and 59% of the human genes have alternative splice forms [7, 8, 9, 10], thus providing the first evidences that alternative splicing is widespread in the human as well as in other genomes. Furthermore, corrected for the different number of ESTs for different species, it seems that alternative splicing is equally frequent from human to worm [11], although this has been discussed controversially [12]. ESTs were also used to predict tissue-specific [13, 14] and cancer-specific splice variants [15].

Another way to monitor alternative splicing in a genome-wide manner is to use specific DNA microarrays (reviewed in [16]). One common approach designs oligonucleotide probes that are specific to the interior of exons and probes that are specific for certain splice junctions (Figure 1.4). The intensity of the different probes is used to elucidate which splice events are present and what the ratio between alternative splice events is. Microarray based studies found 74% of the human genes to be alternatively spliced [17]. Furthermore, microarrays have been used to detect and analyze tissue-specific splice events [18, 19], to study the regulation of alternative splicing [20], and to correlate splicing with human disease [21].

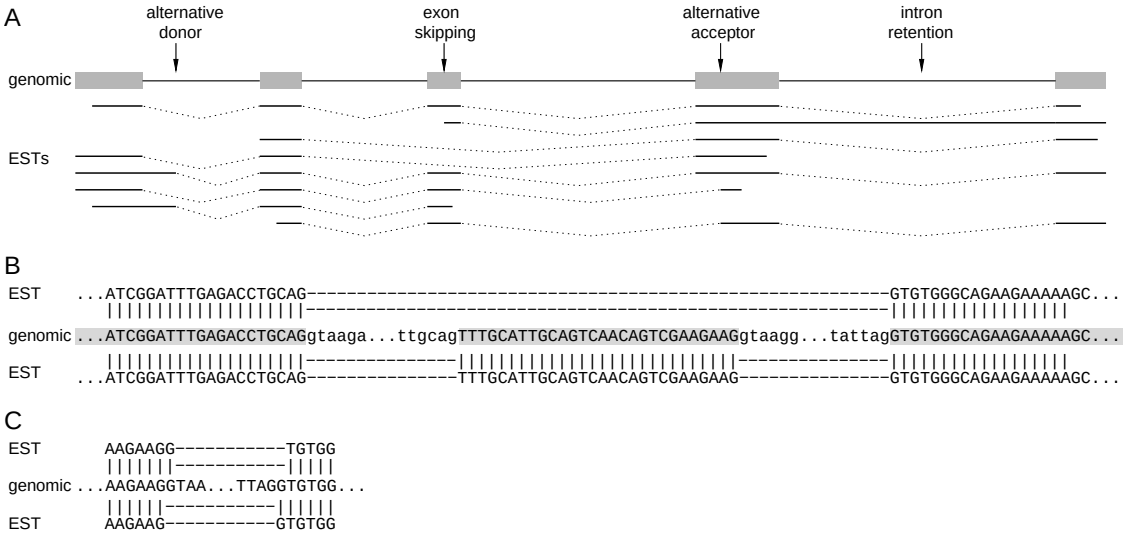

Figure 1.3: Illustration of EST to genome alignments.
(A) Several ESTs (horizontal lines) are aligned to genomic sequence of a gene with five exons (grey boxes). ESTs that span multiple exons provide evidence for the respective splicing pattern (dotted lines). Since ESTs are only fragments of the full-length transcripts they often start and end within the gene. Nevertheless, they allow the detection of diverse alternative splice events (illustrated here for exon skipping, alternative donor and acceptor sites, and intron retention).
(B) The spliced alignment of two ESTs against the genome sequence reveals skipping of exon 3 for the transcript NM_001001392. Exonic nucleotides are in upper case letters, intronic ones in lower case letters. Exons are highlighted in grey.
(C) The given EST sequence can be aligned to the junction of exon 3 and 4 of NM_001001392 in two ways. The incorrect alignment above results in an intron with TA-GG boundaries, while the alignment below leads to the correct intron annotation with canonical GT-AG splice sites.

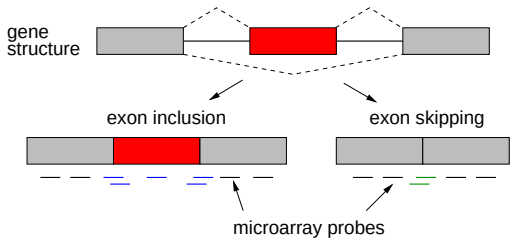

Figure 1.4: Design of microarrays to detect alternative splicing illustrated for an exon skipping event.
Horizontal black lines indicate the location of probes that are specific for the constitutive exons. Blue probes are specific for the exon inclusion event, while the green probes are specific for the exon skipping event.

1.2.2 Regulation of alternative splicing

The spliceosome recognizes introns with a high fidelity. In contrast to yeast, the basic splicing signals (splice sites and branch point) of higher organisms do not contain all the information that is required for an accurate intron recognition since these signals are rather degenerate [22]. Moreover, pseudo splice sites (sequences that resemble real splice sites but that are never used) outnumber real splice sites by an order of magnitude in pre-mRNA transcripts [23]. Furthermore, alternative splicing can be specific for

- a tissue or cell-type,
- a developmental stage,
- or an external stimulus like heat shock or stress conditions [24].

Therefore, additional signals must be involved in the splicing process in general and in the regulation of alternative splicing in particular.

A major contribution to the splicing fidelity comes from additional sequence motifs that are located in exons and introns. Motifs that promote splicing are called *enhancers*, while those that inhibit splicing are called *silencers*. According to their location, they are classified as exonic and intronic splicing enhancers and silencers (abbreviated ESE, ESS, ISE, ISS) [25, 26, 27]. Enhancer motifs are frequently bound by the group of serine/arginine rich (SR) proteins. Binding of SR proteins to enhancer motifs mostly exerts a positive effect on splice site recognition and stimulates the spliceosome assembly [28]. These positive effects can be antagonized by heterogeneous nuclear ribonucleoproteins (hnRNPs) that usually bind silencer elements [29]. However, it should be noted that the same sequence motif can act as an enhancer or silencer, depending on its position with respect to splice sites [30]. Additional splicing signals and splicing factors allow to discriminate between real and pseudo splice sites [27]. Therefore, they are essential for alternative as well as constitutive splicing [31]. Since splicing motifs are abundant in exons [32], exon inclusion is often promoted by several splicing factors. Current research indicates that the high fidelity and sometimes strict regulation of the splicing mechanism is achieved by the combinatorial control of multiple splicing factors [33, 34] as well as proofreading mechanisms [35].

The regulation of alternative splicing is highly dynamic and often controlled in a tissue-specific or stimulus-specific manner. This can be achieved by a different concentration of splicing factors in different environments. Since splicing factors have numerous other mRNAs as potential targets, a change in the concentration of one factor influences the splicing of many transcripts simultaneously. For example, neurons express a specific splicing factor Nova-1 that regulates the splicing of several mRNAs in a neuron-specific manner [36]. In addition to targeting many other mRNAs, most splicing factors use alternative splicing of its own mRNA to autoregulate their own protein level. One such example is the human tra2β splicing factor. In high concentrations, this protein binds

to enhancer elements present in exon 2 of its pre-mRNA, which leads to the inclusion of this exon and a mature mRNA that is not translated into functional proteins [37]. Furthermore, the activity of splicing factors depends on their phosphorylation status. Phosphorylation or dephosphorylation of splicing factors can also lead to a movement into a different subcellular localization (such as from the nucleus to the cytoplasm), where they are unable to affect splicing [24]. These mechanisms also influence the global splicing pattern in a cell. For example, the heat shock mediated dephosphorylation of SRp38 results in a global shutdown of the splicing activity [38]. As splicing occurs co-transcriptionally, the promoter architecture and the transcriptional speed affect the regulation of alternative splicing by kinetic effects and differential recruitment of splicing factors [39]. Finally, the secondary structure of mRNAs plays a role in alternative and constitutive splicing as well [40]. A detailed introduction into the latter point is given in section 3.1.

Extensive knowledge about splicing motifs and factors is the basis to understand how alternative splicing is regulated. To this end, much biological research focuses on the identification of splicing factor binding sites. Moreover, several computational analyses have extended the current list of known splicing motifs.

Fairbrother et al. compared the counts of all 4,096 hexamers in exonic vs. intronic sequences and in exons with weak vs. exons with strong splice sites [41]. Hexamers that were significantly enriched in exons with weak splice sites are assumed to act as enhancers, consistent with the view that exons with weak splice sites are not accurately spliced without the aid of additional enhancer elements. They predicted a total of 238 hexamers as ESEs and demonstrated the enhancer function of representative motifs experimentally. Avoiding potential biases due to protein coding sequences, Zhang and Chasin compared internal non-coding exons vs. pseudo exons and internal non-coding exons vs. the untranslated regions (UTRs) of one-exon genes [42]. Here, pseudo exons are intronic parts that are bounded by sequences resembling real splice sites but that never become exonic. Pseudo exons and the UTRs of intronless genes should have only few ESEs but frequently ESSs. Thus, motifs that are significantly enriched in real internal exons and rare in both control groups are assumed to be ESEs, while motifs that are more frequent in pseudo exons and in UTRs of one-exon genes are assumed to be ESSs. With this strategy, they identified 2,060 octamers as ESEs and 1,019 octamers as ESSs. Other computational studies predicted ISE motifs [43] as well as motifs associated with brain-specific alternative splicing [44] and exon skipping [45]. Very recently, comparative genomics was used to identify splicing motifs [30].

1.2.3 Impact of alternative splicing

One of the most unexpected findings by the human genome project was the surprisingly small number of protein-coding genes. It is estimated that the human genome harbors

only 22,000 genes [2]. That means, humans have only slightly more genes than less complex species like the nematode *Caenorhabditis elegans* with 20,000 genes. This indicates that the gene number is not correlated with the complexity of an organism. The finding that most genes in higher eukaryotes express several alternative splice forms partially resolved this apparent discrepancy. Indeed, alternative splicing is currently considered to be a major mechanism for producing a complex proteome from a limited number of genes [46].

Protein isoforms, produced by alternative splicing, can differ in various physiological aspects including ligand binding affinity, signaling activity, protein domain composition, subcellular localization, and protein half-life [47, 48, 49, 50]. Often the protein function is altered by inserting/deleting functional units like protein domains, transmembrane (TM) helices, or signal peptides. Bioinformatics analysis have shown that alternative splicing has a tendency to remove certain protein domains like protein-protein interaction or DNA binding domains [51]. For example, alternative splicing frequently removes the protein-protein interaction domain of Kruppel family transcription factors [51]. The alternative protein isoforms will still bind to the target DNA site but do not initiate transcription anymore, thus exerting a dominant negative effect by blocking the binding site. Interestingly, alternative splicing tends to insert/delete complete functional units instead of affecting parts of a unit [52]. Moreover, many proteins occur in a soluble as well as in a membrane bound form. One way to produce soluble isoforms is to skip exons that encode the TM helices. Indeed, computational analyses found that 40-50% of the proteins with one TM helix have a splice form that specifically removes the single TM domain [53, 54]. The regulation of alternative splice events plays a role in several biological processes such as the formation and function of synapses [36], axon guidance in the fruit fly *Drosophila melanogaster* [55], and T-cell activation [56]. Finally, it should be noted that alternative splicing in the UTR regions can have an effect by influencing mRNA stability or translation efficiency [57].

During gene expression it comes undoubtedly to a low rate of errors yielding erroneous transcripts. Such transcripts may be translated into truncated proteins that are harmful. Therefore, cells have evolved a mechanism called nonsense-mediated mRNA decay (NMD) to degrade these transcripts [58]. In humans, cells use a rule based on the 'splicing history' to determine which transcripts are degraded by NMD. According to this rule, mature transcripts with a premature termination codon (PTC) more than 50 nt upstream of the last exon-exon junction are candidates for NMD. Usually these transcripts are degraded rapidly, so that little or no protein is produced. Alternative splicing can lead to a transcript with a PTC, for example by introducing a frameshift or including an exon encoding an in-frame stop codon. Interestingly, computational studies found that 35% of the alternative splice forms contain a PTC. Therefore, it was suggested that alternative splicing together with NMD provides a mechanism for the regulation of

the protein level independent of the transcription level [59, 60]. Indeed, this seems to be exploited in the autoregulation of many splicing factors such as PTB [61]. However, it should be noted that the extent of this mechanism is discussed controversially [62].

Defects in alternative and constitutive splicing are causative for a number of human diseases [63, 64]. Noteworthy, splicing mutations have been suspected to be the most frequent cause of hereditary diseases [65]. For example, a polymorphism in the *PTPRC* gene that is associated with multiple sclerosis destroys an exonic splicing silencer and abolishes the skipping of exon 4 [66]. Furthermore, changes in the normal splicing pattern are thought to contribute to cancer development [67, 15]. Thus, alternative splicing is also of therapeutic interest [68].

1.2.4 Evolution of alternative splicing

Since most genes and gene structures are conserved between human and mouse, the question arises to which extent alternative splice events are conserved? Furthermore, it is important to assess how many splice events are functionally important for an organism and how many are due to aberrant splicing or noise in the splicing process [69]. Given that human and mouse diverged from a common ancestor about 75 million years ago, conservation of a splice event is a strong indication of functionality.

Focusing on exon skipping as the most frequent splice event, conservation of alternative splicing between human and mouse was investigated by several bioinformatics studies. Although these studies are hampered by differences in EST coverage between species, they found that only a small fraction (between 5 and 15%) of the human alternative splice events are conserved in mouse [70, 71, 72]. These percentages represent a lower bound since, in addition to conservation of the exon at the sequence level, it was required that the respective splice event is also confirmed by ESTs from both species. Therefore, additional splice events that currently lack EST confirmation in one organism are likely to be conserved. However, splice events are clearly species-specific if the alternative exon is not conserved at the sequence level. Surprisingly, this is the case for about 50% of the human and mouse alternative exons [72].

These studies found many characteristic differences between conserved and species-specific alternative exons. The characteristic features of conserved alternative exons are as follows:

- Most of these exons are '*peptide-cassettes*', which means their length is a multiple of 3 nt and they do not encode an in-frame stop codon [69, 73, 74]. Therefore, exon skipping simply results in the removal of a part of the protein without changing the reading frame.
- Compared with constitutive exons, the alternative exon itself and its intronic flanking regions exhibit a much higher sequence conservation between human and mouse [75, 76]. This can be used to predict alternative splice events (see section 2.2).

- They are smaller with an average length of 87 nt compared to 116 nt for non-conserved alternative exons [69].
- They have a high inclusion level as measured by the number of ESTs that show inclusion vs. skipping [77, 69].

These features indicate that most of these alternative exons have a function that is conserved during mammalian evolution.

The frequent species-specific exons are either the result of the deletion of ancestral exons in one species or the creation of new exons. Genome-wide comparisons with the genome of a third species (in this case rat) found that exon creation is much more frequent than deletion [77]. The major source for new exons are mobile DNA elements like Alu elements [78]. Alu elements are inserted into random positions in a genome and only a few mutations are required to create a new alternatively spliced exon [79, 80]. Other mobile DNA repeat elements have contributed to exon creation in human and mouse as well. Furthermore, many mutually exclusive alternative splicing events have evolved after the duplication of an exon [81, 82].

Alternatively spliced exons are subjected to a relaxed evolutionary selection pressure to preserve the protein coding sequence and this relaxation is even more pronounced in exons that are only rarely included [83]. The same even holds for entire genes that express alternative splice forms [84]. Thus, alternative splicing is associated with a faster protein evolution. While mutations that change the coding sequence accumulate in rarely included alternative exons, it is remarkable that these exons are subjected to much stronger constraints to preserve the RNA sequence at translationally silent positions [83]. A likely explanation for these (at first glance) inconsistent findings is that these alternative exons are enriched in regulatory splicing motifs assuring that they are rarely and possibly tissue-specifically included. This leads to the constraint to conserve certain exonic regions that are required for their proper alternative splicing. On the other hand, other regions are rather free to evolve to new functional protein sequences, which is reflected by the reduced selection pressure at the coding sequence.

Based on these results, it was proposed that an alternative exon represents an 'internal paralog' of a gene [77]. A newly created exon is likely to be included in only a minority of transcripts since this exon was previously not required for protein function, thus its splice sites and splicing motif composition should not be selected for a high inclusion level. Furthermore, such exons might only be recognized by the spliceosome in the specific environment like in a certain tissue. Alternative splicing of the gene would still produce much of the ancestral mRNA, yielding a sufficiently high level of the functional protein. The minor splice form might be translated to a protein that currently has no function. This protein is now free to evolve to a new function, which is indicated by the increased rate of amino acid mutations in the alternative exon [78]. Thus, as proposed by Gilbert 1978 [1], alternative splicing is a very important mechanism that allows a

gene to evolve a new function without destroying its existing function. Apart from alternative splicing, new functions arise by gene duplication ('external paralogs') followed by divergence of the gene copies. Consistent with the view that an alternative exon represents an internal paralog, genes that exist in only a single copy have a significantly higher level of alternative splicing than genes that exist in large families with several genomic copies [85].

Chapter 2

Non-EST based prediction of alternative splice events using Pfam information

One major goal of research in the post-genomic area is the elucidation and characterization of the entire spectrum of alternative splice forms. Most of the known alternative splice events have been detected by the comparison of ESTs and cDNAs. Although EST based approaches are powerful, not all splice events are represented in EST databases since ESTs have several biases. Furthermore, these methods are limited to genomes having a sufficiently high EST coverage. Therefore, it is of great interest to apply non-EST based methods to predict alternative splice events *ab initio*.

Despite the existence of many algorithms for a related problem - the prediction of gene structures in genomic DNA - there are only few methods for the prediction of alternative splice events. These methods mainly exploit information of conservation patterns.

In the first part of this thesis, we address the problem of the *ab initio* prediction of alternative splice events with a novel strategy that is solely based on the annotation of protein domains. As the Pfam (Protein domain families) database is one of the most comprehensive collection of functional protein domains, we use Pfam domains for the splice event prediction. In contrast to existing methods, our approach is independent of the existence of orthologous sequences. To apply this approach in a genome-wide manner, we develop an efficient algorithm to reduce the computational complexity. This algorithm was designed to predict exon skipping as well as intron retention events.

We applied our approach to all human RefSeq transcripts and demonstrate that our predictions are very reliable. Subsequent analysis of splice events within Pfam domains revealed a significant preference of alternative exon junctions to be located at the protein surface and to avoid secondary structure elements. Thus, splice events within Pfams are likely to alter the structure and function of a domain, which makes them highly interesting for detailed biological investigation. As Pfam domains are annotated in many other species, our strategy to predict exon skipping and intron retention events might be important for species with a lower number of ESTs. In summary, our algorithm complements a growing list of bioinformatics tools for non-EST based splice event prediction.

Plan of the chapter

We give an overview of approaches to detect alternative splicing on a genome-wide scale in section 2.1. Then, we briefly discuss existing approaches for non-EST based alternative splice event prediction in section 2.2. In section 2.3, we provide the fundamental principle of our Pfam-based approach by searching for features that discriminate between alternative and constitutive exons. A naive prediction algorithm and a more efficient algorithmic solution is given in section 2.4 and 2.5. We apply this algorithm in a genome-wide manner and evaluate the results in section 2.6. Finally, we conclude with a discussion in section 2.7.

2.1 Genome-wide detection of alternative splice events

Almost all large-scale bioinformatics studies of alternative splicing use the wealth of information stored in EST databases and most alternative splice forms are detected by the alignment of EST sequences to the genome and to other ESTs/cDNAs [11, 86, 7, 10]. Despite more than seven million human ESTs in dbEST (release July 2006), not all existing splice variants are represented in this database due to several reasons.

1. The expression level of a transcript must be sufficiently high to be sampled as an EST. Therefore, lowly expressed splice forms are underrepresented. However, minor splice forms can be very important. For example, the *RAC1* gene produces a minor splice variant (Rac1b) that constitutes a large portion of activated Rac1 proteins in a cell and might play a role in tumorgenesis [87].
2. Alternative splicing can be highly specific for a tissue or a cell type, a developmental stage, or an external stimulus [24]. Such specific splice forms can only be detected if ESTs are sampled from the right tissue, at the right time, and under the right condition. Moreover, the tissue distribution of ESTs is strongly biased [13]. Currently, brain has the highest number of human ESTs, which presumably reflects the research focus. Additionally, lowly expressed variants have a tendency to be tissue-specific [77], which makes their detection even more difficult.
3. ESTs are biased towards the ends of transcripts, especially towards the 3' end. For example, the first exons of *CFTR* or *NRXN2* are not covered by a single EST, whereas their 3' UTR is covered by 31 and 13 ESTs, respectively.
4. About 70% of the human ESTs are sampled from tumor libraries. In some cases, this led to gene annotations based on tumor specific transcripts, although another (possibly unknown) predominant splice form is expressed in normal tissue [15].
5. Due to the single read nature, ESTs are error-prone and false positive predictions may be included in alternative splice databases [88].

Apart from ESTs, microarrays with specific exon-exon junction probes have been used to find alternative exons in a genome-wide scale [17]. Specific microarrays have also been used to detect a variety of alternative splice events including exon skipping, alternative donor/acceptor sites, and mutually exclusive exons by searching for tissue-specific changes in the responses of certain microarray probes [18]. Despite the power of microarrays, the main problem remains since it is very hard to test all combinations of tissues, developmental stages, and external stimuli. Furthermore, events like intron retention and alternative donor/acceptor sites or additional exons that are located in introns (relative to the given exon structure of the gene for which probes are designed) can only be detected if intronic probes are included in the microarray design.

Furthermore, a large number of algorithms for the prediction of gene structures based on genomic DNA exists [89]. Except for a few methods [90, 91], these algorithms only compute a single optimal gene structure. Since it is unknown to which extent suboptimal gene structures correspond to alternative splice forms, their use for alternative splice event prediction is limited. Consequently, our current view of alternative splicing is still incomplete and non-EST based methods for the prediction of splice variants are needed to complete our knowledge of the human transcriptome.

2.2 Related work

Recently, Sorek et al. described a non-EST based method that uses characteristic features of alternative exons to discriminate between constitutive and alternative ones [92]. The most discriminative single-feature is a high conservation of alternative exons and their flanking intron regions in mouse [76]. Additional features are an exon size divisible by three, differences in tri-mer counts, and the composition of the splice sites [93]. Comparative genomics was also successfully used to predict exon skipping events in Drosophila [94]. Yeo et al. described an approach ACESCAN that is able to identify conserved exon skipping events in both human and mouse [71]. This approach also uses exonic and intronic conservation as well as splice site scores, exon and intron lengths, and oligonucleotide composition. Ohler et al. demonstrated that even alternative exons that are completely missed in current gene annotations can be discovered by applying a pair hidden Markov model algorithm to orthologous human-mouse introns [95]. Finally, Raetsch et al. used a support vector machine to predict alternative exons [96]. These studies demonstrate that a classifier based on characteristic genomic features can reliably predict exon skipping events *ab initio*.

2.3 Finding discriminating features between alternative and constitutive exons

Our new prediction algorithm is solely based on the annotation of Pfam domains. Since we aim at a highly specific prediction approach, we first have to detect discriminating features between alternative and constitutive exons. Although the boundaries of (alternative and constitutive) exons correlate with Pfam domain boundaries in general [97], a considerable fraction of the alternative splice events (28%) occurs within protein domains [52]. Therefore, we investigated the differences in the contribution of alternative and constitutive exons to Pfam domains.

We constructed a set of 213 known alternative and 5,728 constitutive exons where each exon encodes a complete Pfam domain or a part of it. Consistent with other studies [76], we considered an exon as constitutive if it has at least six ESTs that show inclusion and no EST that shows skipping. An alternative exon is skipped in at least three ESTs. First, we only considered exons that do not introduce a frameshift or a premature termination codon (PTC) when skipped (denoted '*peptide-cassette*' exon). Then, we compared the Pfam score between the proteins with and without such an exon and counted the number of cases where exon skipping results in an increase of the Pfam score. We found that significantly fewer constitutive exons have this property compared to alternative exons (Table 2.1). Furthermore, the average score increase observed for alternative exons is significantly higher than the average increase for the constitutive ones (Table 2.1). Therefore, we searched for a minimum score increase that leads to an even better separation of constitutive and alternative exons. We decided to use 10 as a threshold value, since only a tiny fraction (0.1%) of the constitutive exons results in a Pfam score increase of at least 10 when skipped, in contrast to 9% of the alternative exons (Table 2.1). This suggests that a genome-wide search for exons with this characteristic property can be used to predict alternative exons with a high specificity. How the skipping of a peptide-cassette can result in a score increase is illustrated by two examples in section 2.6.1 Figure 2.9.

Up to now, we have only considered peptide-cassette exons. However, exons that are not peptide-cassettes can also result in a Pfam score increase.

- The skipping of such an exon can lead to a frameshift and the new protein sequence downstream can encode a longer C-terminus of a Pfam domain or a completely new domain.
- The skipping of an exon that encodes PTCs can elongate the reading frame (respective examples are given in section 2.6.1 Figure 2.10).

Such exons most likely are alternative ones since Pfam domains have a high sequence specificity. Consequently, it is very unlikely that the protein sequence in the other reading frame or downstream of the PTC has a high similarity to a Pfam domain just by chance.

	constitutive	alternative	P-value
total number of exons	5,728	213	-
Pfam score increase when skipped [a]	99 (1.7%)	34 (16%)	P<0.0001 [b]
average score increase[c]	2.9	13.4	P<0.0001 [d]
Pfam score increase of 10 when skipped [e]	6 (0.1%)	19 (9%)	P<0.0001 [b]

Table 2.1: Different contribution of alternative and constitutive exons to Pfam domains.
[a] counting the number of exons for those skipping yield a higher Pfam score
[b] Fisher's exact test was used to analyze a 2x2 contingency table
[c] average score increase was computed only for those exons for which skipping yield a higher score
[d] t-test was used to compare two means
[e] counting the number of exons for those skipping yield a Pfam score increase of at least 10

Apart from exon skipping, a retained intron can also encode a new part of a Pfam domain or result in a frameshift, and thus increase the score (see section 2.6.3 Figure 2.12). Therefore, we extend our strategy to include skipping of non-peptide-cassette exons and retention of introns.

2.4 General approach

Our approach can be summarized as follows. Given the exon structure of a transcript and its pre-mRNA sequence, we search for exon skipping and intron retention events that increase the Pfam score for the respective protein of at least 10 (Figure 2.1). To make sure that the Pfam annotation is highly reliable, we only considered domains with a score above the 'gathering cut-off' value that is given in the Pfam database. Without the input of additional splice sites like (known or predicted) alternative donors or acceptors, only the prediction of exon skipping and intron retention events is possible.

For the prediction, we want to consider all hypothetical splice variants and their respective proteins. Given the pre-mRNA sequence and the exon-intron structure of a transcript, a simple algorithm is as follows:

- generate all putative splice forms by allowing the skipping of exons as well as the retention of introns,
- translate each splice form,
- search the Pfam database for domain hits for each translated sequence,
- compare the score of the Pfam domains for the hypothetical protein to the score for the protein that corresponds to the given transcript,
- if a splice form leads to a score increase of at least 10, output the respective splice event.

In principle, one has to correct for multiple testing since Pfam domains are scored many times during this procedure. However, consistent with genome-wide Pfam domain anno-

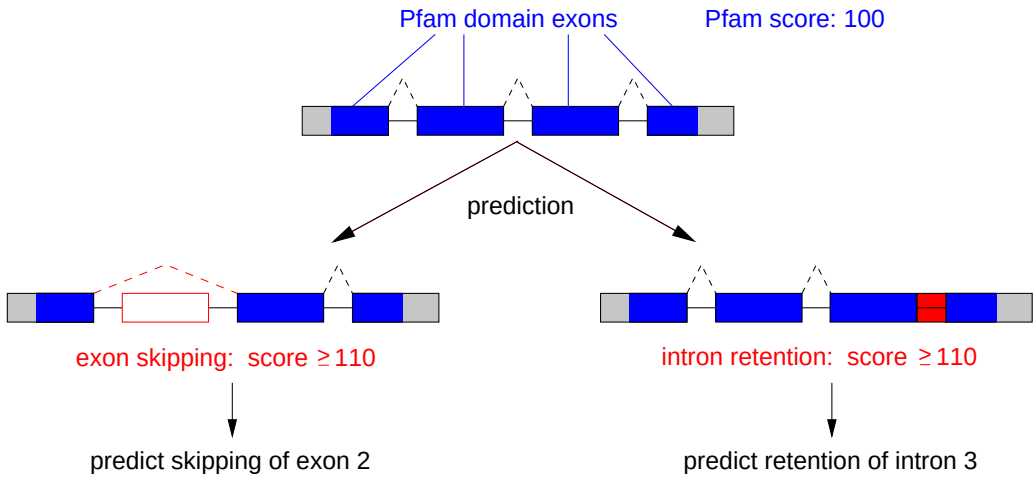

Figure 2.1: Schema of the non-EST based prediction approach.
Four coding exons and the respective intron sequences are given. Assume a Pfam domain is encoded by exons 1-4 and the respective Pfam score is 100. For the prediction, hypothetical novel splice variants are checked to find those with a higher Pfam score of at least 10. Exons are shown as boxes; dashed lines indicate the splicing patterns; open red box: skipped exon; filled red box: retained intron.

tations (like in Ensembl [98]), we consider only Pfam hits that exceed the very stringent gathering cut-off scores recommended in the Pfam database. Therefore, we do not attempt to correct for multiple testing in the following.

2.5 An efficient prediction algorithm

Since most Pfam domains have a length that exceeds the length of a typical exon (in amino acids), domains are often encoded by multiple exons. Consequently, we cannot investigate each hypothetical splice event independently. For example, the skipping of exon 3 and 5 together might result in a sufficiently high score increase but not the skipping of only one of these exons. Therefore, we have to investigate entire splice forms as the concatenation of exons and possibly introns.

The average human gene has about ten exons [2]. Considering only internal exons as candidates for alternative exons, there are $2^8 = 256$ possible combinations that arise by skipping/including of these eight exons. Thus, for a typical human gene about 256 putative splice forms have to be checked. For the human *TTN* gene with 178 exons [9], this would yield an astronomically high number of putative variants. Including intron retention further increases this exponentially high number of variants. Thus, the simple algorithm proposed in section 2.4 is computationally infeasible.

Furthermore, it is not sufficient to consider only those Pfam domains that are annotated for the protein of the given transcript. This is important because new Pfam hits

may arise from exon skipping and intron retention (section 2.6.1 Figure 2.9B). Therefore, we have to consider a larger number of Pfams, which additionally increases the computation time.

As our goal is to predict splice events in a genome-wide manner, we have to cope with this computational problem. To this end, we reformulate the given problem: Instead of searching for all splice forms yielding a higher Pfam score, we **only search the splice form yielding the highest score for a Pfam domain**. If this splice form leads to a score increase of at least 10, we will output the corresponding splice event. In the following, we show that this problem can be solved in polynomial runtime by extending the Viterbi algorithm.

2.5.1 Pfam architecture and the classical Viterbi algorithm

A Pfam domain is described as a profile hidden Markov model (HMM) [99, 100]. Profile HMMs represent a special type of HMMs and are widely used in bioinformatics [100, 101, 102]. A profile HMM is a probabilistic model of a multiple sequence alignment and contains match, insert, and delete states. While match and insert states emit characters (emitting states), delete states do not (silent states). The profile HMMs in the Pfam database allow a compact representation of the sequence information from many instances of a protein domain. One major purpose is to determine whether a given sequence belongs to a Pfam domain family or not. Profile HMMs can yield a more specific and sensitive answer compared to the pairwise alignment of the given sequence to a member of the domain family.

To find out if a given sequence belongs to the domain family, two algorithms are commonly used: the *forward algorithm* and the *Viterbi algorithm* [103]. The forward algorithm computes the sum of the scores for all paths through the HMM that output the given sequence. The Viterbi algorithm computes the score of the best path that outputs the given sequence. This best path corresponds to an alignment of a sequence to an HMM. A membership to the Pfam domain is assumed if the score of the forward or Viterbi algorithm exceeds a threshold value.

Since the software package HMMER [104] that is an inherent part of the Pfam database uses the Viterbi algorithm as the standard method to annotated Pfam domains in a sequence, we focus on the Viterbi algorithm in the following. The Viterbi (as well as the forward) algorithm is an instance of the class of dynamic programming (DP) algorithms. Common to DP algorithms is to compute partial solutions and to use these partial results to compute the solution for a bigger problem until the complete task is solved. In the Viterbi algorithm, the partial solution is the best path from the begin state to another state in the HMM.

Let $V_j(i)$ be the log-odds score of the best path through the HMM ending at state j and at the sequence position i. We denote by $E_x(y)$ the log-odds score that state x

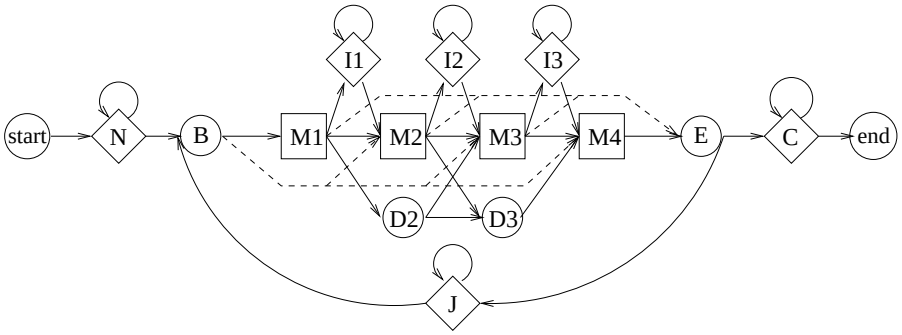

Figure 2.2: Architecture of a four match state plan7 profile HMM.
Match states are shown as squares, insert states as diamonds, and silent states as circles. The dashed lines indicate direct entry and exit transitions. Note that the architecture demands that at least one character is emitted in each iteration $B \rightarrow M_i \rightarrow \ldots \rightarrow E \rightarrow J \rightarrow B$.

emits character y and by $A_{x,y}$ the log-odds score for the transition from state x to state y. Let $\mathcal{S}$ be the input sequence with length L. We write $P(x)$ for the set of predecessors of state x, i.e. all states that have a direct transition to x. The Viterbi algorithm is:

- initialization:

$$V_j(0) = \begin{cases} 0 & \text{if } j = start \\ -\infty & \text{otherwise} \end{cases}$$

- recursion ($\forall$ states j and $i = 1 \ldots L$):

$$V_j(i) = E_j(\mathcal{S}[i]) + \max_{p \in P(j)} (V_p(i-1) + A_{p,j})$$

- termination:

$$Sc = \max_j (V_j(L) + A_{j,end})$$

where *start* is the start-state, *end* is the end-state, and Sc is the final score for the best path. Consistent with the Pfam-HMM architecture, we assume here that *start* and *end* are just silent states representing the start and the end of the model. These two states are not considered in the recursion equation.

The architecture of the profile HMMs in the Pfam database is called *plan7* (Figure 2.2). Plan7 means that direct transitions between insert and delete states and vice versa are not allowed. The main model is separated by two silent states (B- and E-state). Furthermore, there are three special insert states: one before the main model (N-state), one after the main model (C-state), and one allowing multiple iterations through the main model (J-state). Note that the special insert states only emit characters on the loop transition (the transition to itself). Moreover, there are direct entry transitions from B-state to any match state as well as direct exit transitions from any match state to E-state (dashed lines in Figure 2.2).

2.5.2 Nucleotide-level HMM target

As our primary goal is to apply the extended Viterbi algorithm to Pfam profile HMMs, we will specify the recursion equations for the plan7 architecture. Of course, the algorithm is not restricted to plan7. The sophisticated part of the algorithm is the necessity to handle frameshifts occurring during the skip process and to assure the existence of an open reading frame (ORF). Thus, for ease of understanding, we first describe the algorithm for a nucleotide-level HMM (i.e. an HMM that is built from a set of nucleotide sequences) and give the description for a protein-level HMM in section 2.5.3. Furthermore, we first consider only exon skipping events and show the extension to intron retention events in section 2.5.4.

Let us formalize the problem. We are given a transcript Tr with n exons $e_1, \ldots, e_n$ and an HMM H. The binary vector $s = (s_1, \ldots, s_n)$ denotes a *splice form* with $s_i = 1$ if exon i is included and 0 if exon i is skipped. Furthermore, $splice = \{s \mid s = (s_1, \ldots, s_n),\ s_i \in \{0,1\}, 1 \leq i \leq n\}$ is the set of all possible splice forms. Let $mRNA(s)$ be the concatenated mRNA sequence of all exons that are included in s and $Sc(H, mRNA(s))$ the Viterbi log-odds score of the sequence $mRNA(s)$ and the HMM H. Our algorithm computes the splice form that maximizes the Viterbi score $Sc(H, mRNA(s))$, that is

$$s_{max} = \underset{s \,\in\, splice}{\operatorname{argmax}} \{Sc(H, mRNA(s))\}.$$

The basic idea is to include exon skipping during the calculation of the dynamic programming matrix. Since an HMM can be divided into emitting and silent states, we have to determine which states allow for exon skipping. Clearly, exon skipping has to be handled for all emitting states. In contrast, a silent state always has an emitting state as (indirect) predecessor where the current character is emitted. Hence, we can use standard recursions for silent states and only extend the recursion equations for emitting states.

We denote by $\mathcal{S}$ the concatenated nucleotide sequence of all exons of Tr. Let $\mathcal{B}_l = \{b_2^l, ..., b_n^l\}$ be the set of left boundaries for the exons 2 to n where b_i^l is the position of the first base of exon i in $\mathcal{S}$. Let $\mathcal{B}_r = \{b_1^r, ..., b_{n-1}^r\}$ be the set of right boundaries for exon 1 to $n-1$ where b_i^r is the position of the last base of exon i in $\mathcal{S}$. The sets of left and right boundaries correspond exactly to the set of splice sites (Figure 2.3). Our algorithm requires $\mathcal{S}$, $\mathcal{B}_l$ and $\mathcal{B}_r$ as input.

The recursion equation for the extended Viterbi algorithm is:

$$V_j(i) = \begin{cases} E_j(\mathcal{S}[i]) + back(j, i-1) & \text{if } i \notin \mathcal{B}_l \text{ and } j \text{ emitting} \\ E_j(\mathcal{S}[i]) + \underset{\substack{r \,\in\, \mathcal{B}_r, \\ r \,<\, i}}{\max} back(j, r) & \text{if } i \in \mathcal{B}_l \text{ and } j \text{ emitting} \\ back(j, i) & \text{if } j \text{ silent} \end{cases}$$

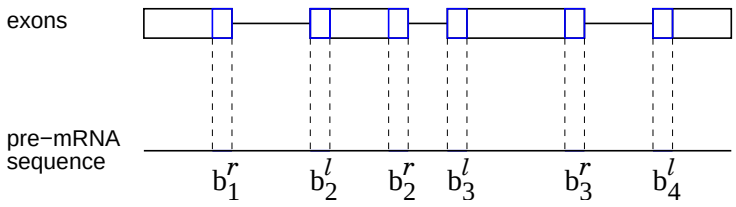

Figure 2.3: Illustration of the left and right boundaries.
Exons are shown as boxes. The first (left boundary) and last position (right boundary) is highlighted in blue. Note that the beginning of the first exon and the end of the last exon are no splice sites. Thus, they are not contained in the set of boundaries $\mathcal{B}_l$ and $\mathcal{B}_r$.

where $back(j,i) = \max_{p \in P(j)} \{V_p(i) + A_{p,j}\}$.

Note that the definition of $V_j(i)$ implies only that sequence position i is reached, not that the complete subsequence $\mathcal{S}[1 \ldots i]$ is emitted. Thus, $V_j(i)$ gives the score for the best alignment, which ends at state j, of the best concatenation of upstream exons up to sequence position i.

Since this algorithm guarantees that only disjoint sequence parts (bounded by elements from $\mathcal{B}_l$ and $\mathcal{B}_r$) are concatenated, it finds the best non-overlapping concatenation of exons. This means that $\mathcal{B}_l$ and $\mathcal{B}_r$ can be extended by alternative 5' and 3' ends of exons, respectively, to allow for alternative donors and acceptors.

2.5.3 Protein-level HMM target

Now, we describe how the algorithm can be modified for an amino acid level HMM so that frameshifts as well as correct open reading frames can be handled simultaneously. Since not all exon lengths are multiples of three nucleotides, frameshifts occur during exon skipping.

According to section 2.5.2, now the problem is the computation of

$$s_{max} = \operatorname*{argmax}_{s \in splice} \{Sc(H, AA(s))\}.$$

where $AA(s)$ is the translated mRNA sequence $mRNA(s)$. To switch to protein level, we consider the current sequence position in $\mathcal{S}$ as the third codon position and translate the codon consisting of the current and the two previous nucleotides. Then, the step length is set to three, i.e. we access $V_j(i-3)$ when computing $V_j(i)$. We extend the Viterbi algorithm to include all three reading frames. It follows that exon skipping is allowed if the current sequence position is not more than 2 nt away from a left exon boundary. Furthermore, each exon skipping variant can lead to a different codon and thus to a different amino acid (illustrated in Figure 2.4)

Let H be a plan7 profile HMM with m match states, $m-1$ insert states, and $m-2$ delete states. According to the notation in [105], $V_j^M(i)$ is the log-odds score of the best

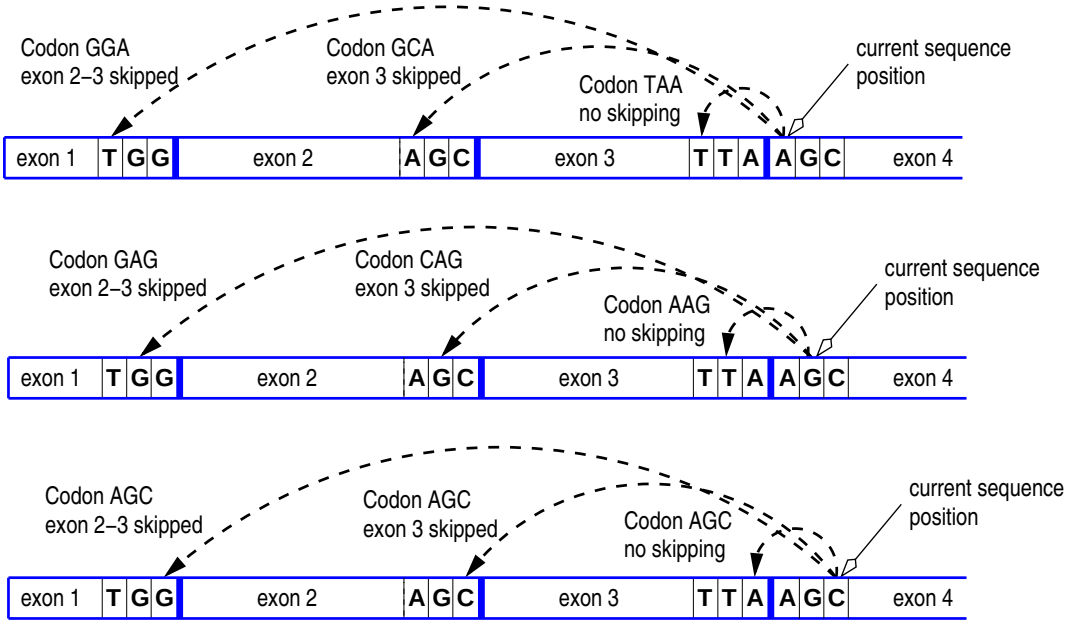

Figure 2.4: Handling of exon skipping and frameshifts in the recursion equation.
The figure illustrates how exon skipping and the three reading frames are handled for positions with a maximal distance of 2 nt to the left exon boundary. Different codons arise from different exon skipping events. While computing the DP matrix from left to right, we access already precomputed entries (to the left). The arrows indicate the sequence position where the DP matrix is accessed during the recursion.

path through the HMM ending at match state j at sequence position i. Similarly, $V_j^I(i)$ and $V_j^D(i)$ are defined for insert and delete states, respectively, and $V^X(i)$ for the special states where $X \in \{start, N, B, J, E, C, end\}$.

With $\mathcal{B}_l \oplus 1$ ($\mathcal{B}_l \oplus 2$) we denote the set $\{b_2^l+1, ..., b_n^l+1\}$ ($\{b_2^l+2, ..., b_n^l+2\}$). Furthermore, we write $\text{codon}_{i,j,k}^{\mathcal{S}}$ for the amino acid that corresponds to the codon $\mathcal{S}[i]\mathcal{S}[j]\mathcal{S}[k]$. Hence, we get the following recursion equation for the M-states.

$$V_j^M(i) = \begin{cases} \max\limits_{\substack{r \in \mathcal{B}_r, \\ r < i}} \left\{ E_{M_j}(\text{codon}_{r-1,r,i}^{\mathcal{S}}) + \text{back}^M(j, r-2) \right\} & \text{if } i \in \mathcal{B}_l \\ \max\limits_{\substack{r \in \mathcal{B}_r, \\ r < i-1}} \left\{ E_{M_j}(\text{codon}_{r,i-1,i}^{\mathcal{S}}) + \text{back}^M(j, r-1) \right\} & \text{if } i \in \mathcal{B}_l \oplus 1 \\ \max\limits_{\substack{r \in \mathcal{B}_r, \\ r < i-2}} \left\{ E_{M_j}(\text{codon}_{i-2,i-1,i}^{\mathcal{S}}) + \text{back}^M(j, r) \right\} & \text{if } i \in \mathcal{B}_l \oplus 2 \\ E_{M_j}(\text{codon}_{i-2,i-1,i}^{\mathcal{S}}) + \text{back}^M(j, i-3) & \text{otherwise} \end{cases}$$

$$\text{where back}^M(j,r) = \max \begin{cases} V^M_{j-1}(r) + A_{M_{j-1},M_j} \\ V^I_{j-1}(r) + A_{I_{j-1},M_j} \\ V^D_{j-1}(r) + A_{D_{j-1},M_j} \\ V^B(r) \ + A_{B,M_j} \end{cases}$$

The recursion equation for the I-states is

$$V^I_j(i) = \begin{cases} \max\limits_{\substack{r \in \mathcal{B}_r, \\ r < i}} \left\{ E_{I_j}(\text{codon}^{\mathcal{S}}_{r-1,r,i}) + \text{back}^I(j, r-2) \right\} & \text{if } i \in \mathcal{B}_l \\ \max\limits_{\substack{r \in \mathcal{B}_r, \\ r < i-1}} \left\{ E_{I_j}(\text{codon}^{\mathcal{S}}_{r,i-1,i}) + \text{back}^I(j, r-1) \right\} & \text{if } i \in \mathcal{B}_l \oplus 1 \\ \max\limits_{\substack{r \in \mathcal{B}_r, \\ r < i-2}} \left\{ E_{I_j}(\text{codon}^{\mathcal{S}}_{i-2,i-1,i}) + \text{back}^I(j, r) \right\} & \text{if } i \in \mathcal{B}_l \oplus 2 \\ E_{I_j}(\text{codon}^{\mathcal{S}}_{i-2,i-1,i}) + \text{back}^I(j, i-3) & \text{otherwise} \end{cases}$$

$$\text{where back}^I(j,r) = \max \begin{cases} V^M_j(r) + A_{M_j,I_j} \\ V^I_j(r) \ + A_{I_j,I_j} \end{cases}$$

The recursion equation for the special insert state C is a little tricky, since C acts as both a silent and a non-silent state. Characters are only emitted via the loop transition, so exon skipping will only be handled for the $C \rightarrow C$ transition. This yields the following equation:

$$V^C(i) = \max \begin{cases} \begin{cases} \max\limits_{\substack{r \in \mathcal{B}_r, \\ r < i}} \left\{ E_C(\text{codon}^{\mathcal{S}}_{r-1,r,i}) + \text{back}^C(r-2) \right\} & \text{if } i \in \mathcal{B}_l \\ \max\limits_{\substack{r \in \mathcal{B}_r, \\ r < i-1}} \left\{ E_C(\text{codon}^{\mathcal{S}}_{r,i-1,i}) + \text{back}^C(r-1) \right\} & \text{if } i \in \mathcal{B}_l \oplus 1 \\ \max\limits_{\substack{r \in \mathcal{B}_r, \\ r < i-2}} \left\{ E_C(\text{codon}^{\mathcal{S}}_{i-2,i-1,i}) + \text{back}^C(r) \right\} & \text{if } i \in \mathcal{B}_l \oplus 2 \\ E_C(\text{codon}^{\mathcal{S}}_{i-2,i-1,i}) + \text{back}^C(i-3) & \text{otherwise} \end{cases} \\ V^E(i) + A_{E,C} \end{cases}$$

where $\text{back}^C(r) = V^C(r) + A_{C,C}$

A graphical illustration of these extended recursion equations is given in Figure 2.5.

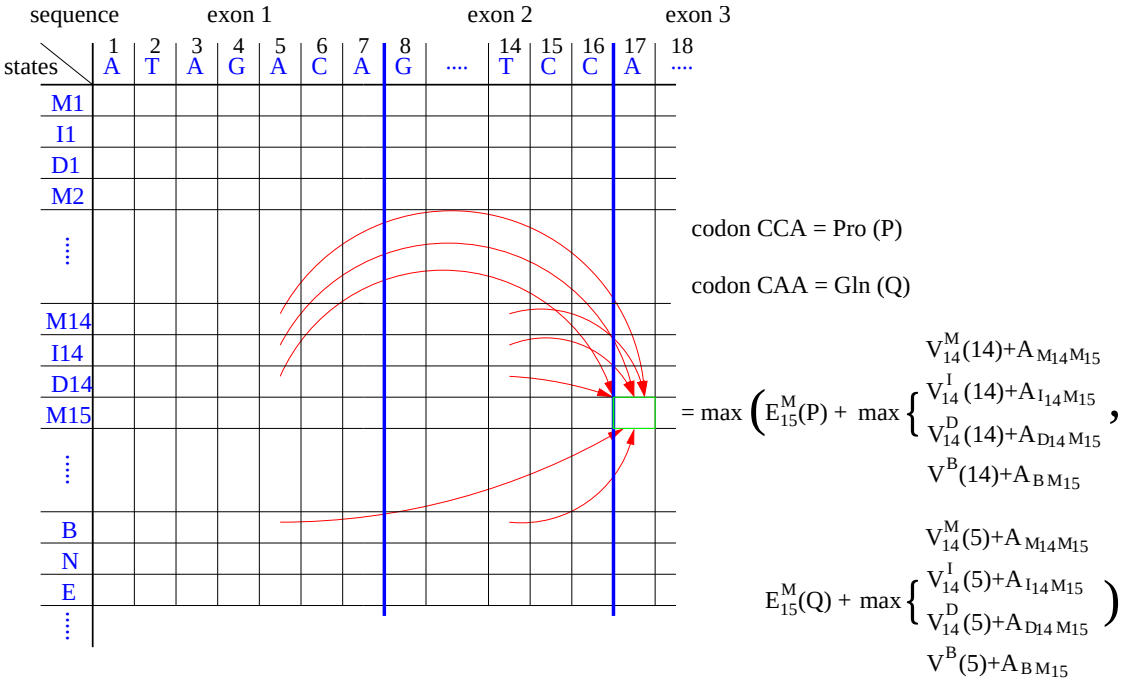

Figure 2.5: Illustration of the extended recursion equations.
The figure shows a part of the DP matrix with the HMM states as lines and the nucleotides as columns. Vertical blue lines represent exon-exon boundaries. The current sequence position is the first position of exon 3 (highlighted in green). Red arrows indicate precomputed matrix entries that are accessed.

The equations for the N-state and the J-state are similar to the C-state equation and are not shown. For completeness, we show the recursions for the silent states:

$$V_j^D(i) = \max \begin{cases} V_{j-1}^M(i) + A_{M_{j-1},D_j} \\ V_{j-1}^D(i) + A_{D_{j-1},D_j} \end{cases}$$

$$V^B(i) = \max \begin{cases} V^N(i) + A_{N,B} \\ V^J(i) + A_{J,B} \end{cases}$$

$$V^E(i) = \max_{j=1,\ldots,m} \{V_j^M(i) + A_{M_j,E}\}$$

$$V^{end}(i) = V^C(i) + A_{C,end}$$

Of course, not all possible splice forms will form an ORF since some of them might lack a start and/or stop codon. However, the start and stop codon condition can easily be included in the algorithm. Since matrix entries for the *start*-state are not computed but initialized, we set all of them to $-\infty$ except for the positions where a start codon begins:

$$V^{start}(i) = \begin{cases} 0 & \text{if} \quad \mathcal{S}[i+1]\mathcal{S}[i+2]\mathcal{S}[i+3] = \text{ATG} \\ -\infty & \text{otherwise} \end{cases}$$

Per definition, an ORF ends at the first stop codon. To take this into account, we define

$$E_x(\text{codon}^{\mathcal{S}}_{i,j,k}) = -\infty \quad \text{if} \quad \mathcal{S}[i]\mathcal{S}[j]\mathcal{S}[k] \in \{\text{TGA,TAA,TAG}\}$$

for all non-silent states x. The emission score for the 20 amino acids are taken from the Pfam database. Furthermore, we have to compute the set of possible *end positions*, taking into account that a stop codon can be assembled on exon boundaries. Therefore, during the dynamic programming procedure, we compute all positions after which a stop codon occurs and denote this set as $EndPos$

$$EndPos = \left\{ \begin{array}{lcl} r-2 & | & \mathcal{S}[r-1]\mathcal{S}[r]\mathcal{S}[i] \in \{\text{TGA,TAA,TAG}\} : \\ & & \quad r \in \mathcal{B}_r,\, r < i,\, i \in \mathcal{B}_l \\ r-1 & | & \mathcal{S}[r]\mathcal{S}[i-1]\mathcal{S}[i] \in \{\text{TGA,TAA,TAG}\} : \\ & & \quad r \in \mathcal{B}_r,\, r < i-1,\, i \in \mathcal{B}_l \oplus 1 \\ r & | & \mathcal{S}[i-2]\mathcal{S}[i-1]\mathcal{S}[i] \in \{\text{TGA,TAA,TAG}\} : \\ & & \quad r \in \mathcal{B}_r,\, r < i-2,\, i \in \mathcal{B}_l \oplus 2 \\ i-3 & | & \mathcal{S}[i-2]\mathcal{S}[i-1]\mathcal{S}[i] \in \{\text{TGA,TAA,TAG}\} : \\ & & \quad i \notin \mathcal{B}_l \cup \mathcal{B}_l \oplus 1 \cup \mathcal{B}_l \oplus 2 \end{array} \right\}$$

Finally, the highest Viterbi score Sc (implicitly considering all possible splice forms) is given by

$$Sc = \max\, \{V^{end}(i) \,|\, i \in EndPos\}.$$

A normal traceback determines s_{max}. Backtracking from other positions in $EndPos$ and choosing suboptimal paths on a traceback can be used to find suboptimal splice form predictions.

2.5.4 Including intron retention events

To allow for intron retention events, the sequence $\mathcal{S}$ is now the pre-mRNA sequence (i.e. the sequence of all exons and introns). Given a position at the left boundary of exon i, the recursion equations in section 2.5.3 allow only to access matrix columns that correspond to the end of the upstream exons $1 \ldots i-1$. To include intron retention, we extend the equations by adding the possibility to access the position 3 nt upstream (Figure 2.6). Thus, the equation for the M-states is

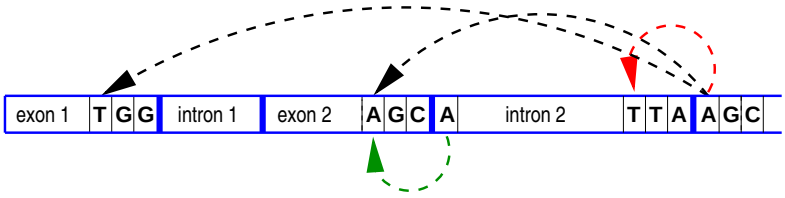

Figure 2.6: Extending the equations to allow intron retention events.
The red arrow indicates the access of matrix columns that correspond to the end of the intron immediately upstream. Please note that the beginning of the intron is not part of $\mathcal{B}_l$, thus only a position within the upstream exon is possible (green arrow).

$$V_j^M(i) = \begin{cases} \max \left\{ \begin{array}{l} \max\limits_{\substack{r \in \mathcal{B}_r, \\ r < i}} \left\{ E_{M_j}(\text{codon}_{r-1,r,i}^{\mathcal{S}}) + \text{back}^M(j, r-2) \right\} \\ E_{M_j}(\text{codon}_{i-2,i-1,i}^{\mathcal{S}}) + \text{back}^M(j, i-3) \end{array} \right. & \text{if } i \in \mathcal{B}_l \\ \max \left\{ \begin{array}{l} \max\limits_{\substack{r \in \mathcal{B}_r, \\ r < i-1}} \left\{ E_{M_j}(\text{codon}_{r,i-1,i}^{\mathcal{S}}) + \text{back}^M(j, r-1) \right\} \\ E_{M_j}(\text{codon}_{i-2,i-1,i}^{\mathcal{S}}) + \text{back}^M(j, i-3) \end{array} \right. & \text{if } i \in \mathcal{B}_l \oplus 1 \\ \max \left\{ \begin{array}{l} \max\limits_{\substack{r \in \mathcal{B}_r, \\ r < i-2}} \left\{ E_{M_j}(\text{codon}_{i-2,i-1,i}^{\mathcal{S}}) + \text{back}^M(j, r) \right\} \\ E_{M_j}(\text{codon}_{i-2,i-1,i}^{\mathcal{S}}) + \text{back}^M(j, i-3) \end{array} \right. & \text{if } i \in \mathcal{B}_l \oplus 2 \\ E_{M_j}(\text{codon}_{i-2,i-1,i}^{\mathcal{S}}) + \text{back}^M(j, i-3) & \text{otherwise} \end{cases}$$

Likewise, the equations for I-, C-, N-, and J-states are extended. The retention of an intron always implies that both its upstream and its downstream exon are also included in the mature mRNA. This condition is fulfilled since only left boundaries allow to skip an upstream sequence part and the beginning of an intron is not included in the set of left exon boundaries $\mathcal{B}_l$. Thus, the retention of intron i always leads to the inclusion of exon $i-1$ and exon i (Figure 2.6).

2.5.5 Runtime analysis

The runtime of the algorithm is as follows. L is the length of $\mathcal{S}$ and $n-1$ is the number of boundaries. Let $M = m + (m-1) + (m-2) + 7$ be the number of states in the plan7 profile HMM.

If a sequence position is not in $\mathcal{B}_l$, $\mathcal{B}_l \oplus 1$, or $\mathcal{B}_l \oplus 2$, the runtime for one matrix entry is $O(1)$ for all states except for the E-state with $O(m)$. Thus, the total runtime for one such column in the DP matrix is $(M-1) \cdot O(1) + O(m) = O(M+m)$. Since $n << L$ there are $L - 3(n-1) \approx L$ sequence positions that do not allow for skipping. Thus, the

total runtime for all these position is approximately $O((M + m) \cdot L)$, which is also the runtime of the standard Viterbi algorithm for a plan7 HMM.

If a sequence position allows exon skipping, one matrix entry can be computed in $O(n)$ for all emitting states (M, I, C, J, N), in $O(m)$ for the E-state and in $O(1)$ for all silent states (D, B, *end*). One such matrix column is computed in $(m + m - 1 + 3) \cdot O(n) + O(m) + (m) \cdot O(1) = O(m \cdot n + m) = O(m \cdot n)$. The overall runtime for the $3(n - 1)$ matrix columns is $O(m \cdot n \cdot n) = O(m \cdot n^2)$.

The total runtime of the extended Viterbi algorithm is $O((M + m) \cdot L + m \cdot n^2)$. Despite the number of 2^{n-2} hypothetical splice forms, the runtime of our algorithm is only quadratic with respect to the number of exons.

Compared to the Viterbi algorithm for a Pfam HMM, our algorithm is about three-times slower, since we are working at the nucleotide and not at the protein level. Furthermore, we use the pre-mRNA sequence, which results in much larger input sequences, since introns are on average 23-times longer than exons [9].

2.5.6 Validation of the algorithm

To test the ability of our algorithm to identify real alternative exons, we constructed a test set of alternative exons that are skipped in a RefSeq transcript and retained in other EST/cDNA sequences. This test set consists of 202 peptide-cassette and 195 non-peptide-cassette exons. Inclusion of these exons results in a Pfam score decrease of at least 10. Then, given the exon structure of the transcript including the alternative exon, we used the algorithm to find the splice form with the highest Pfam score. In 392 (99%) cases (200 peptide-cassette and 192 non-peptide-cassette exons), the predicted splice form was equal to the RefSeq transcript. This includes 18 cases where more than one exon was skipped in a RefSeq transcript (four cases with two consecutive exons and 14 cases with two non-consecutive ones). In the five remaining cases, other exons were skipped in addition to the expected exon, which gives an even higher score. Thus, all alternative exons in our test set can be found by this algorithm. This demonstrates that the algorithm can be used to predict alternative splice events that result in a Pfam score increase with a high sensitivity.

2.6 Genome-wide prediction of alternative splice events

To predict exon skipping and intron retention events in the entire human genome, we applied this approach to all 18,572 RefSeq transcripts (August 2004). We only considered novel splice forms that are not candidates for nonsense-mediated mRNA decay (NMD) [58], since the rationale behind our strategy is that the novel splice variant is expressed to be translated into a functional protein. To get highly confident Pfam annotations, we only considered predictions with a Pfam score above the 'gathering cut-off'

	number of predictions	confirmed[a]		different event confirmed[b]		unconfirmed	
single exon skipping	183	119	65%	25	14%	39	21%
multiple exon skipping	57	14	25%	16	28%	27	47%
intron retention	67	37	55%	28	42%	2	3%
hidden exon event	5	-	-	5	100%	-	-
complex event	9	-	-	6	67%	3	33%
sum	321	170	53%	80	25%	71	22%

Table 2.2: Summary of the genome-wide scan.
[a] exactly the predicted event is confirmed
[b] a different event is confirmed (alternative donor/acceptor, inclusion of an exon that is skipped in the given transcript, alternative transcription start); most of these events involve frameshifts

value as given in the Pfam database.

Despite the efficient algorithm, the computational expense to consider all Pfam domains for 18,572 transcripts is huge. Therefore, we reduced the total runtime by restricting the search for one transcript to those Pfams that match the RefSeq protein up to a rather low significance level (E-value ≤ 10). All Pfams matching the protein with an E-value above 10 were not considered. This procedure is based on the observation that Pfams matching already the RefSeq protein with a moderate score are more likely to yield a hit above the gathering cut-off value for a new splice form.

In this genome-wide scan, we predicted alternative exons and introns for 309 RefSeq transcripts. For the purpose to simplify the following evaluation, we distinguish five cases:

1. the skipping of a single exon,
2. the skipping of multiple consecutive exons,
3. the retention of an intron,
4. hidden exon events,
5. and complex events as any combination of 1-4.

These five cases are shown in Figure 2.7. The results are summarized in Table 2.2.

2.6.1 Single skipped exons

We predicted a total of 183 single RefSeq annotated exons to be alternative. To check if known alternative exons are contained in this set, we used Blast with a 60 nt search string from the flanking exons (30 nt from the upstream and 30 nt from the downstream exon) to search dbEST (December 2004) and cDNAs from GenBank. We denote a predicted alternative splice event as *confirmed* if there is EST/cDNA evidence for it and *unconfirmed* if there is currently no EST/cDNA evidence. This notation takes into account that current EST/cDNA data is incomplete.

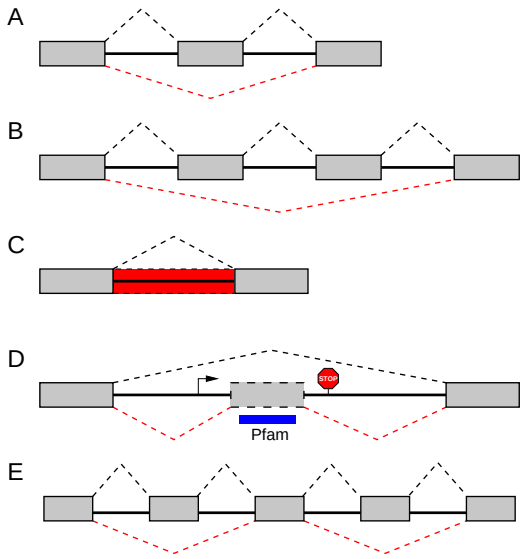

Figure 2.7: Classification of alternative splice events predicted by the algorithm.
(A) Skipping of a single exon. (B) Skipping of multiple consecutive exons (shown here for two exons). (C) Intron retention (shown as red dashed box). (D) Hidden exon: the algorithm found an open reading frame (start and stop codon are indicated) within an intron that encodes a Pfam domain (blue box), which indicates the existence of a hidden exon (grey dashed box). (E) Complex events involve any combination of A-D. Here, we show an example with two skipped exons. Exons are shown as grey boxes and dashed lines indicate the splicing patterns.

We found exon skipping evidence for 119 (65%) of the 183 exons. Three further exons are skipped in addition to alternative donor or acceptor usage of one neighboring exon. As mentioned in section 2.3, a frameshift introduced by exon skipping leads to a new protein sequence, which can encode a longer or a new Pfam domain. While there are generally several possibilities to introduce the frameshift, our algorithm is only able to handle frameshifts caused by exon skipping or intron retention, since no other splice sites are given. However, the same frameshift might be introduced by the usage of alternative donor/acceptor sites or the inclusion of exons that are skipped in the RefSeq transcript (Figure 2.8, see also Figure 2.14). Therefore, we examined frameshift predictions in detail and found that in 22 cases the EST confirmed frameshift is not caused by exon skipping but by a different splice event. Remarkably, the target reading frame of the predicted shift is always identical to the confirmed one. Thus, a frameshift prediction should be taken as a strong hint that a frameshift event exists in the vicinity of the skipped exon. These 22 predictions are not considered further. Altogether only 39 (21% of the 183) predictions remain that cannot be confirmed by existing expressed sequences.

Then, we compared the number of ESTs that match the upstream and downstream exon of confirmed and unconfirmed predictions to see whether the exon skipping events in both groups have an equal chance to be detected. The downstream exon of the 119

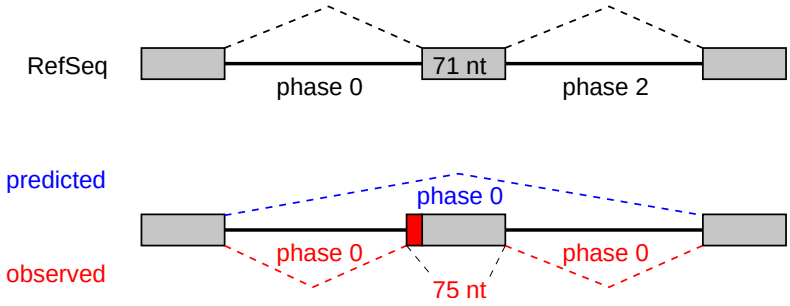

Figure 2.8: Illustration of multiple possibilities to introduce a frameshift.
The annotated RefSeq transcript contains a 71 nt exon. Inclusion of this exon leads to phase 2 for the downstream intron. Our algorithm is only able to introduce a frameshift by exon skipping (shown in blue), resulting in phase 0 for the downstream intron. However, ESTs indicate that the same frameshift is introduced by using an alternative acceptor site 4 nt upstream (shown in red). This alternative acceptor leads to a 75 nt exon and consequently also to phase 0 for the downstream intron. In principle, the same frameshift can be introduced by any other splice event. Likewise, our algorithm might introduce a frameshift by skipping multiple exons.

confirmed alternatives is covered on average by 81 ESTs, which is four-times higher than the average coverage of 20 for the unconfirmed predictions (median 14 vs. 5). The upstream exon has similar EST counts in both groups (average 77 vs. 13). This suggests that insufficient EST coverage may be the reason for the current lack of confirmation. Furthermore, we found that the unconfirmed exons are on average 688 nt further upstream of the 3' mRNA end. Given the average EST length of 530 nt and that most ESTs are sampled from the 3' end, this may contribute to their lower EST coverage.

To check which percentage of single exon skipping events can be expected by chance, we randomly chose 2,828 Pfam domain exons. To exclude exons with an EST coverage too low for detection of skipping events, we only considered exons with at least 20 hits for the up- and downstream exon, giving a median coverage of 48 (note that this is very conservative compared to 14 matches to the downstream exon of confirmed single exon skipping events). We only found for 15% (424 of 2,828 exons) EST/cDNA evidence for exon skipping. In contrast, 75% (119 of 158, excluding 25 with a different confirmed event) of the predicted single exons are EST/cDNA confirmed. This indicates that our predictions are significantly enriched in real alternative exons (Fisher's exact test: $P<0.0001$).

Then, we compared the number of ESTs/cDNAs that contain or miss a confirmed single exon. On average, these exons are skipped in 39 cases and included in only eight (5:1 skipping-inclusion ratio), which contributes to the high confirmation rate for predicted single exon events. However, the inclusion in several transcripts and at least one RefSeq demonstrates that these exons are real. Alternative exons with a low inclusion rate are often not conserved in mouse and such exons are the result of exon creation or

loss [77]. Therefore, we searched for their existence in the mouse genome by inspecting the exons as well as introns of the orthologous mouse loci. For 15 single exons we failed to identify either an orthologous mouse gene or the exons that flank the single exon. For the remaining 104 exons, we only found an orthologous mouse exon for 45 (43%), which is in agreement with [77]. In recent studies, Sorek et al. and Yeo et al. predicted a total of 952 and 2,092 exons to be alternative, respectively [92, 71]. Only 18% (21 of 119) of the confirmed single exons predicted here are contained in this combined exon set, which may be attributed to the fact that 42% (19 of 45) of the orthologous human-mouse exon pairs have sequence identities of less than 95% (this cut-off was used in [92]). Moreover, unlike our predictions, the exons predicted by Yeo et al. have a tendency not to overlap InterPro domains. Thus, the exons addressed by our Pfam based approach and the comparative methods have different characteristics and both approaches complement each other.

Finally, we analyzed how confirmed exon skipping events can lead to a Pfam score increase. We found that most of the peptide-cassette exons are aligned to gaps in the Pfam alignment. Thus, exon skipping results in a score increase by reducing the number of gaps. Such an example is shown in Figure 2.9A. Strikingly, we also found that the skipping of a peptide-cassette exon can result in the creation of a new Pfam domain. Such an example is the RefSeq transcript NM_024565 where the skipping of exon 4 results in the creation of a new 'Cyclin, N-terminal domain' (PF00134) (Figure 2.9B). The Pfam parts encoded by the individual exons 3 and 5 have scores (0.3 and 9.6) that are far below the gathering cut-off score of 17. Thus, it is likely that the inclusion of this exon leads to the destruction of this functional domain. Motivated by these examples, we performed genome-wide searches (that are not described in this thesis) and found other functional protein features such as transmembrane helices and post-translational modification sites that can be destroyed by exon inclusion. Taken together, this represents a novel mechanism of alternative splicing to modulate protein function. This mechanism creates a functional protein domain by putting together two non-consecutive exons and destroys it by inserting an exon in its body (exon 4 in the above example). Such alternative exons might only perform the function of a 'spacer'.

In contrast to peptide-cassette exons, the skipping of a non-peptide-cassette exon leads to a longer protein sequence by removing a stop codon or a new protein sequence by introducing a frameshift. This protein sequence can encode a new or a longer Pfam domain. Two representative examples are shown in Figure 2.10A and B.

2.6.2 Multiple skipped consecutive exons

In the genome-wide scan, we predicted 57 multiple exon skipping events and found EST/cDNA evidence for 14 of them. Similar to single exons, further 16 frameshift predictions are confirmed by different splice events (see Figure 2.8). The remaining

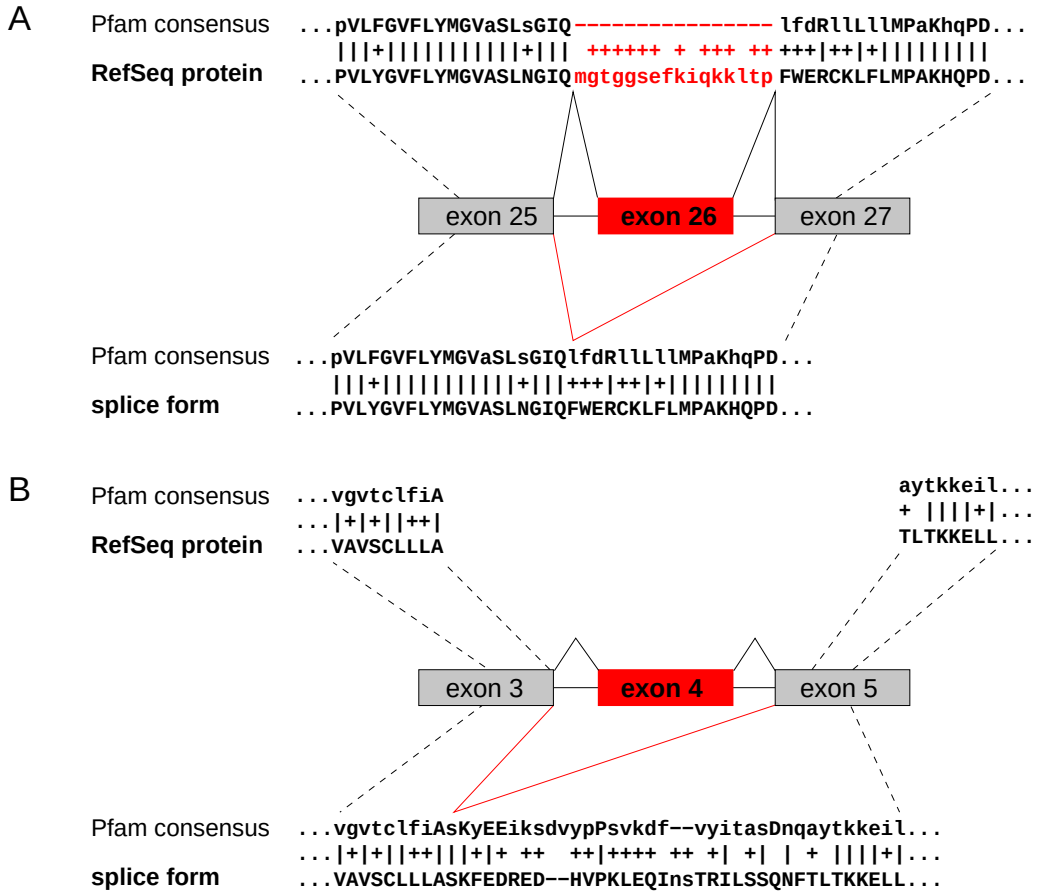

Figure 2.9: Effect of alternative peptide-cassette exons on Pfam domains.
(A) *SLC4A5* (NM_033323): Exon 26 disrupts the Pfam domain PF00955 as shown by the gaps in the alignment. Skipping of the exon increases the Pfam score from 1183 to 1208.
(B) *FLJ14166* (NM_024565): Skipping exon 4 results in the creation of a new Pfam domain, since the score of 52.9 for the exon skipping splice form exceeds the threshold of 17.
Alternative exons are depicted in red. Pfam alignments for the RefSeq protein are shown at the top, for the alternative splice form at the bottom.

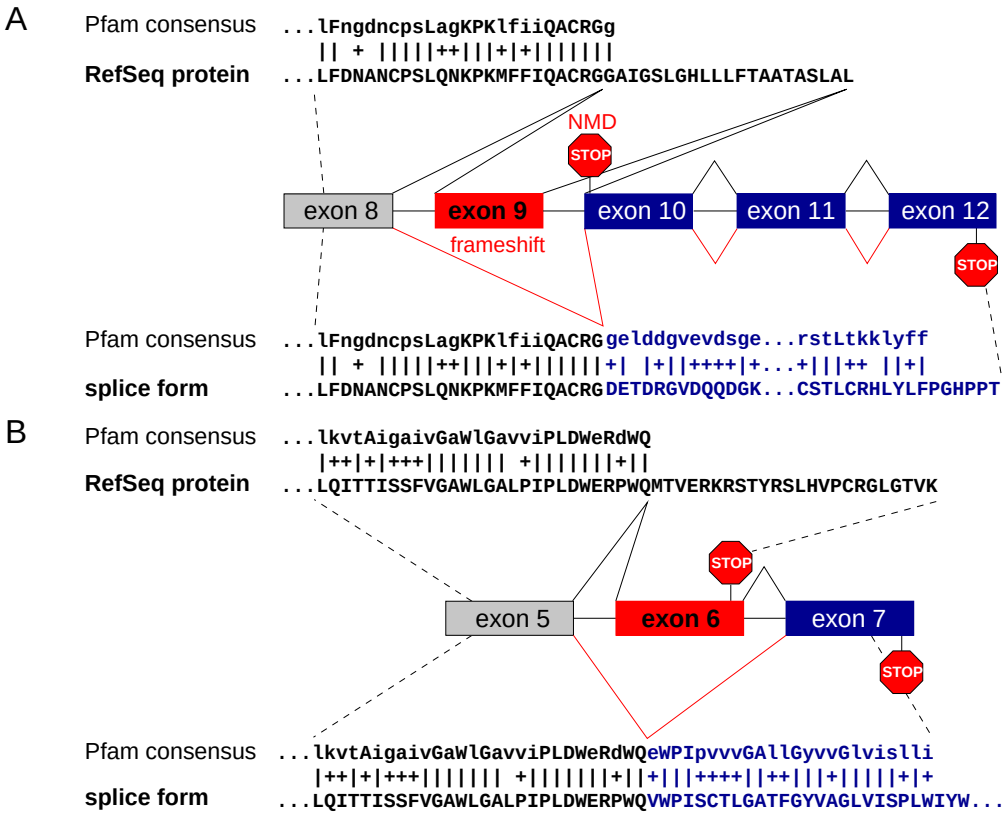

Figure 2.10: Effect of skipping of non-peptide-cassette exons on Pfam domains.
(A) *CASP2* (NM_001224): Inclusion of exon 9 results in a reading frame with a stop codon in exon 10 and this transcript should induce NMD. The skipping of exon 9 leads to a frameshift and a new C-terminal part of the Caspase domain PF00656 (score increase from 174 to 317).
(B) *PIGF* (NM_173074): Exon 6 encodes an in-frame stop codon 33 nt upstream of the last exon-exon junction, which should not elicit NMD. Skipping of exon 6 results in a new C-terminus encoded by exon 7 and a score increase for PF06699 from 299 to 362.
Alternative exons are depicted in red, exons that become coding in the alternative splice form are depicted in blue. Pfam alignments for the RefSeq protein are shown at the top, for the alternative splice form at the bottom.

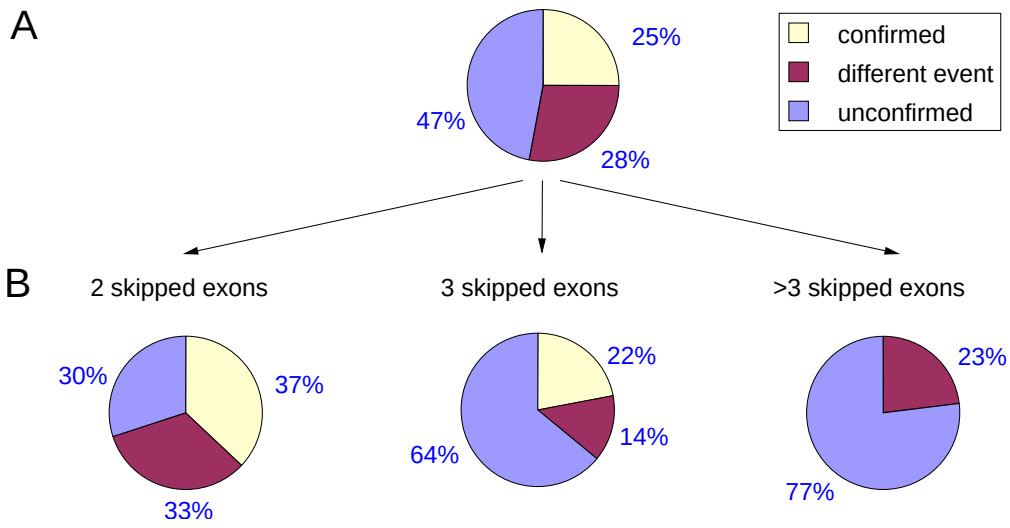

Figure 2.11: Percentage of confirmed multiple exon skipping events.
(A) Percentage of confirmed, confirmed by a different splice event, and unconfirmed predictions with more than one skipped exon. (B) Percentage of confirmed, confirmed by a different splice event, and unconfirmed predictions divided into the number of skipped exons.

27 predictions are unconfirmed (Figure 2.11A). Again EST coverage of the downstream exon is higher for the confirmed predictions compared to unconfirmed ones (average 42 vs. 23, median 25 vs. 7). Of all 57 predictions 30 events have two skipped exons, 14 three skipped exons, and 13 more than three exons. We found that no prediction with more than three exons is confirmed and that the percentage of unconfirmed predictions increases with the number of skipped exons (Figure 2.11B). Thus, it is conceivable that some predictions are false positives and that the Pfam score is increased just by chance. This holds especially for predictions with many skipped exons since the number of possible exon-exon combinations goes up. Indeed, we found that the average Pfam score increase for the unconfirmed predictions is lower than for the confirmed predictions (19 vs. 28), which suggests that an increase of the threshold value with the number of skipped exons should eliminate many false positive predictions.

2.6.3 Retained introns

We predicted 67 intron retention events that increase the Pfam score by encoding a new part of a domain or introducing a frameshift (an example is given in Figure 2.12). We found EST/cDNA evidence for 65 (97%) of these events. Only ESTs with a spliced intron up- or downstream were accepted to reduce the possibility of partially spliced ESTs. We found that 36 (54% of 67) of these introns do not have consensus splice sites (GT-AG or GC-AG). These introns can be the result of annotation or mapping errors of the RefSeq transcripts or the consequence of allele-specific splicing [106], since some

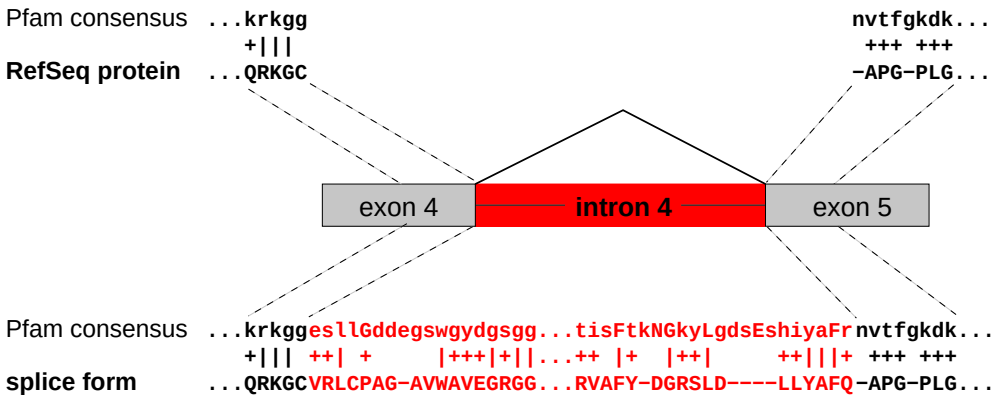

Figure 2.12: Effect of intron retention on Pfam domains.
RNF39 (NM_170769): Intron 4 encodes the middle part of the SPRY domain (PF00622). Intron retention results in a score increase from 10 to 47.
Pfam alignments for the RefSeq protein are shown at the top, for the alternative splice form at the bottom.

of those have splice sites that diverge from the consensus in only a single mutation (e.g. an AA instead of an AG acceptor site). Therefore, some of the predicted events do not involve real introns, which may contribute to this extremely high confirmation rate. However, 25 (81%) retention events of the remaining 31 introns with consensus splice sites are confirmed by other RefSeq transcripts, which indicates that they are real and not artifacts.

We classified predicted intron retentions into three groups (Figure 2.13). In case of '*I-introns*' the internal region of a Pfam is encoded by the intron and both neighboring exons also contribute to the domain. '*N-introns*' encode a novel N-terminal domain part and thus, only the downstream exons add to the Pfam. Likewise, '*C-introns*' encode a novel C-terminal Pfam part and only the upstream exons contribute to the domain. Of the 67 predictions, 23 are I-introns and all are experimentally confirmed. Of the 15 N-intron predictions, 13 are confirmed by at least partial EST matches to one intron-exon boundary. Eleven of them do not have a continuous open reading frame (i.e. an in-frame stop codon), which is a strong indication for the existence of alternative acceptor sites further upstream of the Pfam encoding exons. Indeed, ten of those have a confirmed alternative acceptor and we found one alternative transcription start. The remaining two N-introns with a continuous reading frame are confirmed by EST matches. Finally, all of the 29 predicted C-intron retention events are confirmed (twelve intron retentions and 17 alternative donors).

Most predicted intron retention events involve the last intron in the transcript since we excluded NMD candidate splice forms and a PTC due to the retention of the last intron cannot trigger NMD. Consequently, these splice events result in protein isoforms with an altered Pfam domain at their C-terminus.

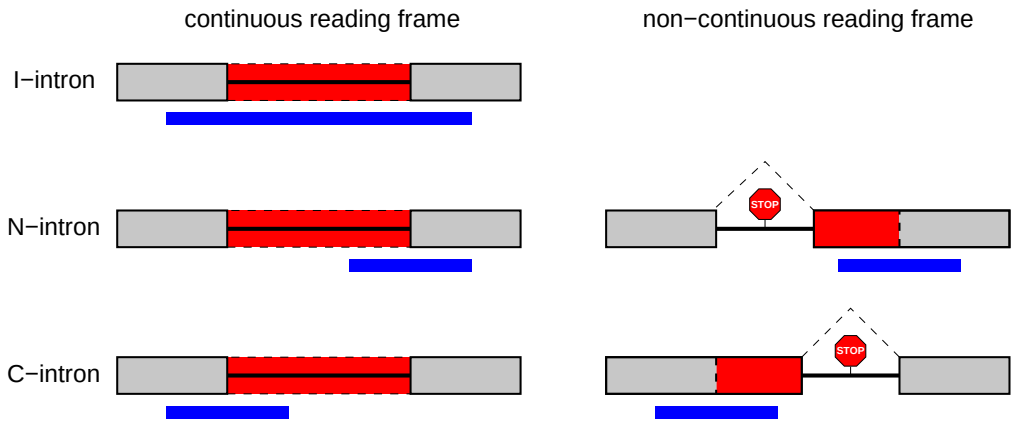

Figure 2.13: Classification of intron retention events.
I-introns have a continuous reading frame and both neighboring exons also encode the Pfam domain. For *N*- and *C-introns* only the downstream and upstream exon encode the Pfam domain, respectively, and they may not have a continuous reading frame (indicated by the stop codon). Non-continuous reading frames are a strong indication of alternative donor/acceptor sites within the intron.
Exons are shown as grey boxes with solid lines, introns as a line and a retained (partial) intron as a red box with dashed lines. The position of the Pfam domain is shown as a blue box below the gene structure. Stop codons and splicing patterns (dashed line) are indicated.

2.6.4 Hidden exon events

In the genome-wide scan, we also found seven predictions that involve introns containing an open reading frame that encodes the complete or a part of a Pfam without the neighboring exons. Thus, it is possible that an exon, which is skipped in the given transcript, is 'hidden' in the respective intron. Therefore, we examined these hits and found for five of them EST confirmation of hidden exons. For example, intron 5 of the NM_013954 transcript of *NOX1* contains seven alternative exons that encode parts of the 'Ferric reductase like transmembrane component' domain (PF01794). These exons are included in another transcript of *NOX1* (NM_007052). Manual inspection of the remaining two unconfirmed predictions (NM_152476 intron 10, NM_206894 intron 5) with the Ensembl genome browser revealed that these RefSeq transcripts falsely span two non-overlapping genes and that the predicted intronic parts are exons of the downstream genes. Thus, these two cases are likely due to annotation errors and were excluded. (It should be mentioned that a very recent study demonstrated that such "transcription-induced chimerism events" can really occur in human cells [107]).

2.6.5 Complex events

We also predicted nine complex events. In each case, the given transcript is a clear NMD candidate and our prediction aims at maintaining a reading frame. For six of

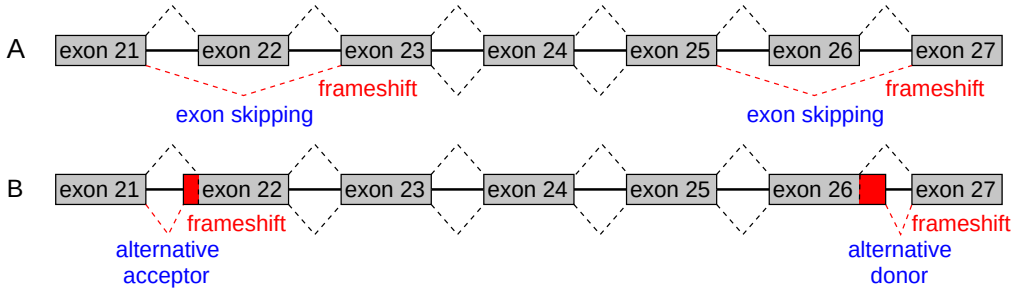

Figure 2.14: Example of a predicted complex splice event for the *CPT1B* gene.
The predicted splice event (skipping of two exons) is shown in (A), the observed event (alternative donor and acceptor usage) in (B).

the nine cases, manual inspection revealed other splice events like the multiple usage of alternative donor and acceptor sites. For example, our prediction for the NMD candidate NM_152247 of *CPT1B* is to skip exon 22 and 26 to restore the reading frame. Instead of skipping two exons, an alternative acceptor 5 nt upstream of the beginning of exon 22 and an alternative donor 169 nt downstream of exon 26 is used in another transcript of *CPT1B* (NM_152246) to produce a non-NMD splice form (Figure 2.14).

2.6.6 Experimental verification of unconfirmed predictions

In collaboration with the genome analysis group of Matthias Platzer, we tested eleven randomly chosen unconfirmed single exon skipping events in a pool of 16 human tissues. In 27% (three of eleven) the predicted exon skipping was observed (*DHRS* exon 7, *CDH2* exon 11, and *MYO9* exon 6). Since multiple exon skipping events have a lower EST confirmation rate compared to single exon events and no case of a four-exon skipping is EST confirmed, we selected three two-exon, one three-exon, and two four-exon skipping events for experimental verification. Furthermore, the two unconfirmed N-introns were tested. We did not observe the expected splice variants for these eight predictions. In general, our experiments may suffer from some of the problems mentioned above for ESTs, since specific splice events can be restricted to narrow windows in space and time.

2.6.7 Location of alternative peptide-cassette exons within Pfam domain structures

Alternative splicing has a tendency to coincide with domain boundaries and to avoid the interior of functional and structural domains [108, 52]. Since our single exon events might interfere with the Pfam domain structure as indicated by the low inclusion rate, we were interested in finding out where confirmed peptide-cassette exons are located with respect to the secondary structure and the protein surface. We computed the secondary structures and the surface accessibility of residues from known three-dimensional struc-

tures of Pfam domains using the *pdb2pfam* function of the Pfam web sites. Since in each case the structure does not include the alternative exon, we consider the location of the exon-exon junction of both neighboring exons in the following. We mapped 28 alternative exon junctions and, as a control group, 80 constitutive exon junctions to these secondary structures. The residue at the exon junction was classified as being located in an alpha-helix, in a beta-sheet, or in a non-regular element. We found a significant difference between the alternative and constitutive junctions (χ^2 test was used to analyze a 2x3 contingency table, P=0.034) with a striking preference of alternative junctions for non-regular elements and the avoidance of helices (Figure 2.15A).

To rule out that this result is biased by inaccuracies in the secondary structure assignment, which is sometimes problematic at the end of structural elements, we considered a broader context (± one residue) around the exon junction. We classified the context to be '*inside a structural element*' if all three residues are either in a helix or in a sheet. If the three residues of the context are in two different structural elements or if all are inside a non-regular element, the context is classified as '*outside a structural element*'. Again we found a significant preference of the alternative exon junctions to be located outside structural elements (Fisher's exact test: P=0.043) (Figure 2.15B). An interesting example is the BAR domain that consists of four long helices. While the constitutive junctions of all exons of *BIN1* that encode this domain are located within these helices, the position of the alternative junction is in the loop between two helices (Figure 2.15C). Furthermore, alternative junctions have a tendency to be located at the protein surface (± one residue context, average 2.96 vs. 2.36, higher values indicate exposed residues). This clearly shows that alternative exon junctions are non-randomly distributed within Pfam domain structures. The preferred position at the surface and between secondary structure elements argues against a destructive role of most of these splice events.

2.7 Discussion

We describe a novel approach that uses information about Pfam domains to predict exon skipping and intron retention events *ab initio*. Only the genomic sequence and gene structure annotation are required. Our approach is able to predict alternative exons regardless whether their size is divisible by three and is independent of the existence of orthologs in another species. We have shown that this approach can reliably identify exon skipping and intron retention events *ab initio* and that it complements existing comparative methods. Our approach has two limitations. Firstly, it is restricted to the regions of a gene that encode Pfam domains. However, Pfam is one of the most comprehensive descriptions of functional domains as Pfam domains match 75% of all proteins in Swiss-Prot/TrEMBL and cover 53% of all residues [109]. Apart from Pfam domains, the general approach can use other functional motif descriptions like those

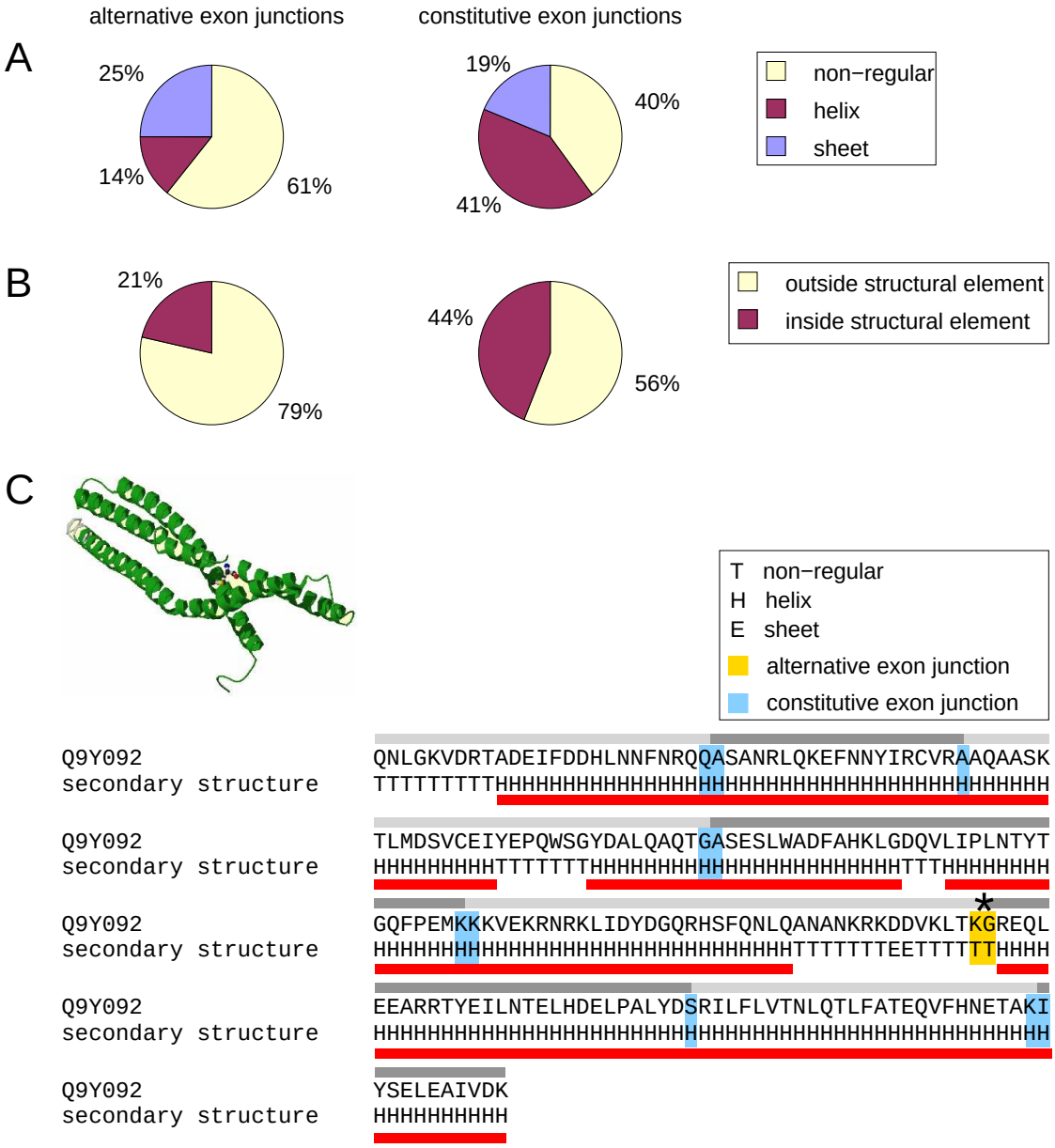

Figure 2.15: Location of alternative and constitutive exon junctions within Pfam domains. (A) Analysis of residues at the exon junction with respect to location in a helix, in a sheet, or in non-regular elements. (B) Analysis of the ± one amino acid context around exon junctions. (C) BAR domain PF03114 as an example: The exon junctions of *BIN1* (NM_19345) are mapped to the secondary structure of the BAR domain using the known structure (PDB 1uru) of CG8604-PA (Swiss-Prot Q9Y092) as a template. The three-dimensional structure is shown in the top-left corner of (C), the secondary structure at the bottom. While all constitutive exon junctions are located within helices, the alternative junction is in the loop. If the exon junction splits a codon, only this amino acid is highlighted and otherwise, if the junction is between two codons, both residues are highlighted. Red bars indicate the four helices. The positions of the exons are indicated by boxes in alternating grey. The star indicates the position where the alternative exon is inserted.

contained in the InterPro database. Additionally the constant growth of these databases will lead to a higher coverage and more predictions. Secondly, our approach is restricted to cases where the Pfam score is increased because it is unlikely that this occurs just by chance. Many splice events result in a deletion of functional domains, which decrease the overall Pfam score. Such events cannot be predicted by this approach since a strategy that arbitrarily predicts an exon to be alternative will also result in a lower Pfam score.

In this study, we considered a total of 18,572 human RefSeq transcripts and made a prediction for 307 (1.7%) of them. We only predicted exon skipping and intron retention events as no other putative alternative splice sites are given. However indirectly, for a number of predictions that result in a frameshift, we found an alternative donor/acceptor site or an exon that is skipped in the given transcript. These alternative splice events cause the same frameshift that is predicted by our algorithm. As written in section 2.5.2, our algorithm can also handle alternative donor/acceptor sites if they are given in $\mathcal{B}_l$ and $\mathcal{B}_r$. To test this, we evaluated the prediction of the algorithm in five cases where the positions of the additional splice sites were given. As expected, in each case the algorithm uses the additional splice site and produces a splice form that equals the known alternative splice form. Moreover, C-intron retentions and hidden exon predictions were only found for the last intron in the transcript, since most of them do not have a continuous reading frame and we excluded hypothetical splice forms that are NMD candidates. Numerous of these events in other introns can be found by relaxing the NMD criterion. Again, such events can be predicted if the corresponding alternative splice sites are included. Consequently, it is promising to include other splice sites, for example those derived from predicted suboptimal exons, which can be found by gene prediction programs such as Genscan [90]. This will increase the number of predictions with alternative donor/acceptor sites as well as exons that are hidden in introns and whose inclusion is not seen in available expressed sequences.

We have shown that sequence inserts inside Pfam domains prefer to be located at the protein surface and strongly avoid a position within secondary structure elements. This is in line with their negative impact on a Pfam domain (based on the score), their low inclusion level, and their low conservation in mouse. This suggests that most of these inserts alter the domain structure and function, which is in contrast to other alternative splice variants that delete an entire Pfam domain or an essential part of it [52]. A likely evolutionary scenario is the exonization of a part of an intron followed by a selection pressure assuring that the novel exon is rarely included to produce enough amount of the functional protein. If the inclusion of this exon has no drastic consequences for the domain structure, it might acquire a function. Indeed, we found examples in the literature where such splice events have important functional consequences. For example, a splice form of *TRAF2* with a seven amino acid insert into a Ring finger domain acts as a dominant negative inhibitor of *TNFR2*-dependent *NFκB* activation [110]. Alternatively spliced

inserts modulate the structure of loops at a protein interaction surface of neurexin Iβ, which influences the binding of protein ligands [49]. However, even small inserts may result in a change of the overall protein fold. For example, insertion of nine residues into the C_2 domain of Piccolo due to inclusion of exon 15 leads to a rearrangement of the β-sheets, which explains the drastic differences in Ca^{2+} affinity for both splice forms [48]. A 17 amino acid insert for *UAP1* modifies the architecture of the active site and alters substrate specificity [111]. We believe that many of the splice forms found in this study are biologically interesting as they affect a protein domain and presumably alter its structure and function.

Intron retention seems to be a rare splice event with an estimated frequency of 6% [112]. Furthermore, they are difficult to detect because of unspliced or partially spliced ESTs. Most of the intron retentions predicted here contribute to Pfam domains and the retention is confirmed by the existence of another RefSeq transcript. Therefore, they are likely to represent important alternative splice forms [113]. Since nearly all predicted intron retention and hidden exon events are EST confirmed, we conclude that intronic open reading frames encoding Pfam domains are very likely to become exonic in another transcript.

Due to a high number of human ESTs and intensive biomedical research, the human transcriptome presumably is the best characterized one. In contrast, the number of ESTs is much lower for other species, for example, chicken has less than 600,000 ESTs and Drosophila less than 500,000 (release July 2006). Even in the well-annotated genome of *C. elegans*, there are thousands of genes without EST/cDNA support [114]. As alternative splicing is assumed to be equally frequent in other species [11], *ab initio* prediction should be very useful for species with low EST numbers. Therefore, we believe that the application of our approach to other organisms will lead to the discovery of numerous novel alternative splice events.

Chapter 3

General influence of mRNA secondary structure on splicing

The second part of this thesis deals with the influence of RNA secondary structure on splicing. Although mRNA is often considered as a linear sequence of codons, its secondary structure features are important for a number of maturation processes including splicing. Furthermore, most proteins that affect splicing decisions are equipped with domains that bind single-stranded RNA and the sequestration of a binding site into a double strand was reported to prevent protein binding.

In this chapter, we analyze the secondary structure of an extensive set of experimentally determined enhancer and silencer motifs in their natural context. We found that the binding sites of splicing factors are significantly more single-stranded and tested this principle experimentally. Since splicing is regulated by many splicing factors binding to multiple sites, this finding argues for a general importance of mRNA secondary structures for splicing. Our results can have far reaching implications from the interpretation of mutagenesis experiments to the *in silico* prediction of splicing motifs.

Another important implication is that secondary structures can help to discriminate real from spurious protein binding sites in *de novo* motif finding. As knowledge about the binding motif is a crucial step to understand the function of an RNA-binding protein, secondary structures should not be neglected when searching for the binding motif of proteins that bind single-stranded RNA. To this end, we developed and implemented MEMERIS, a novel approach for searching sequence motifs in a set of RNA sequences and simultaneously integrating information about secondary structures. To abstract from specific structural elements, MEMERIS precomputes position-specific values measuring the single-strandedness of all substrings of an RNA sequence. These values are used as prior knowledge about the motif starts to guide the motif search. We performed extensive tests with artificial and biological data and demonstrated that MEMERIS is able to identify motifs in single-stranded regions even if a stronger motif located in double-stranded parts exists. The general principle to use prior knowledge about putative motif start positions can be extended to other applications.

Plan of the chapter

An introduction about the influence of mRNA secondary structures on splicing is given in section 3.1. In section 3.2, we investigate the single-strandedness of splicing factor binding sites. The MEMERIS algorithm and its application to artificial and real data are described in section 3.3. Finally, we discuss our results and potential implications in section 3.4.

3.1 Functions of mRNA secondary structures

For a long time it has been thought that mRNAs contain only the linear information of the codon sequence. In contrast, numerous other RNA species like transfer RNAs or ribosomal RNAs have conserved secondary structures that are essential for their function. However, several studies demonstrated that mRNA is more than simply a carrier for codons. Indeed, the secondary and tertiary structures of mRNAs play important roles in a number of processes including editing, splicing, localization, stability, and translation [115, 116, 117]. For example, the formation of double-stranded parts is essential for adenosine to inosine (A to I) editing by double-stranded RNA-specific editases [118]. Translation of ferritin mRNAs is controlled by a small hairpin structure (the iron responsive element) [119] and some viral as well as cellular transcripts contain IRES (internal ribosome entry site) elements in their 5' UTR that enable a cap-independent translation initiation [120]. The interpretation of a TGA codon as a codon for selenocysteine depends on a downstream hairpin structure (the selenocysteine insertion sequence) [121].

There is emerging evidence that mRNA secondary structure plays a role in alternative splicing as well [40]. For example, the skipping of exon 10 of the *MAPT* gene is correlated with the stability of a stem structure that sequesters the donor site [122]. The splicing of two mutually exclusive exons of *FGFR2* is regulated by use of a conserved secondary structure [123] and there is evidence that the tight regulation of a cluster of 48 mutually exclusive exons in the Drosophila *DSCAM* gene is achieved by formation of secondary structures [124]. Moreover, it has been proposed that sequences surrounding alternative exons might form structures that loop out the exon and prevent its recognition [45, 125]. Interestingly, such loops can also be formed by hnRNP A1 or PTB dimers that bind motifs up- and downstream of an exon [126, 127].

SR proteins and hnRNPs are equipped with single-stranded RNA binding domains [128]. Indeed, several of these proteins were shown to bind sequence motifs in a special structural context. For example, Nova-1 binds the sequence TCAT only when located in a hairpin loop [129]. Furthermore, SRp55 and hnRNP A1 proteins bind specific single-stranded sequences in hairpin loops [130, 131]. Thus, the sequestration of a binding site into a double-stranded part can prevent the binding of single-stranded RNA binding

proteins [132]. One such example is the mouse fibronectin EDA exon. The deletion of a silencer in this exon results in a shift of a critical ESE from a position in a hairpin loop to a position in a stem, which leads to complete exon skipping [133]. However, it is currently unclear to which extent mRNA secondary structures form in natural environments since studies have shown large differences between *in vitro* and *in vivo* experiments [134, 135]. These differences might be explained by co-transcriptional formation of secondary structures as well as a rapid binding of proteins to the nascent transcript. Moreover, it has been reported that RNA helicases influence splicing [136, 137, 138, 139]. It is conceivable that these enzymes rapidly resolve most of the existing secondary structures *in vivo* so that in a nearly single-stranded mRNA only the sequence of the splicing motifs and not their structural context is important. Thus, it is currently unknown if the structural context of binding sites for splicing proteins is important in general.

3.2 Higher single-strandedness for experimentally verified splicing motifs

3.2.1 Measurement of single-strandedness

The following analysis investigates the single-strandedness of substrings of an RNA sequence, therefore we first define how to measure the single-strandedness. In contrast to the three-dimensional structure of a protein, RNAs often have a more flexible and dynamic structure, which hampers the experimental structure determination. Therefore, the secondary structure of an RNA, which is the set of all base pairings, is often considered. RNA secondary structures can be efficiently predicted by energy minimization [140]. Since the natural occurring secondary structure is not always the structure with the predicted lowest free energy and a single RNA sequence can adopt more than one structure and perform more than one function [141], a set of suboptimal structures is usually computed. Structures with a similar free energy can differ greatly and it is often arbitrary to set an energy threshold value up to which suboptimal structures are considered. Therefore, we decided to use the equilibrium partition function and the base pair probabilities for our purpose [142]. Base pair probabilities represent the likelihood that two bases form hydrogen bonds in the ensemble of all possible secondary structures. Since there is no standard method to compute the single-strandedness of a substring of an RNA sequence, we introduce three different measurements:

- the Expected Fraction of bases in the substring that are unpaired (denoted EF),
- the Probability that all bases in the substring are Unpaired (denoted PU),
- and the Energy Difference between the ensemble of all structures and the ensemble of those structures that do not have base pairs in the substring (denoted ED).

Note that PU values have also been used in other studies [143, 144].

Let us consider the substring in a given RNA sequence between positions a and b. $EF_{a,b}$ is defined as

$$EF_{a,b} = 1 - \frac{\sum_{i=a}^{b} \sum_{j=1}^{L} p_{i,j}}{b - a + 1}$$

where L is the length of the RNA sequence and $p_{i,j}$ is the probability that bases i and j are paired. The base pair probabilities $p_{i,j}$ are computed with the RNAfold program [145]. $PU_{a,b}$ is defined as

$$PU_{a,b} = e^{\frac{E^{all} - E^{unpaired}_{a,b}}{RT}}$$

where E^{all} is the free energy of the ensemble of all structures, $E^{unpaired}_{a,b}$ is the free energy of the ensemble of all structures that have the complete substring unpaired, R is the universal gas constant, and T is the temperature. We compute E^{all} and $E^{unpaired}_{a,b}$ using the partition function version of RNAfold. For $E^{unpaired}_{a,b}$, we assure that the region a to b is unpaired by applying additional constraints (RNAfold parameter -C). Recently, a more efficient computation of PU values was implemented in the RNAup program [143]. $ED_{a,b}$ is defined as

$$ED_{a,b} = E^{all} - E^{unpaired}_{a,b}.$$

Higher values for PU and EF indicate a higher single-strandedness of the motif. In contrast, the higher the ED value, the more stable structures have at least one base pair in the substring, which contributes to a higher energy difference between the two ensembles. Thus, lower ED values indicate a higher single-strandedness. These measurements have the advantage that they account for all possible structures and that the values for two motifs of equal length can be directly compared. Furthermore, the measurements are based on the free energies, thus differentiating between stable (C-G) and less stable base pairings (A-U, G-U). This is reasonable because the break of base pairs by helicases or by protein binding should be more difficult for stable pairings. Finally, it is advantageous that these measurements abstract from specific structural elements.

3.2.2 Data set of experimentally verified splicing motifs

To investigate the structural context of splicing motifs, we extracted a set of 165 experimentally determined motifs with their natural mRNA sequence context from the AEDB database [146]. This set comprises exonic and intronic enhancers and silencers from human, mouse, rat, chicken, Drosophila, and several viruses. To get a high-quality data set, we checked the consistency of the listed genes, species, and motif sequences with the respective publications. Only motifs that were demonstrated to influence splicing in their natural context were considered. Three-dimensional structures of single-stranded RNA binding proteins indicate that they usually contact only a few residues. Therefore, we

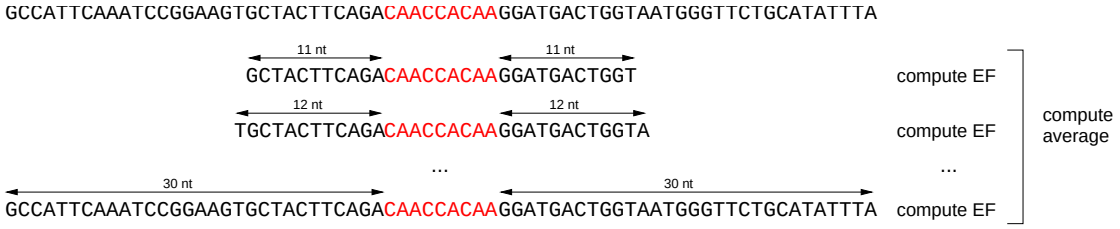

Figure 3.1: Scheme for computing the single-strandedness of a verified splicing motif. The ESE CAACCACAA (red) of exon 8 of the *CD44* gene is shown together with 30 nt of its mRNA sequence context. We compute EF values of the splicing motif for all context lengths from 11 nt up to 30 nt. To get a single value that measures the single-strandedness of this ESE, we compute the average of these 20 values.

discarded 88 motifs with a length of more than 9 nt. These motifs are likely to contain a core binding site at a location that was not experimentally determined. The remaining 77 motifs with a length equal or smaller than 9 nt were selected for further analysis. For convenience, we denote these motifs also as '*verified motifs*'.

3.2.3 Folding window

It is unclear to which extent mRNA is free to fold in natural environments, but several lines of evidence argue that *in vivo* secondary structures can only be formed in a rather narrow window. Firstly, mRNAs are bound by numerous proteins and protein binding should influence their ability to fold freely. Secondly, the formation of secondary structures occurs co-transcriptionally, therefore short range base pairs are favored compared to long-range base pairs, which is consistent with the results of kinetic folding algorithms [147]. Thirdly, the hybridization of the nascent mRNA with the DNA strand should elicit RNAse H degradation, thus such a hybridization must be prevented either by local mRNA secondary structures or protein binding. Fourthly, experiments suggested that mRNA folding is limited to a region of about 50 nt downstream of the transcribing polymerase [148]. Finally, for practical reasons, the accuracy of RNA folding programs drops for long sequences.

In light of these uncertainties, we decided to consider all symmetrical context lengths from 11 up to 30 nt up- and downstream of the motif. Thus, for a motif with length 6 nt, we considered sequences with a total length from 28 nt (for context length of 11 nt) to length 66 nt (for context length of 30 nt). A minimum context length of 11 nt was chosen because a certain sequence length is necessary to allow the formation of energetically stable structures (free energy less than 0 kcal/mol). We computed the EF value of the splicing motif for all 20 context lengths. Then, we averaged these 20 values to get a single EF value for each splicing motif (Figure 3.1). We assume that this procedure is more unbiased compared to the simpler way of computing the single-strandedness from just one

fixed context length. To get an overall measure for all splicing motifs, we computed the average EF value for the 77 verified splicing motifs. Likewise, this procedure is repeated with the PU and ED value. It should be mentioned that we use the same context lengths for all following tests to assure the comparability of average EF, PU, and ED values. An example of the base pair probability matrix and these three (EF, PU, ED) measurements for a sequence containing two PTB binding sites is given in Figure 3.2.

3.2.4 Experimentally determined splicing motifs are preferably located in single-strands

The average EF, PU, and ED values for the 77 verified splicing motifs are given in Table 3.1. To assess whether verified splicing motifs have a preference for single strands, we constructed several null models. As PU and to a lower extent EF and ED depend on the length of the motif, it is necessary to use the length distribution of the verified splicing motifs in all null models.

Firstly, we randomly chose a new motif of the same length in the 150 nt up- and downstream flanks of the natural context for all 77 motifs. We repeated this 100 times to obtain 100 sets, each with 77 randomly chosen motifs from the natural context (denoted null model 1). This null model has the advantage to account for possible biases in the selection of genes or exons since it uses the same sequences. The P-value was computed as the fraction of sets having a higher average single-strandedness compared to the set of verified motifs (for example, if all 100 random sets have a lower EF value than the verified motifs, the P-value is less than 0.01 (1 of 101 total sets)). We obtained significant P-values of 0.01 for EF and PU and a P-value of 0.04 for ED (Table 3.1).

Secondly, we repeated the procedure of null model 1 but replaced the randomly chosen motif by the verified one and got essentially the same results (null model 2, Table 3.1). This null model also accounts for the sequence bias of the motifs.

Thirdly, since the dinucleotide composition influences the stability of secondary structures, we used dinucleotide shuffling [150] to modify the up- and downstream flanks of verified motifs, while preserving the motif sequence. Again, 100 different sets were constructed and the P-value was computed as described above (null model 3). For all measurements, we got P-values of 0.01 (Table 3.1).

Fourthly, we randomly selected a new motif in a set of 10,000 randomly chosen internal exons (null model 4) and introns (null model 5) using the length distribution of the verified motifs. Since exons and introns have differences in their nucleotide composition [9], we split the 77 verified motifs into 50 exonic and 27 intronic ones according to their location in the exon-intron structure. Then, we compared the average EF, PU, and ED values for the 10,000 random exonic (intronic) motifs and the verified exonic (intronic) motifs. Using the t-test, we found significant differences between verified and random motifs except for exonic ED values (Table 3.1).

```
GGCCCUGUCUUCGCACCUCAGCCUCUCCUUCUCUCUGCUUCUCUCUCGCUGGCCCUUAGGAGGAAGGUGGAUGUCA
(((.(..(((((((.....)).)).(((((....(((.(((.........)).)).....)))))))))..)..))). -18.40
(((.(..(((((((.....)).)).(((((.....(((.(((.........)).)).....)))))))))..)..))). -18.16
(((.((..(((((((.....)).)).(((((...((.(((.........)).)).....)))))))))..))..))). -18.10
(((.(..(((((((.....)).)).(((((.....(((.(((.........)).)).....)))))))))..)..))). -18.10
(((.(..(((((((.....)).)).(((((.....(((((.........)))).....)))))))))..)..))). -18.10
(((.((..(((((((.....)).)).(((((.....(((.(((.........)).)).....)))))))))..))..))). -17.86
```

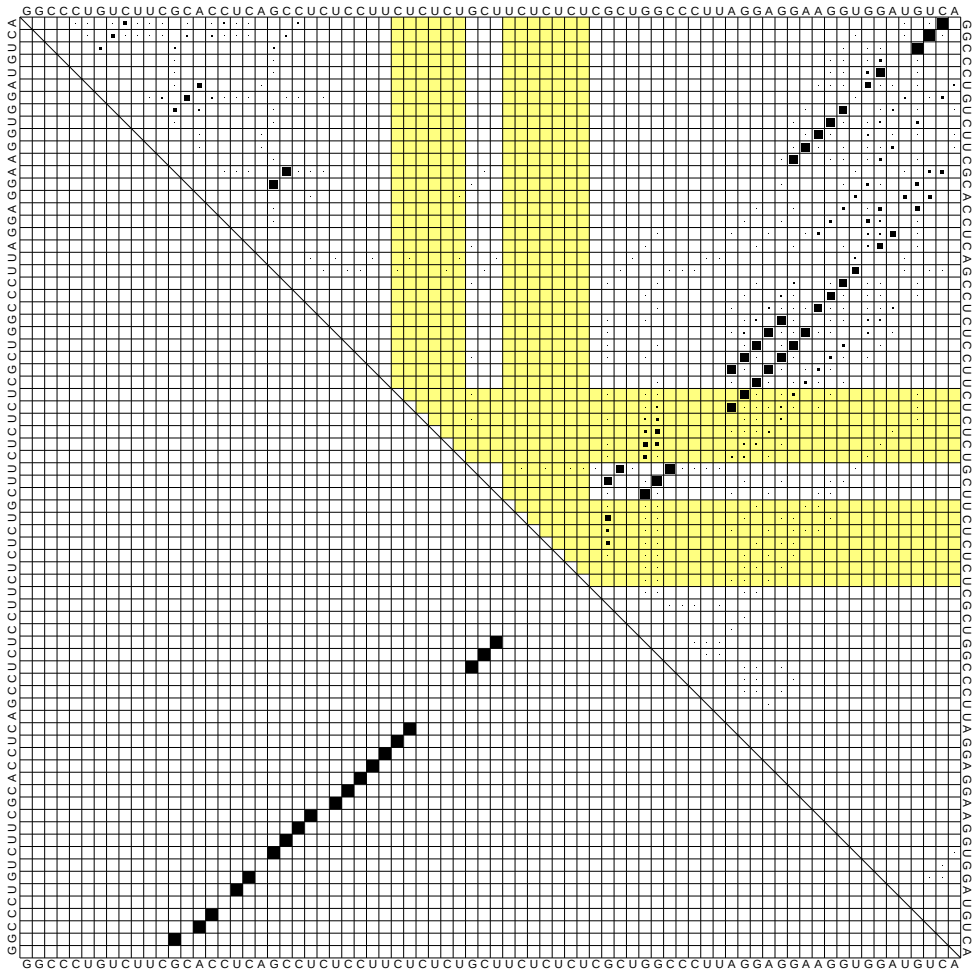

Figure 3.2: Example of the secondary structure context of two PTB binding sites in the intron upstream of the mouse c-src exon N1 [149]. The base pair probability matrix for the sequence comprising the two PTB motifs (CUCUCU and UCUCUCU) and the 30 nt context up- and downstream is shown. Both binding sites are highlighted yellow. The base pair probabilities for the complete structure ensemble are contained in the upper right part of the matrix (big dots represent high probabilities). The optimal and the five best suboptimal structures and their free energies in kcal/mol are shown above the matrix (computed with RNAsubopt [145], parentheses represent base pairs, dots unpaired bases). It can be seen that both binding sites are found in mostly single-stranded regions. Accordingly, the measurements of single-strandedness are far below the average for all null models (average for context lengths 11-30 nt: CUCUCU: EF=0.93, PU=0.71, ED=0.38; UCUCUCU: EF=0.95, PU=0.60, ED=0.32; compare with values in Table 3.1).

To exclude the possibility that the maximal context length of 30 nt is inappropriate, we repeated the entire analysis with average EF, PU, and ED values for context lengths 11 to 20 nt as well as 11 to 50 nt. Since we assume that the existence of long range but not short range base pairs in a cell is questionable, shorter context lengths were always included. Basically, we observed the same results as for context lengths 11 to 30 nt (Table 3.1). These findings indicate that verified splicing motifs have a significant trend to be located in single-stranded regions. Furthermore, the results from null models 2 and 3 that use the same motif sequences indicate that the higher single-strandedness is attributed to the flanks of those motifs rather than to the motifs themselves.

3.2.5 Higher single-strandedness for splicing motifs cannot be explained by lower GC content

The GC content of a sequence has an influence on the stability of secondary structures, since C-G base pairs allow more stable base pairings. Indeed, we found that a higher GC content leads to a lower single-strandedness in general. Therefore, we have to exclude that the observed higher single-strandedness for verified motifs is attributed to a lower GC content of the motifs and/or their flanks. The average GC content for the motif and the 30 nt up- and downstream flanks is virtually identical for the verified motifs and null models 1-3 (Table 3.1). Consistent with the observation that introns are less GC-rich than exons [9], our null model 5 is very conservative with an average GC content that is 7% below that for intronic verified motifs (Table 3.1).

However, exon sequences from null model 4 have an about 2% higher GC content compared to verified exonic motifs, which might contribute to the observed higher single-strandedness for the latter group. To exclude this possibility, we generated a new set of 10,000 exonic motifs with almost equal average GC contents and one with lower GC contents. Again, we observed more single-strandedness for the exonic splicing motifs (Table 3.1). Thus, even conservative controls with a lower GC content yield significant differences. For completeness, we generated two sets with 10,000 exonic and intronic motifs having a higher GC content (about 6% higher) and, as expected, the single-strandedness decreased. To summarize, we observed a higher single-strandedness for splicing motifs for EF, PU, and ED values in all null models. We conclude that the GC content influences our measurements, but does not explain the observation that verified splicing motifs are more single-stranded.

		EF						PU						ED						average GC content		
		11...20 nt		11...30 nt		11...50 nt		11...20 nt		11...30 nt		11...50 nt		11...20 nt		11...30 nt		11...50 nt				
	number of motifs	ave[a]	P<[b]	ave	P<	ave	P<	ave	P<	ave	P<	ave	P<	ave	P<	ave	P<	ave	P<	motif	30 nt up-stream	30 nt down-stream
splicing motifs																						
all	77	0.69		0.65		0.62		0.29		0.25		0.21		2.17		2.49		2.77		0.49	0.48	0.49
exonic	50	0.67		0.62		0.59		0.23		0.19		0.16		2.41		2.77		3.02		0.49	0.47	0.48
intronic	27	0.73		0.71		0.67		0.40		0.35		0.29		1.72		1.96		2.32		0.50	0.51	0.49
null model 1	100[c]	0.63	0.01	0.59	0.01	0.55	0.01	0.18	0.01	0.15	0.01	0.13	0.01	2.59	0.03	2.86	0.04	3.16	0.02	0.48	0.48	0.49
null model 2	100[c]	0.64	0.02	0.60	0.01	0.57	0.01	0.21	0.01	0.18	0.01	0.15	0.01	2.60	0.04	2.91	0.05	3.19	0.04	0.49	*0.49*	0.49
null model 3	100[c]	0.63	0.01	0.59	0.01	0.55	0.01	0.18	0.01	0.15	0.01	0.12	0.01	2.61	0.01	2.91	0.01	3.22	0.02	0.49	0.48	0.48[d]
null model 4	10,000	0.60	0.008	0.57	0.029	0.53	0.013	0.14	0.008	0.12	0.007	0.09	0.004	2.83	*0.131*	3.12	*0.228*	3.40	*0.183*	*0.50*	*0.50*	*0.50*
lower GC content	10,000	0.61	0.023	0.58	*0.075*	0.54	0.039	0.15	0.019	0.12	0.020	0.10	0.014	2.67	*0.327*	2.95	*0.520*	3.22	*0.450*	0.47	0.46	0.47
equal GC content	10,000	0.61	0.017	0.57	*0.057*	0.53	0.025	0.15	0.019	0.12	0.018	0.10	0.012	2.73	*0.241*	3.02	*0.381*	3.31	*0.304*	0.49	0.47	0.48
higher GC content	10,000	0.59	0.001	0.55	0.005	0.51	0.002	0.12	0.001	0.10	0.001	0.08	0.001	3.07	0.020	3.37	0.036	3.68	0.021	*0.57*	*0.53*	*0.53*
null model 5	10,000	0.62	0.004	0.58	0.001	0.54	0.001	0.18	0.001	0.15	0.001	0.12	0.001	2.52	0.032	2.83	0.025	3.14	0.042	0.43	0.43	0.43
higher GC content	10,000	0.58	0.001	0.54	0.001	0.50	0.001	0.13	0.001	0.11	0.001	0.08	0.001	3.26	0.001	3.60	0.001	3.97	0.001	*0.60*	*0.55*	*0.55*

Table 3.1: EF, PU, ED values, P-values and GC content for verified splicing motifs and all null models.
The values for the different context lengths are indicated in the second line as '11..20 nt', '11..30 nt', and '11..50 nt'. The P-values for null models 1-3 were computed as the fraction of test sets with a higher average single-strandedness compared to the all 77 splicing motifs. The P-values for null models 4 and 5 were computed by the t-test (null model 4 and 5 was compared to exonic and intronic splicing motifs, resp.). The GC content for null models 1-3 was computed by averaging over the 100 test sets. P-values in italics are not significant at the 0.05 level. GC content values in italics are higher for the null model than for the verified splicing motifs.
[a] average values for the given context lengths
[b] means 'P value <'
[c] means 100 test sets with each 77 motifs
[d] The slight change in the GC content is due to shuffling of dinucleotides that are located at the ends of the 30 nt contexts.

3.2.6 Testing the effect of single- and double-stranded splicing motifs experimentally

In collaboration with the group of Stefan Stamm, we performed *in vivo* minigene experiments to verify these computational results. We selected the SXN minigene construct, which contains an artificial alternative exon between two constitutively spliced exons. Then, we designed oligonucleotide sequences that contain an ESE or an ESS sequence in single- or double-stranded conformation. We used the ESS TAGGGT, the ESE CAACCACAA, and the ESS CAAGG (Figure 3.3A). TAGGGT is a strong binding site for hnRNP A1, CAACCACAA was demonstrated to enhance splicing in the *CD44* gene, and CAAGG acts as an exonic silencer in the fibronectin gene. Except for a shift of 1 nt in the CAACCACAA constructs, the single- and double-stranded motifs are located exactly at the same position. Therefore, any effect due to a variation in the position of the splicing motif can be excluded [30].

The oligonucleotides were inserted into the alternative exon of the minigene, the minigene was transfected into cells, and the mRNA was amplified by RT-PCR. As shown in Figure 3.3B, the inclusion level differs between the constructs having the motif in the loop or in the stem of the designed secondary structure. According to our hypothesis, the ESS TAGGGT should allow more efficient binding in single-stranded conformation, which should result in a higher level of exon skipping. Indeed, the loop construct results in only 15% exon inclusion, which is lower than 37% for the stem construct. Likewise, we expect that the ESE CAACCACAA leads to a higher inclusion level when located in a single-strand. Consistently, the loop construct results in 70% inclusion compared to only 7% for the stem construct. Unexpected, the sequence CAAGG leads to 75% inclusion for the loop construct and to 34% inclusion for the stem construct. Thus, CAAGG seems to act as an enhancer in our test system, although it was described as a silencer in the fibronectin gene [133]. Interestingly, this pentamer occurs in 18 of the 2,042 ESE octamers predicted in [42], but it occurs in none of the 1,019 predicted ESS octamers (Fisher´s exact test: P=0.0015). Thus, it is conceivable that CAAGG can also act as an enhancer.

It should be mentioned that these experiments cannot exclude that the observed effects are influenced by other splicing motifs that are contained in the designed sequences, since it was nearly impossible to avoid the occurrence of any other known splicing motif. Furthermore, during the experiments with these constructs, another study described numerous new splicing motifs [30] and it is likely that additional motifs remain to be discovered. However, although the ESS TAGGGT and the ESE CAACCACAA loop constructs have 98% identity, we observed a great difference in the exon inclusion level. Therefore, we assume that the motif in the loop exerts a dominant role. Although further studies and experiments are needed, we conclude that the secondary structure of a splicing motif influences alternative splicing.

A

TAGGGT

ATCCATGGGGCTGGATGTGACGTAGTAGGGTATACGTCACATAGCTTCCTCTCATGA
...(((((((..((((((((((((.......))))))))))).....)))))))))).
CTACCCTACGCATGATACGCATGCGTAGGGTAGCACTGCATGAGCTTCCTCACGTTT
((((((((((((((.....))))))))))))))....((.((((....)))).))..

CAACCACAA

ATCCATGGGGCTGGATGTGACGTACAACCACAATACGTCACATACTTCCTCTCATGA
...(((((((..((((((((((((.........)))))))))).))...)))))))).
ATGATGGGTATGTGCGTTGCTTCGGCAACCACAACTCATCGCATACTTCCTCTCATGA
..(((((((.((((.(((((....)))))))))))))).................

CAAGG

ATCCATGGGGCTGGATGTGACGTAACAAGGCATACGTCACATAGCTTCCTCTCATGA
...(((((((..((((((((((((.......))))))))))).....)))))))))).
CTACCTTGCGCATGATACGCATGCGCAAGGTAGCACTGCATGAGCTTCCTCACGTTT
((((((((((((((.....))))))))))))))....((.((((....)))).))..

B

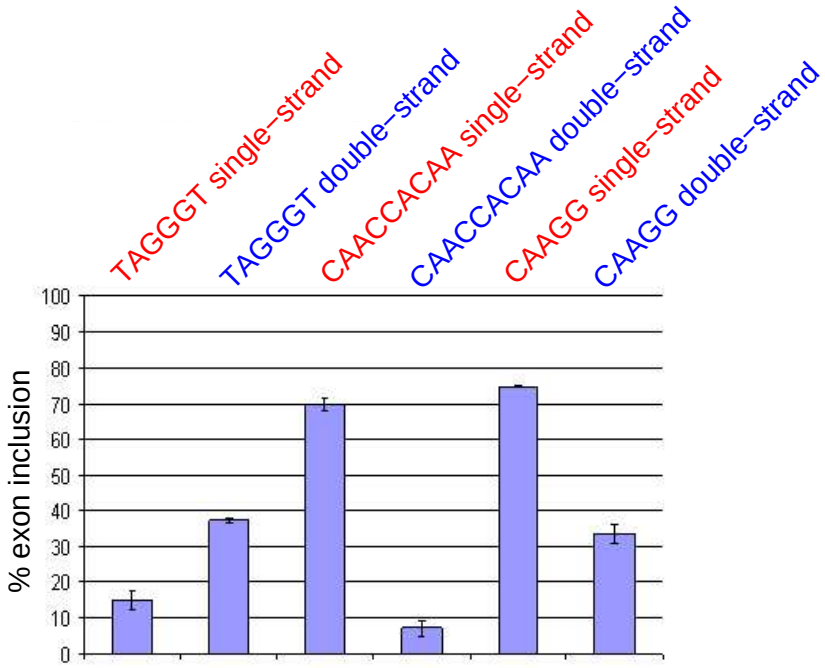

Figure 3.3: Results of the minigene experiments.
(A) We designed three pairs of oligonucleotides containing the splicing motif in single-stranded or double-stranded conformation. The splicing motif is highlighted yellow. The predicted optimal secondary structure is shown below the sequences. (B) The experimentally determined exon inclusion level for the six constructs is shown in the bar chart.

3.3 RNA sequence motif finding in single-stranded regions

3.3.1 Motivation

In the previous section, we found that a location in single-stranded conformation is a characteristic feature of known splicing factor binding sites. This indicates that information about secondary structures might be utilized in an important computational problem: the detection of unknown motifs in a given set of sequences. In the following, we assume that the given sequences are bound by a protein and that the binding sites have a certain sequence similarity. If the binding sites are known, one can build a model of the binding motif.

A crucial step to understand the function of a splicing factor is to elucidate its binding motif and to identify target mRNAs. One common experimental approach is the *selection of ligands by exponential enrichment* (SELEX), which is routinely used to identify the binding motif of splicing factors [151, 152, 153]. The result of a SELEX experiment is a set of sequences that are bound by the specific protein at one (or more) yet unknown binding sites. To identify these binding sites and the binding motif, motif finding programs are usually applied to this set of sequences. These motif finders expect that these sequences are enriched in a similar motif since all are bound by a specific protein. In light of our findings, we suggest that including additional information about secondary structures can be beneficial when searching for these binding sites.

However, currently existing motif finding programs like MEME [154, 155, 156] or Gibbs sampler [157] only work at the sequence and not at the structure level. On the other hand, several programs exist that search for sequence-structure motifs in RNA sequences [158, 159, 160, 161] or that perform RNA sequence-structure alignments [162, 163, 164]. However, these methods expect that the motif consists of specific sequence-structure elements such as a stem-loop structure possibly with additional sequence constraints. Hence, they would not be able to find a sequence motif with a general structural property such as being located in single-stranded parts of arbitrary structure elements. Such an example has been described for the hnRNP K protein, where the sequence motif is found in the loop of a hairpin or in the single-stranded part between two stems [165]. Consequently, existing methods cannot be used to find sequence motifs with the general characteristic property of being single-stranded.

Apart from splicing factors, the binding affinity of other proteins such as the mouse Prrp [166] or the αCP-2KL and hnRNP K proteins [165] is affected by the secondary structure context of the binding sites. Another example is the HuR protein that influences mRNA stability by binding to the motif NNTTNNTTT [132]. It has been demonstrated that the HuR binding affinity correlates with the single-strandedness of its binding motif and that the sequestration of its binding site in a double-strand abolishes protein binding. Interestingly, small antisense oligonucleotides that are designed to bind outside the HuR

motif can influence mRNA stability by modulating the secondary structure of the binding site [132, 144]. These experimental findings agree with our results from the previous section and further support the need for a motif finding method that includes knowledge about RNA secondary structures.

3.3.2 Overview of MEMERIS

We introduce an approach for searching sequence motifs that are preferably located in any single-stranded conformation. This approach is implemented as an extension of the widely used MEME motif finder and is called MEMERIS - MEME in RNAs Including secondary Structures. MEMERIS precomputes EF or PU values (defined in section 3.2.1) to characterize the single-strandedness of all putative motif occurrences in all given input sequences. Then, these values are used to guide the motif search towards single-stranded regions. The fundamental idea behind MEMERIS is to replace the uniformly distributed prior distribution for the motif starts used in MEME by a distribution that depends on the EF or PU values. We proceed to introduce the MEME algorithm. Then, we describe the MEMERIS algorithm in detail and emphasize the differences to MEME.

3.3.3 MEME

MEME is a program for finding motifs in a set of n unaligned nucleotide or protein sequences (denoted $X = X_1, X_2, \ldots, X_n$). A motif is described as a position-specific probability matrix $\Theta_1 = (P_1, P_2, \ldots, P_W)$, where W is the length of the motif and the vector P_j is the probability distribution of the letters at position j. A given input sequence X_i is modeled as consisting of different parts:

- zero, one, or more non-overlapping motif occurrences sampled from Θ_1 and
- random samples from a background probability distribution $\Theta_0 = P_0$ for the remaining sequence positions.

We denote $\Theta = (\Theta_0, \Theta_1)$. The number of motif occurrences depends on a user-specified model. MEME considers three different models:

- exactly one motif occurrence per sequence (OOPS model),
- zero or one motif occurrence per sequence (ZOOPS model),
- zero or more motif occurrences per sequence (two-component mixture (TCM) model).

To find a motif, MEME uses an expectation maximization (EM) algorithm to perform a maximum likelihood (ML) estimation of the model given the data [167]. EM algorithms have many applications in bioinformatics (for example, reconstructing full-length transcripts from EST fragments [168]) and are commonly used for ML estimations where a part of the complete data is not given or '*hidden*'. The EM algorithm iteratively

- computes the expectation of the hidden variables using the current model (E-step) and
- performs a ML estimation of the model parameters on the joint probability of the complete data (M-step).

In MEME, the complete data is the set of sequences (given data) and the start positions of the motif occurrences (hidden data). The hidden data is described by indicator variables $Z_{i,j}$ with $Z_{i,j} = 1$ if a motif occurrence starts at position j in sequence X_i, and $Z_{i,j} = 0$ otherwise.

3.3.4 The OOPS model in MEMERIS

MEME makes no assumption about the start position of a motif occurrence in a sequence. Thus, MEME uses a uniform probability distribution

$$P(Z_{i,j} = 1) = \frac{1}{m} \quad \forall j$$

where $m = L - W + 1$ is the number of possible start positions for a given motif length W in a sequence of length L (just for convenience, we assume that all sequences have the same length). Per definition

$$\sum_{j=1}^{m} P(Z_{i,j} = 1) = 1$$

which is the assumption of the OOPS model that there is exactly one motif occurrence per sequence.

The additional information about the single-strandedness of each substring of length W can be considered as an informative prior about putative motif starts, since single-stranded sequence parts are more likely to be real motif occurrences than parts that are sequestered in a double-stranded region. We integrate the single-strandedness by replacing the uniform probability distribution by a distribution that depends on the EF or PU values. For convenience, we focus on PU values in the following, although everything below holds for EF values too. Instead of $\frac{1}{m}$ as in MEME, the prior probabilities for the OOPS model in MEMERIS are

$$P(Z_{i,j} = 1|PU_i) = \frac{PU_{i,j} + \pi}{\sum_{k=1}^{m}(PU_{i,k} + \pi)}$$

where PU_i is the vector of PU values for sequence X_i and π is a user-given pseudocount that is used to smooth the distribution (Figure 3.4). The PU values are precomputed for all substrings (i.e. all putative motif occurrences) of a fixed length W of the input sequences. The higher the PU value for position j, the higher is the prior probability of being a motif start position $P(Z_{i,j} = 1|PU_i)$. Despite X_i is used to compute the

PU values, we assume that these values are given as prior knowledge. By definition $\sum_{j=1}^{m} P(Z_{i,j} = 1|PU_i) = 1$, thus the underlying model assumption (one motif occurrence per sequence) remains unchanged.

E-step In iteration t of the EM algorithm, the expected values $Z_{i,j}^t$ of the hidden variables $Z_{i,j}$ are computed given the parameters Θ^t and W. Here, Θ^t are the parameters in iteration t. MEMERIS requires a fixed motif length W during one application of the algorithm, thus the model parameters reduce to Θ^t (note that W is also fixed for one run in MEME [156]). Using the definition of the expectation and Bayes law, the E-step in MEMERIS is

$$\begin{aligned}
Z_{i,j}^t &= E_{Z|X,\Theta^t,PU}[Z_{i,j}] \\
&= 0 \cdot P(Z_{i,j} = 0|X_i, \Theta^t, PU_i) + 1 \cdot P(Z_{i,j} = 1|X_i, \Theta^t, PU_i) \\
&= P(Z_{i,j} = 1|X_i, \Theta^t, PU_i) \\
&= \frac{P(X_i|Z_{i,j} = 1, \Theta^t, PU_i)P(Z_{i,j} = 1|\Theta^t, PU_i)}{P(X_i|\Theta^t, PU_i)} \\
&= \frac{P(X_i|Z_{i,j} = 1, \Theta^t, PU_i)P(Z_{i,j} = 1|\Theta^t, PU_i)}{\sum_{k=1}^{m} P(X_i|Z_{i,k} = 1, \Theta^t, PU_i)P(Z_{i,k} = 1|\Theta^t, PU_i)}.
\end{aligned}$$

Since $P(Z_{i,j} = 1|\Theta, PU_i)$ does not depend on Θ, we can write $P(Z_{i,j} = 1|PU_i)$. Furthermore, we write $P(X_i|Z_{i,j} = 1, \Theta)$ instead of $P(X_i|Z_{i,j} = 1, \Theta, PU_i)$ since the probability of a sequence does not depend on the PU values. Thus, the final E-step equation is

$$Z_{i,j}^t = \frac{P(X_i|Z_{i,j} = 1, \Theta^t)P(Z_{i,j} = 1|PU_i)}{\sum_{k=1}^{m} P(X_i|Z_{i,k} = 1, \Theta^t)P(Z_{i,k} = 1|PU_i)}.$$

Both simplifications are also used in the equations given below.

Assuming that a sequence X_i contains at positions j_1 and j_2 the same substring, only the prior probabilities constitute the difference for the expected values Z_{i,j_1}^t and Z_{i,j_2}^t in the E-step equation. Hence, it is an advantageous property of the prior probabilities that the ratio of the PU values for j_1 and j_2 is preserved in $P(Z_{i,j_1} = 1|PU_i)$ and $P(Z_{i,j_2} = 1|PU_i)$ if $\pi = 0$, since

$$\frac{P(Z_{i,j_1} = 1|PU_i)}{P(Z_{i,j_2} = 1|PU_i)} = \frac{\frac{PU_{i,j_1}}{\sum_{k=1}^{m} PU_{i,k}}}{\frac{PU_{i,j_2}}{\sum_{k=1}^{m} PU_{i,k}}} = \frac{PU_{i,j_1}}{PU_{i,j_2}}.$$

The pseudocount π is used to reduce this ratio. The higher π, the more this distribution equals the uniform distribution of MEME (Figure 3.4).

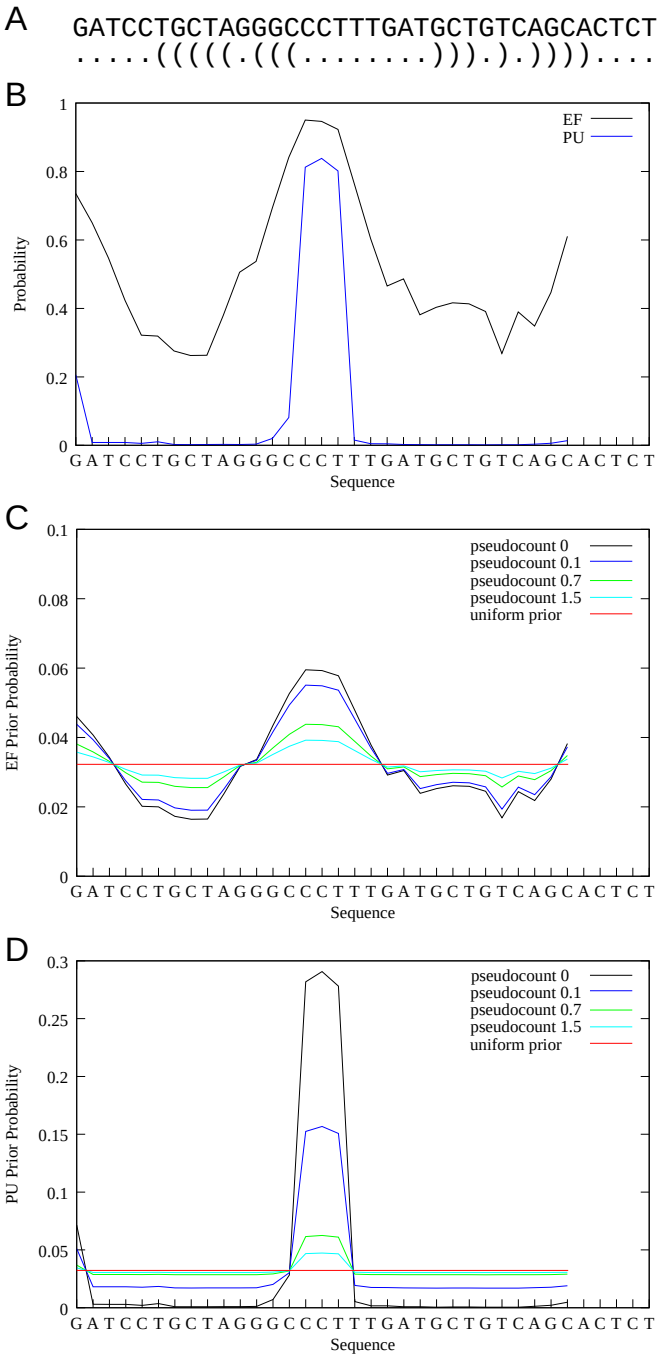

Figure 3.4: Effect of the pseudocount π on the prior probability distribution.
The figure shows a randomly chosen sequence and its optimal secondary structure (A), the EF and PU values for a motif length of 6 nt (B), and the prior probability distribution for the OOPS and ZOOPS model using EF (C) and PU values (D) with different pseudocounts. Each data point represents the value for the motif starting at the respective position. The uniform prior is $P(Z_{i,j} = 1) = \frac{1}{31}$ (sequence length is 36 nt).

M-step The M-step is not affected by the modified prior distribution, which is shown in the following. In the M-step, the model parameters are determined that maximize the joint log likelihood of the complete (given and hidden) data. Again, the motif length W is fixed and therefore not considered in the maximization. For the OOPS model, the joint log likelihood is defined as

$$
\begin{aligned}
&\log P(X, Z|\Theta, PU) \\
&\quad = \log \prod_{i=1}^{n} P(X_i, Z_i|\Theta, PU_i) \\
&\quad = \sum_{i=1}^{n} \log P(X_i, Z_i|\Theta, PU_i) \\
&\quad = \sum_{i=1}^{n} \log \Big(P(X_i|Z_i, \Theta, PU_i) P(Z_i|\Theta, PU_i) \Big) \\
&\quad = \sum_{i=1}^{n} \log \prod_{j=1}^{m} \Big(P(X_i|Z_{i,j} = 1, \Theta, PU_i) P(Z_{i,j} = 1|\Theta, PU_{i,j}) \Big)^{Z_{i,j}} \\
&\quad = \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{i,j} \Big(\log P(X_i|Z_{i,j} = 1, \Theta, PU_i) + \log P(Z_{i,j} = 1|\Theta, PU_{i,j}) \Big) \\
&\quad = \sum_{i=1}^{n} \sum_{j=1}^{m} \Big(Z_{i,j} \log P(X_i|Z_{i,j} = 1, \Theta, PU_i) + Z_{i,j} \log P(Z_{i,j} = 1|\Theta, PU_i) \Big) \\
&\quad = \sum_{i=1}^{n} \sum_{j=1}^{m} \Big(Z_{i,j} \log P(X_i|Z_{i,j} = 1, \Theta) + Z_{i,j} \log P(Z_{i,j} = 1|PU_i) \Big).
\end{aligned}
$$

Here, we used the basic assumption of the OOPS model that $Z_{i,j} = 1$ for exactly one position j in X_i and $Z_{i,j} = 0$ for the remaining positions.

Since $Z_{i,j}$ are hidden, the expected values computed in the E-step are used instead

$$
\begin{aligned}
&E_{Z|X,\Theta^t,PU}[\log P(X, Z|\Theta, PU)] \\
&= E_{Z|X,\Theta^t,PU} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \Big(Z_{i,j} \log P(X_i|Z_{i,j} = 1, \Theta) + Z_{i,j} \log P(Z_{i,j} = 1|PU_i) \Big) \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \Big(E_{Z|X,\Theta^t,PU}[Z_{i,j}] \log P(X_i|Z_{i,j} = 1, \Theta) + E_{Z|X,\Theta^t,PU}[Z_{i,j}] \log P(Z_{i,j} = 1|PU_i) \Big) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \Big(Z_{i,j}^t \log P(X_i|Z_{i,j} = 1, \Theta) + Z_{i,j}^t \log P(Z_{i,j} = 1|PU_i) \Big).
\end{aligned}
$$

Since the prior probabilities for the motif start positions do not depend on Θ, we can skip the second term $(Z_{i,j}^t \log P(Z_{i,j} = 1|PU_i))$. Thus, one has to solve

$$
\Theta^{t+1} = \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{i,j}^t \log P(X_i|Z_{i,j} = 1, \Theta)
$$

which is the same as for MEME [156].

3.3.5 The ZOOPS model in MEMERIS

The ZOOPS model takes into account that a sequence might have no motif occurrence. To this end, a new hidden variable Q_i is introduced for each sequence with

$$Q_i = \sum_{j=1}^{m} Z_{i,j}.$$

Thus, $Q_i = 1$ if sequence X_i contains one motif occurrence and $Q_i = 0$ otherwise. The probability $P(Q_i = 1)$ is denoted as γ, which is equal for all sequences.

E-step We use the same prior probabilities as for the OOPS model, thus the E-step for the ZOOPS model in MEMERIS becomes to

$$\begin{aligned}
&Z_{i,j}^t \\
&= E_{Z|X,\Theta^t,\gamma^t,PU}[Z_{i,j}] \\
&= P(Z_{i,j} = 1|X_i, \Theta^t, \gamma^t, PU_i) \\
&= \frac{P(X_i|Z_{i,j} = 1, \Theta^t, \gamma^t, PU_i) P(Z_{i,j} = 1|\Theta^t, \gamma^t, PU_i)}{P(X_i|\Theta^t, \gamma^t, PU_i)} \\
&= \frac{P(X_i|Z_{i,j} = 1, \Theta^t, \gamma^t, PU_i) P(Z_{i,j} = 1|\Theta^t, \gamma^t, PU_i)}{P(X_i|Q_i = 0, \Theta^t, \gamma^t, PU_i)(1-\gamma^t) + \sum_{k=1}^{m} P(X_i|Z_{i,k} = 1, \Theta^t, \gamma^t, PU_i) P(Z_{i,k} = 1|\Theta^t, \gamma^t, PU_i)} \\
&= \frac{P(X_i|Z_{i,j} = 1, \Theta^t) P(Z_{i,j} = 1|Q_i = 1, PU_i)\, \gamma^t}{P(X_i|Q_i = 0, \Theta^t)(1-\gamma^t) + \sum_{k=1}^{m} P(X_i|Z_{i,k} = 1, \Theta^t) P(Z_{i,k} = 1|Q_i = 1, PU_i)\, \gamma^t}
\end{aligned}$$

where γ^t denotes the respective parameter in iteration t. In the last step, we used that the probability of a sequence with and without a motif does not depend on γ. Furthermore, we used in the last step that

$$\begin{aligned}
P(Z_{i,j} = 1|\gamma, PU_i) &= P(Z_{i,j} = 1|Q_i = 0, \gamma, PU_i)(1-\gamma) + P(Z_{i,j} = 1|Q_i = 1, \gamma, PU_i)\, \gamma \\
&= 0(1-\gamma) + P(Z_{i,j} = 1|Q_i = 1, \gamma, PU_i)\, \gamma \\
&= P(Z_{i,j} = 1|Q_i = 1, \gamma, PU_i)\, \gamma \\
&= P(Z_{i,j} = 1|Q_i = 1, PU_i)\, \gamma.
\end{aligned}$$

M-step The M-step for the ZOOPS model in MEMERIS is the same as in MEME, which is shown in the following. $\phi = (\Theta, \gamma)$ is the complete parameter set for the ZOOPS

model. The joint log likelihood is

$$
\begin{aligned}
&\log P(X, Z|\phi, PU) \\
&\quad= \sum_{i=1}^{n} \log P(X_i, Z_i|\phi, PU_i) \\
&\quad= \sum_{i=1}^{n} \log \Big(P(X_i|Z_i, \phi, PU_i) P(Z_i|\phi, PU_i) \Big) \\
&\quad= \sum_{i=1}^{n} \log \Bigg[\Big[P(Q_i = 1) \prod_{j=1}^{m} \Big(P(X_i|Z_{i,j} = 1, \phi, PU_i) P(Z_{i,j} = 1|Q_i = 1, PU_i) \Big)^{Z_{i,j}} \Big]^{Q_i} \\
&\qquad\qquad \cdot \Big[P(Q_i = 0) P(X_i|Q_i = 0, \phi, PU_i) \Big]^{1-Q_i} \Bigg] \\
&\quad= \sum_{i=1}^{n} \Bigg[Q_i \Big[\log \gamma + \sum_{j=1}^{m} Z_{i,j} \Big(\log P(X_i|Z_{i,j} = 1, \phi, PU_i) + \log P(Z_{i,j} = 1|Q_i = 1, PU_i) \Big) \Big] \\
&\qquad\qquad + (1 - Q_i) \Big[\log(1 - \gamma) + \log P(X_i|Q_i = 0, \phi, PU_i) \Big] \Bigg] \\
&\quad= \sum_{i=1}^{n} \Bigg[Q_i \log \gamma + \sum_{j=1}^{m} \Big(Q_i Z_{i,j} \log P(X_i|Z_{i,j} = 1, \phi, PU_i) \Big) \\
&\qquad\qquad + \sum_{j=1}^{m} \Big(Q_i Z_{i,j} \log P(Z_{i,j} = 1|Q_i = 1, PU_i) \Big) \\
&\qquad\qquad + (1 - Q_i) \log(1 - \gamma) + (1 - Q_i) \log P(X_i|Q_i = 0, \phi, PU_i) \Bigg] \\
&\quad= \sum_{i=1}^{n} \Bigg[Q_i \log \gamma + \sum_{j=1}^{m} \Big(Z_{i,j} \log P(X_i|Z_{i,j} = 1, \Theta) \Big) \\
&\qquad\qquad + \sum_{j=1}^{m} \Big(Z_{i,j} \log P(Z_{i,j} = 1|Q_i = 1, PU_i) \Big) \\
&\qquad\qquad + (1 - Q_i) \log(1 - \gamma) + (1 - Q_i) \log P(X_i|Q_i = 0, \Theta) \Bigg]
\end{aligned}
$$

In the third step, we used the basic assumption of the ZOOPS model that a sequence contains either zero or one motif occurrence. In the last step, we used $Q_i\, Z_{i,j} = Z_{i,j}$, which is easy to show

$Z_{i,j}$	Q_i	$Z_{i,j}\, Q_i$
0	0	0
0	1	0
1	1	1

and because $Z_{i,j} = 1$ and $Q_i = 0$ is not possible per definition.

From the definition of Q_i, it follows that

$$E_{Z|X,\phi^t,PU}[Q_i] = E_{Z|X,\phi^t,PU}[\sum_{j=1}^{m} Z_{i,j}] = \sum_{j=1}^{m} E_{Z|X,\phi^t,PU}[Z_{i,j}] = \sum_{j=1}^{m} Z_{i,j}^t = Q_i^t$$

which is used in the following equation. Since $Z_{i,j}$ are hidden, the expected values of the hidden variables are used to compute the joint log likelihood

$$\begin{aligned}
&E_{Z|X,\phi^t,PU}[\log P(X, Z|\phi, PU)] \\
&= \sum_{i=1}^{n} \Big[E_{Z|X,\phi^t,PU}[Q_i] \log\gamma + \sum_{j=1}^{m} E_{Z|X,\phi^t,PU}[Z_{i,j}] \log P(X_i|Z_{i,j}=1,\Theta) \\
&\quad + \sum_{j=1}^{m} E_{Z|X,\phi^t,PU}[Z_{i,j}] \log P(Z_{i,j}=1|Q_i=1,PU_i) \\
&\quad + (1 - E_{Z|X,\phi^t,PU}[Q_i]) \log(1-\gamma) + (1 - E_{Z|X,\phi^t,PU}[Q_i]) \log P(X_i|Q_i=0,\Theta) \Big] \\
&= \sum_{i=1}^{n} \Big[Q_i^t \log\gamma + \sum_{j=1}^{m} \Big(Z_{i,j}^t \log P(X_i|Z_{i,j}=1,\Theta) \Big) \\
&\quad + \sum_{j=1}^{m} \Big(Z_{i,j}^t \log P(Z_{i,j}=1|Q_i=1,PU_i) \Big) \\
&\quad + (1-Q_i^t) \log(1-\gamma) + (1-Q_i^t) \log P(X_i|Q_i=0,\Theta) \Big].
\end{aligned}$$

This equation is reordered into two terms, where *term1* is

$$\sum_{i=1}^{n} \Big[\sum_{j=1}^{m} \Big(Z_{i,j}^t \log P(X_i|Z_{i,j}=1,\Theta) \Big) + (1-Q_i^t) \log P(X_i|Q_i=0,\Theta) \Big]$$

and *term2* is

$$\sum_{i=1}^{n} \Big[Q_i^t \log\gamma + \sum_{j=1}^{m} \Big(Z_{i,j}^t \log P(Z_{i,j}=1|Q_i=1,PU_i) \Big) + (1-Q_i^t) \log(1-\gamma) \Big].$$

Now, *term1* only depends on Θ, while *term2* only depends on γ. Thus, the maximization of the log likelihood can be done separately by maximizing *term1* with respect to Θ and *term2* with respect to γ. Finding the Θ that maximizes *term1* is the same as for MEME [156]. The maximization of *term2* with respect to γ is as follows. The first derivative of *term2* with respect to γ is

$$\frac{1}{\gamma} \sum_{i=1}^{n} Q_i^t - \frac{1}{1-\gamma} \sum_{i=1}^{n} (1 - Q_i^t).$$

Setting the first derivative to 0 gives

$$\begin{aligned}
\frac{1}{\gamma}\sum_{i=1}^{n} Q_i^t - \frac{1}{1-\gamma}\sum_{i=1}^{n}(1-Q_i^t) &= 0 \\
(1-\gamma)\sum_{i=1}^{n} Q_i^t &= \gamma\sum_{i=1}^{n}(1-Q_i^t) \\
\sum_{i=1}^{n} Q_i^t - \gamma\sum_{i=1}^{n} Q_i^t &= \gamma n - \gamma\sum_{i=1}^{n} Q_i^t \\
\sum_{i=1}^{n} Q_i^t &= \gamma n \\
\gamma &= \frac{\sum_{i=1}^{n} Q_i^t}{n}.
\end{aligned}$$

This is the same result as for MEME [156]. Thus, the M-step is not affected by using the modified prior probability distribution.

3.3.6 The TCM model in MEMERIS

For a TCM model, there can be zero, one, or more than one motif occurrences per sequence. In MEME, the probability of starting a motif at position j in sequence X_i is

$$P(Z_{i,j} = 1) = \lambda$$

which is independent on the position. The expected number of start positions per sequence is

$$\sum_{j=1}^{m} \lambda = \lambda m.$$

To integrate the single-strandedness, we have to project λm expected motif occurrences onto the prior probability distribution derived from the PU values $\frac{PU_{i,j}+\pi}{\sum_{k=1}^{m}(PU_{i,k}+\pi)}$. To this end, we convert λ into position-specific parameters $\lambda_{i,j}$ with

$$\lambda_{i,j} = P(Z_{i,j} = 1|\lambda, PU_i) = \lambda m\Big(\frac{PU_{i,j}+\pi}{\sum\limits_{k=1}^{m}(PU_{i,k}+\pi)}\Big).$$

Naturally, this conversion is limited to cases where

$$\lambda_{i,j_{max}} = \lambda m\Big(\frac{PU_{i,j_{max}}+\pi}{\sum\limits_{k=1}^{m}(PU_{i,k}+\pi)}\Big) \leq 1 \tag{3.1}$$

with $j_{max} = \text{argmax}_j\ PU_{i,j}$. If this condition is not fulfilled, one can either smooth the probability distribution or reduce the expected number of motif starts. We decided to

smooth the prior probability distribution by choosing a pseudocount π for sequence X_i and the respective EM iteration, which fulfills equation 3.1. In the MEMERIS implementation, in cases where equation 3.1 is not fulfilled, we compute

$$\pi = \frac{-PU_{i,j_{max}}\lambda m + \sum_{k=1}^{m} PU_{i,k}}{m(\lambda - 1)}.$$

Using this new pseudocount results in $\lambda_{i,j_{max}} = 1$. This assures that the expected number of motif occurrences is not affected even though the ratio $\frac{\lambda_{i,j_1}}{\lambda_{i,j_2}}$ might be changed in an iteration-specific manner. However, during extensive MEMERIS tests on artificial and biological data (see sections 3.3.7 and 3.3.8), we found π to be unaffected in the great majority of tests and only slightly changed in the remaining cases.

E-step Since a sequence can have more than one motif, we consider in the TCM model not the complete sequence X_i but the substring of length W that starts at position j (denoted as $X_{i,j}$). The E-step in iteration t for the TCM model in MEMERIS is

$$\begin{aligned}
Z_{i,j}^t &= E_{Z|X,\Theta^t,\lambda^t,PU}[Z_{i,j}] \\
&= P(Z_{i,j} = 1|X_{i,j}, \Theta^t, \lambda_{i,j}^t, PU_i) \\
&= \frac{P(X_{i,j}|Z_{i,j} = 1, \Theta^t, \lambda_{i,j}^t, PU_i)P(Z_{i,j} = 1|\lambda_{i,j}^t, PU_i)}{P(X_{i,j}|\Theta^t, \lambda_{i,j}^t, PU_i)} \\
&= \frac{P(X_{i,j}|\Theta_1^t)\,\lambda_{i,j}^t}{P(X_{i,j}|\Theta_1^t)\,\lambda_{i,j}^t + P(X_{i,j}|\Theta_0^t)(1 - \lambda_{i,j}^t)}
\end{aligned}$$

where λ^t denotes the respective parameter in iteration t. Note that $Z_{i,j} = 1$ implies that $X_{i,j}$ is sampled from Θ_1 and $Z_{i,j} = 0$ implies that $X_{i,j}$ is sampled from Θ_0. Again, the substring $X_{i,j}$ does not depend on λ and on the PU values.

M-step In contrast to the OOPS and ZOOPS model, the M-step in the TCM model is different between MEME and MEMERIS. $\phi = (\Theta, \lambda)$ is the complete parameter set for the TCM model.

The joint log likelihood of all sequences in MEMERIS is

$$
\begin{aligned}
&\log P(X, Z|\phi, PU) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \log P(X_{i,j}|Z_{i,j}, \phi, PU_i)P(Z_{i,j}|\phi, PU_i) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \log \Big[\Big(P(X_{i,j}|Z_{i,j}=0, \phi, PU_i)P(Z_{i,j}=0|\phi, PU_i) \Big)^{1-Z_{i,j}} \\
&\qquad\qquad \cdot \Big(P(X_{i,j}|Z_{i,j}=1, \phi, PU_i)P(Z_{i,j}=1|\phi, PU_i) \Big)^{Z_{i,j}} \Big] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \Big[(1-Z_{i,j}) \Big(\log P(X_{i,j}|Z_{i,j}=0, \phi, PU_i) + \log P(Z_{i,j}=0|\phi, PU_i) \Big) \\
&\qquad\qquad + Z_{i,j} \Big(\log P(X_{i,j}|Z_{i,j}=1, \phi, PU_i) + \log P(Z_{i,j}=1|\phi, PU_i) \Big) \Big] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \Big[(1-Z_{i,j}) \Big(\log P(X_{i,j}|\Theta_0) + \log(1-\lambda_{i,j}) \Big) + Z_{i,j} \Big(\log P(X_{i,j}|\Theta_1) + \log \lambda_{i,j} \Big) \Big].
\end{aligned}
$$

Using the expected values of the hidden variables $Z_{i,j}^t$

$$
\begin{aligned}
&E_{Z|X,\phi^t,PU}[\log P(X, Z|\phi, PU)] \\
&\quad = \sum_{i=1}^{n}\sum_{j=1}^{m} \Big[(1-Z_{i,j}^t) \Big(\log P(X_{i,j}|\Theta_0) + \log(1-\lambda_{i,j}) \Big) \\
&\qquad\qquad + Z_{i,j}^t \Big(\log P(X_{i,j}|\Theta_1) + \log \lambda_{i,j} \Big) \Big] \\
&\quad = \sum_{i=1}^{n}\sum_{j=1}^{m} (1-Z_{i,j}^t) \log P(X_{i,j}|\Theta_0) + \sum_{i=1}^{n}\sum_{j=1}^{m} Z_{i,j}^t \log P(X_{i,j}|\Theta_1) \\
&\qquad + \sum_{i=1}^{n}\sum_{j=1}^{m} (1-Z_{i,j}^t) \log(1-\lambda_{i,j}) + \sum_{i=1}^{n}\sum_{j=1}^{m} Z_{i,j}^t \log \lambda_{i,j}.
\end{aligned}
$$

The first two terms only depend on Θ and finding the Θ that maximizes these terms is the same as for MEME [156]. Writing $\lambda m \sigma_{i,j}$ for $\lambda_{i,j}$ with $\sigma_{i,j} = \dfrac{PU_{i,j} + \pi}{\sum_{k=1}^{m}(PU_{i,k} + \pi)}$, the last two terms become a function of λ

$$
f(\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{m} (1-Z_{i,j}^t) \log(1-\lambda m \sigma_{i,j}) + \sum_{i=1}^{n}\sum_{j=1}^{m} Z_{i,j}^t \log(\lambda m \sigma_{i,j}).
$$

However, finding the λ that maximizes this function is different in MEMERIS. We set the first derivative of $f(\lambda)$ with respect to λ to zero

$$
-\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(1-Z_{i,j}^t) m \sigma_{i,j}}{1-\lambda m \sigma_{i,j}} + \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{Z_{i,j}^t}{\lambda} = 0.
$$

This equation cannot be analytically solved for λ. Therefore, we apply the Newton-Raphson method to find the approximation of the root of the first derivative. We iteratively compute

$$\lambda_{k+1} = \lambda_k \frac{f'(\lambda_k)}{f''(\lambda_k)}$$

where $f'(\lambda)$ and $f''(\lambda)$ are the first and second derivatives of $f(\lambda)$, respectively. The starting point is set to $\lambda_0 = \lambda^t$. The second derivative of $f(\lambda)$ is

$$f''(\lambda) = -\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(1 - Z_{i,j}^t)m^2\sigma_{i,j}^2}{(1 - \lambda m\sigma_{i,j})^2} - \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{Z_{i,j}^t}{\lambda^2}.$$

In practice, this method finds the root in a few iterations with a precision of 10^{-10}. Furthermore, using λ^t as the starting point contributes to the quick convergence since λ is usually only slightly changed.

Runtime

Except for the offset due to the computation of the secondary structure values and the Newton-Raphson method in the TCM model, MEMERIS has the same runtime like MEME. In practice, MEMERIS is reasonable fast for typical and even large data sets. For example, searching two motifs of length 6 nt in a set of 120 sequences of length 50 nt takes less than a minute on a workstation with a 2 GHz CPU.

3.3.7 Testing MEMERIS on artifical data sets

To test whether the secondary structure information integrated into MEMERIS is able to guide the motif search towards single-stranded regions, we started with tests using artificial data sets.

Design of artificial test sets

The basic principle of the design of the artificial data sets is as follows. Each test set consists of 20 sequences that contain motifs either as a fixed string or as a sample from a position-specific probability matrix (PSPM) in single- and/or double-stranded conformation. Each artificial test sequence consists of a random sequence part at the 5' and 3' end and a stem-loop structure that contains a single-stranded motif (called *ssMotif*) in the hairpin loop and a double-stranded motif (called *dsMotif*) on either side of the stem (Figure 3.5A). We generated random RNA sequences by sampling from the uniform distribution (probability of 0.25 for A,C,G, and T). We allowed base pairs between A and T, C and G, G and T. With a probability of 0.5, we changed one position from the complementary part of the double-stranded motif so that it cannot form a base pair anymore. This mutation and the possibility of base pairs between G and T assure

that the complementary part of the dsMotif is not a fixed string, which might otherwise interfere with the motif finding. The stem consisted of 12 base pairs, the total length of the random sequence up- and downstream was set to 20 nt. We set the motif length to 6 nt, which is a typical motif length for an RNA binding protein. All test sets are also described in Figure 3.5.

For the tests below, we set the pseudocount π to 0.1 for test sets 1-6 and 9 and π to 0.01 for test sets 7 and 8. For MEME and MEMERIS, we set the background letter probability distribution to a uniform distribution (parameter -bfile), since the sequences are too small for an accurate frequency estimation and the artificial sequences were sampled from a uniform distribution.

OOPS model

First, we tested the OOPS model by comparing MEME with MEMERIS. We asked whether the EF or PU values influence which motif is found in the first pass, given that one motif is rather single-stranded, while the other one is rather double-stranded. This will be important if a user wants to discover only a single motif. We designed sequences containing both a ssMotif and a dsMotif as a fixed string (test set 1, Figure 3.5B). We found that MEME always detects the most upstream motif in the first sequence in the first pass. Of course, this motif can be the dsMotif, depending on the order of the sequences. In contrast, MEMERIS using EF or PU values always detects the ssMotif first, independent of the order of the sequences.

Next, we sampled the motifs from two PSPMs (test set 2), where the second PSPM was derived from the first one by randomly permutating the letter probabilities. This procedure yields two PSPMs with an equal information content. The information content of a PSPM measures the strength of the motif and is computed as

$$\sum_{i=1}^{W} \sum_{j} f_{i,j} \log_2(\frac{f_{i,j}}{q_j})$$

where $f_{i,j}$ is the probability of the jth letter in the alphabet at position i of the motif and q_j is the background probability of the jth letter. Again, MEMERIS but not MEME detects the ssMotif in the first pass. Furthermore, these results are not affected by increasing the total sequence length from 50 to 80 nt.

Next, we asked whether MEMERIS also detects the ssMotif in the first pass, even if the ssMotif is weakened by introducing a single mutation in 25% of its occurrences (test set 3) or by sampling from a PSPM with a lower information content (test set 4, Figure 3.5B). While MEME detects the stronger dsMotif in the first pass, MEMERIS identifies the weaker ssMotif first. To exclude that these findings are affected by some unknown bias in the motif or in the sequences, we repeated all tests two times with

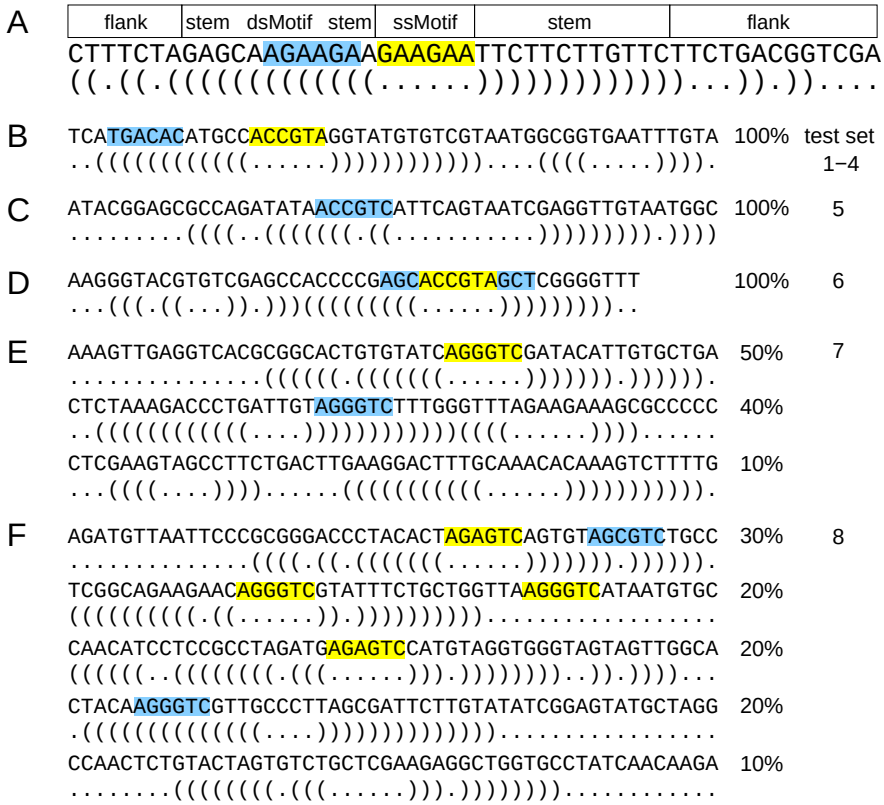

Figure 3.5: Overview of the artificial test sets.
(A) The figure shows an artificial sequence with a single-stranded motif (ssMotif, highlighted yellow) and a double-stranded motif (dsMotif, highlighted blue) together with its optimal secondary structure. The general scheme for constructing sequences is (i) to randomly sample an up- and downstream flank with a total length of 20 nt, (ii) to generate a stem of 12 base pairs that contains the dsMotif, and (iii) to insert the ssMotif as the hairpin loop. The dsMotif can occur on either side of the stem. (B) The sequences in test sets 1-4 contain a ssMotif as well as a dsMotif. For test sets 1 and 3, we used a fixed string as the ssMotif (ACCGTA in this example, yellow) and a permutation of it as the dsMotif (TGACAC, blue). These motifs are sampled from two PSPMs for test sets 2 and 4. In test set 3, a single mutation is introduced in 25% of the ssMotifs. (C) Test set 5 contains only one motif in double-stranded conformation (sampled from a PSPM). (D) Sequences in test set 6 contain a 12 nt motif as a fixed string, where only the 6 nt in the middle of the motif (yellow) are single-stranded. (E) Sequences in test set 7 contain either a ssMotif, a dsMotif, or no motif (sampled from a PSPM). (F) Test set 8 contains sequences with a ssMotif and a dsMotif, with two ssMotifs, with one ssMotif, with one dsMotif, and without a motif (sampled from a PSPM). The percentages indicate to which fraction sequences with the respective features are contained in the data set.

new random sequences and different motifs and found consistent results. In general, PU values performed equally well or better than EF values in these tests.

To illustrate the effect of varying the pseudocount π, we designed a test set containing only a dsMotif (test set 5, Figure 3.5C). MEMERIS using PU values detects the dsMotif, if the pseudocount π is higher than 0.22 (Figure 3.6). Lower values for π lead to the detection of other motifs with a higher average single-strandedness (defined as the average of the PU values for all detected motif occurrences). These motifs differ from the dsMotif and are therefore weaker and less significant (as indicated by a lower information content of the resulting motif matrix in Figure 3.6). Thus, the pseudocount π provides an easy means for the user to adjust the importance of the secondary structures. Naturally, PU values (the probability that a complete substring is unpaired) are stricter than EF values (the fraction of the substring that is not involved in base pairing). Thus, MEMERIS using PU values will favor single-stranded regions stronger than MEMERIS using EF values (Figure 3.4). Consistently, for test set 5, MEMERIS using EF values discovers the dsMotif independently of the pseudocount.

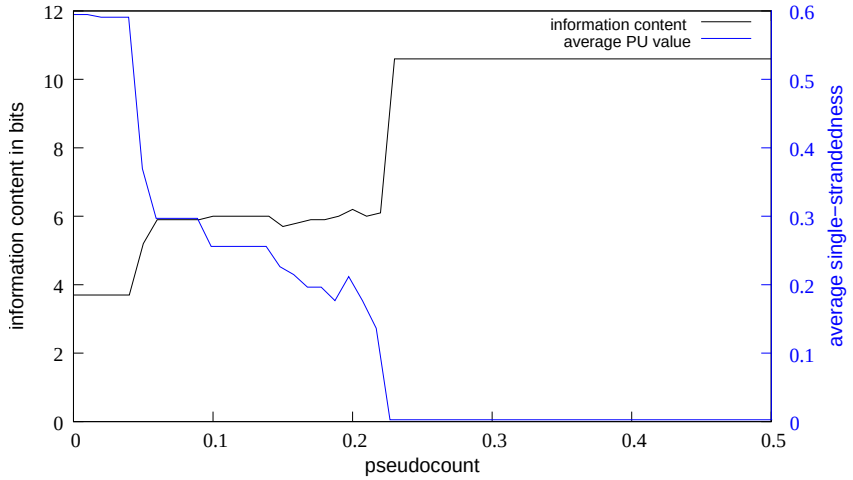

Figure 3.6: Effect of varying the pseudocount.
The figure shows the information content of the motif PSPM found by MEMERIS in bits (black curve) and its average single-strandedness (average PU values of all motif occurrences, blue curve) for pseudocounts from 0 to 0.5 in steps of 0.01. Test set 5 that contains sequences with only one dsMotif (10.6 bits, average single-strandedness 0.003) was used. This motif is found by MEMERIS for a pseudocount greater than 0.22. In general, the lower the pseudocount, the higher is the average single-strandedness of the motif occurrences.

Next, we tested the ability of MEMERIS to identify the single-stranded part of a longer sequence motif as the potential protein binding site. We designed a test set containing a 12 nt motif, whose three positions at the beginning and at the end form base pairs (test set 6, Figure 3.5D). Setting the motif length to 6 nt, MEME identifies the first 6 nt of this 12 nt motif, while MEMERIS exactly finds the 6 nt that are not involved in base pairing.

We conclude that MEMERIS preferably selects single-stranded motif occurrences and that it is able to identify a weaker over a stronger motif, if the average single-strandedness is sufficiently higher.

ZOOPS and TCM model

In addition to identifying the motif locations, the ZOOPS and TCM model have to solve a further question: How many motif occurrences are in the data set? We intended to integrate the secondary structure information in a way that guides but not restricts the motif search to single-stranded regions. Therefore, this additional question is only marginally affected in MEMERIS. Up to which single-strandedness a motif occurrence is believed to be a real protein binding site is hard to determine in an automatic manner since this would necessitate statistical measures that take the motif sequence as well as its structural properties into account. Furthermore, the requirement for single-strandedness certainly depends on the dataset and on the binding protein. However, we propose a simple procedure that requires the user to decide according to the motif sequence and the EF or PU values how many occurrences are there in the given data set.

1. Run MEMERIS using a rather high pseudocount π, which mimics a MEME run and leads to the detection of motif hits nearly independent of the single-strandedness.
2. Inspect the sequence and the single-strandedness of all detected motif hits and determine the number of motif occurrences.
3. Run MEMERIS again with a low pseudocount π and a fixed number of motif occurrences (parameter -minsites and -maxsites).

The second MEMERIS run with a fixed number of motif occurrences should result in the identification of single-stranded occurrences and thus a refinement of the final motif matrix.

For the ZOOPS model, we tested this on a data set that contains sequences with either (i) one ssMotif, (ii) one dsMotif, or (iii) without a motif (test set 7, Figure 3.5E). For the TCM model, we applied this procedure to a data set consisting of sequences having either (i) one ssMotif and one dsMotif, (ii) two ssMotifs, (iii) one ssMotif, (iv) one dsMotif, or (v) no motif (test set 8, Figure 3.5F). Since point 2 above involves manual inspection, we have to avoid any bias arising from our knowledge about the PSPM and the data set. Thus, we assessed the number of motif occurrences in an automatic manner by simply counting the number of motif hits having an EF or PU value greater than 0.5. Comparing MEME and MEMERIS with a given number of motif hits, MEMERIS identifies the single-stranded motif occurrences, even if this leads to a lower information content of the motif. One example for the TCM model is shown in Figure 3.7. Again, PU values often led to better results than EF values.

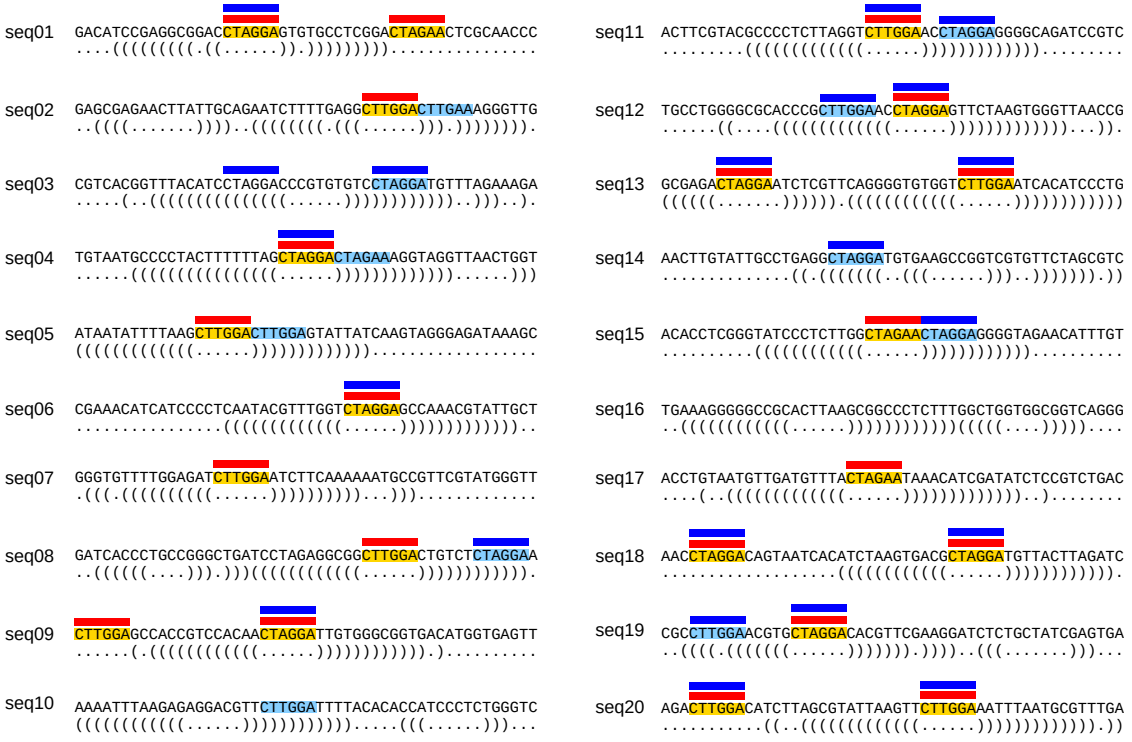

Figure 3.7: Comparison of MEME and MEMERIS for test set 8 (testing the TCM model). The figure shows 20 sequences that contain ssMotifs (highlighted yellow) and/or dsMotifs (highlighted light blue). The optimal structure is shown below each sequence. Red and blue bars indicate the position of the motif occurrences found by MEMERIS and MEME, respectively. While MEMERIS detects all ssMotifs and no dsMotif leading to an information content of the motif matrix of 10.4 bits, MEME identifies a stronger motif (11.1 bits) but detects eight dsMotif occurrences. MEMERIS results are shown for PU values and a pseudocount of 0.01. The number of motif hits was set to 21 for MEME and MEMERIS.

Motifs in single-strands of arbitrary structures

By design, the above test sets contain the ssMotif in the loop of a hairpin and the dsMotif in the stem. In contrast to programs that search for RNA sequence-structure elements, MEMERIS should be able to identify a single-stranded motif independent of the structural element in which it is contained. To test this, we designed a test set where the motif is located either (i) in a hairpin loop, (ii) in an internal loop, (iii) in a single-stranded part of a multiple loop, or (iv) between two stems (test set 9). While MEMERIS (and MEME too) clearly detects the motif, two RNA motif finders, RSMatch [160] and CMfinder [161] (that are not designed for this task) are not able to discover any motif in this test set.

3.3.8 Testing MEMERIS on biological data sets

To evaluate MEMERIS in a more realistic way, we applied it to real biological data sets, like SELEX data.

SELEX data

We tested MEMERIS on SELEX data that are found to contain sequence motifs in single-stranded conformations. Buckanovich et al. identified 33 TCAT or ACAT repeats in the hairpin loops of the SELEX winner sequences of the neuron-specific splicing factor Nova-1 [129]. Searching for 33 motif occurrences with a TCM model and a motif length of 4 nt, MEMERIS exactly identifies those 33 TCAT and ACAT hits that are described in [129] (Figure 3.8). MEME also detects the correct motif, but at least two of its motif hits are located outside the hairpin loop and are presumably no Nova-1 binding sites (SB2 and SB4 in Figure 3.8).

Next, we tested MEMERIS on the SELEX data of the splicing factors SF2/ASF and SC35 [151, 152]. Comparing MEMERIS results with the results given in the respective publications, we found similar motifs although the positions of the motif hits differ in several sequences. However, a detailed evaluation turned out to be difficult due to the following reasons. Firstly, these splicing factors bind to degenerate sequence motifs, thus motif occurrences at different positions can result in highly similar motif matrices. Secondly, the location of the binding sites in all SELEX winner sequences have not been determined experimentally. Thus, the real binding sites in these sequences remain unknown and we do not know whether the MEMERIS motif hits are correct or not. Therefore, we further evaluate MEMERIS on other biological sequences with known binding sites.

Protein binding sites in cis-acting RNA elements

Cis-acting elements in the UTR regions of mRNAs can be important for mRNA stability and translation efficiency by providing binding sites for regulatory proteins. These elements are often conserved at the sequence and secondary structure level. Thus, they are fundamentally different compared to the randomly generated SELEX sequences. To test the ability of MEMERIS to identify protein binding sites in the larger context of conserved sequence-structure elements, we selected the sequences of cis-acting RNA elements having a defined secondary structure and a known protein binding site from the Rfam database [169]. Redundant sequences with complete identity were taken only once.

The iron responsive element (IRE, Rfam entry RF00037) located in the 5' UTR of mRNAs is essential for the expression of proteins that are involved in the iron metabolism [170]. The IRE consists of a stem-loop structure and the nucleotides in the hairpin loop are bound by iron-regulatory proteins. MEMERIS detects the hairpin loop as the

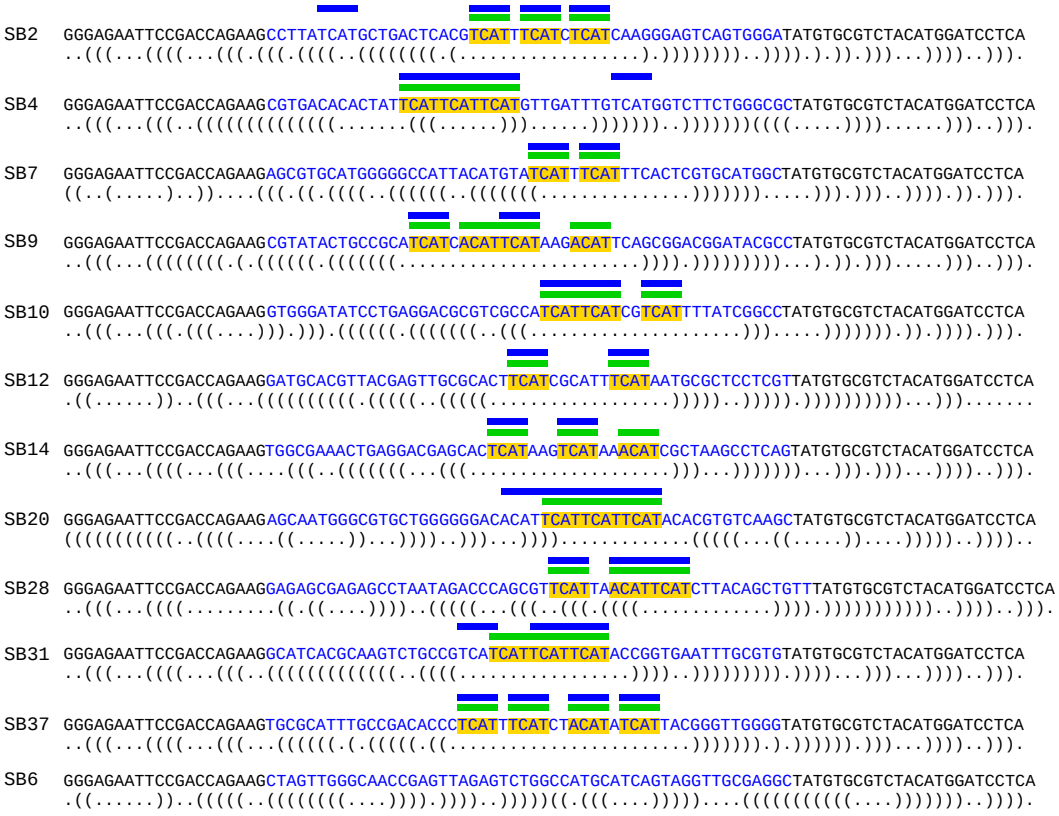

Figure 3.8: Comparison of MEME and MEMERIS for the SELEX sequences of Nova-1.
The figure shows the sequences and labels of the individual clones described in [129]. The random oligonucleotides are in blue letters. The primer binding sites (black letters) were included in the RNA secondary structure prediction but not in the motif search. The optimal secondary structure is shown below each sequence. The TCAT and ACAT motifs identified in [129] are highlighted yellow. Blue and green bars indicate the positions of the motif hits found by MEME and MEMERIS, respectively. The motif matrix found by MEME has an information content of 7.6 bits, the MEMERIS motif matrix has 7.4 bits. MEME and MEMERIS were run with the TCM model and the number of motif hits was set to 33. MEMERIS results are shown for PU values and a pseudocount of 0.01.

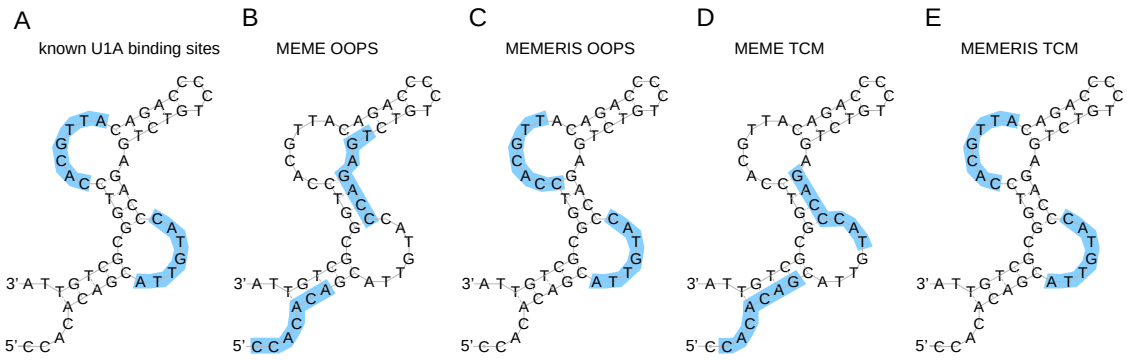

Figure 3.9: Results of MEME and MEMERIS for the PIE Rfam (RF00460) data set.
The figure shows the consensus sequence and consensus structure of the PIE RNA. The U1A protein binds the single-stranded sequences in the two asymmetrical internal loops in a cooperative manner (A). Using the OOPS model, MEME finds two motifs (14 and 13.3 bits, resp.) that do not overlap the real binding sites (B), while MEMERIS finds the real upstream binding site exactly (11.8 bits) and the downstream site (10.5 bits) with a shift of one position (C). Since both individual binding sites are very similar, we used the TCM model to search for a motif with two occurrences in each sequence. Again MEME finds a different motif (11.6 bits) (D), while MEMERIS detects the correct protein binding sites (10.7 bits) (E). The known binding sites and the predicted motifs are highlighted in blue. The motif length was set to 7 nt. For MEMERIS, the PU values were used with a pseudocount of 0.01.

motif hit in all sequences, leading to a motif matrix with an information content of 10 bits, while MEME discovers a stronger motif (10.8 bits) that is moved to one position upstream. In addition, MEME identifies a different motif occurrence in the upstream stem in two sequences.

The polyadenylation inhibition element (PIE) contains two binding sites for U1A proteins [171]. U1A binding leads to an inhibition of the poly(A) polymerase and a reduced mRNA stability and translation efficiency due to a shortened poly(A) tail. Interestingly, U1A autoregulates itself by binding to a PIE in its own 3' UTR. PIE consists of a stem structure with two asymmetric internal loops that represent U1A binding sites. Both internal loops are identified by MEMERIS using the TCM model or searching for two distinct motifs with the OOPS model (Figure 3.9). MEME detects stronger motifs in both models that are different from the known binding sites.

The trans-activation response (TAR) element of the HIV-1 virus is required for efficient transcription [172, 173]. The hairpin loop of this element is bound by a heterodimer consisting of Tat and CycT1 proteins. MEMERIS clearly identifies the motif in the hairpin loop, while MEME detects a stronger motif located in the stem (Figure 3.10). The Tat protein also binds the pyrimidine-rich 3 nt bulge loop of the TAR element. However, neither MEMERIS nor MEME is able to identify this binding site because this motif is too degenerate and in several TAR elements this bulge consists of only two nucleotides.

The stem-loop destabilizing element (SLDE) consists of three stems located in the 3' UTR of *CSF3* mRNAs and is used to regulate the stability of the mRNA [174]. The

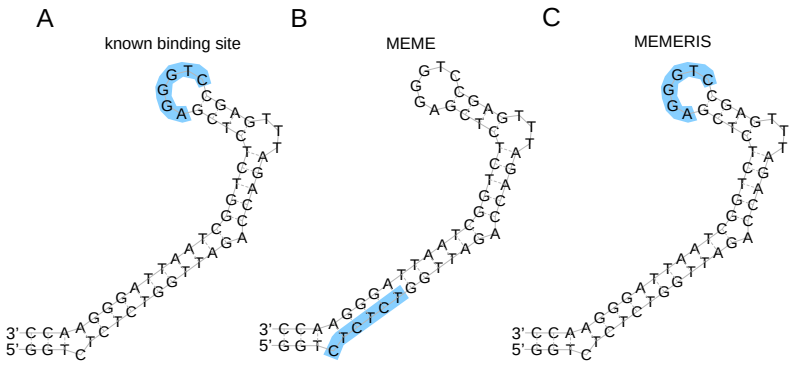

Figure 3.10: Results of MEME and MEMERIS for the TAR Rfam (RF00250) data set. The figure shows the consensus sequence and structure of the TAR element. The hairpin loop is bound by the Tat protein (A). We searched for one binding site in each sequence (OOPS model) with MEME (B) and MEMERIS (C). MEME detects a motif (12 bits) that does not overlap the known binding site, while MEMERIS identifies the binding site, although the respective motif is noticeably weaker (10 bits). The known binding sites and the predicted motifs are highlighted in blue. The motif length was set to 6 nt. For MEMERIS, the PU values were used with a pseudocount of 0.01.

hairpin loop sequence of the third stem is essential for the function of this element and assumed to be bound by an unknown protein. Again, MEMERIS detects this loop as the motif, while MEME finds a different motif (Figure 3.11).

3.4 Discussion

We analyzed the secondary structure of an extensive set of 77 experimentally verified splicing motifs in their natural context and found that they have a higher single-strandedness. These results were also confirmed by minigene experiments demonstrating that single-stranded splicing motifs exert a stronger effect on the splicing pattern. Although long time thought to contain only the codon sequence, it has become clear that the coding sequence of an mRNA is superposed with additional signals that determine its fate during the numerous processing steps. These signals comprise editing sites, binding sites for proteins regulating constitutive and alternative splicing, and additionally (if not located in UTR regions) sites that determine mRNA localization, export, stability, and translation. Since splicing motifs are abundant in exons [32], we argue that a typical exon is subjected to at least three different selection pressures:

- preserving the coding sequence,
- preserving the splicing motifs,
- and preserving an appropriate structural context for these splicing motifs.

We observed a lower single-strandedness for exonic motifs compared to intronic ones. This might be due to the strong selection pressure to preserve the coding sequence that

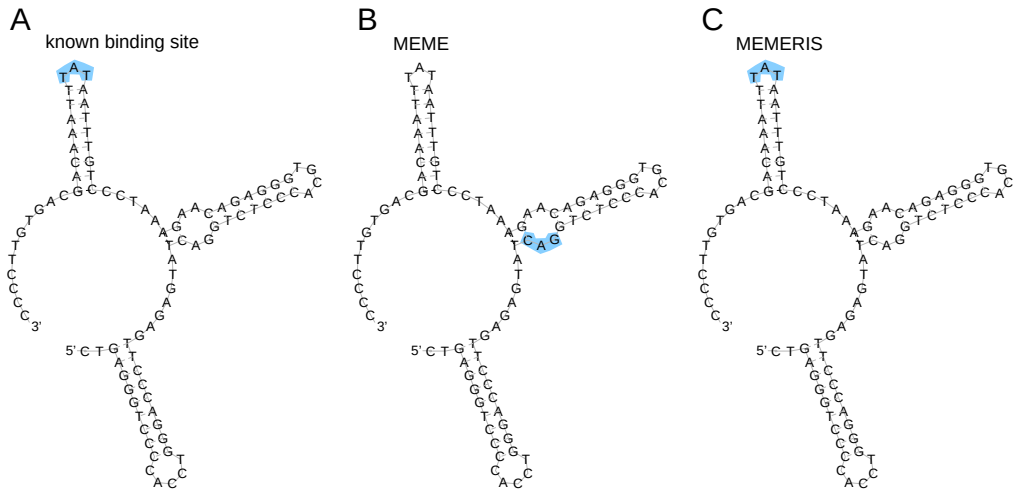

Figure 3.11: Results of MEME and MEMERIS for the SLDE Rfam (RF00183) data set. The figure shows the consensus sequence and structure of the SLDE element. The hairpin loop of the essential third stem is bound by an unknown protein factor (A). MEME detects a CAG motif, which does not overlap the binding site (B). In contrast, MEMERIS identifies the TAT sequence of the hairpin loop as the motif (C). Both motif matrices have an information content of 6 bits. The known binding sites and the predicted motifs are highlighted in blue. The motif length was set to 3 nt. For MEMERIS, the PU values were used with a pseudocount of 0.01.

prevents an extensive selection for a high single-strandedness for exonic motifs. If more splicing motifs are experimentally determined, it would be interesting to compare the structural context of exonic motifs in the UTR with intronic ones.

Our study indicates that the structural context can play a role in mutagenesis experiments. Usually, single or multiple residues were mutated or deleted and the observed variation in the splicing pattern is directly interpreted as a function of the affected residues. However, the mutation or deletion of residues might also change the secondary structure context of an adjacent motif and, as in case of the fibronectin EDA exon [133], this would provide another explanation for the observed effects. Moreover, the again and again observed 'context effects' in such experiments might be partially explained by structural changes. We propose that secondary structures should be taken into consideration for such experiments, especially if the exact location of a binding site or the interplay of two binding sites in close proximity is the subject of investigation. Our measurements of single-strandedness can easily be used to plot the values for all motifs of a fixed length over the complete sequence. The comparison between the wild-type and the mutated sequence can give a hint if the introduced mutations result in drastic structural changes. The secondary structure should also be taken into consideration when interpreting the results of mutagenesis studies. For example, the exon 8 of the rat beta-tropomyosin gene contains three purine rich ESE candidates, but only the second and the third candidate were shown to act as an enhancer [175]. These experimental results correlate well with

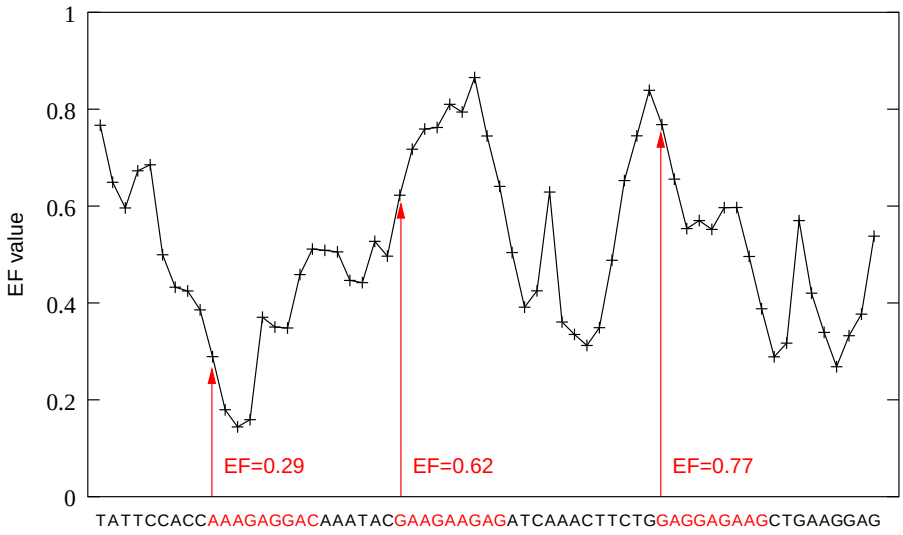

Figure 3.12: Plot of EF values for exon 8 of the rat beta-tropomyosin gene.
Each point in this plot gives the EF value for the 9-mer sequence starting at this position. RNA folding was done using a 30 nt genomic context upstream and downstream. The three purine rich ESE candidates are shown in red. While the first motif is more double-stranded, the second and third one are found in a rather single-stranded context (as indicated by a higher EF value). This agrees well with experimental results as the mutation of the second and third but not the first motif affects splicing [175].

the predicted structural context of these three candidates (Figure 3.12).

In light of our results, it might also be interesting to re-evaluate the effect of single nucleotide polymorphisms (SNPs) on splicing. If a SNP (especially a translationally silent SNP) leads to a change in the splicing pattern, this SNP is often assumed to destroy or create a splice site or a splicing motif [176, 63]. However, SNPs are also reported to change the mRNA secondary structure [177]. Thus, it is tempting to speculate that the influence on splicing for some SNPs is not due to the direct impact on a splicing motif, but due to a change in secondary structure that sequesters a previously single-stranded binding site (or vice versa).

As secondary structures are one aspect for discriminating real from false-positive binding sites, it might be beneficial to consider secondary structure contexts in computational approaches to predict splicing motifs. Likewise, using this information in splicing simulation and gene prediction algorithms may result in a higher accuracy [178].

Despite the significant preference of verified splicing motifs for single-strands, individual motifs are located in a rather double-stranded context. It would be interesting to find out how these motifs are bound by trans-acting factors. In this situation, the predicted secondary structure might be wrong due to protein binding in the proximity of the motif, which can result in a disruption of a hindering structure. Alternatively, the computed structures might be incorrect since the considered context lengths are too

short or too long for the respective motif loci. As another hypothesis, RNA dependent helicases might be responsible for a resolution of hindering structures [136, 138, 139].

As a misregulation of alternative splicing is the basis for many human diseases, alternative splicing is an important therapeutic target [63, 179]. Antisense oligonucleotides that target splice sites can correct the splicing pattern of a gene [68]. Recently, small antisense oligonucleotides were shown to influence mRNA stability by modulating the binding affinity of the regulatory protein HuR [132, 144]. These oligonucleotides hybridize outside the HuR binding site and exert their effect by altering the secondary structure of the binding site. Such designed antisense oligonucleotides may also change the secondary structure of splicing enhancers or silencers, thus providing an alternative way to correct the splicing pattern. Interestingly, by increasing or decreasing the single-strandedness of a splicing motif, this mechanism would allow a positive as well as a negative regulation [132, 144].

The basic splicing signals (splice sites, branch point) do not contain sufficient information to allow accurate intron and exon definition [22]. Further information comes from enhancer and silencer motifs, but there is still a noticeable gap to the information that is needed to achieve the accuracy of the spliceosome machinery (Figure 4 in [22]). This gap might be narrowed by additional information coming from the secondary structures. We believe that our findings provide another piece for the decoding of the mRNA splicing code [178].

Using the defined measurements of single-strandedness, we developed a new motif finding approach MEMERIS that simultaneously searches a sequence motif and integrates information about secondary structures. In contrast to other algorithms, MEMERIS abstracts from specific structural elements by using EF or PU values. Performing tests with artificial and biological data, we have shown that MEMERIS is able to identify single-stranded sequence motifs, which often represent the known protein binding motif. Compared to MEME, MEMERIS achieves a higher accuracy in our tests, which demonstrates that the additional information about the single-strandedness of putative motif occurrences is useful. To maintain a secondary structure, a mutation in a base pair often requires a compensatory mutation. This may result in a stronger selection pressure for double-stranded compared to single-stranded sequence regions. Consistently, RNA-binding proteins may bind degenerate sequences [151]. Therefore, it is a valuable property that MEMERIS is able to select a weaker over a stronger motif, if this motif has a higher average single-strandedness (exemplified in Figures 3.9 and 3.10). We believe that the application of MEMERIS to SELEX data of splicing factors and other RNA binding proteins can help to identify the real binding motif. To summarize, RNA secondary structure properties are important for distinguishing real from spurious protein binding sites and should be considered when searching for the binding motif of a protein.

The general principle of MEMERIS to include prior knowledge about the motif start sites can be extended to other applications. It is straightforward to search for sequence motifs in double-stranded structure parts, for example by computing the expected fraction of bases that are paired (1-EF) or the probability that the complete motif occurrence is paired. A further application can be the search for transcription factor binding sites in DNA promoter sequences. If information is available that a DNA motif is preferably located in proximity to the transcription start site, the prior start site distribution can be adjusted to have higher probabilities for the 3' sequence ends of promoter sequences. Since highly condensed DNA regions are inaccessible to transcription factors [180], prior knowledge about chromatin condensation and higher-order chromosomal structures can be used to prevent the detection of motifs in inaccessible regions. Interestingly, an informative prior was applied to a related idea recently. Narlikar et al. used information about the structural class of a transcription factor to identify DNA regions that are more likely bound by this protein [181].

In future, it would be desirable to automatically determine the number of single-stranded motifs in a ZOOPS or TCM model. This is challenging because the degree to which a real binding site has to be single-stranded certainly depends on the respective protein. Furthermore, this requirement for single-strandedness may be affected by the presence of RNA helicases that are involved in several important processes like splicing and translation. Finally, the statistical models, which currently only evaluate the motif sequence, need to be extended to account for the sequence and the secondary structure context of a motif.

Chapter 4

Genome-wide bioinformatics analysis of alternative splicing at tandem splice sites

Much research about alternative splicing focused on larger alternative splice events that often result in noticeable effects on the function of a protein. On the other hand, there are incidental reports of alternative splice events that lead to the production of very similar protein isoforms. However, these events received little attention in the past.

In the third part of this thesis, we performed a genome-wide analysis of splice events that result in only minor changes of the mRNA and the respective protein. We show that most of these subtle events are produced by alternative splicing at so called NAGNAG or tandem acceptors. We found several significant biases indicating that NAGNAG acceptors are non-randomly distributed at the genome and protein level, suggesting that these splice events are subjected to selection pressures during evolution. This view is supported by our finding that a subset of NAGNAG acceptors is conserved in mouse and that these splice events can be regulated in a tissue-specific manner. To study the NAGNAG splice mechanism and to check a potential disease relevance of the splice events, we also investigated human polymorphic NAGNAGs. Given that thousands of human genes have NAGNAG acceptors, these splice events represent one major mechanism to increase the diversity of the human proteome.

Then, we extended our analysis to alternative splicing at so called GYNGYN or tandem donors. Investigating what distinguishes alternatively from not alternatively spliced GYNGYN donors, we found differences in the binding to U1 snRNA, overrepresented sequence motifs, and a higher conservation of the exonic and intronic flanks between human and mouse.

Apart from human, we found alternative GYNGYNs and NAGNAGs in many other eukaryotic species. We continued to develop a relational database, TassDB, which stores the wealth of data we had collected about alternative tandem splice sites. This database facilitates further experimental studies and large-scale bioinformatics analyses that are required to address important open questions concerning these subtle splice events.

Plan of the chapter

An introduction into subtle alternative splice events and the motivation for the following bioinformatics analysis is given in section 4.1. The genome-wide study of NAGNAG acceptors is described in section 4.2. In section 4.3, we investigate polymorphic NAGNAG acceptors. We extend our tandem acceptor analysis to the special type of U12 introns in section 4.4. The genome-wide analysis of tandem donors is described in section 4.5. Section 4.6 contains a description of the TassDB database. Finally, we summarize our findings and discuss implications and open questions in section 4.7.

4.1 The impact of subtle alternative splice events?

Exon skipping is one of the most frequent types of alternative splicing [182]. As the average length of an alternative exon is 137 nt [182], its inclusion or skipping affects a larger region of the protein (about 45 amino acids). Studies have shown that such alternative splice events often result in functionally different protein isoforms. These isoforms can differ in various aspects including ligand binding affinity, signaling activity, protein domain composition, subcellular localization, and protein half-life [47, 51, 52, 53]. Furthermore, the introduction of a frameshift by skipping an exon with a length that is not divisible by three can lead to non-functional proteins or to a degradation of the transcript [59]. The effect of larger splice events on the proteins, their regulation, and their frequency has been the subject of numerous experimental and computational studies [183, 51, 74, 184, 185].

On the other hand, there are incidental experimental reports of a special alternative splicing event, which inserts or deletes a single NAG triplet (N stands for A,C,G, or T) in mRNAs [186, 187, 188, 189]. Furthermore, during transcript analysis of human disease genes, additional cases were observed by the genome analysis group of Matthias Platzer and one cDNA based study [190] suggested that these events are not rare. Such splice events happen when one of the two splice acceptor AGs contained in the sequence NAGNAG is chosen by the spliceosome. In the following, we term these sequence motifs NAGNAG or tandem acceptors.

The insertion/deletion (indel) of a NAG triplet in the coding sequence (CDS) probably has only a subtle effect since the reading frame is not changed and both the transcript and protein isoforms are highly similar. It seems that only few studies have thoroughly investigated such splice events. The fact that such small splice events are difficult to observe experimentally (for example, a 3 nt difference between two bands is barely visible on an agarose gel) contributes to this. Furthermore, most bioinformatics studies performing EST-to-genome or EST-to-EST alignments apply threshold values that do not allow to detect these splice events. Consequently, little is known about their frequency in

eukaryotic genomes, their regulation, and their splice mechanism. Furthermore, it is of interest to find out if these events can have functional consequences for the proteins and what these consequences are. However, as indicated by the question mark in the section heading, this is particularly difficult to answer in a larger scale since the putative effect of a triplet indel in the coding sequence is by far not as obvious as the impact of most larger splice events.

Here, we address these open points and try to shed light on this largely unstudied research area. Of course, the elucidation of functional effects of subtle splice events requires detailed experimental investigation of several cases, which is infeasible for us. However, genome-wide computational studies can be used to address the following questions:

- Are these splice events overrepresented or underrepresented in certain genes, proteins, or protein regions?
- Is there evidence for specific selection pressures that act on these splice events?
- Is there evidence for a regulation of these splice events, for example in a tissue-specific manner?
- Are these events evolutionary conserved?

Thus, bioinformatics can provide a global view of these subtle splice events and can help to guide experimental efforts that clearly have to complement and prove the computational findings.

We start with analyzing NAGNAG acceptors, proceed with investigating genetic variations at these sites, and extend our studies to tandem donor sites. In collaboration with Matthias Platzer´s group, we tested our computational results experimentally.

4.2 Widespread occurrence of alternative splicing at NAGNAG acceptors

4.2.1 Genomic view of NAGNAG acceptors

We started our analysis with a screen of the human genome for acceptor sites with a NAGNAG pattern. We used the RefSeq transcript annotations (release October 2003) in the UCSC Genome Browser as an extensive collection of human transcripts. To assure a high quality data set, we discarded all transcripts with an erroneous open reading frame or with ambiguous characters in their sequence. The exon-intron structures of these transcripts were taken from the RefSeq to genome alignments.

This RefSeq scan revealed for 5% (8,105 of 152,288) of the human acceptors a NAGNAG motif. According to the CDS annotation of the RefSeq transcripts, 627 and 152 of these belong to introns that are exclusively located in the 5' and 3' UTR, respectively. As our prime interest was the impact of alternative splicing at tandem acceptors for the proteome diversity, we considered all 7,326 remaining NAGNAG acceptors that are located

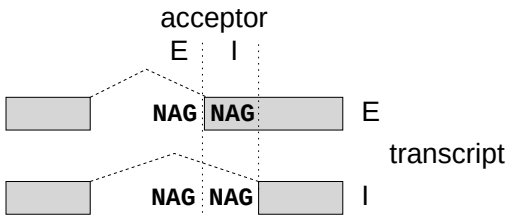

Figure 4.1: Proposed nomenclature for NAGNAG acceptors and transcripts.
E: 3' half of the NAGNAG motif becomes part of the exon; I: the NAGNAG motif is completely retained in the intron.

upstream of an exon annotated as part of the protein coding sequence. However, it is important to note that in numerous cases a UTR exon in RefSeq becomes part of the CDS in an alternative transcript, either as the result of a different mRNA maturation or as the consequence of alternative promoter usage. Thus, several of these NAGNAGs in the UTR might be located in the CDS in another transcript.

To create a senseful nomenclature, we denote the upstream AG in the NAGNAG motif as the '*E acceptor*' giving rise to the '*E transcript*', since part of the tandem will be exonic. The downstream AG is denoted as '*I acceptor*' and its usage results in an '*I transcript*', while the whole tandem is intronic (Figure 4.1).

To find out which fraction of the tandem acceptors is known to be alternatively spliced, we used 30 nt from the upstream and downstream exon to compile two search strings that represent the E and I transcript. Then, we used Blast to search dbEST. Since this alternative splice event is rather small, we applied very stringent filtering criteria (only one gap or one mismatch and exact identity in the region 27-33 of the search string, resulting in an E-value $<$ 1e-20) to exclude putative EST artifacts that mimic NAGNAG splicing. We consider a NAGNAG acceptor as alternatively spliced (also denoted as '*confirmed*') if both E and I transcript are matched by at least one EST and/or RefSeq transcript. The remaining NAGNAG acceptors are called '*unconfirmed*' with the notion that they are enriched in tandem acceptors that are not alternatively spliced. We found evidence for alternative splicing for 878 of the 7,326 (12%) NAGNAGs. Remarkably, the majority of NAGNAGs is confirmed by multiple ESTs. Thus, NAGNAG acceptors occur frequently in the human genome and a considerable fraction is known to be alternatively spliced.

4.2.2 Characteristic features of NAGNAG splicing

Next, we asked which NAGNAG acceptors are preferably alternatively spliced. We divided all NAGNAG acceptors into the 16 possible motifs. As shown in Table 4.1, 48% (533 of 1,111) of the YAGYAGs (Y stands for C or T) are confirmed. In contrast, only 1.5% (65 of 4,430) of the NAGGAG acceptors are confirmed. This agrees with the nu-

motif	observed	confirmed	
AAGAAG	54	20	37%
AAGCAG	164	69	42%
AAGGAG	199	4	2%
AAGTAG	32	15	47%
CAGAAG	888	104	12%
CAGCAG	720	343	48%
CAGGAG	2,882	41	1%
CAGTAG	96	28	29%
GAGAAG	9	1	11%
GAGCAG	227	10	4%
GAGGAG	15	2	13%
GAGTAG	45	2	4%
TAGAAG	366	59	16%
TAGCAG	258	142	55%
TAGGAG	1,334	18	1%
TAGTAG	37	20	54%
sum	7,326	878	12%

Table 4.1: Number of observed and confirmed tandem acceptors divided into the 16 motifs.

cleotide preference at position -3 for standard acceptor splice sites, where C and T are most frequent and G is very rare [182].

Next, we calculated scores measuring the strength of the E and I acceptor splice site with the Genesplicer program [191]. For confirmed tandem acceptors, the E acceptor has a significantly higher average score than the I acceptor (4.5 vs. 1.2, t-test: $P<0.0001$). This is in agreement with the EST count, where the E acceptor has on average 28 EST hits more than the I acceptor (40 vs. 12, t-test: $P<0.0001$). Thus, in general, the E acceptor is used preferentially. Moreover, this coincides with the observation that AG is the most avoided dinucleotide in the -20 to -2 region of an acceptor [23]. These data indicate that the common splicing machinery is operating at tandem acceptors and that the sequence context of the E and I acceptors is one of the discriminating factors in the splice site choice.

4.2.3 Effect on the proteins

The position of an intron relative to codons is referred to as the intron phase. An intron in phase 0 is located between two codons. Introns in phase 1 start after the first codon base and introns in phase 2 start after the second codon base. The consequence of a NAG indel in the CDS depends on the phase of the affected intron. The effects on the proteins are indels of a single amino acid (aa for short), the exchange of a dipeptide and an unrelated single aa, or the creation/destruction of a stop codon (Figure 4.2). Remarkably, there are eight different single aa events:

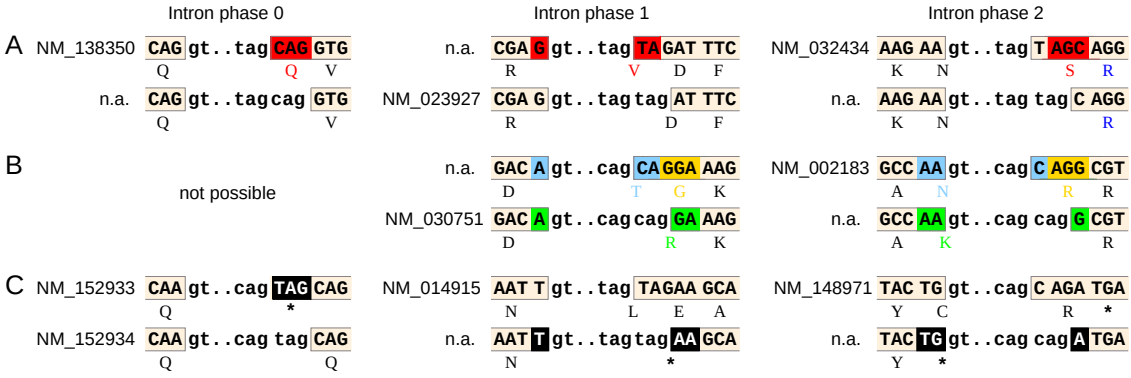

Figure 4.2: Protein variability caused by alternative splicing at tandem acceptors.
(A) Single amino acid indels. (B) Exchange of a single aa and an unrelated dipeptide (only possible for intron phase 1 and 2). (C) Indel of a stop codon. Exonic nucleotides are in upper case, intronic nucleotides are in lower case. RefSeq ID is given for annotated transcripts, n.a. means not annotated.

- Glu, Lys, and Gln in phase 0,
- Ala, Glu, Gly, and Val in phase 1,
- Arg and Ser in phase 2.

The dipeptide events are even more diverse (Figure 4.2). Thus, despite the rather simple genomic structure, NAG indels lead to a surprisingly high diversity at the protein level.

We found 51 tandem acceptors that create/destroy a stop codon and ten of these are confirmed. The transcripts with the premature stop codon can be either candidates for NMD, which results in the downregulation of the transcript and protein level or they are translated to a truncated protein. In any case, the creation/destruction of a stop codon by such a small alternative splice event can have a dramatic effect on the protein.

4.2.4 Non random distribution of NAGNAG acceptors

The intron phase distribution of all human introns is 46%, 33%, and 21% in phases 0, 1, and 2, respectively [192]. In contrast, 40% of the introns harboring confirmed tandem acceptors are in phase 0, 43% in phase 1, and 17% in phase 2, which is a significant difference (χ^2: $P<0.0001$). The striking bias towards phase 1 is a strong hint that tandem acceptors are not randomly scattered in the human genome.

The exchange of a dipeptide and an unrelated single aa is only possible for intron phase 1 and 2. Nevertheless, we found that 92% of phase 1 and 91% of phase 2 tandem acceptors result in single aa indels. To assess whether confirmed tandem acceptors are enriched in single aa indel events, we simulated the 3 nt indels using non-NAGNAG splice acceptors as a null model. We extracted 12,448 non-NAGNAG acceptors from phase 1 introns and 8,553 non-NAGNAGs from phase 2 introns. All those acceptors have

the consensus NAG. In order to obtain 'simulated' alternative transcripts, we elongated the respective exons by 3 nt upstream, thus including the downstream splice site NAG into the mature mRNA. The modified mRNAs were translated and the percentage of single aa indel events was determined. We found that null model acceptors result in significantly fewer single aa indels (87% for phase 1, 81% for phase 2, Fisher's exact test: P=0.0009 and P=0.0007, respectively). Thus, confirmed NAGNAGs in phase 1 and 2 are significantly enriched in single aa indels. We suppose that this process is driven by a higher compatibility of single aa indels with essential functions of the affected protein.

Nevertheless, an indel of a charged aa might be a dramatic event for a protein. Among the single aa indels, Arg and Lys are positively charged and Glu is negatively charged. Using the isoelectric point (pI) as a measure for the charge, we observed that the ±10-aa context of 126 confirmed Glu-indels is significantly enriched in negatively charged residues (average pI=6.22). Accordingly, the contexts of 66 Lys- and 33 Arg-indels are already basic (Lys: pI=7.47; Arg: 7.64). In contrast, charged amino acids in non-NAGNAG genes are in more neutral contexts (Glu: pI=6.93, Lys: 7.25, Arg: 7.46) and exon-exon junctions do not introduce a bias, since 32,097 junctions from non-NAGNAG genes are on average neutral (pI=7.06). For Glu, the observed pI difference is significant (t-test: P=0.0046). Thus, the indel of a positively (negatively) charged aa by a NAGNAG splice event mostly happens in a local environment that is positively (negatively) charged. This is expected to result in less dramatic effects on the proteins. These findings further support our view that tandem acceptors evolve to introduce subtle protein changes.

The average hydrophobicity (using Kyte-Doolittle values) of exon-exon junctions of confirmed NAGNAGs for the ±10-aa context is -0.73 (negative values indicate hydrophilic amino acids). In contrast, non-NAGNAG exon junctions are significantly more hydrophobic with an average of -0.36 (t-test: P<0.0001). As the protein surface is generally more polar, this indicates that the affected protein regions are often located at the surface and not in the structural core. Moreover, this might also indicate that the respective protein regions are frequently involved in protein-protein interactions, since polar residue hot spots have been observed at protein-protein binding sites [193].

4.2.5 Non random distribution of NAGNAG proteins

To test the hypothesis that genes with NAGNAG acceptors are involved in protein-protein interactions, we considered the DIP database that contains data on experimentally verified protein interactions. While 4% (548 of 14,209) of the proteins expressed from genes without NAGNAG acceptors are found in this database, this percentage is significantly higher for proteins expressed from genes with confirmed tandem acceptors (7%, 71 of 1,054; Fisher's exact test: P<0.0001).

We further analyzed the distribution of Pfam domain families between proteins expressed from genes with confirmed and without tandem acceptors. We found that the

Gene		dbEST				cDNA library				
Symbol	intron	n_E	n_I	f_E	L	name	n_E	n_I	P	tissue
MRPS11	1	70	50	0.58	1	S2SNU668s1*	0	12	0.0003	stomach
FKBP8	4	147	81	0.64	4	MGC-99	6	17	0.0018	lymph
C1orf77	3	44	61	0.42	1	L4SNU368s1*	0	12	0.0163	liver
C1orf144	1	46	49	0.48	1	MGC-109	10	1	0.0442	ovary
CAP1	2	344	55	0.86	4	MGC-20	7	7	0.0126	skin

Table 4.2: EST mining for tissue specificity of alternative splicing at NAGNAG acceptors.
n_E: number of ESTs confirming the E acceptor.
n_I: number of ESTs confirming the I acceptor.
f_E: frequency of dbEST hits for the E acceptor, $f_E = n_E/(n_E + n_I)$.
L: number of cDNA libraries matching the criteria of at least eleven ESTs hits for the NAGNAG.
P: P-value for obtaining an equally or more extreme number of E and I matching ESTs was determined by using the binomial distribution: $\binom{n_E+n_I}{n_E} f_E^{n_E}(1-f_E)^{n_I}$ where n_E and n_I are the number of ESTs for the library. The P-value was corrected for multiple testing by multiplication with the number of examined cDNA libraries (11).
*: annotated in dbEST as normalized cDNA libraries.

distribution of certain Pfam domain families is significantly biased towards either of these protein classes. The Pfam PF00001 "7 transmembrane receptor (rhodopsin family)" is completely missing in genes with confirmed tandem acceptors (Fisher's exact test: P=0.0016, corrected for multiple testing), while PF00076 "RNA recognition motif" is preferably coded by these genes (P=0.0022). We conclude from these results that tandem acceptors prevalently occur in genes coding for proteins that interact with other macromolecules.

4.2.6 Tissue-specific regulation

Alternative splicing is often controlled in a tissue or developmental stage specific manner [13]. Tissue or cell specificity of alternative splicing is usually taken as evidence that these events are biologically important.

To find out if alternative splicing at NAGNAG acceptors can be tissue-specific, we first carried out an *in silico* EST mining on human dbEST. From all confirmed tandem acceptors, we selected those having at least 44 EST matches for both the E and I acceptor (15 NAGNAGs). We defined the overall percentage of E acceptor usage by the percentage of EST that match the E acceptor in the entire dbEST. Then, we extracted information about the cDNA libraries (the tissue or cell line sources from which the ESTs were sampled) from the dbEST annotation. For all cDNA libraries that match a NAGNAG with at least eleven EST hits, we determined the library specific number of E and I matching ESTs. For five of the 15 NAGNAGs, we found one cDNA library with a number of E and I matching ESTs that is significantly different from the expectation (Table 4.2).

	ITGAM			*SMARCA4*			*BTNL2*		
tissue	E	I	f_E	E	I	f_E	E	I	f_E
Leucocytes	164	85	0.66	51	21	0.71	75	0	1.00
Liver	59	25	0.70	$+^a$	$+^a$		31	0	1.00
Pancreas	23	27	0.46	$+^a$	$+^a$		8	0	1.00
Brain	45	28	0.62	$+^a$	$+^a$		41	0	1.00
Small intestine	0	59	0	-	$+^b$		0	24	0
LCL0844	16	0	1.00	n.d.	n.d.		n.d.	n.d.	

Table 4.3: Quantification of tissue-specific expression of E and I transcripts by resequencing of RT-PCR subclones.
f_E: fraction of E transcripts.
[a]: no subcloning, direct sequencing revealed pattern E+I.
[b]: no subcloning, direct sequencing revealed pattern I.
LCL: lymphoblastoid cell line.
n.d.: not determined.

To verify these *in silico* findings, we experimentally investigated the tandem acceptors for the genes *ITGAM* (intron 13), *SMARCA4* (intron 30), and *BTNL2* (intron 1) in different tissues (Table 4.3). Although *ITGAM* and *SMARCA4* produce E and I transcripts in several tissues, we found exclusively the I variant in small intestine (Figure 4.3). For *BTNL2*, we also observed only I transcripts in small intestine, while all other tissues express only E transcripts. Thus, we conclude that splicing at tandem acceptors can be tissue-specific.

4.2.7 NAGNAG acceptors in other species

To find out if tandem acceptors occur in other genomes, we first searched the literature and found reports of alternative NAGNAG splicing in rat [194, 195], mouse and chicken [196], rabbit [197], ruminants [198], and tomato [199]. To investigate the frequency of NAGNAG acceptors in other genomes, we considered the RefSeq and EST databases for mouse (*Mus musculus*), fruitfly (*Drosophila melanogaster*), and nematode (*Caenorhabditis elegans*). NAGNAG acceptors occur frequently in these species and many of them have EST evidence for alternative splicing (Table 4.4). Remarkably, *C. elegans* has a higher number of RefSeqs and ESTs as *D. melanogaster*, but only a very low fraction of confirmed tandem acceptors. This may reflect unique features of the splicing machinery in *C. elegans* [200], whose introns typically lack both branch point and polypyrimidine tract consensus sequences, preventing an extensive utilization of tandem acceptors in this species.

4.2.8 Conservation of NAGNAG acceptors in mouse

We further asked if purifying selection is acting on NAGNAG acceptors. We constructed a large set of 77,414 orthologous acceptor pairs from human and mouse. In this set, we

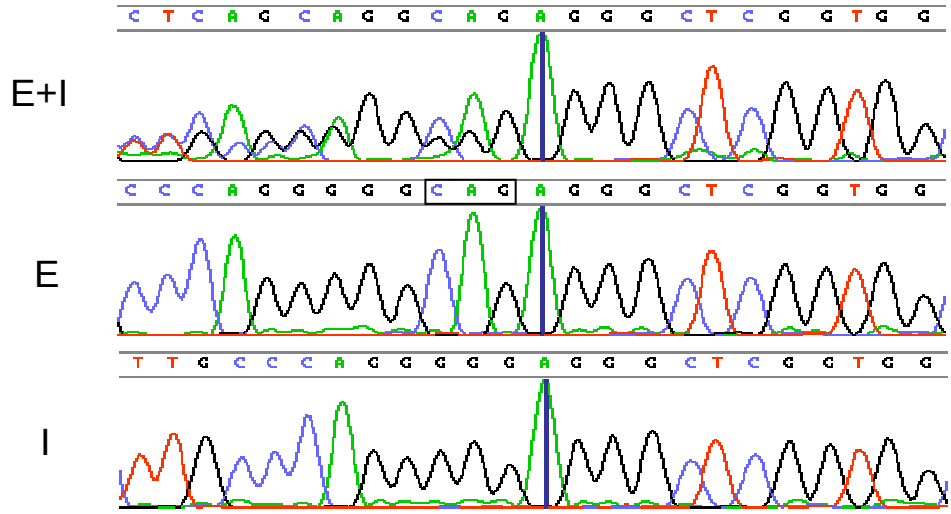

Figure 4.3: Tissue-specific expression of E and I transcripts of *ITGAM*.
The figure shows partial electropherograms from direct DNA sequencing of RT-PCR products for the exon 12 - exon 13 junction (reverse sequencing direction).
(E+I) Simultaneous expression of E and I transcripts results in a superposition of two sequences after crossing the exon boundary (vertical line). (E) Expression of only the E transcript gives clear sequencing results including the E acceptor NAG (boxed). (I) Expression of only the I transcript gives clear sequencing results lacking the NAG.

identified 3,861 human tandem acceptors. Of these, the NAGNAG motif is conserved in mouse in 2,806 (73%) cases. This high conservation rate is consistent with the recent observation that NAGNAG acceptors represent 45% of all human-mouse conserved alternative 3' splice sites [75].

Furthermore, we used the set of 77,414 orthologous acceptor pairs to determine the local base-specific sequence conservation. For all human NAGNAG acceptors, we determined the fraction of identical bases in the orthologous mouse sequences. According to the annotated acceptor, the following alignment positions (indicated by a '?') were studied:

	intronic	exonic
H. sapiens	`nagnagNNN`	`nnnnagNAG`
M. musculus	`n??nagNNN`	`nnnnagN??`

As a control, we analyzed the local base conservation for A in AH and G in BG dinucleotides (H stands for A,C,T; B stands for C,G,T) in non-NAGNAG acceptor pairs:

	intronic A	intronic G	exonic A	exonic G
H. sapiens	`nahnagNNN`	`nbgnagNNN`	`nnnnagNAH`	`nnnnagNBG`
M. musculus	`n?nnagNNN`	`nn?nagNNN`	`nnnnagN?N`	`nnnnagNN?`

Intronic extra AGs of NAGNAG acceptors show higher base identity rates between orthologs than the control cases (72% vs. 59%, Fisher's exact test: $P<0.0001$ for A; 65%

				acceptors				
					NAGNAG			
species	RefSeqs	ESTs	$\frac{\text{ESTs}}{\text{RefSeqs}}$	all	observed		confirmed	
H. sapiens	20,213	5,483,952	271	152,288	7,326	4.8%[a]	878	12.0%[b]
M. musculus	16,960	4,056,273	239	127,954	5,100	4.0%	629	12.3%
D. melanogaster	18,960	274,367	14	44,026	1,484	3.4%	97	6.50%
C. elegans	23,461	298,805	13	101,562	4,487	4.4%	30	0.7%

Table 4.4: Observed and confirmed tandem acceptors in human, mouse, fly, and worm.
[a] observed vs. all.
[b] confirmed vs. observed.

vs. 55%, P=0.0004 for G). Moreover, the identity rate of the AG dinucleotides is higher than the expected value from the combination of the base identity frequencies (observed 52%, expected $72\% \cdot 65\% = 47\%$). In contrast, the nucleotides of exonic extra AGs are slightly but significantly less conserved than bases in control acceptors (89% vs. 91%, Fisher's exact test: P=0.003 for A; 84% vs. 87%, P<0.0001 for G). Here, the AG dinucleotide has a conservation rate as expected from combinatorics (observed 74.6%, expected 74.8%). This may be explained by much stronger coding constraints that superimpose the exonic alternative splice site. Furthermore, many exonic I acceptors are GAG (Table 4.1), which are unlikely to be used as an acceptor. We conclude that a subset of tandem acceptors is conserved between human and mouse.

4.3 SNPs in tandem acceptors influence alternative splicing

In our further analysis, we investigated genetic variation of NAGNAG acceptors in the human population. In particular, we focused on single nucleotide polymorphisms (SNPs), since they make up more than 90% of the human sequence variations and more than eight million are listed in dbSNP [201] (release January 2005). This analysis seems to be promising since SNPs affecting NAGNAG acceptors

- may influence alternative splicing,
- may contribute to human diseases,
- and represent 'natural knock-out experiments', which allow us to study the NAGNAG splicing mechanism.

4.3.1 Genome-wide screen for polymorphic NAGNAG acceptors

We started with a genomic screen to identify all known human SNPs that affect the NAGNAG motif of a tandem acceptor. First, we extracted all annotated SNPs from the UCSC Genome Browser that are located within the last six nucleotides of an intron or within the first three nucleotides of an exon, given intron-exon boundaries from RefSeq

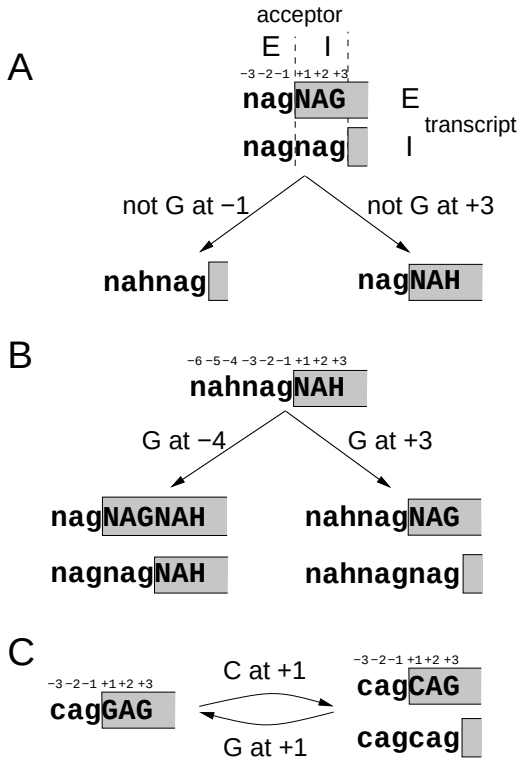

Figure 4.4: Schematic illustration how SNPs affect splicing at NAGNAG acceptors.
(A) SNP alleles at position -2, -1, +2, or +3 of a NAGNAG acceptor destroy this motif by affecting the E (left) or I acceptor (right), thus preventing alternative splicing. (B) SNP alleles at intron positions -5, -4 can create a novel E acceptor (left) and at exon positions +2, +3 a novel I acceptor (right), thus yielding a NAGNAG motif. Acceptors at these alleles may allow alternative splicing as indicated by the two transcripts (E transcript above, I transcript below). (C) SNP alleles at position -3 or +1 can convert a NAGNAG acceptor that allows alternative splicing (right) to a NAGNAG that only allows the expression of one transcript (left), or vice versa. Positions refer to a standard intron-exon boundary. H stands for A, C, or T. Upper and lower case letters indicate exonic and intronic nucleotides, respectively. Exonic nucleotides are boxed. For simplicity, only variations at the G of acceptor AGs are shown in (A) and (B).

transcripts. Then, we selected the SNPs that affect a NAGNAG acceptor. With respect to the human reference genome sequence (which represents one allele), the other SNP allele can create or destroy a NAGNAG acceptor by affecting one of both AGs (Figure 4.4A and B). Since the nucleotide upstream of a standard acceptor AG is usually C or T [182] and a change at this position is likely to alter alternative splicing at a tandem acceptor, we also considered SNPs at the N positions in an existing tandem (Figure 4.4C). We found a total of 121 SNPs affecting NAGNAG acceptors.

To check if these polymorphic NAGNAGs are alternatively spliced, we searched dbEST for the existence of E and I transcripts. We found EST evidence for 19 of these 121 (16%) tandem acceptors. However, this percentage must be considered as a lower

	intron phase					
	0		1		2	
confirmed NAGNAGs[a]	349	39.8%	379	43.2%	150	17.0%
plausible NAGNAGs[b]	1,111	42.5%	1,099	42.0%	405	15.5%
implausible NAGNAGs[b]	2,568	54.5%	1,466	31.1%	677	14.4%
all introns [c]		46%		33%		21%

Table 4.5: Phase distribution of human introns and NAGNAG acceptors.
Only NAGNAGs that are located upstream of a coding exon are considered.
[a] the 878 confirmed NAGNAG acceptors (see section 4.2.1).
[b] the 7,236 CDS NAGNAG acceptors (see section 4.2.1).
[c] genome-wide frequencies [192].

bound, since in addition to the general limitations of an EST based evaluation of alternative splicing (mainly insufficient EST coverage, see section 2.1), the allele frequencies of the NAGNAG alleles as well as population biases in EST sampling introduce further constrictions. Thus, we believe that more polymorphic NAGNAGs are alternatively spliced than the current data suggest. Interestingly, for nine of these 19 confirmed polymorphic NAGNAGs the human reference genome sequence is identical with the non-NAGNAG allele, thus the confirmed NAGNAG acceptor sequence is not 'visible' in genome browsers.

4.3.2 Extracting splicing relevant SNPs

Since more alternatively spliced polymorphic NAGNAGs are probably contained in the total set, we searched for a way to subdivide all 121 SNPs into those that are likely and unlikely to affect alternative splicing, respectively. Considering the NAGNAG motif, 18 of the 19 (95%) confirmed tandem acceptors match the consensus HAGHAG (H stands for A, C, or T). Thus, from 68 polymorphic HAGHAGs, 18 (26%) are EST confirmed, whereas from the 53 acceptors carrying G at one or both variable positions of the NAGNAG motif only 1.9% (one of 53) are EST supported. This is in line with our genome-wide observation, where 30.6% of the HAGHAGs but only 1.7% of the remaining NAGNAGs are confirmed (Table 4.1).

Based on these differences in the degree of confirmation by EST data, we propose to subdivide all tandem acceptors into

- '*plausible*' (HAGHAG),
- and '*implausible*' (GAGHAG, HAGGAG, GAGGAG) acceptors.

Here, plausible has the meaning 'likely to be alternatively spliced', whereas implausible means 'unlikely to be alternatively spliced'. Further support for this classification comes from the genome-wide observation that all plausible NAGNAGs have the same bias towards intron phase 1 as EST confirmed NAGNAGs (section 4.2.4), while the introns with implausible tandem acceptors are not biased towards phase 1 (Table 4.5).

dbSNP ID	gene symbol	genotypes					
		homozygous NAGNAG		heterozygous		homozygous non-NAGNAG	
		genomic[a]	cDNA[b]	genomic	cDNA	genomic	cDNA
rs2245425	*TOR1AIP1*[c]	3	E+I	6	E+I	2	I
rs2275992	*ZFP91*[c]	1	E+I	7	E+I	4	E
rs1558876	*KIAA1001*	0	-	6	E+I	6	E
rs2290647	*KIAA1533*	0	-	4	E+I	8	E

Table 4.6: Correlation between acceptor genotypes and the appearance of E and I transcripts.
[a] number of probands with the respective genotype.
[b] E+I: presence of both E and I transcripts; E: only E transcripts; I: only I transcripts.
[c] see also Figure 4.5.

Applying this classification to the 121 SNPs, 68 (56%) SNPs affect a plausible NAGNAG. However, four of those convert a plausible into another plausible NAGNAG, which has presumably no drastic consequence for NAGNAG splicing, even though we cannot exclude the possibility of changes in the ratio of E and I transcripts or changes in tissue specificity. As our primary goal is to extract SNPs that affect alternative NAGNAG splicing, we consider the remaining 64 SNPs as relevant for NAGNAG splicing.

4.3.3 The NAGNAG motif is necessary and sufficient for alternative splicing

SNPs that lead to NAGNAG and non-NAGNAG acceptor alleles represent 'knock-out experiments made by nature'. We took this opportunity to prove the assumed correlation between NAGNAG acceptor genotypes and the appearance of E and I transcripts. Such a study seemed reasonable since so far it has been performed in artificial splicing systems only [202]. We selected four SNPs with a minor allele frequency of greater than 0.2 that affect EST confirmed HAGHAG acceptors for genotyping and detection of transcript forms. We consistently observed E and I transcripts in cells with at least one HAGHAG allele, while cells that do not have a HAGHAG acceptor allele produced only one transcript (Table 4.6). This strict correlation between NAGNAG alleles and alternative splicing is illustrated for *ZFP91* and *TOR1AIP1* in Figure 4.5. These results confirm that a NAGNAG acceptor motif is necessary for this type of alternative splicing.

An even more challenging question is whether the NAGNAG motif is also sufficient for alternative splicing. This question can be addressed by investigating NAGNAG acceptors that have been created in recent human evolution. With regard to the human reference genome sequence, in 36 of 64 cases (56%) a novel NAGNAG is created, in 18 cases (28%) a known NAGNAG is destroyed by affecting an AG, and ten (16%) N positions are changed. Since the appearance of a SNP allele in the human genome sequence does not reflect its evolutionary history, the best reference for the question of gain vs. loss of NAGNAG

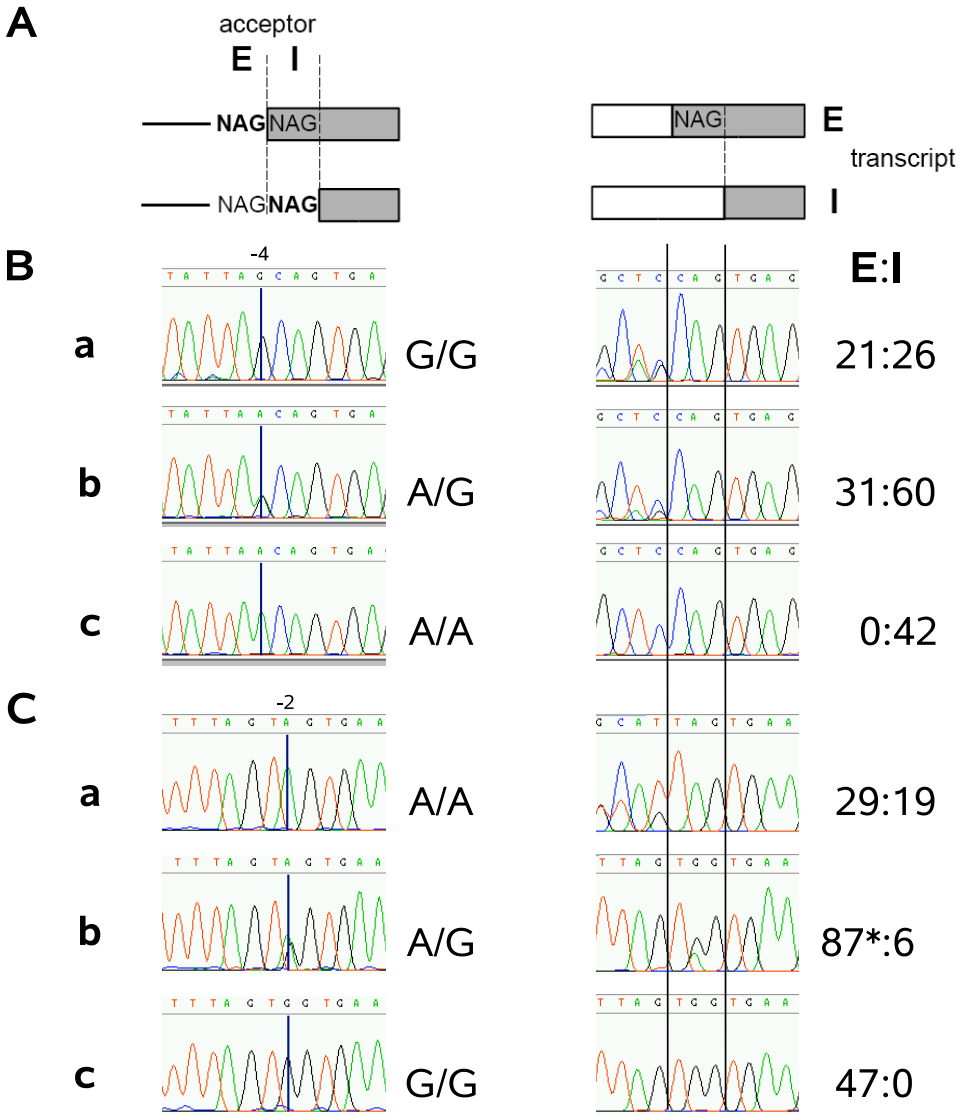

Figure 4.5: SNPs that affect plausible NAGNAG acceptors as 'natural knock-out experiments'.
(A) Representation of a NAGNAG acceptor at the genome level (left) and representation of the splice event at the transcript level (right).
(B) rs2245425 affecting the E acceptor of intron 2 of *TOR1AIP1* leads to the exclusive expression of the I transcript from the A allele (NAGNAG position -4).
(C) rs2275992 affecting the I acceptor of intron 4 of *ZFP91* leads to the exclusive expression of the E transcript from the G allele (position -2).
a: homozygous NAGNAG allele;
b: heterozygous;
c: homozygous non-NAGNAG allele;
left: genomic with genotypes;
right: cDNA with E:I transcript ratio determined by counting subcloned and sequenced RT-PCR fragments;
*: E transcripts can be assigned to the SNP alleles in the I acceptor (A=15, G=72).

acceptors is the chimpanzee genome sequence. Comparing the sequence context of the 64 plausible NAGNAG affecting SNPs, for 61 (95%) the orthologous chimpanzee nucleotide is identical to one of both human alleles, which we therefore consider as the ancestral allele [203]. In 43 cases the plausible NAGNAG is gained (non-ancestral) and in 18 cases it is lost (ancestral).

We found EST evidence for alternative splicing for seven of the 43 (16%) non-ancestral NAGNAGs, indicating that the non-ancestral SNP alleles enable alternative NAGNAG splicing. To provide further experimental support, we selected two non-ancestral plausible NAGNAGs without EST evidence. In individuals that are heterozygous or homozygous for the NAGNAG allele of the SNP rs5248, we observed the expression of E and I transcripts in the ratios 4:14 and 11:7, respectively. In case of rs17105087, we were unable to identify the non-ancestral allele in our Caucasian population sample. By analyzing the human-chimpanzee genomic sequence context of the seven EST confirmed non-ancestral NAGNAGs, we found three cases where both genomes are identical in a long range (rs2287800 identical nucleotides -140 to +123 from the SNP position, rs3765018 -130/+95 nt, and rs2290647 -105/+70 nt). Since most splicing enhancers function only in a distance of less than 100 nt from the affected splice site [31], these findings indicate that NAGNAG motifs are sufficient for alternative splicing in the context of a previously non-NAGNAG acceptor.

4.3.4 Evolutionary aspects of SNPs in NAGNAG acceptors

We observed striking differences in the numbers of SNPs that affect the AG of the E and I acceptor in ancestral plausible and implausible NAGNAGs, respectively. For the 16 ancestral HAGHAGs, the E acceptor is affected in eleven and the I acceptor in five cases. In contrast, for 22 implausible HAGGAGs (one ancestral GAGGAG and two GAGHAGs were not considered since the number of cases is too small), the E acceptor is affected in five and the I acceptor in 17 cases (Fisher's exact test: P=0.008). Interestingly, we observed the same trend by comparing all 138 human NAGNAGs that are not conserved in the chimpanzee genome (one GAGGAG and seven GAGHAGs were omitted). The I acceptors of 79 HAGHAGs are affected in 44 cases (56%), while the GAG of 59 HAGGAGs is affected in 49 cases (83%, Fisher's exact test: P=0.0009).

Since tandem acceptors are non-randomly distributed in the human genome with a bias towards intron phase 1 and towards single amino acid indels in phase 1 and 2, we questioned whether the non-ancestral plausible NAGNAGs are also biased. Indeed, these NAGNAGs show the same bias towards intron phase 1 and they also have a strong tendency to result in single aa indels (Table 4.7). Thus, the process of establishing SNPs that are relevant for alternative NAGNAG splicing in the human population seems to be a non-random process, which is subjected to the same evolutionary forces as the maintenance of the tandem acceptors themselves.

	intron phase						single aa events	
	0		1		2		phases 1 and 2	
non-ancestral NAGNAG alleles [a]	12	31.6%	16	42.1%	10	26.3%	24	92.3%
non-polymorphic confirmed NAGNAGs [b]	349	39.8%	379	43.2%	150	17.0%	487	92.1%

Table 4.7: Intron phase distribution and single aa events of non-ancestral plausible NAGNAG acceptors.
Only NAGNAGs that are located upstream of a coding exon are considered.
[a] plausible polymorphic NAGNAGs where the chimpanzee acceptor has no NAGNAG.
[b] EST confirmed NAGNAGs.

4.3.5 Potential disease relevance of NAGNAG SNPs

Alternative splicing at tandem acceptors can result in the gain or loss of a premature stop codon in the mRNA. Among SNPs affecting plausible NAGNAGs, the G allele of the SNP rs9644946 changes the acceptor context of intron 7 of *GOLGA1* from AAATAG to AAGTAG. Since intron 7 resides in phase 0, an in-frame TAG insertion would be the consequence if the novel E acceptor is used. Interestingly, the gene codes for an autoantigen associated with Sjogren's syndrome (OMIM 270150). As the E acceptor is preferred in alternative NAGNAG splicing (section 4.2.2), the novel AAG acceptor is likely to be functional. The resulting E transcript is a candidate for nonsense-mediated mRNA decay. Thus, the AAGTAG allele would result in a lower protein expression. Alternatively, it is possible that the mRNA containing the premature stop codon escapes degradation and that the truncated protein exhibits autoantigenic properties. It remains to be elucidated in populations with a sufficiently high allele frequency whether alternative splicing at the AAGTAG acceptor contributes to the disease.

A literature search revealed an example that demonstrated the disease relevance of a NAGNAG SNP for the *ABCA4* gene [204]. In this work, Maugeri et al. describe a NAGNAG mutation (2588G→C, changing the acceptor site TAGGAG→TAGCAG) that has a much higher frequency in patients with Stargardt disease 1 (STGD1; OMIM 248200). This mutation is assumed to cause STGD1 in combination with a severe *ABCA4* mutation. By experimental analysis of the splice patterns of two STGD1 patients carrying the mutation and one control individual, they found that only the alleles with the plausible tandem acceptor (TAGCAG) produce two splice forms. Our study exactly predicts this mutation outcome.

4.4 Tandem acceptors in U12 introns

The great majority of introns has GT-AG or GC-AG termini and is spliced by the major U2-dependent spliceosome that requires the snRNPs U1, U2, U4, U5, and U6 for

splicing. However, another class of introns often having AT-AC termini exist in higher eukaryotes [205, 206]. These introns are spliced by the minor U12-dependent spliceosome that requires the snRNPs U11, U12, U4atac, and U6atac, while the U5 snRNP is shared by both spliceosomes. U12-dependent introns (U12 introns for short) exhibit a nearly invariant donor splice site with the sequence (A/G)TATCCTTT. Furthermore, the branch point sequence TTCCTTAAC is very strict and the bulged adenosine (underlined) is mostly located 10-26 nt upstream of the acceptor [207]. Remarkably, the acceptor dinucleotide in U12 introns is highly diverse including AC, AG, AA, CG, and TT [208, 207]. Despite these U12 introns are very rare, they occur in numerous species and their splice sites are often highly conserved [206]. Furthermore, mutations in these introns have been linked to human diseases [209].

There is evidence that the fidelity in the acceptor recognition is lower for U12 compared to U2 introns. Mutations in the donor site are reported to activate other acceptor sites in the vicinity of the annotated acceptor [208, 209] and small acceptor variations are also reported in an EST based study [210].

Motivated by the widespread alternative splicing at NAGNAG acceptors (which almost exclusively occur in U2 introns), we investigated whether acceptor sites in U12 introns exhibit similar three nucleotide splice variations. Based on the RefSeq transcript annotations, we scanned all human introns for donor sites with the pattern (A/G)TATCC. For these introns, we searched dbEST for evidence that putative acceptor sites ± 3 nt from the annotated acceptor are used in the splicing process. Due to the diversity of the U12 acceptor, we did not restrict the search to certain acceptor dinucleotides. From the 896 U12 introns, 13 (1.5%) exhibit three nucleotide variations at the acceptor site (Table 4.8). A U12 branch point sequence is found for all 13 introns, suggesting that they are real U12 introns. For none of these 13 acceptors we found SNPs in the acceptor vicinity, excluding the possibility of allele-specific splicing. In two cases (*GBL*, *GMFB*) the alternative acceptor is conserved and confirmed in mouse, suggesting that these splice events are real and not EST artifacts.

Consistent with previous studies, the alternative acceptor dinucleotides are highly variable, although AT and AG are preferred (Table 4.8). Only one of these 13 U12 introns has a NAGNAG acceptor (*GBL*). Finally, it should be noted that our scan specifically searched for ± 3 nt splice variants. However, in contrast to U2 introns, U12 introns frequently produce small out-of-frame splice variants by choosing other acceptor dinucleotides located in a range of 1 to 6 nt from the 'normal' acceptor, which is likely to be the consequence of a higher error rate in acceptor recognition [210]. Whether the 13 identified ± 3 nt splice variants for U12 introns have a biological function or represent splicing errors, remains unclear.

gene symbol	RefSeq	intron	splice sites [a]	acceptor pattern [b]	E:I [c]	ratio [d]
XPO4	NM_022459	19	AT-AG	cag|AAT	9:2	18.2%
SLC12A6	NM_005135	11	GT-AG	tggtag|	1:14	6.7%
GMFB	NM_004124	4	AT-AC	aaaaac|	4:75	5.1%
C10ORF61	NM_015631	8	GT-GG	tgg|ATG	23:1	4.2%
POLR2E	NM_002695	5	GT-AG	tcgcag|	15:348	4.1%
GBL	NM_022372	2	GT-AG	cag|CAG	208:8	3.7%
WDR10	NM_052985	25	GT-AG	cggcag|	1:33	2.9%
STX6	NM_005819	2	AT-AC	acctac|	1:35	2.8%
E2F5	NM_001951	3	AT-AC	cagtac|	1:38	2.6%
KCMF1	NM_020122	4	AT-AT	tat|GAT	60:1	1.6%
TUSC3	NM_006765	7	GT-AG	ctgcag|	1:104	1.0%
SMS	NM_004595	6	AT-AC	gcggac|	1:115	0.9%
RBM8A	NM_005105	4	GT-AG	cag|GGG	298:1	0.3%

Table 4.8: Three nucleotide variations at acceptor sites of U12 introns.
[a] donor-acceptor splice site dinucleotides of the annotated intron.
[b] | denotes the annotated intron-exon boundary. Upper and lower case letters indicate exonic and intronic nucleotides, respectively.
[c] number of ESTs for the E and I transcript.
[d] EST ratio of the minor splice form.

4.5 Alternative splicing at tandem donors

Having observed that NAGNAG acceptors are frequently alternatively spliced, we questioned whether donor splice sites with the motif GTNGTN also allow the expression of two splice forms differing only by a GTN triplet (note that only U2 intron donor sites can have a GTNGTN motif). This is of interest since the recognition of donor and acceptor splice sites is entirely different. While the acceptor AG and its preceding polypyrimidine tract is recognized by the U2AF heterodimer [211], the donor splice site has an extended consensus sequence CAG|GTRAGT (| is the exon-intron boundary, R stands for A or G) that is recognized by base pairing with the 5' end (nucleotides 2-10) of the U1 snRNA [212]. Remarkably, two donor sites that are only 3 nt apart would result in overlapping U1 snRNA binding sites and the GTNGTN motif differs from the donor consensus sequence at the two conserved positions +4 and +5.

4.5.1 Genomic view of tandem donors

Consistent with the proposed nomenclature for NAGNAG acceptors, we termed the upstream donor '*i donor*' that renders the complete GTNGTN motif to be intronic. Likewise, the downstream donor is called the '*e donor*' since the upstream GTN becomes exonic (Figure 4.6A). Note that inversely to NAGNAG acceptors, the e donor is located downstream to the i donor. We use lower case letters for the two donor sites and upper case letters for the two acceptor sites to distinguish between the transcripts that

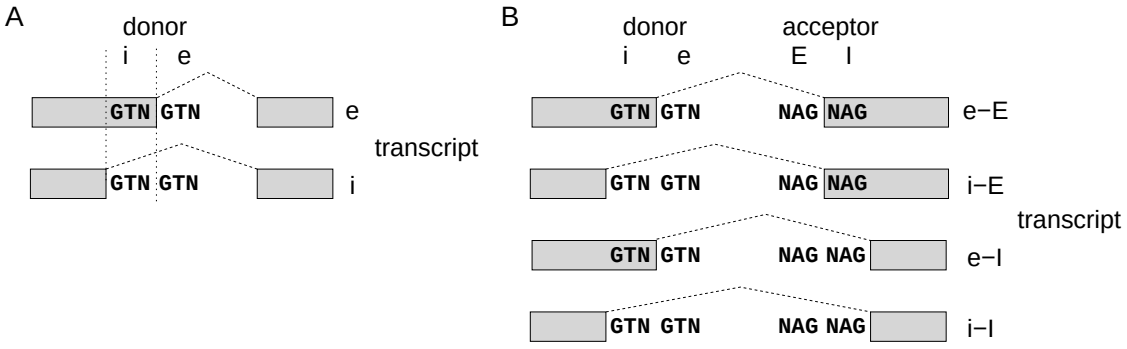

Figure 4.6: Nomenclature for tandem donor sites and transcripts.
(A) Splicing at the downstream e donor makes the upstream GTN exonic, while splicing at the upstream i donor makes the complete GTNGTN motif intronic. (B) Simultaneous usage of e or i donor and E or I acceptor results in four different transcripts (e-E, i-E, e-I, and i-I transcript).

arise by alternative splicing at tandem donors or acceptors and between combinations of alternative donor and acceptor usage (Figure 4.6B, see also section 4.7.2).

The search for tandem donors as well as the Blast against dbEST was done similar to the NAGNAG analysis described in section 4.2.1. Again, our genome-wide analysis is based on the RefSeq transcript annotations (November 2005). In agreement with the donor consensus sequence that shows no GT dinucleotide 3 nt up- or downstream of the donor site, we only found 4,152 (2.5%) tandem donors from the total of 165,295 annotated donor sites. By searching dbEST and the human mRNAs from GenBank, we identified evidences for alternative splicing at 81 (2% of 4,152) tandem donors. We term these tandem donors '*confirmed*', whereas the remaining 4,071 donors are called '*unconfirmed*'.

Further supporting the EST-derived confirmation of these alternative splice events, we performed RT-PCR in several human tissues. We selected six genes with confirmed GTNGTNs and found e and i transcripts for all six tandem donors. We detected no variation among the tissues, suggesting that these tandem donors are not regulated in a tissue-specific manner.

It has been reported that SNPs in the vicinity of donor sites lead to a shift in the splice site [213, 214]. To exclude that there is a general trend that confirmed GTNGTNs might be influenced by SNPs in their genomic flanks, thus giving rise to allele-specific splice forms [106], we selected all SNPs from dbSNP that are mapped to the 100 nucleotide context up- and downstream of these tandem donors. We found that 59% (48 of 81) of the confirmed GTNGTNs do not have an annotated SNP in this 206 nucleotide region. As a control, we randomly selected 500 unconfirmed GTNGTNs and likewise found no SNP for 59% (294 of 500), suggesting that most of the confirmed tandems are not associated with allele-specific splice forms. We also analyzed the splicing at one tandem donor (intron

21 of *STAT3*) in leucocytes of six individuals and consistently observed both transcripts. This agrees with the *in silico* finding that tandem donor splicing in general does not depend on specific genotypes and further excludes the possibility that a peculiarity of the spliceosome or its components is causal for the two splice forms.

We proceed with a characterization of these splice events. However, it should be noted that the small number of confirmed GTNGTNs does not allow to perform all the analyses done for NAGNAGs.

4.5.2 Characteristic features of confirmed GTNGTN donors

A or G is strongly preferred at intron position +3 for standard donor sites GTN, while T and C have lower frequencies [215]. We classified the confirmed GTNGTN donors according to their pattern into three groups:

- GTRGTR (R = A or G),
- GTTGTR, GTRGTT or GTTGTT,
- and GTCGTN or GTNGTC.

The GTRGTR pattern is clearly preferred as 86% (70 of 81) of the confirmed GTNGTN donors belong to this group. A smaller fraction has one or two Ts at the N-positions (eight of 81, 10%) and the third group is very rare with only three cases. These findings indicate that the common splicing machinery is operating at these sites.

Furthermore, we generated a sequence logo for the genomic context of

- confirmed tandem donors,
- unconfirmed GTNGTNs where either the e or i donor is confirmed,
- and donor sites without a GTNGTN motif (Figure 4.7).

The three nucleotides up- and downstream of confirmed tandem donors are non-randomly distributed (Figure 4.7B), consistent with the observation that both donor sites are alternatively used in the splice process. In contrast, either the upstream or downstream side of unconfirmed GTNGTNs is more randomly distributed. The higher conservation of the AG upstream of the unconfirmed GTNGTN motifs with annotated i donor indicates that the non-consensus intronic sequence is compensated by a more stringent match to the exonic part of the donor consensus sequence (compare Figure 4.7C with A).

4.5.3 Differences in U1 snRNA binding for confirmed and unconfirmed GTNGTN donors

The U1 snRNA determines the donor site by base pairing with the mRNA [212]. To define the strength of a donor site, we calculated the average free energy of the binding of the nucleotides 2-10 of U1 snRNA to the donor site [216]. In general, the e donor of confirmed GTNGTNs has a higher strength compared to the i donor (average -4.96

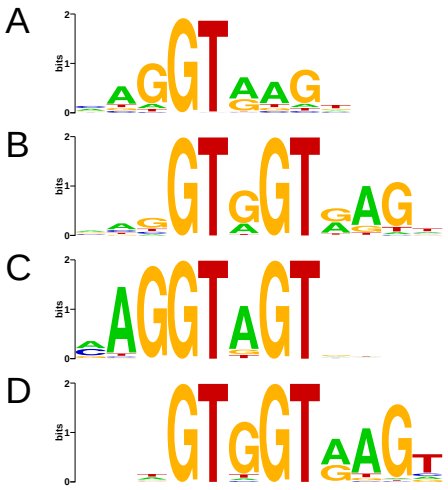

Figure 4.7: Sequence logos of the 12 nt donor context.
(A) Donors without a GTNGTN motif. (B) Confirmed GTNGTN donors. (C) unconfirmed GTNGTNs with annotated i donor. (D) unconfirmed GTNGTNs with annotated e donor. The y axis is given in bits.

vs. -3.68 kcal/mol). In agreement with that, the e donor is annotated in 73% (59 of 81 confirmed GTNGTNs) in RefSeq. Furthermore, the e donor is represented by an average of 233 ESTs, which is about tenfold higher than the average of 24 ESTs for the i donor. These findings can be explained by a stronger consensus sequence downstream of a standard GT donor compared to the three upstream positions (Figure 4.7A).

Nevertheless, there are 17 of the 81 confirmed GTNGTN tandems with more ESTs for the i donor than the e donor. Therefore, we compared the free energy values and found that 15 of these 17 cases (88%) have a lower free energy for the i donor, thus allowing a more stable U1 binding (Figure 4.8A). Likewise, 56 of the remaining 64 confirmed GTNGTNs (88%) with more ESTs for the e donor have a lower free energy for the e donor. The same trend was observed for the annotated donor of unconfirmed GTNGTNs (Figure 4.8B). In agreement with other experimental and computational studies [217, 80], the free energy of the U1 snRNA binding generally determines the donor that is used more frequently.

Since only a small fraction of all human tandem donors are confirmed, we searched for differences between confirmed and unconfirmed ones. Comparing the average free energies, we found that the e as well as the i donor of confirmed GTNGTNs is significantly stronger than the respective unannotated donor of unconfirmed GTNGTNs (Table 4.9, t-test: $P<0.00001$). In contrast, the annotated donor of unconfirmed GTNGTNs is significantly stronger than the respective donor of confirmed GTNGTNs (Table 4.9, t-test: $P<0.00001$). We repeated this analysis using the number of base pairs between donor sites and U1 snRNA [216] and the maximum entropy scores [218] to measure the

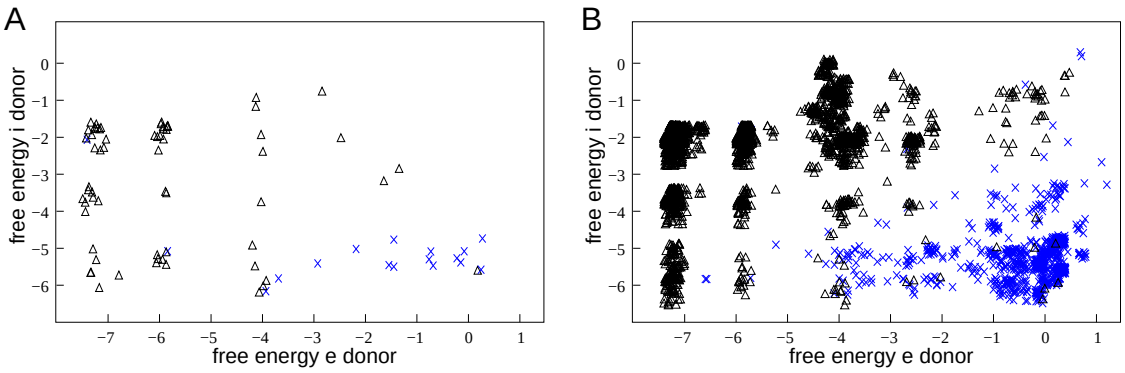

Figure 4.8: The free energy values (kcal/mol) of U1 snRNA binding to the e and i donor. (A) confirmed GTNGTNs. (B) unconfirmed GTNGTNs.
In (A) black triangles represent tandem donors with more ESTs for the e donor, blue crosses tandem donors with more ESTs for the i donor. In (B) black triangles represent annotated e donors, blue crosses annotated i donors. To better illustrate the distribution of the free energies, we added a random number between -0.1 and 0.1 to each value (necessary since many donor sites have the same 9 nt context pattern).

strength of a donor site and found consistent results (Table 4.9, t-test: all P-values $<$ 0.00001). Thus, unconfirmed tandem donors are characterized by a strong donor that successfully competes for U1 snRNA binding with the much weaker donor. The smaller difference between both donors for confirmed tandems probably allows U1 binding to both sites, leading to the observed splice variants.

We assumed that the strength of both donors might be a criterion to distinguish alternatively from non-alternatively spliced tandem donors. To test this experimentally, we selected nine unconfirmed GTNGTNs with a low free energy for both donor sites for experimental verification. For none of the nine candidates, we found evidence for alternative splicing at the tandem donor, suggesting that the majority of unconfirmed GTNGTNs is presumably not alternatively spliced. We conclude that

- stable U1 binding is necessary but not sufficient for alternative tandem donor splicing,
- the currently confirmed GTNGTNs represent a large fraction of all alternatively spliced tandem donors,
- and alternatively spliced GTNGTNs are not easily predictable.

4.5.4 Confirmed tandem donors have overrepresented sequence motifs in their intron flanks

Since the free energy of U1 binding seems not to be the only discriminative criterion, we searched for other differences between confirmed and unconfirmed GTNGTNs. The regulation of alternative splicing often involves auxiliary exonic and intronic splicing

	average					
	free energy (kcal/mol)		no. of base pairs		maximum entropy score [a]	
	i	e	i	e	i	e
unconfirmed, e annotated	-2.3	-5.93	3.87	7.31	-14.62	7.85
unconfirmed, i annotated	-5.21	-0.53	6.41	4.13	4.11	-10.95
confirmed	-3.68	-4.96	4.84	6.64	-6.47	3.89

Table 4.9: Characteristics of U1 snRNA binding to confirmed and unconfirmed GTNGTN donors.
[a] higher values indicate stronger splice sites

enhancer and silencer elements (ESE, ESS, ISE, and ISS). Previous computational studies followed by experimental verification identified 238 hexamers as ESEs [41], 2,060 octamers as ESEs and 1,019 octamers as ESSs [42], and 133 hexamers as ISE motifs in the vicinity of donor sites [43].

We used these motifs to compare their average frequency between both groups. Since most unconfirmed GTNGTNs are probably not alternatively spliced, this large group constitutes an appropriate null model. The 100 nt exonic flanks of confirmed GTNGTNs are statistically indistinguishable from unconfirmed ones when comparing the frequency of ESE and ESS motifs. However, we found a significantly higher frequency of ISE motifs in the 100 nt intron flanks for confirmed GTNGTNs (average 10 vs. 8, t-test: P=0.0174). Repeating this analysis with a shorter intronic context of 50 nt, leads to consistent results.

To find out if specific ISE hexamers are statistically overrepresented, we used a resampling strategy. We randomly sampled 10,000 sets, each comprising 81 intron flanks from unconfirmed GTNGTNs. We estimated the P-value as the fraction of random sets with a higher frequency of a given ISE hexamer compared to the observed frequency in confirmed tandem donors. CGGGGT is the only one among the 133 ISE motifs that is significantly overrepresented in the vicinity of confirmed GTNGTN donors as all 10,000 random sets have a lower frequency ($P < 1/10,000 \cdot 133 = 0.0133$ to correct for multiple testing).

To find out if other sequence motifs are overrepresented in the intron flanks of confirmed tandem donors, we repeated this procedure with tetramers. A word length of 4 nt was chosen to account for the rather small set of confirmed GTNGTNs. We expect that such motifs occur (i) at least with the expected frequency assuming random sequences and (ii) with a significant higher frequency in the flanks of confirmed GTNGTNs compared to unconfirmed GTNGTNs. There are 97 overlapping tetramers in a 100 nt sequence, thus we analyze a total of $81 \cdot 97 = 7,857$ tetramer occurrences. For complete random sequences, each tetramer should occur $7,857/256 = 30.7$ times. To fulfill (i), we considered a total of 119 tetramers that occur 30 times or more in the flanks of confirmed GTNGTNs and multiply the P-value with 119 to correct for multiple testing. Point (i)

prevents the detection of overrepresented but rare motifs that presumably do not explain why most confirmed GTNGTNs are alternatively spliced.

We found a significant overrepresentation for GGGT and CGGG (both have a higher frequency in only two random sets, $P< 3/10,000 \cdot 119 = 0.0357$). Since both GGGT and CGGG are substrings of the overrepresented ISE CGGGGT, no new sequence motifs were found. The common feature of the overrepresented sequence motifs is the G triplet. Interestingly, this motif occurs in 82 of the 133 ISEs [43] and is a known splicing enhancer [219]. Since both splice sites of confirmed GTNGTNs are weaker compared to the annotated splice site of unconfirmed ones (Table 4.9), the G triplets might simply be associated with weak splice sites. To exclude this possibility, we compared the average GGG frequency with unconfirmed GTNGTNs having a low U1 binding potential for both e and i donor (average free energy -3 kcal/mol for the e donor, -2.2 for the i donor) and still found an overrepresentation in the intron flanks of confirmed GTNGTNs (average 4.4 vs. 2.6 G triplets per intron flank). Since this triplet was found to be more frequent in shorter introns [220], we divided our confirmed and unconfirmed datasets into short and long introns using 200 nt as a cut-off. Consistently, the GGG is more frequent in the flanks of short as well as long introns with confirmed GTNGTNs (average 8.3 vs. 4.4 G triplets per short intron, average 3.4 vs. 2.7 per long intron). Thus, the occurrence of G triplets is another discriminating criterion between confirmed and unconfirmed tandem donors.

4.5.5 Effect on the proteins

72 of the 81 (89%) confirmed GTNGTNs are located downstream of a coding exon. As for NAGNAGs, the effect on the protein depends on the phase of the intron as well as the sequence of the i donor and the up-/downstream exon. Of the 72 GTNGTNs, 60 result in the following single aa events:

- Val in phase 0 (encoded by the i donor GTN),
- Gly, Arg, and Ser in phase 1,
- Trp, Cys, and Tyr in phase 2.

Apart from eight observed dipeptide events, alternative splicing at four GTNGTNs results in the indel of a stop codon. In two cases, the splice form with the premature stop codon is a clear candidate for NMD. In the other two cases, the tandem donor affects the last intron of the transcript, thus the stop codon-containing splice variant should be translated into a protein with a shortened C-terminus.

Next, we compared the frequency of single aa events in phase 1 and 2 for confirmed and, as a control, for unconfirmed GTNGTNs. While only 42% (495 of 1180) of unconfirmed tandem donors in phase 1 result in a single residue indel, this percentage is significantly higher for confirmed tandems (64%, 18 of 28, Fisher's exact test: $P = 0.02$).

species	no. of donors [a]	GTNGTN			
		observed		confirmed	
H. sapiens	165,295	4,152	2.51%[b]	81	1.95%[c]
M. musculus	125,332	3,188	2.54%	49	1.54%
R. norvegicus	53,631	1,424	2.66%	12	0.84%
G. gallus	19,793	554	2.80%	1	0.18%
D. rerio	29,091	619	2.13%	5	0.81%
D. melanogaster	40,811	1,274	3.12%	19	1.49%
C. elegans	92,938	3,195	3.44%	26	0.81%
A. thaliana	112,684	3,541	3.14%	36	1.02%

Table 4.10: GTNGTN donors in eight investigated species.
[a] total number of all unique donor sites annotated in RefSeq transcripts; for *A. thaliana* total number of unique donor sites based on the CDS feature annotation of GenBank
[b] no. of observed GTNGTNs / no. of all donors
[c] no. of confirmed GTNGTNs / no. of observed GTNGTNs

The small number of phase 2 tandems does not allow a significant result, although the same trend is visible (100%, five of five confirmed tandems; 76%, 431 of 566 unconfirmed tandems). Similar to NAGNAG acceptors, this argues for a selection pressure towards single aa indels. However, we cannot exclude the possibility that this is an indirect consequence of a sequence bias of the GTNGTN motif and its context for confirmed tandems that primarily aims at a more stable U1 snRNA binding.

4.5.6 Tandem donors in seven other species

Next, we asked whether alternative splicing at GTNGTN donors occurs in other species. Therefore, we extended our analysis to the RefSeq transcripts of mouse, fruitfly, and nematode (release November 2005). Due to an improved RefSeq annotation and an increased number of ESTs, we extend this analysis also to three other species (rat (*Rattus norvegicus*), chicken (*Gallus gallus*), and zebrafish (*Danio rerio*)).

The percentage of GTNGTN motifs in all donor sites is similar in all species and ranges from 2.1% to 3.4% (Table 4.10). Except for chicken, we found EST evidence for alternative splicing for 0.8% to 1.5% of all GTNGTN donors. Finally, we searched tandem donors in the plant *Arabidopsis thaliana* using the CDS annotation from GenBank and detected 36 confirmed GTNGTNs. Thus, all investigated species are able to produce e and i transcripts at tandem donors by alternative splicing.

4.5.7 Conservation of exonic and intronic flanks in mouse

Having observed several alternative GTNGTN splice events in human and mouse, we found conservation of the GTNGTN motif for 53 (65.4%) of the 81 human confirmed GTNGTNs. To find out whether this percentage is high or not, we counted GTNGTN conservation for the 3,909 unconfirmed tandems (162 of the 4,071 unconfirmed ones have

no orthologous locus in mouse) and found a very similar percentage of 65.5% (2,561 of 3,909). The fraction of tandem donors that have a completely identical GTNGTN pattern in mouse is also equal: 40 of 81 (49.4%) confirmed, 1,939 of 3,909 (49.6%) unconfirmed. Thus, there is no evidence for a general selection pressure to maintain a confirmed tandem donor since the divergence of the human-mouse ancestor.

However, a considerable fraction (ten of 53, 19%) of the conserved and confirmed human GTNGTNs is also confirmed in mouse. For example, the GTAGTT donor of intron 21 of human *STAT3* is conserved in the orthologous mouse gene *Stat3* and both e and i transcripts are supported by mouse ESTs. As in humans, we performed RT-PCR in mouse tissues and found experimental evidence for alternative splicing at the *Stat3* tandem donor. Interestingly, the ratio of e and i transcripts estimated by the EST counts are virtually identical: 57 of 74 (77%) human ESTs and 55 of 69 (79.7%) mouse ESTs are e transcripts. To accurately quantify the ratio of e and i transcripts in one selected tissue, we counted individual sequenced clones and found a remarkable agreement in the transcript ratio: 82.8% of the human clones indicate splicing at the e donor, which is highly similar to 85.3% in mouse. Moreover, this tandem donor is conserved in several other mammals and the e:i ratio is very similar (9:2 ESTs for rat, 12:3 for cow, 9:1 for dog). This suggests that in addition to the tandem donor putative regulatory elements are conserved.

The intronic flanks of alternative exons are significantly more conserved in mouse compared to the flanks of constitutive exons, which is presumably attributed to the force to maintain regulatory elements [76]. From genomic human-mouse alignments, we calculated a per-position identity value for the region 30 nt up- and downstream of the human GTNGTNs. For a specific position, this value is the fraction of identical nucleotides in all pairwise alignments [75, 76]. We calculated per-position identities for three groups:

(i) confirmed human tandem donors with a conserved GTNGTN motif in mouse,
(ii) the subset of (i) that is also confirmed in mouse,
(iii) and unconfirmed human tandems.

Plotting these average values, it can be seen that group (i) and in particular group (ii) have noticeably higher identities for both the exonic and intronic side compared to the control group (iii) (Figure 4.9). The exonic identities for the ten human and mouse confirmed and conserved tandem motifs exceed 90% for most positions, a feature that is also typical for alternative exons [92]. Furthermore, the GTNGTN pattern with 3 nt up- and downstream is completely identical between both species for these ten tandems and average identities of more than 80% are observed for the first 13 intronic positions.

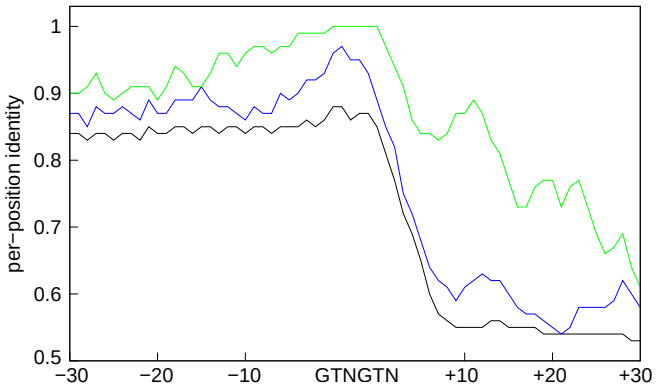

Figure 4.9: Per-position identity values in human-mouse alignments for the region 30 nt up- and downstream of the GTNGTN motif.
The black line represents unconfirmed human GTNGTN donors, the blue line confirmed human tandem donors with a conserved GTNGTN motif in mouse, and the green line conserved GTNGTNs that are confirmed in human and mouse. To avoid large variations due to low case numbers, we plotted for each position the average of this and the three positions up- and downstream.

splice donor pattern	observed		confirmed	
GTNGTN	4,152	2.51%[a]	81	1.95%[b]
GTNGCN	856	0.52%	14	1.64%
GCNGTN	3,510	2.12%	15	0.43%
GCNGCN	32	0.02%	0	0.00%
GYNGYN	8,550	5.17%	110	1.29%

Table 4.11: Human tandem donor sites divided into the four different GYNGYN patterns.
[a] percent of all 165,295 annotated donor sites
[b] no. of confirmed / no. of observed

4.5.8 Alternative splicing at GCNGTN and GTNGCN donors

Although the great majority of donor sites has a GT dinucleotide, a small fraction of 0.76% has a GC dinucleotide [9]. Thus, we questioned whether donor sites with a GCNGTN, GTNGCN, and GCNGCN motif also allow alternative splicing. Together with GTNGTNs, we call these donors GYNGYN donors (Y stands for C or T). A search in the human species revealed numerous such donors (Table 4.11). Surprisingly, we found EST evidence for alternative splicing for 15 GCNGTNs and 14 GTNGCNs (Table 4.11). No confirmed GCNGCN donor was found, presumably because this motif is very rare and the weaker GC donor requires a more stringent sequence context.

Extending the search to the seven other species, we consistently found confirmed GCNGTNs and GTNGCNs but no confirmed GCNGCN (Table 4.12). Interestingly, three of the human confirmed GCNGTNs/GTNGCNs are conserved and confirmed in other species, suggesting that the respective splice events are real.

species	no. of donors [a]	GCNGTN and GTNGCN			
		observed		confirmed	
H. sapiens	165,295	4,398	2.66%[b]	29	0.66%[c]
M. musculus	125,332	3,237	2.58%	12	0.37%
R. norvegicus	53,631	1,440	2.69%	3	0.21%
G. gallus	19,793	553	2.79%	2	0.36%
D. rerio	29,091	699	2.40%	1	0.14%
D. melanogaster	40,811	1,906	4.67%	5	0.26%
C. elegans	92,938	2,838	3.05%	1	0.04%
A. thaliana	112,684	2,091	1.86%	8	0.38%

Table 4.12: GCNGTN and GTNGCN donors in eight investigated species.
[a] total number of all unique donor sites annotated in RefSeq transcripts; for *A. thaliana* total number of unique donor sites based on the CDS feature annotation of GenBank
[b] no. of observed GCNGTNs and GTNGCNs / no. of all donors
[c] no. of confirmed GCNGTNs and GTNGCNs / no. of observed GCNGTNs and GTNGCNs

4.5.9 Comparison of GYNGYN donors and NAGNAG acceptors

To provide a complete genomic view of alternative splicing at GYNGYN donors and NAGNAG acceptors, we updated the NAGNAG analysis to the seven species having a RefSeq annotation in the UCSC Genome Browser using the same data as for GYNGYNs (Table 4.13). In general, the percentage of confirmed NAGNAGs is one order of magnitude higher compared to GYNGYN donors (compare Table 4.10 and 4.12 with 4.13). This can be explained by large differences in the mechanisms of donor and acceptor site recognition. While the acceptor AG is bound by the U2AF35 protein, the donor site is recognized by base pairing with the U1 snRNA. In contrast to the acceptor, the binding site of U1 comprises a larger range. This imposes more sequence constraints on a tandem donor site and prevents the extensive use of potential e and i donors compared to potential E and I acceptors. Apart from human and mouse, the fruitfly has a high percentage of confirmed NAGNAG sites, which is probably due to the higher percentage of tandem acceptors with the HAGHAG pattern that preferably allow alternative splicing. In contrast, a very low fraction of the NAGNAG acceptors of *C. elegans* is confirmed, which is particularly striking since *C. elegans* has the highest fraction of HAGHAG acceptors. This rareness of alternative splice events at NAGNAG acceptors is not due to differences in the EST coverage (see Table 4.4) and *C. elegans* has a similar percentage of confirmed tandem donors compared to the other species. This finding, based on a larger data set, confirms our initial result that NAGNAGs in *C. elegans* are rarely alternatively spliced.

4.6 A relational database of tandem splice sites

Although tandem splice sites are frequent in many species, neither existing databases on alternative splicing nor genome browsers provide easy and comprehensive access to

species	no. of acceptors[a]	NAGNAG		HAGHAG		confirmed NAGNAG		confirmed HAGHAG	
H. sapiens	164,841	9,465	5.7%[b]	3,530	37.3%[c]	1,511	16%[d]	1,373	90.9%[e]
M. musculus	125,233	7,116	5.7%	2,662	37.4%	1,087	15.3%	1,022	94.0%
R. norvegicus	53,598	3,080	5.7%	1,098	35.6%	215	7.0%	202	94.0%
G. gallus	19,794	1,069	5.4%	401	37.5%	97	9.1%	92	94.8%
D. rerio	29,067	1,540	5.3%	484	31.4%	132	8.6%	118	89.4%
D. melanogaster	39,441	1,584	4.0%	859	54.2%	177	11.2%	170	96.0%
C. elegans	92,867	4,184	4.5%	2,637	63.0%	33	0.8%	33	100.0%

Table 4.13: NAGNAG acceptors in seven species.
[a] total number of all unique acceptor sites annotated in RefSeq transcripts
[b] no. of NAGNAG acceptors / no. of all acceptors
[c] no. of HAGHAG acceptors / no. of NAGNAG acceptors
[d] no. of confirmed NAGNAG acceptors / no. of NAGNAG acceptors
[e] no. of confirmed HAGHAG acceptors / no. of confirmed NAGNAG acceptors

this phenomenon. Since we had collected a wealth of data about the splice events at GYNGYN donors and NAGNAG acceptors, we developed a database TassDB (TAndem Splice Site DataBase) to provide public access to these data. TassDB consists of a relational database (PostgreSQL 8.0.3) and a web interface to retrieve the data. With this database, we aim at facilitating further large-scale bioinformatics as well as experimental analysis of tandem splice sites.

TassDB contains tandem splice sites of eight species (the seven species listed in Table 4.13 and in addition the dog (*Canis familiaris*)). We stored the following data:

- the tandem splice site motif,
- the maximum entropy scores for both splice sites [218],
- the genomic locus,
- location in the transcript (5'/3' UTR or intron phase 0/1/2),
- the impact of the splice event on the protein,
- the sequences and length of the up- and downstream exon and the intron,
- and information about the ESTs and mRNAs that match the E/I and e/i transcript, respectively.

In addition, the database annotates the 121 NAGNAG acceptor SNPs (identified in section 4.3.1). As for 51 polymorphic tandem acceptors the NAGNAG pattern is not visible in the genome reference sequence, TassDB always stores the allele sequence with the NAGNAG acceptor.

The basic design of this database was driven by the idea to separate splice site specific data from transcript specific data. For example, the GYNGYN and NAGNAG motif, the genomic locus, and the splice site scores are independent of the transcript annotation. However, features such as intron phase, protein impact, and EST confirmation depend on

PHF1 PHD finger protein 1		
acceptor CAGCAG (plausible tandem)		
locus	chr6:33491313-33491376	
sequence context	gactttccccactccaacccCAGCAGCCCCATCCGGATGTTTGCTT	
splice site scores E/I	6.093 / 1.942	
transcript	NM_024165	NM_002636
exon number	14	13
annotated splice site	E	E
number of E/I transcripts	25 / 1	2 / 1
transcript / protein impact	intron phase 2 / indel S	intron phase 1 / indel A
position in protein	aa: 444	aa: 409
exon/intron/exon context	nt: 95 / 260 / 81	nt: 179 / 445 / 81

Figure 4.10: TassDB result table for the CAGCAG acceptor of intron 13 of *PHF1*.

the annotation as well as the exon-intron structure of the transcript. Thus, one tandem splice site can have multiple transcript specific data. For example, the intron 13 of the human *PHF1* gene that contains a CAGCAG acceptor is in intron phase 2 according to the annotation of NM_024165. Due to skipping of the upstream 95 nt exon, this intron is in phase 1 according to the annotation of another transcript NM_002636. Thus, the protein impact of the CAGCAG is indel S in NM_024165 but indel A in NM_002636. In such cases, TassDB will show a result table with more than two columns (Figure 4.10).

The database has three web interfaces to retrieve data that allow:

1. to search for all (confirmed and unconfirmed) GYNGYNs and NAGNAGs of a gene given its symbol or transcript accession,
2. to search for genes containing tandem splice sites with specific features (splice site pattern, the number of ESTs/mRNAs that match both splice forms, location in the UTR or in the CDS, and the protein impact),
3. and to perform complex searches by sending a user-defined SQL query to the database.

Thus, TassDB can be used to retrieve large datasets for further computational analysis of tandem splice sites. Furthermore, it allows biologists to search for their genes of interest and to get all relevant information about the respective tandem splice sites. The database is available at http://helios.informatik.uni-freiburg.de/TassDB/.

4.7 Discussion

4.7.1 Functional consequences of tandem splice sites

We have performed the first detailed analysis of alternative splicing at GYNGYN donors and NAGNAG acceptors. Addressing the four questions on page 85 in a genome-wide analysis, we detected the following characteristics of NAGNAG acceptors:

- a bias towards introns in phase 1,
- a bias towards single aa indels,
- the indel of a charged aa happens in a protein context that is similarly charged,
- exon-exon junctions of confirmed NAGNAGs are enriched in polar residues,
- genes with NAGNAG acceptors are frequently involved in protein-protein interactions,
- genes with NAGNAG acceptors exhibit a special distribution of Pfam domains,
- a subset of NAGNAG acceptors is conserved in mouse,
- splicing at NAGNAG acceptors can be tissue-specific.

These points strongly indicate that NAGNAG acceptors are subjected to selection pressures during evolution. In particular, as signs of negative selection, the protein-related biases show that the NAGNAG-derived variability is deleterious for certain proteins or protein regions.

Noteworthy, our findings concerning tissue specificity and conservation were confirmed and extended by other groups later. Tadokoro et al. detected tissue-specific variations of E and I transcripts for several genes [202]. Akerman and Mandel-Gutfreund confirmed the sequence conservation of the NAGNAG motif and additionally found a high conservation of intronic flanking regions [221]. Furthermore, they detected several overrepresented sequence motifs in the vicinity of confirmed NAGNAGs, which might contribute to the regulation of these sites.

As conservation and tissue-specific regulation is usually taken as evidence for biological or functional relevance of splice events, we performed a literature search and found several cases where NAGNAG-derived alternative splicing events result in functionally different protein isoforms:

- *IGF1R* isoforms (Thr-Gly vs. Arg) have different signaling activities [188],
- *DRPLA* isoforms (Gln indel) have differences in subcellular localization [202],
- mouse *Pax-3* isoforms (Gln indel) have different DNA binding affinities [189],
- isoforms of the *A. thaliana* U11-35K protein (Gln indel) have different protein binding affinities [222],
- alternative NAGNAG splicing in the UTR of mouse *Ggt1* affects the translational efficiency [223].

Furthermore, a mutation in the *ABCA4* gene that enables alternative NAGNAG splicing is relevant for Stargardt disease 1 [204]. It is tempting to speculate that further functional differences can be caused by changes in recognition sequences for post-translational modifications (such as phosphorylation) or by a variation in the distance between functional domains in proteins.

For tandem donors, the smaller data set does not allow to perform all the statistical analyses done for NAGNAG acceptors. In contrast to NAGNAGs, the GTNGTN pattern is not significantly conserved and tandem donor splicing seems not to be tissue-specific. However, a considerable fraction of the human confirmed and evolutionary conserved tandem donors is also confirmed in other species, and tandem donors exhibit the same bias towards single aa indels as NAGNAGs. Interestingly, one tandem donor with functional consequences is also described in the literature. Alternative splicing at a GTAGTG donor of the human glucocorticoid receptor gene *NR3C1* leads to an Arg insertion between two zinc fingers, and the respective isoform has an activity reduced to 48% [224, 225].

As a major difference to alternative splicing in general, which often severely affects the protein function, tandem splice sites provide a mechanism to create subtle changes. This is supported by our finding that confirmed GTNGTNs and NAGNAGs are significantly enriched in single aa events and that for NAGNAGs the indel of a charged aa happens in a protein context that is similarly charged. As discussed above, these subtle changes may be of functional relevance.

Apart from functional consequences, it is conceivable that many other of these subtle splice events might have no functional implications. Except for dramatic events such as stop codon indels, the amino acid indels might simply be tolerated by the cells. Thus, similar to genetic variants, tandem donor and acceptor splice variants may be neutral with respect to physiological protein properties or may result in phenotypic differences. Consequently, these splice variants represent another large playground of molecular evolution [226, 77], with purifying selection acting to remove deleterious variants and positive selection reinforcing beneficial variants. Which fraction of NAGNAG acceptors and GYNGYN donors plays a role in biological functions deserves further research.

We and others [202] found a subset of NAGNAG acceptors to be regulated in a tissue-specific manner. Other NAGNAG acceptors as well as GYNGYN donors do not exhibit this property [227]. The latter group of splice events is likely to contain cases, which are the result of a stochastic or 'noisy' binding of the spliceosome to the neighboring splice sites [228]. This is supported by the finding that the strength of a splice site is an important factor that determines its frequency (section 4.2.2, 4.5.3 and [228]). However, it should be noted that noise is important for many biological processes such as cell divergence in Dictyostelium, pili expression in bacteria, and neuronal firing [229, 230], leading to the model of 'cultivated noise' [230]. Another example is noisy alternative splicing of the Drosophila *DSCAM* gene. This gene has four clusters of mutually exclusive

exons and may express a total of 38,016 different transcripts. In a stochastic manner each single cell produces only a small number of the 38,016 possible transcripts [231]. This noisy process results in an efficient mechanism to individualize cells, which is important for proper axon guidance in Drosophila [55]. Although it has to be proven, it is tempting to speculate that noise arising by orchestrated splicing at tandem splice sites provides another stochastic mechanism, maybe for cell individualization.

4.7.2 A mechanism to increase the protein diversity

One of the main findings is that the rather simple structures of tandem splice sites allow highly diverse protein events. Confirmed tandem donors and acceptors are able to insert twelve of the 20 different amino acids by single aa events and the dipeptide exchanges are even more diverse. Furthermore, stop codon indels were observed for GYNGYNs and NAGNAGs.

Currently, there are more than 1,500 confirmed human NAGNAG acceptors. Due to a limited EST coverage, we expect more NAGNAGs to be also alternatively spliced. Since a large number of genes is affected, alternative splicing at tandem splice sites is a major mechanism to increase the proteome diversity. Additionally, the number of possible protein variants is strongly increased, if one gene harbors more than one tandem acceptor. We found several genes with more than one confirmed tandem acceptor (up to five for the human *NCOR1* gene). If alternative splice events from the five tandem acceptors are freely combined, this would result in $2^5 = 32$ *NCOR1* protein isoforms.

The simultaneous use of a GYNGYN donor and a NAGNAG acceptor for one intron further contributes to variability (Figure 4.6B). Such an example is intron 9 of *BRUNOL4* for which we found 14 e-E, three i-E, and six e-I transcripts in dbEST that result in protein forms with a GPA, AA, and GP peptide, respectively.

Another dimension of variability of the protein level comes from translationally non-silent SNPs that affect NAGNAG acceptors. Of the 64 SNP detected in section 4.3.2, 15 (23%) are non-silent and thus change the I acceptor and the amino acid sequence of the E protein. While homozygotes express either one or two isoforms, heterozygosity results in even three different proteins (Figure 4.11). The amino acid change can be dramatic, as for example from Glu to the oppositely charged Lys in *PAPSS2*. Moreover, it is conceivable that some of the non-silent SNPs may confer a heterozygous advantage.

Apart from tandem splice sites, there are other mechanisms to introduce subtle protein changes by alternative splicing. Very similar mutually exclusive exons can lead to similar but functionally different proteins. This was observed for the *RAB6A* gene [232], many ion channels [82], and the above mentioned Drosophila *DSCAM* gene. Comparing mutually exclusive exons with tandem splice sites, the latter provide a simpler way to introduce minor protein changes, probably explaining their higher frequency.

Figure 4.11: A SNP that affects the I acceptor and the amino acid sequence of the E protein (rs2275992 in *ZFP91*).
(A) Homozygosity of the G allele without a NAGNAG results in the expression of one protein, (B) homozygosity of A allele with the NAGNAG results in two, and (C) heterozygosity in three isoforms. All three transcripts are confirmed by at least four ESTs/mRNAs. The two allele variants are highlighted in blue and green. Amino acids are shown below the second codon position. Upper and lower case letters indicate exonic and intronic nucleotides, respectively. Exons are boxed.

4.7.3 Mechanism of tandem donor and acceptor splicing

NAGNAG acceptors prevalently occur in genes coding for proteins that interact with other proteins and RNA molecules. In agreement with that, genes involved in splicing are equipped with tandem acceptors, for example *PRPF3*, *PRPF8*, *PRPF31*, *PRPF4B*, *SFRS11*, and the neuronal polypyrimidine tract-binding protein *PTBP2*, which is of particular interest for alternative splicing in the nervous system [233]. These tandem acceptors are conserved between human, mouse, and rat. Moreover, tandem acceptors raise interesting questions about the 3' splice site selection [234]. Their alternative usage requires some flexibility in the interaction of the splicing factors that recognize the branch point, the polypyrimidine tract, and the splice site AG. This flexibility of the splicing machinery may be enhanced by isoforms of its protein components. U2AF35 is known to be alternatively spliced [235]. Interestingly, both U2AF35 as well as its interacting partner U2AF65 exhibit a tandem acceptor. Tandem acceptor derived isoforms of U2AF subunits might promote flexibility in the spatial architecture of the spliceosome with possible functional consequences for the splicing process and even for splicing at NAGNAG acceptors.

Most confirmed GYNGYNs have a low free energy of U1 snRNA binding to both the e and i donor, suggesting that the U1 snRNA can stably bind to both sites. However, there are a few exceptions where one donor is much stronger than the other one in a confirmed tandem (Figure 4.8A). The mechanism of splicing at these sites remains unclear, but there are several hypotheses that might guide future investigations. For example, it has been reported that U6 snRNA rather than U1 snRNA determines a donor site in the human *FGFR1* gene [236]. Moreover, there is *in vitro* evidence that splicing can occur without U1 snRNA binding to the donor site [237, 238]. Furthermore, other protein factors can

influence the splice site choice and/or (de)stabilize U1 snRNA binding [239, 240]. We believe that a further experimental investigation of confirmed tandem splice sites may help to elucidate further details of the splicing process.

4.7.4 Polymorphic NAGNAG acceptors

We identified 121 SNPs that may affect alternative splicing by creation, destruction, or changing NAGNAG acceptors. In order to improve the specificity of our prediction, we classified NAGNAG acceptors into plausible and implausible ones. This subdivision of the tandem acceptors, primarily based on the degree of confirmation by mRNA and EST data, is further supported by

- the fact that GAG acceptors are very rare [182],
- our genome-wide observation that only plausible but not implausible NAGNAGs have the same bias towards intron phase 1 as EST confirmed NAGNAGs,
- and the observed differences in the number of SNPs that affect the AGs of ancestral plausible and implausible NAGNAGs, respectively.

The latter indicates, that the selection pressure to maintain the E acceptor for HAGGAGs is higher than the pressure to preserve the coding sequence since destruction of the HAG acceptor will leave a GAG that is unlikely to act as an acceptor site. In contrast, for plausible HAGHAGs, destruction of either AG is much less deleterious as the other will still function as an acceptor. Nevertheless, it represents an experimental and bioinformatics challenge to elucidate what makes the rare cases of confirmed implausible NAGNAG acceptors. Focusing on SNPs that affect NAGNAG acceptors, our approach for the identification of SNPs that result in variations in alternative splicing patterns is highly effective.

We used SNPs in NAGNAG motifs as 'knock-out experiments by nature' to confirm that the disruption of a plausible NAGNAG acceptor abolishes the expression of alternative transcripts. Then, we asked whether NAGNAG motifs created by the non-ancestral SNP alleles allow alternative splicing or not. Usually, the introduction of an AG anywhere in the pre-mRNA does not create a functional acceptor site since a polypyrimidine tract upstream and possibly enhancer sequences are required for recognition by the spliceosome. However, we suppose that the creation of a second AG three bases up- or downstream of an existing acceptor is very likely to result in a functional tandem acceptor since the splice relevant sequence context is already present. Referring to the chimpanzee genome as the reference for ancestral SNP alleles, we found EST and RT-PCR evidences that novel plausible NAGNAGs are most likely functional. This implies that a change of a normal acceptor to a plausible NAGNAG acceptor by a single mutation is sufficient to enable alternative splicing. Although the mechanism of NAGNAG splicing is not understood in detail, our findings argue against a general involvement of other signals than the

NAGNAG motif itself. However, additional signals might be necessary for tissue-specific regulation of alternative splicing at tandem acceptors [221]. We conclude that SNPs in plausible NAGNAGs have an influence on splicing at ancestral as well as non-ancestral NAGNAGs.

Alternative splicing is a major source of proteome diversity and is therefore relevant as a therapeutic target [68]. The subtle effects of alternative splicing at tandem acceptors on multiple proteins simultaneously might be of importance in the pathogenesis of complex diseases. For example, four of the six genes known to cause obesity by single-gene mutations contain NAGNAG acceptors (*LEPR*, *POMC*, *PCSK1*, and *LEP*). Furthermore, SNPs in NAGNAG acceptors might be associated with human diseases. The SNP rs1650232 within a NAGNAG acceptor is associated with the respiratory-distress syndrome [241] and for the *ABCA4* gene the disease relevance of a NAGNAG mutation was demonstrated [204]. Moreover, we found that 18 of the 64 (28%) SNPs that affect plausible NAGNAGs (see section 4.3.2) occur in known disease genes. Thus, they are preferable candidates for more detailed functional analyses and association studies to link alternative splicing with diseases.

4.7.5 Alternative splicing at tandem donors

Extending our analysis to the 5' intron end, we found alternative donor usage for GTNGTN, GTNGCN, and GCNGTN motifs in eight investigated eukaryotic species. Since only about 1% of all GYNGYN donors are confirmed, we have to exclude that the observed events are attributed to EST artifacts. Several lines of evidence indicate that the majority of confirmed GYNGYN splice events are real:

- our experimental verification of six human and one mouse GTNGTN donor,
- numerous GTNGTNs are confirmed by multiple ESTs/mRNAs and for several of these events both e and i transcripts are deposited in the RefSeq database,
- the existence of orthologous tandem donors that are confirmed in two or more species,
- confirmed GTNGTNs have a higher conservation of the exonic and intronic flanking regions, a situation that is typical for conserved alternative splice events [76, 75],
- all of the six investigated human individuals express e and i transcripts for *STAT3*, thus excluding the possibility of allele-specific instead of alternative splicing [106],
- by manual examination of all human confirmed GYNGYNs, we excluded the existence of paralogs or processed pseudogenes that could mimic alternative splicing at a tandem donor.

The percentage of donor sites with a GYNGYN motif as well as the percentage of tandem donors that are confirmed is very similar between the eight investigated species (tolerating some variation probably due to differences in the number of ESTs and mRNAs). Given the large evolutionary distance between *A. thaliana*, *C. elegans*, and *H. sapiens*, it is likely

that all species that have alternatively spliced genes are able to produce e and i transcripts at certain tandem donor sites. The detection of 44 alternatively spliced tandem donors in *A. thaliana* is consistent with the recent finding that alternative splicing in plants is not as rare as thought for a long time [242, 243].

Although only a fraction of the tandem donors is confirmed, we found features that distinguish confirmed from unconfirmed ones. Since the non-annotated donor of unconfirmed tandems does not allow a sufficiently stable binding to the U1 snRNA, the other donor is used exclusively in the splice process. For confirmed tandem donors, both sites allow a stable binding to U1 snRNA. However, in most of the confirmed cases one donor has a better strength and this results in its preferred usage as measured by the EST ratio between both transcripts. The second discriminative feature is the overabundance of G triplets in the intronic flanks of confirmed GTNGTNs, especially for introns shorter than 200 nt. This triplet is the core of many known ISE motifs [43, 219] and was demonstrated to function in splice site definition [220]. Interestingly, in the human alpha-globin gene, GGG elements were shown to exert their effect by binding to the nucleotides 8-10 (5'-CCT-3') of the U1 snRNA [219]. We have searched for overrepresented tetramers and found a significantly higher frequency of CGGG and GGGT. Strikingly, the nucleotides 7-11 of U1 snRNA are 5'-ACCTG-3'. The CGGG as well as the GGGT motif are complementary to this part of U1, thus it is tempting to speculate that these motifs bind to U1 snRNA with four instead of three base pairs. Since CGGG and GGGT are more frequent in the intronic flanks of confirmed tandem donors, they may be involved in alternative splicing at these donor sites. If U1 snRNA is a critical factor, we do not expect much variation in splicing between tissues since U1 is ubiquitously expressed in high amounts. Consistent with this notion, the six experimentally investigated tandem donors exhibit similar e to i transcript ratios in all tissues.

Chapter 5

Outlook

In the first part of this thesis, we described a novel Pfam domain based approach for predicting exon skipping and intron retention events without using EST data. We developed an efficient algorithm to overcome the computational complexity and demonstrated in a genome-wide application that this method yields highly reliable predictions.

As EST data is often the limiting factor in splice event discovery, it should be noted that there are only seven species having more than one million ESTs in dbEST. Even model organisms that have been studied for a long time can have a surprisingly low number of ESTs, for example *C. elegans* has less than 350,000 ESTs. Thus, it is likely that many splice events in these organisms remain to be detected. Noteworthy, despite the existence of about eight million human ESTs, we were still able to predict and verify novel human splice events, indicating that even millions of ESTs are not enough.

Furthermore, as a precondition for the accurate detection of conserved elements by comparative genomics [244], numerous genome sequencing projects are in progress, providing additional raw genome sequences in future. It is unlikely that these genomes will be complemented by as many ESTs and mRNAs as there are for the human genome. Thus, the annotation of genes, transcripts, and proteins needs computational tools. Genome browsers and databases like Ensembl or the UCSC Genome Bioinformatics Site use complex pipelines to address the first step, which is the prediction of genes in genomic sequences [98, 245]. These pipelines use information from *ab initio* gene prediction algorithms, comparative methods, and available transcript and protein sequence data. Next, a second pipeline can predict the entire set of transcripts. Although numerous algorithms for *ab initio* gene prediction exist, approaches for *ab initio* splice event prediction are rather recent developments, thus their number is small. The second pipeline, currently mainly using EST data [246], needs to be extended by these non-EST based prediction methods and, of course, data from available microarray experiments. Based on the predictions of this second pipeline, a third pipeline can annotate the protein sequences and predict their function. Noteworthy, the knowledge about proteins and their function requires information about the transcripts. For completeness, it should be mentioned that

this simple three step view (set of genes → set of transcripts → set of proteins) describes only a very rough picture, which misses many important aspects such as mRNA editing, mRNA stability, or post-translational protein modifications.

We have shown that our splice event prediction approach is highly reliable in human. As Pfam domains are annotated in many other species, we expect that this method also yields reliable predictions when applied to other genomes. Furthermore, our approach complements existing methods since our predictions have little overlap with the predictions of other methods. In summary, we hope that our method will become part of a larger set of bioinformatics approaches for non-EST based splice event prediction and that these approaches will be routinely and successfully used in future.

In the second part of this thesis, we provide evidence that the secondary structure of splicing factor binding sites affects the splicing process. Then, we developed a new motif finding algorithm that integrates the additional knowledge about secondary structures to better discriminate real from spurious protein binding sites.

Like cells use a code for translation, they have a code for (constitutive and alternative) splicing [178]. It is of great interest to decipher this splicing code, which apparently seems to be much harder compared to the translation code. Much experimental and computational research focuses on the characterization of regulatory splicing proteins and their binding motifs. Often, only the sequence of a binding site but not its secondary structure context is considered. Our findings argue that the sequestration of binding sites in double-strands can hamper or abolish binding of splicing factors, which seems to hold for any single-stranded RNA binding protein as well [132]. This general principle has many fascinating implications from interpreting mutagenesis experiments or the effect of splicing relevant SNPs to potential new therapeutic ways to correct splicing defects. As secondary structures are dynamic and small sequence changes may induce large structure changes [177], it is tempting to speculate that nature uses mRNA structural changes to produce different splicing patterns under different conditions (such as developmental stages or tissues) or even during evolution of species [133]. Most importantly, this mechanism provides another piece in the complex puzzle of the splicing code.

We demonstrated that information about secondary structures integrated in the MEMERIS algorithm helps to identify the real binding motif. However, an evaluation of MEMERIS on SELEX data for splicing factors turned out to be difficult due to degenerate sequence motifs and an unknown location of the real binding sites in these sequences. Thus, how useful MEMERIS is to identify binding motifs of splicing factors awaits further investigations and experimental validations of the predicted binding sites.

Despite evidence that real binding sites have a higher single-strandedness, we do not know in which regions mRNA is free to fold and to which extent base pairs within the binding sites can be tolerated by the binding proteins. A better understanding of

these points will help to make better predictions and to design better computational approaches.

In the third part of this thesis, we focus on a barely investigated group of alternative splice events that occur at tandem donors and acceptors. We provide strong evidences that tandem splice sites represent an important mechanism to increase the proteome diversity in a wide range of species. Recently, these subtle splice events have attracted attention by several other research groups [202, 221, 227, 228, 247].

The elucidation which splice variants have functional consequences and what these consequences are is one big area of biological research and many experimental studies address this topic. Furthermore, computational analyses contribute to the global picture of the diverse functions of alternative splicing. However, the subtle splice events at tandem splice sites were often not considered in experimental as well as computational studies. Nevertheless, few cases are known where alternative splicing at tandem splice sites result in functionally different protein isoforms. Thus, for a complete characterization of a transcriptome, it is necessary to include these subtle splice events.

However, which of these subtle splice events play a biological role and which are due to noise that is tolerated by cells remain challenging, since the putative effect of NAG indels on transcripts and the effect of single amino acid indels on proteins is mostly not obvious. Therefore, it is important to extract promising candidates for further experiments. To this end, knowledge about conservation and tissue-specific regulation can be used. We have identified functional gene groups that are enriched or depleted in NAGNAG acceptors. Further analyses are needed to refine and extend these protein regions exhibiting over- or underrepresentation of tandem splice sites. As enrichment may indicate positive selection, these tandem splice sites may be interesting experimental candidates too. Likewise, SNPs that create tandem acceptors in protein regions where NAGNAGs are usually depleted may be interesting for association studies or for linking genetic variations to phenotypic effects. Finally, we hope that our tandem splice site database TassDB will be useful for computational and experimental studies.

As a more speculative functional impact, the model of 'cultivated noise' [230] may apply to some tandem splice sites. We believe that it was important to shed light on tandem donors and acceptors and that ongoing studies will provide fascinating new insights.

Bibliography

[1] Gilbert W. Why genes in pieces? *Nature* **271**(5645), 501, 1978.

[2] Consortium IHGS. Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011), 931–45, 2004.

[3] Gooding C, Clark F, Wollerton MC, Grellscheid SN, Groom H, et al. A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. *Genome Biol* **7**(1), R1, 2006.

[4] Burge CB, Tuschl T, Sharp PA. *Splicing of precursors to mRNAs by the spliceosomes*, 525–560. The RNA World II, R. F. Gesteland and T. R. Cech and J. F. Atkins, (eds.). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1999.

[5] Early P, Rogers J, Davis M, Calame K, Bond M, et al. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**(2), 313–9, 1980.

[6] Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8**(9), 967–74, 1998.

[7] Mironov AA, Fickett JW, Gelfand MS. Frequent alternative splicing of human genes. *Genome Res* **9**(12), 1288–93, 1999.

[8] Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, et al. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* **474**(1), 83–6, 2000.

[9] Consortium IHGS. Initial sequencing and analysis of the human genome. *Nature* **409**(6822), 860–921, 2001.

[10] Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research* **29**(13), 2850–9, 2001.

[11] Brett D, Pospisil H, Valcarcel J, Reich J, Bork P. Alternative splicing and genome complexity. *Nat Genet* **30**(1), 29–30, 2002.

[12] Kim H, Klein R, Majewski J, Ott J. Estimating rates of alternative splicing in mammals and invertebrates. *Nat Genet* **36**(9), 915–6, 2004.

[13] Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. *Genome Biol* **5**(10), R74, 2004.

[14] Xu Q, Modrek B, Lee C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Research* **30**(17), 3754–66, 2002.

[15] Xu Q, Lee C. Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Research* **31**(19), 5635–43, 2003.

[16] Lee C, Roy M. Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol* **5**(7), 231, 2004.

[17] Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**(5653), 2141–4, 2003.

[18] Le K, Mitsouras K, Roy M, Wang Q, Xu Q, et al. Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Research* **32**(22), e180, 2004.

[19] Sugnet CW, Srinivasan K, Clark TA, O'brien G, Cline MS, et al. Unusual Intron Conservation near Tissue-Regulated Exons Found by Splicing Microarrays. *PLoS Comput Biol* **2**(1), e4, 2006.

[20] Blanchette M, Green RE, Brenner SE, Rio DC. Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila. *Genes Dev* **19**(11), 1306–14, 2005.

[21] Nagao K, Togawa N, Fujii K, Uchikawa H, Kohno Y, et al. Detecting tissue-specific alternative splicing and disease-associated aberrant splicing of the PTCH gene with exon junction microarrays. *Hum Mol Genet* **14**(22), 3379–88, 2005.

[22] Lim LP, Burge CB. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl. Acad. Sci. USA* **98**(20), 11193–8, 2001.

[23] Zhang XHF, Heller KA, Hefter I, Leslie CS, Chasin LA. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res* **13**(12), 2637–50, 2003.

[24] Stamm S. Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. *Hum Mol Genet* **11**(20), 2409–16, 2002.

[25] Blencowe BJ. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends in Biochemical Sciences* **25**(3), 106–10, 2000.

[26] Ladd AN, Cooper TA. Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol* **3**(11), reviews0008, 2002.

[27] Wang Z, Xiao X, Van Nostrand E, Burge CB. General and specific functions of exonic splicing silencers in splicing control. *Mol Cell* **23**(1), 61–70, 2006.

[28] Graveley BR. Sorting out the complexity of SR protein functions. *RNA* **6**(9), 1197–211, 2000.

[29] Pozzoli U, Sironi M. Silencers regulate both constitutive and alternative splicing events in mammals. *Cell Mol Life Sci* **62**(14), 1579–604, 2005.

[30] Goren A, Ram O, Amit M, Keren H, Lev-Maor G, et al. Comparative analysis identifies exonic splicing regulatory sequences–the complex definition of enhancers and silencers. *Mol Cell* **22**(6), 769–81, 2006.

[31] Schaal TD, Maniatis T. Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Mol Cell Biol* **19**(1), 261–73, 1999.

[32] Wang J, Smith PJ, Krainer AR, Zhang MQ. Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Research* **33**(16), 5053–62, 2005.

[33] Smith CW, Valcarcel J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends in Biochemical Sciences* **25**(8), 381–8, 2000.

[34] Han K, Yeo G, An P, Burge CB, Grabowski PJ. A Combinatorial Code for Splicing Silencing: UAGG and GGGG Motifs. *PLoS Biol* **3**(5), e158, 2005.

[35] Mayas RM, Maita H, Staley JP. Exon ligation is proofread by the DExD/H-box ATPase Prp22p. *Nat Struct Mol Biol* **13**(6), 482–90, 2006.

[36] Ule J, Ule A, Spencer J, Williams A, Hu JS, et al. Nova regulates brain-specific splicing to shape the synapse. *Nat Genet* **37**(8), 844–52, 2005.

[37] Stoilov P, Daoud R, Nayler O, Stamm S. Human tra2-beta1 autoregulates its protein concentration by influencing alternative splicing of its pre-mRNA. *Hum Mol Genet* **13**(5), 509–24, 2004.

[38] Shin C, Feng Y, Manley JL. Dephosphorylated SRp38 acts as a splicing repressor in response to heat shock. *Nature* **427**(6974), 553–8, 2004.

[39] Nogues G, Kadener S, Cramer P, de la Mata M, Fededa JP, et al. Control of alternative pre-mRNA splicing by RNA Pol II elongation: faster is not always better. *IUBMB Life* **55**(4-5), 235–41, 2003.

[40] Buratti E, Baralle FE. Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol Cell Biol* **24**(24), 10505–14, 2004.

[41] Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**(5583), 1007–13, 2002.

[42] Zhang XHF, Chasin LA. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* **18**(11), 1241–50, 2004.

[43] Yeo G, Hoon S, Venkatesh B, Burge CB. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc. Natl. Acad. Sci. USA* **101**(44), 15700–5, 2004.

[44] Brudno M, Gelfand MS, Spengler S, Zorn M, Dubchak I, et al. Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucleic Acids Research* **29**(11), 2338–48, 2001.

[45] Miriami E, Margalit H, Sperling R. Conserved sequence elements associated with exon skipping. *Nucleic Acids Research* **31**(7), 1974–83, 2003.

[46] Maniatis T, Tasic B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**(6894), 236–43, 2002.

[47] Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, et al. Function of alternative splicing. *Gene* **344C**, 1–20, 2005.

[48] Garcia J, Gerber SH, Sugita S, Sudhof TC, Rizo J. A conformational switch in the Piccolo C(2)A domain regulated by alternative splicing. *Nat Struct Mol Biol* **11**(1), 45–53, 2004.

[49] Rudenko G, Nguyen T, Chelliah Y, Sudhof TC, Deisenhofer J. The structure of the ligand-binding domain of neurexin Ibeta: regulation of LNS domain function by alternative splicing. *Cell* **99**(1), 93–101, 1999.

[50] Kamatkar S, Radha V, Nambirajan S, Reddy RS, Swarup G. Two splice variants of a tyrosine phosphatase differ in substrate specificity, DNA binding, and subcellular location. *Journal of Biological Chemistry* **271**(43), 26755–61, 1996.

[51] Resch A, Xing Y, Modrek B, Gorlick M, Riley R, et al. Assessing the impact of alternative splicing on domain interactions in the human proteome. *J Proteome Res* **3**(1), 76–83, 2004.

[52] Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, et al. Increase of functional diversity by alternative splicing. *Trends in Genetics* **19**(3), 124–8, 2003.

[53] Xing Y, Xu Q, Lee C. Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. *FEBS Lett* **555**(3), 572–8, 2003.

[54] Cline MS, Shigeta R, Wheeler RL, Siani-Rose MA, Kulp D, et al. The effects of alternative splicing on transmembrane proteins in the mouse genome. In *Pacific Symposium on Biocomputing*, 17–28. 2004.

[55] Wojtowicz WM, Flanagan JJ, Millard SS, Zipursky SL, Clemens JC. Alternative Splicing of Drosophila Dscam Generates Axon Guidance Receptors that Exhibit Isoform-Specific Homophilic Binding. *Cell* **118**(5), 619–33, 2004.

[56] Lynch KW. Consequences of regulated pre-mRNA splicing in the immune system. *Nat Rev Immunol* **4**(12), 931–40, 2004.

[57] Roberts AG, Redding SJ, Llewellyn DH. An alternatively-spliced exon in the 5'-UTR of human ALAS1 mRNA inhibits translation and renders it resistant to haem-mediated decay. *FEBS Lett* **579**(5), 1061–6, 2005.

[58] Maquat LE. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol* **5**(2), 89–99, 2004.

[59] Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. USA* **100**(1), 189–92, 2003.

[60] Baek D, Green P. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc. Natl. Acad. Sci. USA* **102**(36), 12813–8, 2005.

[61] Wollerton MC, Gooding C, Wagner EJ, Garcia-Blanco MA, Smith CWJ. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol Cell* **13**(1), 91–100, 2004.

[62] Pan Q, Saltzman AL, Kim YK, Misquitta C, Shai O, et al. Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev* **20**(2), 153–8, 2006.

[63] Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3**(4), 285–98, 2002.

[64] Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. *Genes Dev* **17**(4), 419–37, 2003.

[65] Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett* **579**(9), 1900–3, 2005.

[66] Lynch KW, Weiss A. A CD45 polymorphism associated with multiple sclerosis disrupts an exonic splicing silencer. *Journal of Biological Chemistry* **276**(26), 24341–7, 2001.

[67] Kalnina Z, Zayakin P, Silina K, Line A. Alterations of pre-mRNA splicing in cancer. *Genes Chromosomes Cancer* **42**(4), 342–57, 2005.

[68] Sazani P, Kole R. Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing. *J Clin Invest* **112**(4), 481–6, 2003.

[69] Sorek R, Shamir R, Ast G. How prevalent is functional alternative splicing in the human genome? *Trends in Genetics* **20**(2), 68–71, 2004.

[70] Thanaraj TA, Clark F, Muilu J. Conservation of human alternative splice events in mouse. *Nucleic Acids Research* **31**(10), 2544–52, 2003.

[71] Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB. Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl. Acad. Sci. USA* **102**(8), 2850–5, 2005.

[72] Pan Q, Bakowski MA, Morris Q, Zhang W, Frey BJ, et al. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends in Genetics* **21**(2), 73–7, 2005.

[73] Resch A, Xing Y, Alekseyenko A, Modrek B, Lee C. Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Research* **32**(4), 1261–9, 2004.

[74] Magen A, Ast G. The importance of being divisible by three in alternative splicing. *Nucleic Acids Research* **33**(17), 5574–82, 2005.

[75] Sugnet CW, Kent WJ, Ares MJ, Haussler D. Transcriptome and genome conservation of alternative splicing events in humans and mice. In *Pacific Symposium on Biocomputing (PSB 2004)*, 66–77. 2004.

[76] Sorek R, Ast G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res* **13**(7), 1631–7, 2003.

[77] Modrek B, Lee CJ. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet* **34**(2), 177–80, 2003.

[78] Zhang XHF, Chasin LA. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc. Natl. Acad. Sci. USA* **103**(36), 13427–32, 2006.

[79] Sorek R, Ast G, Graur D. Alu-containing exons are alternatively spliced. *Genome Res* **12**(7), 1060–7, 2002.

[80] Sorek R, Lev-Maor G, Reznik M, Dagan T, Belinky F, et al. Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. *Mol Cell* **14**(2), 221–31, 2004.

[81] Kondrashov FA, Koonin EV. Origin of alternative splicing by tandem exon duplication. *Hum Mol Genet* **10**(23), 2661–9, 2001.

[82] Copley RR. Evolutionary convergence of alternative splicing in ion channels. *Trends in Genetics* **20**(4), 171–6, 2004.

[83] Xing Y, Lee C. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl. Acad. Sci. USA* **102**(38), 13526–31, 2005.

[84] Cusack BP, Wolfe KH. Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons. *Mol Biol Evol* **22**(11), 2198–208, 2005.

[85] Kopelman NM, Lancet D, Yanai I. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat Genet* **37**(6), 588–9, 2005.

[86] Clark F, Thanaraj TA. Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum Mol Genet* **11**(4), 451–64, 2002.

[87] Matos P, Collard JG, Jordan P. Tumor-related alternatively spliced Rac1b is not regulated by Rho-GDP dissociation inhibitors and exhibits selective downstream signaling. *Journal of Biological Chemistry* **278**(50), 50442–8, 2003.

[88] Modrek B, Lee C. A genomic view of alternative splicing. *Nat Genet* **30**(1), 13–9, 2002.

[89] Mathe C, Sagot MF, Schiex T, Rouze P. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research* **30**(19), 4103–17, 2002.

[90] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**(1), 78–94, 1997.

[91] Reese MG, Kulp D, Tammana H, Haussler D. Genie – gene finding in Drosophila melanogaster. *Genome Res* **10**(4), 529–38, 2000.

[92] Sorek R, Shemesh R, Cohen Y, Basechess O, Ast G, et al. A Non-EST-Based Method for Exon-Skipping Prediction. *Genome Res* **14**(8), 1617–23, 2004.

[93] Dror G, Sorek R, Shamir R. Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics* **21**(7), 897–901, 2004.

[94] Philipps DL, Park JW, Graveley BR. A computational and experimental approach toward a priori identification of alternatively spliced exons. *RNA* **10**(12), 1838–44, 2004.

[95] Ohler U, Shomron N, Burge CB. Recognition of unknown conserved alternatively spliced exons. *PLoS Comput Biol* **1**(2), e15, 2005.

[96] Ratsch G, Sonnenburg S, Scholkopf B. RASE: recognition of alternatively spliced exons in C.elegans. *Bioinformatics* **21 Suppl 1**, i369–i377, 2005.

[97] Liu M, Grigoriev A. Protein domains correlate strongly with exons in multiple eukaryotic genomes – evidence of exon shuffling? *Trends in Genetics* **20**(9), 399–403, 2004.

[98] Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, et al. The Ensembl automatic gene annotation system. *Genome Res* **14**(5), 942–50, 2004.

[99] Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology* **235**(5), 1501–31, 1994.

[100] Eddy SR. Profile hidden Markov models. *Bioinformatics* **14**(9), 755–63, 1998.

[101] Zhang Z, Wood WI. A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics* **19**(2), 307–8, 2003.

[102] Weinberg Z, Ruzzo WL. Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. *Bioinformatics* **20 Suppl 1**, I334–I341, 2004.

[103] Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* **13**(2), 260–269, 1967.

[104] Eddy S. Hmmer user's guide, 1998. Version 2.1.1, see http://hmmer.wustl.edu.

[105] Durbin R, Eddy S, Krogh A, Mitchison G. *Biological sequence analysis - Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK, 1998.

[106] Nembaware V, Wolfe KH, Bettoni F, Kelso J, Seoighe C. Allele-specific transcript isoforms in human. *FEBS Lett* **577**(1-2), 233–8, 2004.

[107] Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, et al. Transcription-mediated gene fusion in the human genome. *Genome Res* **16**(1), 30–6, 2006.

[108] Homma K, Kikuno RF, Nagase T, Ohara O, Nishikawa K. Alternative Splice Variants Encoding Unstable Protein Domains Exist in the Human Brain. *Journal of Molecular Biology* **343**(5), 1207–1220, 2004.

[109] Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. The Pfam protein families database. *Nucleic Acids Research* **32 Database issue**, D138–41, 2004.

[110] Brink R, Lodish HF. Tumor necrosis factor receptor (TNFR)-associated factor 2A (TRAF2A), a TRAF2 splice variant with an extended RING finger domain that inhibits TNFR2-mediated NF-kappaB activation. *Journal of Biological Chemistry* **273**(7), 4129–34, 1998.

[111] Peneff C, Ferrari P, Charrier V, Taburet Y, Monnier C, et al. Crystal structures of two human pyrophosphorylase isoforms in complexes with UDPGlc(Gal)NAc: role of the alternatively spliced insert in the enzyme oligomeric assembly and active site architecture. *EMBO J* **20**(22), 6191–202, 2001.

[112] Thanaraj TA, Stamm S. Prediction and statistical analysis of alternatively spliced exons. *Prog Mol Subcell Biol* **31**, 1–31, 2003.

[113] Galante PAF, Sakabe NJ, Kirschbaum-Slager N, de Souza SJ. Detection and evaluation of intron retention events in the human transcriptome. *RNA* **10**(5), 757–65, 2004.

[114] Wei C, Lamesch P, Arumugam M, Rosenberg J, Hu P, et al. Closing in on the C. elegans ORFeome by cloning TWINSCAN predictions. *Genome Res* **15**(4), 577–82, 2005.

[115] St Johnston D. Moving messages: the intracellular localization of mRNAs. *Nat Rev Mol Cell Biol* **6**(5), 363–75, 2005.

[116] Winkler WC, Nahvi A, Roth A, Collins JA, Breaker RR. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428**(6980), 281–6, 2004.

[117] Kozak M. Structural features in eukaryotic mRNAs that modulate the initiation of translation. *Journal of Biological Chemistry* **266**(30), 19867–70, 1991.

[118] Blow M, Futreal PA, Wooster R, Stratton MR. A survey of RNA editing in human brain. *Genome Res* **14**(12), 2379–87, 2004.

[119] Addess KJ, Basilion JP, Klausner RD, Rouault TA, Pardi A. Structure and dynamics of the iron responsive element RNA: implications for binding of the RNA by iron regulatory binding proteins. *Journal of Molecular Biology* **274**(1), 72–83, 1997.

[120] Hellen CU, Sarnow P. Internal ribosome entry sites in eukaryotic mRNA molecules. *Genes Dev* **15**(13), 1593–612, 2001.

[121] Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, et al. Characterization of mammalian selenoproteomes. *Science* **300**(5624), 1439–43, 2003.

[122] Grover A, Houlden H, Baker M, Adamson J, Lewis J, et al. 5' splice site mutations in tau associated with the inherited dementia FTDP-17 affect a stem-loop structure that regulates alternative splicing of exon 10. *Journal of Biological Chemistry* **274**(21), 15134–43, 1999.

[123] Muh SJ, Hovhannisyan RH, Carstens RP. A Non-sequence-specific double-stranded RNA structural element regulates splicing of two mutually exclusive exons of fibroblast growth factor receptor 2 (FGFR2). *Journal of Biological Chemistry* **277**(51), 50143–54, 2002.

[124] Graveley BR. Mutually Exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures. *Cell* **123**(1), 65–73, 2005.

[125] Lian Y, Garner HR. Evidence for the regulation of alternative splicing via complementary DNA sequence repeats. *Bioinformatics* **21**(8), 1358–64, 2005.

[126] Nasim FUH, Hutchison S, Cordeau M, Chabot B. High-affinity hnRNP A1 binding sites and duplex-forming inverted repeats have similar effects on 5' splice site selection in support of a common looping out and repression mechanism. *RNA* **8**(8), 1078–89, 2002.

[127] Oberstrass FC, Auweter SD, Erat M, Hargous Y, Henning A, et al. Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science* **309**(5743), 2054–7, 2005.

[128] Matlin AJ, Clark F, Smith CWJ. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* **6**(5), 386–98, 2005.

[129] Buckanovich RJ, Darnell RB. The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. *Mol Cell Biol* **17**(6), 3194–201, 1997.

[130] Shi H, Hoffman BE, Lis JT. A specific RNA hairpin loop structure binds the RNA recognition motifs of the Drosophila SR protein B52. *Mol Cell Biol* **17**(5), 2649–57, 1997.

[131] Damgaard CK, Tange TO, Kjems J. hnRNP A1 controls HIV-1 mRNA splicing through cooperative binding to intron and exon splicing silencers in the context of a conserved secondary structure. *RNA* **8**(11), 1401–15, 2002.

[132] Meisner NC, Hackermuller J, Uhl V, Aszodi A, Jaritz M, et al. mRNA openers and closers: modulating AU-rich element-controlled mRNA stability by a molecular switch in mRNA secondary structure. *Chembiochem* **5**(10), 1432–47, 2004.

[133] Buratti E, Muro AF, Giombi M, Gherbassi D, Iaconcig A, et al. RNA folding affects the recruitment of SR proteins by mouse and human polypurinic enhancer elements in the fibronectin EDA exon. *Mol Cell Biol* **24**(3), 1387–400, 2004.

[134] Goguel V, Wang Y, Rosbash M. Short artificial hairpins sequester splicing signals and inhibit yeast pre-mRNA splicing. *Mol Cell Biol* **13**(11), 6841–8, 1993.

[135] Solnick D, Lee SI. Amount of RNA secondary structure required to induce an alternative splice. *Mol Cell Biol* **7**(9), 3194–8, 1987.

[136] Guil S, Gattoni R, Carrascal M, Abian J, Stevenin J, et al. Roles of hnRNP A1, SR proteins, and p68 helicase in c-H-ras alternative splicing regulation. *Mol Cell Biol* **23**(8), 2927–41, 2003.

[137] Jankowsky E, Gross CH, Shuman S, Pyle AM. Active disruption of an RNA-protein interaction by a DExH/D RNA helicase. *Science* **291**(5501), 121–5, 2001.

[138] Reenan RA, Hanrahan CJ, Barry G. The mle(napts) RNA helicase mutation in drosophila results in a splicing catastrophe of the para Na+ channel transcript in a region of RNA editing. *Neuron* **25**(1), 139–49, 2000.

[139] Honig A, Auboeuf D, Parker MM, O'Malley BW, Berget SM. Regulation of alternative splicing by the ATP-dependent DEAD-box RNA helicase p72. *Mol Cell Biol* **22**(16), 5698–707, 2002.

[140] Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research* **9**(1), 133–48, 1981.

[141] Schultes EA, Bartel DP. One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* **289**(5478), 448–52, 2000.

[142] McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**(6-7), 1105–19, 1990.

[143] Muckstein U, Tafer H, Hackermuller J, Bernhart SH, Stadler PF, et al. Thermodynamics of RNA-RNA binding. *Bioinformatics* **22**(10), 1177–82, 2006.

[144] Hackermuller J, Meisner NC, Auer M, Jaritz M, Stadler PF. The effect of RNA secondary structures on RNA-ligand binding and the modifier RNA mechanism: a quantitative model. *Gene* **345**(1), 3–12, 2005.

[145] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, et al. Fast folding and comparison of RNA secondary structures. *Monatshefte Chemie* **125**, 167–188, 1994.

[146] Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, et al. ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Research* **34**(Database issue), D46–55, 2006.

[147] Flamm C, Fontana W, Hofacker IL, Schuster P. RNA folding at elementary step resolution. *RNA* **6**(3), 325–38, 2000.

[148] Eperon LP, Graham IR, Griffiths AD, Eperon IC. Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? *Cell* **54**(3), 393–401, 1988.

[149] Chan RC, Black DL. The polypyrimidine tract binding protein binds upstream of neural cell-specific c-src exon N1 to repress the splicing of the intron downstream. *Mol Cell Biol* **17**(8), 4667–76, 1997.

[150] Katz L, Burge CB. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res* **13**(9), 2042–51, 2003.

[151] Liu HX, Zhang M, Krainer AR. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* **12**(13), 1998–2012, 1998.

[152] Liu HX, Chew SL, Cartegni L, Zhang MQ, Krainer AR. Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol Cell Biol* **20**(3), 1063–71, 2000.

[153] Cavaloc Y, Bourgeois CF, Kister L, Stevenin J. The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA* **5**(3), 468–83, 1999.

[154] Bailey T, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28–36, 1994.

[155] Bailey TL, Elkan C. Unsupervised learning of multiple motifs in biopolymers using EM. *Machine Learning* **21**(1-2), 51–80, 1995.

[156] Bailey TL. *Discovering motifs in DNA and protein sequences: The approximate common substring problem.* Ph.D. thesis, University of California, San Diego, 1995.

[157] Lawrence C, Altschul S, Boguski M, Liu J, Neuwald A, et al. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science* **262**(5131), 208–14, 1993.

[158] Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, et al. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Research* **29**(22), 4724–35, 2001.

[159] Pavesi G, Mauri G, Stefani M, Pesole G. RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. *Nucleic Acids Research* **32**(10), 3258–69, 2004.

[160] Liu J, Wang JTL, Hu J, Tian B. A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC Bioinformatics* **6**(1), 89, 2005.

[161] Yao Z, Weinberg Z, Ruzzo WL. CMfinder – a covariance model based RNA motif finding algorithm. *Bioinformatics* **22**(4), 445–52, 2006.

[162] Klein RJ, Eddy SR. RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics* **4**(1), 44, 2003.

[163] Siebert S, Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* **21**(16), 3352–9, 2005.

[164] Backofen R, Will S. Local sequence-structure motifs in RNA. *Journal of Bioinformatics and Computational Biology (JBCB)* **2**(4), 681–698, 2004.

[165] Thisted T, Lyakhov DL, Liebhaber SA. Optimized RNA targets of two closely related triple KH domain proteins, heterogeneous nuclear ribonucleoprotein K and alphaCP-2KL, suggest distinct modes of RNA recognition. *Journal of Biological Chemistry* **276**(20), 17484–96, 2001.

[166] Hori T, Taguchi Y, Uesugi S, Kurihara Y. The RNA ligands for mouse proline-rich RNA-binding protein (mouse Prrp) contain two consensus sequences in separate loop structure. *Nucleic Acids Research* **33**(1), 190–200, 2005.

[167] Dempster A, Laird N, Rubin D. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society* **39**, 1 – 38, 1977.

[168] Xing Y, Yu T, Wu YN, Roy M, Kim J, et al. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Research* **34**(10), 3150–60, 2006.

[169] Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* **33 Database Issue**, D121–4, 2005.

[170] Hentze MW, Kuhn LC. Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc. Natl. Acad. Sci. USA* **93**(16), 8175–82, 1996.

[171] Varani L, Gunderson SI, Mattaj IW, Kay LE, Neuhaus D, et al. The NMR structure of the 38 kDa U1A protein - PIE RNA complex reveals the basis of cooperativity in regulation of polyadenylation by human U1A protein. *Nat Struct Biol* **7**(4), 329–35, 2000.

[172] Richter S, Cao H, Rana TM. Specific HIV-1 TAR RNA loop sequence and functional groups are required for human cyclin T1-Tat-TAR ternary complex formation. *Biochemistry* **41**(20), 6391–7, 2002.

[173] Richter S, Ping YH, Rana TM. TAR RNA loop: a scaffold for the assembly of a regulatory switch in HIV replication. *Proc. Natl. Acad. Sci. USA* **99**(12), 7928–33, 2002.

[174] Putland RA, Sassinis TA, Harvey JS, Diamond P, Coles LS, et al. RNA destabilization by the granulocyte colony-stimulating factor stem-loop destabilizing element involves a single stem-loop that promotes deadenylation. *Mol Cell Biol* **22**(6), 1664–73, 2002.

[175] Selvakumar M, Helfman DM. Exonic splicing enhancers contribute to the use of both 3' and 5' splice site usage of rat beta-tropomyosin pre-mRNA. *RNA* **5**(3), 378–94, 1999.

[176] Pagani F, Baralle FE. Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet* **5**(5), 389–96, 2004.

[177] Shen LX, Basilion JP, Stanton VPJ. Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc. Natl. Acad. Sci. USA* **96**(14), 7871–6, 1999.

[178] Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, et al. Systematic identification and analysis of exonic splicing silencers. *Cell* **119**(6), 831–45, 2004.

[179] Garcia-Blanco MA, Baraniak AP, Lasda EL. Alternative splicing in disease and therapy. *Nat Biotechnol* **22**(5), 535–46, 2004.

[180] Pedersen AG, Baldi P, Chauvin Y, Brunak S. The biology of eukaryotic promoter prediction–a review. *Comput Chem* **23**(3-4), 191–207, 1999.

[181] Narlikar L, Gordan R, Ohler U, Hartemink AJ. Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics* **22**(14), e384–92, 2006.

[182] Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, et al. An alternative-exon database and its statistical analysis. *DNA Cell Biol* **19**(12), 739–56, 2000.

[183] Liu S, Altman RB. Large scale study of protein domain distribution in the context of alternative splicing. *Nucleic Acids Research* **31**(16), 4828–35, 2003.

[184] Xing Y, Lee CJ. Protein Modularity of Alternatively Spliced Exons Is Associated with Tissue-Specific Regulation of Alternative Splicing. *PLoS Genet* **1**(3), e34, 2005.

[185] Xing Y, Lee C. Alternative splicing and RNA selection pressure - evolutionary consequences for eukaryotic genomes. *Nat Rev Genet* **7**(7), 499–509, 2006.

[186] Yee D, Lebovic GS, Marcus RR, Rosen N. Identification of an alternate type I insulin-like growth factor receptor beta subunit mRNA transcript. *Journal of Biological Chemistry* **264**(36), 21439–41, 1989.

[187] McKenzie AN, Culpepper JA, de Waal Malefyt R, Briere F, Punnonen J, et al. Interleukin 13, a T-cell-derived cytokine that regulates human monocyte and B-cell function. *Proc. Natl. Acad. Sci. USA* **90**(8), 3735–9, 1993.

[188] Condorelli G, Bueno R, Smith RJ. Two alternatively spliced forms of the human insulin-like growth factor I receptor have distinct biological activities and internalization kinetics. *Journal of Biological Chemistry* **269**(11), 8510–6, 1994.

[189] Vogan KJ, Underhill DA, Gros P. An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity. *Mol Cell Biol* **16**(12), 6677–86, 1996.

[190] Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, et al. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res* **13**(6B), 1290–300, 2003.

[191] Pertea M, Lin X, Salzberg SL. Genesplicer: a new computational method for splice site prediction. *Nucleic Acids Research* **29**(5), 1185–90, 2001.

[192] Long M, Deutsch M. Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. *Mol Biol Evol* **16**(11), 1528–34, 1999.

[193] Ma B, Elkayam T, Wolfson H, Nussinov R. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci. USA* **100**(10), 5772–7, 2003.

[194] Blasband AJ, Rogers KT, Chen XR, Azizkhan JC, Lee DC. Characterization of the rat transforming growth factor alpha gene and identification of promoter sequences. *Mol Cell Biol* **10**(5), 2111–21, 1990.

[195] Hosoda H, Kojima M, Matsuo H, Kangawa K. Purification and characterization of rat des-Gln14-Ghrelin, a second endogenous ligand for the growth hormone secretagogue receptor. *Journal of Biological Chemistry* **275**(29), 21995–2000, 2000.

[196] Rogina B, Upholt WB. The chicken homeobox gene Hoxd-11 encodes two alternatively spliced RNA species. *Biochem Mol Biol Int* **35**(4), 825–31, 1995.

[197] Takeuchi M, Fujisawa H. New alternatively spliced variants of calmodulin-dependent protein kinase II from rabbit liver. *Gene* **221**(1), 107–15, 1998.

[198] Ferranti P, Lilla S, Chianese L, Addeo F. Alternative nonallelic deletion is constitutive of ruminant alpha(s1)-casein. *J Protein Chem* **18**(5), 595–602, 1999.

[199] Li L, Howe GA. Alternative splicing of prosystemin pre-mRNA produces two isoforms that are active as signals in the wound response pathway. *Plant Mol Biol* **46**(4), 409–19, 2001.

[200] Zhang H, Blumenthal T. Functional analysis of an intron 3' splice site in Caenorhabditis elegans. *RNA* **2**(4), 380–8, 1996.

[201] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**(1), 308–11, 2001.

[202] Tadokoro K, Yamazaki-Inoue M, Tachibana M, Fujishiro M, Nagao K, et al. Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: the case of Gln in DRPLA affects subcellular localization of the products. *J Hum Genet* **50**(8), 382–94, 2005.

[203] Watanabe H, Fujiyama A, Hattori M, Taylor TD, Toyoda A, et al. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429**(6990), 382–8, 2004.

[204] Maugeri A, van Driel MA, van de Pol DJ, Klevering BJ, van Haren FJ, et al. The 2588G–>C mutation in the ABCR gene is a mild frequent founder mutation in the Western European population and allows the classification of ABCR mutations in patients with Stargardt disease. *Am J Hum Genet* **64**(4), 1024–35, 1999.

[205] Burge CB, Padgett RA, Sharp PA. Evolutionary fates and origins of U12-type introns. *Mol Cell* **2**(6), 773–85, 1998.

[206] Patel AA, Steitz JA. Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol* **4**(12), 960–70, 2003.

[207] Dietrich RC, Peris MJ, Seyboldt AS, Padgett RA. Role of the 3' splice site in U12-dependent intron splicing. *Mol Cell Biol* **21**(6), 1942–52, 2001.

[208] Dietrich RC, Fuller JD, Padgett RA. A mutational analysis of U12-dependent splice site dinucleotides. *RNA* **11**(9), 1430–40, 2005.

[209] Hastings ML, Resta N, Traum D, Stella A, Guanti G, et al. An LKB1 AT-AC intron mutation causes Peutz-Jeghers syndrome via splicing at noncanonical cryptic splice sites. *Nat Struct Mol Biol* **12**(1), 54–9, 2005.

[210] Levine A, Durbin R. A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Research* **29**(19), 4006–13, 2001.

[211] Wu S, Romfo CM, Nilsen TW, Green MR. Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature* **402**(6763), 832–5, 1999.

[212] Zhuang Y, Weiner AM. A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell* **46**(6), 827–35, 1986.

[213] Barbaux S, Niaudet P, Gubler MC, Grunfeld JP, Jaubert F, et al. Donor splice-site mutations in WT1 are responsible for Frasier syndrome. *Nat Genet* **17**(4), 467–70, 1997.

[214] Valentonyte R, Hampe J, Huse K, Rosenstiel P, Albrecht M, et al. Sarcoidosis is associated with a truncating splice site mutation in BTNL2. *Nat Genet* **37**(4), 357–64, 2005.

[215] Abril JF, Castelo R, Guigo R. Comparison of splice sites in mammals and chicken. *Genome Res* **15**(1), 111–9, 2005.

[216] Carmel I, Tal S, Vig I, Ast G. Comparative analysis detects dependencies among the 5' splice-site positions. *RNA* **10**(5), 828–40, 2004.

[217] Roca X, Sachidanandam R, Krainer AR. Determinants of the inherent strength of human 5' splice sites. *RNA* **11**(5), 683–98, 2005.

[218] Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology* **11**(2-3), 377–94, 2004.

[219] McCullough AJ, Berget SM. An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5' splice sites. *Mol Cell Biol* **20**(24), 9225–35, 2000.

[220] McCullough AJ, Berget SM. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol Cell Biol* **17**(8), 4562–71, 1997.

[221] Akerman M, Mandel-Gutfreund Y. Alternative splicing regulation at tandem 3' splice sites. *Nucleic Acids Research* **34**(1), 23–31, 2006.

[222] Lorkovic ZJ, Lehner R, Forstner C, Barta A. Evolutionary conservation of minor U12-type spliceosome between plants and humans. *RNA* **11**(7), 1095–107, 2005.

[223] Joyce-Brady M, Jean JC, Hughey RP. gamma -glutamyltransferase and its isoform mediate an endoplasmic reticulum stress response. *Journal of Biological Chemistry* **276**(12), 9468–77, 2001.

[224] Ray DW, Davis JR, White A, Clark AJ. Glucocorticoid receptor structure and function in glucocorticoid-resistant small cell lung carcinoma cells. *Cancer Res* **56**(14), 3276–80, 1996.

[225] Rivers C, Levy A, Hancock J, Lightman S, Norman M. Insertion of an amino acid in the DNA-binding domain of the glucocorticoid receptor as a result of alternative splicing. *J Clin Endocrinol Metab* **84**(11), 4283–6, 1999.

[226] Ast G. How did alternative splicing evolve? *Nat Rev Genet* **5**(10), 773–82, 2004.

[227] Tsai KW, Lin WC. Quantitative analysis of wobble splicing indicates that it is not tissue specific. *Genomics* 2006.

[228] Chern TM, van Nimwegen E, Kai C, Kawai J, Carninci P, et al. A simple physical model predicts small exon length variations. *PLoS Genet* **2**(4), e45, 2006.

[229] Fedoroff N, Fontana W. Genetic networks. Small numbers of big molecules. *Science* **297**(5584), 1129–31, 2002.

[230] Rao CV, Wolf DM, Arkin AP. Control, exploitation and tolerance of intracellular noise. *Nature* **420**(6912), 231–7, 2002.

[231] Neves G, Zucker J, Daly M, Chess A. Stochastic yet biased expression of multiple Dscam splice variants by individual cells. *Nat Genet* **36**(3), 240–6, 2004.

[232] Echard A, Opdam FJ, de Leeuw HJ, Jollivet F, Savelkoul P, et al. Alternative splicing of the human Rab6A gene generates two close but functionally different isoforms. *Mol Biol Cell* **11**(11), 3819–33, 2000.

[233] Markovtsov V, Nikolic JM, Goldman JA, Turck CW, Chou MY, et al. Cooperative assembly of an hnRNP complex induced by a tissue-specific homolog of polypyrimidine tract binding protein. *Mol Cell Biol* **20**(20), 7463–79, 2000.

[234] Chen S, Anderson K, Moore MJ. Evidence for a linear search in bimolecular 3' splice site AG selection. *Proc. Natl. Acad. Sci. USA* **97**(2), 593–8, 2000.

[235] Pacheco TR, Gomes AQ, Barbosa-Morais NL, Benes V, Ansorge W, et al. Diversity of vertebrate splicing factor U2AF35: identification of alternatively spliced U2AF1 mRNAS. *Journal of Biological Chemistry* **279**(26), 27039–49, 2004.

[236] Brackenridge S, Wilkie AOM, Screaton GR. Efficient use of a 'dead-end' GA 5' splice site in the human fibroblast growth factor receptor genes. *EMBO J* **22**(7), 1620–31, 2003.

[237] Crispino JD, Blencowe BJ, Sharp PA. Complementation by SR proteins of pre-mRNA splicing reactions depleted of U1 snRNP. *Science* **265**(5180), 1866–9, 1994.

[238] Crispino JD, Sharp PA. A U6 snRNA:pre-mRNA interaction can be rate-limiting for U1-independent splicing. *Genes Dev* **9**(18), 2314–23, 1995.

[239] Hastings ML, Krainer AR. Pre-mRNA splicing in the new millennium. *Current Opinion in Cell Biology* **13**(3), 302–9, 2001.

[240] Chen JY, Stands L, Staley JP, Jackups RRJ, Latus LJ, et al. Specific alterations of U1-C protein or U1 small nuclear RNA can eliminate the requirement of Prp28p, an essential DEAD box splicing factor. *Mol Cell* **7**(1), 227–32, 2001.

[241] Karinch AM, deMello DE, Floros J. Effect of genotype on the levels of surfactant protein A mRNA and on the SP-A2 splice variants in adult humans. *Biochem J* **321**, 39–47, 1997.

[242] Kazan K. Alternative splicing and proteome diversity in plants: the tip of the iceberg has just emerged. *Trends Plant Sci* **8**(10), 468–71, 2003.

[243] Iida K, Seki M, Sakurai T, Satou M, Akiyama K, et al. Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences. *Nucleic Acids Research* **32**(17), 5096–103, 2004.

[244] Eddy SR. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol* **3**(1), e10, 2005.

[245] Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, et al. The UCSC Known Genes. *Bioinformatics* **22**(9), 1036–46, 2006.

[246] Eyras E, Caccamo M, Curwen V, Clamp M. ESTGenes: alternative splicing from ESTs in Ensembl. *Genome Res* **14**(5), 976–87, 2004.

[247] Lai CH, Hu LY, Lin Wc. Single amino-acid InDel variants generated by alternative tandem splice-donor and -acceptor selection. *Biochem Biophys Res Commun* **342**(1), 197–205, 2006.

Abbreviations

aa	amino acid
cDNA	complementary DNA
CDS	protein coding sequence
CPU	central processing unit
dbEST	database for Expressed Sequence Tags
dbSNP	database for Single Nucleotide Polymorphisms
DP	dynamic programming
dsMotif	double-stranded motif
ED	energy difference
EF	expected fraction of unpaired bases
EM	expectation maximization
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
EST	expressed sequence tag
HIV	human immunodeficiency virus
HMM	hidden Markov model
hnRNP	heterogeneous nuclear ribonucleoprotein
indel	insertion/deletion
IRE	iron responsive element
IRES	internal ribosome entry site
ISE	intronic splicing enhancer
ISS	intronic splicing silencer
ML	maximum likelihood
mRNA	messenger RNA
NMD	nonsense-mediated mRNA decay
nt	nucleotides
OMIM	online mendelian inheritance in man
OOPS	One motif Occurrence Per Sequence model
ORF	open reading frame
PDB	protein data bank
peptide cassette exons	an exon with a length that is a multiple of 3 nt and that do not encode an in-frame stop codon
Pfam	protein domain family
PIE	polyadenylation inhibition element
pI	isoelectric point
PSPM	position-specific probability matrix
PTC	premature termination codon
PU	probability that a substring is unpaired
Rfam	RNA family
RT-PCR	reverse transcription polymerase chain reaction
SELEX	SElection of Ligands by EXponential enrichment
SLDE	stem-loop destabilizing element
SNP	single nucleotide polymorphism
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein particle
SQL	structured query language
SR protein	serine/arginine rich protein
ssMotif	single-stranded motif

TAR	trans-activation response
TCM	Two-Component Mixture model
TM	transmembrane
UCSC Genome Browser	University of California Santa Cruz, Genome Browser
UTR	untranslated region
ZOOPS	Zero or One motif Occurrence Per Sequence model

Used IUPAC nucleotide codes

B	C, G, or T/U
H	A, C, or T/U
R	A or G
Y	C or T

Statistical tests

We used the following standard statistical tests in this thesis.

- The t-test was used to compare to means.
- Fisher´s exact test was used to test independence in a 2x2 contingency table.
- The χ^2 test was used to test independence in a contingency table larger than 2x2.