Computational Analysis of Alternative Splicing in Human and Mice

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

seit 1558

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät
der Friedrich-Schiller-Universität Jena

von MSc Ralf H. Bortfeldt
geboren am 16. Februar 1977 **in** Berlin

Die vorliegende Arbeit wurde am Lehrstuhl für Bioinformatik der biologisch-pharmazeutischen Fakultät der Friedrich Schiller Universität unter Leitung von Prof. Dr. Stefan Schuster angefertigt.

Gutachter:

1. Prof. Dr. Stefan Schuster, Friedrich Schiller Universität Jena
2. PD Dr. Matthias Platzer, FLI - Leibniz-Institut für Altersforschung Jena
3. PD Dr. Jürgen Kleffe, Institut für Molekularbiologie und BioInformatik, Charité Campus Benjamin Franklin, Berlin

Tag der öffentlichen Verteidigung: 04. Mai 2009

Summary

The presented dissertation focuses on the computational analysis of the molecular biological mechanism called Alternative Splicing and integrates the topics *„Prediction and validation of alternative splice forms“*, *„Characterization and functional impact of short 5′ splice site (ss) variations“* and *„Petri Net modeling of the spliceosomal assembly pathway“*. In the first part different features were analyzed in their potential to distinguish two classes of splice sites, alternative and reference splice sites. These were computationally derived earlier from transcript alignments of „Expressed Sequence Tags“ (ESTs) and mRNA and stored in the EASED database.

A central question was to what extent different features contribute to discrimination of both types of splice sites. In particular the features „splice site score“, „transcript coverage“ and „splicing factor binding sites“ in vicinity of the considered splice sites were investigated. The applied scoring system describes the agreement of the considered splice site to an entropy model of human splice sites and thus makes it possible to capture for example the complementarity of the donor site to the recognizing motif within the small nuclear RNA (snRNA) of the spliceosomal U1 complex. For comparison reference splice sites were collected from transcripts that were not predictive of alternative splicing, and subjected to the same analyses. The results showed that alternative splice (AS) forms identified by the EASED pipeline were in two of three analyzed features statistically separable.

Alternative splice sites could especially be discriminated from constitutive splice sites by a lower splice site score and an increased presence of splicing factor binding motifs in the vicinity of alternative splice sites. Additionally, a positive correlation within the compared classes of alternative and reference splice forms could be observed, in the dependency between transcript coverage and the splice site score. This dependency shows that the confidence in the AS form predicting

process raises with the number of available transcripts and similar as it can be observed within the set of reference splice forms.
The results have been published in the book series *„Modelling and Simulation in Science, Engineering and Technology“* in Birkhäuser / Springer (1).

With the ongoing development of prediction and annotation techniques for alternative splice forms, distinct splice patterns have been discovered and analyzed in more detail. Most comprehensively the class of skipped exons (SE) has been studied over the past decade, followed by alternative $5'$ and $3'$ splice site exons. Nevertheless only recently, the interesting concept termed *subtle* alternative splicing has been introduced based on the findings of a special pattern of human alternative acceptor splice forms of NAGNAG type (2). In conjunction with this finding, the second part of the presented thesis focuses on a subtle alternative splicing pattern with the motif GYNNGY at the donor site, which occurs almost as frequently as splice sites with the NAGNAG motif. In contrast, splicing at GYNNGY tandem donors implies an mRNA variation, which is not a multiple of the codon length and hence indicates a more intricate role in mRNA turnover. For example, changing the reading frame may result in a premature stop codon in the matured transcript, potentially draining the mRNA from the pool of expressed transcripts. The motif GYNNGY itself also serves as recognition site for U1 and U5 snRNA motifs, hence the splicing machinery has to distinguish between these two overlapping donor sites.
The analysis of such small splicing variations requires a dataset of higher resolution and sensitivity near exon intron junctions. To this end the Hollywood database provided a thoroughly prepared dataset of alternative splice patterns (3). Among the alternative exons, overlapping donor splice sites constituted more than 80% of the total set of predicted alternative $5'$ splice sites. Among these, 41% of the alternative donor sites showed the motif GYNNGY, thus being the most frequently tandem donor variant. *In silico* and experimental characterization of alternative splice variants of this GYNNGY type showed distinct properties in splice site strength, sequence conservation, presence of splicing enhancer motifs and functional consequences according to utilization levels of the alternative overlapping donor splice sites. It turned out that those AS forms which predominantly utilized the downstream (proximal) donor over the rare upstream (distal) donor (type-I) were clearly distinguishable from constitutive exons. In contrast, the group of AS forms predominantly utilizing

the upstream donor over the alternative downstream donor (type-II), was more similar to constitutive exons with respect to the analyzed features. Summarizing, the results point to the existence of a biological mechanism, which involves frame shifted alternative splice variants, continuously produced at low levels for the cause of either a directed nonsense-mediated mRNA decay or truncated proteins. The mere observation of this phenomenon questions the absolute effectiveness of the NMD mechanism and connetcs to new theories about the processing of primary transcripts beyond the purpose of coding for functional proteins.
This work is published in the journal *BMC Genomics* (4).

In the last part of this thesis, a computational basis for investigating the concurrent processes of spliceosome assembly was established. The spliceosome is the catalytically active key component of the splicing process, whereupon a functional macro complex is assembled from proteins and snRNA via a multitude of signaling reactions within the nucleus of eukaryotic cells. Petri net theory was applied to model the reactions and interactions of the spliceosome assembly pathway. Initially a high degree of biological knowledge has to be screened in order to draft a Petri Net model but the resulting graph structure allows an easier validation and exploitation of the accumulating experimental data. Calculation of semi positive t-invariants serves as a means of validating existing - and uncover new signal flows within the spliceosomal assembly process. The biologically and mathematically validated model proves as being suitable for simulation and classification of signaling pathways. A set of 71 invariant pathways could be delineated from a network of reaction, and being subjected to further structural classification by means of clustering and decomposition into partial pathways with high degree of shared information flow. This approach serves well to integrate experimental data from literature and structure molecular biological information. As consequence, Petri nets help to uncover weaknesses of existing partial models, to uncover inconsistent hypothesizing and finally to generate new insights into molecular biological mechanism.

Zusammenfassung

Die vorliegende Dissertation konzentriert sich auf bioinformatische Analysen zum molekularbiologischen Mechanismus des *Alternativen Spleißens* und verbindet die Themen *„Vorhersage und Bewertung von alternativen Spleißformen“*, *„Charakterisierung und funktionelle Auswirkungen von kleinen Variationen des 5′ Spleißsignals“* und *„Modellierung der Bildung des spleißosomalen Proteinkomplexes mittels Petri-Netzen“*.

Im ersten Teil der Arbeit wurden verschiedene Merkmale auf ihr Potential hin analysiert, zwei Typen von Spleißstellen, alternative und Referenzspleißstellen, zu unterscheiden. Diese wurden zuvor *in silcio* aus Transkript Alignments von „Expressed Sequence Tags“ (ESTs) und mRNA abgeleitet und mit zusätzlichen Informationen in der EASED-Datenbank gespeichert. Eine zentrale Frage war, in welchem Maße verschiedene Merkmale zur Diskriminierung beider Klassen von Spleißereignissen beitragen. Besonders die Merkmale „Spleißsignalstärke“, „Transkripthäufigkeit“ und „Spleißfaktor-Bindstellen“ im Bereich der betrachtetenSpleißstellen wurden analysiert. Das verwendete Scoringmaß für die Spleißstellen beschreibt die Übereinstimmung der betrachteten Spleißstelle zu einem Entropiemodell von humanen Spleißstellen und ermöglicht so z.B. die Komplementarität des Donor-Motivs zur Erkennungssequenz innerhalb der *small nuclear RNA* (snRNA) der spleißosomalen Untereinheit U1 zu bewerten. Zum Vergleich wurden Referenzspleißstellen aus Transkripten zusammengestellt, welche keine Spleißvariationen zeigten, und auf die gleichen Merkmale untersucht. Die Ergebnisse belegten, dass durch den EASED Algorithmus generierte AS Formen in zwei der drei untersuchten Merkmale von den Referenzspleißformen unterschieden werden konnten. Alternative Spleißstellen liessen sich dabei von konstitutiven Spleißstellen besonders durch einen niedrigeren Spleißstellen-Score und vermehrtes Auftreten von Bindemotiven für SR Proteine in der näheren Umgebung der Spleißstellen abgrenzen. Zusätzlich konnte eine positive Korre-

lation in beiden Vergleichsklassen beobachtet werden, zwischen der Häufigkeit von gespleißten Transkripten und dem Spleißstellen-Score. Diese Abhängigkeit zeigt, dass das Vertrauen in die vorhergesagten alternativen Spleißformen mit der Anzahl verfügbarer Transkripte steigt, und zwar in gleicher Weise, wie es bei konstitutiven Transkripten beobachtet werden kann.
Die Ergebnisse dieser Studie wurden in der Buchserie *„Modelling and Simulation in Science, Engineering and Technology“* des Birkhäuser / Springer Verlags (1) veröffentlicht.

Mit der voranschreitenden Entwicklung von Techniken zur Vorhersage und Annotation von alternativen Spleißformen wurden diverse Spleißmuster entdeckt und detaillierter untersucht. Am umfangreichsten wurde über die letzten Jahre hinweg Spleißmuster untersucht, welche vollständige Exons in Transkripten übergehen oder einfügen („Kassettenexons“). Weitere gut untersuchte Beispiele existieren für Spleißformen, welche Exons an der angrenzenden $5'$ Spleißstelle (Donor) oder $3'$ Spleißstelle (Akzeptor) variieren. Erst kürzlich wurde das interessante Konzept *subtilen* alternativen Spleißens vorgestellt. Dieses Konzept basiert auf Beobachtungen von Spleißvorgängen an sogenannten Tandem-Akzeptor Spleißsignalen mit dem Motiv NAGNAG, welche im Humanentrankriptom aber auch in Transkripten anderer Eukaryoten gefunden wurde (2). In Anknüpfung an diese Entdeckung befasst sich der zweite Teil der vorliegenden Arbeit mit einem weiteren subtilen alternativen Spleißmuster, welches ähnlich häufig, jedoch an Donor-Spleißsignalen auftritt. Im Unterschied zu der drei Nukleotide umfassenden Variation am Akzeptor lässt die vier Nukleotide betreffende Donorvariation auf eine komplexe Rolle im RNA-Reifungsprozess schließen. Durch die Verschiebung des Leserasters kann z.B. ein deutlich häufigeres Auftreten von vorzeitigen Stopcodons in reifen mRNAs beobachtet werden. Das Spleißsignalmotiv GYNNGY selbst dient dabei als Erkennungssequenz für die zwei wichtigen spleißosomalen Ribonukleoproteinkomplexe U1 und U5, deren Bindung durch die zwei überlappenden Motive beeinflusst wird. Die Analyse dieses speziellen Spleißmusters erforderte einen Datensatz mit einer nukleotidgenauen Auflösung der alternativen Spleißstellen, um die kurzen Variationen charakterisieren zu können. Ein entsprechender Datensatz wurde aus der HOLLYWOOD-Datenbank (3) gewonnen, welche im Gegensatz zur EASED Datenbank mehr Spleißmuster klassifiziert und auch Transkriptvariationen von weniger als zehn 10 Nukleotiden berücksichtigt. Die Daten zeigten, dass kurze Variationen

mit überlappenden Donormotiven mehr als 80% aller nachweisbaren alternativen Donor Spleißereignisse ausmachen. Dabei tritt unter diesen überlappenden Donoren mit 41% am häufigsten das Motiv GYNNGY auf. Mittels *in silico* Analysen und experimenteller Validierung wurde die Plausibilität dieses subtilen Spleißmusters bestätigt. Sequenzanalysen zeigten ausserdem ausgeprägte Charakteristika bezüglich des Spleißstellen-Scores (Nukleotidzusammensetzung), Konservierung, Vorliegen von potentiellen Bindemotiven für Spleißfaktoren sowie funktionelle Auswirkungen in Abhängigkeit der Nachweishäufigkeit beider überlappender Donoren in gespleißten Transkripten. Ein Ergebnis deutet darauf hin, dass besonders die GYNNGY Spleißformen, welche seltener am ersten (distalen) Donor gegenüber dem zweiten (proximalen) Donor gespleißt werden, in allen oben genannten Merkmalen deutlich von konstitutiven Spleißformen (mit ähnlichem Donor Motiv) unterschieden werden können. Gleiches konnte für den umgekehrten Fall, bei dem die erste Spleißstelle häufiger gespleißt wird als die zweite Spleißstelle, nur bedingt nachgewiesen werden.
Zusammenfassend deuten die Ergebnisse auf einen biologischen Mechanismus hin, welcher alternative Spleißvarianten erzeugt, die das Leseraster verschieben und kontinuierlich in geringen Mengen generiert werden, entweder zum gezielten Auslösen des NMD Mechanismus oder zur Erzeugung verkürzter Proteine. Die alleinige Beobachtung dieses Phänomens stellt die uneingeschränkte Wirksamkeit des „Nonsense Mediated RNA Decay“ (NMD) Mechanismus in Frage und knüpft an neue Theorien über die Verarbeitung von Primärtranskripten an, jenseits ihrer klassischen Funktion funktionelle Proteine zu kodieren. Diese Arbeit wurde im Journal *BMC Genomics* veröffentlicht (4).

Das Spleißosom ist die katalytisch aktive Schlüsselkomponente des Spleißens, wobei durch eine Vielzahl von Signalreaktionen ein funktioneller Makrokomplex aus Proteinen und kleinen nukleären RNAs an Primärtranskripten im Zellkern von Eukaryoten aufgebaut wird. Die spleißosomalen Signalreaktionen wurden aus experimenteller Fachliteratur zusammengestellt und mit Hilfe der Petri Netz Theorie modelliert. Im Gegensatz zu bisherigen Anwendungen von Petri Netzen zur Analyse metabolischer Netzwerke, wurde in diesem Ansatz, bedingt durch das Fehlen stöchiometrischer Daten, ein binärer Entscheidungsbaum modelliert. Dabei entsprechen Token auf den Knoten des Netzwerks einem Informationsgehalt oder Zustand, der, wenn gegeben, eine Reaktion ermöglicht. Durch diese Interpretation wird es möglich, Kanten im Netzwerk konstant mit einem Gewicht

von eins zu modelieren. Im Gegensatz zu ungerichteten Protein-Protein Wechselwirkungsnetzwerken, erforderte dieser Ansatz die Einordnung von Protein-Protein sowieo Protein-RNA Wechselwirkungen in eine Abfolge zeitlich geordneter Reaktionen.

Die Darstellung des spleißsomalen Signalwegs als Petri Netz eröffnet unter der Annahme eines *Steady State Systems* die Berechnung von mimimalen, semipostivien T-Invarianten, welche zur Validierung des Modells herangezogen werden können. Die darauf aufbauenden Analysemöglichkeiten sowie die leichte Erweiterbarkeit des Modells rechtfertigen den Modellierungsansatz im Hinblick auf die stetig wachsende Menge experimenteller Daten. Die T-Invarianten beschreiben strukturelle Eigenschaften des spleißosomalen Signalnetzwerks und bilden eine Grundlage für die Identifizierung neuer Signalwege. Insbesondere die Eigenschaft nebenläufige Prozesse abbilden zu können ermöglichte es, mit Hilfe des Petrinetzes die verschiedenen Varianten der Initiierung des spleißsomalen Signalnetzwerks darzustellen und zu vergleichen.

Das Modellnetzwerk eignet sich darüberhinaus sowohl für die animierte Simulation der Reaktionen des Spleißosom-Aufbaus als auch für die Klassifizierung von Signalwegen, um wichtige Knotenpunkte im Netzwerk zu identifizieren. Zusamenfassend eröffnet dieser Modellansatz die Möglichkeit, die großen Mengen an Proteininteraktionsdaten aus öffentlichen Datenbanken und Publikationen sinnvoll für die Erstellung neuer- und zur Analyse bestehender Hypothesen einzusetzen.

Contents

Abbreviations

5′ss	5′ splice site
3′ss	3′ splice site
AS, as	alternative splicing, alternatively spliced
asf	alternative splice form
ass	alternative splice site
A3E	alternative 3′ss exon
A5E	alternative 5′ss exon
ACE	alternative conserved exon
AFE	alternative first exon
ALE	alternative last exon
AP	alternative polyadenylation
BP	Branchpoint
CE	constitutive exon
cDNA	complementary DNA
CDS	coding sequence
css	constitutive splice site
dΨ4	pseudo Δ4 tandem donor site without evidence of AS; located upstream of a constitutively spliced donor
DΔ4 (dΔ4)	distal-major (distal-minor) donor
DTC	delayed termination codon
E5	exon 5′ end
E3	exon 3′ end
ESE	exon splicing enhancer
ESS	exon splicing silencer
EST	expressed sequence tag
hsp	high scoring segment pair
I5	intron 5′ end

I3	intron 3′ end
ISS	intron splicing silencer
MCTS	maximal common transition set
MXE	mutually exclusive exon
NMD	nonsense mediated RNA decay
nt	nucleotide
pΨ4	pseudo Δ4 tandem donor site without evidence of AS; located downstream of a constitutively spliced donor
PΔ4 (pΔ4)	proximal-major (proximal-minor) tandem donor
PN	Petri net
PolyA	polyadenylated tail of mature mRNA
PPT	polypyrimidine tract
PTC	premature termination codon
RI	retained intron
rss	reference splice site
RNP	ribonucleoprotein
RRM	RNA recognition motif
SE	skipped exon (cassette exon)
snRNA	small nuclear RNA
SRp	proteine (splicing factors) with Serine-aRginine rich domains

Chapter 1

Validation of Alternative Splice Forms Detected *in silico*

1.1 Introduction

1.1.1 Alternative Splicing

The central dogma of molecular biology postulates the directed flow of genomically encoded information in the nucleus via messenger RNA (mRNA) to the cytoplasm where the message is translated into proteins .In higher organisms, this information flow becomes dynamically modulated by RNA maturation processes among which *splicing* is a prominent example. Splicing describes the process by which parts (introns) of the pre-cursor messenger RNA are catalytically excised by a ribozyme complex while remaining pieces (exons) are ligated and constitute the coding sequence. Furthermore, some parts of the spliced transcript remain untranslated at the $5'$ and $3'$ end of the mRNA. The recognition of the donor and acceptor splice sites at the exon-intron boundaries are sensitive steps in initializing the splicing reaction. Only a few years after discovery of the splicing mechanism in 1977 (5), it was clear that mRNAs from the same genomic region could differ in length and nucleotide composition. Successively, it was shown that alternative splice reactions can take place when splicing signals are switched on and off or occur in competing proximity to each other (6). Today, alternative splicing (AS) has become a paradigm to explain the increasing morphological complexity as compared to genome size, that is observable in eukaryotes from protozoans via nematodes, arthropods to vertebrates (7, 8).

AS events are categorized in different patterns of splice site selection and one

can distinguish the four basic types: exon-skipping (SE), in which mRNA isoforms differ by the inclusion/exclusion of an exon; alternative 5′ss exon (A5E) or alternative 3′ss exon (A3E), in which isoforms differ in the usage of a 5′ss or 3′ss, respectively; and retained-intron types (RI), in which isoforms differ by the presence/absence of an un-spliced intron (9). These types are not necessarily mutually exclusive and more complex types of AS events can be constructed from such canonical types. In addition, AS holds the possibility to control gene expression at the post-transcriptional level via the non-sense mediated mRNA decay (NMD) pathway. To prevent aberrantly or deliberately incorrectly spliced transcripts that prematurely terminate translation, NMD ensures that only correctly spliced mRNAs that contain the full (or nearly so) message are subsequently utilized for protein synthesis. Therefore, NMD scans newly synthesized mRNA for the presence of one or more premature-termination codons (PTCs), and, if detected, can selectively degrade defective mRNAs (10). Furthermore, pathological complexity can be ascribed to misregulated splicing with cancer as one of the most disastrous examples (11). The mechanism of alternative splicing has attracted a wide range of scientific research addressing the problem with computational strategies and tools (12, 13, 14, 15, 16). A variety of databases have been designed to collect alternative splice forms (17) along with accompanying experimental conditions like source tissue, developmental stage or pathological information (*cf.* Table 1.1).

Purification of spliceosomal components (see chapter 3.1.1) and *in vitro* splicing assays showed that regulatory proteins are involved in initiating the splice mechanism and maintaining spliceosomal activities (18, 19, 20). Most of the known splice regulatory proteins are members of a protein family that share serine-arginine domains, called *SR proteins*. Additionally, SR proteins can bind via RNA recognition domains to specific sequence motifs to promote alternative splice site usage and exon definition (7, 21). Considering the growing percentage of genes that are affected by alternative splicing transcripts there must be a dense regulatory network, which receives and forward signals required to perform the surgical task of splicing. Consequently eukaryotic cells have a means to react appropriately to different environmental conditions by triggering the splicing of different isoforms. Over the last decade experimental studies have amply provided examples of alternative splice events whose regulation depends on the presence and distribution of cis-elements and the concentrations of their recognizing trans-factors (6, 22). These preliminary works paved the way for a variety

of computational studies, some of their covered topics briefly summarized in the following list:

i. splice site modeling (23),

ii. splice site plasticity (2),

iii. exon/intron sequence composition (24)

iv. splicing pattern statistics (25) and

v. splicing enhancer and silencer prediction (26, 27).

1.1.2 History of Computational Analyses on Alternative Splicing

The computational detection of alternative splice forms based on sequence analysis is tightly connected to the identification of genes within genomes. That is because mature transcripts still bear a strong resemblance to the genomic *loci* from which they were transcribed. Hence, the comparison of processed (and thus assumed to be matured) RNA transcripts to genomic sequence can serve to identify intronic regions which were removed from the primary transcript during the splicing process. One of the first approaches to find gene regions consisted of aligning complete (full-length) RNA transcripts against genomic sequence and thus pinpointing the location of genes. UniGene cluster for example successively incorporate transcripts (mRNA and EST sequences) into initial alignments forming gene centered clusters of spliced sequences (28). Since then, spliced alignments have been included into several gene prediction pipelines (12, 29) as for example the Transcript Assembly Program (12). They constitute an extrinsic or indirect approach to gene prediction as contrasted by *ab initio* methods which make directly use of statistical properties of the genomic sequence such as the GC content (30). From the initial estimates of 30%-40% of the human genes being alternatively spliced (31, 32), by improving experimental and computational techniques this estimate has constantly increased over the past years, ranging presently at more than 70% (33, 34, 35). Along with the improvement of detecting alternative splice patterns, databases and web application were developed for storing and analyzing the wealth of data. Two main types of alternative splicing data repositories can be distinguished: *i*) AS

databases created by text mining, storing manually curated or collected AS events reported in the literature and *ii*) computationally detected and annotated AS databases.(*cf.* Table 1.1).

The quality of EST sequences is often compromised by sequencing errors. Additionally, ESTs can be overrepresented in the $5'$ or $3'$ region of gene, posing an unbalanced transcript coverage. This encouraged scientists to explore also other possibilities for corroborating the evidence on alternative splice patterns. One strategy pursues the interspecies comparison of splice patterns to confirm AS events by their sequence conservation in other species (36, 37). However, this is often not feasible on a genome wide scale, due to the incomplete sequence coverage in non-human species. Another approach is the artificial construction of splice junction probes based on reference mRNA sequences and their hybridization with RNA cell extracts (which can be from various tissues) on microarray chips (35, 38). While producing a large amount of information confirming specific AS patterns and giving insights into their regulation, microarray and high throughput sequencing approaches constitute a considerable financial effort. Beside genomic and mRNA sequence information, also the wealth of protein sequences has been used to predict and analyze AS forms (39, 40). A third possibility, reported by Hiller et al. uses protein family domains (PFAM) to reconstruct AS events (41). This approach classifies novel AS events by evaluating the PFAM score after computationally removing exons or retaining introns in reference transcript structures and translating them into protein sequence.

Table 1.1: Historical outline of alternative splicing databases. AFE=alternative first exon; ALE=alternative last exon; AP=alternative polyadenylation. For further explanation of acronyms see references and glossary; na = no information available.

Year	Database	Data Description	Collected Types of AS Patterns	Number of AS Sequences	Ref.
1999	ASDB	Collection of protein and DNA sequence entries, labeled as "VARSPLIC" (Swiss-Prot) and "Alternative Spliced" (GenBank)	n/a	1,200 DNA Sequences	(42, 43)
2000	ASFINDER	BLAST of EST against cDNA sequences	Inserts, Deletions	2,747 AS cDNAs (1,797 AS genes)	(44)
2001	SPLICEDB	Collection of (non)canonical splice site pairs from GenBank sequences	n/a	28,468 ss pairs	(45)
2002	SPLICENEST	Aligning ESTs from UniGene cluster alignment against genomic sequence (SIM4); filtering and delinieation of AS patterns	SE, RI, A5E, A3E	14,900 AS UniGene clusters	(46, 47)
2002	PALSDB	Comparing UniGene cluster transcripts	inserts, deletions	14,106 AS genes	(48)
2003	PROSPLICER	Alignment EST and mRNA (SIM4) and protein (TBLASTN) sequences to genomic sequence (ENSEMBL)	SE, A5E, A3E	n/a	(49)
2003	ASAP	Mapping UniGene sequences to genome (BLAST); analyzing EST (mRNA) cluster with GeneMiner	SE, A5E, A3E	11,717 AS genes	(34, 50, 51)
2004	EASED	Aligning mRNAs to dbEST (BLAST); filtering and mapping of AS mRNAs to ENSEMBL genes	Inserts, Deletions	18,308 AS transcripts (14,792 AS genes)	(52)
2004	ASD	Manually curated EST (mRNA) alignments (BLAST) against genomic sequence	CE, A5E, A3E, SE, RI, MXE	2,581 AS Events (ALTEXTRON); 8,314 AS genes (ALTSPLICE)	(53, 54)
2005	FASTDB	EST (mRNA) alignments (SIM4) against protein coding gene sequences (ENSEMBL)	CE, A5E, A3E, SE, RI	11,071 AS genes	(55)
2005	SPLICEINFO	AS forms derived from PROSPLICER and from mRNA-protein sequence comparison to genomic sequence (ENSEMBL)	SE, A5E, A3E, RI	6,309 AS genes	(56)
2005	MAASE	Semi-automated AS analysis of manually provided gene IDs by BLAT and SIM4	A5E, A3E, SE, RI, MXE	1,007 AS genes	(57)
2006	ASTRA	Mapping of mRNAs (UniGene) to genome by MEGABLAST and ALN	AFE, ALE, A5E, A3E, SE, RI, MXE	12,470 AS cDNAs (4,931 AS genes)	(58)
2006	HOLLYWOOD	Aligning mRNA vs genomic sequence (ENSEMBL); subsequent alignments of EST against the mRNA hits (SIM4) within the genome	CE, A5E, A3E, SE, RI, MXE	10,800 AS genes[a]	(3)
2006	TASSDB	Aligning ESTs (BLAST) against RefSeq splice site junctions which show specific splice cite patterns	A3E (NAGNAG), A5E (GYNGYN)	10,995 GYNGYN, 11,964 NAGNAG patterns	(16)
2006	ALTTRANS	Extending the ALTSPLICE DB for alternative PolyA sites	AP	2,053 AP genes	(59)
2007	SPLICEMINER	Web Frontend to the "Evidence Viewer" DB of non redundant human splice variants with complete CDS	no classification	na	(60)
2007	BIPASS	Two-step alignments of EST (dbEST) and mRNA (GenBank) to genomic sequence by BLAST and SIM4	AFE , ALE, SE, A5E, A3E	na	(61)

[a] number referring to AS events of internal exons

1.1.3 Splicing Factors and their Role in Regulating Alternative Splicing

Many proteins have been found to influence alternative splicing events. Especially SR proteins, a family of proteins that share serine-arginine rich conserved domains, are frequently located near sites of active splicing. These proteins can interact via their RS domains and thus provide bridging functions in spliceosome assembly and splice site definition (62). Furthermore SR proteins possess RNA recognition domains and have been shown to recognize a variety of *cis*-elements. According to location and effect on the splicing process, these *cis*-elements are classified as exonic or intronic splicing enhancer (ESE, ISE) or -silencer (ESS, ISS) elements. Understanding the role of SR proteins in AS regulation is complicated by the fact that the recognized sequence motif alone is often insufficient to determine their function as positive or negative splicing regulator. Studies have shown that more than one copy of a high affinity SR protein binding site can efficiently activate splicing as was shown for three sequential SRp binding motifs binding ASF/SF2 and SRp40 (63, 64). This finding was explained by the action of cooperative binding of several splicing factors of the same type resulting in a higher specificity that helps in outcompeting other trans-factors with lower RNA binding affinity in this specific region. Contrary to the single type factor binding model other studies have shown that also combinations of different SR proteins can enhance splicing by recognizing and binding to similar adjacent enhancer motifs (65). A variation of this model is shown by another example where exon 5 and 6 of the human *caldesmon* gene showed multiple purine rich repeats, sharing at least 32nt, which promoted alternative splicing of an internal 5′ss within exon 5 (66). Since most of the reported ESEs show lengths between 5-19 nucleotides (62), longer motifs may consist of several overlapping sub motifs that increase the binding affinity of its recognizing trans-factors (67). Even the possibility of a composite splice regulatory element has been demonstrated, consisting of different motifs within exon 5 of an transcript of the CD44 gene. These motifs were required in a directly adjacent location to activate a functional splice complex, also suggesting interactions of the respective binding proteins (68).

1.2 Methods

1.2.1 The EASED Database

All analyses in this chapter are based on the „Extended Alternative Spliced EST Database“ (EASED)[1], which was developed as a repository of annotated putative alternative splice forms (52). The database uses mRNA and gene sequence material of the ENSEMBL database[2] version 19.34a and ESTs from the dbEST database[3] (December 2003) (69) and is currently restricted to the species *Homo sapiens*. The information is organized in a top-down order from genes via transcripts to individual alternative splice events and their associated features. In general the identification of alternative splice forms (asf) is based on an mRNA - EST alignments with a stringent set of parameters (see 1.2.2). The EASED project constitutes one of the earliest efforts to combine an algorithm for locating alternative splice forms with annotation of a variety of secondary, but medically relevant information such as associated diseases (for example cancer), prevalence for a specific tissue or developmental state(52). An overview on the performed analyses is given in Figure 1.1.

1.2.2 Locating Alternative Splice Forms

Spliced transcripts, present as ESTs or fully sequenced mRNAs, are composed of exons after excision of introns from a transcribed precursor transcript. Hence, an alignment of mRNAs against the genomic sequence reveals intronic regions as gaps within the spliced (mature) mRNA (Figure 1.2). Compared to full-length mRNAs, much ore ESTs are available, providing snapshots of the outcome of different splicing processes. Since a direct alignment of ESTs against genomic sequence is in many ways impractical, because of the large search space and possibility of unspecific hits, the available ENSEMBL transcripts (∼31.500 as of ENSEMBL version 19.34a) were used to restrict the analysis to genomic regions of known genes. Thus, full length transcripts were aligned against all available human ESTs from dbEST (∼5.5 Million as of December 2003), using the WU-BLAST program[4]. In order to prevent unspecific hits and to reduce the compu-

[1] http://eased.bioinf.mdc-berlin.de
[2] http://www.ensembl.org
[3] http://www.ncbi.nlm.nih.gov/dbEST
[4] http://blast.wustl.edu/

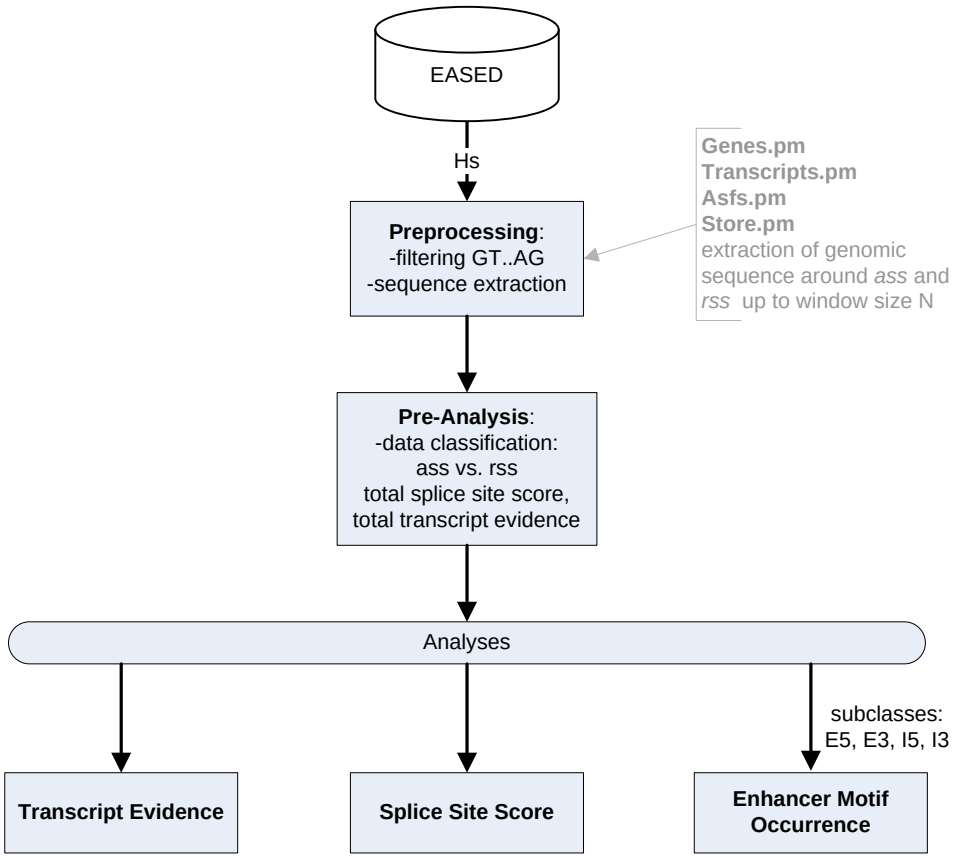


Figure 1.1: Workflow of data preparation and analyses performed with EASED dataset. The grey box to the right indicates the Perl modules programmed for pre-processing the EASED data. Donor and acceptor splice site context were partitioned into E3=exon 3′ end, I3=intron 5′ end, I3=intron 3′ end and E5=exon 5′ end.

tational load, known repetitive sequences as collected in RepBase[5]) were masked in all mRNA transcripts prior to the alignments. The longest available ENSEMBL transcript of each gene *locus* provided an initial exon intron annotation, denoting a set of „constitutive“ or reference donors and acceptors. Taking the longest mRNA as a reference holds the chance to observe a maximal number of alignment gaps against the available EST sequences and hence, to increase the number of identified alternative splice junctions. Each mRNA-EST alignment was filtered against a stringent set of quality criteria. The aligning blocks (hsps) had to show at least 98% identity over a length of 100 nt to qualify as exons of further use (see details in (52)). Differences between exon structures of ESTs and the exon locations of the reference transcripts were used as indicators of alternative splice

[5]http://www.girinst.org/repbase

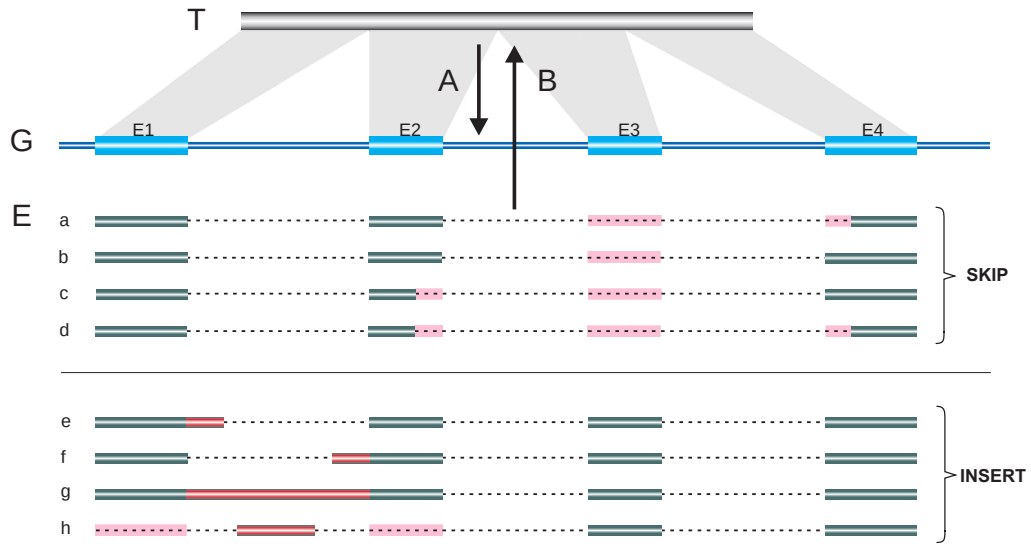


Figure 1.2: Modeling alternative splice patterns by transcript alignments: **(A)** a full-length reference transcript (T) with known exon-intron structure serves as anchoring region within the genome (G); **(B)** ESTs (E) are aligned against T to annotate alternative splicing patterns. Two main types of alternative splice events are distinguished in the EASED database which are SKIP and INSERT events (defined relative to the reference transcript). SKIP events are further distinguished into the AS pattern *a-d* resulting in deleted exonic sequence relative to T. INSERT events fall into categories *e-h* showing different pattern of inserted exonic sequence. Each AS type creates relative to T one ore more alternative splice sites (ass); splice sites in T and E mapping to the same position in G are considered as reference splice sites (rss); exonic regions in the genome are labeled E1-E4; alternative parts as indicated by the transcript alignments are shown in pink and red color if they are missing or inserted relative to T, respectively

sites. According to the alignments, exon boundaries differing between ESTs and the reference transcripts were treated as alternative splice sites while the remaining exon boundaries being conform with the reference transcript where classified as reference splice sites. Hence, all EST boundaries matching a reference splice site were counted as evidence for the set of reference donor and acceptor sites. These splice sites are often labelled „constitutive“. However, present lack of evidence of AS near a constitutive splice site does not preclude that an alternative splice site becomes activated in this region under specific conditions. Being aware of this uncertainty, in the following I will refer to these splice sites as „reference“ instead of „constitutive“.

1.2.3 Data Preparation and Refinement

For creating the initial dataset, the database was queried with a threshold of at least 5 asf (EST) per gene to increase the chance of capturing a true alternative splice event. Next, all genes whose supporting EST did not indicate alternative splice events utilizing the canonical splice site dinucleotides GT..AG, were discarded. After refinement the dataset consisted of 2624 genes with an average number of 8 alternative splice sites per gene and an average support of 3 EST per alternative splice site. In contrast, ∼26 ESTs per gene were indicative of reference splice sites. Due to the filtering steps the number of initially available ESTs and annotated splice events decreased considerably as documented in Box 1.2.1.

Box 1.2.1 EASED figures
Filtering steps and discarded sequences from the initial towards the final dataset; (*) number of asf satisfying prediction criteria (**) number of asf strictly meeting the criterion of canonical GT..AG splice sites

Processing Step	Number	Source
blasted ESTs	5,427,257	dbEST Database
Ensembl Genes (ENSG)	23,531	Ensembl Database (19.34a)
Ensembl Transcripts (ENST)	31,609	
Ensembl Exons (ENSE)	225,897	
EST, matching ≥ 1 ENST	3,947,548	
asf predicting EST*	428,474	
predicted as sites	102,104	Eased Database
ENST matching ≥ 1 asf	21,044	
ENSG matching ≥ 1 asf	15,426	
ENSG matching ≥ 5 asf	2624	
assAcceptors**	3862	
assDonors**	3705	Perl scripts
rssAcceptors**	25526	
rssDonors**	25103	

As a further parameter, the genomic sequence window around each donor and acceptor site was set to 50 nt up and downstream of the splice site (Figure 1.3). The window size was chosen for locating known splice regulatory motifs and restricted to 50 nt after inspecting the exon length distribution of transcripts present in the Eased database. More than 70% (90%) of the represented exons (introns) are longer than 100 nt (1.2), allowing to span a window of 50 nt into 5′ and 3′ direction) around each splice site which for the majority of exons and introns does not overlap between the donor and acceptor site. These clearly separable regions were used for the enhancer motif analysis. Finally each splice site was stored as alternative- (ass) or reference (rss) splice site together with a number of supporting transcripts (mRNA and/or EST). The main relationship *one gene → n complete transcripts (mRNA) → m partial*

transcripts (ESTs) $\rightarrow$ *x alternative splice events* is reflected in the EASED table structure (Figure A.1, Table A.1). Supporting tables contain additional data as for example corresponding protein information on mRNAs or tissue and developmental information on ESTs. The analyses were restricted to *Homo sapiens* as this species provides the most abundant sequence information. Data for subsequent analyses was arranged and complemented via Perl scripts and the EASED database tables (Table A.1)

Table 1.2: Number and lengths (L) of EASED cDNA exons and introns

Counts	**Exons**	**Introns**
total	219,388	189,145
$L < 100$ nt	64,062	15,760
$L \geq 100$ nt	155,326 (71%)	173,385 (92%)
mean (L)	262	5564
GT..AG	-	169,732 (97.9%)
GC..AG	-	422 (0.24%)

1.2.4 Splice Site Scoring

In order to compare different splice sites, it is reasonable to apply a measure of information. The information content can be expressed as entropy, which is equivalent to the uncertainty to observe a specific event, for example, a specific nucleotide within a sequence motif. The entropy is calculated via probabilities, reflecting an increasing certainty to observe a specific event. The more probable the event, the lower the uncertainty or entropy. Thus the entropy becomes close to zero if there is a high certainty about the event. For example, position +1 and +2 in eukaryotic splice sites are to >98% conserved to the bases guanine and thymine (33), lowering the uncertainty about this dinucleotide at human donor positions to ∼0.14. According to the maximum entropy principle, of all possible distributions in the hypothesis space, the distribution that is the best approximation of the true distribution given what is known, is the one with the largest Shannon entropy (Equation 1.1, (23)):

$$H = -\sum p \cdot \log_b(p) \tag{1.1}$$

where p is the probability to observe an event. For example, the four bases A, C, G, T that are possible at each splice site position represent k observable events and the absolute entropy of a splice site motif of length k corresponds to

an information content I (Equation 1.2). Consequently, splice site information scores are expressed in units of bits.

$$I = -\sum_{i=1}^{k} p_i \cdot \log_2(p_i) \ [Bit] \tag{1.2}$$

Often one does know only little about the true nature of a given set of sequences and has to compare the splice sites of different sets in order to capture meaningful and interpretable differences. In this case it is more practical to calculate the relative entropy, also termed *transinformation* or *Kullback-Leibler divergence* (70), which estimates the „distance“ between an observed (p) to an expected (q) frequency distribution (Equation 1.3). The background distribution was taken as $(q_1, q_2, q_3, q_4) = \{A, G, C, T\} = (0.3, 0.2, 0.2, 0.3)$, owing to a GC content of 41% on average observed in the human genome (33).

$$KL = \sum_{i=1}^{k} p_i \cdot \log_2(\frac{p_i}{q_i}) \tag{1.3}$$

Finally, the information content of a splice site or sequence motif is the sum of KL over all considered sequence positions (Equation 1.4) which for the donor site comprises n=9 nucleotides (position -3 to +6 in exon-intron orientation) and n=23 nucleotides (position -20 to +3) for acceptor sites.

$$S = \sum_{j=1}^{n} KL_j \tag{1.4}$$

Measuring the splice site information content can be further improved by including marginal constraints. These can be imposed through neighborhood relations between nucleotide positions and their observed frequencies estimated from experimental data (23). This scoring scheme is implemented in the program MaxEntScan[6] and was used to calculate the splice site scores in all analyses. Alternative splice sites were treated as a joint dataset, that is not further separated into up- and downstream site per acceptor and donor. The score distribution of alternative donor and acceptor sites was compared to the respective set of reference splice sites, using the Wilcoxon rank-sum test.

[6]http://genes.mit.edu/burgelab/maxent/

1.2.5 Scanning for Splice-Enhancing Motifs

To estimate the frequency of high scoring exon splicing enhancer (ESE) motifs within a defined context around the investigated splice sites, scoring matrices implemented in the software ESEFINDER (27) were used . The matrices contain nucleotide frequencies of short motifs that were shown to bind the SR proteins SF2/ASF (7mer), SC35 (8mer), SRp40 (7mer) and SRp55 (6mer). Experiments have demonstrated the applicability of ESE consensus motifs in explaining splice events as exon skipping and mis-splicing due to point mutations within those motifs (71). Nevertheless, due to the degenerated nature of SRp binding motifs, the frequency of ESE motif hits was not considered alone but their occurrence in the context of additional characteristic information (splice site score, transcript support) evaluated. Binding motifs for these four SR proteins were scanned in a window of 100 nucleotides, composed of a two 50 nt regions up- and downstream of the predicted *ass* and *rss* (Figure 1.3). Especially when parts of introns are

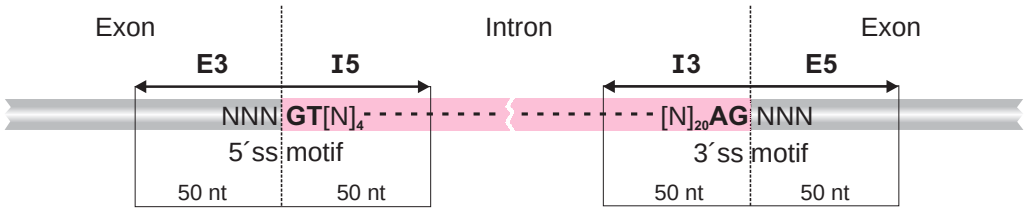


Figure 1.3: Sequence windows used for determining the ESE frequencies around the scored splice sites. The four validation classes are defined as I5, I3, E5 and E3 with the following convention: E = exon, I = intron, 3 = downstream region (3´), 5 = upstream region(5´). Each sequence region spanned 50 nt; intron-exon borders are defined by the donor- and acceptor splice site motifs which are also used for the splice site scoring.

spliced into the mature transcript, one can expect to find ESE motifs at both sides (up- and downstream) of splice sites. However, the scan window was kept with 100 nucleotides in a size that covers a distance to the splice site in which experimental data have shown the location of ESEs (72, 73). Reducing the window size may certainly cause the loss of regulatory motifs for the analyses but enables to include more and especially shorter introns into the analyses (*cf.* Table 1.2) while maintaining non-overlapping scan windows. There exists cases where regulatory motifs are located more than 100 nucleotides downstream of a donor site (74) and it is known that proximity of motifs to its target splice sites is governed by the RNA structure (75). However, looking in the close sequence proximity of splice sites should increase the chance to find splice enhancer motifs regulating exactly the alternative splice site under scrutiny.

1.2.6 Data Partitioning for Statistical Analysis

As shown in Figure 1.3, four classes (data subsets or partitions) were created from the EASED data. These comprise the sets of donor (acceptor) splice sites of *ass* and *rss* splice forms. Due to the up- and downstream distinction of the splice site environment, the data subsets can also be analyzed with respect to differences between exon and intron specific features as for example the frequency of SR protein binding motifs.

1.3 Results

1.3.1 Analysis of EASED Splice Forms

Up to now no perfect *ab initio* prediction algorithm for alternative splice events exists and an observed sequence feature if taken isolated may not be strong enough to validate a splice site to certainty. Based on the EASED database (52) different attributes were compared between putative alternative and constitutive splice forms. Genomic sequence features that are characteristic for spliced transcripts were used to evaluate the available set of annotated alternative splice sites. A central question was, whether multiple ESEs or enhancer repeats may contribute to the discrimination of these ass. The combination of several classes of information such as splice site score (S_{ss}), frequency of (splice enhancing) SRp binding motifs (f_{ese}), transcript support (f_t), showed to be a strong ensemble for differentiating between the processes of both constitutive and alternative splicing. Given the fact that a higher transcript coverage (EST, mRNA) exists for the set of reference splice sites, their attributes are a reliable and informative source for the comparative description of putative alternative splice forms. Additionally, annotations inherent to EST records such as tissue type, disease and developmental state have previously been shown to discriminate reference from alternative splicing (13) and thus can be considered as supportive criterions for „multiple-feature“ analysis strategies.

1.3.2 General Characterization of Splice Site Attributes

Splice Site Score

Comparison of the main attributes *splice site score*, *transcript support* and *ESE frequency* was done first for the complete distribution of both splice site types (donors/acceptors) independent from partitioning into reference (rss) and alternative (ass) splice sites. The results indicate that the S_{ss} between the whole unpartitioned distributions of donor and acceptor site vary significantly ($p < 10^{-3}$, Wilcoxon-, t- and f-test) in their means and variances. The median of both distributions resides at S_{ss} of ~ 8.5 bit but donor splice sites tend to lower scores above the median compared to acceptor sites (Table 1.3a).

Table 1.3: Distributional characteristics of splice site score (S_{ss}) (**a**) and transcript support (f_t) (**b**) compared between donor and acceptor splice sites

(a) Splice site score

	5′ss	**3′ss**
Minimum	-44,399	-46,658
1st Quantile	6,640	6,483
Median	8,456	8,450
Mean	6,976	7,671
3rd Quantile	9,652	10,141
Maximum	11,807	15,589

(b) Transcript support

	5′ss	**3′ss**
Minimum	1	1
1st Quantile	1	1
Median	16	16
Mean	41	37
3rd Quantile	68	63
Maximum	466	326

Transcript Support

The number of transcripts that support donor and acceptor splice sites (f_t = frequency of transcripts, that utilize a specific splice site) show different distributions in particular at higher transcript numbers per splice site (Figure 1.4b). Means and variances of the splice site type dependent f_t distributions vary significantly ($p < 10^{-3}$) with a tendency to more transcripts at donor splice sites. The mode of both donor and acceptor f_t distributions reside between 1 and 20 transcripts. The tendency of higher transcript support of donor sites means that more ESTs (mRNAs) show a clean donor site without a downstream acceptor site (transcript containing the first exon) than *vice versa* a clean acceptor without an upstream donor site (transcripts containing the last exon) when mapped to the genomic sequence. This is indicative of partial transcripts which more often locate in the upstream than in the downstream region of a gene.

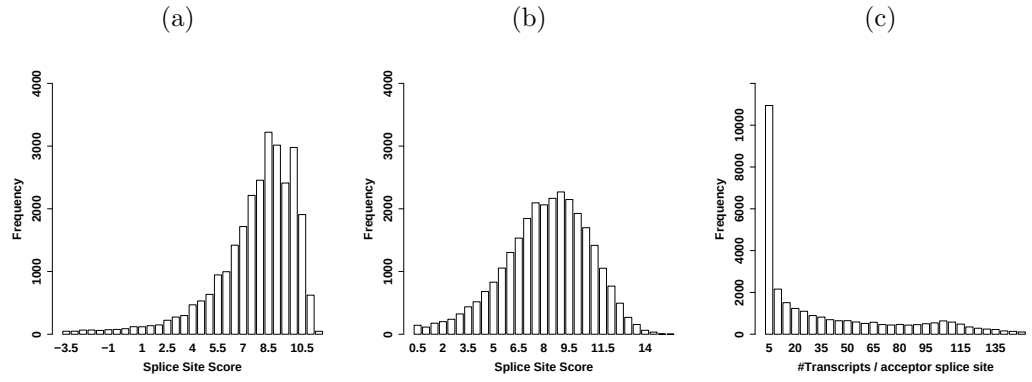


Figure 1.4: Histograms of **(a)** donor splice site scores (S_{ss}) and **(b)** acceptor splice site scores; **(c)** histogram of transcript frequencies (f_t) confirming the analysed acceptor sites. Each bin shows how many splice sites are confirmed by the given number of observed transcripts.

Frequency of SRp Binding Motifs

In investigating the repeated occurrence of SRp binding motifs (f_{ese}), at first the total number of detected motifs were considered. Comparison of this combined ESE frequencies showed no significant difference ($p > 0.1$, t-test) between the distributions in the exonic donor (E3)- and acceptor (E5) region but a significant difference between the intronic donor (I5)- and acceptor (I3) region ($p < 10^{-4}$). Also the occurrence of ESE motifs between adjacent exonic (E5) and intronic (I3) splice site flanks differed significantly at acceptor sites ($p < 10^{-3}$). Looking at the motifs of the specific SR protein ASF/SF2 one observes similar f_{ese} distributions in the exonic part of 5´- and 3´ splice sites but though they show significant difference in their variances ($p < 10^{-4}$), their means are not different ($p > 0.1$, Wilcoxon test). In contrast, the number of ASF/SF2 motifs in the intron flanks of the splice sites vary significantly in their variances and means ($p < 10^{-4}$) and the same effect was found between adjacent exonic and intronic splice site environments. A summary of variances and means, characterizing the f_{ese} distributions of all four tested SR proteins is listed in Table 1.4.

Table 1.4: Summary of ESE motif frequency (f_{ese}) characteristics considering the whole set of splice site while distinguishing the intronic and exonic parts of donor and acceptor splice sites; $p(\mu;\sigma)$ denotes p-values in comparing means and variances between the motif frequencies at the specified exonic/intronic region; p-values of significantly ($p \leq 0.0001$) different motif frequencies between the compared splice site regions are marked with (*)

SRp	**Splice Site Region**	**Median**	**Mean**	**Variance**	**E3** $p(\mu;\sigma^2)$	**I3** $p(\mu;\sigma^2)$
ASF/SF2	E3	4	3,934	5,285	–	–
	E5	4	3,933	5,016	< 1; $< 10^{-4*}$	$< 10^{-4*}$; $< 10^{-4*}$
	I3	2	2,502	4,229	–	–
	I5	3	3,359	6,298	$< 10^{-4*}$; $< 10^{-4*}$	$< 10^{-4*}$; $< 10^{-4*}$
SC35	E3	4	3,901	3,989	–	–
	E5	4	3,873	3,949	$< 10^{-1}$; < 1	$< 10^{-4*}$; $< 10^{-4*}$
	I3	3	2,928	3,467	–	–
	I5	3	3,622	4,560	$< 10^{-4*}$; $< 10^{-4*}$	$< 10^{-4*}$; $< 10^{-4*}$
SRp40	E3	4	3,796	3,145	–	–
	E5	4	3,794	3,095	< 1; < 1	$< 10^{-4*}$; $< 10^{-4*}$
	I3	3	2,823	2,522	–	–
	I5	3	3,452	3,183	$< 10^{-4*}$; < 1	$< 10^{-4*}$; $< 10^{-4*}$
SRp55	E3	2	2,344	2,277	–	–
	E5	2	2,334	2,236	< 1; < 1	$< 10^{-4*}$; $< 10^{-4*}$
	I3	1	1,626	1,612	–	–
	I5	2	1,947	1,965	$< 10^{-4*}$; $< 10^{-4*}$	$< 10^{-4*}$; $< 10^{-4*}$

Occurrence of SRp Binding Motifs at Low and High Scoring Splice Sites

Additionally, and in order to validate f_{ese} independently from our definition of ass and rss, all splice sites were separated by their S_{ss} distributions (Figure 1.4). Furthermore, only the tails of the score distribution (< 0 and > 10) were selected. Beside the question whether the f_{ese} distribution vary significantly between the low and high scoring sets, this served also to measure the enrichment of predicted asf in both score dependent data subsets (discussed in the section of ass and rss specific analyses). The total donor/acceptor splice sites partition in 6%/4% with $S_{ss} < 0$ contrasted by 19%/25% with a score above 10 and thus almost consensus quality. Between these two sets of extreme-scoring splice sites significant differences in the means of their total (exonic) f_{ese} distributions were found . In particular, at both donor- and acceptor sites the motif distribution of the protein SC35 tends on average to at least one motif more at low scoring than at high scoring splice sites. For two other SR proteins this effect was found only either at donor splice sites (SRp55) or at acceptor splice sites (ASF/SF2).

1.3.3 Characterizing Specific Attributes of Alternative and Reference Splice Sites

Splice Site Score and Transcript Support

Comparing S_{ss} between reference and alternative donors shows a significant difference in the mean values ($p < 10^{-4}$, Wilcoxon test), although the mode of the distribution still resides at scores between 5 and 10. The same observation applies for the acceptor sites. Interestingly, the difference of the means between donor- and acceptor score distributions is higher in ass than in rss (underlined in Table 1.5). Considering the number of transcripts that support the splice sites, the defined rss are in generally better covered with transcripts than the predicted ass (at both donor and acceptor sites). The mode of both distributions can be found on a class level of 1-20 transcripts though the overall number of transcripts is approximately six times higher in the respective class level of the reference dataset. Between score and transcript support a strong dependency in both ass and rss ($p < 10^{-4}$, χ^2-test) were found whereupon better scoring splice sites clearly show a better transcript support.

Table 1.5: Splice site score characteristics in the *ass* and *rss* datasets. Underlined values emphasize the significant difference between mean values of *ass*- and *rss* score distributions

Splice Site Type	#Total Counts	Median	Mean	SD	Variance
all donors	28808	8.456	6.976	5.475	29.971
ass donors	3705	5.263	0.862	10.219	104.438
rss donors	25103	8.626	7.878	3.557	12.652
all acceptors	29388	8.450	7.671	4.446	19.765
ass acceptors	3862	5.997	3.356	8.035	64.562
rss acceptors	25526	8.678	8.324	3.122	9.747

Table 1.6: Analysis of variances (f-test) and means (Wilcoxon test) between ESE motif frequencies (f_{ese}) distributions at *ass* and *rss* splice sites in different pre-mRNA contexts (*cf.* Figure 1.3); μ = sample mean, m = sample median, σ^2 = sample variance; p-values significant at $\alpha \leq 0.05$ are marked with (*)

SRp	Region	*ass* $\mu/m/\sigma^2$	*rss* $\mu/m/\sigma^2$	$ass \leftrightarrow rss$ $p(\mu)$	$ass \leftrightarrow rss$ $p(\sigma^2)$
ASF/SF2	E3	4.148 / 4 / 5.862	3.903 / 4 / 5.192	$< 10^{-4*}$	$< 10^{-4*}$
	E5	4.187 / 4 / 5.382	3.894 / 4 / 5.950	$< 10^{-4*}$	$< 10^{-3*}$
	I3	2.910 / 3 / 4.268	2.440 / 2 / 4.194	$< 10^{-4*}$	$> 10^{-1}$
	I5	3.864 / 4 / 6.580	3.285 / 3 / 6.213	$< 10^{-4*}$	$< 5 \cdot 10^{-2*}$
SC35	E3	4.026 / 4 / 4.068	3.883 / 4 / 3.975	$< 10^{-4*}$	$> 10^{-1}$
	E5	4.076 / 4 / 3.949	3.842 / 4 / 3.941	$< 10^{-4*}$	$> 10^{-1}$
	I3	3.254 / 3 / 3.434	2.878 / 3 / 3.454	$< 10^{-4*}$	$> 10^{-1}$
	I5	3.914 / 4 / 4.477	3.578 / 3 / 4.558	$< 10^{-4*}$	$> 10^{-1}$
SRp40	E3	3.769 / 4 / 3.299	3.800 / 4 / 3.122	$> 10^{-1}$	$< 5 \cdot 10^{-2*}$
	E5	3.893 / 4 / 3.191	3.779 / 4 / 3.079	$< 10^{-3*}$	$> 10^{-1}$
	I3	3.081 / 3 / 2.583	2.784 / 3 / 2.501	$< 10^{-4*}$	$> 10^{-1}$
	I5	3.629 / 3 / 3.232	3.426 / 3 / 3.170	$< 10^{-4*}$	$> 10^{-1}$
SRp55	E3	2.334 / 2 / 2.253	2.346 / 2 / 2.281	$> 10^{-1}$	$> 10^{-1}$
	E5	2.347 / 2 / 2.327	2.332 / 2 / 2.222	$> 10^{-1}$	$< 10^{-2*}$
	I3	1.695 / 2 / 1.719	1.616 / 1 / 1.595	$< 10^{-3*}$	$< 10^{-2*}$
	I5	2.098 / 2 / 2.101	1.924 / 2 / 1.941	$< 10^{-4*}$	$< 10^{-2*}$

SRp Binding Motifs in ass and rss Splice Site Environments

Considering f_{ese} at exonic and intronic flanks of splice sites, one observes only subtle differences between *ass* and *rss*. In case of the ASF/SF2 motif f_{ese} at the exonic $5'$ and $3'$ region has its mode at a frequency of three motifs (except for the exonic $5'$ end of *ass*). Nevertheless, there is a significant difference in means and variances between the number of ASF/SF2 motif at the exonic part of *ass* and *rss* (Table 1.6). In contrast, the intronic regions of *ass* and *rss* show generally a higher ASF/SF2 motif abundance which is unexpected since this motif – as

exonic splicing enhancer – should occur more frequently in exonic regions. There is no significant difference in the variances of the f_{ese} distributions between intronic ass and rss acceptor sites but also here the means vary significantly. Considering the size of the data set, the results clearly indicate a tendency to more ASF/SF2 motifs in the flanking regions of predicted ass than in the set of rss. Table 1.6 summarizes the SRp specific f_{ese} distributions found for ass and rss, being compared also between the splice site flanking pre-mRNA regions. The ASF/SF2 and SC35 motif frequencies show in the exonic flanks of $5'$- and $3'$ss a tendency to more motifs in neighborhood of the ass than at rss splice sites. The same result was found for SRp40 motif frequencies with exception of the exon $3'$ flanks. SRp55 motifs appeared significantly more frequent downstream of alternative $5'$ss (I5) than downstream of rss. The means of the f_{ese} distribution of SRp SC35 show a significant difference between ass and rss, although the variances do not convey this information. Based on these tests one can conclude that f_{ese} is different between mRNA regions around splice sites of ass and rss. While these motif modules might be individually subtle (e.g. between two splice sites) they appear to be significant on the whole dataset.

1.4 Discussion

For a dataset of computationally predicted alternative splice sites (ass) it was shown how inherent information can be utilized to validate the predictions by applying statistics on different features typical for splice sites. These features were compared between a set of predicted ass and splice sites arising from a set of mRNAs not predictive of ass by current experimental knowledge (rss). The results suggests that in spite of not predicting ass, the reference transcripts and their splice sites share similar characteristics to the alternative ones as for example the overlapping region in the splice site score (S_{ss}) distributions (ranging from +5 to -10 bit) demonstrate. However, in the low scoring region both ass and rss separate clearly with the ass exhibiting more frequently scores below zero at both donor and acceptor sites. Thus, a significant part of the predicted ass possess motifs incongruent to the splice site motifs found for human GT..AG reference splice sites. In fact, this observation could still be due to pseudo splice sites but as the test of S_{ss} against the transcript support (f_t) indicates, there exists a clear dependency between the number of transcripts that utilize these splice sites and the pertinent score in both the ass and rss.

As a promising splice site feature, the binding motif frequency (f_{ese}) of splice-enhancing SR proteins (SR_p) was investigated in context of exonic and intronic splice site flanks and compared between ass and rss. Firstly, the donor/acceptor site specific occurrence of SRp motifs was analyzed independent from the classification into ass and rss. For both the exonic and intronic flanks a higher variance for f_{ese} of the SR protein ASF/SF2 was found at donor and acceptor sites but only in comparison between the intronic splice site flanks on average 1-2 motifs more were found at the donor site. For the other SR proteins a similar trend towards more binding motifs at intronic donor compared to acceptor regions can be observed. This suggests a higher presence of these motifs at the intronic flank of donor sites, a surprising effect since the motifs were initially determined by consensus sequences made as „exon“ splicing enhancer (27). Nevertheless, since the rss make up the major fraction in these donor/acceptor - intronic /exonic datasets, this observation needed to be further investigated to derive conclusions on an effect that is present also in predicted ass.

Hence, in the next step the differences in f_{ese} of exonic donor and -acceptor sites between ass and rss were analyzed. The ASF/SF2 motif was found to occur in a significant fraction of predicted alternative splice sites (exonic flanks) more

frequently than in the set of reference splice sites. The median of both frequency distributions indicates as much as four ASF/SF2 motifs at both alternative and reference splice site flanks. However, the mean and variance indicate that at both $5'$ and $3'$ splice sites the pattern of ESE occurrence deviates between ass and rss with the tendency to at least one additional motif in the flanking region of alternative splice sites. For example, one observes that ASF/SF2 binding motifs occur in average more frequently in the intron flanks of ass compared to rss, with about the same variance around the mean of four and three motifs respectively. Diverging characteristics (similar variances around different means) in the ass and rss specific motif frequencies of the three other types of SR proteins suggests that both datasets exhibit little variation around significantly different pattern of ESE distribution. It will be interesting to follow up investigations on modules of SR protein binding motifs between predicted ass and rss to test computational predictions on multimerizing enhancer (or silencer) protein complexes which may bind to specific combination classes of cis-elements.

Chapter 2

Analysis of Overlapping Donor Splice Sites

2.1 Introduction

2.1.1 Subtle Splice Variants

Compared with skipped exons as the most prevalent type of AS in human and mammalian cells, A3Es and A5Es are thought to create more subtle changes, by affecting the choice of the 3′ss or 5′ss, respectively. Here, splice site usage gives rise to two types of exon segments – the 'core' common to both splice forms and the 'extension' that is present in only the longer isoform. Both types of AS events have been shown to play decisive roles during development, e.g., sex determination and differentiation in *Drosophila melanogaster* (76) or developmental stage-related changes in the human *CFTR* gene (77), but also in human disease, e.g. 5′ss mutations in the `tau` gene (78). A3Es and A5Es are thought to be regulated by splicing-regulatory elements in exons and nearby exon flanking regions, as well as *trans*-acting antagonistic splicing factors, which bind them and affect the choice of splice sites in a concentration dependent manner (79, 80). Interestingly, computational studies showed that for both A3Es and A5Es the distribution of extensions, $f(E)$, is markedly skewed toward short-range splice forms (81). In particular, alternative splice sites that are separated by the three-nucleotide long motif NAG/NAG/ (where '/' marks an inferred splice site) make up a predominant proportion of A3E events in a mammals, extending to invertebrates and plants (82, 2). The frequent occurrence of the NAGNAG acceptor motif in the human genome, which can be observed at intron-exon borders of

∼30% of human reference transcripts when mapped to the genome, introduced the concept of *subtle alternative splice events* emerging from *tandem splice sites* and raised a series of questions. How can two splicing signals in such an extreme proximity be differentiated by the splicing machinery? Which function can alternative mRNAs provide, being only different in a triplet? In contrast, the triplet variation GYNGYN at the donor site of exons is much less frequently found (83). The fraction of transcripts confirming this tandem donor is with 1.4% a multiple lower compared to the fraction of transcripts confirming alternative splicing at NAGNAG acceptors (17.6%)(84). However, there is a subtle variation at the donor site which occurs strikingly frequent. The repertoire of donor splice site variations is severely dominated by a four nucleotide variation (81), which occurs more frequently than alternative splicing at the GYNGYN or at the NAGNAG motif. The donor splice site itself forms a consensus motif, which in higher eukaryotes contains two splice signals forming the core motif GYNNGY. As the 5′ss is the first signal that is recognized by U1 snRNP during spliceosome assembly it is conceivable that this donor motif is under selection pressure to maintain the crucial complementarity to U1 snRNA. However, the question arises whether this complementarity alone can explain the shift from splicing at the distal (upstream) to splicing at the proximal (downstream) GY within this motif. Support from experimental studies regarding alternative splice forms that emerge from tandem splice sites is still very sparse. Hence, similarities and differences between overlapping, non-overlapping and constitutive splice sites remain to be delineated and have been addressed by the work in this chapter.

2.1.2 Proposed Mechanisms for Regulating the Donor Splice Site

In addition to the general models of splice site selection which make a basic distinction between the pairing of splice sites across the exon („exon-definition“) or the intron („intron-definition“) (85), several other models have been reported. One model considers the concentration of the antagonistic splicing factors SF2/ASF and hnRNP A1, such that higher doses of SF2/ASF enhance simultaneously U1 snRNP binding to two competing splice sites (leading to use of the downstream ss) whereas higher doses of hnRNP A1 decrease U1snRNP affinity at both donor sites (86, 87). Also normal concentrations of SF2/ASF were shown to be sufficient in binding of U1-snRNP to both splice sites if the 5′ss are close to consensus. In case of high hnRNP occupation U1-snRNP can still bind a splice site (because of higher affinity) but is shifted to that 5′ss that is closer to the nearest high-affinity enhancer binding site. This model cannot be applied directly to splicing at overlapping donor splice sites, since sterical hindrance would prevent any case of double occupancy. However, it is conceivable, that high affinity enhancer binding sites exist exclusively in proximal polarity (near to the weaker downstream donor of two overlapping donor sites), pulling U1-snRNP to this end of the tandem donor motif. This pull could even be increased if exon silencer locate in distal location to the weak donor, impairing U1-snRNP binding to the stronger splice site. Similar observations were made by Bai *et al.* at the 3′ss where ASF/SF2 promoted use of a proximal 3′ss and hnRNP A1 the distal 3′ss in a CGRP transcript in vivo (88).

Another model of 5′ss selection was suggested based on a competition assay between 5´ss subclasses of strong, intermediate and weak strengths of 5´ss and taking into account the free binding energy of U1 snRNA as well as exon- and intron binding splice factors (89). They demonstrated that splicing efficiencies and 5′ss selection are dependent on whether the competing donor sites belong to different subclasses and upon their ability to form G·ψ base pairs. However, their analysis also implied decreased splice efficiency for 5´ss that are less than 40 nt apart due to the effect of sterical hindrance of U1 snRNP binding. Finally a mechanism of oriented scanning has been proposed as a model for 5´ss selection (90) where within intron 7 of the human F7 gene several monomeric repeats duplicate an authentic donor site, yet the most upstream donor remains the only selected 5′ss. The distance between these donor sites comprised 37 nucleotides

and thus appearing rather as non-overlapping than overlapping donor sites. The authors showed that after mutating the wild type donor the next available and most upstream located pseudo splice site became activated. This example states an interesting contrast to the above mentioned models and suggest a model of reconstituting or sustained competition between alternative donor sites, both with high complementarity to the U1 snRNA recognition site.

2.2 Methods

2.2.1 Data Set of Alternative Exons

Exons of human and mouse genes were extracted from the HOLLYWOOD database (3). For two different transcripts aligned to a genomic locus, alternative 5′ss exons (A5Es) matched at their 3′ss, but exhibited exactly one short and one long splice form resulting from variation at the 5′ss. Alternative 3′ss exons (A3Es) matched at their 5′ss, but exhibited exactly one short and one long splice form resulting from variation at the 3′ss. Constitutive exons (CEs) were defined as exons of multi-exon genes that have as of date no transcript-supported evidence for undergoing any type of AS. In all AS events, A5Es, A3Es and CEs are internal exons, i.e., all exons of a transcript except the first and last one, because these are flanked only by one splice site and may contain additional regulatory sequences of splicing initiation and termination. Each exon had to be flanked by /GT or /GC type splice sites at the donor site and AG/ type splice sites at the acceptor site. U12-type introns were excluded from this analysis, because of their low fraction (less than 1% of the human introns). Figure 2.1 gives an overview of the data preparation and applied methodology of this chapter.

2.2.2 Spliced-Alignments

Manual inspection of A5Es with short extensions ($E \leq 6$ nucleotides), originally excluded in HOLLYWOOD, revealed a substantial amount of putative alignment artifacts due to misaligned nucleotides close to exon-intron junctions (see Appendix B.1). Alignments were derived for ESTs by the SIM4 program (91), and were corroborated in a recent performance study of spliced-alignment algorithms (92). In particular, examples were found, where SIM4 introduced shifts of EST nucleotides between genomic donor and acceptor sites at genomic loci that encode short varying alternative exon (*cf.* Figure 2.2). To decrease the number of spurious alignments in the dataset of A5Es and A3Es, the original ESTs were used and created new transcript-to-genomic alignments, by utilizing two different algorithms: *i*) BLAT (93), as stored in the UCSC database (http://genome.ucsc.edu); and *ii*) EXALIN (92), with the parameter set $(m, n, q, r, x) = (25, 25, -25, -25, \text{ and } -25)$. Manual inspection of control samples in the alignment results confirmed a clearly improved quality in the correct exon-intron boundary recognition. In all, about 35% of all initial A5E

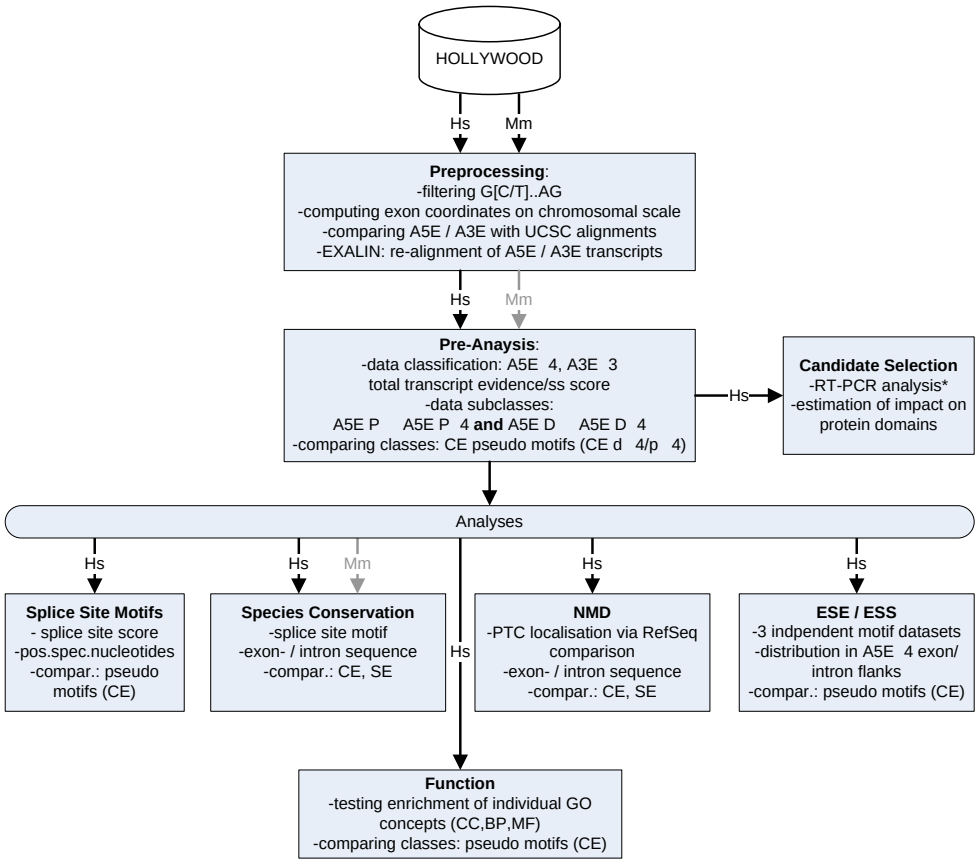


Figure 2.1: Overview of methodology applied for analysis of subtle donor splice sites

predictions (9%) of A5EΔ4 splicing exons could be confirmed with very similar numbers by both BLAT and EXALIN alignments (Table 2.1), showing at least one qualitative aligning transcript. Increasing the stringency to require at least two transcripts confirming each donor variant, reduced the number of EXALIN refined A5E further to 27% (504/1,868) with a fraction of 7.7% (39/504) of A5EΔ4 splicing exons. Subsequent analyses were performed using the subset confirmed by all three alignment methods, with at least one transcript for the minor variant, to keep the dataset in a reasonable size.

2.2.3 Classification of Major and Minor Tandem Donors

The number of transcripts that aligned either to the distal $N(\mathrm{d})$ or proximal $N(\mathrm{p})$ donor was used to classify A5Es. To this end, *i*) the ratio R ($0 < \mathrm{R} < 1$) of the lower over the higher transcript coverage was calculated as

$$R = \frac{N(\mathrm{d})}{N(\mathrm{p})} \text{ if } N(\mathrm{d}) < N(\mathrm{p}) \tag{2.1}$$

or

$$R = \frac{N(\mathrm{p})}{N(\mathrm{d})} \text{ if } N(\mathrm{p}) < N(\mathrm{d}) \tag{2.2}$$

where cases of R = 1 if $N(\mathrm{d}) = N(\mathrm{p})$ were discarded, because accounting for only 2.3% *ii*) each donor was defined as „major“ if it was oberved to be spliced in at least twice as much transcripts as the adjacent alternative splice site, which then became the „minor“ splice site per definition. Hence, in equation 2.1 all donors reaching a treshold of $R_T \leq 0.5$, defined the proximal donor as major splice site, while the same treshold in equation 2.2 defined the distal donor as major splice site. Further, in this analysis a minimal coverage of at least one transcript was required.

2.2.4 Statistical Analysis of Splice Sites

The deviation of splice sites from the consensus is quantified by a maximum-entropy scoring model, implemented in MaxEntScan and publicly available (94) (see also chapter 1.2.4). The 5′ss model incorporates the last three (first six) nucleotides of the exon (intron), and the 3′ss model incorporates the last 20 (first three) nucleotides of the intron (exon). Sequence logos and pictograms were computed and displayed using the Weblogo tool with finite-sample size correction (95).

P-values of splice site frequencies where calculated as follows: 1) frequencies of occurrences at the considered at PΔ4 and PΨ4 splicing exons, as well as DΔ4 and DΨ4 splicing exons, where compared by a 4x2 contingency table and χ^2-test; 2) statistically significant positions were selected at $P < 0.05$; 3) at the same position, the nucleotide (maximally two nucleotides) with the largest difference of the frequency of occurrence between two types (e.g., PΔ4 and PΨ4) was subsequently tested against the remaining nucleotides by 2x2 contingency table and χ^2-test, where $P < 0.05$ was considered as statistically significant.

The information along a sequence was calculated as the relative entropy, as described in Equation 1.3 (Chapter 1.2.4).

2.2.5 Identification of Non-Sense Codons

For each A5EΔ4 splicing exon, the longest cDNA that mapped to the corresponding gene with annotated CDS start and end position was taken as a reference sequence. In most cases such a reference was only available for either the proximal or distal alternative splice form. Identification of mRNAs with the potential to trigger NMD was performed, by comparing the reading-frame after splicing at each tandem donor. Tandem events led to a new reading-fame, the first downstream non-sense codon of which was detected and analyzed for PTCs occurring more than 50 nucleotides upstream of the last exon-exon junction to elicit NMD (10, 96).

2.2.6 Detection of Sequence Conservation

The core of A5EΔ4 splicing exons was matched against mouse genomic DNA (version *mm03*), using BLAST with parameter values *-a2 -gT -W10 -q-2 -r3 -e0.001*. Significant matches of similarity were filtered for canonical splice sites and the exon flanking regions of 200 nucleotides were extracted from the genomic sequence. Subsequently, orthologous human and mouse intron regions were aligned using the DNA BLOCK ALIGNER (97), with *-nomatchn -gap 0.02 -blockopen 0.2 -umatch 0.05 -pff*, which detects block of conserved sequences located at possible different positions relative to splice junction. The sequence position of detected blocks of conservation was parsed and recorded with the script DBA-PARSER (Holste, unpublished data) and plotted in a region of 100 nucleotides, with a moving-average of ten nucleotides. Exon conservation was determined by the score (S_{ort}) from CLUSTALW alignments, self-alignment of the larger exons to yield the score S_{id}, and calculation of the normalized score $S_{tot} = S_{ort}/S_{id}$.

2.2.7 Experimental Assay

Experimental confirmation of tandem splice forms was performed at the Fritz-Lipmann-Institute (FLI) Jena by Stefanie Schindler and Karol Szafranski, according to the following protocol (4):

i. RT-PCR amplification: For validation of splice variants, nested PCR was performed using 100 ng cDNA templates from the Human Multiple Tissue cDNA Panels I and II (BD Biosciences). Splice variants were enriched for EST originating from different cDNA libraries and, for a given gene, suitable

tissues were chosen according to the origin of ESTs for the minor splice variant or the expression profile found in the Stanford SOURCE data base (98). Primers were obtained from Metabion. Nested RT-PCR reactions were set up with ReadyToGo PCR beads (Amersham) and 10 pmol primer in 25 μl total volume, according to the manufacturer's instructions. The thermocycle protocol was 1 min 30 sec initial denaturation at 93°C, followed by 25 cycles of 40 sec denaturation at 93°C, 40 sec annealing at 55°C, 1 min extension at 72°C, and a final 4 min extension step at 72°C. In the second round of nested PCR, 2 μl first-round product was amplified for 30 cycles. Ethanol-precipitated PCR products were directly sequenced using target-specific forward and reverse primers;

ii. Sanger sequencing: Reactions were set up with 200 ng template DNA, 10 pmol primer, and BigDye v3.1 (Applied Biosystems) in 10 μl final volume, according to the supplier's instructions. The thermocycle protocol was 5 min initial denaturation at 95°C, followed by 29 cycles of 30 s denaturation at 95°C, 10 s annealing at 55°C, 4 min extension at 60°C. After ethanol precipitation, automated sequence separation and detection was done on an ABI 3730XL sequencer. Electropherograms were processed by PHRED (99). After automated assembly (Staden package, (100)), sequence variations were verified by manual inspection using GAP4 (Staden package).

2.2.8 Presence of Predicted Splicing-Regulatory Elements

Searching for splicing regulatory elements in exon flanking regions was performed by using the following data sets: 176 predicted exonic splicing silencers identified in Wang *et al.* (101), 753 predicted intronic enhancers and/or silcencers identified in Yeo *et al.* (102), and 1,013 putative exonic splicing silencers identified in Zhang *et al.* (103). All elements were searched for in a region of 100 nucleotides flanking proximal tandem donors, and exact matches were counted in non-overlapping sequence windows of 20 nucleotides.

2.2.9 Gene Ontology (GO) Annotations

GO-terms for genes with A5EΔ4 splicing exons (358 GO terms), A5Es (1,414), and CEs (3,655) were obtained from the Ensembl database (www.emsembl.org), corresponding to 129 and 1,283 genes with A5EΔ4 splicing exons or A5Es, respec-

tively, and 8,664 genes of a control set. GO annotations for A5EΔ4 splicing exons of 129 of 166 genes (representing the total set of 171 A5EΔ4 splicing exons) were mapped, and the most frequent category annotations „*molecular function*“ and „*biological process*“ were selected; in decreasing order: „*ATP binding*“, „*Zinc ion binding*“, „*Regulation of transcription, DNA-dependent*“ , „*Transferase activity*“, „*Signal transduction*“, „*Hydrolase activity*“, „*RNA binding*“, „*Protein binding*“, „*Transcription factor activity*“ and „*DNA binding*“. In order to compare the GO annotations of A5EΔ4 genes against a control, 10,000 genes with at least one pseudo splice site, dΨ4 or pΨ4 splicing exons (each comprising 129 genes) were sampled and the frequency of occurrence of a certain GO term was computed. The statistical significance (P-value) was calculated analogous to (102), by assessing the frequency of occurrence that a certain GO-term was present in the control more frequently than in the A5EΔ4 gene set, divided by 10,000. The outcome showed the following categories as significant at the 0.05 percent level: „*Signal transduction*“ (PΔ4/dΔ4 vs 5´ss/dΨ4, 0.07; DΔ4/pΔ4 vs 5´ss/pΨ4, 0.15), „*RNA binding*“ (0.0004; 0.003), „*GTP binding*“ (0.02; 0.04), „*Electron transport*“ (0.02; 0.03), „*Protein biosynthesis*“ (0.01; 0.03), „*Signal transducer activity*“ (0.04; 0.08). To correct for multiple testing, a (conservative) Bonferroni correction (104) were applied, the P-value chosen was divided by the number of performed tests, and GO-terms occurring with $P_c < 0.05/10 = 0.005$ were considered as significant.

2.3 Results

2.3.1 Biased Extensions of Alternative 5′ss and 3′ss Exons

Exon-skipping is the most prevalent AS type produced by the human spliceosome, as well as by all other mammals investigated to date, when averaged across different organ systems and cell types that can exhibit tissue-enriched splice forms (13, 105). Internal alternative exons that involve exclusively either the 3′ss (A3Es) or the 5′ss (A5Es) are also abundantly produced, while the simultaneous alteration of 3′ss and 5′ss (producing exons that overlap but match neither splice site) are markedly less frequent. For A5Es the most distal splice site defines the exon core, while proximal sites (if more than one alternative choice is possible) are exon extensions only included in selected mRNAs.

Out of a collection of $\sim 37,400$ transcript-inferred human alternative exons maintained in the HOLLYWOOD database (3), AS events of about 10,300 A5Es and 9,200 A3Es were filtered for short/long exon splice variants of solely one proximal/one distal 5′ss, while being constitutively spliced at the opposite site, and resulted to 5,275 A5Es and 4,497 A3Es. Stringent alignment criteria were imposed on all transcripts: 1) ESTs were required to overlap at least one the co-aligned cDNAs; 2) the first and last aligned segments of ESTs were required to be at least 30 nucleotides in length with 90% sequence identity; 3) the entire EST sequence alignment was required to extend over at least 90% of the length of the EST with at least 90% sequence identity; and 4) realignments of ESTs with two other algorithms were required to agree in three out of all three independent alignments (see below, as well as Methods). The resulting dataset of identical computational inferences of three methods contained 1,868 ($\sim$ 18%) A5Es and 3,301 ($\sim$ 36%) A3Es.

Alternative exons were subdivided into their core and extension parts, where the latter is the sequence between the distal and proximal splice sites. The extension (E) included lengths up to about 250 nucleotides, with quickly decreasing transcript coverage/utilization as E increases. Larger extensions existed, albeit with barely more than a few transcripts. For the sake of simplicity, the boundary between A5E (A3E) overlapping and non-overlapping splices was defined at $E > 6$ ($E > 18$) nucleotides and the distribution $f(E)$ for $E = 1, 2, ..., 18$ nucleotides was displayed in a window across the boundary region. Noticeably, the obtained distribution $f(E)$ for both A5Es and A3Es was highly biased for extensions with overlapping splice sites. Figure 2.2 shows (in the upper-left

panel) that for extensions at the 5′ss the bias is caused predominantly by a peak at $E = 4$ nucleotides. It further shows for A5Es that short extensions exhibit a small but persistent pattern periodically occurring at $E = 6, 9, 12, 15,$ and 18 nucleotides, all multiples of three, and thus, preserving the reading-frame. These patterns of AS for short extensions were in accord, both qualitatively and in good approximation quantitatively, in an independent, comparative analysis for the mouse *Mus musculus* (Figure 2.2, lower-left panel). Overall, the median sizes of inferred alternative exons showed that SEs and A5Es tend to be shorter than CEs and A3Es, while overlapping and skewed to larger sizes (Figure 2.3).

Somewhat unexpectedly, Figure 2.2 was also indicative that different splice-alignment algorithms gave rise to strikingly different outcomes, particularly when faced with alignments involving short extensions. The fraction of short extensions ($E \leq 6$) ranges between 17-38% among several standard algorithms, with. SIM4 predicting most liberally almost 40% of A5Es as such donor variations. This alignment algorithm also suggests a strong tendency toward $E = 4$ nucleotides (27% *cf.* Table 2.1). A conservative approach was taken to substantiate the identified A5E events, by realigning all corresponding transcripts to the same genomic sequence with two other algorithms, EXALIN and BLAT (the latter lacks an explicit splice site model). The results showed that for $E = 4$ the proportion of A5E events derived from SIM4 ($\sim$28%) was markedly higher than alignments derived from EXALIN or BLAT - yet the bias for extensions was consistently shown at $E = 4$ nucleotides, though with a lower proportion of $\sim$9% (Table 2.1). Manual inspection of selected SIM4 alignments showed apparent sequence inconsistencies, when compared to the secondary alignments (see Figure B.1). In all, 1,868 of 5,275 A5Es were taken for further analysis, where $\sim$9% (171/1,868) accounted for $E = 4$ nucleotides extensions.

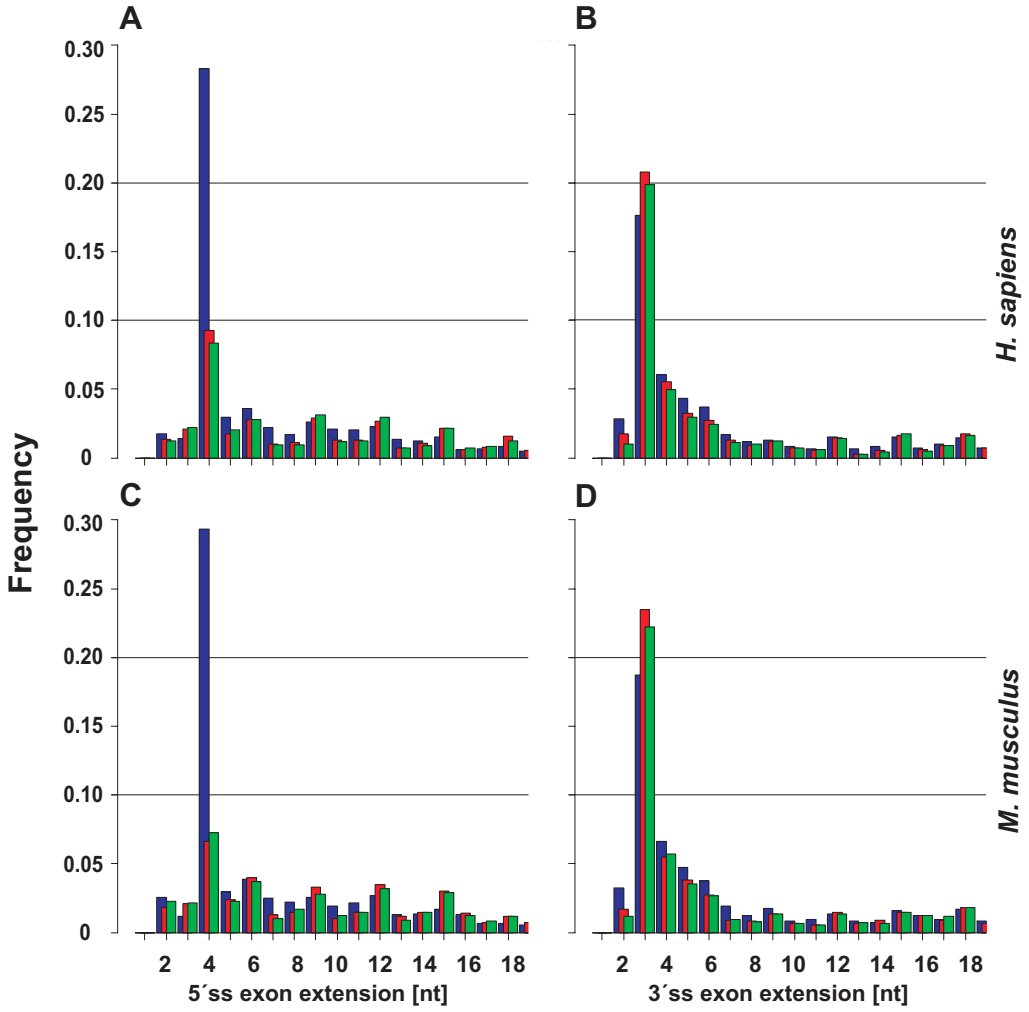


Figure 2.2: Occurrence of extensions ($E = 1, 2, ..., 18$ nucleotides) for A5Es **(A,C)** and A3Es **(B,D)**, with human and mouse exons in the top and bottom panels, respectively. Extensions were inferred from three different alignment algorithms (colored as blue, SIM4; red, BLAT; and green, EXALIN) of cDNAs/ESTs to genomic DNA. The distribution $f(E)$ for A5Es was markedly biased for extensions (E) with overlapping splice sites, with a peak at $E = 4$ nucleotides. Exon extensions exhibited relatively smaller but persistent periodic peaks at $E = 6, 9, 12, 15,$ and 18 nucleotides. $f(E)$ for A3Es also displayed a bias for overlapping splice sites, with a peak at $E = 3$ nucleotides and smaller peaks at 4-6 nucleotides. The program SIM4 predicted significantly more extensions at $E = 4$ nucleotides as compared to BLAT and EXALIN predictions of the same initial set of cDNAs/ESTs, which was indicative of spurious alignments. A comparative analysis of alternative exons in *M. musculus* corroborated the above patterns.

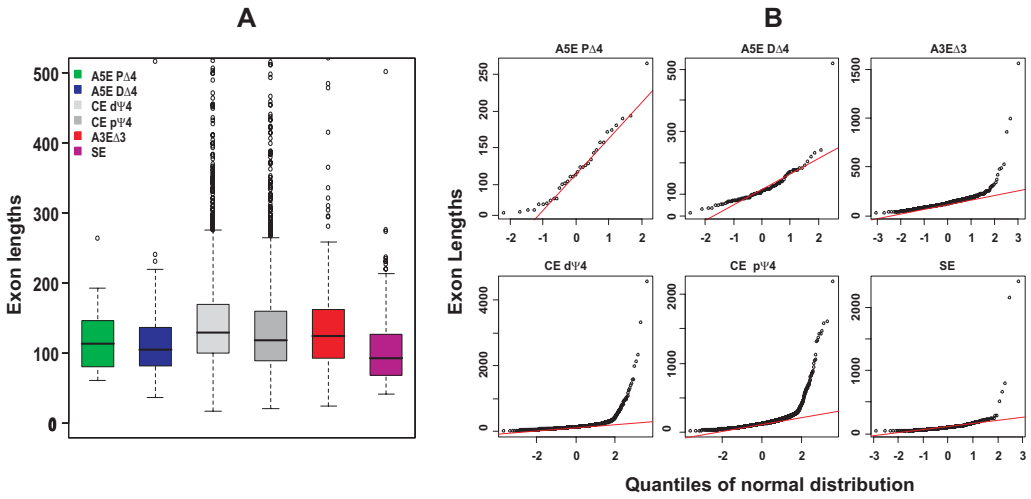


Figure 2.3: Length distribution of human exons. **(A)** box-plots of constitutive and alternative exon lengths (median and interquartile range, „whiskers“ denote the 1.5-fold interquartile range). Major-proximal (PΔ4) and major-distal (DΔ4) splicing exons, constitutive 5′ss with distal pseudo (dΨ4) and proximal pseudo (pΨ4) splicing exons, and skipped exons (SEs). **(B)** quantile-quantile-plots of exon lengths (x-axes, quantiles of normal distribution; y-axes, empirical data).

Table 2.1: Characterization of alternative donor splice events, involving exactly two different exon isoforms. Column one describs the initial dataset from the HOLLYWOOD database (aligned with SIM4, column two and three the same dataset after refinement by **i)** realigning the transcript isoform with a more recent spliced alignment algorithm (EXALIN) and **ii)** by computing the overlap of exon isoforms with those stored in the UCSC repository (computed as BLAT alignments), respectively. tx = transcript; E = extension = nucleotide difference between two A5E isoforms.

	HOLLYWOOD (SIM4)		**UCSC** (BLAT)		**FSU** (EXALIN)	
	min. 1 tx	>1 tx	min. 1 tx	>1 tx	min. 1 tx	>1 tx
total	5,275	1,284	1,926	NA	1,868	504
$E \leq 6$	2,011 (38.1%)	555 (43.2%)	333 (17.3%)	NA	324 (17.3%)	92 (18.2%)
$E = 4$	1,493 (28.3%)	406 (31.6%)	179 (9.3%)	NA	171 (9.2%)	39 (7.7%)
$E = 3$	76 (1.4%)	18 (1.4%)	40 (2.1%)	NA	40 (2.1%)	14 (2.7%)

In order to compare these findings with A3E events, the distribution of short extensions was obtained and a similar, albeit distinctively different patterns identified (Figure 2.2, upper-right panel). Figure 2.2 shows that $f(E)$ exhibits clear peak at $E = 3$ nucleotides, with successively smaller peaks at $E = 4, 5,$ and 6 nucleotides. Again, these AS patterns were corroborated in a comparative analysis for *M. musculus* (Figure 2.2, lower-right panel). The extension preference of alternative 5′ss and 3′ss exons is in accord with previous studies, where in particular $E = 3$ nucleotides for A3Es had been examined and found to obey the splicing pattern at the NAG/NAG/ motif.

2.3.2 Tandem Donors and Acceptors

Patterns of A5Es and A3E extensions with overlapping splice sites are interesting in their own context, because they are 1) possibly differently regulated than non-overlapping alternative donor or acceptor splice site exons (89, 106); and 2) predictive of different downstream effects of AS, resulting in differentially preferred modes of alternative splicing at the 5′ss (predominantly out-of-frame splicing) and the 3′ss (predominantly in-frame splicing). For overlapping 5′ss and 3′ss are mainly represented by extensions of four and three nucleotides, respectively, hereafter they are denoted by „A5EΔ4“ tandem donors with $E = 4$ and similarly by „A3EΔ3“ tandem acceptors with $E = 3$ nucleotides. This study focuses on tandem donors with respect to sequence features that are known to be involved in the recognition of 5′ss, and compares them to 3′ss of alternative exons as well as constitutive exons including potential pseudo splice sites.
Generally, the basic recognition and binding to 5′ss incorporates intronic (involving positions from 1 to 6) and exonic nucleotides (positions from -3 to -1). The consensus motif for 5′ss of mammalian genes is known as CAG/GTRAGT (at positions $P_{-3}P_{-2}P_{-1}/P_1P_2P_3P_4P_5P_6$) with R standing for purine bases. This nine nucleotide-long motif is highly degenerated and, in fact, in the present data set of human exons only proportions of ∼0.9% (966/113,386) and ∼1.3% (1,431/113,386) of inferred constitutive exons exhibited exact matches to the motifs CAG/GTAAGT or CAG/GTGAGT, respectively. Figure 2.4 illustrates splice sites and utilization of tandem donors for three selected human genes:

A. The gene *RAD9A* (Ensembl gene-identifier *ENSG00000172613*) is a homolog conserved from yeast (*S.pombe*) to human, which encodes a cell cycle-check point control protein that is required for cell-cycle arrest and DNA damage repair. The primary transcript sequence of *RAD9A* exhibited two alternative, overlapping 5′ss at exon E8, identified as CAG/**G<u>C</u><u>A</u><u>G</u>**/GT at the distal 5′ss and CAG/GTA<u>GT</u>T at the proximal 5′ss that extends E8 (non-consensus nucleotides are underlined; exon extension bolded). The distal and proximal 5′ss gave rise to three and 17 mRNAs, respectively, which aligned to the primary transcript structure of *RAD9A*. In addition to the tandem donor pattern, Figure 2.4 shows the splice site strength, quantified by the MaxEntScan score (see Methods 2.2.4), and the conservation profile across exons and intron, quantified by the Phastcon score (107) computed across several genomes (from *P. troglodytes* to *T. rubripes*). Lo-

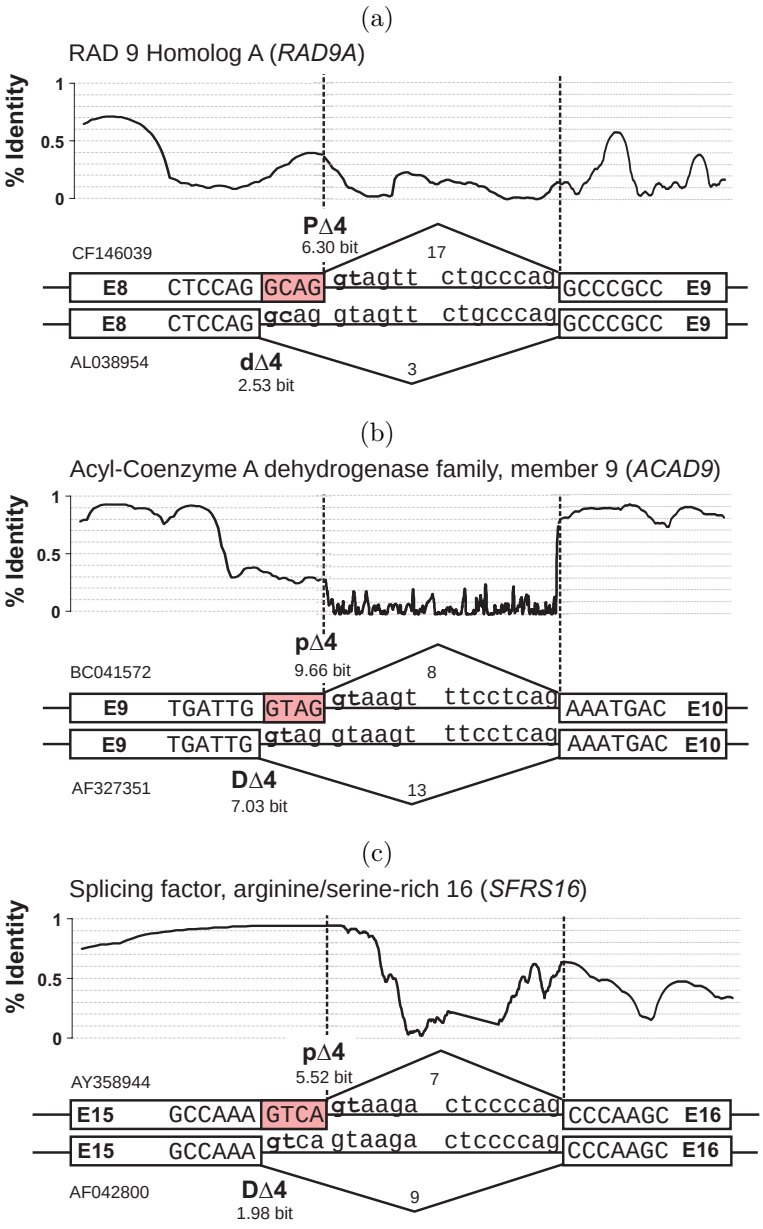


Figure 2.4: Illustrative examples of inferred tandem donors. White boxes denote exon and lines intron nucleotides; exon numbers (E#) corresponded to 5′-to-3′ enumerated RefSeq annotations, the splice site score as measured by MaxEntScan, and the transcript coverage of the proximal and distal donor site corresponded to the number of aligned sequences. In **(a)**, E8 of the *RAD9A* gene shows a tandem donor with extension /GCAG/; in **(b)** E9 of the *ACAD9* gene shows a tandem donor with extension /GTAG/; in **(c)**, E15 of the *SFRS16* gene shows a tandem donor with extension /GTCA/. Tandem donors in **(a)** and **(c)** were preferentially included in transcripts. The conservation plot (Phastcon scores, not in scale with the stated exon and intron nucleotides) covers A5EΔ4 splicing exons, as well as adjacent introns and downstream exons, and shows alternating patterns of high/low levels across all three examples.

cal regions of high levels of sequence conservation for exons compared with the intron are apparent.

B. A tandem donor was detected for E9 (TTG/**GTAG**/GT and TAG/GTAAGT) of the *ACAD9* gene (*ENSG00000177646*), which encodes a member of the acyl-CoA dehydrogenase gene family and plays a role in lipid catabolism. The distal and proximal 5′ss gave rise to 13 and eight mRNAs, respectively. Figure 2.4 shows for E9 consistently elevated levels of sequence conservation.

C. The arginine/serine-rich splicing factor 16 (*ENSG00000104859*) showed a tandem donor at E15 (AAA/**GTCA**/GT and TCA/GTAAGA). Distal and proximal 5′ss choice gave rise to nine and six mRNAs of *SFRS16*, respectively. Figure 2.4 shows that the level of sequence conservation of E15 steadily rises toward the 3′-terminus and extends well across the exon-intron junction to I16, before it rapidly decays, which is indicative of conservation with splicing-regulatory function (102).

2.3.3 Experimental Validation of Tandem Donors

Having obtained sufficient evidence from stringent transcript alignments, validating the functional utilization of tandem splice sites from independent lines of evidence was pursued. To this end, first publicly available literature[1][2] was searched for AS events involving short 5′ss extensions. Yet only a very limited number of reported cases of splice variants with short extensions that could be traced back to tandem donors was found. The human *Clasp* gene (known synonyms are *SFRS16*, *SWAP2*), for instance, encodes the Clk4-associating arginine/serine-rich (SR)-related protein that binds to the family of CDC2-like kinases (108, 109). The 5′ss of E15 of the *Clasp/SFSR16* is an alternative tandem donor, which gives rise to the splice forms *ClaspS* (with the extension **GTCA**) and *ClaspL* (without). Both isoforms differ by 246 nucleotides, where *ClaspS* carries a PTC due to out-of-frame splicing and thereby omits a third RS-domain encoded by *Clasp/SFSR16*. Both isoforms were tissue-enriched in the mice brain and testis, and displayed different intra-nuclear locations, possibly controlled by the third RS-domain (108). Another AS event involving tandem splice sites has been detected in the human growth hormone (GH) gene cluster, whose expression is

[1] http://www.pubmed.org
[2] http://apps.isiknowledge.com/

developmentally controlled. The gene *GH-V* differentially expressed three isoforms in the placenta and testis, one of which is due to a tandem donor splice site (/**G<u>TG</u>G**/GT) of exon E4; the tandem site was not sequence-conserved in the remaining four family members (G<u>GG</u>G/GT). The use of the distal out-of-frame splice site caused a reading-frame shift of E5 downstream, which, in turn, overread the original termination codon and utilized a new („delayed“) termination codon further downstream. Overall, the original splice variant and GH-V/Δ4 shared 124/219 and differed by 95/219 amino acids.

Clearly, the detection of alternative tandem splice site exons is hampered due to the high similarity of isoforms and often only detectable by direct sequencing and protein sequence analysis. Consequently, an experimental assay was performed by cooperation partner at the FLI Jena to explore directly the splicing patterns of computationally identified alternative tandem donors. Table 2.2 list the names of a set of 14 genes with tandem acceptors (8% of total), which were manually selected from known genes exhibiting a varying degree of transcript coverage (ranging from one to 35 transcripts for tandem splice site usage) and tested in a battery of human organ systems and cell types by RT-PCR primers targeted to the flanking exons; panels of ten normal tissue samples (from the testis, brain, colon, heart, kidney, small intestine, spleen, thymus, ovary, and leukocytes) were assayed. The products of these 45 RT-PCRs were used to verify the identity of these PCR products by sequencing (see Figure 2.4, as well as Methods 2.2.7). For instance, Figure 2.5 shows for E15 of SFRS16 schematically the gene structure, proximal and distal sites of the tandem donor, and the sequence electropherogram interrogated in samples derived from the human spleen and blood. Upstream of the E15 tandem donor, both transcript sequences identically overlap and thus, cannot be distinguished in the electropherogram; downstream, two nucleotide signals appear above the base line, indicating the presence of two isoforms.

Table 2.2: Summary of the experimental assay for validating computationally inferred human tandem donors. A5EΔ4 splicing exons were selected according to both transcript coverage, concordance of tissues inferred from cDNA-libraries of A5EΔ4 genes, and commercially available samples. RT-PCR primers were targeted to flanking exons, assayed, and sequenced. In the last column, „+“ indicates that the tested A5EΔ4 splicing exon was detected to be present in both splice variants of the corresponding samples, separately for each tested tissue (a bolded „+“ indicates the major form). In all, 7/14 A5EΔ4 splicing exons were verified in panels of nine normal tissues. In the fourth column (PTC), „+“ indicates the presence of a premature termination codon.

Ensembl gene (*ENSG00000#*)	**Gene name**	**Region**	**PTC**	**Transcript coverage** (distal/ proximal)	**Analyzed tissues**	**Confirmed donors** (distal/ proximal)
172613	*RAD9A*; RAD9 homolog	CDS	+	3/17	Kidney; Leukocytes	(+/**+**); (+/**+**)
175605	*ZNF32*, zinc finger protein 32	CDS	+	14/2	Heart; Leukocytes	(**+**/+); (**+**/+)
104859	*SFRS16*, arginine/ serine-rich splicing factor 16	CDS	+	9/7	Leukocytes; Spleen	(+/**+**); (+/**+**)
161574	*CCL15*, small inducible cytokine A15 precursor	CDS	+	35/6	Colon	(**+**/+)
177646	*ACAD-9*, Acyl-CoA Dehydrogenase Family, mitochondrial Precursor	CDS	+	13/8	Brain; Heart	(**+**/+); (+/-)
148459	*PDSS1*, Trans-Prenyltransferase	CDS	+	6/2	Small intestine	(+/+); (**+**/+)
180198	*RCC1*, regulator of chromosome condensation	5´UTR	+	4/2	Small intestine; Testis	(**+**/+); (-/+)
170581	*STAT2*, signal transducer and activator of transcription 2	CDS	+	8/1	Brain; Thymus	(+/-); (+/-)
102878	*HSF4*, heat shock transcription factor 4	CDS	+	6/1	Colona, Braina	(-/+); (-/+)
090061	*CCNK*, cyclin K	CDS	+	17/1	Leukocytes	(+/-)
137502	*RAB30*, Ras-related Protein RAB-30	CDS	+	1/7	Leukocytes	(-/+)
134987	*WDR36*, WD-Repeat Prtoeine 36	CDS	+	1/4	Leukocytes	(-/+)
157911	*PEX10*, peroxisome assembly protein 10	CDS	+	3/18	Brain	(-/+)
049656	*CLPTM1L*, cisplatin resistance related protein CRR9p	CDS	+	2/32	Ovary; Small Intestine	(-/+); (-/+)

Table 2.2 lists the outcome for all 14 genes. In all, 50% (7 of 14 total) of selected A5EΔ4 splicing exons showed PCR-products displaying $E = 4$ nucleotides for the sets of interrogated alternative exons, and the experimentally observed splice

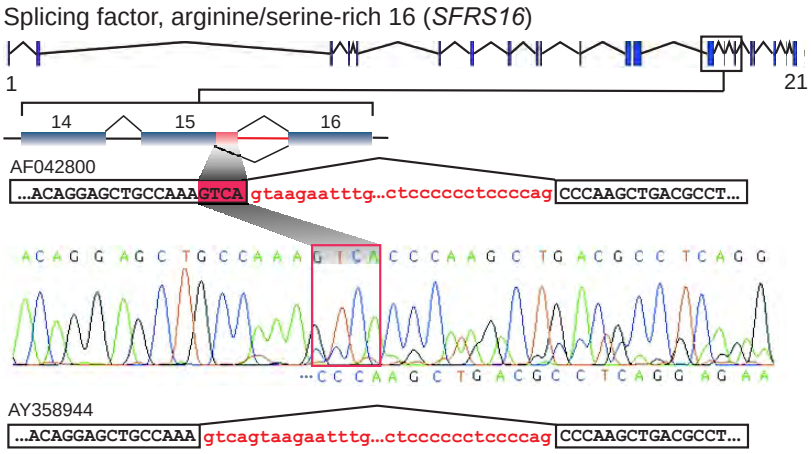


Figure 2.5: Experimental validation of a tandem donor activated in E15 of the *SFRS16* gene using RT-PCR and direct sequencing. The top shows the gene structure of *SFRS16*; in the middle and bottom, E14-16 are schematically extracted and the 3′-end core and full extension sequence of E15 for proximal (TCA/gtaaga) and distal (AAA/**gtca**gt) splicing are shown. Prior to reaching the 5′ss of E15, both mRNA isoforms cannot be distinguished and consequently the electropherogram displays, for each position, one nucleotide signal peak above the base line. After the tandem donor site, two nucleotide signals above the base line become visible, indicating the presence of two isoforms.

ratio between minor and major form was in agreement with the ratio suggested by EST data. Six of seven A5EΔ4 splicing exons could be mapped to protein coding gene sequences and all six CDS affecting alternative exons create a PTC. For human tissues samples were tried to match EST-associated cDNA libraries, using a larger battery of different organ systems and cell types might validate additional A5EΔ4 splicing exons and, therefore, this approach is rather delivering a lower boundary of the presence of AS events involving tandem donors.

2.3.4 Two Distinct Levels of A5E Proximal and Distal Splicing

Studies of the inclusion and exclusion of skipped exons of the human and mouse genomes have shown that SEs can be broadly subdivided into two types: SEs that are included in the majority of transcripts (termed „major-form“), and those that are predominantly excluded („minor-form“). Interestingly, such SEs posses different splicing and phylogenetic properties (110). Here, it was examined whether this property is more generally related to alternative exons, by analyzing the transcript coverage of 1,816 A5Es with one proximal/one distal 5′ss. Figure 2.6 A shows a scatter plot of the distal against proximal 5′ss

transcript coverage for both tandem and non-overlapping donors; the individual transcript coverage of the distal (proximal) splice site is placed above (on the right-hand side). The scatter plot shows that the number of aligned transcripts ranges from a single transcripts up to more than one hundred, with the average centering on ∼13, and is biased toward lower coverage (median value of 2). The ratio of alternative 5′ss usage was defined as R and computed for human, as well as mouse, A5Es. The inset of Figure 2.6 A shows that the histogram of the $\log(R)$ displays a bimodal distribution, which is indicative of the presence of two types (or subpopulations) of alternative 5′ss exons - one, which is characterized by the utilization of the proximal over the distal 5′ss (type-I), and another by the utilization of the distal over the proximal 5′ss (type-II). This is reminiscent of the „major/minor form“ definition of SEs, albeit here it applies to both A5E proximal and distal splice sites. A threshold of $R_c = 2$ was used to group all A5Es into type I and II, or a remaining type, based on the behavior of R (indicated by lines in the scatter plot of Figure 2.6 A, see also Methods). Having two subpopulations of tandem donors, the major site within a tandem is denoted as „PΔ4“ (proximal) or „DΔ4“ (distal), whereas the notation „pΔ4“ (proximal) or „dΔ4“ (distal) adresses the minor donor. Similarly, for non-overlapping alternative 5′ss, the major donor is denoted as „PΔ“ or „DΔ“ and the minor donor as „pΔ“ or „dΔ“, respectively (*cf.* Table 2.3).

Figure 2.6 B shows the scatter plot of the distal against proximal 3′ss transcript coverage. Here, the points are comparatively larger scattered than in Figure 2.6 A and display an „arrow head“ like structure. Using the same threshold as above, no clear distinction between splice sites for A3Es is found. Rather, the data are consistent with a single population of A3Es, and the inset shows the histogram of R as an approximately unimodal shape with values of R in a similar range as observed for A5Es.

In all, tandem and non-overlapping A5Es comprise a set of 1,641 out of 1,868 (∼88%) exons, remaining ∼12% that either exceeded the threshold or were covered by a single transcript. The density of PΔ and DΔ splicing exons was ∼59% (type-I) and ∼41% (type-II), which was in some contrast to PΔ4 and DΔ4 of type-I with ∼26% (44/171) and type-II with ∼69% (118/171) exons, respectively ($P < 0.0001$; Fisher's exact test). Scatter plots, populations, and histograms were corroborated in a comparative analysis of the transcript coverage for A5Es in *M. musculus* (data not shown).

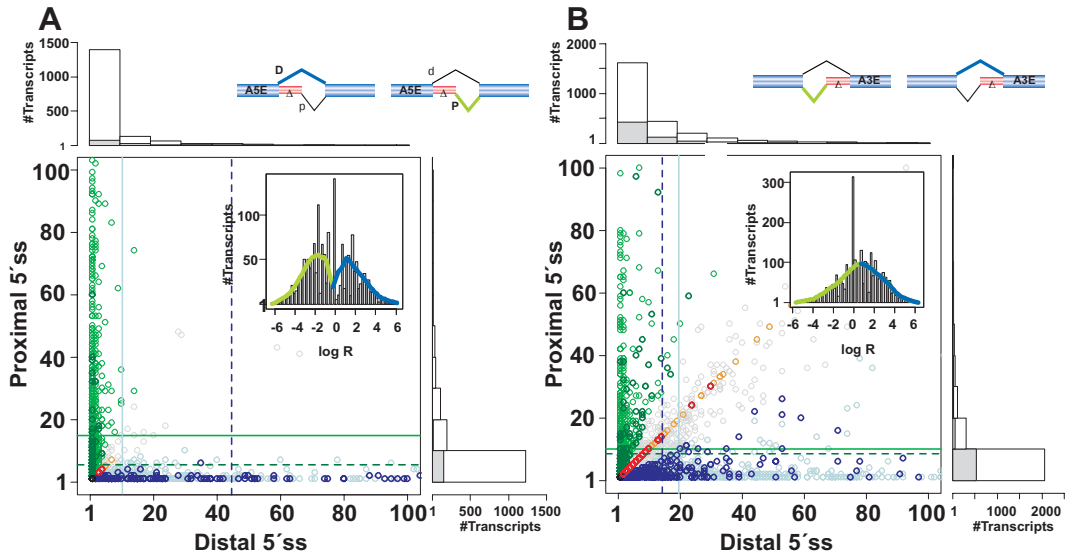


Figure 2.6: Scatter plot of the transcript coverage of non-overlapping and Δ4 tandem donors **(A)** and acceptors **(B)**. Vertical and horizontal axes refer to the coverage of distal and proximal splice sites; solid and dotted lines mark the transcript means; A5EΔ4 and A3EΔ3 splicing exons are bolded, green and blue mark the ΔP and ΔD (major) splicing exons, respectively. The inset shows the histogram of the log-ratio (R) of the coverage of the distal over the proximal 5′ss (3′ss); curves marked in black show the smoothed distribution (splines, R package). In **(A)** the coverage scatters mainly along the vertical or horizontal axis, which is indicative of preferentially including or excluding the exon extension from the core sequence. The coverage pattern was used to partition all A5Es into two main types, I and II, and a remaining type. The inset shows for the histogram of R a bimodal shape, which is indicative of two subpopulations of A5Es with predominant proximal or distal splice site usage. In **(B)** the overlap between distal and proximal tandem acceptor coverage is comparatively broader, and consequently the histogram of R exhibits a unimodal shape consistent with a single population of A3Es.

Table 2.3: Summary of selected features analyzed for A5Es with non-overlapping donors (**a**) and A5EΔ4 splicing exons with tandem donors (**b**), separated into major (PΔ4, DΔ4) and minor (dΔ4, pΔ4) splice forms. Transcript coverage denotes the number of transcripts (full length cDNAs or ESTs) that support either the minor or the major form donor.

(a)

Features of A5Es	**PΔ** (major-form)	**dΔ** (minor-form)	**DΔ** (major-form)	**pΔ** (minor-form)
Number of occurrences	872		598	
in-frame	410 (47%)		257 (43%)	
out-of-frame	462 (53%)		341 (57%)	
Mean core length (nucleotides)	107		126	
Mean extension length (nucleotides)	82		119	
Trancript coverage (mRNA/EST)	3,603 / 19,709	324 / 924	2,186 / 13,126	330 / 556
Average MAXENT score	7.5	-0.5	6.8	4.6

(b)

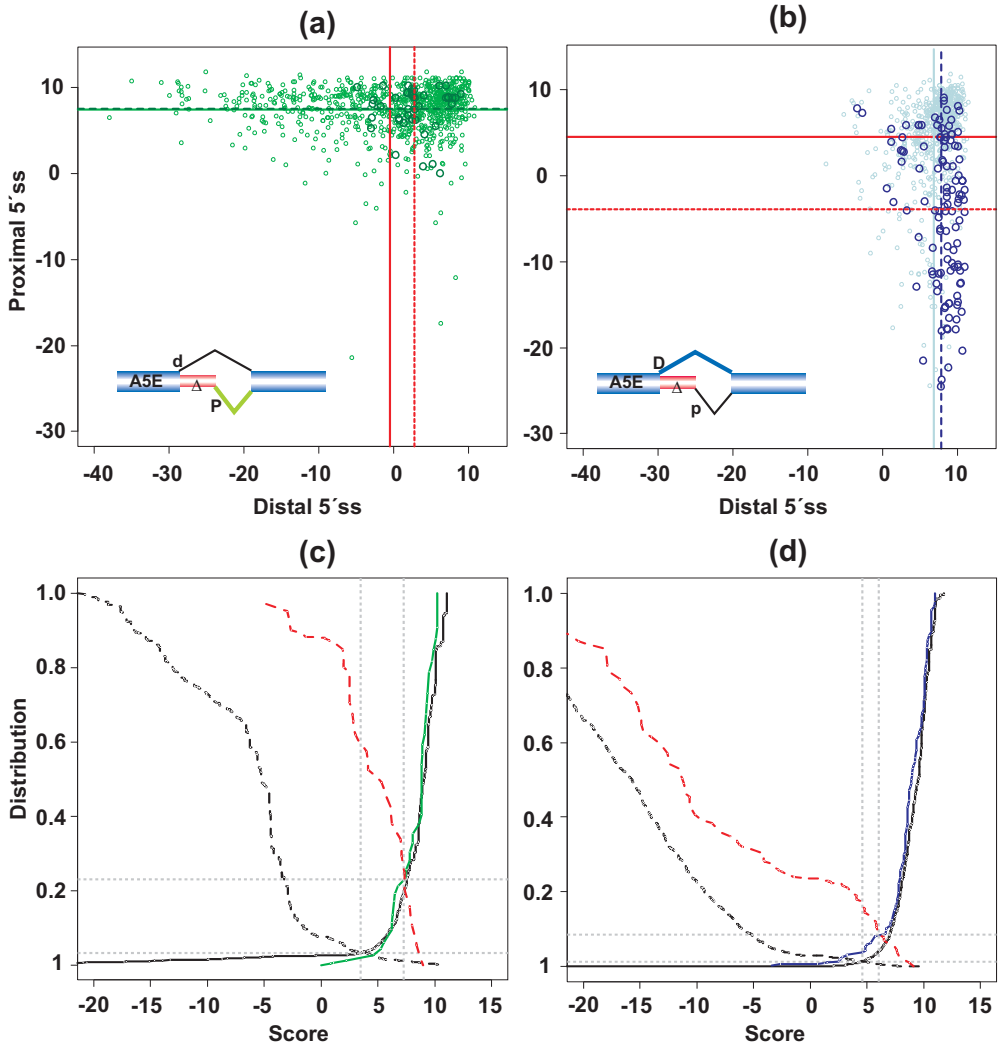
Features of A5EΔ4 exoms	**PΔ4** (major-form)	**dΔ4** (minor-form)	**DΔ4** (major-form)	**pΔ4** (minor-form)
Number of occurrences	44		118	
Mean core length (nucleotides)	122		119	
Mean extension length (nucleotides)	4		4	
Trancript coverage (mRNA/EST)	159 / 619	20 / 46	531 / 7,000	15 / 144
Average MAXENT score	7.5	2.8	7.9	-3.9

2.3.5 Splice Sites of A5Es Score Differently between Type I and II

The relationship between different types of transcript coverage and sequence-complementarity of base pairing to U1 snRNA was analysed by computing the 5′ss score distribution. To this end, a maximum-entropy or Markov-random field, based model was applied, which has been shown to capture additional statistical significant dependencies of splicing signals than standard position-weight matrix representations (111, 94), to score the 5′ss of all A5Es (see also Methods 2.2.4). Figure 2.7A shows for all PΔ and PΔ4 splicing exons of type-I the score distribution, $f(S)$, of the distal against proximal 5′ss. The score is large ($S > 0$) when the splice site is 'close' to the consensus sequence, and small ($S < 0$) when the splice site shows marked deviations from the consensus. For type-I, the scores of most PΔ and PΔ4 splicing exons were positive, ranged up to $S = 12$ (unit of bits), and clustered narrowly around a mean value of $S_{P\Delta} \approx S_{P\Delta 4}$ where $S_{P\Delta 4} = 7.5$ bits (marked by horizontal lines in Figure 2.7A). In contrast, scores of the corresponding dΔ and dΔ4 (the minor-forms) fluctuated more broadly, and mean values were between $\Delta S = 4.5$ and 8 bits weaker than the corresponding

major-form splice site. Interestingly, this trend was reversed for exons of type-II (DΔ, DΔ4), where the score clustered for $S_{D\Delta}$ and $S_{D\Delta4}$ between 7 to 8, yet for minor-forms was again broadly distributed and clustered around $S_{p\Delta} = 4.6$ and $S_{p\Delta4} = -3.9$, respectively. The different patterns of narrow/broad scattering of A5EΔ4 splice site strengths in dependence of their type was corroborated in a comparative analysis of $f(S)$ in *M. musculus* (see Figure B.1).

Observed motifs (/GTNN/GT) of proximal (PΔ4) and distal (DΔ4) tandem splice sites occurred with markedly different proportions (see Table 2.4). To what extent were the observed PΔ4 and DΔ4 splicing exons different from constitutive splicing exons (CEs) with pseudo donors having a „genomic predisposition“ for tandem splicing (but not observed)? This was examined by looking for constitutive 5′ss (/GT) that were flanked by another GT dinucleotide at a distance of four nucleotides either upstream of the authentic 5′ss (denoted as „dΨ4“) or downstream of the authentic 5′ss („pΨ4“). A set of 63,008 constitutive splicing exons (out of 113,386) were searched, that exhibited proximal and/or distal pseudo tandem donors. Assuming position-independent nucleotide concentrations, the expected proportions would be ∼10% (dΨ4) and ∼48% (pΨ4), where the latter reflects the GT motif at positions P_5 and P_6 of the 5′ss consensus. It was found that the frequency of dΨ4 was lower than its expected occurrence and was present only in ∼4% of CEs ($P < 0.001$; z-test), whereas pΨ4 was similar, albeit still significantly different, to the expected occurrence and present in ∼47% of CEs ($P < 0.001$; z-test); a substantial proportion of ∼5% (5,211) was comprised by GYNN/GYNNGY, but was excluded from further analysis to avoid any ambiguity. The distribution $f(S)$ for the above sets showed related differences. The mean scores of PΔ4 and constitutive 5′ss (downstream of dΨ4), $S_{P\Delta4} = 7.5$ and $S_{5'\mathrm{SS}} = 7.9$, were about equally large ($P < 0.13$, Mann-Whitney test), yet $S_{d\Psi4} = -3.6$ was significantly lower as compared with $Sd\Delta4 = 2.8$ ($P < 2.2e-16$). Similarly, the mean scores of DΔ4 and constitutive 5′ss (upstream of pΨ4),$S_{D\Delta4} = 7.9$ and $S_{5'\mathrm{SS}} = 8.7$, were found to be similar, but still significantly different ($P < 0.003$), whereas $S_{p\Psi4} = -10.2$ was significantly lower than $S_{p\Delta4} = -3.9$ ($P < 1.9e-13$). In words, minor splice variants of tandem donors (pΔ4, dΔ4) scored larger than pseudo tandem donors (pΨ4, dΨ4), while lower than 5′ss of constitutive splicing exons, and were consequently sufficiently different from pseudo splice sites, despite the same genomic motif.



(e) PΔ4

29	6	35
38	15	53
67	21	88

(f) dΨ4

4,637	121	4,758
4,789	273	5,062
9,426	394	9,920

(g) DΔ4

96	10	106
108	22	130
204	32	236

(h) pΨ4

50,939	1,256	52,195
51,503	1,915	53,418
102,422	3,171	105,613

Figure 2.7: **(a,b)** Scatter plots of 5′ss scores of non-overlapping and Δ4 tandem donors (*cf.* notation of Figure 2.6). Shown are the individual and mean scores for **(a)** type-I and **(b)** type-II donors respectively, marked by solid/dashed lines for non-overlapping/tandem donors; **(c,d)** Cumulative score distributions. Shown are comparisons of **(c)** PΔ4 and dΔ4 splice sites with constitutive 5′ss and dΨ4 (pseudo distal 5′ss, in black) and **(d)** pΔ4 and DΔ4 splice variants with pΨ4 and 5′ss (pseudo proximal 5′ss, in black). The threshold at which the curves intersect (S^*) marks the accuracy at which sets can be distinguished with equal classification errors on major and minor splice variants. $S^* \approx 78\%$ for PΔ4 versus dΔ4 (PΔ4/dΔ4) and $S^* \approx 92\%$ for pΔ4/DΔ4, and $S^* \approx 95\%$ for dΨ4/5′ss and $S^* \approx 99\%$ for 5′ss/pΨ4. **(e-h)** Tables of exon counts for each splice type above and below S^*; 1st row from left right: TP, FP, TP+FP; 2nd row: TN, FN, TN+FN; 3rd row: TP+TN, FP+FN, P+N (T=True, F=False, P=Positive, N=Negative)

Table 2.4: Summary of the transcript coverage for all possible different motifs of A5EΔ4 splicing exons. The coverage is shown for major-form distal (DΔ4) and proximal (PΔ4) tandem donors. Genes with inferred DΔ4 splicing exons outnumber genes with PΔ4 splicing exons about 2.5-fold, which is reflected in their overall cDNA (about three-fold) and EST (about ten-fold) coverage. In addition to /GT, the /GY motif is shown, if the presence /GC was statistically significant.

Splice site motif	**Distal 5′-splice site (DΔ4)**			**Distal 5′-splice site (DΔ4)**		
	Occurrence	EST	mRNA	Occurrence	EST	mRNA
/GYGA/GT	36	2,555	170	3	16	8
/GYAA/GT	32	2,603	140	4	23	19
/GYAG/GT	27	922	118	19	372	75
/GYAT/GT	7	174	38	1	2	1
/GYGG/GT	6	94	18	11	91	31
/GYAC/GT	2	50	8	2	50	8
/GYCA/GT	2	5	10	-	-	-
/GYGC/GT	2	390	5	1	5	2
/GYGT/GT	2	25	13	-	-	-
/GYTA/GT	2	182	9	-	-	-
/GYTG/GT	1	-	2	3	60	15
/GYCC/GT	-	-	-	-	-	-
/GYCG/GT	-	-	-	-	-	-
/GYCT/GT	-	-	-	-	-	-
/GTTC/GT	-	-	-	-	-	-
/GTTT/GT	-	-	-	-	-	-
	118	**7,000**	**531**	**44**	**619**	**159**

2.3.6 Discriminating A5EΔ4 versus Constitutively Spliced Exons

The difference between the 5′ss score distribution $f(S)$ of major and minor A5EΔ4 splicing exons of tandem donors were used to test, based on the behavior of $f(S)$ alone, how accurate PΔ4 can be distinguished from dΔ4, and DΔ4 from pΔ4 splicing exons. To this end, for type-I the cumulative distribution $F(S^{(n)})$, with $n = 1, 2, ...N$, was computed for the set $\{S_{P\Delta 4}\}$, by *i*) rank-ordering all scores $S^{(n)}$ from the smallest to the largest score, *ii*) calculating $s_N = \sum_{n=1..N} S^{(n)}$, and *iii*) normalizing $F(S^{(n)}) = s_n/s_N$. By construction, $F(S^{(n)})$ is a monotonically increasing function of S and takes on its largest value at $F(S^{(N)}) = 1$. Similarly, $G(S) = 1 - F(S)$ for the set $\{S_{d\Delta 4}\}$ was computed, which is a monotonically decreasing function of S that takes on its largest value at $G(S^{(1)}) = 1$. The intersection of $F(S^*)$ and $G(S^*)$ yields for each set the accuracy at which $\{SP\Delta 4\}$ and $\{Sd\Delta 4\}$ can be distinguished, with smallest probability of error on the classification of both sets (112, 113).

Figure 2.7 C shows for PΔ4/dΔ4 splicing exons the cumulative distributions $F(S)$ and $G(S)$ in the range between -20 and 15 units, together with $F(S)$ and $G(S)$ for constitutive 5′ss and dΨ4 splicing exons for comparison. On the one hand, it was found for PΔ4 and constitutive 5′ss that $F(S)$ collapses to approximately one curve for $S > 0$, and that constitutive 5′ss exhibit a long range of negative scores, which was not seen for tandem donors. $G(S)$ for dΔ4 decays similarly to dΨ4, albeit overall shifted by about ten units toward larger scores, and hence, leads to a greater overlap between the $F(S_{P\Delta 4})$ and $G(S_{d\Delta 4})$ as compared with $F(S_{5'\mathrm{SS}})$ and $G(S_{d\Psi 4})$ for constitutive splicing exons. Consequently, the accuracy $A(S^* = 3.5) > 95\%$ at which one can distinguish constitutive 5′ss from dΨ4 is larger than $A(7.3) = 78\%$ for PΔ4/dΔ4. On the other hand, in Figure 2.7 D for DΔ4 and constitutive 5′ss with pΨ4 similar relationships for $F(S)$ and $G(S)$ were found, with $G(S_{p\Delta 4})$ overall shifted by about five bits toward $G(S_{p\Psi 4})$. Both distributions are wider gapped than observed in Figure 2.7 C, and thus, the accuracy reached $A(6) = 92\%$ for alternative and $A(4.6) = 99\%$ for constitutive splice sites, respectively.

Note that distinguishing the sets above by means of a 5′ss score difference and the log-likelihood difference (LLD), presented in (114), are closely related. This can most easily be seen, by considering splice site scores derived from a standard position specific weight-matrix (PSWM) model with independent nucleotide fre-

quencies: provided the PSWM background model remains unchanged, the slice site score difference is equal to the LLD. For the MAXENT splice site model incorporates higher-order statistical dependencies between nucleotides, this exact relationship is replaced by correlated values.
For this data, the subsets of pΨ4 and dΨ4 splice sites hold an upper limit on the overall number of human tandem donors, where the pseudo splice site remained unobserved or unutilized. Using the threshold scores suggested from discriminating PΔ4 against dΔ4 ($S^* = 7.3$), as well as DΔ4 against pΔ4 ($S^* = 6.0$), one finds that 23 (∼0.5%) of the dΨ4 set and 530 (∼1.0%) exons of the pΨ4 set exceed these thresholds and were putatively classified as unobserved tandem donors.

2.3.7 Nucleotide Conservation around Major and Minor A5EΔ4 Splice Sites

Given existing differences between tandem donors and constitutive splicing exons with either dΨ4 or pΨ4 splice sites, the nucleotide conservation around splice sites was compared and contrasted (*cf.* Table 2.5). To this end, for each splice site position (P_i) the nucleotide frequencies of proximal and distal tandem donors in type-I and II was computed, and their information score I represented by individual sequence logos (95) (see Methods 2.2.4). I is close to zero in the absence of nucleotide conservation with respect to the background, and increases with increasing conservation up to around two bit per sequence position.

Figure 2.8 shows in part A) pictograms for constitutive 5′ss and 3′ss, proximal (PΔ4) and distal (DΔ4) tandem donors, as well as A3EΔ3 splicing exons; in B) the information score difference between PΔ4 and DΔ4 tandem donors to constitutive 5′ss, respectively; and in C) a species comparison of splice site positions of human A5EΔ4 splicing exons that were sequence conserved at positions $P_{-4}P_{-3}$ or P_3P_4 in exon of the orthologous mouse gene. Base frequencies of PΔ4/dΔ4 were compared to constitutive 5′ss/pseudo dΨ4 splice sites, as well as DΔ4/pΔ4 to 5′ss/pΨ4 splice sites, in order to identify differences in the base composition between these classes.

On the one hand, clear statistical differences were found for dΔ4/PΔ4 splicing exons with, e.g., significantly lower levels of C but higher levels of T at P_{-3} ($P < 10^{-4}$, χ^2-test) compared to dΨ4/5′ss splicing exons. Together with P_{-2} and P_{-1}, which show a significant enrichment of G and A ($P < 10^{-4}$, χ^2-test) of

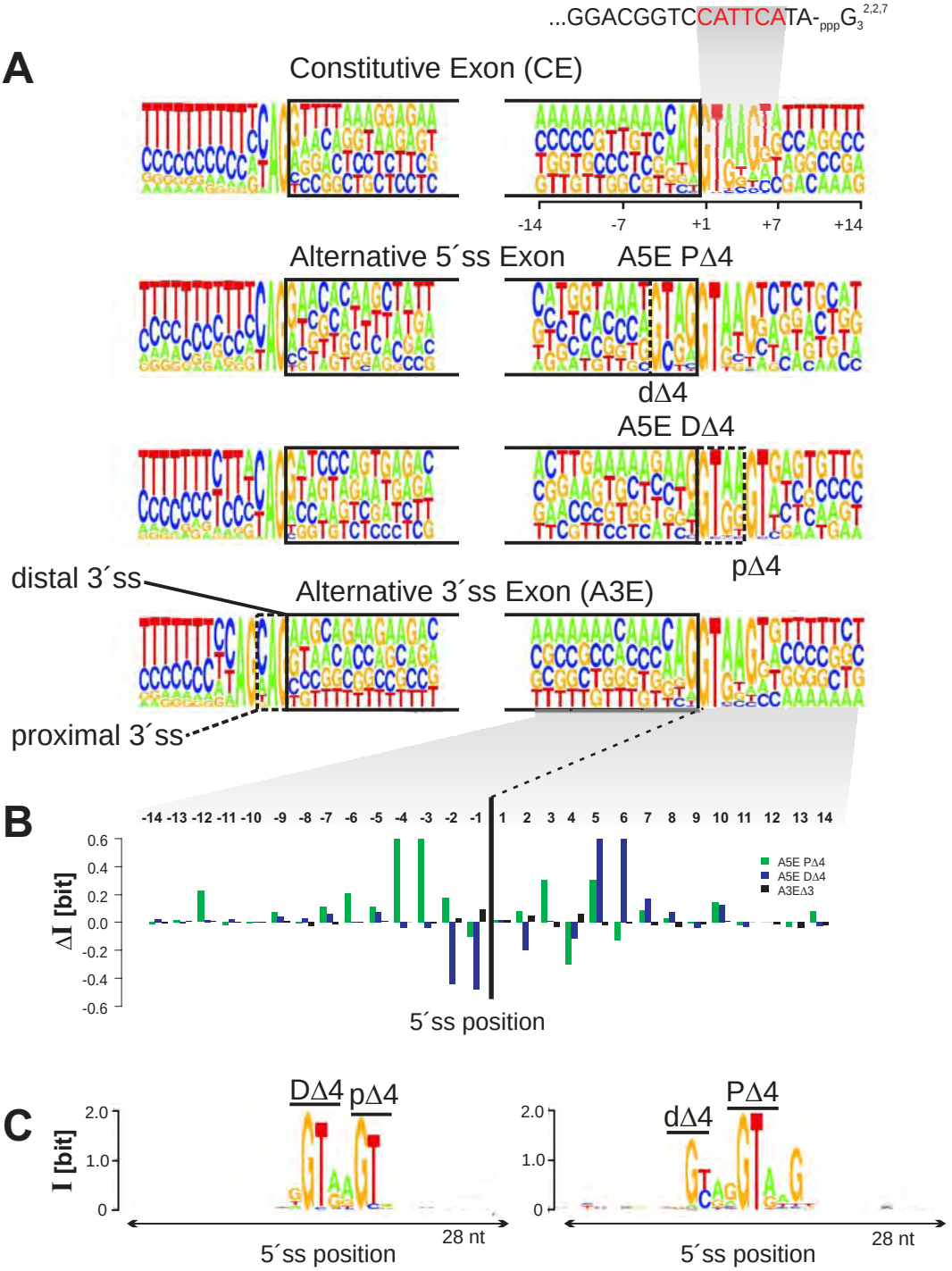


Figure 2.8: Splice site signals and sequence conservation around splice sites. **(A)** Pictograms of 5′ss and 3′ss of constitutive, PΔ4 and DΔ4, and A3EΔ3 splicing exons. The height of a nucleotide represents the frequency of occurrence at a given position, represented in the range of 14 nucleotides around the splice junctions. Above the constitutive 5′ss, the 3′-end of the U1 snRNA is indicated. **(B)** Information score difference (ΔI) between PΔ4 and DΔ4 splicing exons, respectively, and constitutive splice sites, and A3EΔ3 and constitutive exons. For each position, $I > 0 (I < 0)$, indicates more (lack of) information of an alternative compared to a constitutive splice site. **(C)** Sequence conservation of human PΔ4 and DΔ4 splice sites and splice sites of exons of orthologous mouse genes, „anchored“ at major splice sites and with $> 80\%$ exon sequence identity.

Table 2.5: Pseudo tandem donors occurring upstream (dΨ4, distal) or downstream (pΨ4, proximal) of constitutive 5′ss. For constitutive exons, possible dΨ4 motifs are shown in rows 1 and 2, and possible pΨ4 motifs are shown in rows 3 and 4. For alternative A5EΔ4 exons, tandem donors are shown in the last two rows. H=not G, i.e., only A,C,T (R=not C or T, i.e. only A or G) permitted at this motif position.

	Constitutive exons	**PΔ4** proximal major	**PΔ** proximal major	**DΔ4** distal major	**DΔ** distal major
GYNN/**GY**NNYH	4,910 (4%)	-	-	-	-
HNNN/**GY**NNGY GRNN/**GY**NNGY	52,887 (47%)	-	-	-	-
GYNN/**GY**NNGY	5,211 (5%)	-	-	-	-
HNNN/**GY**NNHY GRNN/**GY**NNHY	50,348 (44%)	-	-	-	-
/GYNN/GYNNGY	-	22 (50%)	419 (48%)	26 (22%)	235 (39%)
/GYNN/GYNNHY	-	22 (50%)	453 (52%)	92 (78%)	363 (61%)
	113,356	**44**	**872**	**118**	**598**

dΔ4/PΔ4 over 5′ss/dΨ4 splicing exons, respectively, P_{-2} possibly mismatches to U1 snRNA upon binding to PΔ4, while P_{-3} and P_{-1} possibly support splicing upon binding to dΔ4 due to sequence-complementarity of base pairing with U1 snRNA. Other elevated levels of dΔ4/PΔ4 splicing exons were found for T at P_{-12} ($P < 10^{-4}$), A at P_{-6} ($P < 0.05$), G at both P_{-5} and P_5 ($P < 0.05$), and C or T at P_6 ($P < 10^{-4}$, χ^2-tests). On the other hand, DΔ4/pΔ4 splicing exons showed a significant decrease (increase) of A (T) at P_{-2} ($P < 0.02$) worsening the match with U1 snRNA for both DΔ4 and pΔ4, while an increase of A at P_8 ($P < 0.01$) and T at P_{10} ($P < 0.02$, χ^2-tests) improved the U1 snRNA sequence-complementarity of pΔ4 over pΨ4. In all, several splice site positions were differently depleted or elevated, often with the possibility to enhance the sequence-complementarity to U1 snRNA (115, 116, 117, 118). In particular, G at position P_{-1} has been attributed as crucial for U1 but not U5 snRNA base pairing, creating stacking effects to G at P_1 (119). The conservation of P_{-1} and P_5 observed for A5EΔ4 major-forms, as well as A5Es and CEs but also for dΔ4 splicing exons (type-I), is in agreement with an association of those positions reported by Carmel *et al.* (119). Additionally, P_{-7} and P_{-6} of dΔ4/PΔ4 splicing exons showed elevated levels of A over dΨ4/5′ss and could promote U5 snRNA-dependent base pairing via uridines in the U5 invariant loop, suggested to compensate for weaker U1 snRNA affinity (119) (neither dΨ4/5′ss nor pΔ4 splicing exons showed here elevated levels).

The different conservation levels were in accord with the average information

score that takes into account the frequency of all nucleotides, at a given position, against a background level. Figure 2.8 B shows the difference ΔI between tandem and constitutive 5′ss, which is positive (negative) for higher (lower) scores of tandem against constitutive 5′ss. It was found that dΔ4/PΔ4 splicing exons carried overall more information at P_{-12}, P_{-6}–P_{-2}, and P_{-3}, but as well at P_{-5}, whereas it was found that DΔ4/pΔ4 carried less information at P_{-2} and P_{-1}, but more at P_5 and P_6. Interestingly, Figure 2.8 B shows no marked fluctuations of ΔI between tandem and constitutive 3′ss. Figure 2.8 C supports the above mentioned positional constraints detected for type-I and type-II alternative splicing. It shows the conservation around major (PΔ4, DΔ4) splice sites between human A5EΔ4 splicing exons and mouse exons of orthologous genes, 'anchored' at /GT or /GC splice sites, respectively (the major site, but not the minor site, is conserved by dataset construction). DΔ4/pΔ4 splicing exons only conserved positions P_5 and P_6, whereas dΔ4/PΔ4 showed two recognizable overlapping 5′ss (positions P_{-4}–P_{-2} and P_1–P_6) with nucleotides within the alternative part being complementary to U1 snRNA (119).

2.3.8 Higher Levels of Intron Conservation near Type I Tandem Donors

Exon and flanking sequences of alternative conserved exons, or ACEs, of orthologous human and mouse genes exhibit significant higher levels of sequence conservation. This has most clearly been demonstrated for ACEs that undergo exon-skipping (14, 37, 120), and has also been shown for smaller sets of A5Es and A3Es, including A3EΔ3 tandem acceptors (14, 82). Such conservation could imply the utilization of splicing regulatory signals that are common to orthologous sets of genes.

It was examined whether A5Es and their flanking regions exhibited higher sequence conservation when compared with constitutive exons. To this end, the set of exons arising from tandem (overlapping) and non overlapping alternative splice sites were mapped to exons of orthologous mouse genes. Imposing a level of at least 80% sequence identity and canonical splice sites, matches for about 75% of PΔ4 and 90% of DΔ4 splice variants were obtained. For each species, the sequences of exons and up to 200 nucleotides of their flanking sequences downstream of the donor splice sites were extracted, and the conservation levels for exon and intron regions assessed (see Methods 2.2.6). As control sets, 536/653

A3EΔ3 splicing exons; a randomly selected subset of CEs with 4,145/4,910 and 4,082/4,910 up- (dΨ4) and downstream (pΨ4) pseudo splice sites, respectively; and a randomly selected subset of 2,705/4,910 SEs were mapped. Note that exons of orthologous mouse genes can be constitutive or alternative and, if so, of the same or a different AS type.

Figure 2.9 A shows for PΔ4 test and control sets the exon conservation as a combined score, and the intron conservation in the range between one and 100 nucleotides. Similarly, Figure 2.9 B shows for DΔ4 test and control sets the exon and intron conservation. Test sets have smaller overall sizes than the controls, and therefore possess larger statistical fluctuations. For both exons and introns, the highest level of conservation was observed for the control set of human SEs, which exhibit a clear enrichment over tandem donor A5Es and the remaining controls, in accord with previous analyses (37; 120; 15). On the one hand, for intron flanking regions of PΔ4 splicing exons a markedly higher level of conservation was found as compared with CEs, ranging up to 80 nucleotides (Figure 2.9 A), while for intron flanking regions of DΔ4 splicing exons a conservation level similar to CEs was observed (Figure 2.9 B). On the other hand, Figure 2.9 A and 2.9 B show no marked differences of exon conservation levels between sequences of A5EΔ4 and the control sets (except SEs), and for all investigated exon types the average conservation level was found between 80% and 85%. Previous analyses used datasets enriched by AS events that were specifically conserved between exons of orthologous human and mouse genes (also being smaller sized (14)), and a follow-up study incorporating such data did not distinguish between PΔ4 and DΔ4 splicing exons (121).

2.3.9 Occurrence of Splicing Signals in Exon Flanking Sequences

The above analyses suggested a higher downstream intron conservation of PΔ4 as compared to DΔ4 and constitutive splicing exons, in conjunction with a different splice site score between the major and minor splice variants. It was examined whether the occurrence of splicing-regulatory elements could explain the observed differences (see Methods 2.2.8). To this end, it was searched for over-representations of known oligonucleotides (six to seven-mers) implicated in splicing regulation, which were enriched in A5EΔ4 over constitutive exon flanking regions from one to 100 nucleotides. Four sets of previously computationally

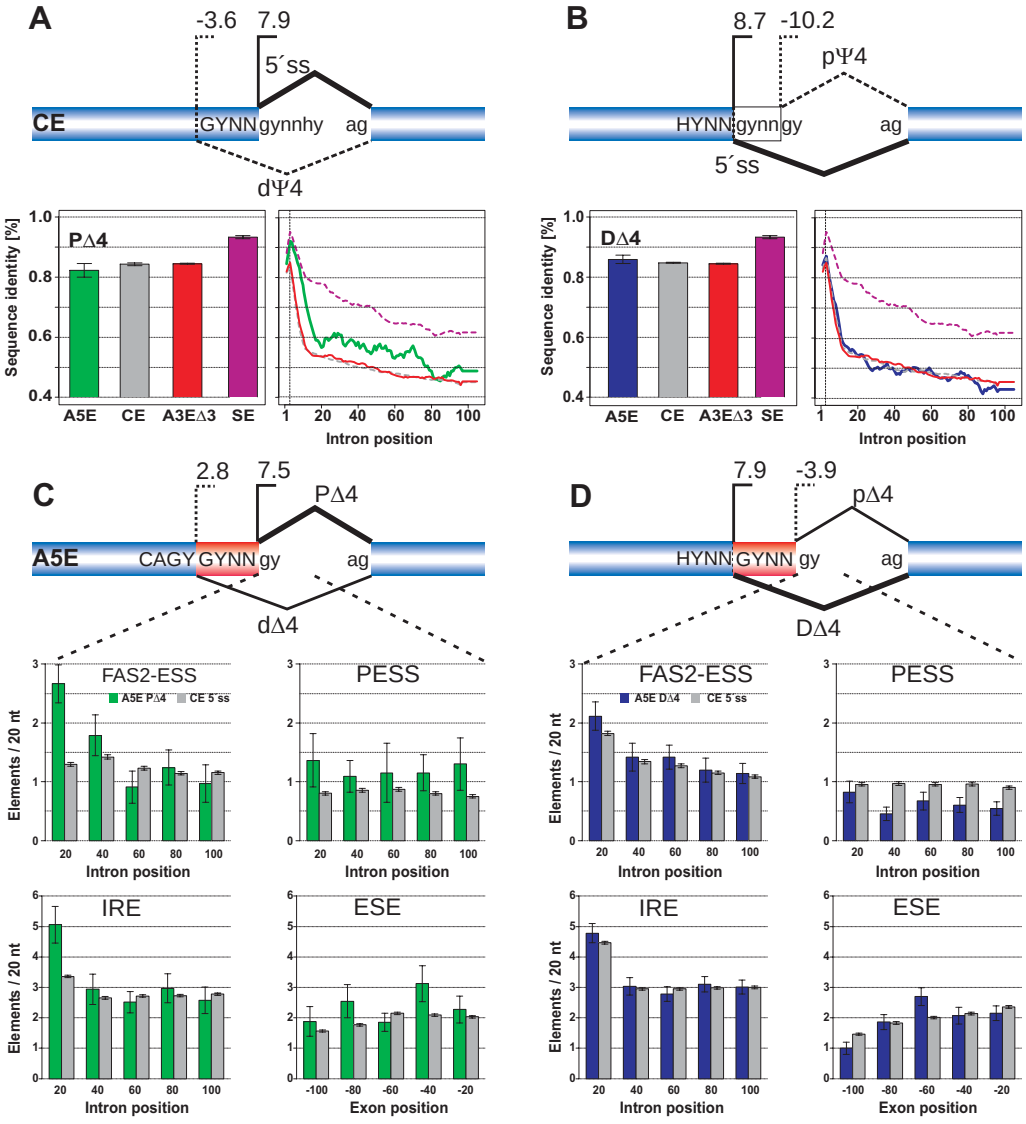


Figure 2.9: Sequence conservation and splicing regulatory elements of A5EΔ4, A3EΔ3, and SEs of orthologous human and mouse genes. **(A)** and **(B)** show for different AS types graphs of the mean exon conservation and of the mean conservation of exon flanking sequences up to 100 nucleotides downstream, respectively. The conservation is shown individually for PΔ4 (panel **A**, green) and DΔ4 (panel **B**, blue) splicing exons; extension regions of A5EΔ4 splicing exons were excluded. **(C)** and **(D)** show plots of occurrences of different splicing regulatory elements, located within the first 100 nucleotides of exon flanking sequences that share $> 80\%$ exon identity and splice site signals with mouse exons.

and/or experimentally identified *cis*-elements were utilized: FAS2-ESS (A) and PESS elements (B), IREs (C), as well as ESE elements (D). Figure 2.9 C compares for PΔ4 splicing exons the frequency of occurrences of all four sets of sequence elements, binned to non-overlapping 20 nucleotide windows and separated for type-I and -II, against the control. Similarly, Figure 2.9 D shows for DΔ4 splicing exons the frequency of occurrences of all four sets of sequence elements. For introns, and both PΔ4 and DΔ4 splicing exons, a generally higher frequency of sequence elements from sets A and C was found, particularly from the start of the splice junction to about 40 nucleotides downstream, while elements of set B are differentially enriched in PΔ4 and suppressed in DΔ4 splicing exons. Sequence elements in exons (set D) were indicative of a general enrichment of ESEs in PΔ4 splicing exons, particularly from about 40 nucleotides upstream to the splice junction, which was not found for DΔ4 splicing exons (except for a peak at about 60 nucleotides upstream the splice junction).
Exon E15 of the gene *SFRS16*, e.g., showed two purine-rich motifs, GGGGGGC and GGTGGG, located at 65 and 87 nucleotides downstream of the 5′ss (contained in sets A and B), respectively. Additional hexamers were located between the positions 117 and 123 nucleotides (GGGAGG), while other sequence elements (set C) occurred often closer to the E15 proximal donor of *SFRS16*, between five and 30 nucleotides. Poly(G)-rich sequence elements are binding sites for the family of hnRNP splicing regulators (122) and have been implicated in the control of 5′ss choice (123, 124, 125). Interestingly, a phylogenetically conserved poly(G)-rich sequence element has previously been reported as involved in the selection of tandem */GTNNNN/GA* splice sites in the splicing of the human *FGFR* gene (126).

2.3.10 A5EΔ4 Splicing Exons often Produce NMD Target Substrates

Inferred AS events of A5EΔ4 and A3EΔ3 splicing exons showed a „splicing dichotomy“ between the 5′ss and 3′ss – while AS events of the latter result in subtle but perhaps biologically significant in-frame variation of a single amino-acid, tandem donors result in out-of-frame shifts downstream of the tandem donor and could thus, lead to a truncated protein with different function or unproductive splicing, depending on position of the exon (Figure 2.10). Indeed, regulated unproductive splicing and translation (RUST) has been proposed to be a mech-

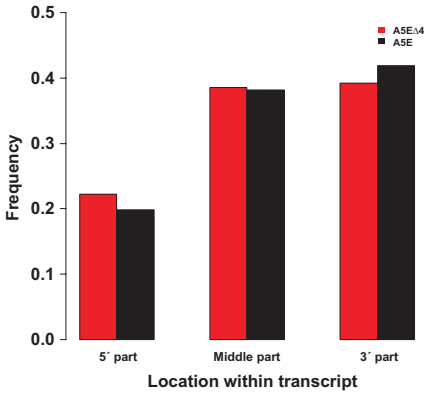


Figure 2.10: Occurrences of A5E and A5EΔ4 exons in REFSEQ sequences. Each sequence was subdivided into three segments of equal size (5′, center, and 3′), and mapped A5E and A5EΔ4 splicing exons were recorded in their respective segments.

anistic link between AS and the NMD quality control pathway (96, 127). What is the proportion of A5EΔ4 splicing exons in the present data set that might be subjected to NMD? To address this question it was *i*) the initially obtained A5E annotation „standardized“ by matching it with REFSEQ-annotated sequences; *ii*) REFSEQ sequences with complete exon-intron structures and annotated start-stop codons of protein coding sequence (CDS) regions identified; and *iii*) proximal and distal splice sites imposed, and the altered reading-frame and stop codon position downstream of A5EΔ4 splicing exons recalculated. Possible compensating AS events were neglected at this step (Figure 2.10).

The detection of in-frame stop codons is schematically sketched in Figure 2.11. In all, 153/171 (∼90%) inferred A5EΔ4 splicing exons were confirmed by at least one REFSEQ sequence, which broke down to 111/153 and 44/153 for distal and proximal donors, respectively. A large majority of A5EΔ4 splicing exons (∼94%) was located in CDS regions, with only marginal proportions in the 5′-untranslated region (5′-UTR) or 3′-UTR. During splicing, choice of the out-of-frame tandem donor will create an mRNA isoform with an in-frame stop codon that introduces a premature termination codon (PTC) and shortens the C-terminus in ∼97% of all considered cases. Tandem splicing of exon E8 of the human *RAD9* gene at the minor distal donor dΔ4, e.g., truncates the *RAD9* domain by 52 amino acids (15% of total length). While possibly still maintaining the domain functionality, the loss of four C-terminal phosphoserines could prevent the interaction with the (9-1-1) cell-cycle checkpoint response complex (128). In all, about three-quarters (78%) of PTCs were located more than 50 nucleotides upstream of the last exon-exon junction, and thus, predicted to produce a marked proportion of

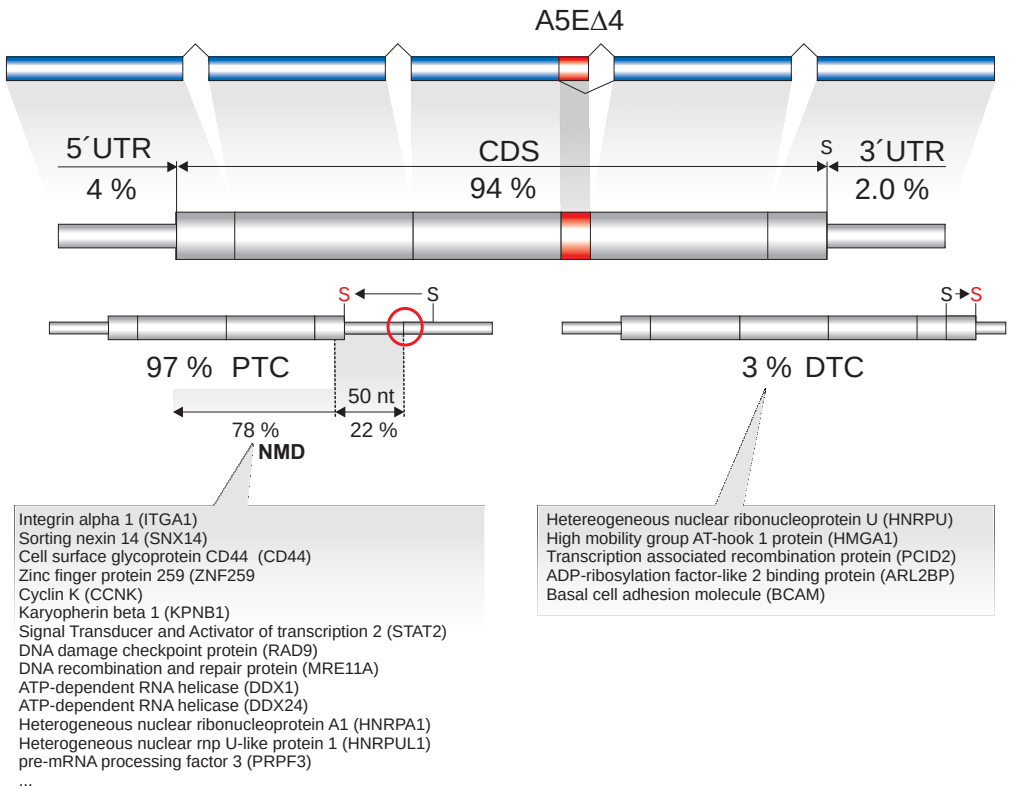


Figure 2.11: Annotation of A5EΔ4 splicing exons in RefSeq genes. Percentages refer to fractions of A5EΔ4 splicing exons located in the 5′-UTR, coding sequence (CDS) region, or 3′-UTR. A black-colored „s“ indicates the position of the stop codon relative to the RefSeq transcript structure, whereas the red-colored version indicates the altered stop codon due to tandem donor splicing. A5EΔ4 splicing exons embedded within CDS regions are broken down into two categories, depending on the creation of a premature (PTC) or delayed termination codon (DTC). PTCs can signal mRNAs as substrates for non-sense mediated decay.

NMD substrates (10). In contexts of type-I and -II, more than twice (∼69%) NMD candidates were produced by DΔ4 splicing exons (where splicing of pΔ4 produced PTCs), as compared with ∼26% PΔ4 splicing exons (where splicing of dΔ4 produced PTCs). The reminder of about 5% of NMD candidates did not stem from type-I or II.

Interestingly, a small number of A5EΔ4 splicing exons (3%) was going to avoid the truncation of the transcript due to the out-of-frame shift but instead extended the message. In close relation to premature termination codons (PTCs), these were termed „delayed“ termination codons (DTCs), and all detected DTCs were produced from utilization of the minor donor (pΔ4). For instance, tandem splicing at the pΔ4 donor of exon E13 of the *HNRPU* gene (*ENSG00000153187*), which encodes the heterogeneous nuclear ribonucleoprotein (hnRNP) U, extended

the CDS regions by 27 amino acids. Due to the frame shift and the occurrence of synonymous and non-synonymous codons, the amino-acid sequence is changed such that the complexity at the protein level (determined by the software SMART (129)) increases at the C-terminal end.

2.4 Discussion

This comparative study distinguishes tandem 5′ss and 3′ss, with three to six nucleotides long extensions, as having unusually high proportions of alternative splicing. The study further analyzed differences and similarities between sets of A5Es, A3Es, and CEs, and focused on a particular type of a pair of alternative donors (A5EΔ4) that are tandemly arrayed and overlapping. These tandem donors were experimentally validated in a panel of different human tissues, and a dichotomy in the splicing frequency at these alternative tandem splice sites is highlighted. The results indicate that human alternative exons spliced at overlapping 5′ss possess features of typical splice variants and could well be beneficial for the cell. Alternative splicing is essential for protein diversification and has recently been suggested as mechanistically linked to post-transcriptional gene regulation via nonsense mediated mRNA decay (NMD) (130). The consequences for protein sequence and function alteration, as well as triggering of the NMD pathway, have been demonstrated for exon-skipping events in several studies (131, 132, 133). While there is further evidence for the functioning and regulation of the remaining types of alternative exons (121), present understanding of their sequence evolution, produced AS patterns, regulation, and functioning still remains vague (134).

Alternative 5′ss exons (A5Es) were computationally inferred from a collection of stringently aligned cDNA and EST sequences to the human genome, and their sequence features were compared to known features involved in RNA splicing. Spliced-alignments were obtained from the three independent algorithms (Sim4, Blat, and Exalin). Exalin detected the smallest number of subtle AS patterns, which are characteristic of tandem donors (involving just a few nucleotides long extensions), most of which were also identified by Sim4 and Blat. For there is no unique „true“ method of inferring AS events, all analyses were based on the subset defined by the intersection of the predictions of all three algorithms. While one cannot rule out misalignments still arising from three methods in some instances, such rigor was taken to produce a confidence-enriched set. In addition, other independent lines of evidence were pursued as for example the experimentally validatation of a subset of 14 human genes with tandem donors across different tissues. The outcome confirmed about 50% A5EΔ4 splicing exons and provided evidence that a substantial fraction of tandem donors detectable in public sequence repositories are not explained by sequence alignment ambi-

guities. Almost one tenth of all human A5Es with exactly one shorter and one longer splice variant, and no other inferred splice type (SE, A3E, or RI), were found as A5EΔ4 splicing exons. Interestingly, Figure 2.2 also shows a small but persistent pattern of higher frequencies at $E = 6, 9, 12, 15$ and 18 nucleotides, which is indicative that non-overlapping splice sites had biased extensions that preserve the reading-frame.

The central outcome of this study points to a splicing dichotomy between human alternative 5′ss and 3′ss exons in that exon extensions were markedly biased toward overlapping splice sites, with A5Es biased for $E = 4$ nucleotides (tandem donors, A5EΔ4), in contrast to A3Es biased for $E = 3$ nucleotides (tandem acceptors, A3EΔ3). Both, A3E and A5E biases in exon length variation have been previously reported (2, 114, 135), but their pertinent features have largely remained hidden. It is important to note that AS at both the 5′ss and 3′ss gives rise to splicing variations with very subtle changes to the encoded protein sequence, but further downstream A5EΔ4 and A3EΔ3 splicing exons lead to very different consequences. While A3EΔ3 splicing exons of the form of NAG/NAG/ have been analyzed in some detail, in part with several controversial interpretations (2, 114), A5EΔ4 splicing exons had not previously been confirmed experimentally and only initially been characterized (135).

In this context, two related questions are whether *i)* such frequently observed subtle changes simply arise by „noise“ of the spliceosome action, and *ii)* the splicing cell has found a way to benefit or neutralize the downstream consequences that arise from such AS events . Provided their biological authenticity, what is the nature of overlapping splice site choice? Several models for splice site choice have been proposed, including *i*) competition between antagonistic splicing factors (for example ASF/SF2 and hnRNP A1) and U1-snRNP (86, 136, 87); *ii*) a scanning mechanism (90); or *iii*) *cis*-acting motifs with different free-energy for binding U1-snRNP and splice factors between competing splice sites (89). These models take into account the binding property of the U1-snRNP and additional factors. Consequently, known features involved in splice site choice were investigated, as well as consequences to the post-transcriptional regulation of A5EΔ4-carrying genes. A5EΔ4 splicing exons were also compared with A3EΔ3 and constitutive splicing exons in the light of existing models for 5′ss selection.

Examined features showed pertinent differences that individually came out subtle, yet taken in concert they were indicative of a spliceosomal distinction of overlapping 5′ss. The data supports that overlapping donors, but not acceptors, can be

distinguished into major-form (DΔ4, PΔ4) and minor-form (dΔ4, pΔ4) splicing exons for both distal (DΔ4, pΔ4 = type I) and proximal (dΔ4, PΔ4 = type II) splice sites, respectively. This is further corroborated by their splice site scores, which within the tandem splice sites correlated with their respective utilization as major or minor splice variant. On the one hand, splice sites deviated most from the consensus for PΔ4 splicing exons at positions P_{-4}, P_{-3}, and P_3 (information score difference $\Delta I > 0$) as well as P_4, P_5 ($\Delta I < 0$), thus, overlapping positions of the U1-snRNA, which are involved in 5′ss selection (89, 123). Additionally, some of these positions have also been related to codon preference (135). Interestingly, P_{-12} deviated too from the consensus. Because of its close proximity to the edge of the U1 snRNA stem-loop it could possibly contribute to U1 binding when donor dΔ4 is spliced. On the other hand, DΔ4 splicing exons deviated differently from the consensus at the positions P_{-2}, P_{-1}, P_2 ($\Delta I < 0$) as well as P_5, P_6 ($\Delta I < 0$). Based on previous experiments on position-specific stabilizing and advancing spliceosomal interactions with the 5′ss, these differences between types I and II are indicative that PΔ4 improve compatibility with U1-snRNA above DΔ4 splicing exons.

Previous computational studies showed that the conservation of sequences flanking ACEs is higher compared to sequences around species-specific or constitutively spliced exons (120, 137). Here, higher levels of conservation around PΔ4, but similar levels for DΔ4 splicing exons were observed, when compared with constitutive exons or the 5′ss of A3EΔ3 splicing exons. Interestingly, the higher level is in accord with a larger number of detected silencing splicing-regulatory elements, often positioned in proximity to A5E tandem donors. In contrast to typical AS events, however, tandem donors are hindered to place regulatory elements between alternative donors. The data show an elevation of ESE elements near dΔ4, in conjunction with an enrichment of ESS elements of flanking introns. This could be interpreted in a model, in which tandem donors restrictively exploit elements in proximal location, i.e., near dΔ4, to attract the U1-snRNP to this site of the tandem donor, or in distal location to dΔ4, to impair U1-snRNP binding to PΔ4 (87).

For the majority of tandem donors was embedded in coding regions, the downstream effects of Δ4 splicing was predictive of producing PTCs. Splicing at pΔ4 produced putative NMD substrates in more than two-thirds of all cases, whereas dΔ4 splicing exons showed about one-quarter, suggesting that pΔ4 and dΔ4 (the minor-forms) were more likely to serve as the corresponding NMD candidates. In-

terestingly, a small set of A5EΔ4-carrying genes avoided PTCs, yet instead was inferred to use DTCs (delayed termination codons) positioned downstream of the original signal. Utilization of the E15 proximal tandem donor of the human *SFRS16* gene, e.g., with significantly high levels of E15 flanking sequence conservation well over 120 nucleotides in I16 (typical of RNA splicing conservation across species (120)), produced a PTC that apparently avoided NMD (138). Using differentially binding antibodies, a previous study (108) showed that *SFRS16* produced two detectable isoforms, which correspond to E15 tandem splicing.
A survey of gene ontology (GO) functions of the categories „*molecular function*“ and „*biological process*“ for genes with PΔ4 and DΔ4 splicing exons showed a significant enrichment in several proteins, while after corrections for multiple testing only the single GO-term „*RNA binding*“ ($P < 0.005$, nonparametric t-test) was significantly enriched, when compared between PΔ4/dΨ4 and DΔ4/pΨ4 splicing exons, respectively (see Methods 2.2.9).
This study substantially affirms the utilization of tandem donors. While in principle complementing earlier findings of previously undetected tandem donor AS events (121, 135), different approaches to the characterization of A5EΔ4 events impede to some extend the comparison of results. For example, Dou *et al.* 2006 find type-I tandem donor splice events more frequently than type-II, which stands in contrast to results of this study. However, they do not provide the exact rules based on which the „dominant“ (major) spliced site in a tandem is defined. This makes their class boundaries not reproducable and hence different class occupation between the studies could be a consequence of different methods in the type-I and II tandem donor classification. Furthermore, examples of cryptic Δ4-type 5′ss have been reported in the literature (111, 139). In contrast, here it is demonstrated that such splice variations are potentially enriched in authentic AS events, and also supported by experimental studies (108, 140). Critically, pertinent data are not yet at hand to make conclusive inference about the specific regulation of A5EΔ4 splicing exons (e.g. controlled expression of species-specific minor/major isoforms), here transcript data acquisition and careful spliced-alignments have added to a higher confidence of tandem donor (and acceptor) utilization, and deeper insight will require different types of data, e.g., from mini-genes in different organ systems and cell types, U1 snRNP mutants, or variations of splicing factor dosages.
In one extreme view, incorporating a mechanistic and dosage-dependent model (89, 87), the selection of AS sites depends on the properties of U1 and/or U6

snRNPs binding interrelated with antagonistic effects mediated by splicing enhancing and suppressing factors. Although, type-II tandem donors show a reduced difference between distal and proximal splice sites, an initially equal chance of U1 snRNP binding to either the proximal or distal donor, could later be compensated by the higher complementarity of the minor distal donor to U6 snRNA positions, complemented by silencing elements upstream of the major proximal donor. Similar it was shown that the choice of a tandem splice site of E10 of the *FGFR* gene can be determined by a higher sequence-compatibility of the E10 proximal splice site (pΔ6) to U6 snRNA (126). In addition, constraints set by secondary mRNA structures (75, 141) have been shown to influence splice site choice. In the opposite extreme, suggested by subsets of tandem donors with strong difference between distal and proximal splice site, splicing at such tandem donors could regularly involve stochastic binding of the splicing machinery without an implicit functional relevance (114), which seems to be supported by type-I isoforms. Either view largely requires the NMD pathway to control deliberatively or aberrantly produced truncated messages.

Earlier works investigating spliceosomal components in 5′ss recognition, suggested co-recognition of 5′ ss by U1 snRNP and tri-snRNP as a mechanism to explain activation of cryptic 5′ ss and U1 independent splicing. Cryptic sites were proposed to substitute for mutationally silenced authentic splice sites, still binding low levels of U1 binding. Those bound U1 snRNPs in turn could recruit tri-snRNPs, which themselves can not bind authentic splice sites but activate nearby sequences via Prp8 or other U5 snRNP components (142). It is interesting to note, that U1-dependent and U1-independent splicing share the same 5′ ss sequence requirements, leaving room for the possibility that especially Δ4 AS events of similar splice site plasticity could be differentially regulated by differently composed spliceosomal complexes (143, 144).

An interesting question regarding the finding of deliberately or aberrantly produced AS variants with out-of-frame shifts and PTCs (either due to A5EΔ4 or other types of AS), and their functional utilization on the transcriptional or translational level, is whether there is a possible benefit of generating flawed mRNA isoforms. If such splice variants would be generally produced across organ systems and cell types, in addition to their normal splice variants, cells would have means of producing low levels of imperfect proteins. Depending on the efficiency of mRNA quality control, a fraction of which is subjected to the NMD pathway during the first pioneer round of translation and degraded, while a remaining

fraction could still misfold and - depending on the quality control of protein synthesis - form defective ribosomal products (DRiPs). Ubiquitin-tagged peptide fragments that originate from DRiPs have recently been identified as a potent source of antigens for display by the MHC class I molecules on the cell surface to cognate $CD8^+$ T-cells, in agreement with a recently suggested mechanism of „immune surveillance“ (145, 146, 147). A motivating example is given by the human Tyrosinase-related protein 1 (TYRP1), which utilizes two different reading-frames to produce the protein gp75 (recognized by IgG) and a truncated 24 amino-acids long peptide. The latter was shown to be the source of an antigenic peptide specifically recognized by T-cells as a tumor rejection antigen (148). It remains to be substantiated whether such antigenic peptides are linked to AS events that produce variants with out-of-frame shifts, such as produced by tandem donors.

Chapter 3

Petrinet Analysis of Spliceosome Assembly

3.1 Introduction

3.1.1 The Spliceosome

Alternative splicing depends on the fundamental processes of the spliceosomal protein complex. The spliceosome is a nuclear megadalton complex which is organized in five major subcomplexes U1, U2, U4, U5 and U6 that are formed by core proteins and additional splicing factors. The presently known number of splicesosomal factors is estimated to 150-300 proteins (149, 150, 151, 152, 153), thus providing the cell with a broad repertoire of splicing regulatory proteins. The spliceosomal subunits are ribonucleoprotein complexes of approximately 1-2 μm^3 which form a structured environment around target splice sites. Only recently it was confirmed that the snRNAs of the U2 and U6 subunits are sufficient for RNA cleavage and re-ligation, rendering the spliceosome effectively a ribozyme (154). Though this suggests that the splicing reaction is a relic from the RNA world, a complex protein machine is wrapped around this mechanism with many proteins being homologous between such different organisms as human and yeast. The outcome of different splicing events shows that the maturation of RNA is a complicated process that can be influenced by different events such as mutations in splicing signals or signal cascades, induced by stimuli outside of the cell (155, 11). Splicing decisions are controlled by two major determinants - **i**) the pre-mRNA sequence and its inherent signals and **ii**) the protein-complement of the spliceosome where signal transduction is frequently maintained via arginine-serine rich

domains of the participating proteins (RS domains) (20). Hence, the dependence of gene expression on developmental stage and/or tissue type is modulated to a major extent by a network of protein-protein and protein-RNA interactions that influence the assembly and localization of active spliceosomes. Noteworthy, the spliceosome is thought to serve not only as the catalyst of the splicing reaction steps, but also to translate the limited information presented by pre-mRNA into splicing specificity and orderliness (156).

3.1.2 The Basic Steps of Spliceosome Assembly

Several hypotheses about the dynamics of spliceosome assembly exist: one model suggests the co-transcriptional assembly of spliceosomes at the nascent pre-mRNA, consistent with the fact that in eukaryotes splicing factors have been found in association with the c-terminal end of the RNA polymerase II (157, 158, 159). In opposite, alternative splicing due to exon skipping can require the presence of an at least partially transcribed pre-mRNA, as the remarkable example of the heteronuclear ribonucleoprotein (hnRNP) A1 shows (79). There, hnRNP A1 binds at intronic sites adjacent to an exon of its own pre-mRNA, and by interacting hnRNP molecules across the exon, the exon is looped out from the splicing plane, thus implying the occurrence of a post-transcriptional splicing mechanism. Additionally, the commonly accepted model of exon definition in higher eukaryotes (85), favors a post-transcriptional splicing mechanism. This is reasonable since the $5'$ splice site ($5'$ss), branchpoint (BP) and $3'$ splice site ($3'$ss) have to be present and recognized by an initial protein complement for subsequent bridging across the exon. Consistently in well studied yeast splicing systems it has been found that due to the transcription kinetics, spliceosomes preferentially assemble post-transcriptionally on short second-exon genes, affecting approximately more than 90% of yeast splicing (160). However, the same study and others (161) observed that U1 snRNP recruitment and splicing of long second exon genes (>1 kilobase) occurs mainly co-transcriptionally.

Opposing the stepwise assembly model, a *„holospliceosome“* model has been proposed, where a tetra- or penta-snRNP complex binds to the introns and modulates RNP-RNP or RNP-pre-mRNA contacts within the complex (162, 163, 164). Spliceosome assembly can be described as intracellular allosteric cascade, where every interaction depends on a previous interaction. Besides the snRNP sub-

complexes and their core components, additional proteins (splicing factors) and enzymes (DExD/H-box proteins) are required to shape the active spliceosome. While the first step - the E-complex assembly - requires no energy for structural rearrangements, subsequent steps of snRNP addition and interaction involve helicase like ATPases that catalyze structural rearrangements in snRNA-pre-mRNA contacts (163). The major events in the stepwise assembly prior to the activated spliceosome include U1 snRNP binding to the 5′ ss, U2 snRNP binding to the branchpoint, addition of a preformed U4/U6·U5 triple snRNP, interaction of U2- and U6 snRNA and concomitant destabilization U1 and U4 snRNP (165). At this point, the spliceosome is reshaped such that the first transesterification reaction can take place. The snRNP and splicing factors arrange the 2′-OH of the branchpoint adenosine in such a close proximity to the 5′ss that it can perform a nucleophilic attack to the 5′ss phosphodiester bond, disrupting the pre-mRNA and forming an intermediate loop (lariate) within the intron. Further, internal rearrangements of snRNA contacts and interaction of snRNP components prepare the second step of catalysis whereupon the 3′-OH end of the upstream exon attacks the 3′ ss phosphodiester bond, finally tethering both exons together and releasing the lariat intron (149). These basic steps are conserved from yeast to mammals and also similar for the minor U12 type spliceosome assembly (166; 167).

Description of the Four Major Stages of Spliceosomal Assembly

Splicecosome assembly is a hierarchical process that progresses through several main stages designated (P→) E → A → B → C (→ P). Each stage (E-C) describes an intermediate complex that is build from the pool (P) of proteins available in the nucleus and which are recycled for repeated rounds of spliceosome assembly. The biochemical processes, which result in the mature spliceosome that is able to catalyze the splicing reaction in eukaryotic messenger RNAs can be summarized as following:

I. E-complex (commitment complex):
The recognition of the 5′ splice site is initiated by early interaction of the U1 specific protein U1C with the 5′ss sequence (168). Additionally, interactions via phosphorylated RS domains of U1-70K with splicing enhancing factors, which bind to nearby located specific enhancer motifs, can direct the U1 snRNP to a specific 5′ss (169). It was shown that 5′ss, which are less complementary to the canonical eukaryotic 5′ss, are still selected

by U1 snRNP due to interactions between U1 snRNP proteins and splice enhancing proteins, such as ASF/SF2 or TIA1 (169, 170). This is in agreement with the earlier observation that donor site selection by U1 snRNP can be rescued by the SR protein, SC35, when the donor recognition site within U1 snRNA is impaired (171). In contrast, a weak 5′ ss might also be selected by variants of U1 snRNA (172). This led to the interesting hypothesis that 5′ss selection within the E-complex might be possible even without the binding of U1 snRNP. Experiments indicate that the presence of higher concentrations of SR proteins enable such a spliceosomal pathway (173). At the 3′ end of the intron, E-complex assembly involves the U2 auxilary factor, U2AF, which is a heterodimer of a 65 and 35 kDa subunit. Together with SF1 (in mammals designated branch point binding protein, mBBP), these factors recognize via their RNA recognition motifs (RRM) the polypyrimidine tract, 3′ss and branchpoint, respectively. This step was shown to occur in a coordinated way involving cooperativity between SF1 and U2AF65 (174), which can be mediated by interactions between the RS domains of these proteins (20). Additional interactions reported during E-complex assembly include the bridging of U1 snRNP at the 5′ss and factors bound to the branchpoint-3′ss, which is mediated by SR protein FBP11 (175). U2 snRNP is present in close proximity to the U1 snRNP and formation of the E-complex depends on U2 snRNP (176, 177).

II. A-complex (pre-spliceosome):

The A-complex contains the stably integrated U2 snRNP, which is assembled out of two heteromeric subcomplexes SF3a and SF3b via an intermediate 17S complex. Prior to their contacts to the U2 snRNA, SF3a and SF3b are formed by several interactions between U2 core components (e.g., SF3b155 with SF3b14 or SF3a120 with SF3a60) (178). In this stage, U2 snRNP interacts with the branchpoint site via base pairing between U2 snRNA and the pre-mRNA under consumption of ATP (177). Additionally, the transition from E to A-complex is supported by the SF3b protein, SF3b155, which binds at both sides of the branchpoint and interacts simultaneously with the U2AF65 subunit and the SF3b14 protein (179, 180). The ATP requirement during A-complex formation has been reasoned by two other U2 snRNP proteins, SF3b125 and hPrp5, which are members of the DExD/H family and may function either as helicases or RNPases (181). Such auxiliary enzymes could be recruited from proximal cajal bodies or

speckles (182).

These helicase like proteins have been shown to function as generic ATPases that bind and hydrolyze NTP to unwind double-stranded RNAs (dsRNAs) (165). In contrast, the ADP-bound form can modulate the annealing of complementary RNA strands. In context of the spliceosome DEAxD/H box helicases are governing structural rearrangements to shape the active complex. However, many of these ATPases function in a generic way hydrolyzing also other RNA species. Thus, they have to be specifically activated for catalyzing the correct structural rearrangements, not least to ensure fidelity of the splicing reaction. It was proposed that additional splicing factors interact with DExD/H box helicases to direct their activity to specific substrates during the assembly process.

SF3b125 was detected only in low amounts in the SF3b subcomplex, but not associated with the 17S U2 snRNP and hPrp5 was present in the 17S U2 snRNP, but not in the SF3b subcomplex (182). The protein hPrp5 exhibits an ATP independent function, which stabilizes U2 snRNP to the BP (183) and was proposed to function as bridge between U1 and U2 snRNPs at the time of A-complex formation (181). Accompanying the U2 snRNP rearrangements, which are catalyzed by hPrp5 under ATP hydrolization, the SF3a60 contacts to the U2 snRNA are significantly reduced upon association of U2 with the BP in the A-complex (178). Another DExD/H box helicase, which is required for U2 snRNP / BP interaction, is UAP56, whose recruitment also depends on U2AF65 (184). The reactions of the individual assembly stages are described and formalized in Table C.1.

III. B-complex (active spliceosome):
In this stage, U4, U5, and U6 snRNPs enter the assembly pathway as tri-snRNP complex. This subcomplex is formed in a separate way, where U4 and U5 snRNP assemble, similar as U1 snRNP, from a family of seven RNA binding proteins, termed Sm proteins (185, 186), whereas the U6 snRNA is bound by a different set of proteins termed Sm-like proteins (187). Subsequently, U4 and U6 snRNP form a duplex via base pairing of their snRNAs, resulting in a structural conformation, which is stabilized by several additional proteins, for example hSnu13, Prp3, Prp4, CypH and Prp31 (186). The U5 snRNP contains several proteins, important for structural rearrangements prior to the first catalytic step of splicing, most important the

DExD/H box helicases Brr2 and Prp28 and the two proteins Snu114 and Prp8 (188, 186). An interaction between Prp6 and Prp31 was proposed to serve as bridging step between U4/U6 and U5 snRNP, preparing the formation of the U4/U6.U5 tri-snRNP complex (189). The additional proteins, Snu66, Sad1 and 27K, stabilize the intermediate tri-snRNP complex. The latter is recruited to the spliceosome which, however, is still catalytically inactive. An intermediate state, designated pre-catalytic B′ complex, shows a more flexible tri-snRNP structure, possibly for integrating other components, for example Prp19 (190).

Prior to the conformational rearrangement required for spliceosome activation, U1 snRNP is dissociating from the 5′ss enabling U6 snRNP to contact donor splice site. This step involves the ATPase, Prp28, which was found to counteract the stabilizing effect of the U1 component U1C with the U1 snRNA (191). Prp5 can leave the spliceosome at this stage as it was demonstrated to function mainly before or during A-complex assembly (182). The U5 snRNP components Brr2, Prp8 and Snu114 are critical for unwinding the U4/U6 snRNA stemloop. This step resembles a G-protein activating mechanism, where Snu114 enters a GTP dependent state, and subsequently activates the helicase Brr2 (192, 193). This conformational change also involves Prp19, a part of the Ninteenth complex (NTC) (194). Brr2 is further involved in interactions with Prp16 and U1-70K (195, 186). After unwinding the U4/U6 duplex, U2 and U6 snRNP establish several interactions via their snRNAs and the pre-mRNA, whereas a binding motif within the U6 snRNA directly contacts the 5′ss. Subsequent release of U1 and U4 snRNP results in the catalytically competent B* complex which forms the catalytic core of the spliceosome (165, 186). The B-complex, composed of U4/U6.U5 tris-nRNP in close contact with the U2 snRNP, performs the first catalytic step of splicing by nucleophilic attack of the branchpoint adenosine to the phosphate ester bond of the 5′ ss, and is subsequently converted into C-complex.

IV. C-complex:

The C-complex contains U2, U5 and U6 snRNA at a stage subsequent to the first chemical step, because splicing intermediates can be found in this complex. The conformation is centered around Prp8 which is thought to serve as a „surgery table“, connecting the already free 5′ and fixating the 3′

ss such that the second transesterification can open the intron-downstream exon connection (192). With the ligation of the free exon end, the intron and bound snRNPs are released as a post-spliceosomal lariat complex. Subsequently, bonds between U2, U6 and U5 snRNA are broken, involving the helicase, Prp43, and U5 dissociates into smaller subcomplexes to finally join another round of spliceosome assembly (151). Since Prp43 can be found in the 17S U2 complex (182), which forms during early A-complex assembly, it is conceivable that this protein remains present over several stages of the spliceosome assembly pathway.

Additional factors support the recycling process, for example Prp24, which reanneals U4 and U6 snRNAs and allows regeneration of the U4/U6 snRNP duplex (196). Two important helicases, Prp16 and Prp22, impose kinetic proofreading activity and can subject suboptimal splicing substrates into a discard pathway (197, 198). It is important to note, that the catalyzing function of the spliceosome can experimentally be reduced to its RNA parts, making it effectively a ribozyme (154). However, the protein scaffold is necessary to form the structural environment (RNA conformations) necessary to enable the splicing reaction. Moreover the participating proteins establish important links to other cellular processes as for example transcription or nuclear export.

3.1.3 Modeling the Spliceosome

One of the major difficulties in the functional characterization of the spliceosome arises from the dynamical interactions between its subcomplexes and the proteins organized within those. Although extensive knowledge has been gathered about the factors involved in spliceosomal function, their functional interplay and regulatory impact is not comprehensively understood and even discussed controversially. For example, it is still not fully understood whether the assembly process occurs mainly co- or posttranscriptional and whether a stepwise assembly (158, 164) rules out the possibility of a pre-assembled holospliceosome (162, 199). in this context the term *„transcription factory model“* has been coined denoting the concerted action of transcription and splicing machinery at the places of active gene transcription (200). The connection and coupling of different protein machines in the cell emphasizes the complex environment in which spliceosomal

assembly is embedded.

The vast amount of experimental data makes it necessary to structure the wealth of information to derive new hypotheses on the underlying signal transduction processes. This task requires the development of *in silico* models which are able to integrate much of the existing data while being suited for rigorous validation and stepwise extension. Further modeling goals can be summarized in *i*) visualization, *ii*) comprehensible data annotation and *iii*) data abstraction, allowing an exact mathematical description of the model.

In particular, alternative splicing often involves different sets of splicing factors in addition to the spliceosomal core components, hence variations of the model network should provide a sound base for testing hypotheses on the regulation of splicing and AS events. Several structural and kinetic factors were proposed to influence splicing patterns of a gene such as *i*) precise balance and concentrations of regulatory proteins , *ii*) nature of interaction like inhibiting or cooperative effects, *iii*) number of interactions which define the connectivity of a network, *iv*) speed of transcriptional elongation or recruitment of splicing factors (201, 202). Additionally, a temporal model component can be considered as the assembly pathway bears a number of timed dependencies.

These keypoints pose an essential base for modeling spliceosomal processes, but, most of them can be not comprehensively applied at present. Incompleteness of experimental data as for example the lack of knowledge about interaction kinetics, concentrations or the exact temporal order of reactions is accompanied by heterogeneity of proposed mechanisms for different stages of spliceosome assembly.

Protein interaction databases such as String (203), cross-referencing to Mint, Hprd, BioGrid, Dip and Reactome, Apid (204) or IntAct(205), already provide an extensive organization and integration of experimental and predicted protein interaction data, but, most of them represent static interactions without providing information about the temporal order and direction of interactions within hierarchical networks. This underscores the need to find ways to access this additional information from literature, incorporating knowledge that can be used to extend the capabilities of network exploration for hierarchical processes, such as spliceosome assembly. This should extensively support the

generation of hypotheses on the progression of signals and protein interactions. For comparison, Figure 3.1 depicts an unordered-undirected protein-protein interaction network as reviewed in (206), and a directed ordered interaction network presented in this work. Due to lack of kinetic properties, composing

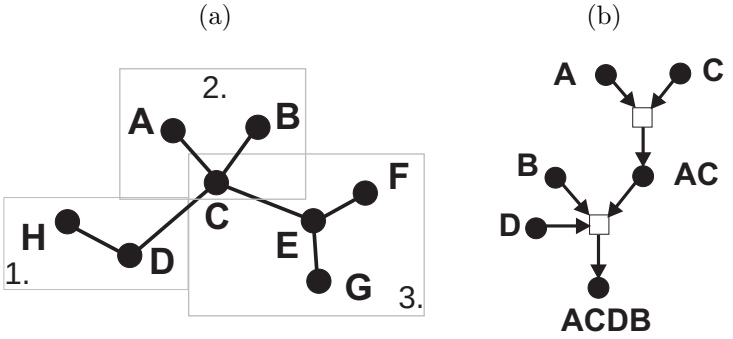


Figure 3.1: Two types of protein interaction networks: **(a)** An undirected network of binary interactions, e.g., modeled in (207) **(b)** Directed hierarchical molecular interaction network as modeled here by Petri net (PN) theory

a discrete structural model was considered as a good initial choice for the computational analysis of the spliceosome. Here, Petri net (PN) theory was chosen, because it offers the advantage to combine visualization, with the possibility of mathematical description and graph-theoretical analyses.

PN theory provides techniques and software tools to model, analyze, and simulate biochemical networks (208, 209, 210, 211). Meanwhile, metabolic networks (212, 213, 214, 210) as well as signal transduction pathways (215, 216) and gene regulatory pathways (217, 218, 219) were successfully modeled using PN. There are parallel approaches to structural modeling such as those based on elementary flux modes of (220, 221) and on extreme pathways (222). They have primarily been applied to metabolic pathways (for reviews, see (223, 224)) but also to signaling networks (225). An interesting question is, whether the model of stoichiometrically quantifiable mass flow inherent to metabolic networks, can be transferred to models of information flow as occurring in signaling networks. Recently, a model of the U1 snRNP subcomplex assembly demonstrated the applicability of PN theory (226). The present work is the first attempt to merge current knowledge about the entire spliceosomal assembly into a network of directed reactions. In doing so, most challenging is the proper translation of a huge body of experimental evidence from the literature into single reactions steps, which can be integrated in a PN. Sorting through these reactions served as a base for gradual refinement, converging in a validated model of spliceosome

maturation.

Box 3.1.1 Allosteric interactions

Allosteric interactions describe interactions between spatially separated regions of a protein and were first described for oxygen binding within hemoglobin. In enzymes, allostery describes the binding of an effector molecule at a site different from the active center, affecting the substrate affinity of the enzyme by conformational changes (227). As a special case of allostery, **cooperative interactions** describe the alleviated binding of a molecule A to its target, after a first molecule B has already bound (174). Cooperativity is further distinguished in homotropic and heterotropic cooperativity, depending on whether effector and substrate are the same molecule or not, respectively. This concept can be generalized and transferred from enyzmes to protein complexes where interactions between protein(domains) dictate conformations which influence the signaling activity of the whole complex (228). Kinetically this means that the equilibrium constant for association of, for example, two dimers B·A (multimerizing to B·A·A·B) is greater than that for A + A → A·A alone see Figure 3.2 (229).

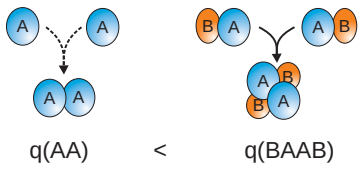


Figure 3.2: In case of cooperativity the equilibrium constant $q = \frac{k_a}{k_d}$ for association of two proteins A is greater after binding of an allosteric effector B. k_a and k_d denote the rate constants of association and dissociation respectively.

The model was designed as signal regulatory network which consists of protein and RNA species relaying a signal rather than a mass flow. As such, this structural model shows dependencies of reactions rather than their kinetics, which for many parts of the network is still unknown. However, spliceosome assembly has been shown to form a complex interacting network of subcomplex formation (151) best described as an allosteric cascade (163) that frequently involves *cooperativity* (174) (see Box 3.1.1).

However, spliceosome assembly has been shown to form a complex interacting network of subcomplex formation (151) best described as an allosteric cascade (163), that frequently involves *cooperativity*(174) (see Box 3.1.1). Already a decade ago, it was proposed that the composition of spliceosomal snRNPs could be different on different splicing substrates in a context-dependent manner (tissue type, developmental stage). This has the consequence that many different combinations of factors could potentially give

rise to different types of active spliceosomes, thus, increasing the potential for splicing regulation (230).

In light of the wealth of proteins that up to now have been identified around the spliceosome , it is necessary to go beyond cataloging the factors. Putting protein and mRNA factors together into the context of a larger network poses a difficult task and requires the integration of a standardized vocabulary to unambiguously describe all possible reactions. While this task will remain a challenge for ontology and text mining specialists, here, it was started to summarize reactions, such that they can be utilized for a PN model and subsequent analyses, e.g. *in silico* testing of hypotheses. A number of analysis strategies accustomed to PN representation of signaling networks is presented in the following sections.

3.2 Methods

3.2.1 Definition of Petri Nets

The spliceosomal assembly network was modeled as a P/T net (see Box 3.2.1). Places correspond to biological objects (e.g. RNA regions, protein factors, protein complexes etc.), whereas transitions correspond to processes, which act upon objects (e.g. protein interaction, phosphorylation reactions, protein-mRNA binding etc.). The direction of arcs defines pre-places (pre-transitions) and post-places (post-transitions). Tokens represent movable objects. They are used to model the equivalent of signal or mass flow units and are symbolized by black dots on places. The maximum number of tokens that a place can hold is defined by its *capacity*. The distribution of tokens over all places is called a *marking*. Each marking defines a certain state of the system.

Box 3.2.1 Petri net definitions: A *Petri net* (PN) is a directed, labeled, bipartite graph consisting of places (circles), transitions (rectangles) and arcs (arrows), such that arcs connect only nodes of different type. A *PN* is a six-tuple $Y = (P, T, F, K, W, M_0)$ and denotes a **Place / Transition -net** if the following definitions hold (231):

i. The tuple (P, T, F) is a net graph N with

(a) $P = \{p_1, p_2, \ldots, p_n\}$: a finite set of *places*

(b) $T = \{t_1, t_2, \ldots, t_m\}$: a finite set of *transitions*

(c) $F \subseteq (P \times T) \cup (T \times P) \neq \emptyset$: the set of *arcs* (flux relations of N)

(d) $P \cup T \neq \emptyset$: union of P and T is never disjunctive

(e) $P \cap T = \emptyset$: sets P and T are disjunctive

ii. $K : P \to \mathbb{N} \cup \{\infty\}$: the *capacity* of places

iii. $W : F \to \mathbb{N}$: the arc weights

iv. $M : P \to \mathbb{N} \cup \{\infty\}$: the *marking* of Y if
$\forall p \in P : M(p) \leq K(p)$ with:

- $M_0 : P \to \mathbb{N}_0$: the *initial marking*
- $M_0 \xrightarrow{w} M'$: the *consecutive marking* given a firing sequence w
- $M_0 \xrightarrow{w}$ defines a firing sequence, which is activated under M_0 for which holds:
  $w = \emptyset \wedge M' = M_0$ or
  $\exists M' \in M(Y) : M_0 \overset{t_1 \ldots t_k}{\to} M'$

Places and transitions are connected via *directed arcs* such that arcs connect only nodes of different type. Tokens represent movable objects. They are used to model the equivalent of signal and/or mass flow units and are symbolized by black dots on places. The maximum number of tokens that a place can hold is defined by its *capacity*. Adjacency relations between nodes are defined by incoming and outgoing arcs. Hence, the set

$$\bullet x := \{y \mid (y, x) \in F\} \tag{3.1}$$

defines *pre-places*, which feed a transition $y \in T$ with tokens. Transitions without pre-places are called *input transitions* and represent external sources.

Accordingly, the set

$$x\bullet := \{y \mid (x, y) \in F\} \tag{3.2}$$

defines *post-places*, which draw tokens from a transition $y \in T$. Transitions without post-places are called *output transitions* represent external sinks. PN are commonly modeled as transition bounded net, where each node $x \in P \cap T$ is a boundary node if $\bullet x = \emptyset \wedge x\bullet = \emptyset$ (232).

The dynamic behavior of the network is realized through the firing of transitions which model the activity of biochemical reactions. A transition can fire (is enabled) if all pre-places are covered by at least one token (see Box 3.2.2). Tokens represent molecules or moles. During *firing* of a transition according to the corresponding arc weights and firing rules, the number of tokens is decreased on the pre-places and increased on the post-places. Consequently, the *marking* of the net is changing, resulting in a new state of the PN. Note that in signal transduction networks without mass flow and reaction stoichiometry, it is reasonable to interpret arc weights rather as „information units“ with a default value of one than as reaction quantities (see Box 3.2.3).
Starting from an initial marking $\mathcal{M}_0$ one can define a *firing sequence* $\mathcal{M}_0 \xrightarrow{w}, w = t_1 \dots t_k \in \mathcal{T}$ (*cf.* Box 3.2.1) as a subset of transitions within the PN, which corresponds to a specific signal propagation through the biological assembly network. For each firing sequence w a frequency vector $\bar{w} = (\#(t_1, w) \dots \#(t_n, w))$ (also called *Parikh-vector*) can be assigned, which indicates how often each transition fires. The change of the net marking can be determined by:

$$\mathcal{M}_0 \xrightarrow{w} \mathcal{M}' = \mathcal{M}_0 + \mathcal{C} \cdot \bar{w} \tag{3.3}$$

whereat $\mathcal{C}$ is the incidence matrix, in which rows and columns correspond to P and T, respectively, and the matrix elements describe the change of token number on a place, when a transition fires. For metabolic networks, the incidence matrix corresponds to the stoichiometric matrix.

Box 3.2.2 PN firing rules

The activity of reactions in a PN system (defined in Box 3.2.1) is simulated by firing of transitions, which is symbolized by a token change. A transition t can only fire, if it is enabled by satisfying following two conditions:

i. $M(t\bullet) \geq W(p_i, t)$: all its pre-places are occupied by at least as many tokens as the weights of the incoming arcs prescribe.

ii. $K(t\bullet) \geq W(t, p_i)$: all its post-places must have at least a capacity as the weights of the outgoing arcs dictate

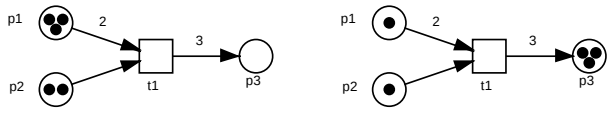


Figure 3.3: Transition firing behavior: **a)** Transition $t1$ is enabled because pre-places $p1$ and $p2$ are occupied by as many tokens as their arc weights prescribe. No indicated arc weight means per default an arc weight of one, **b)** After firing of $t1$ two tokens of $p1$ and one token of $p2$ are consumed and three tokens of $p3$ are produced. Note that in signaling networks, the number of consumed tokens must not necessarily be equal to the number of produced tokens.

The firing sequence w can be determined from the solution of the homogenous equation system $\mathcal{C} \cdot \bar{w} = 0$ which is valid if the signal or information flow within the network is assumed to be conserved. This forms the base to compute systems invariants from the incidence matrix, which can be divided in two different vectors $\bar{w} = \bar{x}$ or $\bar{w} = \bar{y}$ depending on the orientation of $\mathcal{C}$.

The vector $\bar{x}$ is called *non-negative place invariant* (P-invariant) if it solves the homogeneous equation system

$$\mathcal{C}^T \cdot \bar{x} = 0 : x_1 \dots x_n \in \mathbb{N}_0 \tag{3.4}$$

The elements of a P-invariant can be interpreted as *conservation relation* for tokens. For an initial marking $\mathcal{M}_0$ holds:

$$\forall \mathcal{M}' \in [\mathcal{M}_0\rangle : \mathcal{M}' \cdot x = \mathcal{M}_0 \cdot x \tag{3.5}$$

whereat $\mathcal{M}'$ defines a consecutive marking (see Box 3.2.1), that is, a subset of the reachability set $[\mathcal{M}_0\rangle$, which defines all possible states in the net that can be reached from $\mathcal{M}_0$ by firing of w. The solution vector $\bar{y}$ is called *non-negative transition invariant* (T-invariant) if it is a solution of the homogeneous equation

system

$$\mathcal{C} \cdot \bar{y} = 0 : y_1 \dots y_n \in \mathbb{N}_0 \tag{3.6}$$

A T-invariant is a transition sequence that after firing reproduces a certain marking (state) of the network (Equation 3.5). A T-invariants *Parikh-vector*, indicates how often each transition has to fire in order to reach the same state (marking) again.
T-invariants are *minimal* if there exists no smaller positive T-invariant such that $\bar{y}' : (\bar{y} - \bar{y}') > 0$ (231). Hence, minimal T-invariants are not further decomposable into smaller T-invariants. The same holds for minimal P-invariants. In the following the term *„T-invariant“* (*„P-invariant“*) stands as abbreviation for minimal non-negative T-invariant (P-invariant).

T-invariants can be interpreted as flux vectors. In biochemical terms, it was shown, that under *steady state* conditions a metabolic network can be decomposed into sets of minimal meaningful reaction sequences („elementary flux modes“), which may form a variety of flux patterns when expressed as non-negative linear combinations (220, 221). Elementary flux modes correspond to non-negative minimal T-invariants.

Box 3.2.3 PN as signal relay networks

Signaling pathways are often mixed in their types of involved reactions. Commonly known are phosphorylation events where kinases modify special domains of proteins, resulting in the capability to bind proteins. In contrast to metabolic networks, where an enzymatic reaction is clearly defined by its stoichiometry, in signaling events one often observes only a state that coincides with a certain function, without exact knowledge about the amount of the participating signal molecules. For example, the proteins A and B form the dimer AB which may function as the necessary signal to recruit an important third factor C. Reducing the arc weight to one is a meaningful simplification to circumvent the lack of stoichiometric information. It can be interpreted as the minimal *information unit* triggered by an *maintenance concentration* of the necessary molecules, to achieve the minimal requirement for signal transduction.

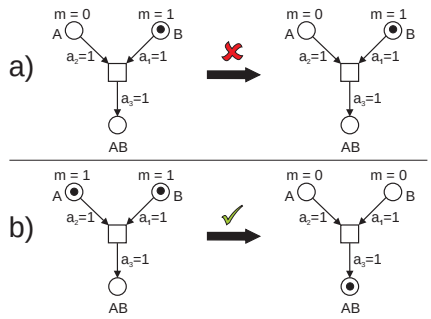


Figure 3.4: Example of signal flow in PN via a reduced arc weight to one, m = number of tokens per place, a = arc weights, A and B = interacting molecules a) only factor A is present, indicated by the token at place B, hence the transition, which models the reaction is not enabled b) both required factors are present, indicated by the token on each place of A and B respectively, hence the transition can fire and generates the signaling complex AB

The spliceosomal assembly net was modeled as transition-bounded net, such that no places without pre- or post transitions exist, but transitions without pre- and post places. The boundary transitions describe reactions that connect to external sources (input transitions) and external sinks (output transitions) of the assembly network, whose reactants can be defined as buffered substances at fixed concentrations (220). In biological interpretation that means, all reactants feeding input transitions or leaving output transitions are considered to be external.

All reactants (places) are modeled in non-limited amounts with a capacity of $K(p) \rightarrow \infty$ and an inital marking of one token per place to enable each transition. The model has been validated using PN analysis techniques. First, by determining the static and dynamic properties, using the programs INA (233) and PINA (234), and second by computing substructures as *maximal common transition sets* (MCTS). The PN model was designed, using the PN editor SNOOPY (235). The initial XML format, was converted into a *pnt* file format which served as input file for the structural analysis by INA using the command sequence given in Table C.2.

With increasing network size and complexity, the number of T-invariants can ex-

ponentially grow. Two approaches were employed to facilitate the validation of the model: *i)* decomposition into disjunctive sub networks (MCTS) and *ii)* decomposition into overlapping sub networks (T-clusters)

Partitioning of T-invariants into MCTS

MCTS are based on a matrix D in which rows and columns correspond to T and X, respectively, with T defining the set of transitions and X defining the set of T-invariants. Each row constitutes a subset $I \subseteq X$ of T-invariants that share a considered transition t. Biologically, this means that a specific reaction is part of a certain number of all possible and minimal *steady state* signaling pathways within the network. All transitions, which in this way are shared exclusively by the same set of T-invariants, form an MCTS $A \subseteq T$ for which holds:

$$\forall t_i, t_j \in T : t_i, t_j \in A \iff I(t_i) = I(t_j). \tag{3.7}$$

The set of all transitions of T-invariant $x \in X$ is called support of x and denoted $supp(x)$. Given Equation 3.7 it follows that each MCTS is either a subset of the support of a T-invariant or does not participate in a T-invariant at all:

$$\forall \vec{y} = x \in X : A \subseteq supp(x) \ \vee \ A \cap supp(x) = \emptyset. \tag{3.8}$$

Clustering of T-Invariants

A clustering of the T-invariants was performed to find similarities imposed by transitions that are shared between different T-invariants. T-clusters define sub networks that can overlap (234). They facilitate the identification of frequently traversed routes, which are formed by common subsets of reactions and are thought to highlight more important structures within the net. The cluster analysis was done by computing a distance measure according to the transformed Tanimoto coefficient (234), which is also known as binary distance or Jaccard index. This measure is well suited for comparing signaling pathways in vector representation, because it is robust against variations in the total number of reactions in a network by considering reactions, missing in both of the compared T-Invariants, as additional evidence of similarity. Hence, three binary states can be specified: M_{11}, the number of reactions present in both compared T-Invariants, M_{10}, the number of reactions present only in T-invariant T_i and M_{10} the number

of reactions present only in T_j. The Jaccard-Tanimoto similarity measure follows as:

$$J(T_i, T_j) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \tag{3.9}$$

and the Tanimoto distance as:

$$D(T_i, T_j) = \frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11}} = 1 - J \tag{3.10}$$

The corresponding distance tree of related T-invariants was constructed by the UPGMA-algorithm (236). A threshold of 80% was chosen to merge T-invariants with less than 20% difference into the same subtree. The same distance measure and clustering algorithm was used to create a *color map* (*cf.* Figure 3.14). A color map is a graphical way of displaying matrices by using colors to represent the numerical values. Due to the binary (on/off = present / non-present) status of transitions within T-invariants, a simplified two color mode was chosen, where dark and light blue tones indicate the presence and absence of an reaction, respectively. The color map also rearranges rows and columns of the distance matrix so that similar rows, and similar columns, are grouped together, according to the distance tree. This representation facilitates to visualize block patterns of transitions, shared by different T-invariants.

3.2.2 Model Refinement

Many biological signaling processes in the human cell are well documented, for example the caspase cascade of apoptosis or the communication network of cytokines between immune cells. Reactions of these signaling pathways can be found in databases as KEGG (237) or TRANSPATH (238). Although the spliceosome is for many years under investigation, no consistent and wholistic network has been published so far. Reactions involved in spliceosome assembly were biochemically described, but not formalized.

Because of the multitudity of involved proteins a high initial effort was necessary to manually review literature that describe or suggest reactions involving these proteins. Known experimental results were taken as base to extract reactions applicable for designing the PN model, but were often only vaguely or contradictory described. Therefore, depending on available data, reactions and its participating factors were summarized or abstracted. A naming convention for reactions was

introduced to give a quick idea of the nature of the reaction (e.g., „_bdg“ for binding „_matur“ for maturation, „_ass“ for assignment). All reactions used for the model are summarized in Table C.1.
A processing scheme was developed (see Figure 3.5), which guides from the extraction of reactions from literature, through their incorporation into the model network to the subsequent validation by T-invariant and MCTS computation. In average, each protein or reaction of the network was supported by at least one experimental observation or thoroughly inferred from such. This time intensive preparation could be more automatized in future, but human inspection for correctness of the pathways will remain an important aspect of quality control of the final model. The processing steps that were iteratively applied to extend and revalidate the model, can be summarized as follows:

i. Collect review articles about E- and A-complex assembly

ii. Identify and extract reactions, and develop the first PN model

iii. Compute T-invariants, MCTS and T-clusters

iv. Check biological meaning of T-invariants/MCTS:
first model check → describe and annotate T-invariants/MCTS
repeated model check → compare previously computed T-invariants/MCTS if necessary, update annotation
implausible result → return to literature, consult independent reports on reactions that appear not to be covered by T-invariants

v. Check composition of T-clusters

vi. Extend model by further reactions reported for B and C-complex formation

This procedure allows the iterative and semi-automated evaluation of the results of the T-invariant / MCTS computation to provide support for the PN model. However, presently the raw output of the PN editor tool (235) has to be manually checked after each processing round, because of the rearrangement of internal node IDs after editing the network. This is a limitation of the present PN editing software. After this intermediate check, all reactions (see Table C.1) were automatically written and formatted into tabular form by a Perl script. Nevertheless, consultation of experimental literature often involve an elaborate search for reports, which describe not only the immediate partners of the reacting molecule,

but if possible also surrounding processes for connecting the reaction appropriately to the existing network.

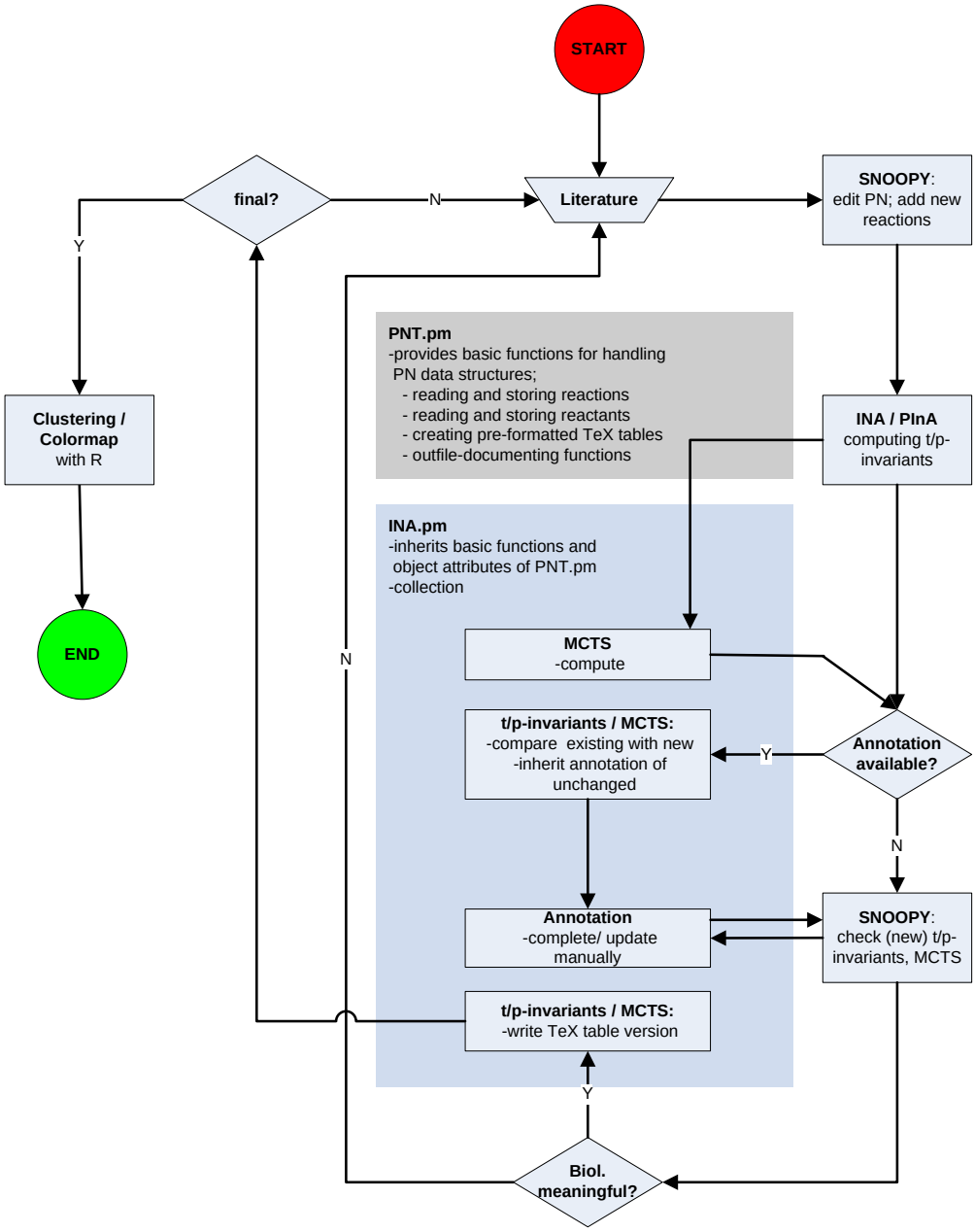


Figure 3.5: Overview of methodology applied to model the spliceosomal assembly pathway. The inner part shows steps that involve in house software (Perl) for reformatting and analyzing the output produced by external tools.

3.2.3 Data Preparation

To handle the output given by the program INA, several perl scripts were written, which run on top of two libraries, collecting functions for analysis of **.pnt* and **.ina* files. These served to:

i. Rewrite summarizing reactions from **.pnt* files, provided by INA (*get_reactions.pl*)

ii. Find inconsistencies within the model by compiling the offending transitions in case the network is not completely covered with transitions (*check_coverage.pl*)

iii. Compute MCTS and print them in two different summaries (1. fractions of T-invariants sharing the MCTS; 2. number of transitions involved in the MCTS) along with some descriptive figures about types and frequencies of T-invariants (*compute_mcts.pl*); this script also creates a core file for annotating the MCTS, from which at a second run the annotation is written into a ready-to-use formatted table

iv. Annotate T-invariants and print them into a formatted table (*print_tinv_report.pl*)

3.3 Results

3.3.1 Application of PN Modules in Splicing Models

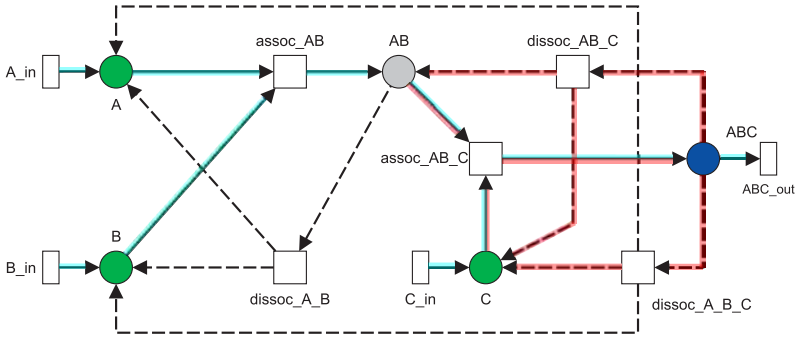
Inspired by previous suggestions (209, 239, 240), a number of smaller net modules were designed at first, which served as building blocks describing different reactions or interactions between spliceosomal components. To reach a valid model, these net modules are useful for testing modeling strategies, which appropriately reflect observed biological behavior of parts of the network. In general, the modeling of biologically meaningful modules within signaling pathways strongly depends on the depth of experimentally verified knowledge of the described mechanism. The following net modules have been used for modeling the basic reactions of spliceosomal assembly:

i. **Allosteric interaction** describes the process in which a protein binds to a specific domain of a target protein, induces a conformational change at a distant site, and hence rendering the target protein itself active or inactive. In spliceosomal processes this concept can be extended to the level of protein complex association, where the binding of special factors is crucial for subsequent progress through intermediate assembly stages (see Box 3.1.1). A module for this biochemical process is decomposable into four T-invariants, two of which being cycles hat describe the repeated association and disintegration of the intermediate complex, AB, and the final complex, ABC (*cf.* Figure 3.6). Note that dissociation is restricted to $AB + C$ or $A + B + C$, since AC or BC are *„forbidden“* by the allosteric rule imposed during complex formation. The same model strongly reduces structural complexity by exhibiting one T-invariant, covering all transitions.

   Further special cases of allostery can be distinguished, and accordingly different net modules were designed.

   (a) **Allosteric inhibition** depends on the presence of a specific domain within a subcomplex, an inhibitor may bind to the complex, inducing either disassembly of the complex or non-functionality of the complex towards specific downstream interactions (*cf.* Figure 3.7). This leads to an extension of the module depicted in Figure 3.6b, where the heterotrimer ABC can either participate in further reactions (ABC_out), dissociates or binds to an inhibitory factor I. After sequestration of

(a) Allosteric cascade with dissociation



(b) Simplified allosteric cascade

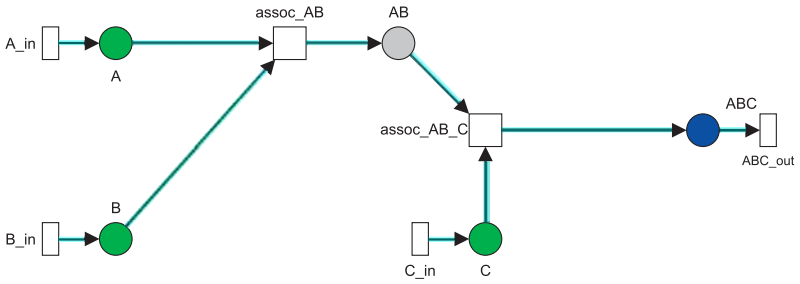


Figure 3.6: Example of a module for protein complex association where two molecules A and B form a heterodimer, AB, which defines a necessary step for binding the factor, C, in progressing the complex assembly through complex ABC. **(a)** PN module with reactions, forming (solid lines) and decomposing (dashed lines) intermediate complexes. Two of four computed T-invariant pathways, completely covering this module are colored and define the main signaling route (blue) and a cycle (red); green = source factors, grey = intermediate factors (complexes); blue = target complex; **(b)** The same model without dissociation reactions strongly reduces structural complexity, leading to one T-invariant.

complex IAB, AB may either dissociate again or remain for a certain time non-functional within IAB. Thus, it is modeled as output transition (IAB_out). The module designed with dissociation reactions again results in several T-invariants (data not shown), including sustained cycles of associations and dissociations. In contrast, the module reduced for dissociation reactions exhibits two minimal T-invariants, reflecting only the important aspects of functional ABC and non-functional IAB formation. The simplified version may suffice in many cases, in particular when time points of protein activities are yet unknown, for example, the time when a specific factor dissociates from an intermediate complex. This net module is easily modified from the basic allosteric cascade shown in Figure 3.6b) by adding the

inhibitor I and introducing an additional edge that purges AB.

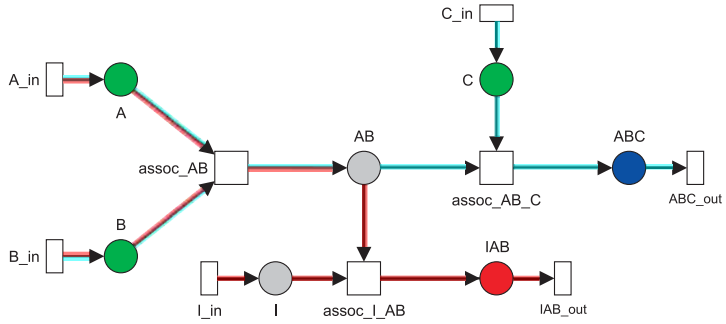
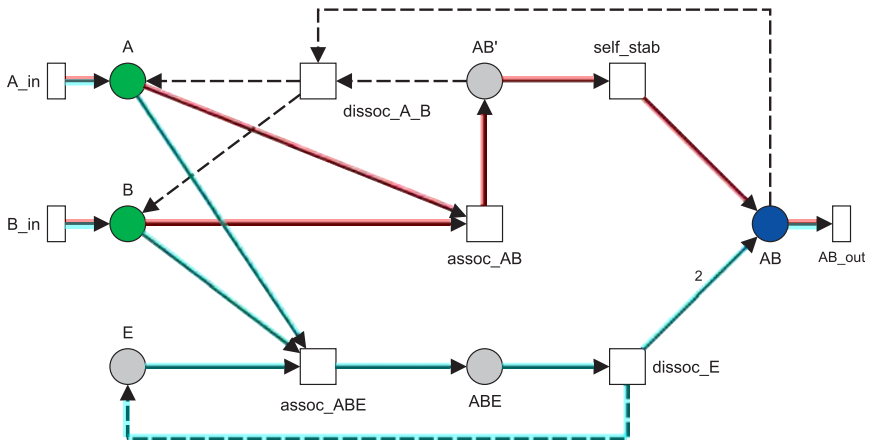


Figure 3.7: Reduced PN module which describes the formation of an inhibiting intermediate complex (IAB). The blue highlighted pathway covers reactions which result in a functional target complex ABC, while the red colored pathway describes a T-invariant covered by reactions which result in a non-functional complex IAB.

(b) **Allosteric enhancement** describes the presence of a specific protein factor, which increases the affinity of other proteins to participate in subsequent reactions (e.g., subcomplex formation, RNA recognition etc.). The structural analysis results in four T-invariants, two of which producing the target complex AB. The interaction of factor A and B can result in a dimerized complex AB (3.8a, red pathway), however, the enhancer may be necessary for the protein (complex) to be active. The model accounts for the presence of enhancer E with a higher output of AB (Figure 3.8a, blue pathway) due to an increased arc weight. Hence, transition AB_out has to fire twice to reproduce the initial marking. Biologically, this can be interpreted as an increased signal transduction as AB reaches a state of higher disposition for participating in downstream reactions. The reduction of the system for the dissociation reaction of the dimer AB decreases the number of T-invariants by one (Figure 3.8b). Two T-invariants involve transition $assoc\_AB$ but only one produces AB, while the other purges AB from the network (Figure 3.8b, red pathway). The enhancer involving pathway via $assoc\_ABE$ again produces an increased amount of AB, thus the reduced model captures all essential aspect of the enhancer dependent complex formation.

ii. **Enzymatic reactions** describe the reactions where a catalytically active enzyme acts on molecular groups of spliceosomal proteins, e.g., kinases

(a) Allosteric enhancement with dissociation



(b) Simplified allosteric enhancement

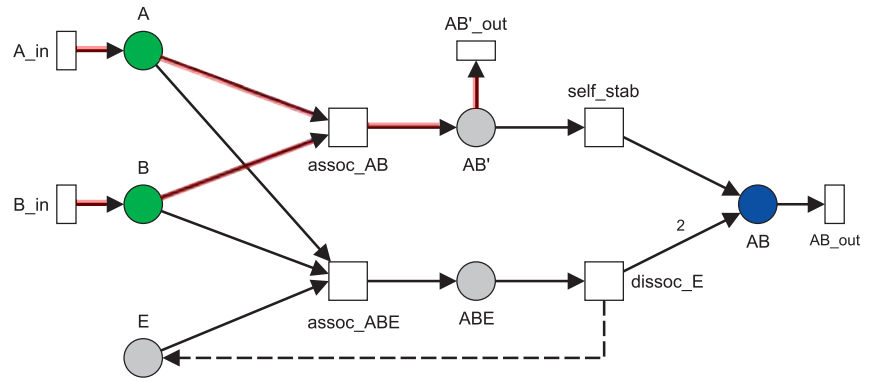
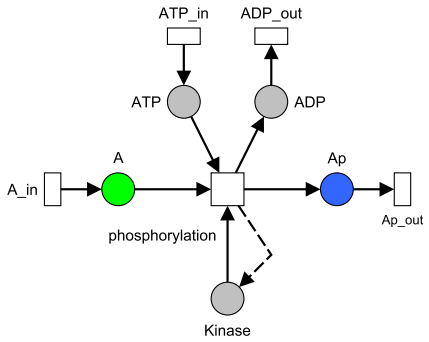


Figure 3.8: PN module for modeling enhanced protein-protein interaction. The presence of enhancer proteins stabilizes complex formation and results in an increased output of dimer, AB, compared to dimerization without the influence of enhancer protein, E. Note that in contrast to Figure 3.6, enhancer E only mediates a temporary effect until the dimer, AB, has stabilized its interaction. **(a)** The module with a dissociation transition for dimer, AB, which results in four T-invariants, including one with a higher output of AB (blue pathway) and one producing AB via a possibly undirected self stabilized interaction (red pathway) **(b)** Replacement of the dissociation reaction by an output transition (red pathway) reduces the number of T-invariants, while preserving the essential model function of enhancer dependent complex formation.

phosphorylate proteins. This behavior was modeled as loop, which preserves the marking of the respective place and results in a simple conservation relation (Figure 3.9a). Also helicase like proteins with DExD/H box domains have been frequently found in purified spliceosomes (149) and were here considered as separate module. This module was extended for another reaction, representing the enhancement of substrate specificity of the helicase (see Figure 3.9). In this context, the rate of NTP hydrolyzation has

been proposed as a crucial parameter for splicing fidelity, since fast kinetics on weak or incorrect protein-substrate (RNA, protein) interactions increase the chance of dissociation of essential spliceosomal proteins. In consequence, such defective substrates could be submitted into a degradation pathway (165). While a putative degradation pathway was integrated as a branch into the PN, the hydrolyzation activity could not be modeled, because of the lack of kinetic parameters. Finally, since several DExD/H box proteins are involved in spliceosome assembly, the total accuracy of splicing may depend on the cumulative success of enzyme modulated signal transduction along the pathway, further complicating the corresponding kinetic model.

(a) Phosphorylation



(b) Helicase specification

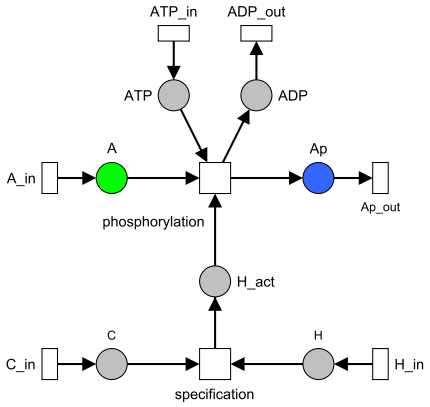


Figure 3.9: PN modules, modeling enzymatic activities found in protein complex assembly. **(a)** Model of protein activation where phosphorylation of a specific domain mediates interacting capabilities, thus, influencing subsequent complex rearrangements; **(b)** Binding of a specific factor to a generic helicase, thus, mediating substrate specificity.

3.3.2 Defining Biological Reactions of the Spliceosome Assembly Pathway

In an iterative process (see Diagram 3.5) literature was inspected to isolate reactions involved in spliceosome assembly. Reactions and proteins focus on knowledge reported for human spliceosomal processes, however some parts where first discovered in yeast and the human homologue proteins discovered later. In this way the model does not discriminate between an explicit human or yeast spliceosomal assembly model. All reactions are ordered according to the four major stages of spliceosome assembly, E-, A-, B- and C-complex (see Table C.1). After

each major stage the structural composition of the assembly network was evaluated by T-invariant analysis. The net was only extended if each modeled reaction was part of at least one T-invariant. Reactions were only included if the factors involved could be integrated in a causal order. As consequence, some spliceosomal factors, e.g., SPF27, CypE, HSP73 and others, are omitted from the model due to lack of evidence for a specific time point at which they participate in the spliceosomal assembly process. Further, it was assumed that all factors for which no evidence of leaving or becoming destabilized in spliceosome assembly is given, implicitly remain in the spliceosomal subcomplexes until finally a dissociation into substructures (e.g., snRNPs) and their recycling takes place. Hence, not all factors modeled with an input transition have an explicit output transition. This is reasonable since many substructures remain intact for repeated rounds of spliceosomal assembly (241) This results in a model with 140 reacting species (places) comprised of RNA, proteins and intermediate complexes was established, which covers about half of the currently known spliceosomal proteins. The network was modeled as transition-bounded network with all places connected to a total of 161 transitions. 92 (57%) of those are boundary transitions, splitting into 68 (74%) input and 24 (26%) output transitions. In total 69 (43%) transitions describe the internal reactions of the assembly network. This biological network was modeled by reactions, which reflect a certain hierarchy characteristic for spliceosome assembly. Therefore and in contrast to PN models for technical processes, the degree of concurrency of reactions is reduced in the network.

3.3.3 Invariant Signaling Pathways in Spliceosome Assembly

Pathways Corresponding to T-Invariants

Structural analyses resulted in a complete coverage of the network by 71 T-invariants (see Table C.3). All T-invariants describe at least one partial pathway within the spliceosomal assembly process and are biologically meaningful. 12 T-invariants (17%) can be considered to be trivial as they describe solely the in- and efflux of the DExD/H box helicases Prp5, Prp28, the proteins hPrp6, SKIP and hDib1, the splicing factors SF1, ASF/SF2, SC35 and the SF3b components SF3b10, SF3b14a, SF3b14b and SF3b49. However, these proteins also constitute a set of spliceosomal components for which it is possible to narrow the putative point of exit from the assembly process, which so far is unknown for the majority

of proteins functioning in spliceosome assembly. For example, the four smaller SF3b proteins are known to be required for SF3b formation, but are not detected at the stage of C-complex formation (*cf.* supplements of (151)). Each set of reactions of a T-invariant provides a solution of the homogenous equation system derived from a matrix of all reactants and reactions under assumption of a steady state system. As result the removal of a single reaction from an invariant signaling pathway is critical for functioning of this pathway. In case of spliceosome assembly, this does not mean that the entire assembly process is stalled. The model clearly illustrates that different results from experimental studies shape up to a network with an inherent redundancy to sustain specific check points which represent crucial intermediate assembly stages. For example, a critical step of early spliceosome assembly is the donor or 5′ splice site recognition, for which several parallel occurring signaling pathways were modeled. These pathways contribute to the intermediate stage of E-complex assembly and are reflected in the model by several sets of T-invariants which are listed below. Note that enumerated T-invariants are prefixed by „i“. T-invariants in brackets describe spliceosomal subpathways that involve 5′ss recognition, but do not proceed via the productive branch of C-complex formation that results in exon ligation:

i. T-invariants i13, i14 (i68, i69) → U1 snRNP independent 5′ss recognition: The central function of this branch is transition *t16.U1_indep_5ss_act*, which models the activity of the SR protein SC35. The presence of SC35 is sufficient to define a 5′ss in absence of a functional U1 snRNP (143) and initiates contacts to the BP occupying proteins SF1 and U2AF. This can result in selection of competing 5′ss, which render this pathway a potential candidate for the production of alternatively spliced mRNAs (171). The remaining transitions of this branch are *t31.U2_BPS_bdg2* and *t32.U6_5ss_bdg2*, which feed U2 snRNP and U4U5U6 tri-snRNP respectively and proceed the assembly pathway to the B-complex stage. Further differences in the T-invariants, sharing this otherwise unique branch of 5′ selection, exists in two different ways of A-complex assembly via early *t13.17S_U2_matur2* (i13, i68) or late *t22.17S_U1_matur1* (i14, i69) action of the enzyme SF3b125. However, for i14 it is not clear at which time SF3b125 leaves the assembly process, hence this T-invariant includes additionally the input reaction *t109.SF3b125_in*. Another variation between these similar T-invariants is the different proceeding during C-complex assembly after the remodeling step *t93.U2_3ss_U6_5ss_U5_remod*. Either a productive spliceosome is

formed via one type of pathway (i13, i14) or the premature disassembly is reflected via another type of discard pathway (i68, i69).

ii. T-invariants i15-i20 (i62-i67) → ASF/SF2 dependent 5′ss recognition:
These T-invariants describe the 5′ss recognition via contacts of U1 factor U170K and the exon-bound splicing factor ASF/SF2 (*t12.ASFp_U170K_bdg*), with subsequent contacts of U1C to intron bound splicing factor TIA1 (*t9.U1C_TIA1_bdg1*) (242, 243, 170).The U1 snRNP protein U170K hereby interacts with ASF/SF2 via its RS domains (169). Subsequently, these T-invariants show some individualities, allowing to form three groups.
(a) Firstly, T-invariants i15, i16 (i66, i67) describe the E-complex formation via U1 contacts to the branchpoint bound splicing factor SF1 (*t58.U1_SF1_bdg*) and the joining of the auxiliary factor U2AF, after its recognition of the polypyrimidine tract *t59.U1_SF1_U2AF_bdg*. Dependent on the U2 maturation via SF3b125, there exist again two different T-invariants for this mode of E-complex formation.

(b) Secondly, the presence of the splicing factor SC35, has been found to facilitate 5′ss recognition (*t17.U170K_U2AF35_bdg*) but required the U1-complex and U2AF ((244)). Hereby, the protein FBP11 helps to bridge U1 and SF1, a constellation which is in agreement with the observation that SF1 and U2AF bind cooperatively to the branchpoint and polypyrimidine tract, respectively (175, 174). This mode of E-complex assembly is reflected by the T-invariants i17, i18 (i64, i65).

(c) Thirdly the T-invariants i19, i20 (i62, i63) describe a pathway of E-complex assembly, in which the splicing factor SC35 can substitute for the requirement of the auxiliary factor U2AF. Again, SC35 bridges the 5′ss and the branchpoint via contacts to U1 and SF1 *t56.U1_SC35_SF1_bdg*, but subsequently U2 is directly bound to this intermediate complex *t152.U2_SC35_bdg*. It should be noted that in the subsequent A-complex formation this pathway contains some ambiguity in that the helicase UAP56, which is required for conformational change in transition from E to A-complex, needs the U2AF component U2AF65 as essential cofactor for its activity (184, 165).

iii. T-invariants i21 - i26 (i56-i61) → 5′ss recognition in the 5′ terminal exon:
The first donor splice site within a transcript follows a different mode of recognition. Here, the interaction of U1 snRNP proteins (U1C) to proteins of the cap binding complex (CBC) via LUC7 has been shown and hence was modeled with a separate reaction (*t10.U1_CBC_5ss_bdg*), which triggers another set of T-invariants. However, except for the interaction with the cap binding complex, these T-invariants are the same as above (**ii a-c**) and may constitute novel alternative pathways for the initial spliceosome assembly at newly synthesized transcripts.

iv. T-invariants i27 - i32 (i50-i55) → 5′ss recognition via U1C contacts to intron bound splicing factor TIA1:
This branch models the 5′ss proximal (intron) binding of the splicing regulator TIA1, which stabilizes contacts of U1 with the 5′ss via interaction with the N-terminal region of the U1-C protein (*t30.U1C_TIA1_bdg2*) (170). As alternative pathways to the proposed joint action of ASF/SF2 and TIA1 in 5′ss selection (as described above), T-invariant i29 (i52) involve SC35 and FBP11 in subsequent A-complex formation or either only SC35 (i27, i54) or only FBP11 (i31, i50) as U1 snRNP binding supporting splicing factors.

v. T-invariants i33 - i38 (i44-i49) → 5′ss recognition without additional splicing factors:
These T-invariants model the 5′ss recognition without the parallel binding of the U1 stabilizing factors ASF/SF2 or TIA1, which could be described as a mode for strong donor splice site selection. However, the transition from E to A-complex may require conditions as described for the T-invariants in **ii a-c**.

vi. T-invariants i39, i40 (i70, i71) → 5′ss binding by U1 snRNP after initiating contact to U2 snRNP via Prp5:
These T-invariants describe a pathway of splice site selection, which deviates from the classical model of initial 5′ss selection. The ATPase Prp5 bridges U1 and U2 snRNP prior to substrate binding, that is U2 is already associated with U1 and Prp5 when it binds to the intron, a fact that is experimentally affirmed by the binding of Prp5 to U1 and U2 also in absence of pre-mRNA and by the ATP dependent requirement for Prp5 for pre-spliceosome assembly (181). The T-invariants therefore contain the reaction *t27.U1_Prp5_U2_bdg* as bridging step, followed by *t28.U1_5ss_U2_U2AF_bdg*

describing the contacts with the 5′ss and the branchpoint/polypyrimidine tract associated proteins. The ATP dependent structural rearrangements towards A-complex assembly are modeled by *t55.unwind1_U2_stl2* a step that releases the U2 factor SF3a60 (165). The UAP56 catalyzed conformational changes in the U1/U2 pre-mRNA complex are modeled by *t29.U1U2_BPS_bdg* and complete the transition from E to A-complex.

Shortly after or in parallel to the recognition of the 5′ss by U1 snRNP, the U2 snRNP joins the assembly pathway and defines the branchpoint region. Maturation of the 17S U2 snRNP was proposed to proceed via two different actions of the putative DExD/H box helicase SF3b125 (182). This enzyme can act at an early stage of 17S U2 formation by catalyzing a conformational change when SF3b is integrated into the 12S U2 snRNP to form the intermediate 15S U2 snRNP subcomplex (*t13.17S_U2_matur2*). Alternatively, SF3b125 may act subsequent to the binding of SF3b, in this way supporting the conformational rearrangement to integrate the SF3a subcomplex into the U2 snRNP (*t22.17S_U2_matur1*). Experimental evidence suggests that this putative enzyme is largely dissociating during 17S U2 assembly (182). Hence, at least one of the alternative reactions was modeled to set SF3b125 free from the U2 snRNP maturation subpathway. Due to the two different U2 snRNP maturation scenarios, the number of T-invariants is doubled for all subsystems, which require the presence of a mature U2 snRNP, thus demonstrating the emergence of combinatorial complexity in the modeled system.

A similar property of the network can be observed during late spliceosome assembly, where a branching of the pathway was modeled according to the proposed function of the Prp16 DExD/H box helicase. Although the model does not reflect the kinetic behavior of Prp16 in detail, the effect of two different possible kinetics can be described. The proper kinetics of Prp16 requires a proper substrate of pre-mRNA and snRNP conformations and may channel spliceosome assembly into a productive pathway of C-complex assembly, such that the second step of splicing and exon ligation can proceed (via *t101.Prp16_remod_step*). In contrast, mutations in the involved RNA species or unfavorable conformations due to missing proteins, can result in a slowed Prp16 kinetics which was proposed to activate a discard pathway (245, 241) reflected by transition *t100.premature_ATP_hydrol*. This must not necessarily be a degradation pathway as some of the involved factors (Spp382, Prp43) are also active in recycling spliceosomal components (165, 197) modeled by *t96.Spp382_hPrp43_act*. Summarizing, two possible and

different major outcomes of spliceosome assembly are captured by the model and cause a doubling of observed T-invariants: *i)* the productive (T-invariants i13-i43) and *ii)* the unproductive (T-invariants i44-i71) branch of late spliceosome assembly, which are combined with all subpathways passing through E- and A-complex assembly.

Conservation Relations Corresponding to P-invariants

Compared to the number of T-invariants the present network structure generates four place invariants. A trivial P-invariant exists for the serin protein kinase 1 (SRPK1), which was modeled as a loop connected to the transition that describes the phosphorylation of the splicing factor ASF/SF2. In contrast to other putative enzymes, which act within the spliceosomal subcomplexes, SRPK1 is active at an early stage, activating individual splicing factors. Thus, it is assumed not to pariticipate in further spliceosome assembly and has been modeled as available in a non-limited amount. Two other P-invariants are related to the factors Prp31 and Prp38, which are present in the B-complex. Prp31 has been shown to bind the U4 snRNA and the U4/U6 snRNA duplex in presence of another factor, Snu13. Hence, Prp31 enters spliceosomal assembly at least in the stage of U4/U6 subcomplex formation (246, 189). Furthermore it was shown that Prp31 is destabilized at the time of catalytic activation of the spliceosome. Thus, it was modeled to leave the spliceosomal main complex with the reaction *t.68.B_complex_act*. This results in seven places, describing the tri-snRNP and B-complex formation, which form a P-invariant for the system conservation of Prp31. The fact that Prp31 is required for successive rounds of tri-snRNP and spliceosome formation suggests this protein to be abundantly available, which in turn justifies a conservation relation. Prp31 is furthermore a crucial factor in spliceosome assembly because mutations in its gene are related to the blindness causing disease *retinitis pigmentosa* (189). Since all T-invariants, which involve the reaction of U4/U6 snRNP association (*t47.U4_U6_bdg*, ∼79% of all T-invariants) depend on the presence of Prp31, the model suggests that more than three quarter of the network would fail if this protein would be knocked out.

Prp38 (yeast ortholog of human protein 27K) forms a similar albeit smaller P-invariant of five places, which except for Prp38, is itself a complete subset of the Prp31 P-invariant. The time of appearance and release of Prp38 is less clear but it was shown to associate with higher affinity (and thus stability)

with the assembled U4/U6.U5 tri-snRNP than with an individual U snRNP (247). Hence, it was modeled to enter tri-snRNP formation at the time of U5 snRNP integration. Its involvement in structural rearrangement of the U4/U6 complex without possessing a DExD/H domain to actively participate in the required hydrolyzation reactions, makes it a putative auxiliary factor for the DExD/H box helicase Prp28 (247, 248). The possible connection between Prp38 and the helicase Prp28, which catalyzes the unwinding of the U4/U6 snRNA duplex upon U2 snRNP integration (247), implies that both proteins exit from the active assembly pathway after this step. In this way, Prp38 might as well as Prp31 form a conservation relation at the stage of B-complex formation.

The fourth P-invariant defines the cycling of the DExD/H box helicase Prp16, which is a crucial determinant of the second catalytic step, by catalyzing the initial conformational changes in transition from first to second catalytic step (165). Prp16 binds transiently, being no integral snRNP component, and leaves the spliceosome upon ATP hydrolyzation (249). The Prp16 P-invariant consists only of three places (including the free protein Prp16), denoting the intermediate complexes *p90.U2_5ss_U6_U5_conf1* and *p100.U2_5ss_U6_U5_conf2*, in which Prp16 unfolds its catalytic activity. In contrast to other helicases involved in structural rearrangements (e.g., Prp28), this enzyme occurs not in different branches of the network, also explaining its appearance in a conservation relation.

In general, the appearance of essential enzymes in conservation relations can be meaningful to reflect their availability for subsequent rounds of spliceosome assembly, which may require a constant presence proximal to the location of spliceosome formation. The identified P-invariants suggests that more spliceosomal factors exists whose relative concentration do not change markedly via repeated rounds of spliceosome assembly. Lack of evidence at which time other catalytically active proteins specifically (re)enter the assembly process or how long they remain associated with the main complex, presently limit the modeling of further P-invariants. In contrast to metabolic networks, where commonly only low molecular substances appear in conservation relations, here also enzymes or intermediate complexes can be conserved within a defined signaling network.

3.3.4 Decomposition of the Spliceosomal Network into Functional Units

Analysis of Maximal Common Transition Sets

Maximal common transition sets (MCTS) have been defined as a more generalized concept of enzyme subsets, which define enzymes in a biochemical network, that operate under steady state conditions always together, in one or several metabolic fluxes. Enzyme subsets further require that the enzymes involved are all regulated in the same direction and that their fluxes behave proportional (250). MCTS, in contrast, relax the constraints imposed on mutually occurring sets of reactions to some extent. Due to missing stoichiometric coefficients the constraint of proportionality does not apply. They describe sets of reactions that are in a maximal number of T-invariants („signaling fluxes“) present, hence being shared by different signaling pathways. Given the correctness or biological plausibility of T-invariants, MCTS emphasize key parts of signaling routes and facilitate the description of functional parts of a network. *Vice versa*, they can be indicative of modeling flaws if they combine transitions without proven biological relationship.

Table 3.1 shows the MCTS computed on the set of all 71 T-invariants. There exist six smaller MCTS composed of only two transitions, which nevertheless represent crucial elements of the assembly pathway. MCTS 1 (*t0.U2AF35_3ss_bdg, t1.3ss_in*) describes the recognition of the 3′ss by the factor U2AF35 which occurs in more than 56% of all T-invariants. Next frequently, MCTS 7 is formed by two reactions shared by 32 T-invariants (45%), which describe the influx of the bridging factor FBP11 (*t15.FBP11_in*) and subsequent binding of the U2 snRNP to the branch site *t24.U2_BPS_bdg1*). Taken together, MCTS 4 and MCTS 7 form a module that covers the FBP11 supported interaction of U1 snRNP with the branchpoint bound factor SF1 and the subsequent joining and structural rearrangement of U2 snRNP with replacement of U2AF at the polypyrimidine tract. The remaining small MCTS occur still in more than one quarter of all invariant pathways and cover the NTC-complex and Prp19 integration (MCTS 16), the late SF3b125 action in 17S U2 maturation (MCTS 9), and the slowed ATP hydrolysis by Prp16 with initiation of the discard pathway (MCTS 18).

The largest set of shared transitions (MCTS 2) covers 54/161 (33%) of all transitions, occurring in more than three quarter of all T-invariants. This MCTS

is composed of several building blocks, which define biological functions at essential stages of spliceosome assembly, for example, branchpoint definition, 15S U2 snRNP assembly, SF3a and SF3b subcomplex formation, U5- and U6 snRNP maturation and the cyclophilin trimer formation. The energy supply by ATP and removal of ADP is part of MCTS 2, which naturally has to be shared by many T-invariants as each stage requires ATP either for phosphorylation or hydrolyzation reactions. The maturation and recycling of the U1 and U4 snRNP is described by individual MCTS (11 and 13), the former consisting of reactions, which are shared by almost three quarter (73%) of all T-invariants. All MCTS further validate the model network, capturing crucial parts of the U1 and U2 snRNP maturation, U1 independent $5'$ ss recognition and the Prp16 involved discard pathway in spliceosome assembly. A more detailed PN model of U1 snRNP assembly was recently published (226).

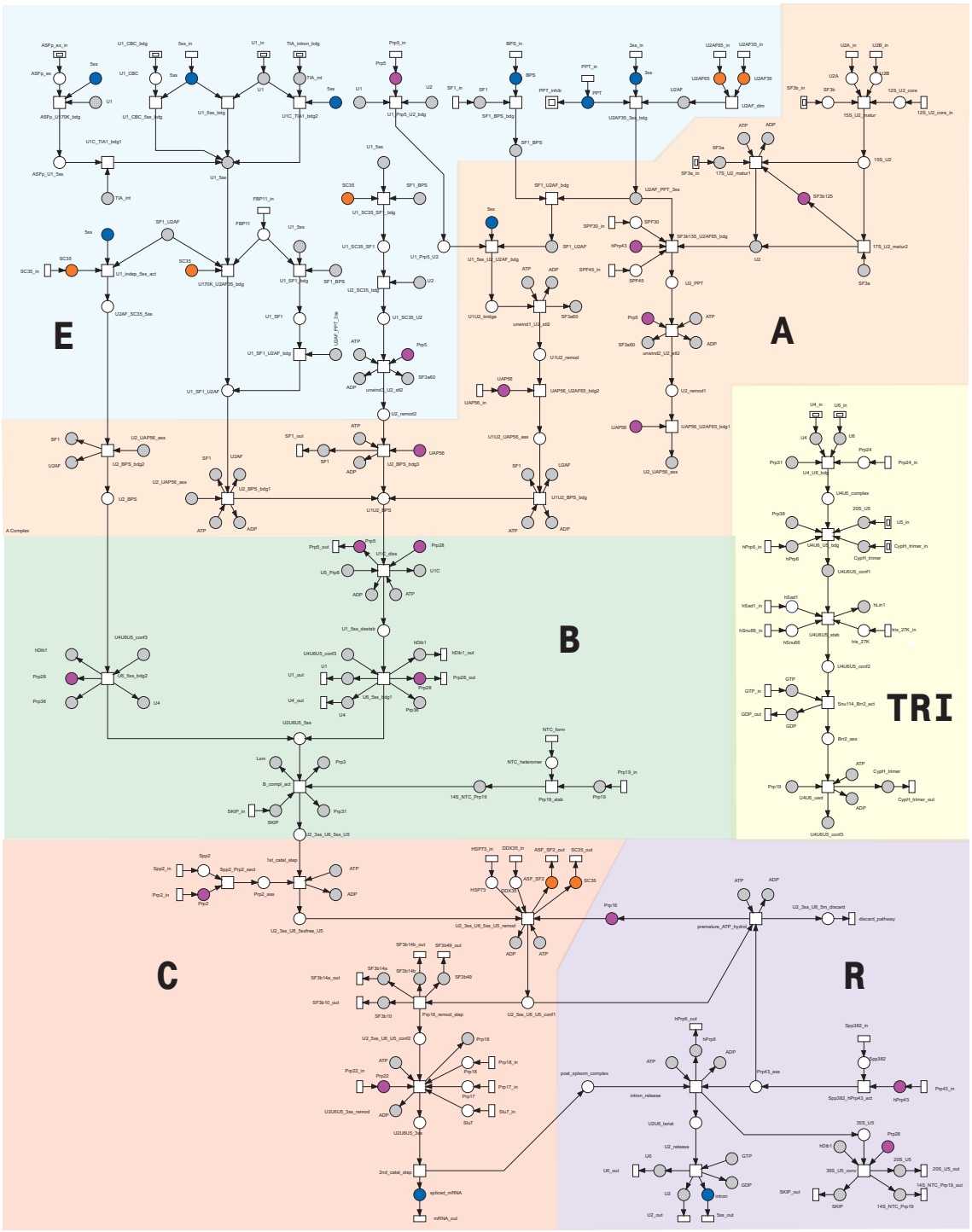


Figure 3.10: The complete network of spliceosome assembly modeled as transition bounded Petri net. The different stages are labeled with capital letters: E = E-complex, A = A-complex, B = B-complex, TRI = tri-snRNP complex, C = C-complex and R = recycling pathways. Places are colored according to different functions: orange = proteins with RS domains; blue = RNA; magenta = DExD/H box proteins functioning as ATP dependent „RNA unwindases“; gray or hatched = logical places, indicating equal places occurring more than once in the figure, but not in the underlying graph structure of the PN. Transitions represented by two squares introduce hierarchical nodes, which connect to further reactions at a lower network level (see Figure 3.12)

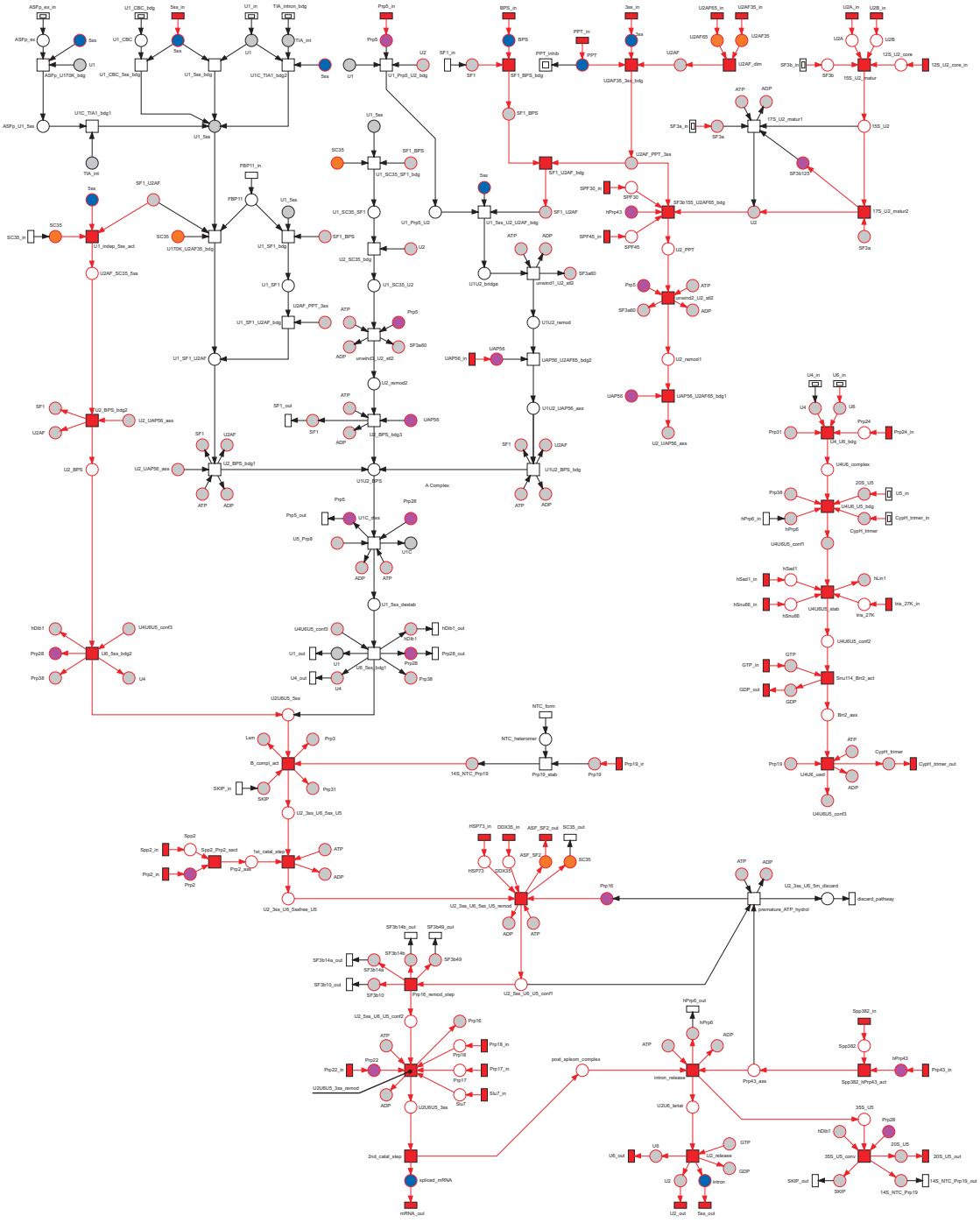


Figure 3.11: Network of spliceosome assembly. The red T-invariant transitions belong to a possible scenario of E-complex assembly, which is part of different concurring redundant subpathways. Here, the 5′ splice site is recognized via an U1 snRNP independent pathway

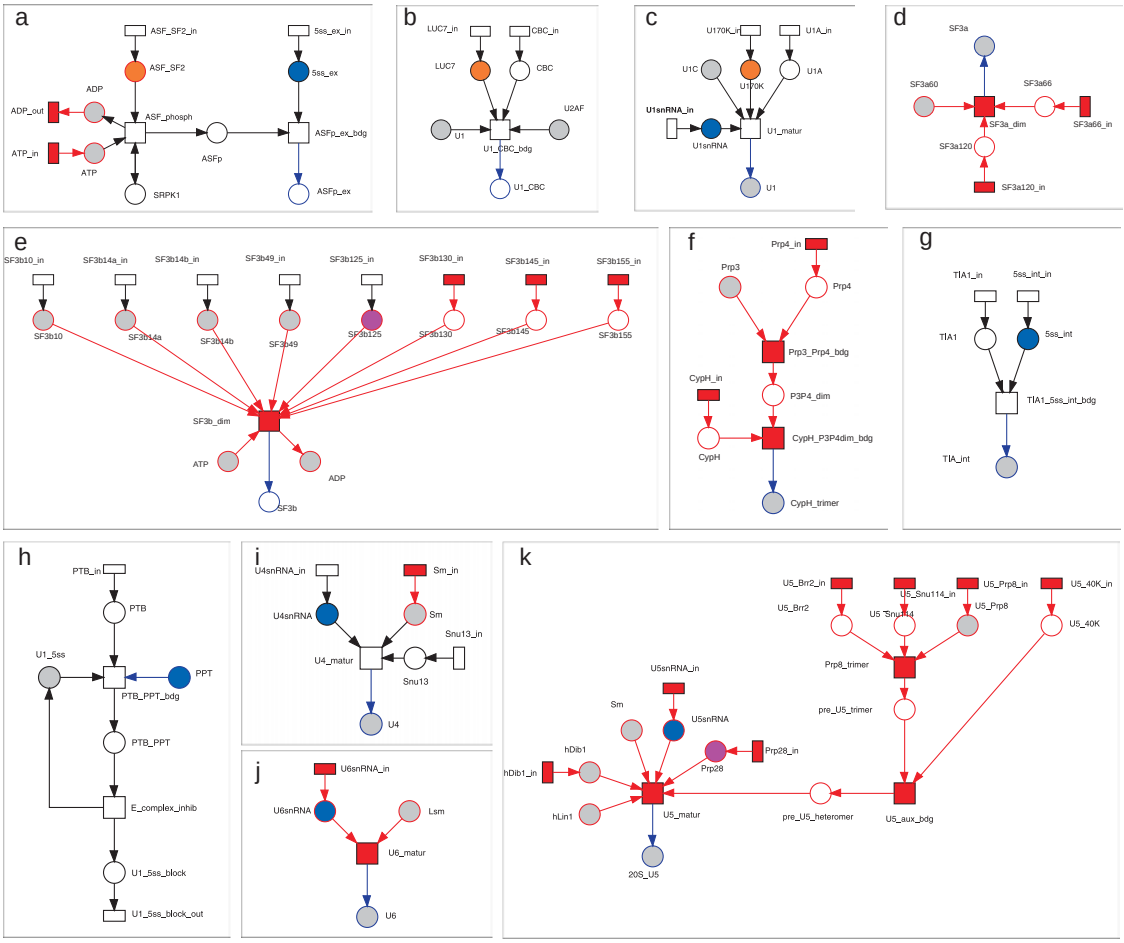


Figure 3.12: (a-k) Subhierarchy levels of the spliceosomal assembly network. Red highlighted transitions showing exemplarily T-invariant 13, which describes one possible scenario of E-complex assembly as shown in Figure 3.11

Table 3.1: Maximal common transition sets (MCTS) as determined from the 71 T-invariants that cover the network. Each MCTS comprises reactions that are exclusively shared by several T-invariants and hence describe frequently used routes through the spliceosomal assembly network.

MCTS	**Transitions**			**T-Invariants**			
ID	**IDs**	**#**	**%**	**IDs**	**#**	**%**	**Biological Interpretation**
1	t0, t1	2	1.24	i13-i18, i23-i26, i29-i32, i35-i40, i44-i47, i50-i53, i56-i59, i64-i71	40	56.34	Recognition of 3′ss by U2AF35
2	t3, t7, t11, t18-t21, t26, t42, t44-t54, t68-t72, t86, t87, t93, t96-t98, t104-t106, t110, t111, t113, t114, t118-t129, t144, t145, t150, t151	54	33.54	i13-i40, i44-i71	56	78.87	Branch point definition, 15S U2 snRNP assembly, SF3a formation, partial SF3b formation, U5- and U6-maturation, CypH trimer formation, ATP-in- ADP-efflux
3	t4-t6	3	1.86	i13-i18, i21-i26, i29-i32, i35-i40, i44-i47, i50-i53, i56-i61, i64-i71	44	61.97	U2AF dimerization
4	t8, t23, t57, t158, t159	5	3.11	i13-i18, i23-i26, i29-i32, i35-i38, i44-i47, i50-i53, i56-i59, i64-i69	36	50.70	Reactions of U2 snRNP remodeling variant 1, involving hPrp43 and UAP56
5	t9, t12, t141-t143	5	3.11	i15-i20, i62-i67	12	16.90	ASF phosphorylation and 5′ss definition via ASF/U1 snRNP/TIA1 interactions
6	t10, t137-t139	4	2.48	i21-i26, i56-i61	12	16.90	5′ Terminal 5′ss definition via U1 snRNP interactions with cap binding complex
7	t15, t24	2	1.24	i15-i18, i23-i26, i29-i32, i35-i38, i44-i47, i50-i53, i56-i59, i64-i67	32	45.07	FBP11 dependent U2 snRNP binding to the branch point
8	t16, t31, t32	3	1.86	i13, i14, i68, i69	4	5.63	U1 snRNP independent 5′ss definition and A complex formation
9	t22, t109	2	1.24	i14, i16, i18, i20, i22, i24, i26, i28, i30, i32, i34, i36, i38, i40, i45, i47, i49, i51, i53, i55, i57, i59, i61, i63, i65, i67, i69, i71	28	39.44	SF3b125 dependent 17S U2 snRNP maturation
10	t25, t27-t29, t55	5	3.11	i39, i40, i70, i71	4	5.63	U1/U2 snRNP bridging by Prp5 and simultaneous binding to 5′ss and branch point
11	t33, t34, t133-t136	6	3.73	i15-i40, i44-i67, i70, i71	52	73.24	U1 snRNP maturation and recycling
12	t35, t56, t152, t155	4	2.48	i19-i22, i27, i28, i33, i34, i48, i49, i54, i55, i60-i63	16	22.54	U2AF independent A complex formation

continued next page

Table 3.1: Maximal common transition sets (MCTS) as determined from the 71 T-invariants that cover the network. Each MCTS comprises reactions that are exclusively shared by several T-invariants and hence describe frequently used routes through the spliceosomal assembly network.

MCTS	**Transitions**			**T-Invariants**			
ID	**IDs**	**#**	**%**	**IDs**	**#**	**%**	**Biological Interpretation**
13	t36, t146, t147, t149	4	2.48	i7	1	1.41	U4 snRNP maturation and recycling
14	t58, t59	2	1.24	i15, i16, i25, i26, i31, i32, i37, i38, i44, i45, i50, i51, i56, i57, i66, i67	16	22.54	FBP11 supported U1 snRNP/SF1 binding and subsequent interaction with U2AF bound PPT-3′ss
15	t60-t63	4	2.48	i3	1	1.41	PTB inhibitory pathway (without *t2.PPT_in*)
16	t65, t66	2	1.24	i41, i44-i71	29	40.85	NTC-complex formation and stable Prp19 integration
17	t76, t77, t81, t82, t84, t85, t88-t92, t94, t95, t101, t160	15	9.32	i13-i40	28	39.44	Prp16 dependent remodeling of C-complex, 2nd catalytic step of splicing, release of ligated exons and disassembly of postspliceosomal complex
18	t99, t100	2	1.24	i44-i71	28	39.44	Prp16 induced and slowed ATP hydrolysis and commitment of C-complex to discard pathway, disassembly supported by hPrp43
19	t130-t132	3	1.86	i15-i20, i27-i32, i50-i55, i62-i67	24	33.80	TIA1 intron binding
$\sum$	-	127	78.86	-	-	-	-

Clustering of T-invariants and MCTS

The computed T-invariants (*cf.* Table C.3) were aligned to determine the individual distance between the signaling pathways. Similar to sequence alignments the multiple comparisons among all T-invariants can be used to built a distance matrix based on which a clustering can be performed. Clusters reflect groups of signaling pathways, which share a given percentage of reactions. Here, a threshold of 80% was chosen to merge T-invariants with less than 20% difference into the same subtree. For example, the T-invariants i13 and i14 show a difference in four reactions in a total pathway length of 92 reactions, i.e. the transition *t13.17S_U2_matur2* is missed in i14 and *t12.ASFp_U170K_bdg, t22.17S_U2_matur1, t109.SF3b125_in* are missed in i13, which makes both invariant to 96% similar (*cf.* Equation 3.9). Comparing different subclusters helps to identify those reactions, which distinguish the T-invariants and which reflect different functions in different stages of spliceosome assembly.

The cluster representation depicts all trivial T-invariants in one group in the lower part of the tree (Figure 3.13 C1-C13, C22, C23), which is reasonable since they share maximal two transitions with the remaining T-invariants (cluster I). Also three short T-invariants, which describe the NTC-complex formation, the subpathway of U4 snRNP maturation and the PTB inhibition pathway group separately, reflecting subpathways, which are not shared by other signaling fluxes. In contrast to the out-group, cluster I combines all T-invariants of at least four reactions.
Subclusters can contain complete or partial MCTS. For example, cluster C17 and C18 together constitute MCTS 10, which is composed of five reactions, describing the subpathway of bridging the U1/U2 snRNP by Prp5 and occurs in four T-invariants. In contrast, T-invariants i15-i20 and i62-i67 share also five reactions involving ASF/SF2 within MCTS 5 but are part of the two different major clusters I and IV. These clusters partition the T-invariants in two sets of reactions: one is reaching the productive end of the spliceosomal assembly pathway (resulting in spliced mRNA) and the other one is representing the discard pathway during C-Complex stage.
This splitting can also be seen by visualizing all invariant pathways and their shared reactions via a color map (*cf.* Figure 3.14). The color map representation was used to aid and accelerate the visual identification of groups of reactions that participate in different T-invariants in conjunction with the dendrogram (Figure

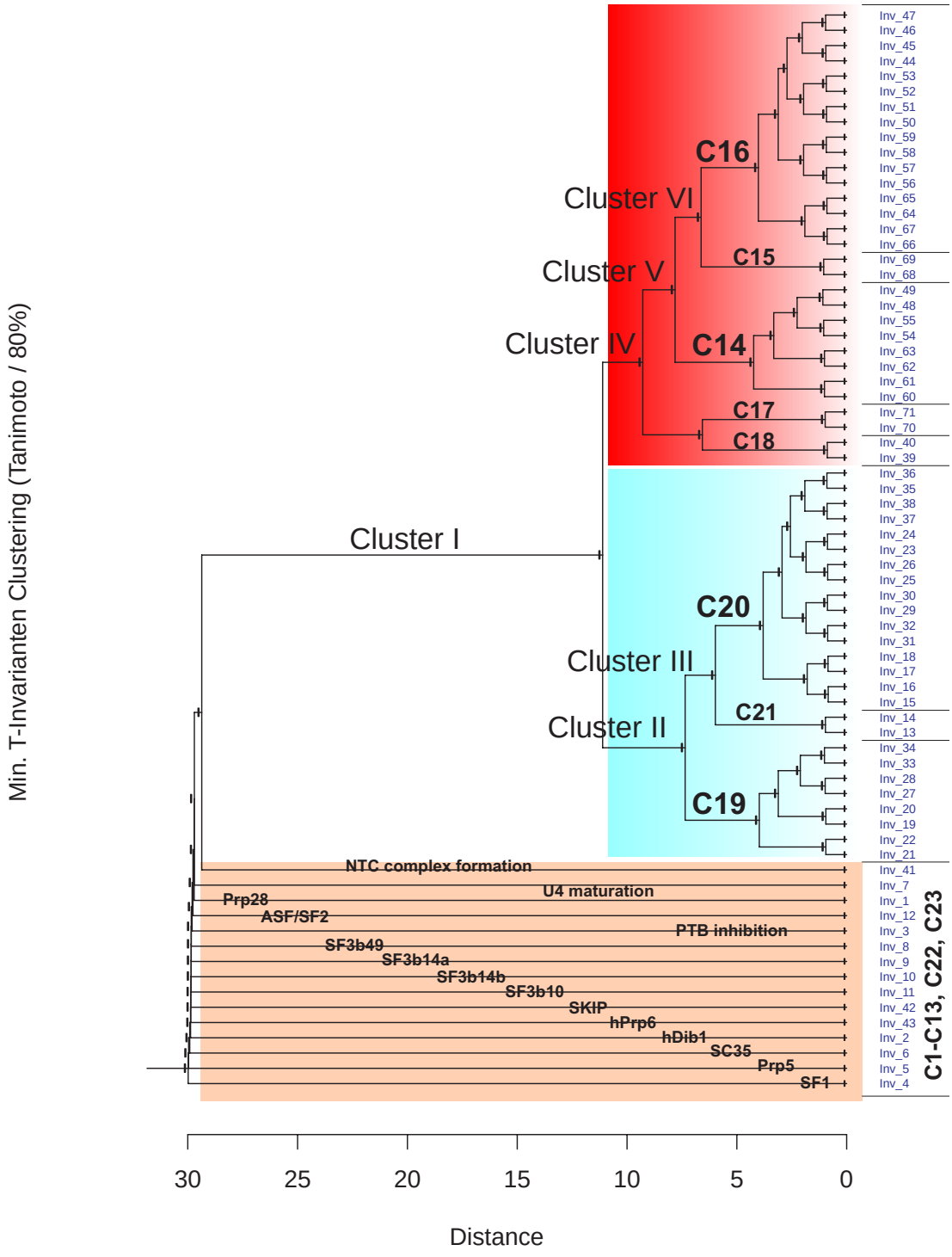


Figure 3.13: The clustering of T-invariants, whereas trivial and small T-invariants are shown in the bottom of the tree. The two main clusters are colored in blue and red. Latin enumeration of clusters refers to the output as obtained from PInA using the Tanimoto distance measure and 80% similarity cutoff. Roman enumeration labels clusters that form larger groups

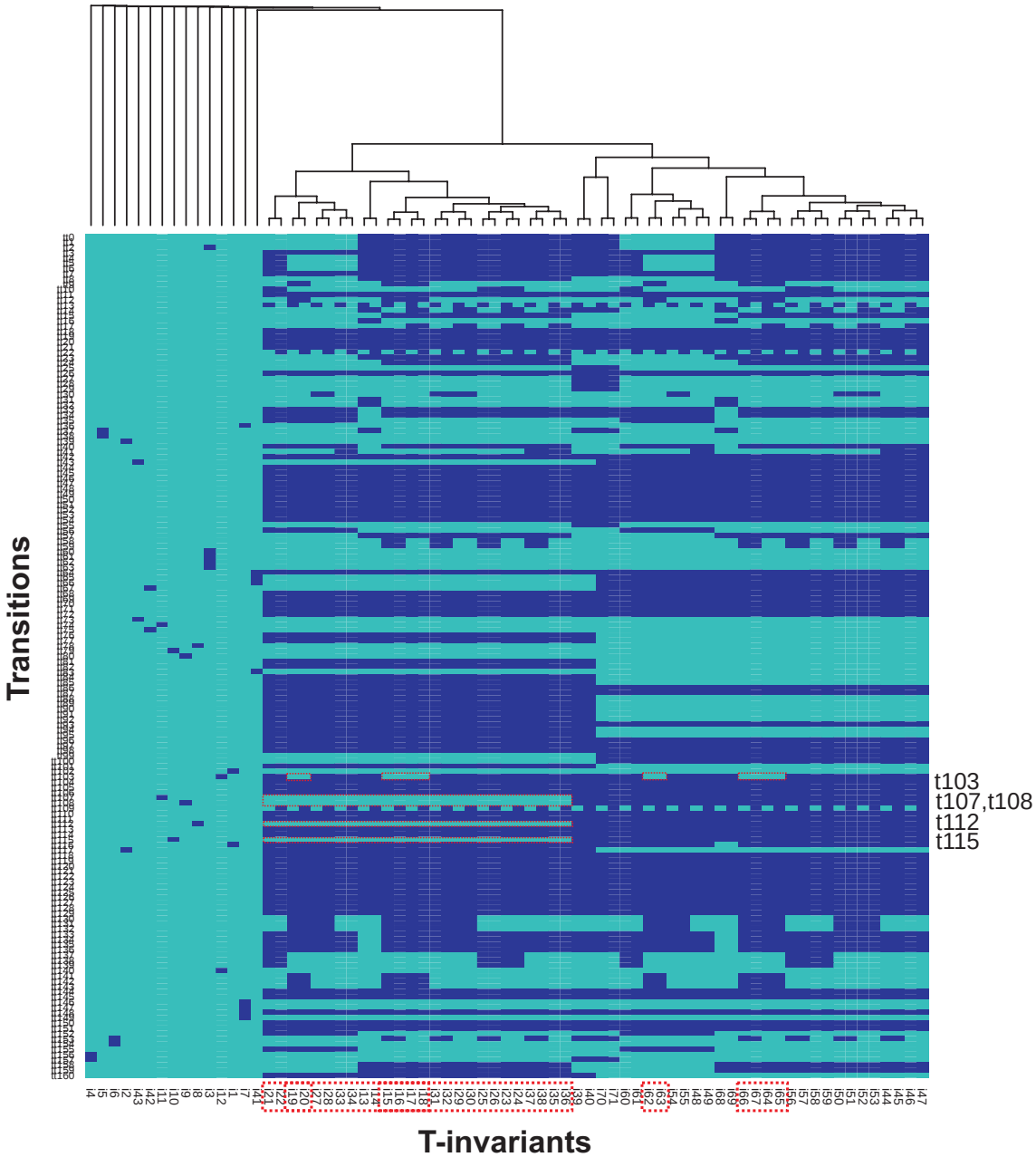


Figure 3.14: Color map (see methods, page 84) to visualize, which transitions occur within similar T-invariants through the network of spliceosome assembly. The similarity can be compared via the dendrogram at top off the figure, which is the same as shown in Figure 3.13. Dark or light blue colors indicate presence or absence of one or several reactions respectively, among the total set of modeled reactions within one or several T-invariants. The red dashed rectangles show an example of groups of T-invariants, which specifically lack the reactions indicated at the right side. It is notable that these groups of T-invariant appear in different clusters and their common features is easier detectable via the color map compared to the dendrogram.

3.13).

It is thought as proposal to introduce a compact representation of the network structure to facilitate the interpretation of differences between signaling pathways. In this case, darker colors symbolize the presence of an reaction within a T-invariant in vertical direction, but participation in several T-invariants in horizontal direction. Bright colors mean that a reaction is not present in T-invariants and consequently also not in MCTS. This representation suggests to extend the analysis of MCTS also for sets of reactions, which form excluded subsets specifically missing in certain MCTS. For example, in Figure 3.14 one can easily recognize two brighter colored horizontal areas within the line of transition *t103.ASF_SF2_out*, which stretches exactly over the columns that contain the T-invariants of MCTS 5 (*cf.* dashed rectangle in Figure 3.14). This means, transition *t103ASF_SF2_out* is specifically *not* involved in the T-invariants i15-i20 and i62-i67.

Further analysis of line *t103.ASF_SF2_out* within the color map reveals that it is present in almost all other T-invariants, as expectedly in i12, which poses the trivial influx and efflux of ASF/SF2. For example, two closely related T-invariants, using this reactions are i13 and i14. They describe the U1 snRNP independent 5′ss definition, which again can occur in the two variants generated by the alternative reactions *22.17S_U2_matur1* (i14) and *t13.17S_U2_matur2* (i13). Interestingly, the T-invariants i13 and i14 do not contain the ASF/SF2 influx reaction (*140.ASF_SF2_in*), in contrast to its efflux reaction (*103.ASF_SF2_out*), which is only missing in MCTS 5. This suggests that the pathways involved in MCTS 5 describe a state in which the spliceosome has entered a repetition cycle, where many factors are already present for recurring assembly. Thus, it is conceivable that U1 snRNP is initially available, but dispensable in subsequent rounds of assembly when a critical amount of SR proteins (e.g., ASF/SF2) are present to enable U1 snRNP independent 5′ss definition. It was previously shown that SR proteins are perfectly capable to organize the cross talk between 5′ss and branchpoint/3′ss by interacting via their SR domains (20), which is reflected in the PN model.

Another observation that could be made from the color map concerns the absence of four reactions, *t107.SF3b10_in, t108.SF3b14a_in, t112.SF3b49_in* and *t115.SF3b14b_in_in*, from a large body of T-invariants (i13-i38), independent of

the trivial T-invariants. These reactions model the influx of four SF3b specific factors. They were modeled as logical places, each with a specified efflux reaction prior to the C-complex stage, because experiments failed to detect these factors within spliceosomal C-complexes (151) (supplementary material). The T-invariants i13-i38 describe the signaling pathways that reach the final stage of spliceosome assembly. Hence, these routes pass the stage, where the SF3b factors are leaving the active assembly process. Among others, the specified SF3b factors are required for subsequent rounds of spliceosome assembly, thus the absence of their influx reactions from T-invariants i13-i38 can be interpreted as a way to remain within range of the spliceosome assembly site. A different scenario occurs for the T-invariants that enter the discard pathway, which is triggered before C-complex formation. Here, no explicit efflux reactions could be adapted from literature for the SF3b factors, hence their influx reaction is part of the T-invariants i44-i71 (*cf.* Figure 3.14). This raises the question, what happens with these factors and when if the discard pathway is activated during spliceosome assembly.

3.4 Discussion

The present work describes a PN model, which combines different scenarios of spliceosome assembly over the basic assembly stages of this multi protein complex. The spliceosome is a nuclear macrocomplex, which is newly built after or as soon as an precursor RNA emerges from the Polymerase II transcription complex. The assembly process involves biochemical reactions, which can be distinguished in enzymatic and association reactions. Unlike in metabolic networks, which commonly model the conversion of low molecular compounds to produce energy or target metabolites, e.g. aminoacids (220, 251), the spliceosome assembly involves many enzymatic reactions, which act on double stranded RNAs as substrate and proteins or NTPs as co-factors (165, 198). This is due to the snRNA containing core components of the spliceosome, which interact via multiple RNA-RNA contacts making it necessary to re-open intermediate conformations at several stages during the assembly process. Additionally, phosphorylation reactions as known from signal transduction networks (252) assure the specificity and localization of splicing factors. Consequently, the model presented here, consists of several types of molecules (RNA, proteins and compounds of both) and reactions, which have been designed and tested individually (*cf.* Chapter 3.3.1) prior to the setup of

the complete network.
Hundreds of individual studies have investigated components of the spliceosome or individual biochemical reactions. The knowledge of almost two decades lab work is available but needs to be translated into a machine readable and human comprehensible language. Thus, one of the main purposes of this work is to channel biochemical knowledge about the spliceosome into a formalized description, suitable for computational analysis. Among the major difficulties is the handling of non-standardized identifiers of the involved proteins, which exacerbates the combination of smaller models, initially devised from indivial reports and successively combined into a larger network of interactions. Thus, the power of predictive modeling will increase as more submodels become integrated, covering more details of the spliceosome assembly pathway.

A network of ordered basic interactions was established, leading to the assembly of an active spliceosome and including also the example of a discard pathway, which was previously suggested (197). In total, about 100 proteins where integrated into the model. Many proteins, participating in the spliceosomal assembly pathway, are themselves alternatively spliced hence may occur in several isoforms. For example, the U2 snRNP specific component SF3b14 shows five predicted alternative splice forms of different types (source ASD (53)). The interactions of SF3b14 are well described (180), including the location of functional domains within the protein sequence. Since increasingly more spliceosomal factors are described in such detail, it should be possible in the near future to estimate the impact of alternative splicing on the spliceosome, which poses an interesting example of combinatorial complexity. Suppose that in four stages of spliceosome assembly occurs only one protein factor in two functional different isoforms, about $2^4 = 16$ different spliceosomes could be assembled and contribute to different alternative splicing decisions (neglecting that some alternative splice forms do not reach the protein level). This is a rough estimate of the lower boundary as many more spliceosomal proteins exist, and most of their genes are likely to be alternatively spliced. Concerning the presented Petri Net modeling approach, one can summarize the following achievements:

i. Translation of different lines of evidence for modular subsystems of the spliceosomal assembly pathway from experimental literature into one unique mathematical formalism.

ii. Compilation of minimal positive T-Invariants (P-invariants) based on the commonly applied steady state assumption for biochemical networks.

iii. Model validation resulting in a network completely covered with T-invariants.

iv. Representation of combined partial pathways, each supported by experimental reports, allowing for model expansion and testing of new hypotheses.

v. Inclusion of special aspects of $5'$ splice site recognition during E-complex formation as well as the potential activation of a discard pathway as simplified model for a kinetic proof reading mechanism during C-complex formation.

vi. Easier identification of discrepancies in current experimental data by the combinatorial arrangement of the subpathways. For example, the activation of UAP56 by U2AF stands in contradiction to the apparent requirement of UAP56 for transition from E- to A-complex within the U2AF independent A-complex assembly pathway.

vii. Comprehensive and condensed visualization of the spliceosomal assembly process, allowing the global inspection of similar and distinct routes. This facilitates the apprehension of a large network such as the spliceosomal assembly pathway and its further extension.

The clustering of the T-invariants representing signaling pathways and participating in spliceosome formation, indicates that there exists a variety of similar pathways leading to the same intermediate complexes. Although each T-invariant, describing one of these routes, is minimal in that it would fail with the loss of one reaction, it is clearly visible that there exists a redundancy in routes leading to the formation of intermediate states. This observation provides the interesting aspect of a backup failure mechanism, ensuring that independently from alternating conditions the spliceosome reaches critical assembly checkpoints with different protein complements. Alternatively, this might extend our view on different spliceosomes in dependence on different cellular or environmental conditions. The existence of a major and a minor spliceosome with different mRNA substrate specificities could support this notion, but the major spliceosome as modeled here suggests for itself a highly dynamic assembly process. The flood of

different mechanistic examples of individual and sometimes quite different intermediate steps makes it clear that there is no *one* spliceosome. Hence, there can be no single model of spliceosome assembly.
Although the current model represents a higher coverage of experimentally supported subsystems in the early (E- and A-complex) in comparison to later assembly stages, it is tempting to hypothesize that the number of different routes increases with the importance of the intermediate complex for the overall assembly process.

Several requirements to future works on spliceosome analysis can be asserted from this work, addressing both, experimental biologists and computational scientists. First, experimental data should immediately be stored in a structured pre-formatted way, making use of existing formalisms and avoiding unnecessary naming morphisms. Experimental data provides precious facts which are necessary to prepare subsequent *in silico* analyses. Second, theoretical and computational contributions can still be improved in the supply of data collection tools as well as integrated pipelines for their global evaluation and analysis. For example, text mining tools at the level of network design and statistical measures at the level of substructure analysis (e.g., T-invariants, MCTS) can enhance the output of this modeling approach.

Finally, one needs to keep in mind, that structural properties depend at first hand on the knowledge put into the model. In light of the wealth of biological information, a direct consequence is the possibility that parts of the model are better covered by data than others, and therefore exhibiting a higher complexity. Consequently, these parts are stronger represented by T-invariants. Nevertheless, the fact to observe a stronger representation of individual aspects of a biological network justifies a model, because it captures relations and trends that are hard to detect using detailed mechanistic studies, which moreover are to numerous for individual analyses.

Conclusions

The present dissertation is structured into three parts, which are connected by the overall topic of *Alternative Splicing* and its computational analysis. The first and second chapter concentrate on sequence based approaches, aiming on how transcripts, which are nowadays abundantly present in public databases, can be used to infer knowledge on the regulation of alternatively splicing. The third chapter adds a systems biological approach by modeling the spliceosome, the biological component, which enables eukaryotic organisms to perform the splicing reaction.

Chapter 1 The outcome of initial preparative sequence processing usually results in repositories of alternatively spliced transcripts such as the EASED database (253). In consequence, a variety of databases storing alternative splice forms exists but although most of them are based on transcript and genome sequences of the same primary databases, the alignment algorithms and filter criteria applied in prediction pipelines vary considerably. Hence, it is difficult to compare alternative splice events from different databases. In essence each dataset should be characterized before being subjected to further analyses. Following this idea, the first part of this work aimed on the investigation of features, which affirm the EASED dataset of splice forms, justifying to pursue new questions based on these data. To this end the dataset was partitioned into sets of splice forms with reference (rss) and alternative splice sites (ass). The distribution of donor- and acceptor signal strength was compared between both sets and between both types of splicing signals. From these comparisons it can be concluded that:

- Donor- and acceptor strengths differ significantly between alternative and reference splice sites, both signals showing a trend to weaker signals in the set of alternative splice sites. However, there exists an overlap in the score

distributions, were reference and alternative splice forms show an equal splice site strength.

- The difference of the average splice site strength ($\Delta S = |S_{rss} - S_{ass}|$) between alternative and reference splice sites is lower at acceptor compared to donor sites.

- Donor splicing signals are weaker than acceptor splicing signals irrespective of the classification into alternative or reference splice sites, being in agreement with the fact that initial donor site selection often requires additional determinants to achieve the necessary level of specificity.

Additionally, the transcript abundance representing alternative and reference splice sites was compared. Here, the EASED dataset showed the general trend that alternative splice sites (pairs of donor- and acceptor sites) are less abundantly supported by transcripts than the reference splice sites. This finding was complemented by the observation that the set of more frequently used splice sites (as derived from the transcript abundance) tend to pose a stronger splicing signal than splice sites of low utilization.
The lower signal strength of alternative compared to constitutive splice sites has been previously reported for other datasets of splice events (254). This agreement was taken to conclude that the splice site strength of alternative splice forms of the EASED database can be sufficiently discriminated from reference splice site. A follow-up analysis on additional splicing regulating motifs was conducted. The weaker signal or information content of alternative splice sites raises the question how trans-acting factors (e.g., the spliceosome) distinguish between these splice sites. As one possible explanation the action of regulatory proteins was proposed (6), which bind to motifs near alternative splice sites and serve as guiding elements for example by interactions between RS domains. Hence, the presence of binding motifs of four well characterized splicing regulating proteins (ASF/SF2, SC35, SRp40 and SRp55) was analyzed and compared between alternative and reference splice sites and between exon and intron splice site context. This led to the following conclusions:

- The four investigated splicing factor binding motifs occur among *ass* and *rss* individually more frequently in the intronic donor than in the intronic acceptor context, but with about the same frequency in the exonic donor region. This indicates that the conservation of these motifs is reduced in

the intron compared to the exon region near donor sites, and even more reduced in the intron region near acceptor sites. Such constraints can be imposed by the coding region, polypyrimidine tract and branchpoint, which are determinants of splice site selection in both, *ass* and *rss*.

- While the investigated motif frequencies are about equal between *ass* and *rss* in the exonic contexts of all four binding motif, the intronic regions show more binding motifs near alternative than near reference splice sites, indicating a higher conservation pressure for splicing factor binding sites for the alternative splice forms.

- For all four SR proteins it is found that although there are less binding motifs in the intron context compared to the exon context near splice sites, the intron context of *rss* is indicative of less binding motifs than the intron context in *ass*.

- The polarization (difference in binding motif frequency between exon and intron context) is for ASF/SF2 stronger in *ass* than in *rss* around both, donor and acceptor sites. SC35 shows almost no polarization and SRp40 and SRp55 only at donor sites, with more binding sites in the exonic than intronic donor context. It can be concluded that although ASF/SF2 has been found to participate in splicing of non-alternative exons, the frequency of its binding motifs is discriminative for alternative and reference splice sites.

Summarizing, there are qualitative differences in the distribution of binding motifs for auxiliary splicing factors, depending on the splice site proximal location within one and the same class (*ass* or *rss*). Additionally, also differences between *ass* and *rss* exists, which stand in line with the compensatory effect that splicing factors convey to the selection of weak alternative splice sites. Although binding motifs for splicing factors are thought to be degenerated in composition, their presence provide an effective means for the cell to regulate splice site selection via concentration gradients of these factors. It is tempting to speculate that an additional layer of specificity in modulating splice site selection may be achieved by combinations of binding motifs, similar as found for transcription factors (255). This analysis did not distinguish between different types of *ass* (an alternative site can either be an alternative acceptor or donor site), however, since eukaryotic exons are relatively short (in average around 100-140nt) the definition of splice

sites occurs predominantly across exons (known as exon definition model (85)). In consequence, the binding of regulatory factors on either site may have an effect on the alternative splicing of both splice sites.

Chapter 2 The EASED database does not resolve different alternative splicing patterns in more detail, for example, exon skipping or specific A5E events. Additionally, the filter parameter of the EASED prediction pipeline was restricted to alignment variations of at least six nucleotides, a limitation of which also other databases suffer. In consequence, minor splice site variations were commonly considered as noise of the alignment algorithm in the past. In 2004, the first reports affirmed the frequent occurrence of subtle variations at human acceptor splice sites and experimentally proved their plausibility as alternative splice forms (2). The second part of this thesis connects to this development by focusing on a special kind of subtle splice events at donor splice sites. Alterations at donor sites were already early shown to pose a crucial element for splice site selection (86). The complementarity of the 5′ss to the 3′ end of the U1 snRNA introduces a naturally favored second donor signal and raises the question whether alternative splicing at this tandem splicing signal leads to a regulatory effect in gene expression. Using the dataset of the HOLLYWOOD database (3), at first the prevalence of alternative donor splice events with focus on single exons was determined, This led to the following conclusions:

- An initial dataset of 5,275 exons, indicating alternative splicing at their downstream flanking donor site, shows a dominant fraction (17 - 38%) of subtle variations (e.g., < 6 bp), which is higher than previous estimates of alignment error rates stated (256).

- Samples indicate that the initial annotation method (SIM4) produces false positive splice forms (see Appendix B.1(a)), which are corrected by a more recent method using dynamic programming in conjunction with a splice site model (EXALIN). This advises to reckon with a higher false positive rate in predicting subtle splice events without additional constraints in the transcript-genome alignment strategy.

- Independent from the prediction method the most frequent alternative donor splice pattern differs by 4 nucleotides (9 - 28% of all A5E), and thus, contrast the most frequent variation (3 nt) observed at acceptors sites.

- The abundance of subtle splice events (< 6 nt), including the most frequently occurring Δ4 isoforms could also be observed in the comparative species *Mus musculus*

The amount of A5EΔ4 splicing is in agreement with the frequency reported previously in a smaller dataset (7.5%), which however, was based on alternative splice events conserved in human and mouse (14). Trusting such rigor, all further analyses and results are based on the most stringent dataset refined by the EXALIN alignment method, pointing to a somewhat higher occurrence of 9% Δ4 isoforms. Based on the transcript evidence, supporting each of two alternative splice events, an obvious difference between donor and acceptor sites is found. While A3E show no clear separation into classes, where either one of the distal or proximal splice sites is preferred, such a dichotomy is observed for the A5E including the Δ4 splice variants. Thus the transcript support suggests that in general two types of alternative donor splice events are distinguishable: a proximal major event (type-I, A5E PΔ4) and a distal major event (type-II, A5E DΔ4). Further, the transcript evidence indicates a higher occurrence of type-II than type-I A5EΔ4 events, a polarization, which appears exactly opposite within the remaining A5E events. This situation suggests a formal distinction between non-overlapping and overlapping (tandem) donor sites, which in similar way was found by other reports (257, 121).
Following this classification scheme, A5EΔ4 splice forms are characterized, leading to the following conclusions:

- The minor form of both A5EΔ4 types exhibits in general a weaker splice site score compared to the major form, however, type-I minor donor splice sites are still closer to the U1 snRNA consensus motif than type-II minor sites, implicating a higher potential for selection of the minor upstream donor (dΔ4) in type-I A5EΔ4 isoforms.

- Pseudo splice sites distanced by four nucleotides from authentic donor splice sites in constitutive exons are more reliably distinguished than minor and major sites of tandem donors in both, type-I and II classification (estimated accuracy of ≥95%). However, type-II A5EΔ4 events reach almost the same accuracy (92%) in distinguishing the minor from the major donor splice site, indicating a risk to enrich also false positive events (pseudo splice sites) in this class.

- Both classes showed a different scheme in position specific nucleotide conservation, which may influence the binding selectivity by snRNAs of the spliceosomal complexes U1, U5 and U6. With respect to the weaker tandem donor sites, type-I A5EΔ4 events show positions conserved, which are crucial for U1 and U6 snRNA binding, while type-II A5EΔ4 events show less conservation and in different positions. The finding of individual splice site nucleotides complement the results determined in the splice site score analysis.

- The presence of cis-elements for binding of splicing silencing proteins, differed between both classes, with a tendency to more silencer elements within the downstream intron of type-I A5EΔ4 events and less silencer elements of type-II A5EΔ4 events compared to introns downstream of constitutive donors. This affirms a functional role for the distal minor donor, while it appears less likely that the distal major donor is silenced to improve the proximal minor donor.

- The comparison of orthologous tandem donor regions indicate that both classes do not reach conservation levels in the flanking intronic region, as observed for skipped exons. However, type-I AS events showed a better intron conservation in the corresponding mouse genes, supporting the presence of evolutionary conserved regulatory elements. Additionally, examples are found, where not only the splice site motif was conserved in the orthologous mouse gene but also both donors are confirmed by transcript data, pointing to regulatory importance of this splicing modus.

- A5EΔ4 splice events are in more than 90% of all cases located within the coding region and, thus, cause a shift in the open reading frame. Type-II AS events introduced in nearly 70% a premature termination codon (PTC) when spliced at the minor donor, while splicing of only one quarter of type-I minor donors caused a PTC. Most of these PTC are located at a distance from the alternative translation stop signal to qualify as triggers of the nonsense mediated decay (NMD) pathway. This suggests especially type-II alternative splicing to be a mechanism of down-regulating protein isoforms within the cell, as contrasted by type-I isoforms, which more frequently may produce truncated proteins. Moreover, type-I splice events occurred significantly more frequently in genes, which encode RNA binding proteins, suggesting a regulatory function at the RNA level of gene expression.

Some questions have been posed by this analysis and remain open for further analysis. The set of tandem donors, showing the highest conservation compared to mouse (for example SFRS16 exon 15, MSF2 exon 3, CASD1 exon 14 or BRSK1 exon 3) can be investigated on a broader range of species to infer knowledge about their evolutional background. Although small, the subset of A5EΔ4 tandem donors with high conserved flanking introns may serve to isolate, in more detail, sequence regions with impact on the U1 or U6 binding behavior. Another extension of this work can investigate to what extent A5EΔ4 events and other frame shifting AS events occur mutually and compensate each other to avoid the NMD pathway. The problem here is that EST inferred AS events are not well suited for such an *in silico* analysis as it has to be assured that two or more AS events derive from the same mRNA. The analysis of conserved positions within tandem donor sites raised the question, to what extend, secondary structures of precursor messenger RNAs contribute to the specific selection of overlapping alternative splice sites. For example, sequence parts downstream of the tandem donor and complementary to the distal donor, may, under specific cellular conditions occupy the distal donor within a stemloop structure, such, that the proximal donor locates in an open loop, and hence, remain accessible for spliceosomal components. This would suggest a model, which controls the regulation of overlapping donor sites by two different layers: *i)* the increased specificity due to the larger sequence context involved in stemloop formation around the splice site and *ii)* the presence of proteins, which control the structural conformation of the pre-mRNA molecule.

It also remains to be investigated to what extent mutations in splice sites result in the creation of A5EΔ4 tandem donors or shift splicing from the proximal to the distal donor and *vice versa*. Aberrant donor splice sites have been previously investigated, but without closer examination of the impact on subtle splice events (258). Some earlier experimental reports indicate the involvement of subtle splice variations in disease formation. For example, a G to A mutation at the first position of intron 10 in the adenosine deaminase (ADA) gene results in a shift from the distal to the proximal donor site. In consequence, a 4 nt insertion can be observed in cDNAs of ADA deficient patients (259). In contrast to this kind of proximal Δ4 donors usage other cases of donor mutations at the first intron base shifted splicing to cryptic donor splice sites located more than 10 nt downstream of the mutation (258). Finally, to address subtle donor splicing in a medical way, the measuring of this pattern over various tissues and by uti-

lizing microarrays, will give further insights in the regulation of these splice forms.

Chapter 3 The DNA and mRNA sequence based analysis of alternative splicing has revealed a magnitude of information, covering different splicing patterns, signals and binding motifs and their evolutionary conservation. However, to understand what initiates and governs the position specific splicing of mRNA requires to understand the network of molecules, which participate in the splicing reaction.
The first and second chapter demonstrated, how sequence inherent signals concur with the occurrence of alternative splicing patterns and how A5EΔ4 splice variants are enriched in the functional category of mRNA-binding factors. Also the importance of the donor site nucleotide composition is clearly a crucial parameter to distinguish AS types. These findings strongly underline the influence of trans-acting elements in the splicing process. Which proteins do interact at which time to realize a specific spliceosome and how? Does a redundancy in the proteomic complement exists, which allows to perform the splicing reaction with high accuracy and reproducibility even under changing conditions ?
Presently, no database can serve with a map of interactions signaling reactions, which describe the spliceosome assembly or summarize and visualize possible nuances of this network. The last chapter aimed on fixating and arranging a large part of data on spliceosomal proteins and ribonucleoprotein complexes into a model for structural analysis.
The achievements of the general model can be summarized as follows:

- The model establishes a network of reactions leading to the assembly of an active spliceosome.
- The model includes more than one hundred molecular components including their ordered and directed interactions, derived by extensive screening of literature.
- The model combines RNA as well as proteins and intermediate compounds of both molecules as reacting and interacting species in the network.
- Components of the spliceosome are modeled in different layers to provide a clear representation. Large biological networks often suffer the problem of overcrowded layout, which impedes orientation. Compartmentalization and

hierarchical layers as shown in Figure 3.10 and Figure 3.11 are approaches to circumvent this problem.

- The network is designed as Petri net model, which allowed the computational validation by decomposition into minimal T-invariants.
- The model network is completely covered by T-invariants, each corresponding to a biological process during spliceosome assembly.
- T-invariants are further grouped into functional clusters, which are a helpful technique to accentuate similarities and differences in long pathways.

Further, as regards biological content, the model allows the following conclusions:

- Excluding the decomposition reactions of the multitude of intermediate complexes prevents the formation of futile cycles.
- The presence of a discard pathways as previously suggested (197), is demonstrative for a regulatory circuit, but also for the problem of combinatorial complexity of spliceosome assembly.
- The different pathways of E-complex assembly indicate redundant modes of spliceosome assembly.
- All stages of spliceosome assembly involve factors, which possess a multitude of interaction partners, which presently cannot be included into the model due to uncertainty about their function and time of entering the assembly process.
- The feedback of non-protein coding alternative splice forms on spliceosome assembly as demonstrated in chapter 2, is presently not clear.

The network of spliceosomal assembly as presented here, serves as a basic scaffold to successively map the occurrence and impact of alternative splice events on the assembly pathway. This may aid in the investigation of new hypotheses about which alternative splice events contribute to which spliceosomal states, allowing to classify spliceosomes in more detail according to their composition and assembly. For example, if a general splicing factor like ASF/SF2 becomes alternatively spliced, such that it can not be phosporylated and hence, participate in splice site recognition anymore, one would expect a spliceosome, which can only recognize

very strong consensus donor sites. Since also other splicing factors, for example, SC35 or TIA1 are known to influence the initial steps during E-complex assembly, redundancy in recognition of pre-mRNA signals by interchangeable factors must be taken into account. This introduces some degree of redundancy, but may assure the fidelity of the splice reaction under different conditions.
Finally, another important aspect for future works, envisioning the step from structural to kinetic modeling, is the consideration of concentration levels of spliceosomal core components and auxiliary factors. For example, RNAi knockouts of transcripts of several DExH/D helicases (Brr2, Prp5, Prp22) affect the exon inclusion levels of the DSCAM exon 4 cluster (201). Since alternative splicing can be associated with weaker splice sites, the kinetics of generating such splice patterns might be more affected by fluctuating levels of spliceosomal proteins, compared to constitutive splice sites. Consistent with this, the concentration ratios of antagonistic splicing factors as ASF/SF2 and hnRNP A1 have been shown to vary over a wide range of tissues (260), which suggests a tissue dependent difference in the kinetics of spliceosome assembly.

Beside the core spliceosome there exist related regulatory networks, which involve essential alternative splicing events, hence being interesting candidate networks for follow-up models to this work. Two examples are given below:

- The Drosophila female-specific splicing regulatory protein Tra2 stabilizes the binding of SR proteins to a splicing enhancer element, which facilitates recruitment of the U2AF heterodimer to the weak female-specific polypyrimidine tract of the *dsx* gene (6). Together with an auto-regulatory splicing mechanism on its own pre-mRNA, this protein controls the production of sex specific alternative isoforms in the fruitfly development. The human orthologue *Tra2ß* is not less important, because it regulates splicing of exon 10 in the *Tau* gene (261), whose missplicing causes a severe variant of frontotemporal dementia (FTPD17).

- The Drosophila SXL protein constitutes a splicing factor, which is functioning prior to *Tra2*. It involves SPF45 as co-regulator and binds cooperatively up and downstream of an alternative exon in its own pre-mRNA (6). The auto-regulatory feedback systems in the sex determination cascade of the fruitfly, constitutes interesting models for system biological analyses (262).

Some conclusions emerged from this work also with respect to technical aspects

of Petri net modeling: *i)* Petri net modeling software, as used here (235), should be improved to overcome the problem of node ID preservation. Presently, the modification of parts of the network resets the IDs of all, even of unchanged, places and transitions. In consequence, it is unnecessary exacerbated, especially in large networks, to compare the outcomes of T-invariant and MCTS analyses between different model scenarios. *ii)* Reactions and reactants should be labeled uniformly in a standardized way. The challenge lies in the balance of mnemonic labels to preserve a clear network representation. Ideally, the model would be converted into a standardized design language (e.g., *systems biology markup language* (263)) to make it comparable and applicable for other analysis tools. *iii)* A great deal of progress in the *ab initio* design of such model networks will be gained, if an automated text mining procedure for timed biological interactions between spliceosomal components is integrated into the design phase. This ascertains a higher information background on biochemical reactions and provides the possibility to estimate a reliability measure via the number of independent reports for each modeled reaction.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel angefertigt habe. Mir ist die geltende Promotionsordnung bekannt und ich habe weder die Hilfe eines Promotionsberaters in Anspruch genommen, noch haben Dritte unmittelbare oder mittelbare geldwerte Leistungen für Arbeit erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die vorgelegte Dissertation wurde noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Weiterhin habe ich mich mit der vorliegenden Arbeit an keiner anderen Hochschule um den akademischen Grad doctor rerum naturalium (Dr. rer. nat.) beworben und weder früher noch gegenwärtig die Eröffnung eines Verfahrens zum Erwerb des o.g. akademischen Grades an einer anderen Hochschule beantragt.

Berlin, 24. November 2008

.....
(Ralf H. Bortfeldt)

Glossary

A

Acceptor: 3′ splice site, that marks the intron end during the splicing process. This boundary is involved in the second step of splicing where the downstream exon „accepts“ the upstream exon when it is handed in for ligation to the downstream exon by the spliceosome

ALN: Dynamic programming algorithm for identifying gene structures by aligning protein sequences or protein homology profiles against codon-encoded DNA. Splicing signals, long gaps, coding potential and frameshift errors are considered in finding optimal matches to the reference sequence

AltExtron: Collection of manually curated and experimentally confirmed alternatively spliced exons, based on annotated GenBank sequences and together with →AltExtron part of the Alternative Transcript Diversity Consortium a the European Molecular Biological Laboratory (EMBL) →*http://www.ebi.ac.uk/asd/altextron*

AltSplice: Production and annotation pipeline for computationally derived alternative splice events based on the Ensembl genome annotation project →*http://www.ebi.ac.uk/asd/altsplice/*

B

Bipartite Graph: A graph $G = \{V, E\}$ whose nodes can be split into disjoint subsets A, B, such that no edges exist between elements of the same set. (V = set of *vertices* or *nodes* and E = set of *edges*).
If $\{v, w\} \in E$ then holds either $v \in A \wedge w \in B$ or $v \in B \wedge w \in A$. $\{A, B\}$ is called *bipartition* of G.

BLAST: Basic local alignment search tool. Generic term for a collection of programs which implement an alogrithm for comparing a new DNA or protein sequence against a database of known sequences. The algorithm returns a set of optimal local alignments given some specified parameter (e.g. gap penalties, seed word length) and their significance as determined by the chance to find the query sequence in a database of random sequences. WU-BLAST is one implementation of the BLAST alogrithm provided by the University of Washington →*http://blast.wustl.edu/*

C

Cajal Body: Organelle of the nucleus, 0.1-2.0 micrometer in diameter. Cajal bodies are considered as sites of assembly and/or modification of the transcription machinery and splicing factors

cDNA: *Copy* DNA. Doublestranded DNA prepared from reversed transcribed RNA. cDNA is more stable than RNA and hence suited to study the modifications a transcript is subjected to during the RNA maturation process

CDS: Coding sequence, part of the mature messenger RNA which encodes the primary amino acid sequence of a protein

css: constitutive splice site, splice site, flanking an exon for which at a given stage of transcript knowledge no alternative splicing is observed

D

dbEST: Division of GenBank that stores → EST with some basic annotations, presently (release August 2008) hosting more than 1,600 organisms, with human (8,137,901 EST) and mouse (4,850,258 EST) being the most frequently represented species →*http://www.ncbi.nlm.nih.gov/dbEST*

Donor: 5′ splice site marking the exon-intron boundary during initiation of the splicing process and commits the upstream exon during initiation of spliceosome assembly

E

Ensembl: Genome Browser of the European Molecularbiological Laboratory (EMBL) and the European Bioinformatics Institute (EBI), based on a gene modelling pipeline that integrates sequence data from primary sources of more than ∼ 40 genomes (release 47, July 2008). Especially suited for bioinformatic analysis due to fast access of sequences and their annotations via a Perl programming interface →*http://www.ensembl.org*

EST: Expressed Sequence Tag; short cDNA sequence read of ∼400-600 nt length, consitituing parts of a mRNA. EST are often prepared from total RNA after cell breakdown, reverse transcription, cloning and sequencing. Oligo deoxy-thymine primer or poly-dT beads are used for selectively enriching messenger RNA from cell extracts, hence EST data may represent only a minor fraction of the total cellular RNA among which mRNA makes up for ∼5% in eucaryotic cells

EvidenceViewer: Graphical visualisation of biological evidence supporting a particular gene model. Linked to the genomic map viewer of GenBank, displaying all RefSeq models, annotated known or potential transcripts of a gene, which align to the area of interest →*http://www.ncbi.nlm.nih.gov/sutils/static/evvdoc.html*

Exon: pre-mRNA part which is spliced into the mature messenger RNA. As determined by the human genome project, the mean exon length of human protein coding genes is 145 bp (median 122 bp), see also

G

GenBank: Primary database of publicly available nucleotide sequences and their protein translations with basic annotations, maintained at the National Center of Biotechnology Information, Bethesda USA. GenBank stores sequences produced by laboratories worldwide and contained with release 166 (June 2008) 88,554,578 sequences of more than 100,000 organisms →*http://www.ncbi.nlm.nih.gov/Genbank/*

GeneMine: Software pipeline for automated analysis of DNA and protein sequences by integrating information of different biological web resources and making them available for mining in a local database →*http://bioinformatics.ucla.edu/genemine*

GENOA: Genome annotation pipeline. Performs spliced alignments of repeat-masked cDNAs (BLASTN) against genomic sequence to find significant *loci* of transcription. Subsequently, cDNAs and ESTs are stringently (re)aligned to these *loci* by the algorithms mRNAvsGen and Sim4 respectively, to infer high quality exons →*http://genes.mit.edu/genoa*

I

Intron: pre-mRNA part which is spliced out and hence missing in the mature messenger RNA. As determined by the human genome project, the mean intron length of human protein coding genes is 3,365 bp (median 1,023 bp), see also

M

MegaBlast: Speed optimised alignment program for only slightly differing nucleotide sequences. By using a "greedy" algorithm it is about 10 times faster compared to BLASTN, when used with larger word sizes of the seed alignments (over 16 nucleotides) and consequently can better handle longer sequences →*http://www.ncbi.nlm.nih.gov/blast/megablast.shtml*

R

RefSeq: Non-redundant, manually curated database of sequences of more than 5,000 different taxa (release 30 - Juli 2008) including genomic DNA, transcripts, and proteins with stable identifiers. RefSeq entries are derived from publicly available archives of primary research data and as such provide a synthesis of information →*http://www.ncbi.nlm.nih.gov/RefSeq/*

RS domain: → SR proteins

rss: reference splice site, synonym for constitutive splice site due to the fact that constitutive splice sites may be involved in alternative splicing, which has just not been verified yet

S

Sim4: Program designed for rapid aligning spliced transcripts against genomic DNA of sizes >100 kB. The algorithms finds matching seeds (12-mers) and extends them to gap-free genomic HSPs, which are subsequently assembled to an exon core chain by dynamic programming, favouring HSP positions at GT..AG or CT..AC intron boundaries.

SQL storage types: Define the data types each field in a SQL table can hold. This is an important issue in design of biological databases, as it is beforehand often unknown how large certain values can become during analysis. For example a field which stores intron sequences (string type) is critically initialized with the MySQL data type *TEXT* as this allows only the storage of values of maximal $L = 2^{16} + 2$ bytes (1 string character = 1 byte).

SR proteins Family of splicing factors that share a modular structure consisting of one or two copies of an N-terminal RRM (RNA-recognition motif) followed by a C-terminal domain rich in alternating serine and arginine residues, termed the *RS* domain

T

TBLASTN: Runtime intensive BLAST due to aligning a query protein sequence against a dynamically translated nucleotide sequence database resulting in a six-fold increased sequence search space because of three possible reading frames on forward and reverse strand →*http://blast.ncbi.nlm.nih.gov/blast/*

U

UCSC Browser: Via a webbrowser accessible graphical representation of many eukaryotic genome sequences and their annotation maintained at the University of California Santa Cruz. Additionally, a table browser allows direct access to the underlying database →*http://genome.ucsc.edu/*

UniGene: Database of Unified clusters of ESTs and full-length mRNA sequences. Transcripts are clustered in a non-redundant and gene-oriented way and annotated with related information (e.g. tissue type). The latest version of the human taxon (#214, June 2008) contains 122,958 clusters representing more than 6.9 million transcripts →*http://www.ncbi.nlm.nih.gov/UniGene*

Bibliography

[1] Bortfeldt, R. H., Herrmann, A., Pospisil, H. & Schuster, S. *Validation of Human Alternative Splice Forms, Using the EASED Platform and Multiple Splice Site Discriminating Features.* in Mathematical Modeling of Biological Systems (Birkhäuser, Boston and Basel) **Part V**, 337–349 (2007).

[2] Hiller, M. *et al. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity.* Nat Genet **36**(12), 1255–1257 Dec (2004).

[3] Holste, D., Huo, G., Tung, V. & Burge, C. B. *HOLLYWOOD: a comparative relational database of alternative splicing.* Nucleic Acids Res **34**(Database issue), D56–D62 Jan (2006).

[4] Bortfeldt, R., Schindler, S., Szafranski, K., Schuster, S. & Holste, D. *Comparative analysis of sequence features involved in the recognition of tandem splice sites.* BMC Genomics **9**(1), 202 Apr (2008).

[5] Berget, S. M., Moore, C. & Sharp, P. A. *Spliced segments at the 5' terminus of adenovirus 2 late mRNA.* Proc Natl Acad Sci U S A **74**(8), 3171–3175 Aug (1977).

[6] Black, D. L. *Mechanisms of alternative pre-messenger RNA splicing.* Annu Rev Biochem **72**, 291–336 (2003).

[7] Graveley, B. R. *Alternative splicing: increasing diversity in the proteomic world.* Trends Genet **17**(2), 100–107 Feb (2001).

[8] Maniatis, T. & Tasic, B. *Alternative pre-mRNA splicing and proteome expansion in metazoans.* Nature **418**(6894), 236–243 Jul (2002).

[9] Ladd, A. N. & Cooper, T. A. *Finding signals that regulate alternative splicing in the post-genomic era.* Genome Biol **3**(11), reviews0008 (2002).

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

[11] Venables, J. P. *Aberrant and alternative splicing in cancer.* Cancer Res **64**(21), 7647–7654 Nov (2004).

[12] Kan, Z., Rouchka, E. C., Gish, W. R. & States, D. J. *Gene structure prediction and alternative splicing analysis using genomically aligned ESTs.* Genome Res **11**(5), 889–900 May (2001).

[13] Xu, Q., Modrek, B. & Lee, C. *Genome-wide detection of tissue-specific alternative splicing in the human transcriptome.* Nucleic Acids Res **30**(17), 3754–3766 Sep (2002).

[14] Sugnet, C. W., Kent, W. J., Ares, M. & Haussler, D. *Transcriptome and genome conservation of alternative splicing events in humans and mice.* Pac Symp Biocomput , 66–77 (2004).

[15] Philipps, D. L., Park, J. W. & Graveley, B. R. *A computational and experimental approach toward a priori identification of alternatively spliced exons.* RNA **10**(12), 1838–1844 Dec (2004).

[16] Hiller, M. *et al. TassDB: a database of alternative tandem splice sites.* Nucleic Acids Res Nov (2006).

[17] Galperin, M. Y. *The Molecular Biology Database Collection: 2005 update.* Nucleic Acids Res **33**(Database issue), D5–24 (2005).

[18] Hertel, K. J. & Graveley, B. R. *RS domains contact the pre-mRNA throughout spliceosome assembly.* Trends Biochem Sci **30**(3), 115–118 Mar (2005).

[19] Bourgeois, C. F., Lejeune, F. & Stévenin, J. *Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA.* Prog Nucleic Acid Res Mol Biol **78**, 37–88 (2004).

[20] Shen, H. & Green, M. R. *A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly.* Mol Cell **16**(3), 363–373 Nov (2004).

[21] Liu, H. X., Zhang, M. & Krainer, A. R. *Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins.* Genes Dev **12**(13), 1998–2012 (1998). hard copy available.

[22] Smith, C. W. & Valcárcel, J. *Alternative pre-mRNA splicing: the logic of combinatorial control.* Trends Biochem Sci **25**(8), 381–388 Aug (2000).

[23] Yeo, G. & Burge, C. B. *Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals.* RECOMB'03 April 10-13 Berlin, Germany (2003).

[24] Clark, F. & Thanaraj, T. A. *Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human.* Hum Mol Genet **11**(4), 451–464 Feb (2002).

[25] Zhang, M. Q. *Statistical features of human exons and their flanking regions.* Hum Mol Genet **7**(5), 919–32 (1998).

[26] Fairbrother, W. G., Yeh, R.-F., Sharp, P. A. & Burge, C. B. *Predictive identification of exonic splicing enhancers in human genes.* Science **297**(5583), 1007–1013 Aug (2002).

[27] Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q. & Krainer, A. R. *ESEfinder: A web resource to identify exonic splicing enhancers.* Nucleic Acids Res **31**(13), 3568–3571 Jul (2003).

[28] Schuler, G. D. *Pieces of the puzzle: expressed sequence tags and the catalog of human genes.* J Mol Med **75**(10), 694–698 Oct (1997). hard copy available.

[29] Haas, B. J. *et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.* Nucleic Acids Res **31**(19), 5654–5666 Oct (2003).

[30] Foissac, S. & Schiex, T. *Integrating alternative splicing detection into gene prediction.* BMC Bioinformatics **6**, 25 (2005).

[31] Mironov, A. A., Fickett, J. W. & Gelfand, M. S. *Frequent alternative splicing of human genes.* Genome Res **9**(12), 1288–1293 Dec (1999).

[32] Brett, D. *et al. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms.* FEBS Letters **474**(1), 83–6 (2000).

[33] Lander, E. S. *et al. Initial sequencing and analysis of the human genome.* Nature **409**(6822), 860–921 Feb (2001).

[34] Modrek, B., Resch, A., Grasso, C. & Lee, C. *Genome-wide detection of alternative splicing in expressed sequences of human genes.* Nucleic Acids Res **29**(13), 2850–2859 Jul (2001).

[35] Johnson, J. M. *et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.* Science **302**(5653), 2141–2144 Dec (2003).

[36] Kan, Z., Castle, J., Johnson, J. M. & Tsinoremas, N. F. *Detection of novel splice forms in human and mouse using cross-species approach.* Pac Symp Biocomput , 42–53 (2004).

[37] Sorek, R. *et al. A non-EST-based method for exon-skipping prediction.* Genome Res **14**(8), 1617–1623 Aug (2004).

[38] Pan, Q. *et al. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform.* Mol Cell **16**(6), 929–941 Dec (2004).

[39] Kriventseva, E. V. *et al. Increase of functional diversity by alternative splicing.* Trends Genet **19**(3), 124–128 Mar (2003).

[40] Nakao, M. *et al. Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals.* Nucleic Acids Res **33**(8), 2355–2363 (2005).

[41] Hiller, M., Huse, K., Platzer, M. & Backofen, R. *Non-EST based prediction of exon skipping and intron retention events using Pfam information.* Nucleic Acids Res **33**(17), 5611–5621 (2005).

[42] Gelfand, M. S., Dubchak, I., Dralyuk, I. & Zorn, M. *ASDB: database of alternatively spliced genes.* Nucleic Acids Res **27**(1), 301–302 Jan (1999).

[43] Dralyuk, I., Brudno, M., Gelfand, M. S., Zorn, M. & Dubchak, I. *ASDB: database of alternatively spliced genes.* Nucleic Acids Res **28**(1), 296–297 Jan (2000). hard copy available.

[44] Brett, D. *et al. EST analysis online: WWW tools for detection of SNPs and alternative splice forms.* Trends Genet **16**(9), 416–418 Sep (2000).

[45] Burset, M., Seledtsov, I. A. & Solovyev, V. V. *SpliceDB: database of canonical and non-canonical mammalian splice sites.* Nucleic Acids Res **29**(1), 255–9 (2001).

[46] Coward, E., Haas, S. A. & Vingron, M. *SpliceNest: visualizing gene structure and alternative splicing based on EST clusters.* Trends Genet **18**(1), 53–55 (2002).

[47] Gupta, S., Zink, D., Korn, B., Vingron, M. & Haas, S. A. *Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing.* BMC Genomics **5**(1), 72 Sep (2004).

[48] Huang, Y.-H., Chen, Y.-T., Lai, J.-J., Yang, S.-T. & Yang, U.-C. *PALS db: Putative Alternative Splicing database.* Nucleic Acids Res **30**(1), 186–190 Jan (2002).

[49] Huang, H.-D., Horng, J.-T., Lee, C.-C. & Liu, B.-J. *ProSplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data.* Genome Biol **4**(4), R29 (2003).

[50] Lee, C., Atanelov, L., Modrek, B. & Xing, Y. *ASAP: the Alternative Splicing Annotation Project.* Nucleic Acids Res **31**(1), 101–5 (2003).

[51] Kim, E., Magen, A. & Ast, G. *Different levels of alternative splicing among eukaryotes.* Nucleic Acids Res **35**(1), 125–131 (2007).

[52] Pospisil, H., Herrmann, A., Bortfeldt, R. H. & Reich, J. G. *EASED: Extended Alternatively Spliced EST Database.* Nucleic Acids Res **32**(Database issue), D70–4 (2004).

[53] Thanaraj, T. A. *et al. ASD: the Alternative Splicing Database.* Nucleic Acids Res **32**(Database issue), D64–9 (2004).

[54] Stamm, S. *et al. ASD: a bioinformatics resource on alternative splicing.* Nucleic Acids Res **34**(Database issue), D46–D55 Jan (2006).

[55] de la Grange, P., Dutertre, M., Martin, N. & Auboeuf, D. *FAST DB: a website resource for the study of the expression regulation of human gene products.* Nucleic Acids Res **33**(13), 4276–4284 (2005).

[56] Huang, H.-D., Horng, J.-T., Lin, F.-M., Chang, Y.-C. & Huang, C.-C. *SpliceInfo: an information repository for mRNA alternative splicing in human genome.* Nucleic Acids Res **33**(Database issue), D80–D85 Jan (2005).

[57] Zheng, C. L. *et al. MAASE: an alternative splicing database designed for supporting splicing microarray applications.* RNA **11**(12), 1767–1776 Dec (2005).

[58] Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M. & Gotoh, O. *Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns.* Bioinformatics **22**(10), 1211–1216 May (2006).

[59] Texier, V. L. *et al. AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation.* BMC Bioinformatics **7**, 169 (2006).

[60] Kahn, A. B. *et al. SpliceMiner: a high-throughput database implementation of the NCBI Evidence Viewer for microarray splice variant analysis.* BMC Bioinformatics **8**, 75 (2007).

[61] Lacroix, Z., Legendre, C., Raschid, L. & Snyder, B. *BIPASS: BioInformatics Pipeline Alternative Splicing Services.* Nucleic Acids Res **35**(Web Server issue), W292–W296 Jul (2007).

[62] Graveley, B. R. *Sorting out the complexity of SR protein functions.* Rna **6**(9), 1197–211 (2000).

[63] Tacke, R. & Manley, J. L. *The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities.* Embo J **14**(14), 3540–51 (1995). hard copy available.

[64] Tacke, R., Chen, Y. & Manley, J. L. *Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer.* Proc Natl Acad Sci U S A **94**(4), 1148–53 (1997). hard copy available.

[65] Ramchatesingh, J., Zahler, A. M., Neugebauer, K. M., Roth, M. B. & Cooper, T. A. *A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer.* Mol Cell Biol **15**(9), 4898–4907 Sep (1995). hard copy available.

[66] Humphrey, M. B., Bryan, J., Cooper, T. A. & Berget, S. M. *A 32-nucleotide exon-splicing enhancer regulates usage of competing 5' splice sites in a differential internal exon.* Mol Cell Biol **15**(8), 3979–88 (1995). hard copy available.

[67] Elrick, L. L., Humphrey, M. B., Cooper, T. A. & Berget, S. M. *A short sequence within two purine-rich enhancers determines 5' splice site specificity.* Mol Cell Biol **18**(1), 343–52 (1998). hard copy available.

[68] König, H., Ponta, H. & Herrlich, P. *Coupling of signal transduction to alternative pre-mRNA splicing by a composite splice regulator.* EMBO J **17**(10), 2904–2913 May (1998). hard copy available.

[69] Boguski, M. S., Lowe, T. M. & Tolstoshev, C. M. *dbEST–database for "expressed sequence tags".* Nat Genet **4**(4), 332–3 (1993). hard copy available.

[70] Gorodkin, J., Heyer, L. J., Brunak, S. & Stormo, G. D. *Displaying the information contents of structural RNA alignments: the structure logos.* Comput Appl Biosci **13**(6), 583–586 Dec (1997).

[71] Cartegni, L., Chew, S. L. & Krainer, A. R. *Listening to silence and understanding nonsense: exonic mutations that affect splicing.* Nat Rev Genet **3**(4), 285–298 Apr (2002).

[72] Hertel, K. J. & Maniatis, T. *The function of multisite splicing enhancers.* Mol Cell **1**(3), 449–55 (1998). hard copy available.

[73] D'Souza, I. & Schellenberg, G. D. *Determinants of 4-repeat tau expression. Coordination between enhancing and inhibitory splicing sequences for exon 10 inclusion.* Journal of Biological Chemistry **275**(23), 17700–9 (2000). hard copy available.

[74] Lou, H., Neugebauer, K. M., Gagel, R. F. & Berget, S. M. *Regulation of alternative polyadenylation by U1 snRNPs and SRp20.* Mol Cell Biol **18**(9), 4977–85 (1998).

[75] Buratti, E. & Baralle, F. E. *Influence of RNA secondary structure on the pre-mRNA splicing process.* Mol Cell Biol **24**(24), 10505–10514 Dec (2004).

[76] Graveley, B. R. *Sex, AGility, and the regulation of alternative splicing.* Cell **109**(4), 409–412 May (2002).

[77] Mouchel, N., Broackes-Carter, F. & Harris, A. *Alternative 5' exons of the CFTR gene show developmental regulation.* Hum Mol Genet **12**(7), 759–769 Apr (2003).

[78] Hutton, M. *et al. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17.* Nature **393**(6686), 702–705 Jun (1998).

[79] Blanchette, M. & Chabot, B. *Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization.* EMBO J **18**(7), 1939–1952 Apr (1999).

[80] Shomron, N., Alberstein, M., Reznik, M. & Ast, G. *Stress alters the subcellular distribution of hSlu7 and thus modulates alternative splicing.* J Cell Sci **118**(Pt 6), 1151–1159 Mar (2005).

[81] Zavolan, M. *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), 1290–1300 Jun (2003).

[82] Akerman, M. & Mandel-Gutfreund, Y. *Alternative splicing regulation at tandem 3' splice sites.* Nucleic Acids Res **34**(1), 23–31 (2006).

[83] Hiller, M. *et al. Phylogenetically widespread alternative splicing at unusual GYNGYN donors.* Genome Biol **7**(7), R65 Jul (2006).

[84] Hiller, M. *et al. Assessing the fraction of short-distance tandem splice sites under purifying selection.* RNA **14**(4), 616–629 Apr (2008).

[85] Berget, S. M. *Exon recognition in vertebrate splicing.* J Biol Chem **270**(6), 2411–2414 Feb (1995).

[86] Eperon, I. C., Ireland, D. C., Smith, R. A., Mayeda, A. & Krainer, A. R. *Pathways for selection of 5' splice sites by U1 snRNPs and SF2/ASF.* EMBO J **12**(9), 3607–3617 Sep (1993).

[87] Eperon, I. C. *et al. Selection of Alternative 5 Splice Sites: Role of U1snRNP and Models for the Antagonistic Effects of SF2/ASF and hnRNP A1.* Molecular and Cellular Biology **20**(22), 8303–8318 Nov (2000). hard copy available.

[88] Bai, Y., Lee, D., Yu, T. & Chasin, L. A. *Control of 3' splice site choice in vivo by ASF/SF2 and hnRNP A1.* Nucleic Acids Res **27**(4), 1126–1134 Feb (1999).

[89] Roca, X., Sachidanandam, R. & Krainer, A. R. *Determinants of the inherent strength of human 5' splice sites.* RNA **11**(5), 683–698 May (2005).

[90] Borensztajn, K. *et al. Oriented Scanning Is the Leading Mechanism Underlying 5' Splice Site Selection in Mammals.* PLoS Genet **2**(9), e138 Sep (2006).

[91] Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. *A computer program for aligning a cDNA sequence with a genomic DNA sequence.* Genome Res **8**(9), 967–974 Sep (1998).

[92] Zhang, M. & Gish, W. *Improved spliced alignment from an information theoretic approach.* Bioinformatics **22**(1), 13–20 Jan (2006).

[93] Kent, W. J. *BLAT–the BLAST-like alignment tool.* Genome Res **12**(4), 656–664 Apr (2002).

[94] Yeo, G. & Burge, C. B. *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.* J Comput Biol **11**(2-3), 377–394 (2004).

[95] Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. *WebLogo: a sequence logo generator.* Genome Res **14**(6), 1188–1190 Jun (2004).

[96] Lewis, B. P., Green, R. E. & Brenner, S. E. *Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans.* Proc Natl Acad Sci U S A **100**(1), 189–192 Jan (2003).

[97] Jareborg, N., Birney, E. & Durbin, R. *Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs.* Genome Res **9**(9), 815–824 Sep (1999).

[98] Diehn, M. *et al. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.* Nucleic Acids Res **31**(1), 219–223 Jan (2003).

[99] Ewing, B., Hillier, L., Wendl, M. C. & Green, P. *Base-calling of automated sequencer traces using phred. I. Accuracy assessment.* Genome Res **8**(3), 175–185 Mar (1998).

[100] Staden, R., Beal, K. F. & Bonfield, J. K. *The Staden package, 1998.* Methods Mol Biol **132**, 115–130 (2000).

[101] Wang, Z. *et al. Systematic identification and analysis of exonic splicing silencers.* Cell **119**(6), 831–845 Dec (2004).

[102] Yeo, G. W., Nostrand, E. L. V. & Liang, T. Y. *Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements.* PLoS Genet **3**(5), e85 May (2007).

[103] Zhang, X. H.-F. & Chasin, L. A. *Computational definition of sequence motifs governing constitutive exon splicing.* Genes Dev **18**(11), 1241–1250 Jun (2004).

[104] Glantz, S. A. *Primer of Biostatistics.* McGraw-Hill Medical, November (2001).

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

[106] Wang, Z., Xiao, X., Nostrand, E. V. & Burge, C. B. *General and specific functions of exonic splicing silencers in splicing control.* Mol Cell **23**(1), 61–70 Jul (2006).

[107] Siepel, A. *et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res **15**(8), 1034–1050 Aug (2005).

[108] Katsu, R., Onogi, H., Wada, K., Kawaguchi, Y. & Hagiwara, M. *Novel SR-rich-related protein clasp specifically interacts with inactivated Clk4 and induces the exon EB inclusion of Clk.* J Biol Chem **277**(46), 44220–44228 Nov (2002).

[109] Lin, C. L., Leu, S., Lu, M. C. & Ouyang, P. *Over-expression of SR-cyclophilin, an interaction partner of nuclear pinin, releases SR family splicing factors from nuclear speckles.* Biochem Biophys Res Commun **321**(3), 638–647 Aug (2004).

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

[111] Roca, X., Sachidanandam, R. & Krainer, A. R. *Intrinsic differences between authentic and cryptic 5' splice sites.* Nucleic Acids Res **31**(21), 6321–6333 Nov (2003).

[112] Holste, D., Grosse, I., Buldyrev, S. V., Stanley, H. E. & Herzel, H. *Optimization of coding potentials using positional dependence of nucleotide frequencies.* J Theor Biol **206**(4), 525–537 Oct (2000).

[113] Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern Classification (2nd Edition).* Wiley-Interscience, November (2000).

[114] Chern, T.-M. *et al. A simple physical model predicts small exon length variations.* PLoS Genet **2**(4), e45 Apr (2006).

[115] Bi, J., Xia, H., Li, F., Zhang, X. & Li, Y. *The effect of U1 snRNA binding free energy on the selection of 5' splice sites.* Biochem Biophys Res Commun **333**(1), 64–69 Jul (2005).

[116] Freund, M. *et al. A novel approach to describe a U1 snRNA binding site.* Nucleic Acids Res **31**(23), 6963–6975 Dec (2003).

[117] Lund, M. & Kjems, J. *Defining a 5' splice site by functional selection in the presence and absence of U1 snRNA 5' end.* RNA **8**(2), 166–179 Feb (2002).

[118] Staley, J. P. & Guthrie, C. *An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p.* Mol Cell **3**(1), 55–64 Jan (1999).

[119] Carmel, I., Tal, S., Vig, I. & Ast, G. *Comparative analysis detects dependencies among the 5' splice-site positions.* RNA **10**(5), 828–840 May (2004).

[120] Yeo, G. W., Nostrand, E. V., Holste, D., Poggio, T. & Burge, C. B. *Identification and analysis of alternative splicing events conserved in human and mouse.* Proc Natl Acad Sci U S A **102**(8), 2850–2855 Feb (2005).

[121] Koren, E., Lev-Maor, G. & Ast, G. *The emergence of alternative 3' and 5' splice site exons from constitutive exons.* PLoS Comput Biol **3**(5), e95 May (2007).

[122] Caputi, M. & Zahler, A. M. *Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family.* J Biol Chem **276**(47), 43850–43859 Nov (2001).

[123] McCullough, A. J. & Berget, S. M. *An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5' splice sites.* Mol Cell Biol **20**(24), 9225–9235 Dec (2000).

[124] Martinez-Contreras, R. *et al. Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing.* PLoS Biol **4**(2), e21 Feb (2006).

[125] Wang, E., Dimova, N. & Cambi, F. *PLP/DM20 ratio is regulated by hnRNPH and F and a novel G-rich enhancer in oligodendrocytes.* Nucleic Acids Res **35**(12), 4164–4178 (2007).

[126] Brackenridge, S., Wilkie, A. O. M. & Screaton, G. R. *Efficient use of a 'dead-end' GA 5' splice site in the human fibroblast growth factor receptor genes.* EMBO J **22**(7), 1620–1631 Apr (2003).

[127] Mansilla, A. *et al. Developmental regulation of a proinsulin messenger RNA generated by intron retention.* EMBO Rep **6**(12), 1182–1187 Dec (2005).

[128] Zhang, J. *et al. Cloning and functional characterization of ACAD-9, a novel member of human acyl-CoA dehydrogenase family.* Biochem Biophys Res Commun **297**(4), 1033–1042 Oct (2002).

[129] Letunic, I. *et al. SMART 5: domains in the context of genomes and networks.* Nucleic Acids Res **34**(Database issue), D257–D260 Jan (2006).

[130] Green, R. E. *et al. Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes.* Bioinformatics **19 Suppl 1**, i118–i121 (2003).

[131] Liu, H. X., Cartegni, L., Zhang, M. Q. & Krainer, A. R. *A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes.* Nat Genet **27**(1), 55–58 Jan (2001).

[132] Wollerton, M. C., Gooding, C., Wagner, E. J., Garcia-Blanco, M. A. & Smith, C. W. J. *Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay.* Mol Cell **13**(1), 91–100 Jan (2004). hard copy available.

[133] Baek, D. & Green, P. *Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing.* Proc Natl Acad Sci U S A **102**(36), 12813–12818 Sep (2005).

[134] Ast, G. *How did alternative splicing evolve?* Nat Rev Genet **5**(10), 773–782 Oct (2004).

[135] Dou, Y., Fox-Walsh, K. L., Baldi, P. F. & Hertel, K. J. *Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site.* RNA **12**, 1–10 Oct (2006).

[136] Chabot, B., Blanchette, M., Lapierre, I. & Branche, H. L. *An intron element modulating 5' splice site selection in the hnRNP A1 pre-mRNA interacts with hnRNP A1.* Mol Cell Biol **17**(4), 1776–1786 Apr (1997). hard copy available.

[137] Sorek, R. & Ast, G. *Intronic sequences flanking alternatively spliced exons are conserved between human and mouse.* Genome Res **13**(7), 1631–1637 Jul (2003).

[138] Buehler, M., Steiner, S., Mohn, F., Paillusson, A. & Muehlemann, O. *EJC-independent degradation of nonsense immunoglobulin-mu mRNA depends on 3' UTR length.* Nat Struct Mol Biol **13**(5), 462–464 May (2006).

[139] O'Neill, J. P., Rogan, P. K., Cariello, N. & Nicklas, J. A. *Mutations that alter RNA splicing of the human HPRT gene: a review of the spectrum.* Mutat Res **411**(3), 179–214 Nov (1998).

[140] Untergasser, G., Hermann, M., Rumpold, H. & Berger, P. *Complex alternative splicing of the GH-V gene in the human testis.* Eur J Endocrinol **139**(4), 424–427 Oct (1998).

[141] Hiller, M., Pudimat, R., Busch, A. & Backofen, R. *Using RNA secondary structures to guide sequence motif finding towards single-stranded regions.* Nucleic Acids Res **34**(17), e117 (2006).

[142] Maroney, P. A., Romfo, C. M. & Nilsen, T. W. *Functional recognition of 5' splice site by U4/U6.U5 tri-snRNP defines a novel ATP-dependent step in early spliceosome assembly.* Mol Cell **6**(2), 317–328 Aug (2000).

[143] Crispino, J. D. & Sharp, P. A. *A U6 snRNA:pre-mRNA interaction can be rate-limiting for U1-independent splicing.* Genes Dev **9**(18), 2314–2323 Sep (1995).

[144] Hwang, D. Y. & Cohen, J. B. *Base pairing at the 5' splice site with U1 small nuclear RNA promotes splicing of the upstream intron but may be dispensable for slicing of the downstream intron.* Mol Cell Biol **16**(6), 3012–3022 Jun (1996).

[145] Kloetzel, P. M. *Generation of major histocompatibility complex class I antigens: functional interplay between proteasomes and TPPII.* Nat Immunol **5**(7), 661–669 Jul (2004).

[146] Yewdell, J. W., Reits, E. & Neefjes, J. *Making sense of mass destruction: quantitating MHC class I antigen presentation.* Nat Rev Immunol **3**(12), 952–961 Dec (2003).

[147] Shastri, N., Schwab, S. & Serwold, T. *Producing nature's gene-chips: the generation of peptides for display by MHC class I molecules.* Annu Rev Immunol **20**, 463–493 (2002).

[148] Wang, R. F., Parkhurst, M. R., Kawakami, Y., Robbins, P. F. & Rosenberg, S. A. *Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen.* J Exp Med **183**(3), 1131–1140 Mar (1996).

[149] Jurica, M. S. & Moore, M. J. *Pre-mRNA splicing: awash in a sea of proteins.* Mol Cell **12**(1), 5–14 Jul (2003).

[150] Neubauer, G. *et al. Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex.* Nat Genet **20**(1), 46–50 Sep (1998).

[151] Makarov, E. M. *et al. Small nuclear ribonucleoprotein remodeling during catalytic activation of the spliceosome.* Science **298**(5601), 2205–2208 Dec (2002).

[152] Rappsilber, J., Ryder, U., Lamond, A. I. & Mann, M. *Large-scale proteomic analysis of the human spliceosome.* Genome Res **12**(8), 1231–1245 Aug (2002).

[153] Zhou, Z., Licklider, L. J., Gygi, S. P. & Reed, R. *Comprehensive proteomic analysis of the human spliceosome.* Nature **419**(6903), 182–185 Sep (2002).

[154] Valadkhan, S., Mohammadi, A., Wachtel, C. & Manley, J. L. *Protein-free spliceosomal snRNAs catalyze a reaction that resembles the first step of splicing.* RNA **13**(12), 2300–2311 Dec (2007).

[155] Stamm, S. *Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome.* Human Molecular Genetics **11**(20), 2409–2416 (2002).

[156] Hwang, D. Y. & Cohen, J. B. *U1 snRNA promotes the selection of nearby 5' splice sites by U6 snRNA in mammalian cells.* Genes Dev **10**(3), 338–350 Feb (1996).

[157] Kornblihtt, A. R., de la Mata, M., Fededa, J. P., Munoz, M. J. & Nogues, G. *Multiple links between transcription and splicing.* RNA **10**(10), 1489–1498 Oct (2004).

[158] Görnemann, J., Kotovic, K. M., Hujer, K. & Neugebauer, K. M. *Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex.* Mol Cell **19**(1), 53–63 Jul (2005).

[159] Listerman, I., Sapra, A. K. & Neugebauer, K. M. *Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells.* Nat Struct Mol Biol **13**(9), 815–822 Sep (2006).

[160] Tardiff, D. F. & Rosbash, M. *Arrested yeast splicing complexes indicate stepwise snRNP recruitment during in vivo spliceosome assembly.* RNA **12**(6), 968–979 Jun (2006).

[161] Kotovic, K. M., Lockshon, D., Boric, L. & Neugebauer, K. M. *Cotranscriptional recruitment of the U1 snRNP to intron-containing genes in yeast.* Mol Cell Biol **23**(16), 5768–5779 Aug (2003).

[162] Stevens, S. W. *et al. Composition and functional characterization of the yeast spliceosomal penta-snRNP.* Mol Cell **9**(1), 31–44 Jan (2002).

[163] Brow, D. A. *Allosteric cascade of spliceosome activation.* Annu Rev Genet **36**, 333–360 (2002).

[164] Behzadnia, N., Hartmuth, K., Will, C. L. & Lührmann, R. *Functional spliceosomal A complexes can be assembled in vitro in the absence of a penta-snRNP.* RNA **12**(9), 1738–1746 Sep (2006).

[165] Staley, J. P. & Guthrie, C. *Mechanical devices of the spliceosome: motors, clocks, springs, and things.* Cell **92**(3), 315–326 Feb (1998).

[166] Tarn, W. Y. & Steitz, J. A. *Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge.* Trends Biochem Sci **22**(4), 132–137 Apr (1997).

[167] Frilander, M. J. & Meng, X. *Proximity of the U12 snRNA with both the 5' splice site and the branch point during early stages of spliceosome assembly.* Mol Cell Biol **25**(12), 4813–4825 Jun (2005).

[168] Du, H. & Rosbash, M. *The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing.* Nature **419**(6902), 86–90 Sep (2002). hard copy available.

[169] Cao, W. & Garcia-Blanco, M. A. *A serine/arginine-rich domain in the human U1 70k protein is necessary and sufficient for ASF/SF2 binding.* J Biol Chem **273**(32), 20629–20635 Aug (1998).

[170] Foerch, P., Puig, O., Martinez, C., Seraphin, B. & Valcarcel, J. *The splicing regulator TIA-1 interacts with U1-C to promote U1 snRNP recruitment to 5' splice sites.* EMBO J **21**(24), 6882–6892 Dec (2002).

[171] Tarn, W. Y. & Steitz, J. A. *SR proteins can compensate for the loss of U1 snRNP functions in vitro.* Genes Dev **8**(22), 2704–2717 Nov (1994).

[172] Kyriakopoulou, C. *et al. U1-like snRNAs lacking complementarity to canonical 5' splice sites.* RNA **12**(9), 1603–1611 Sep (2006).

[173] Crispino, J. D., Mermoud, J. E., Lamond, A. I. & Sharp, P. A. *Cis-acting elements distinct from the 5' splice site promote U1-independent pre-mRNA splicing.* RNA **2**(7), 664–673 Jul (1996).

[174] Berglund, J. A., Abovich, N. & Rosbash, M. *A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition.* Genes Dev **12**(6), 858–867 Mar (1998).

[175] Abovich, N. & Rosbash, M. *Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals.* Cell **89**(3), 403–412 May (1997).

[176] Das, R., Zhou, Z. & Reed, R. *Functional association of U2 snRNP with the ATP-independent spliceosomal complex E.* Mol Cell **5**(5), 779–787 May (2000).

[177] Dönmez, G., Hartmuth, K., Kastner, B., Will, C. L. & Lührmann, R. *The 5' end of U2 snRNA is in close proximity to U1 and functional sites of the pre-mRNA in early spliceosomal complexes.* Mol Cell **25**(3), 399–411 Feb (2007).

[178] Dybkov, O. *et al. U2 snRNA-protein contacts in purified human 17S U2 snRNPs and in spliceosomal A and B complexes.* Mol Cell Biol **26**(7), 2803–2816 Apr (2006).

[179] Gozani, O., Potashkin, J. & Reed, R. *A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site.* Mol Cell Biol **18**(8), 4752–4760 Aug (1998).

[180] Spadaccini, R. *et al. Biochemical and NMR analyses of an SF3b155-p14-U2AF-RNA interaction network involved in branch point definition during pre-mRNA splicing.* RNA **12**(3), 410–425 Mar (2006).

[181] Xu, Y.-Z. *et al. Prp5 bridges U1 and U2 snRNPs and enables stable U2 snRNP association with intron RNA.* EMBO J **23**(2), 376–385 Jan (2004).

[182] Will, C. L. *et al. Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b DEAD-box protein.* EMBO J **21**(18), 4978–4988 Sep (2002).

[183] Perriman, R., Barta, I., Voeltz, G. K., Abelson, J. & Ares, M. *ATP requirement for Prp5p function is determined by Cus2p and the structure of U2 small nuclear RNA.* Proc Natl Acad Sci U S A **100**(24), 13857–13862 Nov (2003).

[184] Fleckner, J., Zhang, M., Valcárcel, J. & Green, M. R. *U2AF65 recruits a novel human DEAD box protein required for the U2 snRNP-branchpoint interaction.* Genes Dev **11**(14), 1864–1872 Jul (1997).

[185] Will, C. L. & Lührmann, R. *Spliceosomal UsnRNP biogenesis, structure and function.* Curr Opin Cell Biol **13**(3), 290–301 Jun (2001).

[186] Liu, S., Rauhut, R., Vornlocher, H.-P. & Lührmann, R. *The network of protein-protein interactions within the human U4/U6.U5 tri-snRNP.* RNA **12**(7), 1418–1430 Jul (2006).

[187] Beggs, J. D. *Lsm proteins and RNA processing.* Biochem Soc Trans **33**(Pt 3), 433–438 Jun (2005).

[188] Laggerbauer, B., Achsel, T. & Lührmann, R. *The human U5-200kD DEXH-box protein unwinds U4/U6 RNA duplices in vitro.* Proc Natl Acad Sci U S A **95**(8), 4188–4192 Apr (1998).

[189] Schaffert, N., Hossbach, M., Heintzmann, R., Achsel, T. & LÃ¼hrmann, R. *RNAi knockdown of hPrp31 leads to an accumulation of U4/U6 di-snRNPs in Cajal bodies.* EMBO J **23**(15), 3000–3009 Aug (2004).

[190] Boehringer, D. *et al. Three-dimensional structure of a pre-catalytic human spliceosomal complex B.* Nat Struct Mol Biol **11**(5), 463–468 May (2004).

[191] Chen, J. Y. *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), 227–232 Jan (2001).

[192] Turner, I. A., Norman, C. M., Churcher, M. J. & Newman, A. J. *Roles of the U5 snRNP in spliceosome dynamics and catalysis.* Biochem Soc Trans **32**(Pt 6), 928–931 Dec (2004).

[193] Small, E. C., Leggett, S. R., Winans, A. A. & Staley, J. P. *The EF-G-like GTPase Snu114p regulates spliceosome dynamics mediated by Brr2p, a DExD/H box ATPase.* Mol Cell **23**(3), 389–399 Aug (2006).

[194] Makarova, O. V. *et al. A subset of human 35S U5 proteins, including Prp19, function prior to catalytic step 1 of splicing.* EMBO J **23**(12), 2381–2391 Jun (2004).

[195] van Nues, R. W. & Beggs, J. D. *Functional contacts with a range of splicing proteins suggest a central role for Brr2p in the dynamic control of the order of events in spliceosomes of Saccharomyces cerevisiae.* Genetics **157**(4), 1451–1467 Apr (2001).

[196] Gottschalk, A., Kastner, B., Lührmann, R. & Fabrizio, P. *The yeast U5 snRNP coisolated with the U1 snRNP has an unexpected protein composition and includes the splicing factor Aar2p.* RNA **7**(11), 1554–1565 Nov (2001).

[197] Villa, T. & Guthrie, C. *The Isy1p component of the NineTeen complex interacts with the ATPase Prp16p to regulate the fidelity of pre-mRNA splicing.* Genes Dev **19**(16), 1894–1904 Aug (2005).

[198] Mayas, R. M., Maita, H. & Staley, J. P. *Exon ligation is proofread by the DExD/H-box ATPase Prp22p.* Nat Struct Mol Biol **13**(6), 482–490 Jun (2006).

[199] Malca, H., Shomron, N. & Ast, G. *The U1 snRNP base pairs with the 5' splice site within a penta-snRNP complex.* Mol Cell Biol **23**(10), 3442–3455 May (2003).

[200] Bentley, D. *The mRNA assembly line: transcription and processing machines in the same factory.* Curr Opin Cell Biol **14**(3), 336–342 Jun (2002).

[201] Park, J. W., Parisky, K., Celotto, A. M., Reenan, R. A. & Graveley, B. R. *Identification of alternative splicing regulators by RNA interference in Drosophila.* Proc Natl Acad Sci U S A **101**(45), 15974–15979 Nov (2004).

[202] House, A. E. & Lynch, K. W. *Regulation of alternative splicing: More than just the ABCs.* J Biol Chem Nov (2007).

[203] von Mering, C. *et al. STRING 7–recent developments in the integration and prediction of protein interactions.* Nucleic Acids Res **35**(Database issue), D358–D362 Jan (2007).

[204] Prieto, C. & Rivas, J. D. L. *APID: Agile Protein Interaction DataAnalyzer.* Nucleic Acids Res **34**(Web Server issue), W298–W302 Jul (2006).

[205] Kerrien, S. *et al. IntAct–open source resource for molecular interaction data.* Nucleic Acids Res **35**(Database issue), D561–D565 Jan (2007).

[206] Zhang, S., Jin, G., Zhang, X.-S. & Chen, L. *Discovering functions and revealing mechanisms at molecular level from biological networks.* Proteomics **7**(16), 2856–2869 Aug (2007).

[207] Ho, Y. *et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.* Nature **415**(6868), 180–183 Jan (2002).

[208] Reddy, V. N., Mavrovouniotis, M. L. & Liebman, M. N. *Petri net representations in metabolic pathways.* Proc Int Conf Intell Syst Mol Biol **1**, 328–336 (1993).

[209] Reddy, V. N., Liebman, M. N. & Mavrovouniotis, M. L. *Qualitative analysis of biochemical reaction systems.* Comput Biol Med **26**(1), 9–24 Jan (1996).

[210] Koch, I., Schueler, M. & Heiner, M. *STEPP–Search Tool for Exploration of Petri net Paths: a new tool for Petri net-based path analysis in biochemical networks.* In Silico Biol **5**(2), 129–137 (2005).

[211] Simão, E., Remy, E., Thieffry, D. & Chaouiya, C. *Qualitative modelling of regulated metabolic pathways: application to the tryptophan biosynthesis in E.coli.* Bioinformatics **21 Suppl 2**, ii190–ii196 Sep (2005).

[212] Hofestädt, R. *Petri Net application of metabolic processes.* J System Analysis, Modeling and Simulation **16**, 113–122 (1994).

[213] Hofestädt, R. & Thelen, S. *Quantitative modeling of biochemical networks.* In Silico Biol **1**(1), 39–53 (1998).

[214] Zevedei-Oancea, I. & Schuster, S. *Topological analysis of metabolic networks based on Petri net theory.* In Silico Biol **3**(3), 323–345 (2003).

[215] Sackmann, A., Heiner, M. & Koch, I. *Application of Petri net based analysis techniques to signal transduction pathways.* BMC Bioinformatics **7**, 482 (2006).

[216] Heiner, M., Koch, I. & Will, J. *Model validation of biological pathways using Petri nets–demonstrated for apoptosis.* Biosystems **75**(1-3), 15–28 Jul (2004).

[217] Matsuno, H., Doi, A., Nagasaki, M. & Miyano, S. *Hybrid Petri net representation of gene regulatory network.* Pac Symp Biocomput , 341–352 (2000).

[218] Marwan, W., Sujatha, A. & Starostzik, C. *Reconstructing the regulatory network controlling commitment and sporulation in Physarum polycephalum based on hierarchical Petri Net modelling and simulation.* J Theor Biol **236**(4), 349–365 Oct (2005).

[219] Matsuno, H., Inouye, S.-I. T., Okitsu, Y., Fujii, Y. & Miyano, S. *A new regulatory interaction suggested by simulations for circadian genetic control mechanism in mammals.* J Bioinform Comput Biol **4**(1), 139–153 Feb (2006).

[220] Schuster, S. & Hilgetag, C. *On elementary flux modes in biochemical reaction systems at steady state.* J Biol Syst **2**, 165–182 (1994).

[221] Schuster, S., Dandekar, T. & Fell, D. A. *Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering.* Trends Biotechnol **17**(2), 53–60 Feb (1999).

[222] Schilling, C. H., Edwards, J. S. & Palsson, B. O. *Toward metabolic phenomics: analysis of genomic data using flux balances.* Biotechnol Prog **15**(3), 288–295 (1999).

[223] Schuster, S., von Kamp, A. & Pachkov, M. *Understanding the roadmap of metabolism by pathway analysis.* Methods Mol Biol **358**, 199–226 (2007).

[224] Papin, J. A. *et al. Comparison of network-based pathway analysis methods.* Trends Biotechnol **22**(8), 400–405 Aug (2004).

[225] Xiong, M., Zhao, J. & Xiong, H. *Network-based regulatory pathways analysis.* Bioinformatics **20**(13), 2056–2066 Sep (2004).

[226] Kielbassa, J., Bortfeldt, R., Schuster, S. & Koch, I. *Modeling of the U1 snRNP assembly pathway in alternative splicing in human cells using Petri nets.* Comput Biol Chem Jul (2008).

[227] Berg, J. M., Tymoczko, J. L. & Stryer, L. *Biochemistry*, volume 5th. W. H. Freeman & Co., New York, NY, (2002).

[228] Pawson, T. *Dynamic control of signaling by modular adaptor proteins.* Curr Opin Cell Biol **19**(2), 112–116 Apr (2007).

[229] Williams, D. H., Maguire, A. J., Tsuzuki, W. & Westwell, M. S. *An analysis of the origins of a cooperative binding energy of dimerization.* Science **280**(5364), 711–714 May (1998).

[230] Fortes, P. *et al. Luc7p, a novel yeast U1 snRNP protein with a role in 5' splice site recognition.* Genes Dev **13**(18), 2425–2438 Sep (1999).

[231] Baumgarten, B. *Petri-Netze, Grundlagen und Anwendungen.* Spektrum Akademischer Verlag, second edition, (1996).

[232] Koch, I. & Heiner, M. *Petri nets in analysis of biological network*, volume 7 of *Wiley Book Series in Bioinformatics.* Wiley, (2008).

[233] Starke, P. *INA - Integrated Net Analyzer Manual, Berlin.* HU Berlin, Dept. of CS , www.informatik.hu–berlin.de/lehrstuehle/automaten/ina/ (1998).

[234] Grafahrend-Belau, E. *et al. Modularization of biochemical networks based on classification of Petri net t-invariants.* BMC Bioinformatics **9**(1), 90 Feb (2008).

[235] Heiner, M. *SNOOPY - Petri net editor and animator.* BTU Cottbus, Dept. of CS , http://www.informatik.tu–cottbus.de/ wwwdssz/ (2004).

[236] Team, R. D. C. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, (2005). ISBN 3-900051-07-0.

[237] Kanehisa, M. *et al. KEGG for linking genomes to life and the environment.* Nucleic Acids Res **36**(Database issue), D480–D484 Jan (2008).

[238] Krull, M. *et al. TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations.* Nucleic Acids Res **34**(Database issue), D546–D551 Jan (2006).

[239] Takai-Igarashi, T. *Ontology based standardization of Petri net modeling for signaling pathways.* In Silico Biol **5**(5-6), 529–536 (2005).

[240] Chaouiya, C. *Petri net modelling of biological networks.* Brief Bioinform **8**(4), 210–219 Jul (2007).

[241] Pandit, S., Lynn, B. & Rymond, B. C. *Inhibition of a spliceosome turnover pathway suppresses splicing defects.* Proc Natl Acad Sci U S A **103**(37), 13700–13705 Sep (2006).

[242] Puig, O., Gottschalk, A., Fabrizio, P. & Séraphin, B. *Interaction of the U1 snRNP with nonconserved intronic sequences affects 5' splice site selection.* Genes Dev **13**(5), 569–580 Mar (1999).

[243] Del Gatto-Konczak, F. *et al. The RNA-binding protein TIA-1 is a novel mammalian splicing regulator acting through intron sequences adjacent to a 5' splice site.* Mol Cell Biol **20**(17), 6287–99 (2000).

[244] MacMillan, A. M., McCaw, P. S., Crispino, J. D. & Sharp, P. A. *SC35-mediated reconstitution of splicing in U2AF-depleted nuclear extract.* Proc Natl Acad Sci U S A **94**(1), 133–136 Jan (1997).

[245] Konarska, M. M. & Query, C. C. *Insights into the mechanisms of splicing: more lessons from the ribosome.* Genes Dev **19**(19), 2255–2260 Oct (2005).

[246] Nottrott, S., Urlaub, H. & Lührmann, R. *Hierarchical, clustered protein interactions with U4/U6 snRNA: a biochemical role for U4/U6 proteins.* EMBO J **21**(20), 5527–5538 Oct (2002).

[247] Xie, J., Beickman, K., Otte, E. & Rymond, B. C. *Progression through the spliceosome cycle requires Prp38p function for U4/U6 snRNA dissociation.* EMBO J **17**(10), 2938–2946 May (1998).

[248] Lybarger, S. *et al. Elevated levels of a U4/U6.U5 snRNP-associated protein, Spp381p, rescue a mutant defective in spliceosome maturation.* Mol Cell Biol **19**(1), 577–584 Jan (1999).

[249] Schwer, B. & Guthrie, C. *PRP16 is an RNA-dependent ATPase that interacts transiently with the spliceosome.* Nature **349**(6309), 494–499 Feb (1991).

[250] Pfeiffer, T., Sánchez-Valdenebro, I., Nuño, J. C., Montero, F. & Schuster, S. *METATOOL: for studying metabolic networks.* Bioinformatics **15**(3), 251–257 Mar (1999).

[251] Schuster, S., Klipp, E. & Marhl, M. *The Predictive Power of Molecular Network Modelling*, volume 3. Springer, discovering biomolecular mechanisms with computational biology edition, (2006).

[252] Pawson, T. & Nash, P. *Protein-protein interactions define specificity in signal transduction.* Genes Dev **14**(9), 1027–1047 May (2000).

[253] Pospisil, H., Herrmann, A., Pankow, H. & Reich, J. G. *A database on alternative splice forms on the integrated genetic map service (IGMS).* In Silico Biol **3**(1-2), 229–234 (2003).

[254] Thanaraj, T. A., Clark, F. & Muilu, J. *Conservation of human alternative splice events in mouse.* Nucleic Acids Res **31**(10), 2544–2552 May (2003).

[255] Kato, M., Hata, N., Banerjee, N., Futcher, B. & Zhang, M. Q. *Identifying combinatorial regulation of transcription factors and binding motifs.* Genome Biol **5**(8), R56 (2004).

[256] Wu, T. D. & Watanabe, C. K. *GMAP: a genomic mapping and alignment program for mRNA and EST sequences.* Bioinformatics **21**(9), 1859–1875 May (2005).

[257] Ermakova, E. O., Nurtdinov, R. N. & Gelfand, M. S. *Overlapping alternative donor splice sites in the human genome.* J Bioinform Comput Biol **5**(5), 991–1004 Oct (2007).

[258] Buratti, E. *et al. Aberrant 5' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization.* Nucleic Acids Res **35**(13), 4250–4263 (2007).

[259] Santisteban, I. *et al. Novel splicing, missense, and deletion mutations in seven adenosine deaminase-deficient patients with late/delayed onset of combined immunodeficiency disease. Contribution of genotype to phenotype.* J Clin Invest **92**(5), 2291–2302 Nov (1993).

[260] Hanamura, A., Cáceres, J. F., Mayeda, A., Franza, B. R. & Krainer, A. R. *Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors.* RNA **4**(4), 430–444 Apr (1998).

[261] Glatz, D. C. *et al. The alternative splicing of tau exon 10 and its regulatory proteins CLK2 and TRA2-BETA1 changes in sporadic Alzheimer's disease.* J Neurochem **96**(3), 635–644 Feb (2006). hard copy available.

[262] Louis, M., Holm, L., Snchez, L. & Kaufman, M. *A theoretical model for the regulation of Sex-lethal, a gene that controls sex determination and dosage compensation in Drosophila melanogaster.* Genetics **165**(3), 1355–1384 Nov (2003). hard copy available.

[263] Hucka, M. *et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.* Bioinformatics **19**(4), 524–531 Mar (2003).

[264] Zhang, X. H.-F., Kangsamaksin, T., Chao, M. S. P., Banerjee, J. K. & Chasin, L. A. *Exon inclusion is dependent on predictable exonic splicing enhancers.* Mol Cell Biol **25**(16), 7323–7332 Aug (2005).

[265] Rosbash, M. & Séraphin, B. *Who's on first? The U1 snRNP-5' splice site interaction and splicing.* Trends Biochem Sci **16**(5), 187–190 May (1991).

[266] Xiao, S. H. & Manley, J. L. *Phosphorylation-dephosphorylation differentially affects activities of splicing factor ASF/SF2.* EMBO J **17**(21), 6359–6367 Nov (1998).

[267] Ma, C.-T. *et al. Ordered multi-site phosphorylation of the splicing factor ASF/SF2 by SRPK1.* J Mol Biol **376**(1), 55–68 Feb (2008).

[268] Wu, J. Y. & Maniatis, T. *Specific interactions between proteins implicated in splice site selection and regulated alternative splicing.* Cell **75**(6), 1061–1070 Dec (1993).

[269] Puig, O., Bragado-Nilsson, E., Koski, T. & Séraphin, B. *The U1 snRNP-associated factor Luc7p affects 5' splice site selection in yeast and human.* Nucleic Acids Res **35**(17), 5874–5885 (2007).

[270] Wu, S., Romfo, C. M., Nilsen, T. W. & Green, M. R. *Functional recognition of the 3' splice site AG by the splicing factor U2AF35.* Nature **402**(6763), 832–835 Dec (1999).

[271] Kent, O. A., Ritchie, D. B. & Macmillan, A. M. *Characterization of a U2AF-independent commitment complex (E') in the mammalian spliceosome assembly pathway.* Mol Cell Biol **25**(1), 233–240 Jan (2005).

[272] Staknis, D. & Reed, R. *SR proteins promote the first specific recognition of Pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex.* Mol Cell Biol **14**(11), 7670–82 (1994).

[273] Sharma, S., Falick, A. M. & Blacks, D. L. *Polypyrimidine tract binding protein blocks the 5' splice site-dependent assembly of U2AF and the prespliceosomal E complex.* Mol Cell **19**(4), 485–496 Aug (2005).

[274] Reed, R. *Mechanisms of fidelity in pre-mRNA splicing.* Curr Opin Cell Biol **12**(3), 340–345 Jun (2000).

[275] Krämer, A. *et al. Structure-function analysis of the U2 snRNP-associated splicing factor SF3a.* Biochem Soc Trans **33**(Pt 3), 439–442 Jun (2005).

[276] Medenbach, J., Schreiner, S., Liu, S., Lührmann, R. & Bindereif, A. *Human U4/U6 snRNP recycling factor p110: mutational analysis reveals the function of the tetratricopeptide repeat domain in recycling.* Mol Cell Biol **24**(17), 7392–7401 Sep (2004).

[277] Teigelkamp, S. *et al. The 20kD protein of human [U4/U6.U5] tri-snRNPs is a novel cyclophilin that forms a complex with the U4/U6-specific 60kD and 90kD proteins.* RNA **4**(2), 127–141 Feb (1998).

[278] Kuhn, A. N., Reichl, E. M. & Brow, D. A. *Distinct domains of splicing factor Prp8 mediate different aspects of spliceosome activation.* Proc Natl Acad Sci U S A **99**(14), 9145–9149 Jul (2002).

[279] Brenner, T. J. & Guthrie, C. *Assembly of Snu114 into U5 snRNP requires Prp8 and a functional GTPase domain.* RNA **12**(5), 862–871 May (2006).

[280] Raghunathan, P. L. & Guthrie, C. *RNA unwinding in U4/U6 snRNPs requires ATP hydrolysis and the DEIH-box splicing factor Brr2.* Curr Biol **8**(15), 847–855 Jul (1998).

[281] Chung, S., McLean, M. R. & Rymond, B. C. *Yeast ortholog of the Drosophila crooked neck protein promotes spliceosome assembly through stable U4/U6.U5 snRNP addition.* RNA **5**(8), 1042–1054 Aug (1999).

[282] Chen, C.-H. *et al. Functional and physical interactions between components of the Prp19p-associated complex.* Nucleic Acids Res **30**(4), 1029–1037 Feb (2002).

[283] Chan, S.-P., Kao, D.-I., Tsai, W.-Y. & Cheng, S.-C. *The Prp19p-associated complex in spliceosome activation.* Science **302**(5643), 279–282 Oct (2003).

[284] Query, C. C. & Konarska, M. M. *Splicing fidelity revisited.* Nat Struct Mol Biol **13**(6), 472–474 Jun (2006).

[285] Zhou, Z. & Reed, R. *Human homologs of yeast prp16 and prp17 reveal conservation of the mechanism for catalytic step II of pre-mRNA splicing.* EMBO J **17**(7), 2095–2106 Apr (1998).

[286] Schwer, B. & Gross, C. H. *Prp22, a DExH-box RNA helicase, plays two distinct roles in yeast pre-mRNA splicing.* EMBO J **17**(7), 2086–2094 Apr (1998).

[287] Aronova, A., Bacíková, D., Crotti, L. B., Horowitz, D. S. & Schwer, B. *Functional interactions between Prp8, Prp18, Slu7, and U5 snRNA during the second step of pre-mRNA splicing.* RNA **13**(9), 1437–1444 Sep (2007).

[288] James, S.-A., Turner, W. & Schwer, B. *How Slu7 and Prp18 cooperate in the second step of yeast pre-mRNA splicing.* RNA **8**(8), 1068–1077 Aug (2002).

Danksagung

Das Material der vorliegenden Arbeit wurde zwischen Oktober 2003 und März 2008 am Lehrstuhl für Bioinformatik der biologisch-parmazeutischen Fakultätin der Friedrich-Schiller Universität Jena erstellt. Mein erster Dank geht daher an Professor Stefan Schuster, der mir in dieser Zeit sehr gute Arbeitsbedingungen ermöglicht hat, stets mit Rat und Tat zur Seite stand und mir mit seinem mathematischen Verständnis der Biologie eine unerschöpfliche Quelle neuer Sichtweisen bot.

Zu diesem Arbeitsumfeld gehörten auch meine engsten Kollegen Beate Knoke, Jörn Behre, Dimitar Kenanov, Swetlana Nikolajewa, Kathrin Schowtka und die Doktoren Ina Weiß, Axel von Kamp, Michael Pachkov und Anja Schröter denen ich für die angenehme Arbeitsatmosphäre danken möchte. Beate Knoke und Doktor Heike Pospisil gilt hierbei mein besonderer Dank für das Korrekturlesen und Feedback zum ersten Kapitel. Meinen Kollegen vom FLI Jena, Stefanie Schindler, Karol Szafranski und Matthias Platzer möchte für die gute Zusammenarbeit und die experimentelle Unterstützung, sowie für die interessanten Diskussionrunden danken.

Meinen wärmsten Dank für viele fachlich interessante Diskussionen und die Eröffnung zahlreicher neuer Perspektiven in Bezug auf meine Arbeit möchte ich Doktor Dirk Holste in Wien ausprechen. Von ihm konnte ich viel zum Alternativen Spleißen und der systematischen Analyse desselben lernen.

Seit vielen Jahren begleitet mich Prof. Ina Koch von der TFH Berlin in der Bioniformatik. Für ihr jederzeit offenes Ohr und die vielen wertvollen Hinweise zur Petri Netz Theorie im dritten Kapitel, möchte ich Ihr meinen herzlichsten Dank aussprechen.

Wenn ich etwas Wertvolles aus Jena mitnehme, so ist es die Liebe einer jungen Frau. Ihre Besonnenheit und Rationalität hat mich vor mancher Verzeiflung gerettet und ohne ihren motivierenden Zuspruch wäre diese Arbeit gegen meine eigenen kritischen Ansprüche nicht zum Abschluß gekommen. Für ihre Liebe und aufopferungsvolle Unterstützung, nicht nur in Form geduldigen Korrekturlesens und reflektierender Diskussionen, bin ich meinem Schatz Berit zu tiefstem Dank verpflichtet.

Allen meinen Freunden danke ich dafür, dass sie trotz der wenigen Zeit, die in den Jahren der Arbeit an der Dissertation blieb, zu mir gehalten und mich hin wieder ermahnt haben, das normale Leben nicht aus den Augen zu verlieren.

Meinen Eltern, die mir das Studium der Biotechnologie und Bioinformatik ermöglichten, aber auch meiner engsten Familie, möchte ich meinen abschließenden Dank widmen -für alle Geduld, Liebe und bedingunslose Unterstützung, für die Möglichkeit in Bergfelde hin und wieder „abschalten“ zu können und für den unerschütterlichen Glauben an eine Zeit nach der Promotion.

Appendix A

Supplements to Chapter 1

A.1 The EASED Database Scheme

Table A.1: Description of tables within the EASED database that where used for the validation analyses. Column two summarizes the main aspects of the data each table holds and column three lists the relations each table shares with other tables. A respective diagram is given in Figure A.1

Table Name	Description	Relations (to other tables)
genes	basic description of each gene: ENSEMBL ID, chromosome and position therein, genomic sequence (slice), number of assigned ENSEMBL transcripts	genes_features (1:n), trans (1:n)
genes_features	cross referencing gene identifier to other databases	genes (n:1)
trans	basic description of ENSEMBL transcripts matching a gene: transcript sequence, CDS start (end) within genomic sequence slice, CDS start (end) within transcript sequence, strand orientation, number of ss pairs (introns), number of matching alternative ESTs,number of matching constitutive ESTs	genes (n:1), exons (1:n), protein_features (1:n), trans_features (1:n), asfs_unique (1:n), hsps (1:n), as_hsps (1:n), trans_tissue (1:n)
trans_features	cross referencing transcript identifier to other databases	trans (n:1)
trans_tissue	referencing numbers of alternative (constitutive) ESTs grouped by tissue types that coincide with an ENSEMBL transcript	trans (n:1), est_tissue (m:1)
protein_features	cross references to functional domain databases, annotation of domain start (end) in the protein sequence	trans (n:1)
exons	basic description of ENSEMBL exons making up a complete transcript: start (end) witin genomic sequence, exon end positions within CDS	
asfs_unique	basic description of AS events: type classification *a-h* (see Figure 1.2), length, quality, position of donor (acceptor) site within transcript and genomic sequence, donor (acceptor) dinucleotides, number of further alternative (constitutive) ESTs coinciding with the AS event	trans (n:1), asfs_tissue (1:n), asfs_ce (1:n), asfs_ae (1:n)
asfs_tissue	referencing numbers of alternative (constitutive) ESTs grouped by tissue types that coincide with an as event	asfs_ae (n:1), est_tissue (m:1)
asfs_ae	extended description of AS events: EST ID, type (insert, skip), quality, start (end) of alternative part within EST sequence (skip/insert length), strand orientation, average EST identity, overlap, reference to the flanking hsp's	asfs_unique (n:1), hsps, asfs_tissue (1:n), ests (n:1)
asfs_ce	extended description of reference splice events: EST ID, quality, start (end) of constitutive part within EST sequence, spanning the alternative part (hence representing the exonic part of the ENSEMBL transcript	asfs_unique (n:1), hsps, asfs_tissue (1:n), ests (n:1)
as_hsps	as table *asfs_ae* but with referene to ENSEMBL transcript ID	trans (n:1), ests (n:1)
hsps	basic description of EST:mRNA (ENSEMBL transcript) alignments: identity, mismatches, gaps, start (end) position within query (EST) sequence, start (end) within subject (mRNA) sequence, evalue, bitscore	trans (n:1), ests (n:1), asfs_ae
ests_tissue	description of available tissue type	asfs_tissue (1:m), trans_tissue (1:m)
ests_library	besic description of EST libraries: UNIGENE library ID, library name- and description, reference to tissue ID, developmental stage- and disease status annotation	ests (1:n), ests_tissue (1:1)

Figure A.1: Entity relationship diagram of the EASED database (v19̇34a). Gene- and Transcript IDs are based on the Ensembl freeze of December 2003

Appendix B

Supplements to Chapter 2

B.1 Sim4 Alignment Errors

Table B.1: A general problem of aligning cDNAs and ESTs against genomic DNA is to correctly map transcript sequences. Posttranscriptional modifications and in vitro transcription errors (1-4% in ESTs (256)) may affect alignment programs and to output genomic coordinates of erroneously aligned transcript blocks. Subsequently, this may introduce a bias in exon datasets derived from such alignment approaches. Generally, *Sim4* is a very fast and capable alignment algorithm that has been successfully used for almost a decade (91). However, a recent comparison of alignment programs revealed apparent weaknesses in the correct alignment of mismatches or insertions/deletion near the splice junction (264). Splicing analysis is crucially depending on correct alignments of transcribed sequences, especially near alternative splice sites: Thus, the two best-performing programs according to a report by Zhang and Gish (264) (*EXALIN* and *BLAT*) were applied for validating initial *SIM4* alignments. In order to demonstrate the filtering capabilities of *EXALIN*, the following examples of an erroneous and correct A5EΔ4 splice site prediction are shown for illustration below

(a) **Erroneous SIM4 alignment:**Filtering alignments for canonical /GT and AG/ splice sites, as frequently applied to prepare data sets of alternatively spliced exons, is not sufficient to effectively remove false-positive alignments. (Incorrect aligned nucleotides are marked in red.

DISTAL SPLICE SITE		
Sequence ID	**SIM4 (True positives)**	**EXALIN (True positives)**
BI037972	CAGAAGCCAAAATG AGGTTGAAGGCTGC	CAGAAGCCAAAATG AGGTTGAAGGCTGC
	|||||||||||||>>> >>>||||||||||||||	|||||||||||||<<<<<< <<<<<<||||||||||||||
ENSG146592	CAGAAGCCAAAATGGTA CAGAGGTTGAAGGCTGC	CAGAAGCCAAAATGGTAAGT TTTCAGAGGTTGAAGGCTGC
PROXIMAL SPLICE SITE		
Sequence ID	**SIM4 (False positives)**	**EXALIN (True negatives)**
BQ367677	AAATGAGTTGGAAG GCTGCATTGACTCA	CAGAAGCCAAAATG A-GTTGGAAGGCTG
	|||||-|-|--||->>> >>>|||||||||||||||	|||||||||||||<<<<<< <<<<<<| ||| ||||||||
ENSG146592	AAATG G T AA GTA AAGGCTGCATTGACTCA	CAGAAGCCAAAATGGTAAGT TTTCAGAGGTT-GAAGGCTG

(b) **Correct SIM4 alignment:** The Δ4 exon extension is marked in bold letters.

DISTAL SPLICE SITE		
Sequence ID	**SIM4 (True positives)**	**EXALIN (True positives)**
BF871212	ACTTCATCCAGTCG GAAGAGAAGATGGA	ACTTCATCCAGTCG GAAGAGAAGATGGA
	|||||||||||||>>> >>>|||||||||||||||	|||||||||||||<<<<<< <<<<<<|||||||||||||||
ENSG102878	ACTTCATCCAGTCGGTA AAGGAAGAGAAGATGGA	ACTTCATCCAGTCGGTAGGT TAAAAGGAAGAGAAGATGGA
PROXIMAL SPLICE SITE		
Sequence ID	**SIM4 (True positives)**	**EXALIN (True positives)**
AW841572	CATCCAGTCGGTAG GAAGAGAAGATGGA	CATCCAGTCGGTAG GAAGAGAAGATGGA
	|||||||||||||>>> >>>|||||||||||||||	|||||||||||||>>>>>> >>>>>>||||||||||||||
ENSG102878	CATCCAGTCGGTAGGTT AAGGAAGAGAAGATGGA	CATCCAGTCGGTAGGTTTGT TAAAAGGAAGAGAAGATGGA

B.2 A5EΔ4 Splice Site Scores in *M. musculus*

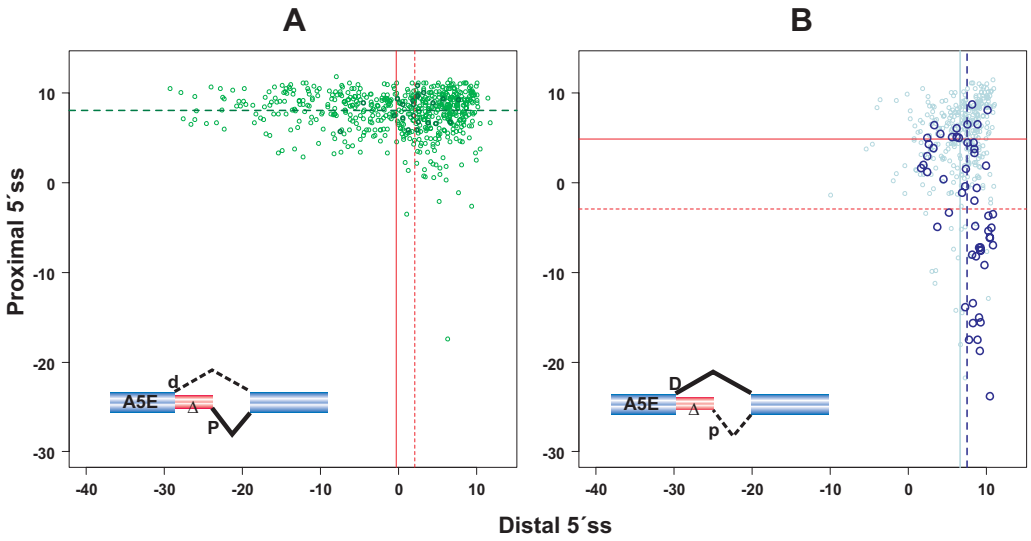
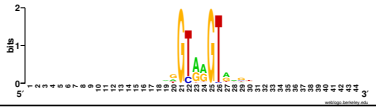
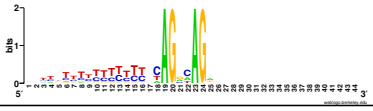
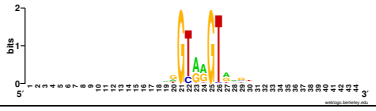
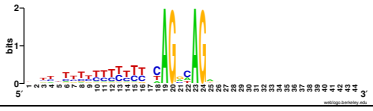
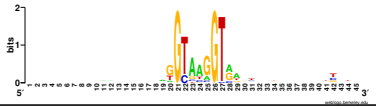
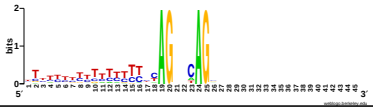
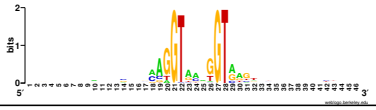
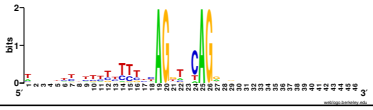
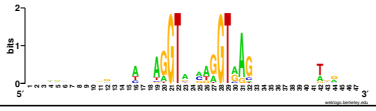
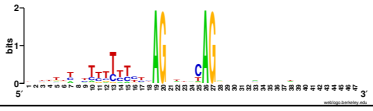
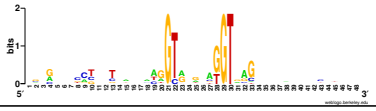
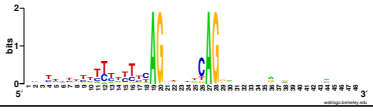
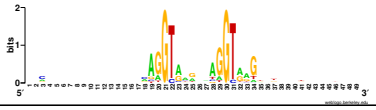
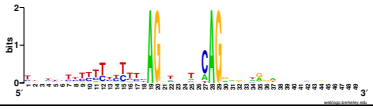
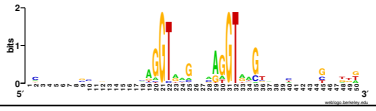
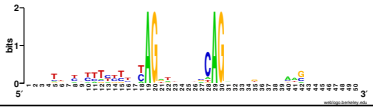
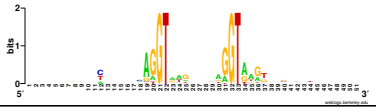
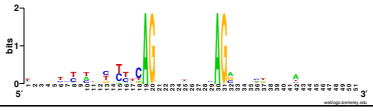
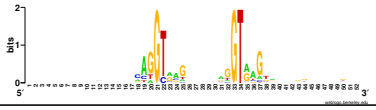
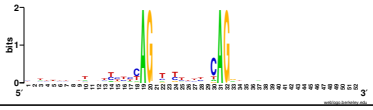
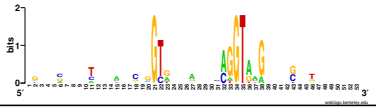
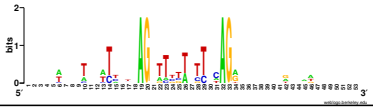


Figure B.1: Scatter plots of 5′ss scores of competitive and tandem donors extracted from mouse *M.musculus*. **(A)** and **(B)** show the individual and mean scores (the latter is marked by solid/dashed lines) for PΔ4 and DΔ4 splicing exons, respectively. Scattered data and mean scores are in accord with human PΔ4 and DΔ4 splicing exons.

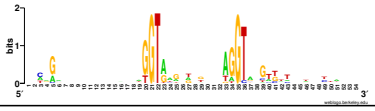
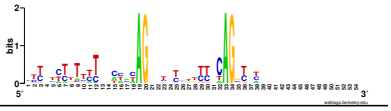
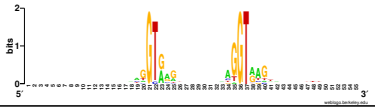
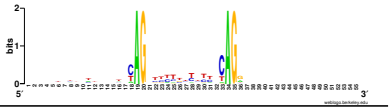
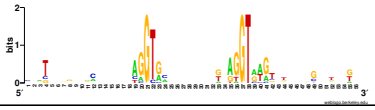
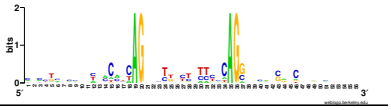
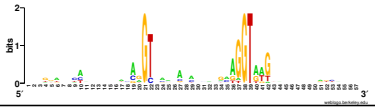
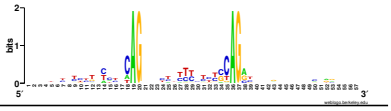
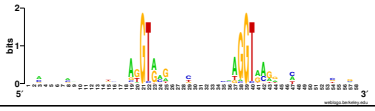
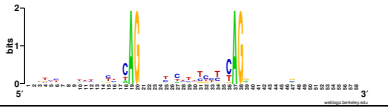
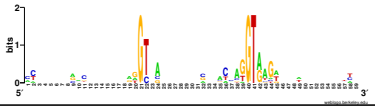
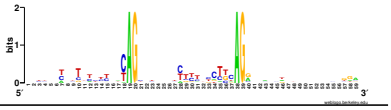
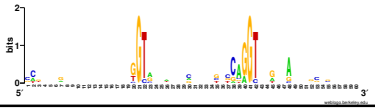
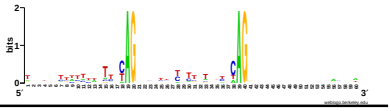
B.3 Base Composition of Tandem Splice Sites

Table B.2: WEBLOGO (95) representations of human A5Es and A3Es with exon extensions $E = 3, 4, \cdots, 20$ nucleotides, in a sequence windows 20nt up- and downstream of the alternative sequence region. The middle column shows for A5Es the high degeneracy of splice sites, while the last column shows for A3Es the degeneracy of the pyrimidine-tract that is increasingly less present with increasing extension for the proximal 3′ss, and more present for $E > 12$ nucleotides for the distal 3′ss.

Exon Extension [nt]	Alternative 5′ Splice Site Exons	Alternative 3′ Splice site Exons
3nt		
4nt		
5nt		
6nt		
7nt		
8nt		
9nt		
10nt		
11nt		
12nt		
13nt		

continued next page

Table B.2: WEBLOGO (95) representations of human A5Es and A3Es with exon extensions $E = 3, 4, \cdots, 20$ nucleotides, in a sequence windows 20nt up- and downstream of the alternative sequence region. The middle column shows for A5Es the high degeneracy of splice sites, while the last column shows for A3Es the degeneracy of the pyrimidine-tract that is increasingly less present with increasing extension for the proximal 3′ss, and more present for $E > 12$ nucleotides for the distal 3′ss.

Exon Extension [nt]	Alternative 5′ Splice Site Exons	Alternative 3′ Splice site Exons
14nt		
15nt		
16nt		
17nt		
18nt		
19nt		
20nt		

Appendix C

Supplements to Chapter 3

C.1 Description of Reactions of the Spliceosomal Model Network

Table C.1: Reactions (transitions) involved in spliceosome assembly (act = activated, ass = assigned, bdg = binding, conv = converted, dim = dimerized, hydrol = hydolized, inhib = inhibited, matur = maturated, phosph = phosphorylated, stab = szabilized, stl = stemloop, uwd = unwound, tri-snRNP = U4/U5/U6 subcomplex)

Description	ID	**Label**	**Reaction**	**References**
E Complex Reactions				
Assembly of U1 specific factors with U1 core components	**1**	*t136.U1_matur*	U1C + U1A + U170K + U1snRNA → U1	(185)
Base pairing between U1 5′ end of the U1 snRNA and strong intron 5′ss (consensus donor) via U1C:5′ ss contacts	**2**	*t41.U1_5ss_bdg*	U1 + 5ss → U1_5ss	(265, 168)
Interaction of phosphorylated ASF/SF2 (regiospecific phosporylation by SRPK1) with U170K enhances recognition of weak 5′ss. Also TIA1 promotes binding of U1 snRNP to weak 5′ ss which may happen in concert or independent of ASF/SF2.	**3**	*t143.ASF_phosph*	ASF_SF2 + ATP + SRPK1 → ASFp + SRPK1 + ADP	(266, 267)
	4	*t141.ASFp_ex_bdg*	ASFp + 5ss_ex → ASFp_ex	(86, 87)
	5	*t12.ASFp_U170K_bdg*	ASFp_ex + U1 + 5ss → ASFp_U1_5ss	(268, 169)
	6	*t9.U1C_TIA1_bdg1*	ASFp_U1_5ss + TIA1_int → U1_5ss	(242, 243, 170)
LUC7 links Cap-Binding-Complex (CBC) to the U1 snRNP via interaction with U2AF65 when splicing the 5′terminal exon	**7**	*t137.U1_CBC_bdg*	CBC + U1 + LUC7 + U2AF → U1_CBC	(230)
	8	*t10.U1_CBC_5ss_bdg*	U1_CBC + 5ss → U1_5ss	(269)
TIA1 binds near 5′ ss and stabilizes U1 snRNP recruitment via direct interaction with U1C when splicing weak 5′ss or in absence of cap on 5′ exon	**9**	*t131.TIA1_5ss_int_bdg*	TIA1 + 5ss_int → TIA1_int	(242, 243)
	10	*t30.U1C_TIA1_bdg2*	U1 + 5ss + TIA1_int → U1_5ss	(170)
U2AF65 dimerizes with U2AF35. U2AF35 recognizes and binds 3′ss (particular for weak PPT). Cooperative binding of U2AF65 to the PPT and SF1(BBP) to the branch point. Interaction between SF1 and U2AF65 after U2AF35 recognized the 3′ss	**11**	*t6.U2AF_dim*	U2AF65 + U2AF35 → U2AF	(270)
	12	*t0.U2AF35_3ss_bdg*	U2AF + 3ss + PPT → U2AF_PPT_3ss	(165, 174, 270)
	13	*t7.SF1_BPS_bdg*	SF1 + BPS → SF1_BPS	
	14	*t14.SF1_U2AF_bdg*	SF1_BPS + U2AF_PPT_3ss → SF1_U2AF	(163, 174)
U1snRNP defines 5′ss and U2AF65 binding to PPT, U2AF65 is anchored as heterodimer via U2AF35 at the 3′ss	**15**	*t59.U1_SF1_U2AF_bdg*	U1_SF1 + U2AF_PPT_3ss → U1_SF1_U2AF	(271)
FBP11/Prp40 bridges U1 and SF1_U2AF and establishes cooperativity between 5′ and 3′ss recognition. SC35 can also be involved in bridging 5′ and 3′ss via interactions of U170K and U2AF35	**16**	*t58.U1_SF1_bdg*	U1_5ss + SF1_BPS + FBP11 → U1_SF1 (E complex)	(175, 271)
	17	*t17.U170K_U2AF35_bdg*	U1_5ss + SC35 + SF1_U2AF + FBP11 → U1_SF2_U2AF	(268, 272)

continued next page

Table C.1: Reactions (transitions) involved in spliceosome assembly (act = activated, ass = assigned, bdg = binding, conv = converted, dim = dimerized, hydrol = hydolized, inhib = inhibited, matur = maturated, phosph = phosphorylated, stab = szabilized, stl = stemloop, uwd = unwound, tri-snRNP = U4/U5/U6 subcomplex)

Description	ID	**Label**	**Reaction**	**References**
U1 and U2 become ATP-independent bridged by DEAD-box protein Prp5. Subsequently U1 interacts with 5′ss and U2 with U2AF. Within the bridged complex U1/SF1 interactions still occur.	**18**	*t27.U1_Prp5_U2_bdg*	U1 + U2 + Prp5 → U1_Prp5_U2	(181)
	19	*t28.U1_5ss_U2_U2AF_bdg*	5ss + U1_Prp5_U2 + SF1_U2AF → U1U2_bridge	
SC35 activates 5′ss, independent of functional U1 snRNP	**20**	*t16.U1_indep_5ss_act*	SC35 + 5ss +SF1_U2AF → U2AF_SC35_5ss	(171, 173, 244)
PTB binds to PPT, outcompeting U2AF and preventing E complex assembly and, thus, splicing of tissue specific exons (e.g. c-src exon N1)	**21**	*t61.PTB_PTT_bdg*	PTB + PPT + U1_5ss → PTB_PPT	(6, 273)
	22	*t62.E_complex_inhib*	PTB_PPT → U1_5ss + U1_5ss_block	
A Complex Reactions				
Assembly of U2 subunit SF3b (components shared with U12 snRNP)	**23**	*t111.SF3b_dim*	SF3b10 + SF3b14a + SF3b14b + SF3b49 + SF3b125 + SF3b130 + SF3b145 + SF3b155 + ATP → SF3b + ADP	(274, 182)
Binding of subunit SF3b to core U2 snRNP	**24**	*t21.15S_U2_matur*	12S_U2_core + SF3b + U2A + U2B → 15S_U2	(275)
Dissociation of SF3b DEAD-box protein SF3b125	**25**	*t13.17S_U2_matur2*	15S_U2 + SF3a → SF3b125 + U2	(182)
Assembly of U2 subunit SF3a	**26**	*t104.SF3a_dim*	SF3a60 + SF3a66 + SF3a120 → SF3a	(275)
Binding of subunit SF3a to 15S U2 snRNP	**27**	*t22.17S_U2_matur1*	15S_U2 + SF3a + ATP + SF3b125 → U2 +ADP	(182)
U2 snRNP component SF3b155 binds to both sides of the BPS and interacts with U2AF65, after U2AF binding to the PPT. hPrp43 was found in 17S U2 complex and proposed to facilitate U2 formation by associating with pre-mRNA as part of the 17S U2 snRNP. hPrp43 was found to be activated in post-spliceosomal complex formation.	**28**	*t8.SF3b155_U2AF65_bdg*	U2AF_PPT_3ss + U2 +hPrp43 → U2_PPT	(179, 182)
Prp5 replaces Prp9 (SF3a60) within U2 snRNP subcomplex and promotes opening of the U2 snRNA stemloop IIa, thus increasing U2 sensitivity for the branch point sequence (BPS)	**29**	*t23.unwind2_U2_stl2*	Prp5 + ATP + U2_PPT → ADP + U2_remod1 + SF3a60	(165, 183, 181)
Specific activation of ATPase UAP56 by U2AF65	**30**	*t57.UAP56_U2AF65_bdg1*	UAP56 +U2_remod1 → U2_UAP56_ass	(184, 165)
	31	*t25.UAP56_U2AF65_bdg2*	UAP56 +U1U2_remod → U2_UAP56_ass	

continued next page

Table C.1: Reactions (transitions) involved in spliceosome assembly (act = activated, ass = assigned, bdg = binding, conv = converted, dim = dimerized, hydrol = hydolized, inhib = inhibited, matur = maturated, phosph = phosphorylated, stab = szabilized, stl = stemloop, uwd = unwound, tri-snRNP = U4/U5/U6 subcomplex)

Description	ID	Label	Reaction	References
Action of Prp5 within bridged U1U2 complex. UAP56 catalyzes stable transition from E to A complex by release of SF1 and binding of strutural rearranged U2 to the branch point	**32**	*t55.unwind1_U2_stl2*	U1U2_bridge + ATP → SF3a60 + ADP + U1U2_remod	(165, 181)
	33	*t29.U1U2_BPS_bdg*	U1U2_UAP56_ass + ATP → SF3a60 + SF1 + U2AF + ADP + U1U2_BPS (A complex)	
ATP dependent (DEAD box helicase/unwindase action) release of SF1 from BPS and binding of U2 to the branch point sequence	**34**	*t24.U2_BPS_bdg1*	U2_UAP56_ass + ATP + U1_SF1_U2AF + → SF1 + U2AF + ADP + U1U2_BPS (A complex)	(165, 163),p345
	35	*t31.U2_BPS_bdg2*	U2AF_SC35_5ss + U2_UAP56_ass → SF1 + U2AF + U2_BPS	
U2AF independent but U1 dependent 5ss recognition (SC35 und Prp5 supported)	**36**	*t56.U1_SC35_SF1_bdg*	SC35 + U1_5ss + SF1_BPS → U1_SC35_SF1	(244)
	37	*t152.U2_SC35_bdg*	U1_SC35_SF1 + U2 → U1_SC35_U2	
	38	*t35.unwind3_U2_stl2*	Prp5 + ATP + U1_SC35_U2 →SF3a60 + ADP + U2_remod2	
	39	*t155.U2_BPS_bdg3*	UAP56 + ATP + U2_remod2 → U1U2_BPS	
B Complex Reactions				
In presence of Snu13(15.5K), Prp31(61K) interacts with U4 and U4/U6 snRNA duplex, but not with U6 snRNA alone. Sm proteins form the U4 snRNP core, while Sm like proteins (Lsm) form the core of U6 snRNP	**40**	*t147.U4_matur*	U4snRNA + Sm + Snu13 → U4	(246, 186)
	41	*150.U6_matur*	U6snRNA + Lsm → U6	(187)
Prp24_ass anneals U4 and U6 snRNA to form snRNA duplex within the U4/U6 snRNP subcomplex. U4 or U4/U6 binds Prp31 in the presence of Snu13	**42**	*t47.U4_U6_bdg*	U6 + U4 + Prp31 + Prp24 → U4U6_complex	(165, 189, 276)
Prp3 and Prp4 interact directly with each other; CypH, Prp4 and Prp3 form a stable RNA free trimeric subcomplex.	**43**	*t127.Prp3_Prp4_bdg*	Prp3+ Prp4 → P3P4_dim	(277, 246, 186)
	44	*t126.CypH_P3P4dim_bdg*	P3P4_dim + CypH → CypH_trimer	
Prp8, Brr2 and Snu114p interact directly with each other. Prp8 interacts also with Snu114 (human ortholog = U5.40K). Brr2p, and Prp8p copurify, along with Snu114 as an RNA-free heterotetrameric complex. U5 snRNA associates with several U5 specific proteins including the DExD/H box helicase Prp28.	**45**	*t118.Prp8_trimer*	U5_Prp8 + U5_Snu114 +U5_Brr2 → pre_U5_trimer	(193, 186)
	46	*t119.U5_aux_bdg*	pre_U5_trimer + U5_40K + Prp28 → pre_U5_heteromer	

continued next page

Table C.1: Reactions (transitions) involved in spliceosome assembly (act = activated, ass = assigned, bdg = binding, conv = converted, dim = dimerized, hydrol = hydolized, inhib = inhibited, matur = maturated, phosph = phosphorylated, stab = szabilized, stl = stemloop, uwd = unwound, tri-snRNP = U4/U5/U6 subcomplex)

Description	ID	Label	Reaction	References
	47	*t123.U5_matur*	Sm + U5snRNA + hDib1 +hLin1 + Prp28 + U5_aux_tetramer → 20S_U5	
Prp3 (CypH_trimer) interacts with Prp24 (recycling factor) prior to U4/U6 snRNA unwinding, specifying Prp24. Prp6 function as bridging factor between human U5 snRNP and U4/U6 snRNPs. Prp38 is suggested to promote a conformational change within the spliceosome needed to expose the U4/U6 helices, for later release of U4.	**48**	*t53.U4U6_U5_bdg*	Prp38 + hPrp6 + CypH_trimer + 20S_U5 → U4U5U6_conf1	(247, 276, 186)
Snu66 (U5.110K) interacts with hPrp3, hPrp6 and hBrr2, contributing to bridging the snRNPs in the tri-snRNP	**49**	*t48.U4U6U5_stab*	U4U6U5_conf1 + hSad1 + hSnu66 + tris_27K → hLin1 + U4U6U5_conf2	(186)
Specific activation of DExD/H box helicase Brr2 by Snu114 and GTP hydrolyzation, preparing unwinding of U4/U6 duplex	**50**	*t46.Snu114_Brr2_act*	U4U6U5_conf2 + GTP → Brr2_ass + GDP	(165, 192, 193)
Activated (specifically directed) helicase Brr2 catalyzes unwinding of U4/U6 stemloop II, involving Prp8. Prp19 is required for tri-snRNP formation (suggested regulatory role in U4/U6 unwinding)	**51**	*t52.U4U6_uwd*	Brr2_ass + Prp19 + ATP → CypH_trimer + U4U6U5_conf3 + ADP	(278, 151, 194)
Prp28 destabilizes U1C, preparing U1 dissociation. Prp5 is not required anymore (see text). Additionally, Prp8 governs activities of the kinases Prp28 (and Brr2).	**52**	*t34.U1C_diss*	Prp28 + U1U2_BPS + ATP + U5_Prp8 → U1C + ADP + U1_5ss_destab + Prp5	(191, 192, 279)
Interaction of Prp19 with FBP11 and U2AF65 (both present in A complex) stabilizes tri-snRNP addition, while U1 and U4 are released from tri-snRNP complex. Dib1 (U5.15K) and Prp28 (U5.100K) are absent in B-complexes after U4/U6 dissociation. This step is assumed to be similar in the U1 independent spliceosomal assembly pathway.	**53**	*t33.U6_5ss_bdg1*	U4U6U5_conf3 (Prp19 present)+ U1_5ss_destab (FBP11, U1_SF1_U2AF present) → U2U6U5_5ss + hDib1 + Prp28 + Prp38 +U1 + U4	(280, 281, 195, 163, 282)
	54	*t32.U6_5ss_bdg2*	U4U6U5_conf3 + U2_BPS → U2U6U5_5ss + hDib1 + Prp28 + Prp38 + U4	(142)
In human, Prp19 is stably associated with several proteins, forming the heteromeric NTC complex	**55**	*t65.Prp19_stab*	NTC_heteromer + Prp19 → 14S_NTC_Prp19	(194)
NTC complex subsequently acts to U4 dissociation stabilizing association of U5 and U6 with the activated spliceosome. NTC destabilizes U6-associated Lsm proteins, but promotes U6:5′ss interactions during activation for 1st step catalysis; Prp31 and Prp3 are destabilized during the catalytic activation step.	**56**	*t68.B_compl_act*	U2U6U5_5ss + SKIP + 14S_NTC_Prp19 → Prp31 + Prp3 + U2_3ss_U6_5ss_U5 (45S activated spliceosome) + free Lsm	(151, 283)
Binding of Prp2 to the spliceosome requires prior binding of Spp2.	**57**	*t70.Spp2_Prp2_act*	Spp2 + Prp2 → Prp2_ass	(165)

continued next page

Table C.1: Reactions (transitions) involved in spliceosome assembly (act = activated, ass = assigned, bdg = binding, conv = converted, dim = dimerized, hydrol = hydolized, inhib = inhibited, matur = maturated, phosph = phosphorylated, stab = szabilized, stl = stemloop, uwd = unwound, tri-snRNP = U4/U5/U6 subcomplex)

Description	ID	**Label**	**Reaction**	**References**
Opening of RNA at 5′ss and ligation of intron 5′ end to branch point adenin (nucleophilic attack).	**58**	*t69.1st_catal_step*	U2_3ss_U6_5ss_U5 + Prp2_ass + ATP → U2_3ss_U6_5ssfree_U5 + ADP	(284, 154)
C Complex Reactions				
Conformational rearrangements by ATP hydrolysis by Prp16, requiring a clear branch point signal. DDX35 and Hsp73 occur together with Prp17 in C-complex, but not in 45S activated spliceosomes (B complex). Critical step due to kinetic proofreading by Prp16 and essential for submitting optimal substrates into the 2nd step of splicing or suboptimal splicing substrates into a discard pathway. Many SR proteins are not observed anymore in purified complexes of this stage.	**59**	*t93.U2_3ss_U6_5ss_U5_remod*	Prp16 + U2_3ss_U6_5ssfree_U5 + ATP + Hsp73 + DDX35 → U2_5ss_U6_U5_conf1 + ADP + ASF_SF2 + SC35	(165, 151, 197)
Proper kinetic speed of Prp16 in ATP hydrolyzation and conformational transition to lariat intermediate, resulting in positioning 5′ss near downstream exon, enabling 3′ss recognition by U5snRNA and Prp8. Four SF3b proteins are not detected in C-complex.	**60**	*t101.Prp16_remod_step*	U2_5ss_U6_U5_conf1 → SF3b49 + SF3b14a + SF3b14b + SF3b10 + U2_5ss_U6_U5_conf2	(285, 151, 197)
DExD/H box ATPase Prp22p binds to the intron just upstream of the 3′ss to promote exon ligation. Prp22 proofreads exon ligation by sensing aberrant substrates before the second transesterification. Hydrolyzation of ATP serves as timer to reject intermediates after a certain time, thus, favoring only fast substrates to be spliced. Prp22 binds additional factors, such as Slu7 and Prp8, which are essential for splicing fidelity and may regulate the ATP dependent activityof Prp22.	**61**	*t92.U2U6U5_3ss_remod*	Prp22 + Prp17 + Prp18 + Slu7 + U2_5ss_U6_U5_conf2 + ATP → Prp16 + U2U6U5_3ss + ADP	(165, 286, 198, 287)
2nd catalytic step by opening the substrate RNA at the 3′ss and ligation of free exon ends along with release of the catenated exons and a post-splicing complex (containing the lariat, composed of intron and associated RNPs).	**62**	*t94.2nd_catal_step*	U2U6U5_3ss → spliced_mRNA + post_splsom_complex	(165)
DExD/H helicase Prp43 replaces Prp22 and promotes release and dissassembly of factors from the postspliceosomal complex. Prp6 is modeled to leave at this stage in order to rejoin another round of tri-snRNP assembly.	**63**	*t95.intron_release*	post_splsom_complex + Prp43_ass + ATP → U2U6_lariat + ADP + hPrp6 + 35S_U5	(288, 241, 193)
Brr2 mediated U2 release separating U2 snRNA from intron, during spliceosome disassembly. Brr2 is still part of the postspliceosomal complex, hence not separately modeled. GTP is required for the Snu114-GTP state, which regulates Brr2 helicase activity.	**64**	*t77.U2_release*	GTP + U2U6U5_lariat → U2 + GDP + U6 + intron	(192, 193)

continued next page

Table C.1: Reactions (transitions) involved in spliceosome assembly (act = activated, ass = assigned, bdg = binding, conv = converted, dim = dimerized, hydrol = hydolized, inhib = inhibited, matur = maturated, phosph = phosphorylated, stab = szabilized, stl = stemloop, uwd = unwound, tri-snRNP = U4/U5/U6 subcomplex)

Description	ID	**Label**	**Reaction**	**References**
35S U5 snRNP is converted into a 20S particle. This conversion involves the release of the Prp19 (NTC) complex and reassociation of Prp6, Prp28 and hDib1.	**65**	*t85.35S_U5_conv*	Prp28 + hDib1 + 35S_U5 → 14S_NTC_Prp19 + SKIP + 20S_U5	(151)
Negative outcome of kinetic proofreading pathway: suboptimal splicing substrates slow conformational rearrangement powered by ATP hydolysis by Prp16 and activate a discard pathway.	**66**	*t100.premature_ATP_hydrol*	U2_5ss_U6_U5_conf1 + ATP → U2_3ss_U6_5m_discard + Prp16 + ADP	(165, 197)
Spp382p-dependent activity in the discard pathway of impaired spliceosomes (in presence of defective RNA substrates). Additionally, this activity can be stimulated byhPrp43, leading to dissociation and reorganization of the spliceosome.	**67**	*t96.Spp382_hPrp43_act*	hPrp43 + Spp382 → Prp43_ass	(241)

C.2 Application of the Integrated Net Analyszer

Table C.2: Analysis sequence in the program INA, applied to the presented PN model. Abbrev.: A,S,Q = primary command switches; W,V,F,E = secondary command switches; Y/N = yes/no decision, e/f = exit/format decision.

Analysis	Decision	Parameter	Function
A			Analyse a Petri net pnt file (name input file *.*pnt*)
	Y		Reset current name options?
		W(V)	Transitions (place) names to be written
	N		Print static conflicts?
S			Compute a basis for all semipositive P/T invariants
	Y		Check current options?
		T/P	Computation of T/P invariants
		F	Outputformat: print non-zero entries with names (#)
	N		Skip a certain number of lines?
	N		Set reduction options?
	(f)		Results by other format?
	N		Reset reduction options?
	Y		Change output format (E) ?
		E	Print non-zero entries only (name output file *.*res*)
	(e)		Results by other format?
	Q		Quit analysis
Q			Quit program
	Y		forget the net?
	Y		Save the commands OPTIONS.ina / Rename SESSION.ina? (name output file *.*ina*)

C.3 T-Invariants Computed for the Spliceosomal Assembly Network

Table C.3: Description of trivial and non-trivialT-invariants, describing signaling pathways during the process of spliceosome assembly

ID	#t	Transitions	Biological Interpretation
1	2	t102, t116	Prp28 (DDX23) influx and efflux (trivial)
2	2	t39, t117	HDib1(U5.15K) influx and efflux (trivial)
3	5	t2, t60-t63	PTB inhibition of branchsite, outcompeting U2AF(65) and blocking of E complex formation(PTB is itself alternatively spliced)
4	2	t156, t157	SF1 (BBP) influx and efflux (trivial)
5	2	t37, t38	Prp5 (DDX46) influx and efflux (trivial)
6	2	t153, t154	SC35 (SFRS2,SRp30b) influx and efflux (trivial)
7	5	t36, t146-t149	U4 snRNP maturation and decay
8	2	t78, t112	SF3b49 (SAP49) influx and efflux (trivial)
9	2	t80, t108	SF3b14a (SAP14) influx and efflux (trivial)
10	2	t79, t115	SF3b14b (PHF5A) influx and efflux (trivial)
11	2	t74, t107	SF3b10 influx and efflux (trivial)
12	2	t103, t140	ASF/SF2 (SFRS2A, SRp30a) influx and efflux (trivial)
13	91	t0-t8, t11, t13, t14, t16, t18-t21, , t23, t26, t31, t32, t37, t42, t44-t54, t57, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t110, t111, t113, t114, t116-t129, t144, t145, t148, t150, t151, t158-t160	U1 independent 5′ss activation, early SF3b125 action (*t13.17S_U2_matur2*) in U2 maturation
14	92	t0-t8, t11, t14, t16, t18-t23, t26, t31, t32, t37, t42, t44-t54, t57, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t109-t111, t113, t114, t116-t129, t144, t145, t148, t150, t151, t158-t160	As i13 but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation
15	105	t0-t9, t11-t13, t15, t18-t21, t23, t24, t26, t33, t34, t40, t42, t44-t54, t57-t59, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t104-t106, t110, t111, t113, t114, t116-t136, t141-t145, t148, t150, t151, t153, t158-t160	SRp (ASF/SF2) supported 5′ss recognition with U1C-TIA1 stabilisation, FBP11 (PRP40) mediated cross-intron bridging to SF1, early SF3b125 action (*t13.17S_U2_matur2*) in U2 maturation
16	106	t0-t9, t11, t12, t15, t18-t24, t26, t33, t34, t40, t42, t44-t54, t57-t59, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t104-t106, t109-t111, t113, t114, t116-t136, t141-t145, t148, t150, t151, t153, t158-t160	As i15, but via late SF3b125 action (*t22.17S_U2_matur1*)
17	104	t0-t9, t11-t15, t17-t21, t23, t24, t26, t33, t34, t40, t42, t44-t54, t57, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t104-t106, t110, t111, t113, t114, t116-t136, t141-t145, t148, t150, t151, t158-t160	SRp (ASF/SF2, SC35) supported 5′ss recognition with U1C-TIA1 stabilization, FBP11 mediated cross-intron bridging to SF1-U2AF heteromer, early SF3b125 action (*t13.17S_U2_matur2*) in U2 maturation
18	105	t0-t9, t11, t12, t14, t15, t17-t24, t26, t33, t34, t40, t42, t44-t54, t57, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t104-t106, t109-t111, t113, t114, t116-t136, t141-t145, t148, t150, t151, t158-t160	As i17, but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation
19	93	t3, t7, t9, t11-t13, t18-t21, t26, t33-t35, t40, t42, t44-t54, t56, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t104-t106, t110, t111, t113, t114, t116-t136, t141-t145, t148, t150-t152, t155, t160	SRp (ASF/SF2,SC35) supported 5′ss recognition, U2AF independent E/A-complex assembly early SF3b125 action (t13) in U2 maturation
20	94	t3, t7, t9, t11, t12, t18-t22, t26, t33-t35, t40, t42, t44-t54, t56, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t104-t106, t109-t111, t113, t114, t116-t136, t141-t145, t148, t150-t152, t155, t160	As i19, but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation

continued next page

Table C.3: Description of trivial and non-trivialT-invariants, describing signaling pathways during the process of spliceosome assembly

ID	#t	Transitions	Biological Interpretation
21	93	t3-t7, t10, t11, t13, t18-t21, t26, t33-t35, t40, t42, t44-t54, t56, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t110, t111, t113, t114, t116-t129, t133-t139, t144, t145, t148, t150-t152, t155, t160	5′ terminal exon 5′ss recognition via U1-CBC interaction (with U2AF), SC35 supported 5′ss recognition, but U2AF independent (!) E-/A-complex assembly, early SF3b125 action (t13) in U2 maturation
22	94	t3-t7, t10, t11, t18-t22, t26, t33-t35, t40, t42, t44-t54, t56, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t109-t111, t113, t114, t116-t129, t133-t139, t144, t145, t148, t150-t152, t155, t160	As i21, but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation
23	101	t0-t8, t10, t11, t13-t15, t17-t21, t23, t24, t26, t33, t34, t40, t42, t44-t54, t57, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t110, t111, t113, t114, t116-t129, t133-t139, t144, t145, t148, t150, t151, t158-t160	5′ terminal exon 5′ss recognition via U1-CBC interaction (with U2AF), SC35 supported 5′ss recognition, FBP11 (PRP40) mediated cross-intron bridging to SF1-U2AF heteromer, early SF3b125 action (t13) in U2 maturation
24	102	t0-t8, t10, t11, t14, t15, t17-t24, t26, t33, t34, t40, t42, t44-t54, t57, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t109-t111, t113, t114, t116-t129, t133-t139, t144, t145, t148, t150, t151, t158-t160	As i23, but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation
25	102	t0-t8, t10, t11, t13, t15, t18-t21, t23, t24, t26, t33, t34, t40, t42, t44-t54, t57-t59, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t110, t111, t113, t114, t116-t129, t133-t139, t144, t145, t148, t150, t151, t153, t158-t160	5′ terminal exon 5′ss recognition via U1-CBC interaction (with U2AF), FBP11 mediated cross-intron bridging between U1 and SF1, early SF3b125 action (t13) in U2 maturation
26	103	t0-t8, t10, t11, t15, t18-t24, t26, t33, t34, t40, t42, t44-t54, t57-t59, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t109-t111, t113, t114, t116-t129, t133-t139, t144, t145, t148, t150, t151, t153, t158-t160	As i25, but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation
27	90	t3, t7, t11, t13, t18-t21, t26, t30, t33-t35, t40, t42, t44-t54, t56, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t110, t111, t113, t114, t116-t136, t144, t145, t148, t150-t152, t155, t160	Intron sided 5′ss recognition via U1C-TIA1 contacts, U2AF independent E/A-complex assembly, early SF3b125 action (t13) in U2 maturation
28	91	t3, t7, t11, t18-t22, t26, t30, t33-t35, t40, t42, t44-t54, t56, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t109-t111, t113, t114, t116-t136, t144, t145, t148, t150-t152, t155, t160	As i27, but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation
29	101	t0-t8, t11, t13-t15, t17-t21, t23, t24, t26, t30, t33, t34, t40, t42, t44-t54, t57, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t110, t111, t113, t114, t116-t136, t144, t145, t148, t150, t151, t158-t160	Intron sided U1 5′ss recognition, FBP11 mediated cross-intron bridging to SF1-U2AF heteromer, early SF3b125 action (t13) in U2 maturation
30	102	t0-t8, t11, t14, t15, t17-t24, t26, t30, t33, t34, t40, t42, t44-t54, t57, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t109-t111, t113, t114, t116-t136, t144, t145, t148, t150, t151, t158-t160	As i29, but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation
31	102	t0-t8, t11, t13, t15, t18-t21, t23, t24, t26, t30, t33, t34, t40, t42, t44-t54, t57-t59, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t110, t111, t113, t114, t116-t136, t144, t145, t148, t150, t151, t153, t158-t160	Intron sided U1 5′ss recognition, FBP11 mediated cross-intron bridging between U1 and SF1, early SF3b125 action (t13) in U2 maturation
32	103	t0-t8, t11, t15, t18-t24, t26, t30, t33, t34, t40, t42, t44-t54, t57-t59, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t109-t111, t113, t114, t116-t136, t144, t145, t148, t150, t151, t153, t158-t160	As i31, but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation
33	87	t3, t7, t11, t13, t18-t21, t26, t33-t35, t40-t42, t44-t54, t56, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t110, t111, t113, t114, t116-t129, t133-t136, t144, t145, t148, t150-t152, t155, t160	Normal U1 5′ss recognition, bridging of U1 via SC35 to SF1-BP, U2AF independent E/A-complex assembly, early SF3b125 action in U2 maturation

continued next page

Table C.3: Description of trivial and non-trivialT-invariants, describing signaling pathways during the process of spliceosome assembly

ID	#t	Transitions	Biological Interpretation
34	88	t3, t7, t11, t18-t22, t26, t33-t35, t40-t42, t44-t54, t56, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t109-t111, t113, t114, t116-t129, t133-t136, t144, t145, t148, t150-t152, t155, t160	As i32, but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation
35	98	t0-t8, t11, t13-t15, t17-t21, t23, t24, t26, t33, t34, t40-t42, t44-t54, t57, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t110, t111, t113, t114, t116-t129, t133-t136, t144, t145, t148, t150, t151, t158-t160	Normal U1 5′ss recognition, FBP11 mediated cross-intron bridging to SF1-U2AF heteromer, early SF3b125 action (t13) in U2 maturation
36	99	t0-t8, t11, t14, t15, t17-t24, t26, t33, t34, t40-t42, t44-t54, t57, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t109-t111, t113, t114, t116-t129, t133-t136, t144, t145, t148, t150, t151, t158-t160	As i35, but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation
37	99	t0-t8, t11, t13, t15, t18-t21, t23, t24, t26, t33, t34, t40-t42, t44-t54, t57-t59, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t110, t111, t113, t114, t116-t129, t133-t136, t144, t145, t148, t150, t151, t153, t158-t160	Normal U1 5′ss recognition, FBP11 mediated cross-intron bridging between U1 and SF1, early SF3b125 action (t13) in U2 maturation
38	100	t0-t8, t11, t15, t18-t24, t26, t33, t34, t40-t42, t44-t54, t57-t59, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t106, t109-t111, t113, t114, t116-t129, t133-t136, t144, t145, t148, t150, t151, t153, t158-t160	As i37, but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation
39	100	t0-t7, t11, t13, t14, t18-t21, t25-t29, t33, t34, t37, t42, t44-t55, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t108, t110-t129, t133-t136, t144, t145, t148, t150, t151, t153, t157, t160	U1 U2 snRNP interaction via Prp5 and subsequent binding to 5′ss and SF-U2AF engaged BP/3′ss, early SF3b125 action in U2 maturation
40	101	t0-t7, t11, t14, t18-t22, t25-t29, t33, t34, t37, t42, t44-t55, t64, t68-t72, t76, t77, t81, t82, t84-t98, t101, t103-t129, t133-t136, t144, t145, t148, t150, t151, t153, t157, t160	As i39, but via late SF3b125 action (*t22.17S_U2_matur1*) in U2 snRNP maturation
41	4	t64-t66, t83	NTC-complex formation
42	2	t67, t75	SKIP (PRPF45) influx and efflux (trivial)
43	2	t43, t73	Prp6 (U5.102K) influx and efflux (trivial)
44	93	t0-t8,t11,t13,t15,t18-t21,t23,t24,t26,t33,t34,t40-t54,t57-t59,t64-t72,t86,t87,t93,t96-t100,t103-t108,t110-t116,t118-t129,t133-t136,t144,t145,t148,t150,t151,t153,t158,t159	Similar to i37, but entering the discard pathway before C-complex stage. Hence, some transitions as *t43.hPrp6_in, t65.Prp19_stab, t66.NTC_form, t67.SKIP_in, t107.SF3b10_in, t108.SF3b14a_in, t112.SF3b49_in* or, *t115.SF3b14b_in*, which leave the assembly process after the critical stage where the spliceosome can turn into the discard pathway, appear in this T-invariant since these factors need to be recylcled (drawn as logical places) in case of the productive outcome of spliceosome assembly, but have no explicit output transition in case of the discard pathway (this applies for all T-invariants covering the discard pathway)