Bioinformatics Analyses of Alternative Splicing

Prediction of alternative splicing events in animals and plants using Machine Learning and analysis of the extent and conservation of subtle alternative splicing

Dissertation

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

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

von

Rileen Sinha

geboren am 06.01.1973 in Pittsburgh, U.S.A

Jena 2009

Die vorliegende Arbeit wurde in der Zeit von April 2006 bis Oktober 2009 am Leibniz Institut für Altersforschung – Fritz-Lipmann-Institut in Jena und am Institut für Informatik, Albert-Ludwigs-Universität Freiburg angefertigt.

Gutachter

1. ....
2. ....
3. ....

Table of Contents

List of abbreviations

AA	alternative acceptor
AD	alternative donor
AS	alternative splicing
AUC	area under the ROC curve
BN	Bayesian Network
cDNA	complementary DNA
DNA	deoxyribonucleic acid
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
EST	expressed sequence tag
hnRNPs	heterogeneous nuclear RNPs
ISE	intronic splicing enhancer
ISRE	intronic splicing regulatory element
ISS	intronic splicing silencers
mRNA	messenger RNA
NCBI	National Center for Biotechnological Information
NGS	next generation sequencing
NMD	nonsense-mediated mRNA decay
nt	nucleotides
PCR	polymerase chain reaction
PPT	polypyrimidine tract
PTC	premature termination codon
RefSeq	Reference sequence Database (of NCBI)
RNA	ribonucleic acid
ROC	receiver operating characteristic
RT-PCR	reverse transcription coupled with polymerase chain reaction
SNP	single-nucleotide polymorphism
snRNP	small nuclear ribonucleoprotein
SR protein	Serine-Arginine (Ser-Arg) protein
SS	splice site
SVM	support vector machine
TassDB	tandem splice site database
UCSC	University of California Santa Cruz
UTR	untranslated region
WGS	whole genome shotgun

Table of Figures

SUMMARY

Zusammenfassung

Alternatives Spleißen (AS) ist ein Mechanismus, durch den ein Multi-Exon-Gen verschiedene Transkripte und damit verschiedene Proteine exprimieren kann. AS trägt wesentlich zur Komplexität und Vielfalt eukaryotischer Transkriptome und Proteome bei. Die Bioinformatik hat in den vergangenen zehn Jahren entscheidenden Beiträge zu unserem Verständnis des AS in Bezug auf Verbreitung, Umfang und Konservierung der verschiedenen Klassen, Evolution, Regulierung und biologische Funktion geliefert. Zum Nachweis des AS im großen Maßstab wurden meist Verfahren zur Genom- und Transkriptom-weiten Alignierung von EST- und mRNA-Daten sowie Microarray-Analysen eingesetzt, die weitestgehend auf bioinformatischen Methoden basieren. Diese wurden durch rechnergestützte Verfahren zur Charakterisierung und Vorhersage von AS ergänzt, die zeigen, wie sich konstitutive und alternative Spleißorte sowie Exons unterscheiden.

Die vorliegende Dissertationsschrift beschäftigt sich mit bioinformatischen Analysen ausgewählter Aspekte des AS. Im ersten Teil habe ich Verfahren zur Vorhersage des AS entwickelt, ohne dabei auf Datensätze exprimierter Sequenzen zurückzugreifen. Insbesondere habe ich Ansätze zur Vorhersage von Kassetten-Exons mittels Bayessches Netze (BN) weiterentwickelt und neue diskriminierende Merkmale etabliert. Diese verbesserten deutlich die Richtig-Positiv-Rate von publizierten 50% auf 61%, bei einer stringenten Falsch-Positiv-Rate von nur 0,5%. Ich konnte zeigen, dass Exons, die als konstitutiv gekennzeichnet waren, denen aber durch das BN eine hohe Wahrscheinlichkeit zugeweisen wurde, alternativ zu sein, in der Tat durch neueste Expressionsdaten als alternativ bestätigt wurden. Bei gleichen Datensätzen und Merkmalen entspricht die Leistungsfähigkeit eines BN der einer publizierten Support-Vektor-Maschine (SVM), was darauf hinweist, dass verlässliche Ergebnisse bei der Klassifikation mehr von den Merkmalen als von der Wahl des Klassifikators abhängen.

Im zweiten Teil habe ich den BN-Ansatz auf eine umfangreiche und evolutionär weit verbreitete Klasse von AS-Ereignissen ausgeweitet, die als NAGNAG-Tandem-Spleißstellen bezeichnet werden und bei denen die alternativen Spleißorte nur 3 Nukleotide (nt) voneinander getrennt sind. Die sorgfältige Zusammenstellung der Trainings- und Test-Datensätze bei der Vorhersage des NAGNAG-AS trug zu einer ausgewogenen Sensitivität und Spezifität von 92% bei. Vorhersagen eines auf dem vereinigten Datensatz trainierten BN konnten in 81% (38/47) der Fälle experimentell bestätigt werden. Im Rahmen dieser Studie wurde damit einer der gegenwärtig umfangreichsten Datensätze zur experimentellen Verifizierung von Vorhersagen des AS generiert. Ein BN, trainiert anhand menschlicher Daten, erzielt ähnliche gute Ergebnisse bei vier anderen Wirbeltier-Genomen. Nur leichte Einbußen bei Vorhersagen für *Drosophila melanogaster* und *Caenorhabditis elegans* weisen darauf hin, dass der zugrunde liegende Spleißmechanismus über weite evolutionäre

Distanzen konserviert zu seien scheint. Schließlich verwendete ich die Vorhersagegenauigkeit der experimentellen Validierung, um die Zahl der noch unentdeckten alternativen NAGNAGs abzuschätzen. Die Ergebnisse deuten darauf hin, dass der Mechanismus des NAGNAG-AS einfach, stochastisch und konserviert ist - unter Wirbeltieren und darüber hinaus. Des weiteren habe ich den BN-Ansatz zur Charakterisierung und Vorhersage von NAGNAG-AS in *Physcomitrella patens*, einem Moos, eingesetzt. Dies ist eine der ersten Studien zur Vorhersage von AS in Pflanzen, ohne dabei auf Datensätze von exprimierten Sequenzen zurückzugreifen. Wir erreichten ähnliche Ergebnisse, wie in unseren anderen Arbeiten zur Vorhersage NAGNAG-AS. Eine unabhängige Validierung mittels 454-NextGen-Sequenzdaten zeigte Richtig-Positiv-Raten von 64%-79% für gut unterstützt Fälle von NAGNAG-AS. Damit scheint der Mechanismus des NAGNAG-AS bei Pflanzen dem der Tiere zu ähneln.

Im dritten Teil habe ich mich an Analysen zur phylogenetischen Konservierung des subtilen AS beteiligt, um die Frage zu beantworten, wieviele subtile AS-Ereignisse von funktioneller Bedeutung sind. Dabei konzentrierten wir uns auf Tandems mit einem Abstand von 3-9 nt. Wir konnten frühere widersprüchliche Ergebnisse zur Konservierung von alternativen und konstitutiven Tandem-Motiven auflösen, indem wir diese auf ein statistisches Paradox (Simpsons Paradox) zurückführten. Anhand von Methoden, die entsprechende Verzerrungen berücksichtigen, wurde gezeigt, dass alternative Tandemmotive stärker konserviert sind als konstitutive. Aus diesen Analysen konnten wir eine konservative Abschätzung der Zahl von Tandem-Spleißorten unter reinigender (negativer) Selektion ableiten.

Schließlich war ich in der Aktualisierung und erheblichen Ausweitung der Tandem-Spleißstellen-Datenbank (TassDB2) beteiligten, die eine umfassende Informationsquelle für Forscher im Bereich des subtilen AS darstellt. TassDB2 enthält sowohl vermeintliche als auch experimentell bestätigte Tandem-Spleißstellen in einer Entfernung von 2-12 nt. Nutzer können nach verschiedenen Kriterien, einschließlich Gen-Namen, Leserahmen-Erhaltung, Anzahl der Transkripte, experimentelle Bestätigung, Isoform-Verhältnis und Konservierung des Tandemmotivs in Maus, Hund, Huhn oder Zebrafisch, suchen.

Insgesamt habe ich in dieser Arbeit sowohl konservierte Kassetten-Exons in Mensch und Maus sowie NAGNAG-AS in fünf Wirbeltier-Genomen, Fliege, Wurm und der Pflanze Physcomitrella vorhergesagt als auch Verbreitung und Konservierung des subtilen AS untersucht.

Summary

Alternative splicing (AS) is a mechanism by which a multi-exonic gene can produce different transcripts and thereby different proteins. AS is a major contributor to the complexity and diversity of eukaryotic transcriptomes and proteomes. Bioinformatics has made significant contributions to research in AS over the past decade. Computational methods have been critical for AS in respect to its abundance, the frequency and conservation of different classes of AS, the evolution of AS, regulation of AS, and its functional impact on various biological processes. Large-scale detection of AS has mostly been performed using alignment of EST and mRNA data to genomes, or microarray data, both of which extensively use bioinformatics methods. These have been complemented by computational methods of characterization and prediction of AS, which show how to distinguish between constitutive and alternative splice sites and exons.

This thesis concerns itself with bioinformatics analyses of selected aspects of AS. In the first part, I predict AS without using expressed sequence information. Specifically, I extend previous studies on predicting conserved cassette exons by using Bayesian Networks (BNs), and several novel discriminative features. This significantly improved the true positive rate from a previously reported 50% to 61%, at a stringent false positive rate of 0.5%. I show that exons which are labelled constitutive but receive a high probability of being alternative by the BN, are in fact alternative exons according to the latest transcript data. When using the same dataset and the same set of features, the BN matches the performance of a support vector machine (SVM) in earlier literature, indicating that good classification depends more on features than on the choice of classifier.

In the second part, I extend the BN approach to AS prediction to an evolutionarily widespread class of AS, the so called “NAGNAG AS”, involving tandem splice sites separated by 3 nucleotides (nt). Careful construction of training and test datasets helped achieve a balanced sensitivity and specificity of ≥ 92% in predicting NAGNAG AS. Predictions by a BN trained on the combined dataset could be experimentally verified in 81% (38/47) of the cases. This constitutes one of the largest sets of experimentally verified predictions of AS to date. A BN learned on human data achieves similar performance on four other vertebrate genomes, while there is only a slight drop for Drosophila and worm, which indicates conservation of the underlying splicing mechanism. Lastly, I use the prediction accuracy according to experimental validation to estimate the number of yet undiscovered alternative NAGNAGs. The results suggest that the mechanism behind NAGNAG AS is simple, stochastic, and conserved among vertebrates and beyond. I then applied the BN approach to characterize and predict NAGNAG AS in *Physcomitrella patens*, a moss, in one of the first reported studies on predicting AS in plants without using expressed sequence

information. We achieve similar results as in our other work on predicting NAGNAG AS, with and independent validation using 454 data resulted in 64%-79% of the well-supported cases of NAGNAG AS being correctly predicted. Thus the mechanism behind NAGNAG AS in plants seems to be similar to that in animals.

In the third part, I contributed to the analyses of phylogenetic conservation to address the question of how many subtle AS events are functionally important. Focusing on tandems with a distance of 3–9 nucleotides, we resolve previous contradicting results on whether alternative or constitutive tandem motifs are more conserved between species by showing that they can be explained by a statistical paradox (Simpson's paradox). The applied methods took biases into account, and found that alternative tandems are more conserved than constitutive tandems. We estimate a lower bound for the number of alternative sites that are under purifying (negative) selection.

Lastly, I was involved in the update and significant extension of the tandem splice site database (TassDB) to create TassDB2, a comprehensive resource for researchers interested in subtle AS. TassDB2 contains both putative and confirmed splice sites separated by distance 2-12 nt. Users can search by many different criteria, including gene name, frame-preservation, number of supporting transcripts for each variant, the ratio of supporting transcripts, and conservation of the splice site pattern in mouse, dog, chicken or zebrafish.

In summary, in this thesis I predict conserved cassette exons in human and mouse, predict NAGNAG AS in five vertebrate genomes, fly, worm, and the plant *Physcomitrella patens*, and study the extent and conservation of subtle AS.

INTRODUCTION

“One consequence of the intronic model is that the dogma of one gene, one polypeptide chain disappears. A gene, a contiguous region of DNA, now corresponds to one transcription unit, but that transcription unit can correspond to many polypeptide chains, of related or differing functions.” – Walter Gilbert [1].

Introduction

The central dogma of molecular biology deals with the directionality of transfer of sequential information. It states that “information cannot be transferred back from protein to either protein or nucleic acid” [2]. The most common transfers of sequential information are DNA to DNA (replication), DNA to RNA (transcription), and RNA to protein (translation). The second and third types of transfer are the two major steps of producing a protein from a gene – first, the DNA of a gene is transcribed to produce mRNA, and this mRNA is then translated into protein. In prokaryotic organisms, transcription occurs in the cytoplasm, and the mRNA is usually not modified, but directly used as a template to read off in steps of three residues (codons), which are then ‘translated’ into amino acids, thus eventually forming a protein. However, in eukaryotes, transcription takes place in the nucleus, where the primary transcript is processed, after which the mature form is exported to the cytoplasm for translation into a protein. These mRNA processing steps are 5’ capping, 3’ polyadenylation, and splicing. Splicing is the process of excision of intervening sequences, called introns, from the primary transcript, which is one of the fundamental differences between the gene architecture of eukaryotes and prokaryotes [3, 4]. The transcribed regions of the vast majority of genes in “higher” eukaryotes can be divided into introns and sequences which are retained in the mature mRNA, called exons. Thus, the protein-coding sequence (CDS) in such eukaryotic genes is mostly split across several parts of the pre-mRNA, and must be brought together before being translated into a protein. Instead of simply doing this in one fixed manner, often variable parts of the pre-mRNA are used, thus giving rise to alternative splicing (AS), a mechanism by which the same gene can produce different mRNAs, and hence eventually different proteins. Both the different mRNAs and the different proteins arising from the same gene are called isoforms. Thus, as foreseen by Walter Gilbert shortly after the discovery of introns, their existence provides an “evolutionary playground” such that new forms of a protein can be experimented with, while still retaining the currently functional variant [1]. AS is one of the ways in which the split gene architecture of eukaryotes can be exploited thus.

Splicing

To understand AS, we must first appreciate the basic mechanism of splicing in general. Splicing is carried out by a huge ribonucleoprotein complex called the spliceosome, which consists of five small nuclear ribonucleoproteins (snRNPs) – U1, U2, U4, U5 and U6 - that are associated with a large number (~300) of proteins known as splicing factors [5-8]. While over 99% of eukaryotic introns are spliced out by this spliceosome, also known as the U2-dependent spliceosome, there also exists a class of introns in higher eukaryotes which are spliced by another spliceosome, called the U12-dependent spliceosome [9-12]. The U12-dependent spliceosome requires the snRNPs U11, U12, U4atac, U6atac, and U5, which is the only snRNP common to both U2 and U12 dependent spliceosomes. The introns spliced out by the two spliceosomes are also referred to as U2 introns and U12 introns. The question of how the spliceosome reliably locates the relatively short exons amidst the vastly larger introns (10-100 times larger in an average vertebrate gene [13]) has been the subject of intensive research for over three decades - and while much has been learned, the quest is far from over. The ends of introns carry signals (Figure 1, Figure 2a) which are degenerate yet highly conserved throughout eukaryotes [13-15]. How the spliceosome specifically selects a particular splice site (SS) depends on the composition of these signals and several other factors. The three basic splicing signals are the SS near the 5' intron end, also called donor, the SS near the 3' intron end, also called acceptor, and the branch point, which is usually located about 40 nucleotides (nt) upstream of the acceptor site and has (for U2 introns) the consensus YTRAY (Y stands for C or T, R stands for A or G, and the branch point adenosine is underlined) [14, 16]. The donor site in mammals has an extended intronic consensus sequence GTRAGT, where the first intronic dinucleotide GT is nearly invariant, as is the last dinucleotide AG at the acceptor or 3' end (Figure 2a). Thus the vast majority of U2 introns have GT-AG termini, and the only noteworthy exception seem to be U2 introns with GC-AG termini, which occur at a frequency of ~1% [12]. On the other hand, U12 introns usually show GT-AG or AT-AC termini, and lack the PPT which is characteristic of U2 introns. The branchpoint in U12 introns has a very strict sequence composition of TTCCTTAAC (the underlined A is bulged adenosine which attacks the 5' SS) and is usually located 10-26 nt upstream of the acceptor, while the dinucleotide at the acceptor, while most often AC, seems particularly diverse in comparison to the nearly invariant AG in U2 introns, and includes AC, AG, AA, CG and TT [17, 18]. For the remainder of this thesis, "spliceosome" shall refer to the U2-dependent spliceosome, unless otherwise specified.

As shown in Figure 1, the splicing reaction begins with the positioning and rearrangement of splicing factors (e.g. SF1) on the pre-mRNA, which eventually leads to the assembly of an active spliceosome that carries out intron removal. The major steps of the splicing of an intron, as depicted (in figure 1) are the following [15, 16, 19]:

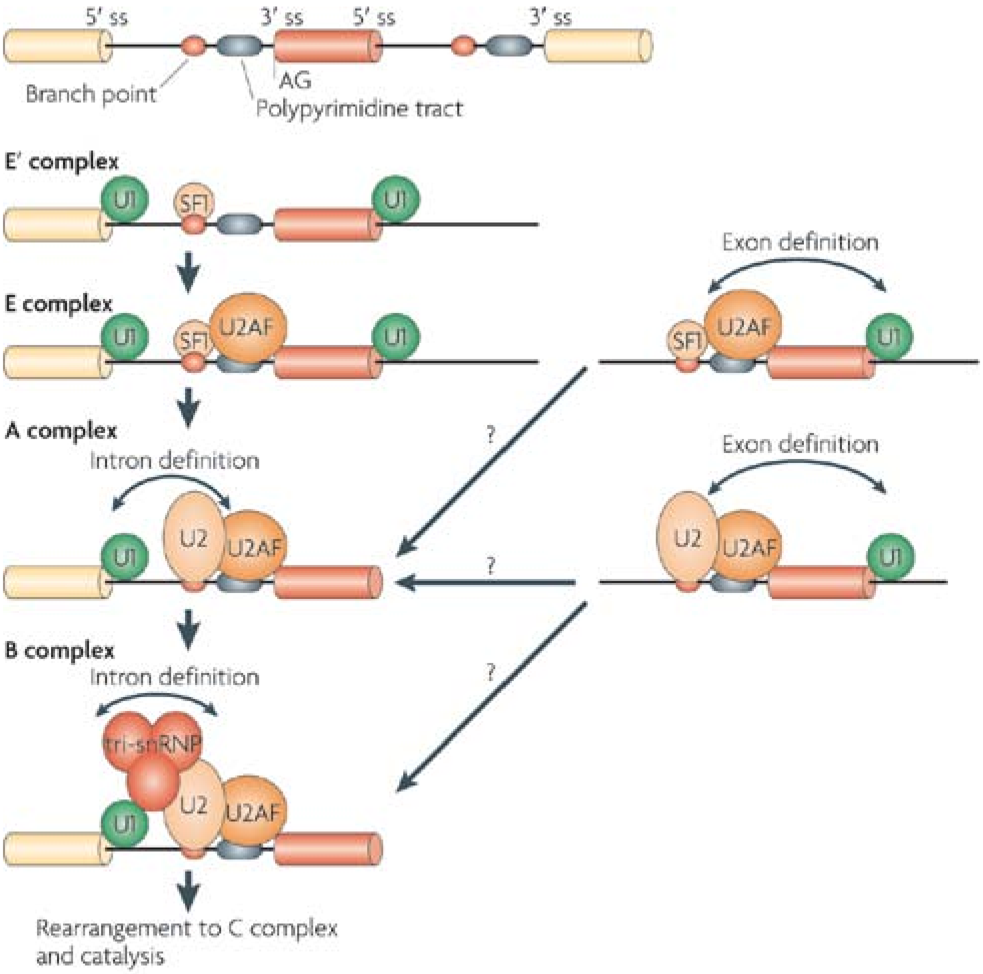


Figure 1. A simplified overview of spliceosome assembly.

First, the U1 snRNA binds to the 5’ splice site (donor) via base-pairing, and splicing factor SF1 binds to the branchpoint in an ATP-independent manner to form the E’ complex. Then the U2 auxiliary factor (U2AF) heterodimer, which consists of the subunits U2AF65 and U2AF35, binds to the polypyrimidine tract (PPT) and the 3’ AG, and the ATP-independent E complex is formed. Replacement of SF1 by the U2 snRNP at the branchpoint converts and the ATP-independent E complex into the and the ATP-dependent A complex. This is followed by the recruitment of the U4/U6-U5 tri-snRNP and the formation of the B complex, containing all spliceosomal subunits involved in pre-mRNA splicing. Following extensive conformational changes and remodelling, the C complex is formed, which is the active spliceosome. Figure taken from [15].

(1) the U1 snRNP binds to the donor site by specific base pairings and SF1 to the branch point,

(2) the protein heterodimer U2AF binds to the polypyrimidine tract and acceptor site [20],

(3) the U2 snRNP binds to the branch site by base pairings and replaces SF1 [21],

(4) the tri-snRNP consisting of U4, U5, and U6 enters the spliceosome,

(5) the U6 snRNP replaces U1 by binding to the donor site, and U1 and U4 are released from the spliceosome,

(6) the mRNA is cleaved at the donor site and the 5’ intron end is attached to the branch point adenosine forming a lariat structure,

(7) the mRNA is cleaved at the acceptor site, the upstream exon is ligated to the downstream exon, and the intron is released.

Alternative splicing

AS, as the name implies, refers to different ways of splicing primary transcripts of a given gene, resulting in different mature transcripts from the same gene. Put another way, splicing may result in a particular donor splice site being used in conjunction with different acceptor splice sites and vice versa. This violated the classical ’one gene, one polypeptide chain’ rule. AS was first reported in 1980 [22], but was considered a relatively rare event, affecting only 5-10% of all genes, for nearly two decades. Bioinformatics, and in particular the analysis of expressed sequence tags (ESTs), has played a big role in sharp upward revisions of estimates of the frequency of AS in the last decade. At first, such EST analyses raised the estimates to 35%-59% [23-25]. Then microarray-based analysis raised the estimate to 74% [26], and the latest estimates, based on data from the latest sequencing platforms such as Illumina/Solexa, are that 92-94% of human genes are alternatively spliced [27, 28].

Their frequency of inclusion in mature transcripts distinguishes constitutive from alternative splice sites. A constitutive splice site is one which is always used in the mature transcripts, while an alternative splice site can be omitted sometimes. The terms constitutive and alternative are similarly applied to exons and to the splicing process in general.

Most alternative splice events can be classified into the following basic types (Figure 2b):

- the inclusion or exclusion of one (or more) exons (denoted exon skipping),
- the usage of alternative donor or acceptor sites,
- the mutual exclusion of exons,
- the retention of an intron.

The frequency of occurrence of these splicing patterns varies across phyla – exon-skipping is the most frequent in mammals, while intron-retention is the most frequent in plants, with alternative acceptors and alternative donors falling in between the two types for both [25, 29-32]. These elementary events consist of binary alternatives. On the other hand, transcript isoforms often represent rather complex splicing patterns, which are combinations of the elementary events and characterized by more than two possible outcomes. Based on full-length cDNA data, as much as 20%-30% of all AS outcomes might be complex [33-35].

Nonetheless, such events remain under-studied and under-appreciated to date, due to a lack of enough transcript data for detailed characterization. Another reason is that complex events are difficult to model, and analysis is simplified when we assume that the splice sites up- and downstream of an AS event with binary alternatives remain unaffected. Transcript coverage of a given locus can be another limiting factor - high coverage is needed to reliably interpret rare and/or complex events. For example, if we consider that the sequences representing each of two splice variants are sampled from the underlying transcriptome according to the binomial distribution and with probabilities corresponding to their relative abundance, then at least 29 sequences containing the respective exon boundaries are required to detect a minor variant occurring at a relative frequency of 10%, with≥ 95% confidence [36, 37]. This shows that even when we restrict ourselves to events with binary alternatives, things are not necessarily simple, and many events may escape detection. Bioinformatics can play a role by predicting likely AS events which have not yet been detected using transcript data. Furthermore, AS can be coupled to transcript variation by alternative promoter or polyadenylation site usage. Since complex processes/phenomena are not addressed in this thesis, whenever I refer to alternative splice events, it means the types described in Figure 2b, unless otherwise mentioned.

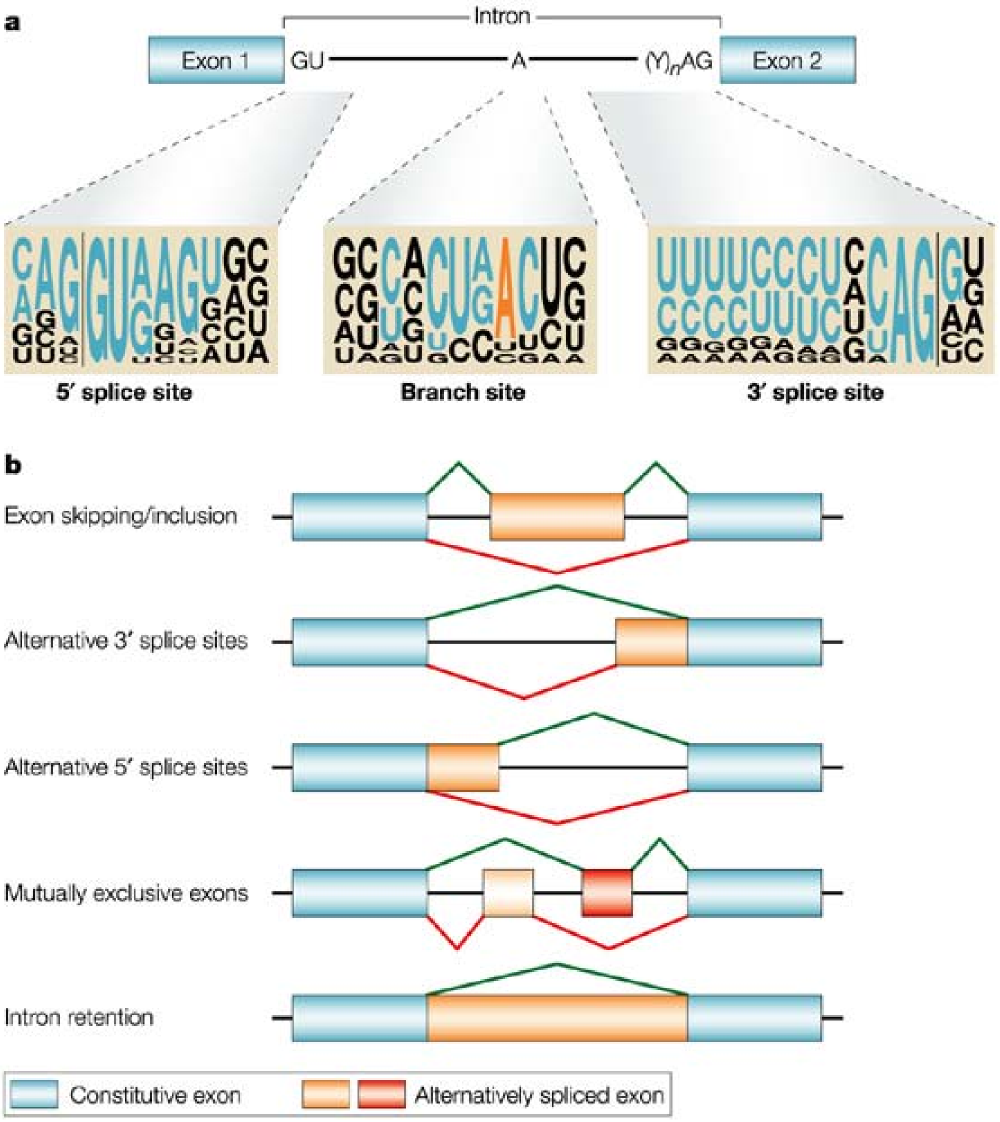


Figure 2. The major splicing signals and most common alternative splicing events.

(a) Conserved motifs at or near the intron ends. The nearly invariant GU and AG dinucleotides at the intron ends, the polypyrimidine tract $(Y)_n$ preceding the 3' AG, and the A residue that serves as a branchpoint are shown in a two-exon pre-mRNA. The sequence motifs that surround these conserved nucleotides are shown below. For each sequence motif, the size of a nucleotide at a given position is proportional to the frequency of that nucleotide at that position. Nucleotides that are part of the classical consensus motifs are shown in blue, except for the branch-point A, which is shown in orange. The vertical lines indicate the exon–intron boundaries. (b) Five common modes of alternative splicing. In each case, one alternative splicing path is indicated in green, the other path in red. In the last example, the alternative pathway corresponds to no splicing. In complex pre-mRNAs, more than one of these modes of alternative splicing can apply to different regions of the transcript, and extra mRNA isoforms can be generated through the use of alternative promoters or polyadenylation sites. Figure taken from [16].

The mechanism of splicing is dependent on the exon-intron architecture of the given gene – when introns are short, splicing seems to proceed via an “intron definition” mechanism, with initial spliceosome assembly occurring around the intron (Figure 1) [38]. However, this mainly seems to happen when introns are shorter than 200 nt – when introns are longer, as is the case with the vast majority of introns in humans and other mammals [13, 39], splicing proceeds via an “exon definition” mechanism (Figure 1) [40, 41]. In this case, the spliceosome first assembles around exons, and later on a conversion to intron definition occurs via cross-intron interactions between the U1 and U2 snRNPs [42, 43]. When both introns and exons are long, the incidence of exons-skipping increases [40], which is also true of cases where exons are very short (say < 30 nt). Splicing in plants proceeds via intron-definition in the majority of cases, and via exon-definition in the majority of cases for vertebrates, which agrees with the findings that while intron-retention is the most common type of AS event in plants, exon-skipping is the most common among vertebrates [25, 29-32, 34, 44-48].

The recognition and utilization of splice sites depends on many factors. The three primary signals – the 5’ SS, the branchpoint, and the 3’ SS (including the PPT) – are very important, but seem to contribute only about 50% of the information needed to discriminate between introns and exons [49]. Therefore the remaining information necessary for the accurate splicing observed in vivo must be contained in exonic and intronic regions. Indeed exons and introns contain binding sites for splicing protein factors that activate or repress splicing. Depending on location (exon/intron) and activity (activation/repression – enhancers activate, while silencers repress), such elements are broadly divided into four categories: exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs). There is, however, overlap between the categories, especially between ESSs and ISSs, which are commonly bound by heterogeneous nuclear RNPs (hnRNPs) [50, 51]. ESEs are usually bound by members of the SR (Ser-Arg) protein family [52-54], whereas ISEs are not as well-characterized as the other three categories, though a few proteins such as hnRNP F, hnRNP H, neuro-oncological ventral antigen 1 (NOVA1), NOVA2, FOX1 and FOX2 (also known as RBM9), have been shown to bind ISEs and to stimulate splicing [55-58].

The impact of alternative splicing

The importance of a biological phenomenon is usually interpreted in terms of its functional impact. AS can result in protein isoforms that differ in various physiological aspects including ligand binding affinity, signalling activity, protein domain composition and sub-cellular localization among others [59]. The ability to produce multiple isoforms with differing properties means that AS has great potential to cause far-reaching changes within the transcriptome and proteome of an organism. To demonstrate this experimentally, a promising beginning has been made by several studies focussed on a small number of proteins, which have shown that AS produces functionally distinct proteins [16, 60-69] (to cite a few). However, large scale demonstrations of functionally different protein isoforms produced by AS are as yet infeasible, though a start has been made recently, based on the analysis of high quality peptide catalogs from the *Drosophila melanogaster* proteome [70]. In the meantime, large-scale bioinformatics studies have given several interesting pointers. They have shown that AS events have a tendency to remove certain protein domains like protein-protein interaction or DNA binding domains, and tend to insert/delete complete functional units instead of affecting parts of a unit [71, 72]. Computational analyses also found that 40-50% of the proteins with one transmembrane helix have a splice form that specifically removes the single transmembrane domain [73, 74], thus creating a soluble protein variant encoded in the exon skipping splice variant and a membrane-bound protein encoded in the exon-inclusion variant. Alternative splice events plays a role in several biological processes such as the formation and function of synapses [75], axon guidance in the fruit fly Drosophila melanogaster [76], and T-cell activation [77]. Moreover, AS in the UTR regions can have an effect by influencing mRNA stability or translation efficiency [78]. Bioinformatics analyses have also indicated that alternative exons tend to evolve faster than constitutive exons, which means that AS can provide a means of tinkering with novel isoforms, trying them out before they are kept or discarded during evolution. When mapped onto protein structures, AS events tend to have deleterious effects on protein structure, which has led to debate over whether such events really result in proteins, or merely transcripts which are degraded before translation [79]. The counterargument is that AS events may play a role in protein structure evolution by facilitating transitions between different folds in the protein sequence-structure space [80]. It has also been shown that nearly a third of AS events result in a premature termination codon, and about 75% of these, or a fourth of all splicing events, are putative targets of a pathway which degrades transcripts with premature stop codons and prevents them from giving rise to deleterious truncated protein isoforms, known as the nonsense-mediated decay (NMD) pathway [81, 82]. Thus, the extent of functional AS is a matter of much healthy debate [83].

Another fact which makes AS very important and interesting is its strong association with disease. Defects in alternative and constitutive splicing are associated with a number of human diseases [16, 84]. Moreover, splicing mutations have been suspected to be the most frequent cause of hereditary diseases [85]. For example, a polymorphism in the *PTPRC* gene that is associated with multiple sclerosis destroys an exonic splicing silencer and abolishes the skipping of exon 4 [86]. Also, changes in the normal splicing pattern are thought to contribute to cancer development [87, 88]. Thus, manipulation of AS to counter disease-causing effects is also of therapeutic interest, and an emerging area of research focuses on treatments which change AS [89-92].

Non-EST based prediction of alternative splicing

When my research for this doctoral thesis started, large scale detection of AS was usually done using ESTs [23, 25] or microarrays (reviewed in [93-95]). Since AS can be highly specific for tissues or developmental stages, these methods can only detect splice events that occur in the underlying probe samples, turning the comprehensive characterisation of the transcriptomes of a complex organism into a challenge exceeding by far the task of determining the respective genome sequence. Moreover, as these approaches were expensive and labour intensive, the data resources are in most cases insufficient, limiting the detection of AS to events with rather high frequencies. Limited transcript coverage also makes it difficult to detect splice events where the minor isoform has a low abundance, even though such events can be of biological importance [96]. In the case of microarrays, the data output is further limited by the array design. Furthermore, at that time Whole Genome Shotgun (WGS) sequencing projects were churning out genomic data at a higher rate than corresponding transcriptome data – the number of ESTs in GenBank Release 161 had increased by 19% in one year, compared to a gain of 39% in the number of contigs in the WGS GenBank division [97]. Thus it was expected that in the foreseeable future, we shall have several genomes without the level of corresponding extensive transcript coverage required to reveal the extent of AS, and hence transcriptomic and proteomic variability. Accordingly, there was a need for *in silico* methods of detecting AS.

More recently, the emergence of new sequencing platforms such Illumina/Solexa, Roche/454, and ABI/Solid has revolutionized the field of molecular biology [98, 99], and splicing is no exception. These technologies enable studies of transcriptomes at an unprecedented level of depth and detail [27, 28]. However, since AS can be specific to tissues, developmental stages, and external stimuli, the number of combinations of situations in which AS can occur is still too large to be exhaustively captured by any existing technology. Furthermore, these technologies are still expensive enough to be out of the reach of many researchers. Thus the need for *in silico* methods of detecting AS, to complement other large-scale methods, remains. Moreover, such methods can provide further stimuli for understanding the mechanisms of AS.

Exon skipping

The most frequent form of AS in humans is exon skipping, whereby an entire exon is either included in, or excluded from the mature transcript [32, 35, 45] – exons which undergo exon-skipping are called skipped, cassette or alternative exons. It has been shown that alternative conserved exons (ACEs) [100] in human and mouse differ from conserved constitutive exons in several ways [100, 101]:

- most ACEs are frame-preserving (also called “symmetrical”, or “peptide cassettes”), that is their length is a multiple of 3, and they do not encode an in-frame stop codon; skipping such an exon only shortens the protein without changing the reading frame;
- both the ACEs as well as their flanking intronic regions tend to show a much higher sequence conservation than their constitutive counterparts:
- ACEs are shorter on average than constitutive exons.

This suggests that features derived from the exon and its flanking introns can be used to predict skipping of exons that are conserved between human and mouse and alternatively spliced in both species (denoted conserved exon skipping events) – an early attempt at such classification used a rule-based classifier and achieved a very high specificity (less than 0.3% false-positive rate), at the cost of a relatively low sensitivity (20%-32%) [101]. The authors also managed to experimentally validate AS in 60% (9/15) of tested exons, including skipping in 40% (6/15) of them. Later, the same authors used more features, including the frequency of dinucleotides and trinucleotides in exonic and flanking intronic sequences, composition of the 5’SS and strength of the PPT, along with a support vector machine (SVM), to achieve a sensitivity of 50% at a false-positive rate of 0.5% [102]. Similar properties were exploited in a genome-wide classification of ACEs using a regularized least-squares classifier [100], and 70% (21/30) of tested predictions were also experimentally validated. SVMs were also successfully used to predict exon skipping in worm, using similar features, but without using sequence conservation [103]. Other approaches to AS prediction used protein domain information [104] and evolutionary conservation [105-107] to detect alternative splice events.

In the first work of my thesis, titled “**Improved identification of conserved cassette exons**” [108], I used Bayesian networks (BNs), a state-of-the-art machine learning method, to predict conserved exon skipping events. BNs are an increasingly popular machine learning approach to data modeling and classification [109-111]. The ability of BNs to cope with features of various value ranges and to learn dependencies between features makes them especially versatile and suited to a large variety of applications. BNs allow multiple dependencies between variables, impose no fixed ordering of variables, allow integration of arbitrary features, and the network structure can be automatically learned. This makes BNs a flexible choice for biological sequence data analysis [112-116]. I introduced several novel features that distinguish alternative exons from constitutive exons, including features based on the single-strandedness of ESEs and ESSs, and features involving intronic splicing regulatory elements (ISRE). By validating our classifiers on various datasets, I identified features which are discriminative irrespective of dataset-specific biases, and provide independent measures of the predictive power of the BNs.

Even though conservation based features have proved to be among the most discriminative features for predicting exon skipping, it is desirable to be able to predict AS using only information from a single genome. Importantly, I could show that our approach can still predict exon skipping without using conservation-based features.

Accurate prediction of NAGNAG alternative splicing

Alternative acceptors are the second most common kind of AS in human, after exon skipping [32, 45, 117]. NAGNAG AS (N stands for any of A,C,G or T), involving tandem acceptors separated by three nucleotides, is a common type of AS, contributing almost half of all cases of conserved alternative acceptor usage [35, 118, 119]. NAGNAG splicing results in two possible splice variants—splicing after the first AG results in the E (exonic, also known as proximal) isoform, whereas splicing after the second AG results in the I (intronic, also known as distal) isoform (Figure 3)—accordingly, we refer to constitutively spliced NAGNAG acceptors as the E- or I-class, and to usage of both acceptors, or AS, as the EI-class. Since the difference between the two isoforms is an inclusion/exclusion of 3 nt, NAGNAG AS does not change the reading frame, and only very rarely results in a premature termination codon (PTC) [118]. Thus, the predominant effect of NAGNAG AS is to produce isoforms which differ from each other in only a single amino acid [118]. The corresponding event at the donor SS, involving donors separated by 3 nt, is called GYNGYN AS (Y = C or T) [120]. The Tandem Splice Site DataBase (TassDB) was the first AS database to extensively characterize NAGNAG and GYNGYN AS in seven genomes [121]. According to the data present in TassDB, 16% (1,815 of 10,740) of human NAGNAG acceptors were alternatively spliced. However, 40% (3,562) of the remaining NAGNAG acceptors had less than ten ESTs each, thus implying that a subset of these NAGNAGs may simply lack evidence of AS due to insufficient sampling of the transcriptome. An accurate predictive method would give us a meaningful estimate of the number of yet undiscovered alternative NAGNAG acceptors. Previous work on predicting alternative 3' splicing, while reporting good results overall, had modest results for NAGNAG AS compared to cases involving larger distances [122]. On the other hand, another previous work reported that a simple model based on splice site strength was enough to explain NAGNAG and other short-distance tandem AS [123], which suggests that short distance tandem AS is mostly due to the spliceosome “slipping” on occasion and stochastically selecting between two nearby competing alternatives – under such a model, subtle AS is mostly a noisy rather than functional process [123, 124]. Another way of looking at this issue is that noise and function are not mutually exclusive – noise is to do with the mechanism, whereas function is what the species makes out of the splicing event.

In the second work of my thesis, titled “**Accurate prediction of NAGNAG alternative splicing**” [125], I sought to improve the prediction of NAGNAG AS, using my Bayesian Networks experience and TassDB [121] to carefully construct our training and test datasets. We achieved a high balanced sensitivity and specificity and good results in extensive experimental validation of predictions [125]. I showed that the performance on a dataset from literature [122] can be improved by a careful consideration of available transcript evidence to include only strongly supported NAGNAGs as constitutive or alternative. Using a BN learned on human data on six genomes (mouse, rat, dog, chicken, *Drosophila Melanogaster* and *C. elegans*), I showed that the performance is comparable or only slightly inferior to that achieved in human. Our results suggest that the mechanism behind NAGNAG splicing is simple, and maintained in evolution.

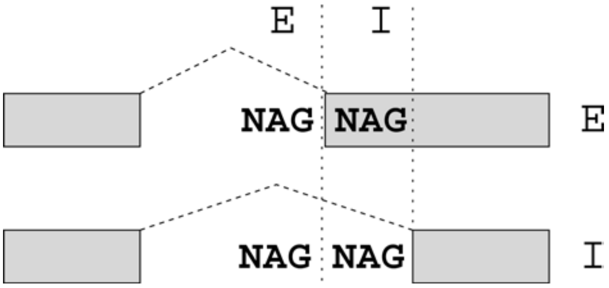


Figure 3. The possible isoforms in NAGNAG splicing.

Splicing at the first AG results in part of the NAGNAG being in the exon, hence the isoform is called exonic; splicing at the second AG creates the I variant, as the entire NAGNAG is now intronic.

Characterization and prediction of NAGNAG alternative splicing in the moss Physcomitrella patens

While there have been numerous experimental as well as computational studies of AS in animals, the study of AS in plants is still in its early stages [31]. While AS is also common in plants, the overall abundance of AS seems lesser than in animals – various studies have estimated that between 20%-30% of plant genes undergo AS [29, 30, 126], while the most recent, high-end estimate based on deep coverage of the Arabidopsis transcriptome is 42%-56% of intron-containing genes[127]. EST-based detection of AS in plants began a few years later in comparison to studies on animals [29, 30], and showed that intron retention appears to be the most common kind of AS event in plants, which is in keeping with the intron definition of model that the majority of plant introns seem to be spliced according to [29, 30, 126]. Similarly, exon-skipping, which is the most common event in animals, is much less frequent in plants. However, alternative acceptors and donors seem to occur at a comparable frequency. In particular, short distance events, or subtle AS events, seem to be just as common, and NAGNAG acceptors are once again widespread, and the most abundant among such events [30]. The model moss *Physcomitrella patens*, the first bryophyte genome to be sequenced, seems to have a distribution of AS events similar to other plants studied so far [128]. In the third work of my thesis "**Identification and characterization of NAGNAG alternative splicing in the moss *Physcomitrella patens***" (manuscript submitted) we extended our previous work on NAGNAG AS by undertaking characterization and prediction of NAGNAG AS in the model moss *Physcomitrella patens*. We analyzed the available EST data using the PASA (Program to Assemble Spliced Alignments) pipeline [129] and found that of 5,031 NAGNAG tandem splice sites with Sanger EST coverage, 295 (5.9%) were alternatively spliced. Furthermore, 4,040 of the constitutive NAGNAGs were covered by < 10 ESTs each, with an average coverage of only 3 ESTs per NAGNAG, indicating that there were potentially many undiscovered alternative NAGNAGs in moss. Use of recently available 454 data increased the total number of alternative NAGNAGs in *Physcomitrella* to 664. Similar to animals, a high *in-silico* accuracy of over 90% was achieved, and independent validation of the classifier (trained sing Sanger EST data) via the 454 data showed that of the well-supported (≥ 2 reads per variant, ≥ 10% of the reads support the minor variant) cases of AS, 64% (80/125) were predicted correctly, which increased to 79% (30/38) if we required ≥ 4 reads per variant while keeping the threshold of minor variant abundance ≥at 10%. When considering the well-supported additional NAGNAG AS events detected using the combined Sanger EST and 454 data, the corresponding numbers were 62% (41/66) and 75% (9/12). On the whole, our results seem to indicate that NAGNAG AS is just as common in *Physcomitrella patens* as it is in *Arabidopsis thaliana* and *Oryza sativa*, and the mechanism of NAGNAG AS is similar in plants and animals, which is also in agreement with recent work showing that NAGNAG AS shares common properties in land plants and animals [130].

Conservation of tandem splice sites

Taken together, AS events involving alternative donor and acceptor splice sites occur at a frequency comparable to that of exon skipping in humans. The majority of these splice site pairs are in close proximity [35, 119, 131, 132], and thus lead to subtle changes in the mRNA changes. In our work involving such events, we analyzed pairs of donor or acceptor sites that are 3–9 nucleotides (nt) apart (Δ3–Δ9 nt) and used the term “tandem sites” to denote these splice site pairs. The most frequent of such subtle events is AS at NAGNAG acceptors [35, 118, 119]. At the donor site, Δ4 tandem splice sites are most prominent, which is to be expected according to the donor consensus sequence GTRAGT which has a second GT four nucleotides downstream [133, 134]. It is likely that the AS mechanism at most tandem sites is based on a stochastic selection of either splice site, so-called “noisy splicing” [123]. It has been shown that the region between the branch point and the acceptor has a strong influence on the splicing ratio of alternatively spliced NAGNAG sites [135]. Targeted experimental studies have revealed functional roles for tandem splice events [136]. For example, AS at conserved tandem acceptors in human and mouse transcription factor genes (NAGNAG – or Δ3 - acceptors in *PAX3* and *PAX7*, Δ6 acceptor in *IRF2*) produces protein isoforms that differ in their ability to activate transcription [68, 137]. Conserved Δ6 donors in human *ALDH18A1* lead to protein variants with different sensitivities to ornithine inhibition [64], and produce protein isoforms of mouse *Fgfr1* that cannot bind FRS2 and are thus unable to activate the Ras/MAPK signalling pathway [62]. A splice event at a conserved Δ6 donor in human *EDA* regulates binding specificity by remodelling the properties of the receptor binding site, such that the longer protein binds only to the EDAR receptor, while the shorter variant binds only to the XEDAR receptor [65, 69]. The Δ9 donor of human *WT1* exon 9 leads to the insertion of three amino acids (KTS), with both splice forms having distinct transcriptional regulation properties - hetero- and homozygous mouse mutants lacking one of the two splice forms show severe defects in kidney development and function [63], while a mutation in this donor motif leads to Frasier syndrome in humans [60].

Though these individual studies demonstrate that several subtle splice events are functionally important, the general extent of functionally relevant subtle AS events remains unknown. Furthermore, there was a discussion whether alternatively spliced tandem sites are better conserved in evolution than constitutively spliced ones [124] since seemingly conflicting results were published for NAGNAG acceptors [118, 123]. Since it was known that alternative and constitutive NAGNAG sites differ in their preferences for specific NAGNAG motifs [118, 138], we considered the possibility that the comparison of two heterogeneous groups caused a known statistical paradox, which is better known as Simpson's paradox. This paradox, frequently encountered in biomedical studies [139], describes a situation in which a trend observed between two groups is reversed when the two groups are split into subgroups [140]. Bickel et al. [141] describe a well-known example of Simpson's paradox, involving university admission data. In this case, while the overall admission rates seemed to indicate a significant bias against female applicants, investigating all departments individually seemed to indicate the opposite, namely, a bias in favor of female applicants. The explanation of this apparent paradox is: “The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into.” [141].

Using statistical tests, we showed in the fourth work of my thesis "**Assessing the fraction of short-distance tandem splice sites under purifying selection**" that previous conflicting conclusions for the evolutionary conservation of NAGNAG acceptors [118, 123] arose from Simpson's paradox caused by substantial conservation differences between specific NAGNAG motifs. Controlling for biases, we found that alternatively spliced NAGNAG acceptors are significantly more conserved than those that are constitutively spliced. We extended the analysis to human tandem donor and acceptor sites that are up to 9 nt apart, and estimated a lower bound for the fraction of tandem sites being under purifying selection, and thus expected to have an evolutionarily advantageous phenotype.

A comprehensive resource for tandem splice sites

Even though subtle AS events involving tandem splice sites separated by a short (2-12 nucleotides) distance are frequent and evolutionarily widespread in eukaryotes, have been either omitted altogether in databases on AS, or only the cases of confirmed AS have been reported. Thus, a database which covers all confirmed cases of subtle AS as well as the numerous putative tandem splice sites (which might be confirmed once more transcript data becomes available), and allows to search for tandem splice sites with specific features and download the results, can be a valuable resource for targeted experimental studies and large-scale bioinformatics analyses of tandem splice sites. TassDB version 1 (Tandem Splice Site DataBase), which stores extensive data about alternative splice events at tandem splice sites separated by 3 nt, was a first effort towards building such a database [121].

In the fifth work of my thesis "**TassDB2 - A comprehensive database of subtle alternative splicing events**" (manuscript under review) I have contributed to the substantial revision and extension of TassDB1 to create TassDB2, containing information about tandem splice sites separated by 2-12 nt for the human and mouse transcriptomes. TassDB2 offers a user-friendly interface to search for specific genes or for genes containing tandem splice sites with specific features as well as the possibility to download large datasets. For example, the users can search for cases of AS where the proportion of EST/mRNA evidence supporting the minor isoform exceeds a specific threshold, or where the difference in splice site scores is specified by the user. The predicted impact (if any) of each event on the protein is also reported, along with information about being a putative target for the nonsense-mediated decay (NMD) pathway [82]. Links are provided to the UCSC (University of California, Santa Cruz) genome browser and other external resources. Available via http://www.tassdb.info, TassDB2 provides comprehensive resources for researchers interested in experimental studies and bioinformatics analyses of short distance (2-12 nt) tandem splice sites. We believe that TassDB2 can be of great help in future studies of subtle alternative splicing.

PUBLICATIONS AND MANUSCRIPTS

RILEEN SINHA
MICHAEL HILLER
RAINER PUDIMAT
ULRIKE GAUSMANN
MATTHIAS PLATZER
ROLF BACKOFEN

Improved identification of conserved cassette exons using Bayesian networks

BMC Bioinformatics 2008, 9:477.

Exon-skipping is the most prevalent form of alternative splicing (AS) in mammals. We used Bayesian networks in combination with discriminative features to classify orthologous human/mouse exons as undergoing either evolutionarily conserved AS, or constitutive splicing. We significantly improved the true positive rate from a previously reported 50% to 61% at a stringent false positive rate of 0.5%. This included the first use of features based on intronic splice regulatory elements and mRNA secondary structure. The improved performance was confirmed by cross validation on an independent dataset. About half of the exons which are labelled constitutive but received a high probability of being alternative by the BN, were in fact alternative exons according to the latest transcript data. We also predicted exon skipping without using conservation-based features, achieving a true positive rate of 29% at a false positive rate of 0.5%. When using identical features, the Bayesian network matched the performance of a support vector machine reported in the literature, showing that for good classification performance, discriminative features are more important than the choice of classifier.

BMC Bioinformatics

Research article

Improved identification of conserved cassette exons using Bayesian networks

Rileen Sinha[1], Michael Hiller[2,3], Rainer Pudimat[2], Ulrike Gausmann[1], Matthias Platzer[1] and Rolf Backofen*[2]

Address: [1]Genome Analysis, Leibniz Institute for Age Research – Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany, [2]Bioinformatics group, Albert-Ludwigs-University Freiburg, Georges-Koehler-Allee 106, 79110 Freiburg, Germany and [3]Department of Developmental Biology, Stanford University, Stanford, CA 94305, USA

Email: Rileen Sinha - rsinha@fli-leibniz.de; Michael Hiller - hillerm@stanford.edu; Rainer Pudimat - pudimat@informatik.uni-freiburg.de; Ulrike Gausmann - ugau@fli-leibniz.de; Matthias Platzer - mplatzer@fli-leibniz.de; Rolf Backofen* - backofen@informatik.uni-freiburg.de

* Corresponding author

Published: 12 November 2008

BMC Bioinformatics 2008, **9**:477 doi:10.1186/1471-2105-9-477

Received: 28 May 2008
Accepted: 12 November 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/477

Abstract

Background: Alternative splicing is a major contributor to the diversity of eukaryotic transcriptomes and proteomes. Currently, large scale detection of alternative splicing using expressed sequence tags (ESTs) or microarrays does not capture all alternative splicing events. Moreover, for many species genomic data is being produced at a far greater rate than corresponding transcript data, hence *in silico* methods of predicting alternative splicing have to be improved.

Results: Here, we show that the use of Bayesian networks (BNs) allows accurate prediction of evolutionary conserved exon skipping events. At a stringent false positive rate of 0.5%, our BN achieves an improved true positive rate of 61%, compared to a previously reported 50% on the same dataset using support vector machines (SVMs). Incorporating several novel discriminative features such as intronic splicing regulatory elements leads to the improvement. Features related to mRNA secondary structure increase the prediction performance, corroborating previous findings that secondary structures are important for exon recognition. Random labelling tests rule out overfitting. Cross-validation on another dataset confirms the increased performance. When using the same dataset and the same set of features, the BN matches the performance of an SVM in earlier literature. Remarkably, we could show that about half of the exons which are labelled constitutive but receive a high probability of being alternative by the BN, are in fact alternative exons according to the latest EST data. Finally, we predict exon skipping without using conservation-based features, and achieve a true positive rate of 29% at a false positive rate of 0.5%.

Conclusion: BNs can be used to achieve accurate identification of alternative exons and provide clues about possible dependencies between relevant features. The near-identical performance of the BN and SVM when using the same features shows that good classification depends more on features than on the choice of classifier. Conservation based features continue to be the most informative, and hence distinguishing alternative exons from constitutive ones without using conservation based features remains a challenging problem.

Background

Eukaryotic primary mRNAs consist of exons and introns. The mature transcript as the substrate for translation is produced by removing introns in a process called splicing. Splicing can be either constitutive, always producing the same mRNA, or alternative, by skipping of variable parts of the primary transcript.

Alternative splicing is a mechanism for producing transcript and protein diversity [1]. It is particularly widespread in higher eukaryotes, especially mammals. Various studies have estimated that up to 74% of all human genes are alternatively spliced. Large scale detection of alternative splicing is usually done using expressed sequence tags (ESTs) [2] or microarrays (reviewed in [3] and [4]). Since alternative splicing can be highly specific for tissues or developmental stages, these methods can only detect splice events that occur in the underlying probe samples with sufficient frequencies and/or are limited to by the microarray design. Furthermore, nowadays Whole Genome Shotgun (WGS) sequencing projects are churning out genomic data at a higher rate than corresponding transcriptome data – the number of ESTs in GenBank Release 161 had increased by 19% in one year, compared to a gain of 39% in the number of contigs in the WGS GenBank division [5]. Thus it can be expected that in the foreseeable future, we shall have several genomes without the level of corresponding extensive transcript coverage required to reveal the extent of alternative splicing, and hence transcriptomic and proteomic variability. Accordingly, there is a need for *in silico* methods of detecting alternative splicing. Moreover, such methods can provide further insights into the mechanisms of alternative splicing.

Exon skipping, whereby a given exon in its entirety is either included in, or excluded from the mature transcript, is the most prevalent form of alternative splicing in humans [6]. It has been shown that sequence-based features, derived from the exon and its flanking introns, can be used to predict skipping of exons that are conserved between human and mouse and alternatively spliced in both species; denoted conserved exon skipping events [7]. Previous studies have used such features with state-of-the-art classifiers such as support vector machines (SVMs) [8,9] and regularized least-squares classifier [10], and achieved success in predicting exon skipping. Other approaches use protein domain information [11] and evolutionary conservation [12-14] to detect alternative splice events.

Here, we use Bayesian networks (BNs), a state-of-the-art machine learning method, to predict conserved exon skipping events. BNs are an increasingly popular machine learning approach to data modeling and classification [15,16]. The ability of BNs to cope with features of various value ranges and to learn dependencies between features makes them especially versatile and suited to a large variety of applications. BNs allow multiple dependencies between variables, impose no fixed ordering of variables, allow integration of arbitrary features, and the network structure can be automatically learned. This makes BNs a flexible choice for biological sequence data analysis [17-21]. We introduce several novel features that distinguish alternative exons from constitutive exons, including features based on the single-strandedness of exonic splicing enhancers and silencers (ESEs and ESSs), and features involving intronic splicing regulatory elements (ISREs). By validating our classifiers on various datasets, we identify features which are discriminative irrespective of dataset-specific biases, and provide independent measures of the predictive power of the BNs.

Even though conservation based features have proved to be among the most discriminative features for predicting exon skipping, it is desirable to be able to predict alternative splicing using only information from a single genome. We show that our approach can still predict exon skipping without using conservation-based features.

Methods

Datasets and genome browser

We used the dataset of [8], henceforth called dataset D1, consisting of 243 alternative and 1,753 constitutive exons, kindly provided by Gideon Dror. In this dataset, constitutive exons are supported by at least four ESTs each in human and mouse with no EST evidence for exon skipping, whereas alternative exons are skipped in both species. The second dataset is the ACESCAN training set [10], henceforth called dataset D2, which comprises 5,069 constitutive and 241 alternative exons. For validation purposes we use the genome builds hg18 for human and mm9 for mouse of the UCSC Genome Browser [22] were used.

Features for machine learning

In total, we used 365 features in this study (Table 1). Thereof, 228 were previously used by [8]: (1) exon length, (2) symmetry, that is, divisibility of exon length by 3; (3) percent identity of the alignment between the exon and its mouse ortholog; (4–7) length of and percent identity of the best local alignment between the up- and downstream 100 nt intronic flanks and their mouse orthologs, which are in total four features; (8–199) trimer counts for the exon and the 100 nt flanking intronic regions, which are a total of 64 × 3 = 192 features; (200) intensity of the polypyrimidine tract (PPT) as the number of pyrimidines in the window -19 to -4 from the 3'ss; and (201–228) nucleotides at the 5'ss positions -3 to -1 and +3 to +6, which are a total of 7 × 4 = 28 features. The features (1–7), used for

Table 1: Features for machine learning used in this study

Feature subset	Number of features	Motivation	First use
Exon: length, symmetry, and identity with mouse ortholog	3	Alternative exons tend to be shorter, frame-preserving, and more conserved compared to constitutive exons	[7]
Conservation of intronic flanks: length/ identity of the best local and identity of the global alignment	2 × 3	Alternative exons tend to have higher conservation in their intronic flanks	[7,10]
Conservation in a 12 nucleotide region spanning the 3' and 5'ss	2	As alternative exons and their intronic flanks are more conserved, this may in particular concern the exon/intron boundaries	This work
PPT intensity	1	Alternative exons tend to have weaker PPTs	[8]
Nucleotides at seven positions flanking the 5'ss	4 × 7	Alternative exons tend to have specific nucleotide preferences near the 5'ss	[8]
Frequency of di- and trimers in the exon and flanking introns	3 × 16 3 × 64	Motifs which are part of splice regulatory motifs might differ in their abundance in alternative and constitutive exons	[8] (trimers), this work (dimers)
Splice site strength of 3'and 5'ss	2	Alternative exons tend to have weak splice sites	[10]
Length of flanking introns	2	Alternative exons tend to be flanked by long introns	[10]
GC content of exon and intronic flanks	3	GC-poor regions tend to promote alternative splicing	This work
Features based on NI scores	24	Alternative exons tend to have fewer ESEs and more ESSs	This work
Features based on PU values	15	Single-stranded motifs are likelier to bind to regulators	This work
PTB-binding sites	6	PTB is a regulator alternative splicing	This work
Features based on ISREs	8	Alternative exons tend to have more ISREs in their intronic flanks	This work
Density of various motifs	22	Several motifs are known to be associated with alternative splicing	This work
Combination features	7	Combining features can capture more information	This work

Note that the total number of features used is 365 whereas the sum of the entries here is 378, because some features have been counted in more than one category (for example, in PU value and NI score related features).

the first time in [7], were kindly provided by Gideon Dror, along with the dataset D1.

We also used six features from [10]: the percent identity of the global alignments between up- and downstream 100 nt intronic flanks and their mouse orthologs, lengths of the upstream and downstream flanking introns, and the strength of the 3' and 5' splice sites (3'ss and 5'ss) computed by MAXENTSCAN [23]. We used the programs "needle" and "water" from the EMBOSS software suite [24] for aligning the exons and the intronic flanks with their mouse orthologs and computing the conservation based features.

Among the new features we added were dinucleotide counts for the exon and the 100 nt intronic flanks, a total of 16 × 3 = 48 features. As it has been shown that exon skipping is more prevalent in regions of low GC content [25], we used the GC content of the exon and the intronic flanks as three additional features.

To use features based on ESEs and ESSs, we applied neighbourhood inference (NI) scores [26]. Briefly, each hexamer has an NI score between -1 and 1, with negative scores indicating a tendency towards acting as an ESS, and a positive score, a tendency to act like an ESE. Hexamers with a score of 1 or -1 are considered "trusted" ESEs and ESSs, respectively, and those with a score of greater than 0.8 or smaller than -0.8 are considered to have "strong" ESE or ESS activity. We used the density of NI scores, defined as the number of hexamers with NI scores 1, ≥ 0.8, > 0, < 0, ≤ -0.8, -1, normalized by the number of hexamers in the exon (6 features). Additionally, the distribution of ESEs and ESSs may have a bearing on splicing as well. Therefore, we used the variance of NI scores for "trusted" and "strong" ESEs and ESSs (2 features). Since the density of ESEs and ESSs near splice junctions has been suggested to be important in determining splicing outcome [4,27-29], we also measured the densities in the first and last 50 nucleotides of the exon (for exons shorter than 50 nt, the entire exon was used; 2 features).

We also designed features using very recently published datasets of conserved ISREs enriched in the upstream and downstream intron flanks of all exons, as well ISREs enriched in upstream and downstream introns flanking alternative exons [30]. We used the density of ISREs from these four lists in both upstream and downstream 100 nt flanking intronic regions, giving us eight novel features.

Secondary structure can influence alternative splicing [31]. The single-strandedness of ESE, ESS or ISRE motifs was characterized using PU (Probability of being Unpaired) values [32], which represent the probability that all the bases in the given motif are unpaired. Since local RNA folding is influenced by the length of the sequence context [33], we minimized potential biases by using 11 to 30 nt symmetrical context lengths up- and downstream of a given hexamer, and computing the average of the 20 PU values thus obtained [34]. We pre-computed PU values in this manner for all the hexamers in the exons, and combined the NI scores with PU values. Various thresholds were used for absolute NI score value (1, = 0.8, > 0) and a PU value of 0.6. Two kinds of combinations were used: (i) a "Boolean" combination, that is, counting the number of hexamers with NI and PU values both above the thresholds; and (ii) the product of NI and PU values (4 features). Similarly, we used PU values in conjunction with ISRE information to characterize the single-strandedness of intronic splicing regulatory elements (4 features).

Mutations around the splice junctions can effect splicing. Therefore, we designed a feature to measure how well the immediate neighbourhood of the splice junctions was conserved. We formed two 12-mers consisting of the bases from positions -6 to +6 around the 3'ss and the 5'ss. The number of identical nucleotides between the human and mouse 12-mers result in two new features.

We also used several motifs from a recent study characterizing conserved motifs associated with constitutive and alternative splicing [35]. However, depending on the partition, these features were either not discriminative or weakly so, indicating that they are important only for a small minority of the alternative exons.

To count the number of binding sites for the Polypyrimidine-tract-binding protein (PTB), a well-studied repressive regulator of alternative splicing [36], we counted the simplest known motifs for its binding sites – UCUU and CUCUCU, as well as the sum. The density of PTB binding sites in the 100 nt intron flanks and the exon gives nine features.

Lastly, we used novel features derived from features already known to be discriminative. For example, while it is known that skipped exons tend, on the average, to be shorter than constitutive exons, it has been shown that long exons can be skipped if flanked by very long introns [37]. Furthermore, it is possible that the shorter the exon is with respect to the flanking introns, the harder it is for the spliceosome to reliably recognize it. Consequently, we used the ratio of upstream and downstream intron length to exon length, as well of the intron lengths, as three features. We also used the pairwise products of human-mouse identity of the exon and each 100 nt intron flank as well as of the exon and both flanks, in order to capture information about simultaneous conservation of the exon and the intronic flanks (four features).

Information gain and information gain ratio

To compare the information content of the features, we used information gain, and information gain ratio, which are established measures of the usefulness of features in the field of machine learning [38]. The formula for information gain is:

$$\mathrm{IG}(\mathrm{Class} \mid \mathrm{Feature}) = \mathrm{H}(\mathrm{Class}) - \mathrm{H}(\mathrm{Class} \mid \mathrm{Feature})$$

where H(Class) is the entropy of the class variable, and H(Class | Feature) is the conditional entropy of the class variable, given the feature. We used the WEKA package [38] for computing information gain and information gain ratio, in order to rank the features according to how informative they were.

Bayesian networks

We used the algorithms for feature selection, model learning and classification as described in [17], and made available *via* the public webserver BioBayesNet [39].

Feature subset selection

Given a training set, we selected features in a three step procedure. First, we use an entropy based method developed by [40] to find partitions of the feature ranges which best separate the given classes (in the following called "discretizer"). Features for which the entire feature range is partitioned into at least two intervals, such that the distributions of the two classes differ significantly in these intervals, are called "discriminative" and they are the basis of further analysis. On the other hand, those features for which no such intervals are found are essentially non-informative, or "non-discriminative" features for our purposes.

Once the discretization algorithm has chosen the set of discriminative features, an optimal (in the local sense) subset can be selected using the sequential floating feature selection (SFFS) method [41]. Briefly, this algorithm starts with an initially empty feature subset, and at each step, adds the feature which most improves a specific quality measure. After this addition, all previously added features are deleted from the subset, unless doing so worsens the quality measure. This is done in order to avoid getting trapped in local minima. The algorithm stops when neither inserting new features nor deleting existing ones improves the quality measure provided by the subset.

Thirdly, one can enforce inclusion or exclusion of any given feature manually. The manual feature selection consists only of removing a few "weak" features (as measured by low information gain, or negligible information loss when they are omitted for classification purpose) as they are unlikely to generalize well to unseen data; and addition of a few "strong" features (as measured by high information gain), which were selected by the discretizer but not by the SFFS algorithm.

Learning the Bayesian network

We restrict the structure of the BNs by using the so-called tree-augmented naïve Bayes (TAN) structure [42]. In a naïve Bayes classifier/network, the attributes are assumed to be independent, given the class, that is, the node representing the class variable is a parent of all other nodes, and there are no other edges in the network. A TAN classifier augments the underlying naïve Bayes classifier by allowing at most one additional parent per node, that is, each node is the child of the class attribute node, and of at most one more node. We use TAN classifiers because while learning the best BN structure, given some training data, is in general an NP-hard problem [43], for TAN networks there exist efficient structure-learning algorithms that reduce the problem of determining the optimal tree structure to finding a maximum-weighted spanning tree [44]. Once the structure of the network has been learned, the (conditional) probability distributions over the feature values of each feature (given the class label and optionally the value of the parent feature) are estimated in a straightforward manner from count statistics derived from learning data. Finally, Bayesian inference of marginal probabilities can be approximately calculated by the efficient technique of variable elimination [45].

Data partition

Given a dataset (D1 or D2), we partitioned the data into three equal parts as carried out in [8]. Then, in turn, we used two-thirds of the data to train the BNs, and the remaining one-third was used for testing. The test set remained untouched while the training set was used for discretization, feature selection, and learning the BN [39]. Finally, the BN which had been learned on the training set was used to classify the samples in the test set. This procedure was repeated twice for the other two one-thirds, and the average of the three runs was taken as the final performance. For comparing 2-fold, 3-fold, 5-fold and 10-fold cross-validation, we used WEKA [38].

Results and discussion

Improved prediction of conserved cassette exons by Bayesian networks

As pointed out by [8], good performance at low false positive rates is especially important for the task of distinguishing alternative exons from constitutive exons on a genome-wide scale, since the latter comprise the majority of exons. Furthermore, a low number of false positives is critical in case of experimental verification of predictions. To this end, we measure the true positive rate (TP) at false positive rate (FP) of 0.5%, and call it $TP_{0.5}$. We also compute the receiver operating curve (ROC) and measure the area under the ROC curve (AUC), which is a standard measure of the quality of a classifier [46].

We used the dataset and the cross-validation scheme described in [8]. This dataset contains 243 alternative and 1,753 constitutive exons and is called D1 in the following. The overall performance obtained, using novel features in addition to those described in the literature (Table 1), was $TP_{0.5}$ = 61%, and AUC = 0.94 (Figure 1), compared to $TP_{0.5}$ = 50%, and AUC = 0.93 reported in [8] using SVMs. This substantial improvement demonstrates that many of the novel features are informative and discriminative for conserved exon skipping events.

Feature selection

The number of features studied in machine learning tasks is often very high, and many (possibly most) of them might be irrelevant, or redundant [38]. Therefore, it is customary to preprocess the data in order to select a useful subset of features – this is called "feature selection". Feature selection can be carried out in three stages within the BioBayesNet framework [39]. Firstly, a "discretizer" applying the algorithm of Fayyad and Irani [40] discards features for which no suitable discriminative intervals are found. Secondly, the sequential feature subset selection (SFFS) algorithm [41] can be applied to select a subset of the remaining features. Thirdly, one can enforce inclusion or exclusion of any given feature manually. The manual feature selection (on D1 and D2) typically involved the addition and removal of 5 or fewer features each, given a feature subset of 20–30 features obtained using the two automated approaches.

The performance on D1 using only feature selection using the "discretizer" was $TP_{0.5}$ = 39%, and AUC = 0.93. Using the SFFS algorithm for further feature selection resulted in $TP_{0.5}$ = 47%, and AUC = 0.94, whereas the manual inclusion/exclusion of features gave the final performance of $TP_{0.5}$ = 61%, and AUC = 0.94. This illustrates that the overall quality of classification, as measure by the AUC, is quite robust, and we get good performance even when only the "discretizer" is used, but the performance at low false positive rates is quite sensitive to small changes in

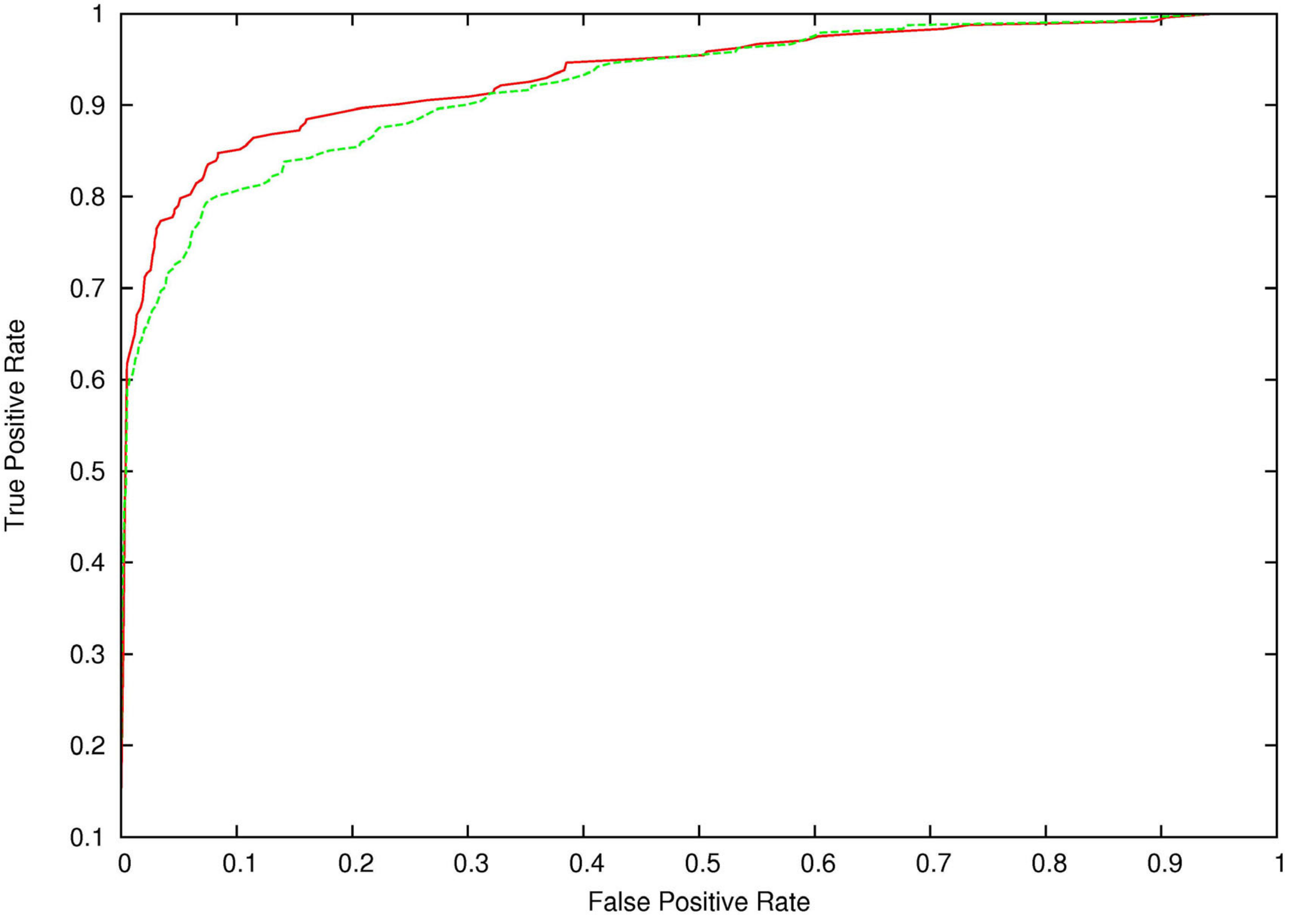


Figure 1
ROC plot showing the average performance of the 3-fold cross-validation on datasets D1 (red line) and D2 (green line).

the feature subset, so the other two methods of feature selection result in significant improvement. We note that manual feature selection is only needed to improve $TP_{0.5}$ – if we consider more global measures of classification performance such the AUC or balanced sensitivity and specificity, the automated feature selection methods suffice. Using only automated feature selection, we routinely achieve AUC values in the 0.93–0.96 range, and balanced sensitivity and specificity in the 87%–91% range.

Discriminative features

ESEs and ESSs are motifs bound by proteins which either enhance or suppress splicing. It has been shown that alternative and constitutive exons differ in the density of ESEs and ESSs [29]. We used Neighborhood Inference (NI) scores to infer ESE and ESS activity for all hexamers [26]. We used the density of ESEs and ESSs, with various thresholds for the NI scores. The constitutive exons have a slightly higher density of ESEs than do alternative exons (median 0.266 vs. 0.254), as well as ESSs (median 0.0694 vs. 0.0679) This was also confirmed using other ESE/ESS datasets [27,29] and is in agreement with previous studies [26,27,29,30]. Depending on the split, the density of ESEs and ESSs was either not discriminative, or weakly so. Varying the threshold of the NI score did not change this. On the other hand, some of the novel features using NI scores were discriminative on most splits – for instance, the average of all positive NI scores, as well as the average of all negative NI scores. Similarly, the average of all "strong ESEs" (NI score ≥ 0.8) and "strong ESSs" (NI score ≤ -0.8) were discriminative features. However, the density of ESEs and ESSs near the splice sites was not found to be discriminative.

Splicing regulatory elements are found in introns as well [47]. Consequently, we also designed features using very recently published datasets of conserved intronic splicing regulatory elements (ISREs) [30]. Similar to ESE and ESS based features, we also used the density of ISREs in the upstream and downstream intronic flanks. Four sets of

ISREs (enriched in the upstream and downstream flank of all exons as well as enriched in the flanks of alternative exons) are given in [30]. We found three of these sets to be discriminative (ISREs from downstream intronic flanks of all exons were not discriminative). For these three discriminative features, alternative exons have a higher density of ISREs than constitutive exons, which agrees well with the finding that the set of ISREs has an overlap with ESSs, and thus many of them may have silencing tendencies [30].

Secondary structure can influence alternative splicing [31]. To the best of our knowledge, existing methods to predict alternative splice events do not use secondary structure related properties. Previously, we found that functionally important splicing motifs are preferentially located in single-stranded mRNA secondary structures [34] and that *ab initio* motif finding benefits from taking the single-strandedness of motif occurrences into account [48]. Thus, we use features based on a measure of the single-strandedness of ESE, ESS or ISRE occurrences, reasoning that single-strandedness of enhancer and silencer occurrences influence the propensity of proteins to bind them. Interestingly, we found that the single-strandedness of ESE motifs is informative. The density of single-stranded ESEs is higher in constitutive than in alternative exons (0.0194 vs. 0.0177, using a PU value of 0.5 for single-strandedness). Moreover, the information gain of this feature was more than that of the ESE density feature (0.0170 vs. 0.0096). As single-stranded motifs are expected binding sites for splicing regulatory proteins, this observation adds to previous evidence that mRNA secondary structures influence alternative splicing [34].

The density of PTB binding sites in the exon and the upstream 100 nt intronic flank were weakly discriminative, indicating that they are important only for a small minority of the alternative exons. The density of PTB binding sites in the downstream 100 nt intronic flank was not discriminative.

The conservation around the splice site, as measured by the number of human-mouse identical positions in a window of 12 nt (6 on either side) around the exon boundaries, is a highly discriminative feature, despite other features already capturing both conservation information as well as splice site strength. It is interesting that while alternative exons have weaker splice sites, they have stronger conservation around the splice junctions. While only 17.6% of the constitutive exons have identical matches from positions -6 to +6 at the 3'ss, the corresponding figure for alternative exons is as high as 54.7%. At the 5'ss, the corresponding numbers are 30.0% and 60.3%, respectively. This is consistent with a previous study [49].

The GC content of the upstream intronic flank was found to be a useful discriminative feature, and was lower for flanks of alternative exons than of constitutive exons (median values 0.39 vs. 0.42), in agreement with previous studies [25]. However, neither the GC content of the exon nor of the downstream intronic flank was found to be discriminative.

We tested if the di- and trimer (2 and 3 nt words) frequency in exons and intron flanks is different for alternative and constitutive exons. We found that the frequency of di- and trimers in exons is often much more discriminative than the intronic di- and trimer frequencies. This suggests that splice regulatory elements governing exon skipping are more common in alternative exons than in introns flanking them.

Apart from introducing novel features, we also used features derived from known features. These combinations were often more informative than the individual features. For example, the ratios of intron lengths to exon length were more informative than the lengths themselves. The ratio of the length of the downstream intron to the exon length was an especially useful feature, suggesting that exon skipping may occur when the spliceosome finds it difficult to accurately "spot" an exon upstream of a relatively much longer intron.

Splice site strength, first used by [10], was also found be a discriminative feature, with alternative exons having both weaker 3'ss as well as 5'ss than constitutive exons (median scores 7.86 vs. 8.76 and 8 vs. 8.68, respectively).

Most informative features

Next we asked which were the most informative features using the information gain, a well established measure in machine learning. Information gain is the reduction in the entropy of the class variable, given the feature. While information gain is a good measure of the quality of features, it tends to prefer features with a large number of possible values [38]. A measure which avoid this is the information gain ratio, which divides the information by the information of the feature itself, thus penalizing features with a high inherent information. The top ten features according to the information gain and the information gain ratio criteria are given in Table 2. Two features, exon identity and length of best alignment in the upstream intron flank, appear in both lists.

Table 3 shows the top ten combination features according to the information gain criterion. Seven of these are more informative than any of the features that were combined to obtain these features, while the other three (the ratio of the intron lengths and the exon length, and the sum of the number of the two types of PTB binding sites) are more

Table 2: Top features according to information gain and information gain ratio (excluding combination features)

Rank	Feature	Information Gain	Feature	Information Gain Ratio
1	Length of best alignment in the upstream intron flank	0.169	Abundance of GA in exon	0.172
2	Upstream intron flank conservation	0.169	Density of single stranded ESEs in exon	0.151
3	Identity of best alignment in the upstream intron flank	0.142	Exon identity	0.128
4	Downstream intron flank conservation	0.138	Average of positive NI scores in exon	0.118
5	Length of best alignment in the downstream intron flank	0.138	Length of best alignment in the upstream intron flank	0.117
6	Exon identity	0.120	Density of AC in exon	0.115
7	Identity of best alignment in the downstream intron flank	0.088	Average of negative NI scores in exon	0.112
8	Exon length	0.080	Density of CT in exon	0.111
9	Matches in 12-mer near 3'ss	0.066	ESE density in exon	0.104
10	Symmetry	0.042	Length of best alignment in the upstream intron flank	0.103

informative than one of the two features, but less than the other. Not surprisingly, the combinations of conservation related features have a very high information gain (top four).

Table 4 shows the ten most informative trimers in the exon and in the intronic flanks according to information gain. Note that the trimers in the exon have a higher information gain, a trend which is also true when looking at all possible 64 trimers in the exon and the intronic flanks. This disagrees with the conclusion of the previous study [8], which used a different feature ranking criterion.

A Bayesian network with an optimized set of 34 features

All three methods of feature selection available in the BioBayesNet framework were used to arrive at an optimized subset of 34 features. A performance of $TP_{0.5}$ = 61% (65%, 61%, and 56% for the 3-fold cross-validation), and AUC = 0.94 (0.94, 0.94 and 0.94) was achieved using the same subset of 34 features with each fold. The BN learned on the entire dataset with the same features, with 34 nodes and 33 edges, can be seen in Fig. 2. We would like to point out some interesting edges in this network which confirm and may extend our biological knowledge of the splicing process:

- *"length of the best local alignment of the upstream intron flank and its mouse ortholog" (node 3) and "density of intronic splice regulatory elements (ISREs) enriched in introns flanking AS exons, in the upstream intron flank" (node 18)*: Since alternative conserved exons (ACE) tend to have longer conserved regions and a higher density of ISREs in their intron flanks, this is a biologically meaningful dependence.

- *"ratio of the lengths of the downstream intron and the exon" (node 5) and "sum of the MAXENTSCAN scores of the 3' and 5' splice sites" (node 12)*: ACEs tend to have high ratios of intron to exon lengths, and weak splice sites, when compared to constitutive exons [10].

- *"density of single-stranded ESEs" (node 16) and "density of TCTT in exon" (node 20)*: ACEs are enriched in exons with multiple occurrences of TCTT, which is a binding site of the splice-repressor, PTB [36], and tend to have a lower

Table 3: Top combination features according to information gain

Rank	Feature	Information Gain
1	Product of identities of exon and both intron flanks	0.208
2	Product of identity of both intron flanks	0.196
3	Product of identities of exon and upstream intron flank	0.181
4	Product of identities of exon and downstream intron flank	0.153
5	Ratio of the downstream intron length to exon length	0.051
6	Ratio of ESE density to ESS density	0.029
7	Sum of splice site scores	0.023
8	Ratio of the upstream intron length to exon length	0.022
9	Ratio of trusted ESE density to trusted ESS density	0.010
10	Density of putative PTB binding sites in exon	0.008

Table 4: Top trimers in the exon and intron flanks according to information gain

Rank	Exon Trimer	Information Gain	Intron	Trimer	Information Gain
1	TCC	0.034	upstream	TTC	0.016
2	ATG	0.031	downstream	AGG	0.014
3	CCT	0.029	downstream	GAG	0.012
4	TCG	0.028	upstream	TTT	0.012
5	CAT	0.028	upstream	TCT	0.012
6	AAG	0.027	downstream	GGA	0.012
7	GTA	0.027	downstream	TTT	0.011
8	GAC	0.026	upstream	GAG	0.011
9	GAT	0.026	upstream	AGG	0.011
10	CAA	0.026	upstream	CAG	0.009

density of single-stranded ESEs when compared to constitutive exons.

- "*MAXENTSCAN score of the 5'ss" (node 11) and nodes representing positions in the 5'ss region (nodes 8 and 9)*: The node representing the strength of the 5'ss has an edge to the node representing the binary variable "Gat5SSplus4", which indicates whether the nucleotide at the position +4 is a G or not, and this node has an edge to the node representing the variable for a T at position +4. Both of these nucleotides are different from the splice site consensus at their respective positions, and thus contribute to lowering the splice site score. Furthermore, it is also known that there are dependencies among the 5'ss positions [50].

- "*density of intronic splice regulatory elements (ISREs) enriched in intronic flanks of AS exons, in the upstream intron flank" (node 18) and "Abundance of CGG in the exon" (node 29)*: ISREs which are enriched in the flanks of alternative exons tend to be CG-rich [30], so the link to CGG motifs in the exon might indicate a subclass of alternative exons found in CG-rich regions.

- "*abundance of CCA in the exon" (node 28) and "average of negative NI scores in exon" (node 13)*: These nodes correspond to features representing the density of the trimer CCA in the exon, and the average NI score of all hexamers with negative NI scores, i.e. ESS-like tendencies. The CCA motif is ~35-fold less frequent in ESSs than ESEs (occurs in 71 of 979 ESEs and only 1 of 496 ESSs), so the BN captures the association of CCA abundance with the average of negative NI scores.

- "*abundance of CGG and GCA in the exon" (nodes 29 and 31) and "ratio of ESEs to ESSs in the exon" (node 18)*: These nodes represent the density of the trimers CGG and GCA in the exon, and the ratio of "trusted" ESEs and ESSs (scores of 1 and -1, respectively). The motif CGG occurs in 7.5% (50 of 666) of the trusted ESEs, but in only 1% (4 of 386) of the trusted ESSs. Similarly, the motif CGG occurs in 10.5% (70 of 666) of the trusted ESEs, but in only 2.1% (8 of 386) of the trusted ESSs. Thus there is a correlation between the abundance of the motifs CGG and GCA in the exon, and the ratio of ESEs to ESSs.

Some of the other edges can be explained in a trivial manner, for instance those involving the density of overlapping motifs (e.g. nodes 28 and 21, and 28 and 34). We note that one must be careful in interpreting the edges, as not all of them may lend themselves to meaningful biological interpretation. While not all edges can be interpreted with biological knowledge, they definitely help in our classification since a classifier omitting all edges (naïve Bayes) performs worse [8].

Comparison of 2-fold, 3-fold, 5-fold, and 10-fold cross-validation

We used 3-fold cross-validation in order to compare our results to [8], who did the same. However, since it is common in machine learning to use 2-fold, 5-fold, 10-fold, or "leave one out" (LOO) cross-validation, we compared the performance of these different approaches on the dataset D1, using the WEKA package [38], and the optimized set of 34 features described above. The results for the 2/3/5/10/LOO cross-validations were: $TP_{0.5}$ = 57%/60%/57%/58%/59%, and AUC = 0.95/0.95/0.95/0.95/0.95.

Performance using the same features as the SVM

To assess the factors behind the improved performance of BNs, we used the same 228 features as reported in [8], and obtained the same overall quality of prediction (AUC = 0.93) and slightly improved $TP_{0.5}$ (51% vs. 50%). This indicates that accurate classification of conserved exon skipping depends more on the features used rather than the choice of classifier.

Performance of Bayesian networks on a second dataset

Next, we tested our approach on a different dataset of conserved exon skipping events, the ACESCAN training set [10] henceforth called dataset D2, which comprises 5,069 constitutive and 241 alternative exons. Using the basic set of 228 features [8], the BN achieved values of $TP_{0.5}$ = 52%,

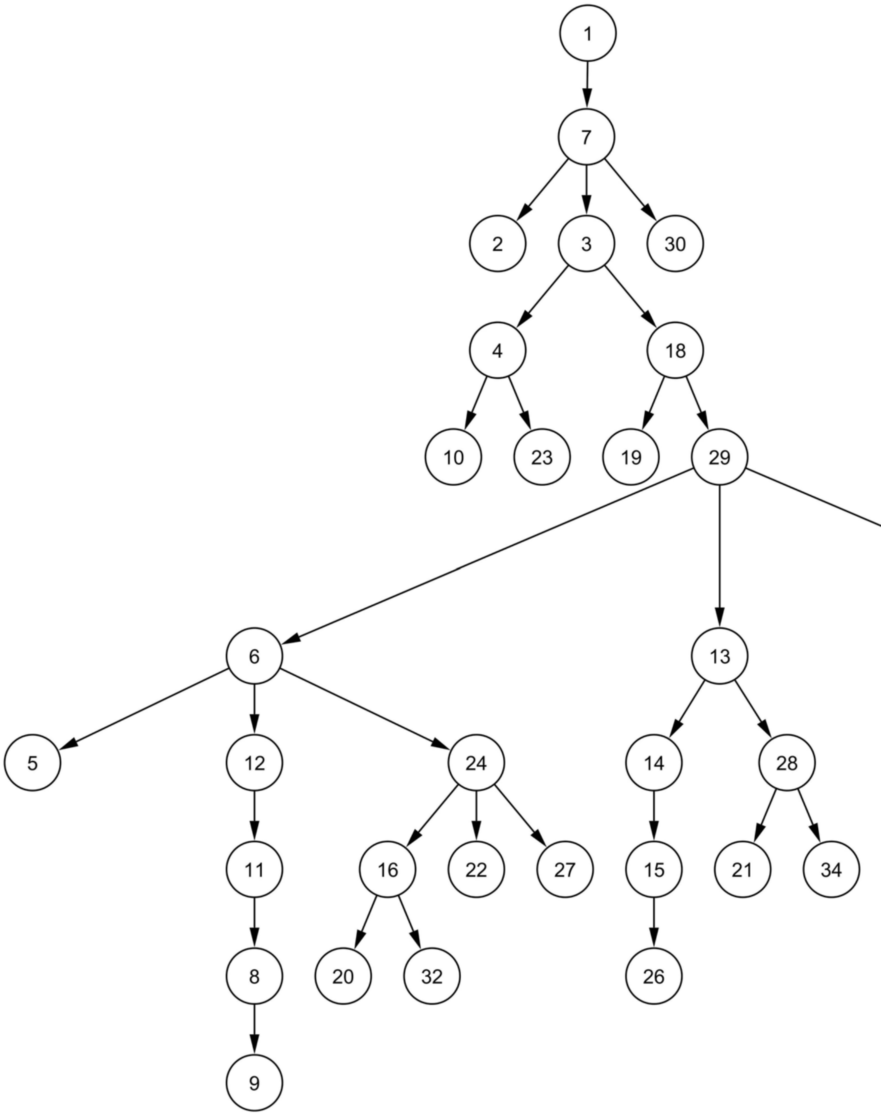


Figure 2

34-feature Bayesian network. Note that BN in fact has 35 nodes. The class node, which has an edge to all other nodes and makes the actual number of edges 67, is omitted for ease of visualization. Thus, this is just the augmenting tree in the TAN classifier. The features associated with the nodes are as follows: 1: 1 if exon length is divisible by 3, otherwise 0. 2: Length of the best alignment in the 3' 100 nt intronic region. 3: Length of the best alignment in the 5' 100 nt intronic region.4: Percent identity of the best alignment in the 5' 100 nt intronic region. 5: Length of the 5' intron. 6: Ratio of the lengths of the 3' intron and the exon. 7: Product of the identities of the exon and both 100-nt intronic flanks with their mouse orthologs. 8: 1 if G at +4 of the 5'ss, otherwise 0. 9: T at +4, 10: A at +6; 11: MAXENTSCAN score of the 5'ss. 12: Sum of the MAXENTSCAN scores of the 3' and 5'ss. 13: Average of the NI scores of all the hexamers with a negative NI score. 14: Variance of the NI scores of all the hexamers with a "strong" (≥ 0.8 or ≤ -0.8) score. 15: Average of the NI scores of all the hexamers with a "strong" (≤ -0.8) negative score. 16: Density of single-stranded (PU value ≥ 0.6), "trusted" ESEs (NI score = 1). 17: Ratio of the number of "trusted" ESEs (NI score = 1) to the number of ESSs (NI score = -1). 18: Density of ISREs enriched in the flanks of AS exons, in the 5'intron flank. 19: Density of single-stranded (PU value ≥ 0.6), intronic splice regulatory elements (ISREs) enriched in the flanks of AS exons, in the 5'intron flank. 20: PTB-binding site TCTT density in the exon. Dimer density in the exon:21:CC, 22: GA; 23: Dimer GA density in the 3' intron flank; Trimer density in the exon: 24: AAG, 25: AGG, 26: ATG, 27: CAA, 28:CCA, 29: CGG, 30: CTC, 31: GCA, 32: GGT, 33: TAG, 34: TCC.

and AUC = 0.92. After incorporating the novel features, and performing feature selection as described above, the best performance achieved on D2 was: $TP_{0.5}$ = 59%, and AUC = 0.93 (Figure 1).

Thus, we achieve a good performance, similar to that on D1, on the dataset D2 as well. However, the number of discriminative features is smaller than for D1. This trend continues with the addition of novel features – of all the 365 features, typically 110–130 are discriminative on a 2/3 split of D1, whereas only 65–80 are discriminative on a 2/3 split of D2. A possible reason for this could be the different criteria used in the construction of the two datasets, resulting in possibly different extents of corruption of the sets of constitutive exons by alternative exons, because the dataset D1 requires 4 identical ESTs for an exon to be considered constitutive, whereas the dataset D2 does not. Furthermore, D1 has more exaggerated differences among the two classes for several features – for example, while 74% of alternative exons preserve the reading frame compared to 37% of the constitutive exons, the corresponding numbers for D2 are 67% and 39%. Thus, the subset of conserved exon skipping events in D1 seems to be characterized by more strongly discriminative features.

Cross-validation by training and testing on the two independently constructed datasets

It is usual in machine learning to divide the available data into training and testing partitions, and optimize the classification using these. It is then assumed that similar performance can be achieved on other datasets of a similar nature. However, given that there are often differences in the way independent datasets are prepared by different groups of scientists, it may be optimistic to presume this. We suggest that testing on an independent dataset is likely to give a better indication of the level of performance that can be expected when scaling to a genome-wide prediction. To use D1 and D2 for this purpose, we removed from D1 the exons already present in D2 – leaving 201 alternative and 1,654 constitutive exons in D1. To minimize any biases introduced by different ratios of the numbers of samples in each class, we then randomly sampled constitutive exons from D2 to have the same ratio (8.23:1) of constitutive to alternative exons, leaving 241 alternative and 1,984 constitutive exons in D2. We then used the optimal feature subsets obtained on D1 and D2 earlier to train BNs on the respective entire datasets. When we used the BN trained on D1 to test D2, the performance achieved was $TP_{0.5}$ = 27%, and AUC = 0.88. The corresponding performance achieved with training on D2 and testing on D1 was $TP_{0.5}$ = 26%, and AUC = 0.91. While an AUC value of 0.91 (or even 0.88) indicates good overall classification, this is less than the 0.94 achieved when tested on unseen data from the same source. The effect on $TP_{0.5}$ is quite dramatic. We think that these figures might be a more accurate estimate of what to expect when a classifier is used to classify independently produced data. Performance will tend to be (at least) slightly worse on independently produced data than on unseen data from the same source, something which is true of all classifiers in general.

Assessing over-fitting

To assess whether our increased performance is due to over-fitting, we randomly permutated the labels 'alternative' and 'constitutive' between the data points and trained the BN on the relabelled datasets D1 and D2. In case of overfitting, we would still expect a good performance, while the AUC value of a random classifier should be close to 0.5 [38].

After relabelling, most features are no longer discriminative. In fact, only 29 features remained discriminative, and these were the same for both datasets – symmetry, and the 28 features describing the positional biases in the 5'ss region. The AUC achieved was 0.51 on dataset D1, and 0.49 on D2. This shows that our approach has no problems with over-fitting.

To further rule out overfitting, we used a random three way split: 60% of the data for training, 20% for validation and optimization, and 20% for testing. We obtained $TP_{0.5}$ = 63% and AUC = 0.94 on the validation set; using the same set of features, the performance on the test set was $TP_{0.5}$ = 59% and AUC = 0.95. Using this "train-validate-test" approach on D2, we obtained $TP_{0.5}$ = 58% and AUC = 0.94 on the validation set, and $TP_{0.5}$ = 60% and AUC = 0.93 on the test set. Since the performance on both datasets is very similar to the performance achieved using our three-stage feature selection approach, we conclude that the improvement is not mainly due to manual feature selection. However, manual selection is not ideal, and an automated feature selection algorithm designed to optimize performance in the low false-positive region would be more satisfying. This is one of the possible future directions of work.

As a first approach to entirely automated feature selection, we performed the following experiment: we randomly chose 75% of D1 for training, and 25% for testing. Feature selection was done using only the training part, and the test part was touched only once at the very end of the procedure. The feature selection was as follows: starting with the full set of features, we iteratively discarded one feature at a time, and performed 10-fold cross-validated classification using a BN (TAN) with the remaining features. Features were discarded in order of increasing information gain, that is, the least informative features were discarded first. We re-inserted a feature only if at least one of $TP_{0.5}$ or AUC decreased as a result of omitting it. This was done only in one pass, and features once discarded, were not considered again. This is clearly not an optimal strategy, and leads to bigger feature subsets than the approach used before, but still yields good results. Using the subset of 50 features thus obtained on D1 led to performance of $TP_{0.5}$ = 54% and AUC = 0.94 on the training set, and $TP_{0.5}$ = 57% and AUC = 0.91 on the test set. On D2, this approach yielded a subset of 35 features and a performance of $TP_{0.5}$ = 56% and AUC = 0.92 on the training set, and $TP_{0.5}$ = 52% and AUC = 0.94 on the test set. Thus, we can also obtain good performance on unseen data using a feature selection strategy which, though suboptimal, is easy to automate.

Moreover, we also used the feature sets obtained in the "train-validate-test" setting with a naïve Bayes classifier (NBC) and obtained $TP_{0.5}$ = 47% and AUC = 0.93 for a 10-fold cross validation on D1, and $TP_{0.5}$ = 43% and AUC = 0.92 for a 10-fold cross validation on D2, which are both better than the performance using NBC reported in [8] ($TP_{0.5}$ = 37% and AUC = 0.89). Compared to the BNs, NBCs achieve a higher sensitivity but lower specificity. This indicates that the novel features help in improving classification performance, and similar improvements should be possible using other classifiers like SVMs, Neural networks and so on.

False positives with high posterior probability are likely true alternative exons

Next, we carefully looked at exons that are labelled constitutive but obtained a high posterior probability of being alternative exons from the BN. Since they seemed to be more similar to ACEs than to other constitutive exons, we hypothesized that newer EST/cDNA data provides evidence for exon skipping, or any other kind of alternative splicing at these exons. Out of 1,753 exons in D1 that were labelled constitutive, 14 were assigned a P(ACE) – posterior probability of being an ACE – of 0.7 or more. A detailed inspection using the UCSC genome browser [22] revealed that seven have EST and/or mRNA evidence of alternative splicing in at least one of human and mouse (six of these seven are cassette exons) and that two of them are alternatively skipped in both species, that is, have evidence of being ACEs. Of the remaining seven exons, one has evidence of being a cassette exon in orangutan (Additional file 1).

The results on D2 are even more impressive – there are 15 exons labelled constitutive and with P(ACE) ≥ 0.7, of which 13 have evidence of exon skipping or another alternative splicing event (seven are cassette exons in at least one of human and mouse; five are ACEs; Additional file 1).

Thus, most FP predictions with high posterior probabilities of being cassette exons in both D1 and D2 datasets are actually alternative despite being labelled constitutive at the time the datasets were prepared. This further demonstrates the good performance of the BN.

Predicting exon skipping without using conservation based features

While conservation based features have proved to be the most discriminative, it is desirable to be able to predict alternative splice events using only features that are available to the spliceosome. The performance on this test is also indicative of our understanding of the process of exon skipping. Hence, we should also aim to predict splicing using only information available from a single genome. We predicted exon skipping omitting all conservation based features – the best performance achieved was $TP_{0.5}$ = 29%, and AUC = 0.86 on dataset D1 and $TP_{0.5}$ = 26%, and AUC = 0.88 on dataset D2.

While this performance is noticeably poorer than that achieved using conservation based features, we would like to note that the datasets D1 and D2 consist of exons that are either constitutively spliced in both human and mouse, or cassette exons in both. Thus, we are still distinguishing only between conserved constitutive splicing and conserved exon skipping, leaving out cases of species-specific splicing, as well as of alternative splicing of species-specific exons, which form the majority of alternative exons [51].

Conclusion

Using Bayesian networks (BNs) and several novel features that emerged from recent studies of alternative splicing, we have achieved considerably improved classification of conserved cassette exons. We were able to improve the performance described in [8] due to the incorporation of novel features. To the best of our knowledge, this is the first time that features involving secondary structure and intronic splice regulatory elements have been employed for distinguishing alternative exons from constitutive ones. We also compared our performance on two datasets, and showed that the BN is able to produce accurate classification on both. However, it is worth noting that these datasets differ with respect to discriminative properties.

One direction of future work would be to consolidate various datasets of constitutive and alternative exons, and compile sets of features, which are discriminative over each of them, and the intersection of these sets, which is discriminative over all datasets. Another interesting line to pursue is to predict other kinds of alternative splicing. Here, we focused on exon skipping, which is the most prevalent form of alternative splicing in human and higher vertebrates. However, other major forms of alternative splicing such as alternative donor and acceptor sites [52-54] are also of biological importance, and it would be worthwhile to develop similarly accurate classifiers for these events.

Ideally, we should be able to predict splicing outcomes without conservation based information, as the information required by the spliceosome is present in the given genome. We report our performance at this task, while it is a promising beginning, clearly there is much work to be done. It should be noted that we have ignored two prominent subclasses of alternative exons – namely orthologous exons which are alternatively spliced in a species-specific manner, and species-specific exons which are alternatively spliced. Both these classes are potentially quite important: as up to 50% of all human alternative exons may be human-specific, and up to 60% of all conserved exons which are alternatively spliced may be alternative in a species-specific manner [51]. Classifiers for these tasks are yet to be developed.

Authors' contributions

RS designed new features, performed feature extraction, used BioBayesNet to perform the classification, did the analyses using WEKA, analyzed results, and drafted the manuscript. MH helped with feature design, extraction of secondary structure related features, analysis of the results, and manuscript development. RP implemented the BioBayesNet webserver, which was the framework for using BNs, and helped by providing various options for feature discretization, feature selection, learning BNs and evaluating performance. UG and MP participated in the analysis of the results, and helped draft the manuscript. RB participated in feature design and helped draft the manuscript. RB and MP supervised the study. All authors have read and approved the final manuscript.

Additional material

Additional file 1

The top false positives in the datasets D1 and D2. *Results of investigating the top false positives (P(ACE) = 0.7) in the datasets D1 and D2, using the UCSC genome browser (human genome build hg18, mouse genome build mm9).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-477-S1.xls]

Acknowledgements

We thank Gideon Dror for providing the dataset D1 as well as a few related files, and helpful discussions. We would like to acknowledge Juan Valcarcel for suggesting the use of PTB binding sites as a feature, Klaus Huse for a critical reading of the manuscript, Karol Szafranski and Stefanie Schindler for stimulating discussions, and Swetlana Nikolaweja for help regarding the BioBayesNet webserver. This work was supported by grants from the Deutsche Forschungsgemeinschaft (SFB604) and the German Ministry of Education and Research (0313652D).

References

1. Graveley BR: **Alternative splicing: increasing diversity in the proteomic world.** *Trends in Genetics* 2001, **17(2):**100-107.
2. Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes.** *Nucl Acids Res* 2001, **29(13):**2850-2859.
3. Blencowe BJ: **Alternative Splicing: New Insights from Global Analyses.** *Cell* 2006, **126(1):**37-47.
4. Lee C, Wang Q: **Bioinformatics analysis of alternative splicing.** *Brief Bioinform* 2005, **6(1):**23-33.
5. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucl Acids Res* 2008, **36(suppl_1):**D25-30.
6. Sugnet CW, Kent WJ, Ares M Jr, Haussler D: **Transcriptome and genome conservation of alternative splicing events in humans and mice.** *Pac Symp Biocomput: 2004* 2004:66-77.
7. Sorek R, Shemesh R, Cohen Y, Basechess O, Ast G, Shamir R: **A Non-EST-Based Method for Exon-Skipping Prediction.** *Genome Res* 2004, **14(8):**1617-1623.
8. Dror G, Sorek R, Shamir R: **Accurate identification of alternatively spliced exons using support vector machine.** *Bioinformatics* 2005, **21(7):**897-901.
9. Ratsch G, Sonnenburg S, Scholkopf B: **RASE: recognition of alternatively spliced exons in C. elegans.** *Bioinformatics* 2005, **21(suppl_1):**i369-377.
10. Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB: **Identification and analysis of alternative splicing events conserved in human and mouse.** *PNAS* 2005, **102(8):**2850-2855.
11. Hiller M, Huse K, Platzer M, Backofen R: **Non-EST based prediction of exon skipping and intron retention events using Pfam information.** *Nucl Acids Res* 2005, **33(17):**5611-5621.
12. Leparc GG, Mitra RD: **Non-EST-based prediction of novel alternatively spliced cassette exons with cell signaling function in Caenorhabditis elegans and human.** *Nucl Acids Res* 2007, **35(10):**3192-3202.
13. Ohler U, Shomron N, Burge CB: **Recognition of Unknown Conserved Alternatively Spliced Exons.** *PLoS Computational Biology* 2005, **1(2):**e15.
14. Philipps DL, Park JW, Graveley BR: **A computational and experimental approach toward a priori identification of alternatively spliced exons.** *RNA* 2004, **10(12):**1838-1844.
15. Beaumont MA, Rannala B: **The Bayesian Revolution In Genetics.** *Nature Reviews Genetics* 2004, **5(4):**251-261.
16. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR: **Inference in Bayesian networks.** *Nat Biotech* 2006, **24(1):**51-53.
17. Pudimat R, Schukat-Talamazzini E-G, Backofen R: **A multiple-feature framework for modelling and predicting transcription factor binding sites.** *Bioinformatics* 2005, **21(14):**3082-3088.
18. Barash YKT, Friedman N, Elidan G: **Proceedings of the 7th International Conference on Research in Computational Molecular Biology (RECOMB).** *The 7th International Conference on Research in Computational Molecular Biology (RECOMB): 2003; Berlin* 2003:28-37.
19. Cai D, Delcher A, Kao B, Kasif S: **Modeling splice sites with Bayes networks.** *Bioinformatics* 2000, **16(2):**152-158.
20. Chen T-M, Lu C-C, Li W-H: **Prediction of splice sites with dependency graphs and their expanded bayesian networks.** *Bioinformatics* 2005, **21(4):**471-482.
21. Deforche K, Silander T, Camacho R, Grossman Z, Soares MA, Van Laethem K, Kantor R, Moreau Y, Vandamme AM, on behalf of the non BW: **Analysis of HIV-1 pol sequences using Bayesian Networks: implications for drug resistance.** *Bioinformatics* 2006, **22(24):**2975-2979.
22. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, *et al.*: **The UCSC Genome Browser Database: 2008 update.** *Nucl Acids Res* 2008, **36(suppl_1):**D773-779.
23. Yeo G, Burge CB: **Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals.** *Journal of Computational Biology* 2004, **11:**2-3.
24. Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open Software Suite.** *Trends in Genetics* 2000, **16(6):**276-277.
25. Clark F, Thanaraj TA: **Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human.** *Hum Mol Genet* 2002, **11(4):**451-464.

26. Stadler MB, Shomron N, Yeo GW, Schneider A, Xiao X, Burge CB: **Inference of Splicing Regulatory Activities by Sequence Neighborhood Analysis.** *PLoS Genetics* 2006, **2(11):**e191.
27. Fairbrother WG, Yeh R-F, Sharp PA, Burge CB: **Predictive Identification of Exonic Splicing Enhancers in Human Genes.** *Science* 2002, **297(5583):**1007-1013.
28. Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G: **Comparative Analysis Identifies Exonic Splicing Regulatory Sequences – The Complex Definition of Enhancers and Silencers.** *Molecular Cell* 2006, **22(6):**769-781.
29. Zhang XHF, Chasin LA: **Computational definition of sequence motifs governing constitutive exon splicing.** *Genes Dev* 2004, **18(11):**1241-1250.
30. Yeo GW, Nostrand ELV, Liang TY: **Discovery and Analysis of Evolutionarily Conserved Intronic Splicing Regulatory Elements.** *PLoS Genetics* 2007, **3(5):**e85.
31. Buratti E, Baralle FE: **Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process.** *Mol Cell Biol* 2004, **24(24):**10505-10514.
32. Muckstein U, Tafer H, Hackermuller J, Bernhart SH, Stadler PF, Hofacker IL: **Thermodynamics of RNA-RNA binding.** *Bioinformatics* 2006, **22(10):**1177-1182.
33. Schroeder R, Grossberger R, Pichler A, Waldsich C: **RNA folding in vivo.** *Curr Opin Struct Biol* 2002, **12:**296-300.
34. Hiller M, Zhang Z, Backofen R, Stamm S: **Pre-mRNA Secondary Structures Influence Exon Recognition.** *PLoS Genetics* 2007, **3(11):**e204.
35. Voelker RB, Berglund JA: **A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing.** *Genome Res* 2007, **17(7):**1023-1033.
36. Spellman R, Smith CWJ: **Novel modes of splicing repression by PTB.** *Trends in Biochemical Sciences* 2006, **31(2):**73-76.
37. Sterner DA, Carlo T, Berget SM: **Architectural limits on split genes.** *PNAS* 1996, **93(26):**15081-15085.
38. Witten IH, Frank E: **Data Mining: Practical machine learning tools and techniques.** Second edition. Morgan Kaufmann, San Francisco; 2005.
39. Nikolajewa S, Pudimat R, Hiller M, Platzer M, Backofen R: **BioBayesNet: a web server for feature extraction and Bayesian network modeling of biological sequence data.** *Nucl Acids Res* 2007, **35(suppl_2):**W688-693.
40. Fayyad UM, Irani KB: **Multi-interval discretization of continuous-valued attributes for classification learning.** *IJCAI* 1993, **2:**1022-1027.
41. Pudil P, Novovicova J, Kittler J: **Floating search methods in feature selection.** *Pattern Recognition Letters* 1994, **15(11):**1119-1125.
42. Friedman N, Geiger D, Goldszmidt M: **Bayesian Network Classifiers.** *Machine Learning* 1997, **29(2):**131-163.
43. Pearl J: **Probabilistic Reasoning in Intelligent Systems.** 2nd edition. Morgan Kauffmann; 1988.
44. Chow CK, Liu CN: **Approximating discrete probability distributions with dependence trees.** *IEEE Transaction on Information Theory* 1968:462-467.
45. Jensen FV: **Bayesian Networks and Decision Graphs.** Berlin: Springer; 2001.
46. Ling C, Huang J, Zhang H: **AUC: a better measure than accuracy in comparing learning algorithms.** *Canadian Artificial Intelligence Conference 2003* 2003:329-341.
47. Ladd AN, Charlet-B N, Cooper TA: **The CELF Family of RNA Binding Proteins Is Implicated in Cell-Specific and Developmentally Regulated Alternative Splicing.** *Mol Cell Biol* 2001, **21(4):**1285-1296.
48. Hiller M, Pudimat R, Busch A, Backofen R: **Using RNA secondary structures to guide sequence motif finding towards single-stranded regions.** *Nucl Acids Res* 2006, **34(17):**e117.
49. Garg K, Green P: **Differing patterns of selection in alternative and constitutive splice sites.** *Genome Res* 2007, **17(7):**1015-1022.
50. Carmel I, Tal S, Vig I, Ast G: **Comparative analysis detects dependencies among the 5' splice-site positions.** *RNA* 2004, **10(5):**828-840.
51. Pan Q, Bakowski MA, Morris Q, Zhang W, Frey BJ, Hughes TR, Blencowe BJ: **Alternative splicing of conserved exons is frequently species-specific in human and mouse.** *Trends in Genetics* 2005, **21(2):**73-77.
52. Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ: **Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site.** *Rna* 2006, **12(12):**2047-2056.
53. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M: **Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity.** *Nat Genet* 2004, **36(12):**1255-1257.
54. Koren E, Lev-Maor G, Ast G: **The emergence of alternative 3' and 5' splice site exons from constitutive exons.** *PLoS Comput Biol* 2007, **3(5):**e95.

Rileen Sinha*
Swetlana Nikolajewa*
Karol Szafranksi
Michael Hiller
Niels Jahn
Klaus Huse
Matthias Platzer
Rolf Backofen

Accurate prediction of NAGNAG alternative splicing

Nucleic Acids Research 2009, Vol. 37, No. 11 3569-3579.

Alternative splicing (AS) involving NAGNAG tandem acceptors is an evolutionarily widespread class of AS. We used Bayesian networks (BN) to predict NAGNAG AS, and achieved a balanced sensitivity and specificity of ≥92%. Genome-wide predictions were followed by one of the most extensive experimental validations of predicted AS so far, and 81% (38/47) of the experimentally tested predictions were verified. A BN learned on human data predicted NAGNAG splicing outcomes in four vertebrate genomes with the same performance as achieved on human data, and with a slight drop for Drosophila and worm. Lastly, using the prediction accuracy according to experimental validation, we estimated the number of yet undiscovered alternative NAGNAGs. Our results suggest that we have identified the major features of the 'NAGNAG-splicing code' within the splice site and its immediate neighborhood, and that the mechanism behind NAGNAG AS is simple, stochastic, and conserved among vertebrates and beyond.

* joint first authors

Published online 9 April 2009 *Nucleic Acids Research, 2009, Vol. 37, No. 11* ***3569–3579***
doi:10.1093/nar/gkp220

Accurate prediction of NAGNAG alternative splicing

Rileen Sinha[1,2], Swetlana Nikolajewa[3,4], Karol Szafranski[1], Michael Hiller[2], Niels Jahn[1], Klaus Huse[1], Matthias Platzer[1] and Rolf Backofen[2,*]

[1]Leibniz Institute for Age Research – Fritz Lipmann Institute, Genome Analysis, Beutenbergstrasse 11, 07745 Jena, [2]Albert-Ludwigs-University, Institute of Computer Science, Bioinformatics Group, Georges-Koehler-Allee 106, 79110 Freiburg, [3]Friedrich-Schiller-University, Faculty of Biology and Pharmacy, Department of Bioinformatics, Ernst-Abbe-Platz 2, 07743 Jena and [4]Leibniz Institute for Natural Product Research and Infection Biology, Hans-Knöll-Institute (HKI), Systems Biology/Bioinformatics, Beutenbergstrasse.11a, 07745 Jena, Germany

Received October 6, 2008; Revised March 17, 2009; Accepted March 19, 2009

ABSTRACT

Alternative splicing (AS) involving NAGNAG tandem acceptors is an evolutionarily widespread class of AS. Recent predictions of alternative acceptor usage reported better results for acceptors separated by larger distances, than for NAGNAGs. To improve the latter, we aimed at the use of Bayesian networks (BN), and extensive experimental validation of the predictions. Using carefully constructed training and test datasets, a balanced sensitivity and specificity of $\geq$92% was achieved. A BN trained on the combined dataset was then used to make predictions, and 81% (38/47) of the experimentally tested predictions were verified. Using a BN learned on human data on six other genomes, we show that while the performance for the vertebrate genomes matches that achieved on human data, there is a slight drop for Drosophila and worm. Lastly, using the prediction accuracy according to experimental validation, we estimate the number of yet undiscovered alternative NAGNAGs. State of the art classifiers can produce highly accurate prediction of AS at NAGNAGs, indicating that we have identified the major features of the 'NAGNAG-splicing code' within the splice site and its immediate neighborhood. Our results suggest that the mechanism behind NAGNAG AS is simple, stochastic, and conserved among vertebrates and beyond.

INTRODUCTION

Alternative splicing (AS) is now well established as a widespread phenomenon in higher eukaryotes and a major contributor to proteome diversity. Over half of the multiexonic human genes are believed to have splice variants (1,2). Large-scale detection of AS usually involves expressed sequence tags (ESTs) or microarray analysis (1,3). However, due to various sampling biases, not all AS events can be detected by these methods; furthermore, exon arrays usually do not probe short distance events. Moreover, nowadays genomic sequence data is being churned out at a much faster rate than transcript data, that is, several genomes have low transcript coverage. Thus, there is a need for independent methods of detecting AS.

Alternative acceptors are the second most common kind of AS in human, after exon skipping (4). NAGNAG AS, involving tandem acceptors separated by three nucleotides, is a common type of AS, contributing almost half of all cases of conserved alternative acceptor usage (5,6). NAGNAG splicing results in two possible splice variants—splicing after the first AG results in the E (exonic, also known as proximal) isoform, whereas splicing after the second AG results in the I (intronic, also known as distal) isoform (Figure 1)—accordingly, we refer to constitutively spliced NAGNAG acceptors as the E- or I-class, and to usage of both acceptors, or AS, as the EI-class. According to the data present in the Tandem Splice Site DataBase TASSDB (7), 16% (1815 of 10 740) of human NAGNAG acceptors are alternatively spliced. However, 40% (3562) of the remaining NAGNAG acceptors have less than ten ESTs each, thus implying that a subset of these NAGNAGs may simply lack evidence of AS due to insufficient sampling of the transcriptome. An accurate predictive method would give us a meaningful estimate of the number of yet undiscovered alternative NAGNAG acceptors. Previous work on predicting alternative 3′ splicing, while reporting good results overall, had modest results for NAGNAG AS compared to cases

*To whom correspondence should be addressed. Tel: +49 (0) 761 203 7461; Fax: +49 (0) 761 203 7462; Email: backofen@informatik.uni-freiburg.de
Present address:
Michael Hiller, Department of Developmental Biology, Stanford University, Stanford, CA 94305, USA.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

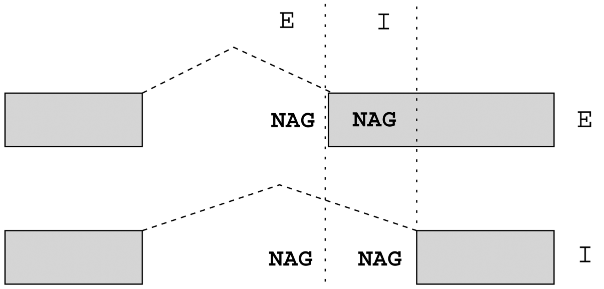


Figure 1. NAGNAG alternative splicing. Nomenclature of NAGNAG AS with E and I sites and isoforms.

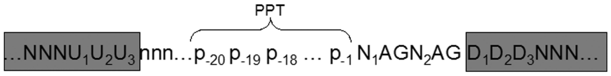

PPT
...NNNU$_1$U$_2$U$_3$nnn...p$_{-20}$ p$_{-19}$ p$_{-18}$ ... p$_{-1}$ N$_1$AGN$_2$AG D$_1$D$_2$D$_3$NNN...

Figure 2. Nomenclature of features used in this study. Nomenclature of sequence features used to analyze NAGNAG splicing. The region used to derive all 42 features is shown, along with the names given to the positional features. Positional features, including the last three nucleotides of the upstream intron, were derived using the database TassDB, which in turn used reference annotations (RefSeq when available, else ENSEMBL).

involving larger distances (8). This seems to contrast with previous work which reported that a simple model based on splice site strength was enough to explain NAGNAG and other short-distance tandem AS (9).

To improve the prediction of NAGNAG AS, we used Bayesian Networks (BN), which are probabilistic graphical models, and TassDB (7) to carefully construct our training and test datasets. BNs are an increasingly popular machine learning approach to data modeling and classification (10,11). We achieved a high balanced sensitivity and specificity and good results in extensive experimental validation of predictions. We show that the performance on a dataset from literature (8) can be improved by a careful consideration of available transcript evidence to include only strongly supported NAGNAGs as constitutive or alternative. Using a BN learned on human data on six genomes from mouse to worm; we show that the performance is comparable or only slightly inferior to that achieved in human. Our results suggest that the mechanism behind NAGNAG splicing is simple, and maintained in evolution.

MATERIALS AND METHODS

Before describing the materials and methods in detail, we note that an overview of the workflow is provided as Supplementary Data (Supplementary Data File 6).

Feature design and extraction

Feature extraction was done using data on NAGNAGs from TassDB (7), using PHP and Perl scripts. The region used for analysis can be seen in Figure 2. Since the composition of the splice site neighborhood influences splicing in general, the base pairs at positions -20 to $+3$ with respect to the NAGNAG were each used as a single feature, as were the two Ns in the NAGNAG motif. The last three positions of the upstream exon were also included, since they can influence both the process of splicing, as well as reflect any influence of codon usage near the exon boundary. Thus, we had a total of 28 features which each represented a nucleotide, and thus had four possible values (A, C, G, T). A weak polypyrimidine tract (PPT) can contribute to AS, and the number of pyrimidines in the 3′ region of the intron is a measure of PPT strength. Therefore, we designed three features related to the pyrimidine content in the 20-bp region upstream of the NAGNAG: ‘Y-content’, which refers to the number of pyrimidines in this region, ‘MaxY-content’, which is the maximal run of consecutive Ys in this region, and their starting position, ‘MaxY-content position’. Additionally, three more PPT-related features were derived from the 50-bp region upstream of the NAGNAG. Following (12), we measured the maximal number of Ys in a 20-nt window, starting from 50-nt upstream of the NAGNAG. Since U and C are not functionally equivalent, PPTs containing 11 continuous Us are the strongest, and the presence of blocks of purines can be detrimental to splicing (13), we also tested two features called ‘T-strength’ and ‘R-strength’, which measured the longest continuous U (Ts in genomic sequences) and R (A or G) strings, starting from 50-nt upstream of the NAGNAG. Since the architecture of the pre-mRNA plays an important role in constitutive and AS (14), the length of the upstream intron (ending in the NAGNAG motif) as well the length of the upstream and downstream exons were taken as features. Splice site strength, being one of the most important determinants of splicing outcome, was also included as a feature—the strength of the two possible splice sites for each NAGNAG exon, as computed using MAXENTSCAN (15), contributed two more features. Lastly, since GC-content can also play a role in splicing, we measured the GC content of the upstream intron as well as the upstream and downstream exons, leading to three more features. In all, 42 features were used (Table 1).

Analyses with dataset D1

The dataset D1 used in (8) was provided by Martin Akerman. To derive the features, we used the genomic coordinates to find the NAGNAGs in TassDB (7), since it contains information about all NAGNAGs in the human genome (as of early 2006). In order to use an SVM for comparison, since that is what was used in (8), we used the WEKA package and the SMO implementation of SVMs therein, using a polynomial kernel. To begin with, we used the labels as provided in D1, and then we replaced the labels according to TassDB, and finally we replaced the samples labeled constitutive by samples with ≥10 ESTs (for one variant only) from TassDB. Leave-one-out cross-validation was used, as in (8). For feature selection within WEKA, we used the method ‘CfsSubsetEval’, as well as manual inclusion and exclusion of features. We also repeated the analysis with a Bayesian network to ensure that BNs are a good choice for this task, and found that the BNs did match the performance of the SVM.

Datasets derived from TassDB

The dataset D2 of human NAGNAG acceptors was extracted from TassDB (7) using the criteria: (i) constitutive: ≥10 ESTs supporting either E or I variant, 0 for the other; (ii) alternative: ≥2 ESTs supporting each variant, ≥10% of ESTs supporting minor variant (Supplementary Data File 2). The remaining human NAGNAGs were used for prediction only (Supplementary Data File 3). NAGNAG acceptors from the mouse, rat, chicken, zebrafish, fly and worm genomes were extracted in the same manner. Only NAGNAG acceptors from transcripts with a correct exon–intron structure as well as a correct open reading frame were used.

Validation of splicing class assignments using next-generation sequence data

Published next-generation transcriptome sequence data (Illumina GA II) was retrieved from the Short Read Archive section of the GEO database at NCBI (accession number GSE12946) (16).The dataset comprised 313 million 32-mer readings, obtained from cDNA from nine different human tissues and five breast cell lines. For the analysis we required exact matches of the readings to one of the isoforms. Matches had to overlap at least 6 nt on both sides of the exon–exon junction and were discarded if the same sequence occurred somewhere else in the human transcriptome (RefSeq transcripts or NAGNAG isoform sequences).

Bayesian networks

We used the algorithms for feature selection, model learning and classification as described in (17), and made available *via* the public webserver BioBayesNet (18). BioBayesNet restricts the structure of the BNs by using the so-called *tree-augmented naïve Bayes* (*TAN*) structure (19). In contrast to a naïve Bayes classifier/network, where the attributes are assumed to be independent, a TAN classifier augments the underlying naïve Bayes classifier by allowing at most one additional parent per node. Feature selection was carried out in three stages. First, a 'discretizer' applying the algorithm of Fayyad and Irani (20) discards features for which no suitable discriminative intervals are found. Secondly, the sequential feature subset selection (SFFS) algorithm (21) was applied. Thirdly, we enforced inclusion or exclusion of features manually.

Experimental validation and quantification of splice variants

For validation and quantification of splice variants, PCRs were performed using 200 pg cDNA templates from the Human Multiple Tissue cDNA Panels I and II (Clontech, Heidelberg, Germany). For each given gene, a suitable tissue was determined from expression data obtained from the Stanford SOURCE database (21). PCR primers were obtained from Metabion (Supplementary Data File 5), each sense primer labeled with 6-carboxyfluorescein (FAM). Reactions were set up with BioMix Red (Bioline, Luckenwalde, Germany) and 10 pmol primer in 25 µl total volume, according to the manufacturer's instructions. The thermocycle protocol was 2 min initial denaturation at 94°C, followed by 42 cycles of 45 s denaturation at 94°C, 50 s annealing at 56°C, 1 min extension at 72°C, and a final 30 min extension step at 72°C. Each product was diluted 1/40, and 1 µl of the dilution mixed with 10 µl formamide (Roth, Karlsruhe, Germany) and 0.5 µl of GeneScan GS500LIZ (Applied Biosystems, Darmstadt, Germany) were heated to 94°C for 3 min. The mixture was than separated on an ABI 3730 capillary sequencer and analyzed with the GeneMapper 4.0 software (Applied Biosystems). If two peaks with about the expected fragment sizes (with a tolerance of ±3 nt) and distance (3 nt) were visible, the isoform ratios were calculated based on the peak areas.

Information gain

Information gain is defined as the reduction in the entropy of the class variable, given the feature. The formula for information gain is:

$$\mathrm{IG(Class \mid Feature) = H(Class) - H(Class \mid Feature)}$$

where H(Class) is the entropy of the class variable, and H(Class|Feature) is the conditional entropy of the class variable, given the feature. Information gain is a well established measure for feature selection in Machine Learning (22). We used the WEKA package (22) for computing information gain, in order to rank the features according to how informative they were. We also used it for prediction based on SVMs, as implemented in the SMO option, and for prediction using Naïve Bayes classifiers.

The BayNAGNAG webserver

We used WEKA to implement the BNs, and C + + code was written to enable the web browser to interact with WEKA, using the features derived from the user's input along with saved BN models to produce the predicted splicing outcome.

Estimating the number of undiscovered alternative NAGNAGs

To estimate the number of alternative NAGNAGs which lack transcript evidence as of now, we used the accuracy of predictions according to the experimental validation, as follows: We computed the average accuracy of prediction in the three probability intervals $f_1 = 0.5–0.69$, $f_2 = 0.7–0.89$ and $f_3 = 0.9–1.0$, according to the experimental results. If f_i is the fraction of experimentally validated predictions in the interval i, and n_i is the number of samples in the test dataset which are currently labeled as constitutive, but predicted to be alternative, then the estimated number of yet undiscovered alternative NAGNAGs is

$$N = n_1 * f_1 + n_2 * f_2 + n_3 * f_3.$$

We used the validation accuracies for two different thresholds (≥1%, and ≥10%) of abundance of the minor variant, leading to two estimates of the number of yet undiscovered alternative NAGNAGs.

RESULTS

Performance on a dataset from the literature

While an SVM reported in (8) succeeded in predicting AS for alternative acceptors separated by up to a distance of 100 nt, NAGNAG acceptors were shown to be the least predictable (8). To understand the reasons behind that, we obtained the underlying dataset from the authors, called D1 in the following. However, in the following we did not use conservation based features, because we aim at predicting AS using information only from a single genome. Using our own set of 42 features (Table 1), we verified that the reported performance is matched by the BN, as well as by an SVM implementation provided in the WEKA package (22). The predicted NAGNAG class is the one which receives the maximum score or posterior probability from the classifier. We computed the receiver operating curve (ROC), which is a plot of the true positive rate versus the false positive rate, and measured the area under the ROC curve (AUC), which is a standard measure of the quality of a classifier (23). An ideal classifier, which makes no errors, would achieve an AUC of 1. By means of the SMO (Sequential Minimal Optimization) implementation of a support vector machine in WEKA and all our features, the AUC obtained for distinguishing EI and E cases is 0.79, the same as reported (8). Using a subset of features (Table 2) yielded by feature selection improves this to 0.82. Similarly, using all 42 features, the AUC obtained for distinguishing EI and I cases is 0.7, the same as reported (8), and this improves to 0.77 using feature selection.

To check whether this relatively modest performance was due to the set of constitutive NAGNAGs in D1 being in fact contaminated by alternative NAGNAGs, we searched the Tandem Splice Site DataBase (TassDB) (7) for the NAGNAGs in the D1 dataset, and replaced the labels ‘alternative’ and ‘constitutive’ according to TassDB. Indeed, this revealed that many NAGNAGs in D1 labeled constitutive were in fact alternative according to the transcript evidence in TassDB—119 of 397 (30%) cases assigned to the E-class, and 104 of 177 (58.8%) cases assigned to the I-class, are in fact alternative (EI-class) according to TassDB. Incorporating this information resulted in improved performance—the AUCs achieved were 0.89 for distinguishing EI cases from E cases, and 0.85 for distinguishing EI cases from I cases (Table 2).

However, such relabeling still allows samples which have very low transcript coverage and are thus potentially mislabeled also in TassDB, and it also changes the ratios of the sizes of the various classes, especially for the EI versus I problem. Therefore, we replaced all samples labeled constitutive in D1 by samples from TassDB which had ≥10 ESTs supporting one splice site, and none for the other. Since there are only 331 such samples in the I-class, we randomly chose 331 (of 5032) samples from the E-class. This new mixed dataset yielded significantly improved performance, with AUC values of 0.97 and 0.94 for EI versus E and EI versus I, respectively.

Table 1. Features for machine learning used in this study

Feature subset	Number of features	Motivation
N_1, N_2, D_1, D_2, D_3 and positions in the PPT	25	NAGNAG splicing is influenced by the NAGNAG motif and its sequence context
U_1, U_2, U_3	3	Potential influence on protein context
Length of neighboring exons and upstream intron	3	The architecture of the pre-mRNA influences AS
GC content of neighboring exons and upstream intron	3	GC content can influence AS
Features related to the pyrimidine content of the PPT	6	Composition of the PPT influences splicing
Splice site strength of E and I splice sites	2	Alternative NAGNAGs tend to have comparable splice site strengths

Table 2. Performance on the dataset D1, using SVMs

Classification problem	Original sample labels		Sample labels according to TassDB	
	AUC	Features[a] used	AUC	Features used
E versus EI	0.82	N_1, N_2, MAXENT-E, MAXENT-I, D_1, p_{-1}, Y-content,	0.89	N_1, N_2, D_1, D_3, U_1, U_2, p_{-8}, p_{-5}, p_{-2}, p_{-1}
I versus EI	0.77	N_1, N_2, MAXENT_E, MAXENT_I, D_1, p_{-2}, p_{-1}, GC-intron,	0.85	N_1, N_2, D_1, D_2, D_3, U_1, U_2, U_3, p_{-19}, p_{-18}, p_{-16}, p_{-13}, p_{-12}, p_{-11}, p_{-10}, p_{-9}, p_{-8}, p_{-6}, p_{-5}, p_{-2}, p_{-4}, p_{-3}, p_{-2}, p_{-1}

[a] For nucleotide nomenclature see Figure 2. *Y-content*: fraction of the 20-bp upstream of the NAGNAG motif that are pyrimidines, *GC_intron*: G + C content of the intron ending with the NAGNAG, *MAXENT_E*, *MAXENT_I*: MAXENT scores for the É and I splice sites.

Removing all NAGNAGs containing a GAG, as done in (8), did not affect the performance drastically, as we obtained AUC values of 0.96 and 0.92 for EI versus E and EI versus I, respectively. Thus, the use of strict thresholds on EST evidence of constitutive splicing greatly reduces the noise in the dataset, and improves the prediction performance. It must be pointed out that we only used transcript evidence for the human genome, that is, some of the alternative cases might be human-specific.

To further validate the relabeling of samples in D1, we analyzed next-generation transcriptome data (Illumina/Solexa GA II), 313 million sequences, obtained from nine different human tissues and five breast cell lines (16) as an additional source of experimental evidence for NAGNAG isoforms. A total of 7509 NAGNAG cases had sequences specifically matching at least one of the isoforms (total of 363009 sequences). We note that the coverage of the transcriptome by these Solexa data is not exhaustive, so there are likely more examples of AS NAGNAGs than thereby supported. We applied stringent filters on the number of sequences supporting an event—these filters had been previously shown to help in the detection of experimentally reproducible AS (24). To consider a NAGNAG to be alternatively spliced, we required at least two supporting sequences for each isoform, and at least 10% of the total sequences to support the minor isoform. A constitutive NAGNAG had to be supported by at least 10 sequences for one isoform, and 0 for the other. We then computed the intersection of this dataset with D1 (Supplementary Data File 8), and compared the labels of the samples. 203 cases of D1 were found in the filtered Solexa dataset—of 142 cases labeled constitutive in D1, 66 (46%) had evidence for being alternatively spliced. When we repeated the comparison after replacing the labels according to TassDB, there were 74 cases labeled constitutive, of which only 12 (16%) were alternative according to the Solexa data. This underscores the need to use thresholds of transcript support for both constitutive and AS as well as confirms our relabeling.

In-silico performance on a TassDB derived dataset

Having seen that sets of constitutive splice events might in fact be significantly corrupted by (not yet detected) alternative acceptors, we decided to take extra care in our selection of human alternative and constitutive NAGNAGs for training data by considering only NAGNAGs which are strongly supported in TassDB. Thus, a NAGNAG was considered constitutive if it had $\geq$10 ESTs supporting one splice site, and none for the other. To be considered alternative, there had to be $\geq$2 ESTs for each splice variant, and $\geq$10% of the ESTs must support the minor variant. Such filtering of alternative events was not required in D1 as another stringent filter—of conserved AS—had already been applied.

This TassDB dataset (called D2 in the following) consists of 5363 constitutive (5032 E, 331 I) and 902 alternative NAGNAGs. We also repeated the comparison with the filtered Solexa dataset (Supplementary Data File 9) as in the previous section—2890 cases of D2 were found in the filtered Solexa data, and of the 2466 cases labeled constitutive, only 37 (1.5%) had evidence of AS in the Solexa data. The much lower number of mislabeled constitutive samples in D2 when compared to the original D1, further justified the choice of stringent filters.

D2 was partitioned into two equal parts, and then, in turn, we used half of the data to train the BNs, and the remaining half was used for testing. The test set remained untouched while the training set was used for discretization, feature selection and learning the BN. Finally, the BN which had been learned on the training set was used to classify the samples in the test set. This procedure was carried out twice, using each half for training and testing in turn, and the average of the two runs was taken as the final performance.

We classified each candidate as belonging to one of the three classes (EI/E/I). The BN achieved AUC of 0.96, 0.97 and 0.98 respectively for identifying EI, E and I variants, as seen in the ROC plot (Figure 3). The balanced sensitivity and specificity obtained was 92%, 95% and 93% (EI/E/I). We would like to note that in contrast to (8), which divided this classification into two sub-tasks, namely predicting EI versus E, and EI versus I, we treat it as a 3-class problem, thus covering all three possible splicing outcomes at the same time.

Another noteworthy difference is that while (8) reported worse performance for distinguishing between EI and I cases, compared to distinguishing between EI and E cases, in the 3-class problem, the highest performance is achieved in predicting the I-class, that is, constitutive usage of the downstream acceptor. This is intuitively easy to grasp, since the scanning mechanism (24) implies that the upstream acceptor is preferentially used, so that constitutive usage of the downstream acceptor is only likely when the upstream splice site is quite weak, for example, when we have a GAGHAG pattern (H = A, C or T). Previous experimental work on 3′ splicing (25), as well *in-silico* analyses of NAGNAG splicing (26,27) have shown that the nucleotide preceding the AG can

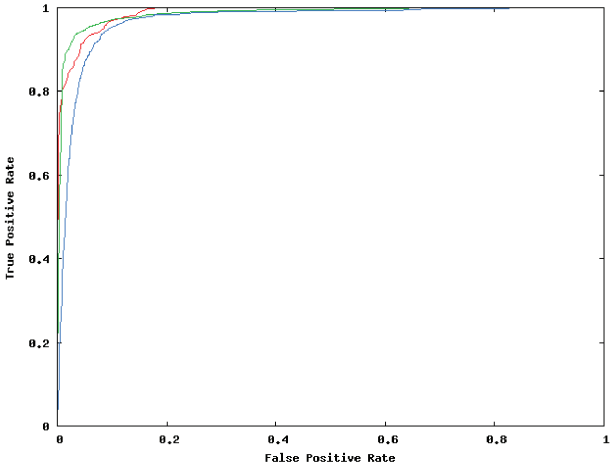


Figure 3. *In-silico* performance of the Bayesian network. ROC plot showing the performance achieved on the 3-class [I-class (red), E-class (green), and EI-class (blue)] classification problem. The I-class is relatively the easiest to predict, whereas the EI-class, or AS, is the hardest.

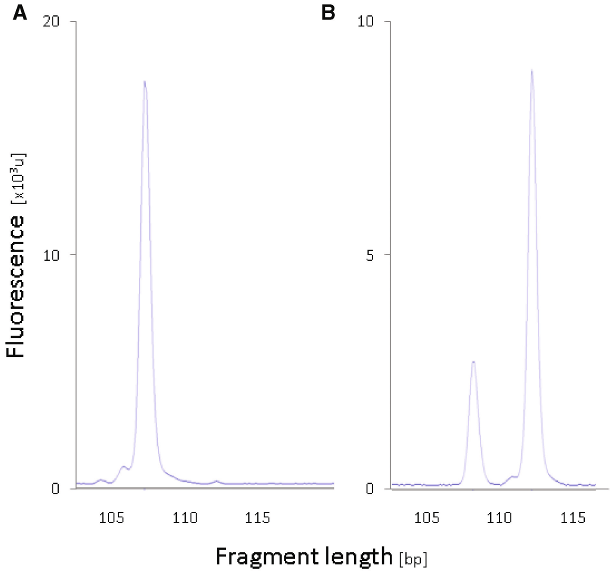


Figure 4. Experimental validation of predictions using RT–PCR and quantification by capillary electrophoresis. Experimental results indicating (**A**) constitutive NAGNAG splicing of VPS13D exon 27 and (**B**) alternative NAGNAG splicing of INPP5E exon 6, minor isoform abundance 24%.

influence the choice of 3′ splice site, with the following order of preference: CAG > TAG > AAG > GAG. Consistent with this, 227 of 331 I cases (68.6%) in the D2 dataset have a GAGHAG pattern. The order of preference is also reflected in the sequence logos (28,29) constructed using the D2 data (Supplementary Data File 1).

Removing all NAGNAGs containing a GAG from the training and test sets results in AUC of 0.90, 0.94 and 0.90 respectively for identifying EI, E and I variants. Removing such NAGNAGs can be considered, as GAGs are believed to rarely serve as functional splice sites (8,30), and therefore such NAGNAGs are considered 'implausible' for the purposes of AS (31). However, since TassDB contains 182 alternative NAGNAGs of this kind (of which 59 have ≥2 ESTs supporting each variant), we decided to include them. The BN achieves AUC of 0.83, 0.98 and 0.99 respectively for identifying EI, E and I variants on the subset of GAG-containing NAGNAG motifs and predicts 6% of the EST-supported ones to be alternatively spliced. On the other hand, among the currently known constitutively spliced GAG-containing NAGNAG acceptors eight (1.2%) are being predicted to be alternative.

Experimental validation

Having established that highly accurate predictions of NAGNAG splicing are possible in-silico, we decided to perform extensive experimental validation of predictions. Experimental validation was performed using RT–PCR followed by capillary electrophoresis with laser-induced fluorescence detection. NAGNAG AS appears in our experimental readout as two fluorescence peaks separated by three nucleotides (Figure 4). To avoid false positive results due to noise, a threshold has to be defined above which the intensity of the minor peak is considered as a robust signal of AS. Accordingly, we measured the accuracy of predictions against the threshold of the isoform ratio, that is, the abundance of the minor transcript (lower peak).

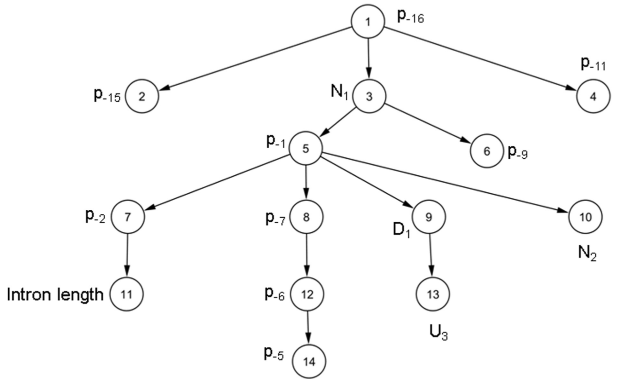


Figure 5. Bayesian network to predict NAGNAG alternative splicing. The 14-feature Bayesian network learned on the D2 dataset. Note that the class node, which has an edge to all other nodes, is omitted for ease of visualization. Thus, this is just the augmenting tree in the TAN classifier.

Candidates for experimental work were chosen from both, the entire D2 dataset described in the previous subsection (termed 'training data'), and the remaining 4475 (913 EI, 3206 E and 356 I) human NAGNAGs in TassDB (termed 'test data'). The BN learned on D2 uses 14 features (Figure 5) and was applied to classify both training as well as test data (Supplementary Data Files 2 and 3, respectively), and candidates were chosen based on the classification results. Besides the prediction of AS for exons with low EST coverage, we decided to select candidates of several different types, as explained in the following, where P(EI) refers to the probability of being alternative.

Class 1: NAGNAGs from the training data, labeled constitutive, but given a P(EI) ≥0.9 by the BN. As control, constitutive training NAGNAGs with a P(EI) ≤0.1 were chosen. With these candidates, we wanted to test whether the BN can find alternative acceptors even within the ones which had strong transcript support in favor of being constitutive. At a minor variant abundance threshold of 4%, the validation rate is 100% for both cases (6/6; Table 3), and controls (2/2). These results indicate that even strong transcript support (EST coverage ≥10) can miss alternative splice events and cannot 'prove' that an exon is indeed constitutive. The highest number of ESTs among these six was 36, for *SNTA1* (NM_003098, exon 8) for which we detected 4% usage of the acceptor that was unsupported by ESTs. The highest observed splice ratio of 43% was obtained for *C3ORF34* (NM_032898 exon 2) which was originally covered by 21 ESTs confirming the E acceptor exclusively.

Class 2: NAGNAGs from the training data, labeled alternative, but given a P(EI) ≤0.1 by the BN. As control, alternative training NAGNAGs with a high P(EI)

Table 3. Accuracy of prediction against threshold of the minor splice variant

Threshold of the minor splice variant (%)	Experimentally confirmed predictions of AS	
	Class 1[a] (%)	Class 3[b] (%)
10	50	60
8	50	90
6	67	90
4	100	100

[a]Six predictions with $P(EI) \geq 0.9$.
[b]Ten predictions with $P(EI) \geq 0.9$.

Table 4. Accuracy of predictions against posterior probability

P(EI)[a]	Accuracy of predictions
0.9–1	100% (10/10)
0.7–0.89	80% (8/10)
0.5–0.69	50% (5/10)

[a]Abundance of the minor splice variant ≥ 4%.

were also chosen. With this class, we wished to identify constitutive NAGNAGs which had erroneous transcript evidence of being alternative. At a minor variant abundance threshold of 4%, the validation rate is 80% (4/5) for cases being constitutive and 100% (6/6) for controls being alternative.

Class 3: NAGNAGs from the test data, labeled constitutive, but given a $P(EI) \geq 0.5$ by the BN. Candidates were chosen from each interval of 0.1 between 0.5 and 1. As controls, two test acceptors, labeled constitutive and with $P(EI) \leq 0.1$ where chosen. The underlying consideration was to test not only the ability of the BN to identify alternative NAGNAGs among the acceptors with low transcript coverage, but also to test whether a higher P(EI) corresponded to a higher accuracy of prediction. The results of the experiments on candidates from this class demonstrates that for a given threshold of isoform ratios, higher posterior probabilities result in more reliable predictions (Table 4). The validation rate for the controls at a minor abundance of 4% is 50% (1/2).

Class 4: NAGNAGs from the test data, labeled alternative, but given a $P(EI) \leq 0.1$ by the BN. As control, alternative NAGNAGs with a $P(EI) \geq 0.9$ were also chosen. Note that the difference when compared to Class 2 is that these are alternative NAGNAGs with relatively weaker transcript support. For Class 4, the validation rates at a minor abundance of 4% are 83% (5/6) for cases being constitutive, and 50% (3/6) for controls being alternative.

In all, 63 NAGNAGs were investigated (Supplementary Data File 4), and the experiments confirmed that the BN can accurately predict NAGNAG-splicing outcome in 81% (38/47) of candidates. Surprisingly, the validation rate for controls was lower with 75% (12/16).

Table 5. Top 10 features according to the information gain

Feature[a]	Information gain
N_2	0.492
MaxEntScan I	0.448
MaxEntScan E	0.252
N_1	0.199
p_{-1}	0.040
D_1	0.020
p_{-2}	0.014
p_{-3}	0.005
G/C content of 3′ exon	0.005
D_3	0.004

[a]For nomenclature see Figure 2.

Most informative features

Next we asked which the most informative features for our classification problem are. By measuring the information gain, we identified the two Ns in the NAGNAG motif, and the splice site scores, to be by far the most informative features (Table 5), which is also in agreement with the literature (8,9,25–27). The downstream N (N_2, Figure 2) is the most informative feature, followed by the splice site score of the I acceptor. The next two most informative features are the upstream N (N_1, Figure 2), and the splice site score of the E acceptor. The nucleotides immediately upstream and downstream of the NAGNAG acceptor (positions −1, +1, −2 and −3) are the next four informative features, and the nucleotide at position +3 is ranked 10, reflecting the highly localized nature of NAGNAG splicing. The feature ranked 9 is the GC-content of the downstream exon—NAGNAGs whose downstream exon has a higher GC-content are enriched in usage of the I acceptor and correspondingly in alternative NAGNAG splicing.

We note that while the splice site scores are very informative, they are not present in the 14 feature BN learned on D2 (Figure 5)—this is because the relevant information is already captured by N_1, N_2 and the immediate neighborhood. The splice site scores are based on information that also uses positions which are relatively distant from the NAGNAG, and likely not strongly influential on the splicing outcome. Moreover, using the splice site scores introduces a systematic bias against the downstream acceptor, since the 'PPT' (polypyrimidine tract) now contains an AG dinucleotide.

Prediction on the mouse, rat, chicken, zebrafish and fly genomes

To test how the BN trained on human performs on data from other species, we first extracted mouse NAGNAG data from TassDB (7), using the same EST-based filtering criteria as for the D2 data above. The performance on the mouse NAGNAG data was nearly identical to that on human (Table 6, Supplementary Data File 10). Encouraged by this, we used the same EST-based filtering in rat, chicken, zebrafish and fly, and predicted NAGNAG splicing using the BN. The performance achieved for the three vertebrates was very similar to

Table 6. Area under the ROC curve for the three classes and six organisms

Organism	AUC		
	EI	E	I
Human	0.967	0.985	0.989
Mouse	0.966	0.982	0.989
Rat	0.967	0.985	0.991
Chicken	0.972	0.983	0.986
Zebrafish	0.967	0.983	0.992
Fruitfly	0.924	0.971	0.952

that on human and mouse, whereas the performance on fly data, while not as high as that on the others, was still quite good (Table 6). Investigating the cause behind the reduction in performance on the Drosophila genome, we found that excluding positions not in the immediate neighborhood of the NAGNAG—in particular, excluding all features except the two Ns in the NAGNAG, and the two nucleotides immediately upstream, lead to a slight improvement on Drosophila data. This simplified BN trained on human D2 data with just four features, also almost matched the performance of the previous BN with 14 features (Figure 5) on the other five genomes, as well as when evaluated by the above outlined experimental results (data not shown).

Prediction on the worm genome

TassDB also contains data from the worm genome, however, there are no examples of constitutive I variants with 10 or more ESTs. We used the 3-class BN with four features (the two Ns in the NAGNAG, and the two nucleotides immediately upstream) trained on human D2 data to predict NAGNAG-splicing outcomes for *Caenarhabditis elegans*, and obtained AUC values of 0.93 for predicting the EI and E-classes. Only one sample was predicted to belong to the I-class. A 2-class BN trained on human D2 data from only the E and EI-classes produced the same AUC values. A closer look at the data revealed that none of the 391 NAGNAGs (369 E, 22 EI) had G as the upstream N (N_1, Figure 2; which is most often the case for constitutive I variants) in the NAGNAG. Thus, it appears that the splice site sequence context is different in NAGNAG splicing in *C. elegans*, compared to vertebrates. This is in agreement with previous studies that identified an extended 3′ splice site consensus in *C. elegans* (28).

Performance using a minimal set of features

Since reducing the number of features lead to an improvement in prediction of NAGNAG AS in Drosophila and worm, we asked how many features we could omit without a significant drop in performance on the human D2 dataset. We found that using only the two Ns in the NAGNAG motif, or only the splice site scores (computed by MAXENTSCAN) led to only slightly worse performance. We also found that using a naïve Bayes classifier instead of a BN (with the same features), led to only a minor drop in performance.In order to compare the impact of leaving out features, we compared the error rates of classification using different feature subsets under a 10-fold cross-validation setting with D2. The results show that the error rate is lowest (5.9%) when using only N_1, N_2, p_1, p_2 and D_1, that is the two Ns in the NAGNAG, and the immediate two upstream and one downstream positions. The error rate using only the MAXENTSCAN scores (7.4%), is higher than that obtained using all features (7.1%), only N_1 and N_2 (6.7%), or the 14-feature BN we used for the experimental validation (6.3%). We would also like to point out that there is practically no difference in the computational cost of using the various models—the cost of extra features in training the models is not much, and more importantly, once trained, the various models take near-identical time to classify new data.

Table 7. Predictions of the 14-feature BN on experimentally studied cases from the literature (30)

Gene	Isoform ratios (E:I) in different tissues (30)	P(EI)	P(E)	P(I)
DRPLA	8:2–9:1	0.76	0.22	0.02
GHRHR	2:8	0.92	0.04	0.05
BAIAP2	1:9–0:10	0.88	0.04	0.07
PTMA	0:10–1:9	0.14	0.33	0.53
IGF1R	7:3–8:2	0.56	0.43	0
PAX3	0:10-10:0	0.72	0.03	0.25
PAX7	0:10–9:1	0.69	0.13	0.18
LEP	1:9–10:0	0.61	0.38	0.02
DNMT1 (Mouse)	4:6–6:4	0.58	0.07	0.35
CAST	9:1–10:0	0.90	0.08	0.03
MAN2B1	0:10–3:7	0.23	0.67	0.10
PSEN2	7:3	0.45	0.55	0
LAP1B	0:10–10:0	0.84	0.15	0.01
NOXO1	0:10–9:1	0.08	0.91	0.01
CCL20	4:6–9:1	0.80	0.18	0.02
SGNE1	4:6–8:2	0.48	0.41	0.11
TGFA	5:5–9:1	0.93	0.04	0.03

Webserver and performance on examples from the literature

To further validate our classifier, we tested it on examples of experimentally studied NAGNAGs from the literature (30), which includes interesting examples of tissue-specific variations of the isoform ratio. As shown in Table 7, the results were promising—13/17 (76%) of the cases were predicted to be alternative. An additional 5/7 cases from (29) were also correctly predicted (data not shown). Thus, the performance on these cases from the literature further underscores the usefulness of our classifier. To enable others to do similar experiments as well as reproduce our results and/or predict NAGNAG AS in candidate acceptors of their interest, we developed a webserver—BayNAGNAG, available at: http://www.tassdb.info/baynagnag/

A user can provide a NAGNAG motif along with the upstream and downstream sequence context, the intron length and the last base of the upstream exon. These are then used to predict the class, and the posterior probabilities of all three classes are provided as output. Predictions using two different BNs are provided—one which uses 14 features (Figure 5) and was used in the experimental validation, and the other trained on MAXENTSCAN (15) scores (of the E and I) splice sites only. Furthermore, we also provide an additional file (Supplementary Data File 7) with the required information for all 10 740 human NAGNAGs used in our study.

Estimating the number of undiscovered alternative NAGNAGs

Using the accuracy of predictions according to the experimental validation, we estimate the number of yet undiscovered alternative NAGNAGs in the human genome (10 740 NAGNAGs, 8925 constitutive, 1815 alternative) to be 258–515. The corresponding estimates for mouse (8735, 7386, 1349), fly (1589, 1411, 178) and worm (4697, 4661, 34) genomes are 214–417, 106–214 and 101–185, respectively.

DISCUSSION

We have demonstrated that BNs can produce highly reliable predictions of NAGNAG-splicing outcomes. Once transcript evidence had been carefully considered to create a training dataset, the BN achieved high performance, not only in-silico with a balanced sensitivity and specificity of $\geq$92%., but also according to extensive experimental validation. Altogether, we investigated the AS of 63 NAGNAGs in one to two tissues and confirmed our predictions in 81% of cases and 75% of controls (4% threshold for the minor isoform). The surprisingly low confirmation rate of controls is primarily due to the 50% (3/6) success rate for low expressed genes (Class 4). Likely, some of these failures are false negatives as AS may take place in other cell types than those tested. In turn, this implies that also some non-confirmed case predictions of AS are false negatives within our experimental setup. Summing up cases and controls with P(EI) $\geq$0.9, the confirmation rate is 89% (25/28) despite that the just discussed problematic Class 4 controls are included. It is natural to ask why ESTs failed to detect the predicted AS in Class 1 candidates, which was successfully validated by our experiments. In our opinion, some of these cases are easily explained by the low minor abundance, which implies that it is not surprising if a relatively low number of ESTs fails to detect AS. For instance, the NAGNAG belonging to the gene *NF1* in Class I has a minor abundance of 0.05, so one would expect to see, on average, 1 EST out of 20 supporting the minor variant. However, since this NAGNAG is only covered by 10 ESTs, it is not surprising that AS is not detected.

To the best of our knowledge, this is the first instance of such extensive validation of in-silico predictions of NAGNAG splicing, and is also among the most extensive experimental validations of non-EST based methods of predicting AS published so far.

The single biggest factor contributing to the performance of the BN was the preparation of the training dataset. As we showed by prediction on the dataset D1 from literature (8), judicious use of transcript evidence, especially a threshold on the number of transcripts required to label an exon as constitutive, makes a big difference. A strict threshold on the EST evidence required to label a splice site as constitutive or alternative is required to minimize the noise inherent in EST databases, and the performance of a classifier can only be as good as the quality of the data that it is trained with.

The most informative features (Table 5) are the two Ns in the NAGNAG motif, and the splice site scores. To some extent, the scores for the upstream and downstream splice sites, and the upstream and downstream Ns can be substituted by each other. The nucleotides immediately neighboring the NAGNAG are the next most important, while other features make only small contributions to the prediction performance. Thus it is evident that most of the information required for prediction is encoded in the immediate splice site neighborhood.

A BN trained on human data achieved near-identical performance on the mouse, rat, chicken and zebrafish genomes, indicating that the determinants of NAGNAG splicing outcome are conserved among vertebrates. Furthermore, the fact that the most informative features were the two Ns in the NAGNAG motif, and its immediate neighboring nucleotides, suggests that the mechanism is simple in nature and maintained in evolution. Given the relatively low transcriptome coverage in rat, chicken and zebrafish, one might ask whether the subset of NAGNAG acceptors we studied for these genomes represent the highly expressed subset of genes and thus likely enriched in conserved alternative events. However, this would not appear to be the case, as we obtain nearly identical results for mouse, which has much higher transcriptome coverage. Thus, our BN should be useful to annotate NAGNAG splicing in animal genomes that currently lack extensive transcript data.

The BN trained on human data was also able to predict NAGNAG AS in the Drosophila genome, though with a drop in performance. However, training using data from Drosophila itself did not improve the performance, indicating that the mechanism may well be conserved between vertebrates and Drosophila. Furthermore, using only four features (the two Ns in the NAGNAG, and the two nucleotides immediately upstream), a BN trained on human data achieved good performance on the worm genome, which contains no instances of the I-class with strong EST support.

This suggests that perhaps what is different in NAGNAG splicing in *C. elegans*, compared to vertebrates is not the mechanism but rather the evolutionary constraints on the splice site sequence context.

Simpler approaches like using only the two Ns in the NAGNAG motif, or only the splice site scores (computed by MAXENTSCAN), or using a naïve Bayes classifier, led to only slightly worse performance, indicating that the other features and the corresponding dependencies learned by the BN are weak in their discriminative power, and in generalization to other datasets. All this points to a simple and stochastic mechanism, at least in as much as predicting the class (EI/E/I) of NAGNAG splicing is concerned. This is in agreement with (9), who proposed a model based on the sequence context from −6 to +6 at the intron-exon boundary, that is, from −3 to +6 with respect to the NAGNAG, or 15 positions in all. We have shown that the class (EI/E/I) of NAGNAG splicing can be predicted in the vast majority of cases with even fewer positions, that is, −2 to +1 with respect to the NAGNAG, or 9 positions in total. However, the prediction of splice ratios and their tissue and/or developmental stage dependent changes has to involve additional cis and/or trans features and can not be based on a simple stochastic mechanistic assumption. We note that the possibility of such a mechanism does not preclude regulation or a biological function (5,32). Stochastic splice site selection might in fact help production of constant splicing ratios, which have been observed in some NAGNAG sites with clear functional implications (5). At a qualitative level, the stronger splice site seems to correspond to the more abundant variant in most cases, thus supporting a model in which the two splice sites compete for binding to the spliceosome. However, quantitative prediction of the precise abundance is much more challenging. Since NAGNAG AS is frame-preserving (and thus not subject to NMD), save for the ~2% of the cases which introduce an in-frame stop codon (25), the vast majority of cases should lead to different proteins Studies so far have found evidence of both cases where such proteins have variations in function, as well as those in which there is no noticeable difference, and thus the AS is apparently just ‘tolerated’ by the cell [(5) and the references therein].

We also estimated that there are up to several hundred undiscovered alternative NAGNAGs in the human, mouse, fruitfly and worm genomes. We note that these numbers could be an underestimate, since we only consider predictions with $P(EI) \geq 0.5$. Given the current level of annotations of the rat, chicken and zebrafish genomes, genomic information about a substantial fraction of NAGNAG acceptors is likely lacking, therefore such estimation would not be meaningful.

Despite the experimentally validated accuracy achieved in predicting the outcome of NAGNAG splicing at the ‘ternary level’ (EI, E or I), the ‘NAGNAG-splicing code’ is not completely solved. Open questions are the isoform ratios and their tissue specificity observed for several NAGNAGs (25,30,33). Here, sequence features may contribute to the isoform ratio although we consider them uninformative for discrimination at the class level, constitutive versus alternative. Prediction of isoform ratios should also address the influence of the sequence context in the intron and in particular of the branch point on the isoform ratios (27). This is a particularly hard task since computational identification of the branch point is an unsolved issue in the splicing field. Finally, the current limitation in studying isoform ratios is that the available transcript data reflect the natural situations with low resolution. In the future a considerably higher amount of transcript data provided by next-generation sequencing technologies might allow an accurate approximation of isoform ratios and ultimately to decipher the splicing code completely.

CONCLUSIONS

BNs can produce highly reliable predictions of NAGNAG-splicing outcomes once transcript evidence had been carefully considered to create training dataset. This indicates that we have identified, on a qualitative level, the most important features of the ‘NAGNAG-splicing code’. As a BN trained on human data achieved near-identical performance on other genomes from mouse to zebrafish and most of the information needed for prediction is encoded in the immediate splice site neighborhood, we conclude that the mechanism is simple in nature and maintained in evolution, as well as that our BN should be useful to annotate NAGNAG splicing in animal genomes that currently lack extensive transcript data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Martin Akerman for providing the dataset D1 and Nicole Ulbricht for skilful technical assistance.

FUNDING

Deutsche Forschungsgemeinschaft (SFB604) and the German Ministry of Education and Research (0313652D, 01GS0426, 01GR0504). Funding for open access charge: University of Freiburg.

Conflict of interest statement. None declared

REFERENCES

1. Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. and Shoemaker,D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
2. Tress,M.L., Martelli,P.L., Frankish,A., Reeves,G.A., Wesselink,J.J., Yeats,C., Olason,P.l., Albrecht,M., Hegyi,H., Giorgetti,A. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *PNAS*, **104**, 5495–5500.
3. Blencowe,B.J. (2006) Alternative splicing: new insights from global analyses. *Cell*, **126**, 37–47.
4. de la Grange,P., Dutertre,M., Correa,M. and Auboeuf,D. (2007) A new advance in alternative splicing databases: from catalogue to detailed analysis of regulation of expression and function of human alternative splicing variants. *BMC Bioinformatics*, **8**, 180.
5. Hiller,M. and Platzer,M. (2008) Widespread and subtle: alternative splicing at short-distance tandem sites. *Trends Gene.*, **24**, 246–255.
6. Sugnet,C.W., Kent,W.J., Ares,M. Jr and Haussler,D. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pacific Symposium on Biocomputing*, **9**, 66–77.
7. Hiller,M., Nikolajewa,S., Huse,K., Szafranski,K., Rosenstiel,P., Schuster,S., Backofen,R. and Platzer,M. (2007) TassDB: a database of alternative tandem splice sites. *Nucleic Acids Res.*, **35**, D188–D192.
8. Akerman,M. and Mandel-Gutfreund,Y. (2007) Does distance matter? Variations in alternative 3′ splicing regulation. *Nucleic Acids Res.*, **35**, 5487–5498.

9. Chern,T.-M., van Nimwegen,E., Kai,C., Kawai,J., Carninci,P., Hayashizaki,Y. and Zavolan,M. (2006) A simple physical model predicts small exon length variations. *PLoS Genetics*, **2**, e45.
10. Needham,C.J., Bradford,J.R., Bulpitt,A.J. and Westhead,D.R. (2006) Inference in Bayesian networks. *Nat. Biotech.*, **24**, 51–53.
11. Beaumont,M.A. and Rannala,B. (2004) The Bayesian revolution in genetics. *Nat. Rev. Genet.*, **5**, 251–261.
12. Zhang,M.Q. (1998) Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.*, **7**, 919–932.
13. Coolidge,C.J., Seely,R.J. and Patton,J.G. (1997) Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res.*, **25**, 888–896.
14. Fox-Walsh,K.L., Dou,Y., Lam,B.J., Hung,S.-p., Baldi,P.F. and Hertel,K.J. (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc. Natl Acad. Sci. USA*, **102**, 16176–16181.
15. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.
16. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
17. Pudimat,R., Schukat-Talamazzini,E.-G. and Backofen,R. (2005) A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics*, **21**, 3082–3088.
18. Nikolajewa,S., Pudimat,R., Hiller,M., Platzer,M. and Backofen,R. (2007) BioBayesNet: a web server for feature extraction and Bayesian network modeling of biological sequence data. *Nucleic Acids Res.*, **35**, W688–W693.
19. Friedman,N., Geiger,D. and Goldszmidt,M. (1997) Bayesian network classifiers. *Machine Learning*, **29**, 131–163.
20. Fayyad,U.M. and Irani,K.B. (1993) Multi-interval discretization of continuousvalued attributes for classification learning. *IJCAI*, **2**, 1022–1027.
21. Pudil,P., Novovicova,J. and Kittler,J. (1994) Floating search methods in feature selection. *Patt. Recognition Lett.*, **15**, 1119–1125.
22. Witten,I.H. and Frank,E. (2005) *Data Mining: Practical machine learning tools and techniques*, 2nd edn. Morgan Kaufmann, San Francisco.
23. Ling,C., Huang,J. and Zhang,H. (2003) *Canadian Artificial Intelligence Conference*. Heidelberg, Springer Berlin, pp. 329–341.
24. Szafranski,K., Schindler,S., Taudien,S., Hiller,M., Huse,K., Jahn,N., Schreiber,S., Backofen,R. and Platzer,M. (2007) Violating the splicing rules: TG dinucleotides function as alternative 3′ splice sites in U2-dependent introns. *Genome Biol.*, **8**, R154.
25. Hiller,M., Huse,K., Szafranski,K., Jahn,N., Hampe,J., Schreiber,S., Backofen,R. and Platzer,M. (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat. Genet.*, **36**, 1255–1257.
26. Akerman,M. and Mandel-Gutfreund,Y. (2006) Alternative splicing regulation at tandem 3′ splice sites. *Nucleic Acids Res.*, **34**, 23–31.
27. Tsai,K.-W., Tarn,W.-Y. and Lin,W.-c. (2007) Wobble splicing reveals the role of the branch point sequence-to-NAGNAG region in 3′ tandem splice site selection. *Mol. Cell Biol.*, **27**, 5835–5848.
28. Hollins,C., Zorio,D.A.R., Macmorris,M. and Blumenthal,T. (2005) U2AF binding selects for the high conservation of the C. elegans 3′ splice site. *RNA*, **11**, 248–253.
29. Tsai,K.-W., Lin, and , (2006) Quantitative analysis of wobble splicing indicates that it is not tissue specific. *Genomics*, **88**, 855–864.
30. Tadokoro,K., Yamazaki-Inoue,M., Tachibana,M., Fujishiro,M., Nagao,K., Toyoda,M., Ozaki,M., Ono,M., Miki,N., Miyashita,T. *et al.* (2005) Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: the case of Gln in DRPLA affects subcellular localization of the products. *J. Hum. Genet.*, **50**, 382–394.
31. Tsai,K.-W., Tseng,H.-C. and Lin,W.-c. (2008) Two wobble-splicing events affect ING4 protein subnuclear localization and degradation. *Exp. Cell Res.*, **314**, 3130–3141.
32. Atkinson,T.P. and Dai,Y. (2007) Activation-induced changes in alternate splice acceptor site usage. *Biochem. Biophys. Res. Commun.*, **358**, 590–595.
33. Schindler,S., Szafranski,K., Hiller,M., Ali,G., Palusa,S., Backofen,R., Platzer,M. and Reddy,A. (2008) Alternative splicing at NAGNAG acceptors in Arabidopsis thaliana SR and SR-related protein-coding genes. *BMC Genomics*, **9**, 159.

Rileen Sinha
Andreas D. Zimmer
Kathrin Bolte
Daniel Lang
Ralf Reski
Matthias Platzer
Stefan Rensing
Rolf Backofen

Identification and characterization of NAGNAG alternative splicing in the moss *Physcomitrella patens*

(submitted to BMC Plant Biology)

Alternative splicing (AS) involving NAGNAG tandem acceptors is widespread in animals, and also common in the model plant *Arabidopsis thaliana* and in the crop *Oryza sativa* (rice). In one of the first studies involving sequence-based prediction of AS in plants, we performed a genome-wide identification and characterization of NAGNAG AS in the model moss *Physcomitrella patens*. We found 591 alternative NAGNAGs in *P. patens* using all currently available transcript evidence. A naïve Bayes classifier trained using judiciously prepared training data predicted NAGNAG AS with a balanced specificity and sensitivity of ≥ 89%. Subsequently, we made genome-wide predictions of NAGNAG splicing outcomes, and 94% (18/19) of the experimentally tested predictions were verified. NAGNAG AS is just as common in *P. patens* as it is in *A. thaliana* and *O. sativa*, and can be predicted with high accuracy. The most informative features are the nucleotides in the NAGNAG and in its immediate vicinity, along with the splice sites, as found earlier for NAGNAG AS in animals. Our results suggest that the mechanism behind NAGNAG AS in plants is similar to that in animals and is largely dependent on the splice site and its immediate neighborhood.

Identification and characterization of NAGNAG alternative splicing in the moss *Physcomitrella patens*

Rileen Sinha[1,5], Andreas D. Zimmer[3], Kathrin Bolte[2,4], Daniel Lang[3], Ralf Reski[3,4,5], Matthias Platzer[6], Stefan A. Rensing[2,4,5*] and Rolf Backofen[1,4,5*]

Addresses:

[1] Bioinformatics group, University of Freiburg, Georges-Koehler-Allee 106, 79110 Freiburg, Germany

[2] Faculty of Biology, University of Freiburg, Hauptstrasse 1, 79104 Freiburg, Germany

[3] Plant Biotechnology, Faculty of Biology, University of Freiburg, Schaenzlestrasse 1, 79104 Freiburg, Germany

[4] Freiburg Initiative for Systems Biology (FRISYS), University of Freiburg, Schaenzlestrasse 1, 79104 Freiburg, Germany

[5] Centre for Biological Signalling Studies (bioss), University of Freiburg, Albertstr. 19, 79104 Freiburg, Germany

[6] Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany

Email addresses:

RS: sinha@informatik.uni-freiburg.de
ADZ: andreas.zimmer@biologie.uni-freiburg.de
KB: kathrin.bolte@biologie.uni-freiburg.de
DL: daniel.lang@biologie.uni-freiburg.de
RR: ralf.reski@biologie.uni-freiburg.de
MP: mplatzer@fli-leibniz.de
SAR: stefan.rensing@biologie.uni-freiburg.de
RB: backofen@informatik.uni-freiburg.de

*Authors for correspondence

Abstract

Background

Alternative splicing (AS) involving tandem acceptors that are separated by three nucleotides (NAGNAG) is an evolutionarily widespread class of AS, which is well studied in *Homo sapiens* (human) and *Mus musculus* (mouse). It has also been shown to be common in the model seed plants *Arabidopsis thaliana* and *Oryza sativa* (rice). In one of the first studies involving sequence-based prediction of AS in plants, we performed a genome-wide identification and characterization of NAGNAG AS in the model plant *Physcomitrella patens*, a moss.

Results

Using Sanger data, we found 295 alternatively used NAGNAG acceptors in *P. patens*. Using 31 features and carefully constructed training and test datasets of constitutive and alternative NAGNAGs, we trained a classifier to predict the splicing outcome at NAGNAG tandem splice sites (alternative splicing, constitutive at the first acceptor, or constitutive at the second acceptor). Our classifier achieved a balanced specificity and sensitivity of 89%. Subsequently, a classifier trained exclusively on data well supported by transcript evidence was used to make genome-wide predictions of NAGNAG splicing outcomes. By generation of more transcript evidence from a next-generation sequencing platform (Roche 454), we found additional evidence for NAGNAG AS, with altogether 664 alternative NAGNAGs being detected in *P. patens* using all currently available transcript evidence. The 454 data also enabled us to validate the predictions of the classifier, with 64% (80/125) of the well-supported cases of AS being predicted correctly.

Conclusion

NAGNAG AS is just as common in the moss *P. patens* as it is in the seed plants *A. thaliana* and *O. sativa* (but not conserved on the level of orthologous introns), and can be predicted with high accuracy. The most informative features are the nucleotides in the NAGNAG and in its immediate vicinity, along with the splice sites scores, as found earlier for NAGNAG AS in animals. Our results suggest that the mechanism behind NAGNAG AS in plants is similar to that in animals and is largely dependent on the splice site and its immediate neighborhood.

Background

Eukaryotic primary mRNAs consist of protein-coding regions (exons) and intervening non-coding regions (introns). The mature mRNA transcript, which acts as substrate for translation into protein, is produced by removing introns in a process called splicing. Splicing can be either constitutive, always producing the same mRNA, or alternative, via variable inclusion of parts of the primary transcript. Alternative splicing (AS) is thus a mechanism that enables multiple transcripts and proteins to be encoded by the same gene, thereby promoting transcript and protein diversity [1]. Furthermore, events of AS can provide an additional level of post-transcriptional gene regulation, e.g. by the production of mRNA isoforms with truncated open reading frames that are subject to degradation by the nonsense mediated decay pathway [2, 3]. It is particularly widespread in higher eukaryotes, especially in mammals – it has been estimated that up to 94% of all multi-exonic *H. sapiens* genes are alternatively spliced [4]. Large-scale detection of AS usually involves expressed sequence tags (ESTs), microarray, or RNA-Seq analysis. However, due to various sampling biases such as bias towards specific tissues and/or developmental stages, not all AS events can be detected by these methods. Furthermore, exon and splice arrays usually do not probe short distance events. Moreover, nowadays genomic sequence data is being churned out at a much faster rate than transcript data, that is, many genomes have low transcript coverage. Thus, there is a need for independent methods of detecting AS.

It has been shown that AS involving alternative donors/acceptors separated by 2-12 nt, also called "subtle alternative splicing" (due to the small difference in length in the transcript isoforms), is an evolutionarily widespread class of AS among animals, and among these, NAGNAG AS, involving acceptors separated by 3 nt, is the most common [5-8]. The terminology "NAGNAG" refers to events of AS that involve two acceptors (two "AG"s) which are preceded by any of the four possible nucleotides (N = A, C, G or T). Hence, the generic pattern for such tandem splice sites is "NAGNAG". NAGNAG AS can result in one of three possibilities (Fig. 1) – constitutive use of the first acceptor (the so-called exonic, or "E" variant), constitutive use of the second acceptor (the so-called intronic, or "I" variant), or use of both acceptors, that is, alternative splicing (the "EI" variant) [5]. NAGNAGs contribute 45% of all conserved alternative acceptors in *H. sapiens* and *M. musculus* [9]. Since the difference between the two isoforms is 3 nucleotides, no frameshift is induced, and the usual impact of a NAGNAG AS event is the insertion or deletion of one amino acid. In a recent study, we predicted the splicing outcome at NAGNAG acceptors in seven animal genomes (human, mouse, rat, dog, chicken, fruit fly and nematode) with a high degree of accuracy, and

83% of the experimentally validated cases agreed with the predictions [10]. In agreement with previous studies [11, 12] the study indicated that, the mechanism behind NAGNAG AS seems to be simple, stochastic and conserved in vertebrates and beyond.

While there have been numerous experimental as well as computational studies of AS in animals, the study of AS in plants is still in its early stages [13]. Although AS is commonly observed in plants, the overall abundance of AS seems to be lower than in animals. Several studies have estimated that between 20%-30% of plant genes undergo AS [14-17], while the current estimate based on deep sequencing of the *Arabidopsis thaliana* transcriptome is 42%-56% of intron-containing genes [18]. In comparison to studies in animals, EST-based detection of AS in plants lagged a few years behind, but revealed that intron retention appears to be the most common kind of AS event in plants [13-16]. Exon-skipping, which is the most common event in animals [19], is much less frequent in plants [14, 16]. The two prevalent models for spliceosome assembly are intron-definition, which applies to short introns (thus to a majority of plant introns) and involves the intron as the initial unit of recognition during spliceosome assembly; and exon-definition, which applies to long introns introns (thus to a majority of animal introns), and involves recognition of the exon as the initial unit for splicing [14, 20-22]. Thus, one would expect inaccurate splicing to result in intron-retention under the intron-definition model, and exon-skipping under the exon-definition model [13]. Hence the results showing that intron-retention is the most common AS event in plants and exon-skipping in animals are consistent with these models of splicing. However, alternative acceptors and donors seem to occur at a comparable frequency [16]. In particular, short distance or subtle AS events, seem to be just as common, and NAGNAG acceptors are widespread and abundant; a study on AS found 953 alternative NAGNAGs in rice and 485 in *A. thaliana* [16].

Initial analyses of the model plant *P. patens*, the first sequenced bryophyte, indicated a distribution of AS events similar to other plants studied so far [23]. Consequently, we here aimed to characterize and predict the extent of NAGNAG AS in *P. patens*. Analysis of the available transcript data indicates that NAGNAG AS is just as common in the moss *P. patens* as in seed plants. We achieved a high level of performance *in silico*, and 64% of the cases of well-supported AS using independently generated 454 data could be correctly predicted. In agreement with a recent study comparing *A. thaliana* and *O. sativa* with mammals [24], our results suggest that the mechanism of NAGNAG AS is similar in plants and animals.

Results and discussion

Identification of alternative NAGNAGs using Sanger ESTs

Since the extent of NAGNAG occurence in *P. patens* had not been reported, we sought to identify genomic NAGNAGs. To find all genomic NAGNAGs (constitutive as well alternative), we looked for all annotated intron-exon boundaries which had an AG at three positions upstream or downstream of the annotated acceptor. This yielded 9,427 NAGNAG motifs, of which 5,031 were covered by Sanger ESTs. Cases where the EST evidence supported only one of the NAG acceptors were called constitutive, whereas cases with EST support for both acceptors were called alternative (here EST support means at least one EST from a high quality alignment, as described in the section "Material and Methods"). 295 (5.9%) of the detected 5,031 NAGNAGs with Sanger EST coverage were alternatively spliced (EI form), while 2,695 (53.6%) were exclusively spliced at the first (intron proximal) acceptor (E form, i.e. part of the NAGNAG is exonic) and 2,041 (40.5%) were spliced only at the second (intron distal) acceptor (I form, i.e. the entire NAGNAG is intronic). Thus, NAGNAG AS is common in *P. patens*. Sequence logos for all NAGNAG splice sites as well as for EI, E and I sites are visualized in Fig. 2.

Gene ontology enrichment analysis

To assess whether genes with NAGNAG AS in *P. patens* are enriched for specific functional categories and whether there is any similarity with *A. thaliana* and *O. sativa* in that sense, we analyzed Gene Ontology (GO) term annotations with GOSSIP [25]. GO terms with a FDR corrected p-value (q-value) less than 0.05 were considered significantly different. We found that 42 genes with the term plastid (GO:0009536, q-value 0.043) are statistically enriched (Table 1) in the set of *P. patens* genes with EST support for an alternative NAGNAG acceptor (225 genes). This could be confirmed by the GOSSIP analysis for the *P. patens* alternative NAGNAG genes supported by Sanger and 454 reads (498 genes, q-value: 8.35E-04). In addition, the terms organelle and mitochondrion (Table 1) were found to be enriched among the NAGNAG genes in *P. patens*. "DNA binding" (GO:0003677) which is reported for *A. thaliana* and *O. sativa* to be enriched in alternative NAGNAG genes [24], could not be observed for *P. patens*. To further examine this inconsistency, the supported alternative NAGNAG genes from *A. thaliana* (combined gene set from [26] and [24]) were subjected to GO enrichment analysis as well (Table 1). This analysis confirms the term "DNA binding" as overrepresented among the *A. thaliana* NAGNAG genes (q-value: 2.28E-04), thus consistent with the analyses for *A. thaliana* and *O. sativa* [24] as well as for mouse [12]. However, “DNA binding” was not found to be enriched in *P. patens* NAGNAG genes. In contrast to the analyses for the seed plants and mouse, Fisher’s exact test with false discovery rate corrected p-values was used here, instead of a chi-square test. However, if the parent term of DNA binding, nucleotide binding (GO:0000166) was also subjected to a chi-square test for *P. patens*, it was found to be enriched (p = 0.02). The fact that “DNA binding” was not found to be enriched in *P. patens* NAGNAGs might be due to the current status of the *P. patens* annotation (v1.2) – e.g. in many cases the gene models lack 5’ and 3’ regions and therefore do not cover the whole protein sequence. On the other hand, mosses and vascular plants diverged more than 450 million years ago and thus *P. patens* alternative NAGNAG acceptor genes might be different. Nevertheless, the GO enrichment analysis in terms of the category “cellular component” reveals that *A. thaliana* as well as *P. patens N*AGNAG genes share a bias towards the term “intracellular organelle”, which includes “nucleus” and “plastid” (Table 1). In addition to the enriched molecular function “DNA binding”, our analysis confirmed the functions “RNA binding”, “transcription factor activity” and “transcription regulator activity” to be also slightly enriched in *A. thaliana* alternative NAGNAG acceptor genes (Table 1), which is coherent with reports for *M. musculus* [12].

Evolutionary conservation of NAGNAG splicing among plants?

Seven clusters of homologous genes with AS at NAGNAG acceptors in the same intron were reported to be conserved between *A. thaliana* and *O. sativa* [24]. In order to check whether AS NAGNAG events are conserved between *A. thaliana* and *P. patens*, a BLAST based single linkage clustering was performed, using all transcripts with a Sanger-supported NAGNAG. Altogether, 1,088 clusters containing *A. thaliana* and *P. patens* genes were identified, of which five clusters contained genes with a NAGNAG motif at the orthologous (as evidenced by numbering from the transcription start site) intron. Five out of the seven *P.patens* genes in these clusters were selected for experimental validation. In all cases only one of the two isoforms could be detected, which is consistent with the support by Sanger ESTs, which in all cases supported only one of the two isoforms. In addition, in all cases the NAGNAG motif itself is not conserved between *A. thaliana* and *P. patens* (table 3). In *A. thaliana*, only one of the NAGNAG motifs contains a GAG, whereas four of five in *P. patens* contain a GAG, and are therefore unlikely to represent alternative NAGNAGs ([5, 10, 12, 27], and section below). Given the assumption that we are looking at orthologous or at least homologous positions and our transcript evidence is sufficient, this observation can be explained by two possible evolutionary scenarios. In the first scenario the alternative NAGNAG sites are ancestral and have diverged in the lineages leading to *A. thaliana* and *P. patens***.** While they might have been inactivated by the introduction of a GAG in the moss *P. patens*, they have been retained functional in *A. thaliana*. In the second scenario these alternative NAGNAG acceptors in *A. thaliana* arose after the divergence of mosses and seed plants. Given the current scarce data, both scenarios appear equally parsimonious. In order to decide which scenario is true, additional taxa would have to be included into the analysis. Given the current data and analyzes there is evidence for conserved NAGNAG AS events between *O. sativa* and *A. thaliana*, but not between *P. patens* and *A. thaliana*. Thus, it appears as if NAGNAG AS is not conserved across several hundreds of millions of years [28] or arose secondarily.

Prediction of NAGNAG AS in *P. patens*

The most crucial prerequisite for good prediction performance is a reliable training dataset. It is critical that the samples are correctly labelled as far as possible. In terms of datasets of alternative and constitutive exons, this means that we should use the available transcript evidence judiciously, in order to minimise mislabelling. In other words, we want to avoid the contamination of the set of constitutive exons by alternative exons which currently lack transcript support for being alternative, as well of alternative exons by potentially erroneously labelled exons. Thus, we used filters on the transcript support to improve the reliability of the labels – as in our previous work on NAGNAG AS prediction in animals [10], a training set was constructed based on the following criteria:

(i) constitutive: ≥ 10 ESTs supporting either E or I variant, 0 for the other;

(ii) alternative: ≥ 2 ESTs supporting each variant, ≥ 10% of ESTs supporting minor variant.

This yielded a training dataset of 833 NAGNAGs - 696 constitutive (424 E, 272 I) and 137 EI, or alternative cases. The classifiers were trained using this dataset. The remaining 4,198 NAGNAGs (2,271 E, 1,769 I, 158 EI) were used as a test set. It is noteworthy that the average coverage per constitutive NAGNAG in this set is only three ESTs (for both E as well as I cases), indicating that there are potentially many undiscovered alternative NAGNAGs in *P. patens*. The training data was used with a classifier (we used naïve Bayes classifiers, Bayesian networks, and support vector machines, all of which yielded very similar performance) in a cross-validation setting. Briefly, the classifier uses part of the training data to learn a model based on the sample labels and the features, and then uses this model to assign posterior probabilities (P(EI), P(E) and P(I) according to the three possible classed) to each sample. The predicted NAGNAG class is the one which receives the maximum score or posterior probability from the classifier. We computed the receiver operating characteristics (ROC), which is a plot of the true positive rate versus the false positive rate, and measured the area under the ROC curve (AUC), which is a standard measure of the quality of a classifier [29]. An ideal classifier, which makes no errors, would achieve an AUC of 1.0. We used 31 features, and achieved an *in silico* performance of AUC = 0.96, 0.99 and 0.98 for the EI, E and I forms, respectively (Fig. 3). This performance was obtained under various cross-validation settings (2-fold, 5-fold, 10-fold, leave-one-out – where n-fold cross-validation means that (n-1)/n of the dataset is used to learn, and the remaining 1/n for prediction – this is repeated n times, and the average performance is reported).

Generation of additional transcript evidence

As mentioned above, average transcript support was found to be low. In order to generate more evidence for alternative acceptors, next generation sequencing was carried out. For this purpose, adult gametophores carrying gametangia (for review of moss tissues see: [30]) were grown, as this tissue was not well represented in the pre-existing ~400,000 Sanger reads. In addition, the cDNA was normalized in order to equalize transcript abundance and thus avoid redundancy. While the ~400,000 Sanger reads map to 19,186 gene models, the ~600,000 454 reads map to 20,161 gene models. The 454 reads map to a total of 2,545 gene models that were not covered previously, and identified 73 alternative NAGNAGs. Even though the 454 data cover only 75% (3,745/5,031) of the NAGNAGs evidenced by Sanger ESTs, they enabled detection of 371 alternative NAGNAGs – 9.9% of the covered NAGNAGs, as compared to 7.5% using Sanger ESTs. Of these 371, 117 were previously identified by Sanger ESTs. There are 42 NAGNAGs which have support for only one acceptor in the Sanger data, and for only the other acceptor in 454 data. Combining the results from Sanger and 454 data, *P. patens* has 664 alternative NAGNAGs. Again these results show that NAGNAG AS is as widespread in the moss *P. patens* as it is in the seed plants *A. thaliana* and *O. sativa*.

Experimental confirmation of the NAGNAG AS

Experiments were performed on 19 candidate NAGNAGs, 14 as controls (seven with AS according to transcript data, and six without AS) to see whether the splicing outcomes according to Sanger and 454 reads could be confirmed by a PCR based approach, and five on the basis of an orthologous alternative NAGNAG intron in *A. thaliana* (see above). Of the seven candidates with support for AS from Sanger or 454 datasets, three were predicted to be alternative spliced with p(EI) values >0.9 (Table 2). Using Sanger sequencing of cDNA based PCR products, all three candidates were indeed verified as being alternatively spliced in *P. patens* protonema and gametophore tissue, respectively. Eight candidate genes were used as potential negative controls, as their p(EI) predictions were 0.365 and lower. All candidates showed support for the single predicted isoform by means of available transcript evidence and consequently only this single isoform could be detected during experimental validation (Table 2). Having support for both variants from either the Sanger or the 454 datasets, but a p(EI) <0.9, four more candidates were chosen to be validated. NAGNAG AS could be confirmed for the gene product Phypa_161321 by Sanger sequencing of cDNA PCR products, although it has a low p(EI) of 0.181 (Table 2). The experimental validation is supported by the Sanger dataset, where 13 "E" variants as well as 27 "I" variants could be identified. This is the only case where prediction from the Naïve Bayes Classifier does not agree with the experimental results. In case of Phypa_74146 and Phypa_199161, only one of the two isoforms could be detected, reflecting the low p(EI) values.

The sensitivity of Sanger sequencing allows detection of AS for ratios of the two isoforms of about 3:1 or lower, meaning that cases in which the minor isoform abundance is < 25%, AS may go undetected even if present. Therefore, validation using fluorescence labeled forward primers and fragment length detection on a capillary sequencer was used to detect the minor isoform abundance for two examples. In case of Phypa_161321 (Table 2) the received data determined by Sanger sequencing of PCR products could be confirmed by the more sensitive detection using the fluorescence labeled forward primers. The two isoforms with three nucleotides difference in length were detected using capillary separation and had a relative abundance of approximately 3:1 (exonic "E" versus intronic "I" variant) (Fig. 4). In case of Phypa_228333, only one of the two isoforms could be detected by Sanger sequencing as well as in the more sensitive validation using fluorescence labeled primers (Table 2). Thus, a low p(EI) prediction for this candidate seems to be correct as is the case for Phypa_74146 and Phypa_199161, for which only one of the two isoforms could be identified as predicted. Detection of both isoforms either in Sanger datasets (Phypa_199161) or in the 454 datasets (Phypa_74146 and Phypa_228333) could be explained by the higher sensitivity of sequencing as compared to the PCR-based approach or by the fact that adult gametophores were used to generate the 454 data, while the validation was carried out in the two principal tissues of the juvenile stage. Thus it cannot not be excluded that these candidates are indeed alternatively spliced.

GAG acceptors

Twelve of the 19 candidate genes possess a GAG in the NAGNAG motif (Table 2). Using the above described methods, all of them are shown to be not alternatively spliced. Therefore, GAG seems not to be used as an alternative acceptor for AS in *P. patens* in most cases, which is in line with the sequence logos (Fig. 2B). Exceptions could be Phypa_199161 and Phypa_228333, which possess both isoforms regarding Sanger and 454 datasets. These two candidates may indeed use GAG as acceptors for AS, but this remains to be proven. Rare usage of GAGs as acceptors in *P. patens* is in agreement with previous work which shows that functional acceptors are only very rarely GAGs – the order of preference for the nucleotide preceding the AG in functional acceptors is C > T > A > G, which has been shown both by experimental work [31] as well as by *in silico* analyses of NAGNAG splicing [5, 12]. When we consider the EST and 454 evidence in *P. patens*, only 4.6% (149/3225) of GAG-containing NAGNAGs are alternative – filtering by transcript support to use only well-supported cases (as described for the preparation of training data in the "Material and Methods" section) further reduces this to 2.6% (14/536). Taken together, this strongly suggests that GAGs function only very rarely as functional acceptors in *P. patens* (if at all).

Using 454 data for independent validation of predictions

The classifier was trained based on previously existing Sanger evidence, the additional 454 evidence was used for independent validation. Combining the 454 and Sanger datasets resulted in 296 additional NAGNAG AS events being detected – of these, 66 had strong support for AS in terms of satisfying the criteria used to define the training dataset (≥ 2 reads for each variant, ≥ 10% of the reads for the minor variant) . 62% (41/66) of these were predicted to be alternative by the Naïve Bayes classifier. If we require ≥ 4 reads per variant while keeping the threshold of minor variant abundance at ≥ 10 %, the correct predictions rise to 75% (9/12). When considering AS according to 454 reads alone, 64% (80/125) of the well-supported cases of AS are predicted correctly, which increases to 79% (30/38) if we require ≥ 4 reads per variant while keeping the threshold of minor variant abundance at ≥ 10%. On the other hand, if we look at cases which are constitutive with a support of ≥ 30 transcripts , according to the combined transcript dataset, only 1/93 E cases and 0/65 I cases are predicted to be alternative. The Naïve Bayes classifier predicts 371 further cases of AS (155 of 2,549 currently labeled E, and 216 of 1,891 currently labeled I) in *P. patens* – the high specificity shown by nearly no predicted AS in strongly supported constitutive NAGNAGs combined with the sensitivity of 62% in detecting newly discovered strongly supported cases of AS shows that there are potentially several hundred as yet undiscovered cases of NAGNAG AS in *P. patens*.

Prediction of NAGNAG AS in *P. patens* by a classifier trained on *H. sapiens* data

We had earlier shown that a classifier trained on only *H. sapiens* NAGNAG data could predict NAGNAG splicing outcomes with near-identical accuracy on other vertebrate genomes (mouse, rat, dog, chicken), and with a slight drop in the case of *D. melanogaster* and *Caenorhabditis elegans* [10]. Therefore, we also tried to predict NAGNAG AS in *P. patens* using a Naive Bayes classifier trained on *H. sapiens* data and achieved an AUC of 0.90, 0.99 and 0.97 for the EI, E and I forms, respectively. This was achieved using five features (the Ns in the NAGNAG, the two positions immediately upstream and the position immediately downstream) and is similar to that achieved on *D. melanogaster* earlier [10], reinforcing the notion that NAGNAG splicing in plants is similar to that in animals.

Conclusions

We here describe the first computational prediction of alternative splicing (AS) in a non-seed plant and found that NAGNAG AS in *P. patens*, a moss, can be predicted with high accuracy. Since the extent of NAGNAGs in *P. patens* had not yet been reported, this work involved both characterization as well prediction of NAGNAG splicing in *P. patens*. Using ESTs, we found that NAGNAG AS is as widespread in the bryophyte *P. patens* as it is in the seed plants *A. thaliana* and *O. sativa.* Thus, NAGNAG AS is likely to be a common feature of AS in all land plants, just as it is in animals. Although we detected homologs with NAGNAG events among the two land plants *P. patens* and *A. thaliana*, NAGNAG splicing seems not to be conserved at the intron level.

Using carefully constructed training and test datasets, an *in silico* performance of AUC = 0.96, 0.99 and 0.98 was achieved for the EI, E and I forms, respectively. The most informative features (according to information gain [32]) were the nucleotides in the NAGNAG and its immediate vicinity, and even a relatively simple classifier like the Naïve Bayes classifier could match the more sophisticated Bayesian network and Support vector machine. The performance achieved by a Naïve Bayes classifier trained on *H. sapiens* data (AUC = 0.90, 0.99 and 0.97 for the EI, E and I forms, respectively) was similar to that achieved on *D. melanogaster* earlier [10]. This indicates that, as in animals, the mechanism behind NAGNAG AS in plants is simple in nature and mostly dependent on the splice site neighborhood. Independent validation of the predictions of the classifier (trained on Sanger EST data alone) using 454 data showed that 64% (80/125) of the well-supported cases of NAGNAG AS could be predicted correctly.

In total, seven candidates were chosen for independent experimental confirmation of the Sanger and 454 evidence of NAGNAG splicing. The experimental confirmation depends on detection of isoforms using sequence electropherograms and is less sensitive than size polymorphism detection using fluorescence-labeled primers. The latter method was used on two of the seven examples and confirmed the results of the previous method. While there is transcript support for alternative use of GAG acceptors this could not be proven in our experimental validation. In addition, a further 12 experiments were performed – six as negative controls, all of which agreed with the predictions, and five to check for possible conserved NAGNAG AS with *A. thaliana*, which could not be detected.

When additional 454 transcript evidence was used to supplement the Sanger EST data, a total of 664 alternative NAGNAGs were found in *P. patens*. Since the average coverage per constitutive NAGNAG was still only ~10 ESTs, this number shall likely continue to rise with deeper coverage of the transcriptome. Nevertheless, the results provide the first evidence that NAGNAG AS is widespread in *P. patens*. Our findings are in agreement with a recent study which showed that NAGNAG AS shares common properties in *A. thaliana* and *O. sativa* and animals [24]. This indicates that the mechanism behind NAGNAG AS in land plants is similar to that in animals. The pervasiveness of NAGNAG AS suggests that it may be a general feature of splicing in animals and plants, and possibly in all eukaryotes.

Materials and Methods

Identification of alternative splicing at NAGNAG acceptors using ESTs

346,871 *P. patens* Sanger EST reads (available at http://www.cosmoss.org) from various developmental stages and tissue types (predominantly protonema and juvenile gametophores) were aligned using GenomeThreader [33]. EST alignments (max. intron length 20,000) with less than 95% identity and 90% EST length coverage were excluded from further analyses to obtain only reliable alternative acceptors. In addition, EST alignments matching a single exon as well as alignments ending at an exon boundary supporting either the E or I site were discarded The sequence regions used for feature extraction (Fig. 5) and EST evidence counts were created using the BioPerl [34] module Bio::DB::SeqFeature::Store.

Sequence logos

Sequence logos were created using the WebLogo software (http://weblogo.berkeley.edu/logo.cgi) [35] with the sequence regions shown in Fig. 5.

Feature design and extraction; classifiers

Feature extraction was done based on annotated data using a Perl script. The region used for analysis can be seen in Fig. 5. Since the composition of the splice site neighborhood influences splicing in general, the base pairs at positions -20 to +3 with respect to the NAGNAG were each used as a single feature, as were the two Ns in the NAGNAG motif. The last three positions of the upstream exon were also included, since they can influence both the process of splicing, as well as reflect influence of codon usage near the exon boundary. Thus, we had a total of 28 features which each represented a nucleotide, and thus had four possible values (A, C, G, T). A weak polypyrimidine tract (PPT) can contribute to AS, and the number of pyrimidines in the 3' region of the intron is a measure of PPT strength. Therefore, we designed a feature called "Y-content", which refers to the number of pyrimidines in the 20 bp upstream of the NAGNAG. Splice site strength, being one of the most important determinants of splicing outcome, was also included as a feature – the strength of the two possible splice sites for each NAGNAG exon, as computed using SpliceMachine [36], contributed two more features. In total, 31 features were used. We used the WEKA package and Bayesian Networks, Naive Bayes classifiers, and Support vector machines [32]. For feature selection within WEKA, we used the method "CfsSubsetEval". In addition, we also used manual inclusion and exclusion of features.

Information gain

Information gain is defined as the reduction in the entropy of the class variable, given the feature [32]. The formula for information gain is:

IG(Class | Feature) = H(Class) - H(Class | Feature)

where H(Class) is the entropy of the class variable, and H(Class|Feature) is the conditional entropy of the class variable, given the feature. Information gain is a well established measure for feature selection in Machine Learning. We used the WEKA package for computing information gain, in order to rank the features according to how informative they were. We also used it for prediction based on SVMs, as implemented in the SMO option, and for prediction using Naïve Bayes classifiers.

Functional annotation and GO enrichment analysis

For every (potential) NAGNAG splicing region an overlapping *P. patens* gene model was assigned using the start and stop coordinates on the genomic scaffolds. The corresponding predicted protein sequences were subjected to BLAST2GO [37] GO term annotation which was extended by various subcellular target prediction and homology-based methods (see http://www.cosmoss.org/annotation/references?cosmoss_ref=1 for details). The resulting GO annotation was mapped to GO slim terms using the Blast2GO internal mapping function using the "goslim_plant.obo" ontology subset. GO enrichment analysis was performed against the complete *P. patens* with the BLAST2GO internal Fisher's exact test/GOSSIP [38] using the two-tailed test, with false discovery rate (FDR) correction and a q-value cut-off < 0.05. The *A. thaliana* alternative NAGNAG splicing gene set was constructed using the alternative NAGNAG acceptor cases identified within the *A. thaliana* genome from [26] and [24]. The resulting alternative NAGNAG acceptor set contains 290 *A. thaliana* proteins. These proteins were subjected to a GO enrichment analysis as described above for *P. patens*. The *A. thaliana* GOA was downloaded from ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.txt (17.11.2009) and mapped to GO slim (goslim_plant.obo) with BLAST2GO.

Candidate selection for evolutionary conserved NAGNAG acceptors

P. patens cosmoss v1.2 and *A. thaliana* TAIR 8 proteins were subjected to a BLAST based single linkage clustering using BLASTCLUST [39]. The parameters were set to 70% length coverage and 70% alignment identity to obtain only highly conserved homologs. In total

1,088 clusters with at least one *P. patens*, respectively *A. thaliana*, protein were found. Five candidates out of seven *P. patens* genes, each sharing a cluster with *A. thaliana* alternative NAGNAG acceptor containing genes [24, 26], were selected for experimental validation. In addition, these *P. paten*s candidate genes contain a potential NAGNAG acceptor in the same intron as the corresponding *A. thaliana* homolog.

Experimental confirmation of splice variants

P. patens total RNA was isolated from protonema and gametophore tissue using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). cDNA synthesis was carried out with 250ng total RNA using Superscript III Reverse Transcriptase (Invitrogen, Karlsruhe, Germany) according to the manufacturers' instructions. For validation of different splice variants, PCR was performed from protonema and gametophore RNA, respectively, using native *Pfu*-Polymerase (Fermentas, St. Leon-Rot, Germany). PCR primers were obtained from Sigma (München, Germany). PCR reactions were carried out using 12 ng cDNA as template. Products were extracted using the QIAquick PCR purification Kit (Qiagen, Hilden, Germany) and directly sequenced (GATC, Konstanz, Germany). Sequences and chromatograms were analysed with ChromasPro Version 1.34. Alternatively, PCR products amplified with carboxyfluorescein (FAM) labeled forward primers were analysed by capillary electrophoresis, where AS was detected as a size difference of three nucleotides in length. PCR products were diluted as appropriate and subjected to capillary electrophoresis for separation and detection. For this purpose, 10 µL HiDi formamide (Applied Biosystems) and 0.5 µL HD400 GS internal size standard were added to each well, and the plate was mounted on a 3100 Genetic Analyzer with Foundation Data Collection software v. 2.0 and Gene Mapper ID software v. 3.2 (Applied Biosystems, Darmstadt, Germany).

Tissue culture and generation of additional transcript evidence

Physcomitrella patens strain Gransden 2004 [23] was cultivated on solidified (1% w/v agar) mineral medium [250 mg L-1 KH2PO4, 250 mg L-1 MgSO4x7-H2O, 250 mg L-1 KCl, 1000 mg L-1 Ca(NO3)2x4H2O, 12.5 mg L-1 FeSO4x7H2O, pH 5.8 with KOH] on 9 cm petri dishes enclosed by laboratory film in a Percival cultivation chamber (CLF, Germany) at 22° C with a 16h light, 8h dark regime under 70µmol*s-1*m-2 white light (long day conditions). Gametophore colonies were grown from single gametophores transferred to the dishes from precultured colonies. Induction of gametangia was performed by placing the dishes under inductive conditions [40], i.e. 20µmol *s-1*m-2 white light and 15° C with a 8h light, 16h dark regime until development of gametangia. After harvesting and freezing, the material was ground under liquid nitrogen and total RNA isolated using the Ambion mirVana miRNA isolation kit (Applied Biosystems, Darmstadt, Germany). RNA isolation and subsequent sequencing pool creation steps were carried out by Vertis Biotechnologie (Freising, Germany). Poly(A)+ RNA was prepared by oligo(dT) chromatography and cDNA was synthesized using a N6 randomized primer. Afterwards, 454 adapters A (CCATCTCATCCCTGCGTGTCTCCGACTCAG) and B (CTGAGACTGCCAAGGCACACAGGGGATAGG) were ligated to the 5' and 3' ends of the cDNA. The resulting N0 cDNA was amplified using PCR (16 cycles) with a proof reading enzyme. Normalization was carried out by one cycle of denaturation and reassociation of the cDNA, resulting in N1-cDNA. Reassociated ds-cDNA was separated from the remaining ss-cDNA (normalized cDNA) by passing the mixture over a hydroxylapatite column. After hydroxylapatite chromatography, the ss-cDNA was amplified with 9 PCR cycles. Finally, the cDNA in the size range of 500–700 bp was eluted from a preparative agarose gel and subjected to GS FLX Titanium sequencing (GATC, Konstanz, Germany), resulting in 631,313 raw reads. After low quality and adapter clipping using LUCY [41] and SeqClean (http://compbio.dfci.harvard.edu/tgi/software/), and polyA-tail removal with trimmest [42], 589,283 reads with a mean length of 343 nucleotides remained. The 454 reads are available at www.cosmoss.org and were mapped against the genome as described above for the *P. patens* Sanger ESTs.

Authors' contributions

RS analyzed the transcript evidences for NAGNAGs, designed and extracted features, used WEKA to perform the classification and analyses, computed the genome-wide prediction, and drafted the manuscript. ADZ obtained the transcript evidences for all genomic NAGNAGs in *P. patens*, helped in feature extraction, performed the GO-analysis, and contributed to writing the manuscript. KB performed the experimental validations and contributed to writing the manuscript. DL participated in the transcript evidence analyses, and helped with the GO-analysis. RR and MP supervised part of the work. RB and SAR contributed to writing the manuscript, conceived of and supervised the study. All authors have read and approved the final manuscript.

Acknowledgments

This work was supported by the German Research Foundation (DFG grant Re 837/10-2 to R.R. and S.A.R.), by the Federal Ministry of Education and Research (BMBF grant FRISYS 0313921 to R.B., R.R. and S.A.R.) and by the Excellence Initiative of the German Federal and State Governments (EXC 294 to R.B., R.R. and S.A.R.). We are grateful to S. Richardt and E. Heupel for skillful technical assistance and to M. Heinrich for analysing the FAM-labeled PCR products.

References

1. Graveley BR: **Alternative splicing: increasing diversity in the proteomic world**. *Trends in Genetics* 2001, **17**(2):100-107.
2. Hughes TA: **Regulation of gene expression by alternative untranslated regions**. *Trends in Genetics* 2006, **22**(3):119-122.
3. Stalder L, Mühlemann O: **The meaning of nonsense**. *Trends in Cell Biology* 2008, **18**(7):315-321.
4. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.
5. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M: **Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity**. *Nat Genet* 2004, **36**(12):1255-1257.
6. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Group RG, Members GSL, Hayashizaki Y, Gaasterland T: **Impact of Alternative Initiation, Splicing, and Termination on the Diversity of the mRNA Transcripts Encoded by the Mouse Transcriptome**. *Genome Res* 2003, **13**(6b):1290-1300.
7. Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ: **Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site**. *RNA* 2006, **12**(12):2047-2056.
8. Ermakova EO, Nurtdinov RN, Gelfand MS: **Overlapping alternative donor splice sites in the human genome**. *Journal of Bioinformatics and Computational Biology* 2007:991–1004.
9. Sugnet CW, Kent WJ, Jr. AM, Haussler D: **Transcriptome and Genome Conservation of Alternative Splicing Events in Humans and Mice**. *Pacific Symposium on Biocomputing* 2004, **9**:66-77.
10. Sinha R, Nikolajewa S, Szafranski K, Hiller M, Jahn N, Huse K, Platzer M, Backofen R: **Accurate prediction of NAGNAG alternative splicing**. *Nucl Acids Res* 2009, **37**(11):3569-3579.
11. Chern T-M, van Nimwegen E, Kai C, Kawai J, Carninci P, Hayashizaki Y, Zavolan M: **A Simple Physical Model Predicts Small Exon Length Variations**. *PLoS Genetics* 2006, **2**(4):e45.
12. Akerman M, Mandel-Gutfreund Y: **Alternative splicing regulation at tandem 3' splice sites**. *Nucl Acids Res* 2006, **34**(1):23-31.
13. Barbazuk WB, Fu Y, McGinnis KM: **Genome-wide analyses of alternative splicing in plants: Opportunities and challenges**. *Genome Research* 2008, **18**(9):1381-1392.
14. Wang B-B, Brendel V: **Genomewide comparative analysis of alternative splicing in plants**. *PNAS* 2006, **103**(18):7175-7180.
15. Wang B-B, O'Toole M, Brendel V, Young N: **Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes**. *BMC Plant Biology* 2008, **8**(1):17.
16. Campbell M, Haas B, Hamilton J, Mount S, Buell CR: **Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis**. *BMC Genomics* 2006, **7**(1):327.
17. Ner-Gaon H, Leviatan N, Rubin E, Fluhr R: **Comparative Cross-Species Alternative Splicing in Plants**. *Plant Physiol* 2007, **144**(3):1632-1641.
18. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong W-K, Mockler TC: **Genome-wide mapping of alternative splicing in Arabidopsis thaliana**. *Genome Research* 2009:-.

19. Kim E, Magen A, Ast G: **Different levels of alternative splicing among eukaryotes**. *Nucl Acids Res* 2007, **35**(1):125-131.

20. Berget SM: **Exon recognition in vertebrate splicing**. *J Biol Chem* 1995, **270**:2411 - 2414.

21. Lorkovic ZJ, Kirk DAW, Lambermon MHL, Filipowicz W: **Pre-mRNA splicing in higher plants**. *Trends in Plant Science* 2000, **5**(4):160-167.

22. Lim LP, Burge CB: **A computational analysis of sequence features involved in recognition of short introns**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(20):11193-11198.

23. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud P-F, Lindquist EA, Kamisugi Y *et al*: **The Physcomitrella Genome Reveals Evolutionary Insights into the Conquest of Land by Plants**. *Science* 2008, **319**(5859):64-69.

24. Iida K, Shionyu M, Suso Y: **Alternative Splicing at NAGNAG Acceptor Sites Shares Common Properties in Land Plants and Mammals**. *Mol Biol Evol* 2008, **25**(4):709-718.

25. Bluthgen N, Brand K, Cajavec B, Swat M, Herzel H, Beule D: **Biological profiling of gene groups utilizing Gene Ontology**. *Genome Inform* 2005, **16**(1):106-115.

26. Schindler S, Szafranski K, Hiller M, Ali G, Palusa S, Backofen R, Platzer M, Reddy A: **Alternative splicing at NAGNAG acceptors in Arabidopsis thaliana SR and SR-related protein-coding genes**. *BMC Genomics* 2008, **9**(1):159.

27. Hiller M, Szafranski K, Sinha R, Huse K, Nikolajewa S, Rosenstiel P, Schreiber S, Backofen R, Platzer M: **Assessing the fraction of short-distance tandem splice sites under purifying selection**. *Rna* 2008, **14**(4):616-629.

28. Lang D, Zimmer AD, Rensing SA, Reski R: **Exploring plant biodiversity: the Physcomitrella genome and beyond**. *Trends in Plant Science* 2008, **13**(10):542-549.

29. Ling C, Huang J, Zhang H: **AUC: a better measure than accuracy in comparing learning algorithms.** In: *Canadian Artificial Intelligence Conference 2003*; 2003: 329–341.

30. Reski R: **Development, genetics and molecular biology of mosses.** . *Botanica Acta* 1998, **111**:1-15.

31. Hollins C, Zorio DAR, Macmorris M, Blumenthal T: **U2AF binding selects for the high conservation of the C. elegans 3' splice site**. *RNA* 2005, **11**(3):248-253.

32. Witten IH, Frank E: **Data Mining: Practical machine learning tools and techniques**, Second. edn: Morgan Kaufmann, San Francisco.; 2005.

33. Gremme G, Brendel V, Sparks ME, Kurtz S: **Engineering a software tool for gene structure prediction in higher organisms**. *Information and Software Technology* 2005, **47**(15):965-978.

34. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H *et al*: **The bioperl toolkit: Perl modules for the life sciences**. *Genome Research* 2002, **12**(10):1611-1618.

35. Crooks GE, Hon G, Chandonia J-M, Brenner SE: **WebLogo: A Sequence Logo Generator**. *Genome Res* 2004, **14**(6):1188-1190.

36. Degroeve S, Saeys Y, De Baets B, Rouze P, Van de Peer Y: **SpliceMachine: predicting splice sites from high-dimensional local context representations**. *Bioinformatics* 2005, **21**(8):1332-1338.

37. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.

38. Bluethgen N, Brand K, Cajavec B, Swat M, Herzel H, Beule D: **Biological profiling of gene groups utilizing Gene Ontology**. *Genome Inform* 2005, **16**(1):106-115.

39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool**. *Journal of Molecular Biology* 1990, **215**(3):403-410.

40. Hohe A, Rensing SA, Mildner M, Lang D, Reski R: **Day length and temperature strongly influence sexual reproduction and expression of a novel MADS-box gene in the moss Physcomitrella patens**. *Plant Biology* 2002, **4**(5):595-602.

41. Chou H-H, Holmes MH: **DNA sequence quality trimming and vector removal**. *Bioinformatics* 2001, **17**(12):1093-1104.

42. Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open Software Suite**. *Trends in Genetics* 2000, **16**(6):276-277.

Figure legends

Figure 1. NAGNAG alternative splicing. Nomenclature of NAGNAG AS with E and I sites and isoforms.

Figure 2. Sequence logos of NAGNAG splice sites – the first three positions represent the last 3 nucleotides (nt) of the upstream exon, followed by the 30 nt upstream of the NAGNAG, the NAGNAG motif itself, and the 10 nt downstream of the NAGNAG (total 49 positions).
A: all splice sites; B: EI sites, C: E sites; D: I sites.

Figure 3. ROC plot depicting the *in silico* performance on the 3-class [I-class (red), E-class (green), and EI-class (blue)] classification problem. The EI-class, or AS, harder to predict (AUC = 0.96) than the two constitutive variants, E and I (AUC = 0.99 for both).

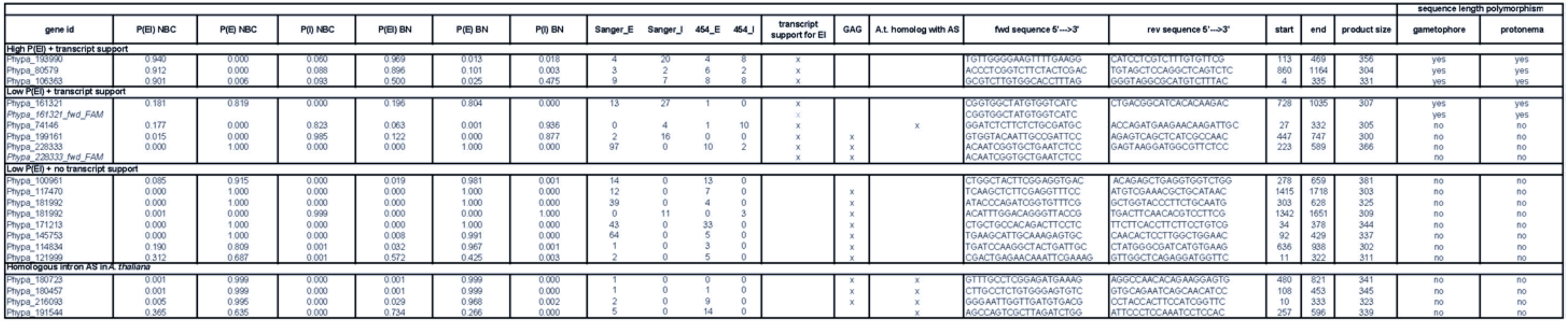
Figure 4. Example of the validation procedures employed.
A: Electropherogram of the direct sequencing of a cDNA PCR product. Starting with the NAGNAG AS site at position 132 the two polymorphic sequence signatures are overlaid.
B: FAM fluorescence intensity peaks of the two polymorphic isoforms (length difference three nucleotides). The lower peak constitutes 40% of the area (ar) of the bigger one, i.e. an approximate ratio of 3:1.

Figure 5. **Nomenclature of features used in this study**
Nomenclature of sequence features used to analyze NAGNAG splicing. The region used to derive all 31 features is shown, along with the names given to the positional features.

Tables

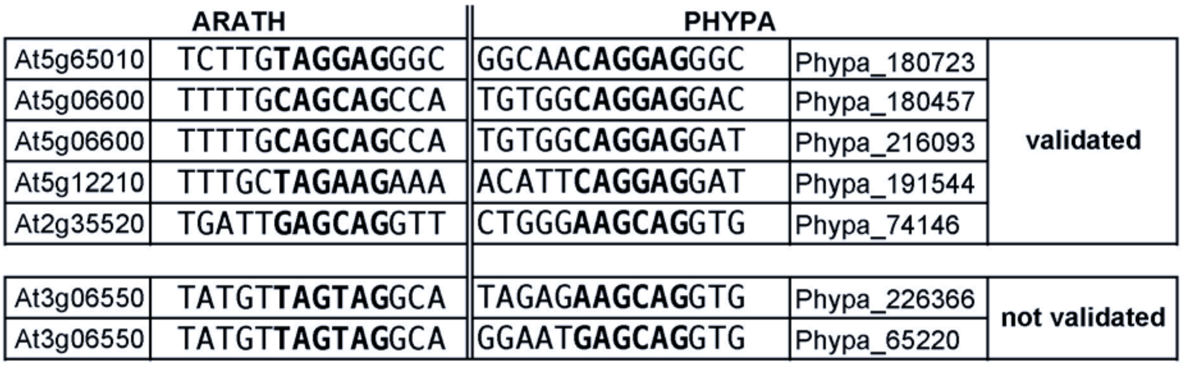
Organism/subset	Name	GO Term	FDR (q-value)	# in test group	# in reference group	# non annot. test	# non annot. reference group
P. patens **alternative NAGNAG genes Sanger support**	cytoplasmic part	GO:0044444	0.013301	78	5407	100	11894
	plastid	GO:0009536	0.042909	42	2607	136	14694
	membrane-bound organelle	GO:0043227	0.042909	89	6778	89	10523
	intracellular membrane-bound organelle	GO:0043231	0.042909	89	6778	89	10523
	intracellular part	GO:0044424	0.049628	105	8388	73	8913
P. patens **alternative NAGNAG genes Sanger and/or 454 support**	cytoplasmic part	GO:0044444	3.52E-04	168	5317	232	11762
	plastid	GO:0009536	8.35E-04	90	2559	310	14520
	membrane-bound organelle	GO:0043227	8.35E-04	196	6671	204	10408
	intracellular membrane-bound organelle	GO:0043231	8.35E-04	196	6671	204	10408
	intracellular part	GO:0044424	0.003794	229	8264	171	8815
	intracellular	GO:0005622	8.35E-04	251	9029	149	8050
	cytoplasm	GO:0005737	3.78E-04	191	6312	209	10767
	mitochondrion	GO:0005739	0.041928	53	1553	347	15526
	thylakoid	GO:0009579	0.023551	12	187	388	16892
	organelle	GO:0043226	0.00436	207	7373	193	9706
	intracellular organelle	GO:0043229	0.00436	207	7373	193	9706
A. thaliana **alternative NAGNAG genes combined (Iida et al., 2008 and Schindler et al., 2008)**	plastid	GO:0009536	0.047395	50	3021	192	17594
	membrane-bound organelle	GO:0043227	2.28E-04	113	6959	129	13656
	intracellular membrane-bound organelle	GO:0043231	2.28E-04	113	6959	129	13656
	intracellular	GO:0005622	5.17E-05	136	8421	106	12194
	organelle	GO:0043226	1.73E-04	119	7287	123	13328
	intracellular organelle	GO:0043229	1.73E-04	119	7287	123	13328
	nucleic acid binding	GO:0003676	5.98E-06	71	3230	171	17385
	nucleus	GO:0005634	2.02E-04	51	2342	191	18273
	DNA binding	GO:0003677	2.28E-04	49	2250	193	18365
	intracellular part	GO:0044424	2.82E-04	124	7881	118	12734
	cell part	GO:0044464	0.002703	159	11257	83	9358
	RNA binding	GO:0003723	0.010001	15	506	227	20109
	binding	GO:0005488	0.011737	132	9231	110	11384
	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GO:0006139	0.01314	47	2608	195	18007
	transcription factor activity	GO:0003700	0.01314	33	1646	209	18969
	transcription regulator activity	GO:0030528	0.043626	34	1852	208	18763

Table 1. GO analyses of genes with alternative NAGNAG acceptor site



																			sequence length polymorphism	
gene id	P(EI) NBC	P(E) NBC	P(I) NBC	P(EI) BN	P(E) BN	P(I) BN	Sanger_E	Sanger_I	454_E	454_I	transcript support for EI	GAG	A.t. homolog with AS	fwd sequence 5'--->3'	rev sequence 5'--->3'	start	end	product size	gametophore	protonema
High P(EI) + transcript support																				
Phypa_193990	0.940	0.000	0.060	0.969	0.013	0.018	4	20	4	8	x			TGTTGGGGAAGTTTTGAAGG	CATCCTCGTCTTTGTGTTCG	113	469	356	yes	yes
Phypa_80579	0.912	0.000	0.088	0.896	0.101	0.003	3	2	6	2	x			ACCCTCGGTCTTCTACTCGAC	TGTAGCTCCAGGCTCAGTCTC	860	1164	304	yes	yes
Phypa_106363	0.901	0.006	0.093	0.500	0.025	0.475	9	7	8	8	x			GCGTCTTGTGGCACCTTTAG	GGGTAGGCGCATGTCTTTAC	4	335	331	yes	yes
Low P(EI) + transcript support																				
Phypa_161321	0.181	0.819	0.000	0.196	0.804	0.000	13	27	1	0	x			CGGTGGCTATGTGGTCATC	CTGACGGCATCACACAAGAC	728	1035	307	yes	yes
Phypa_161321_fwd_FAM											x			CGGTGGCTATGTGGTCATC					yes	yes
Phypa_74146	0.177	0.000	0.823	0.063	0.001	0.936	0	4	1	10	x		x	GGATCTCTTCTCTGCGATGC	ACCAGATGAAGAACAAGATTGC	27	332	305	no	no
Phypa_199161	0.015	0.000	0.985	0.122	0.000	0.877	2	16	0	0	x	x		GTGGTACAATTGCCGATTCC	AGAGTCAGCTCATCGCCAAC	447	747	300	no	no
Phypa_228333	0.000	1.000	0.000	0.000	1.000	0.000	97	0	10	2	x	x		ACAATCGGTGCTGAATCTCC	GAGTAAGGATGGCGTTCTCC	223	589	366	no	no
Phypa_228333_fwd_FAM											x	x		ACAATCGGTGCTGAATCTCC					no	no
Low P(EI) + no transcript support																				
Phypa_100961	0.085	0.915	0.000	0.019	0.981	0.001	14	0	13	0				CTGGCTACTTCGGAGGTGAC	ACAGAGCTGAGGTGGTCTGG	278	659	381	no	no
Phypa_117470	0.000	1.000	0.000	0.000	1.000	0.000	12	0	7	0		x		TCAAGCTCTTCGAGGTTTCC	ATGTCGAAACGCTGCATAAC	1415	1718	303	no	no
Phypa_181992	0.000	1.000	0.000	0.000	1.000	0.000	39	0	4	0		x		ATACCCAGATCGGTGTTTCG	GCTGGTACCCTTCTGCAATG	303	628	325	no	no
Phypa_181992	0.001	0.000	0.999	0.000	0.000	1.000	0	11	0	3		x		ACATTTGGACAGGGTTACCG	TGACTTCAACACGTCCTTCG	1342	1651	309	no	no
Phypa_171213	0.000	1.000	0.000	0.000	1.000	0.000	43	0	33	0		x		CTGCTGCCACAGACTTCCTC	TTCTTCACCTTCTTCCTGTCG	34	378	344	no	no
Phypa_145753	0.000	1.000	0.000	0.008	0.991	0.000	64	0	5	0		x		TGAAGCATTGCAAAGAGTGC	CAACACTCCTTGGCTGGAAC	92	429	337	no	no
Phypa_114834	0.190	0.809	0.001	0.032	0.967	0.001	1	0	3	0		x		TGATCCAAGGCTACTGATTGC	CTATGGGCGATCATGTGAAG	636	938	302	no	no
Phypa_121999	0.312	0.687	0.001	0.572	0.425	0.003	2	0	5	0		x		CGACTGAGAACAAATTCGAAAG	GTTGGCTCAGAGGATGGTTC	11	322	311	no	no
Homologous intron AS in *A. thaliana*																				
Phypa_180723	0.001	0.999	0.000	0.001	0.999	0.000	1	0	0	0		x	x	GTTTGCCTCGGAGATGAAAG	AGGCCAACACAGAAGGAGTG	480	821	341	no	no
Phypa_180457	0.001	0.999	0.000	0.001	0.999	0.000	1	0	1	0		x	x	CTTGCCTCTGTGGGAGTGTC	GTGCAGAATCAGCAACATCC	108	453	345	no	no
Phypa_216093	0.005	0.995	0.000	0.029	0.968	0.002	2	0	9	0		x	x	GGGAATTGGTTGATGTGACG	CCTACCACTTCCATCGGTTC	10	333	323	no	no
Phypa_191544	0.365	0.635	0.000	0.734	0.266	0.000	5	0	14	0			x	AGCCAGTCGCTTAGATCTGG	ATTCCCTCCAAATCCTCCAC	257	596	339	no	no

Table 2. Summarized validation results



ARATH		PHYPA		
At5g65010	TCTTG**TAGGAG**GGC	GGCAA**CAGGAG**GGC	Phypa_180723	validated
At5g06600	TTTTG**CAGCAG**CCA	TGTGG**CAGGAG**GAC	Phypa_180457	
At5g06600	TTTTG**CAGCAG**CCA	TGTGG**CAGGAG**GAT	Phypa_216093	
At5g12210	TTTGC**TAGAAG**AAA	ACATT**CAGGAG**GAT	Phypa_191544	
At2g35520	TGATT**GAGCAG**GTT	CTGGG**AAGCAG**GTG	Phypa_74146	
At3g06550	TATGT**TAGTAG**GCA	TAGAG**AAGCAG**GTG	Phypa_226366	not validated
At3g06550	TATGT**TAGTAG**GCA	GGAAT**GAGCAG**GTG	Phypa_65220	

Table 3. NAGNAG motifs occuring at conserved positions

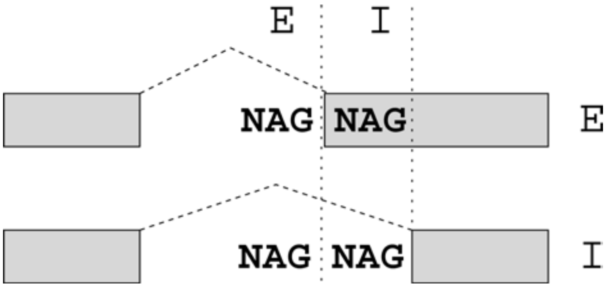


Figure 1. **NAGNAG alternative splicing**

Nomenclature of NAGNAG AS with E and I sites and isoforms.

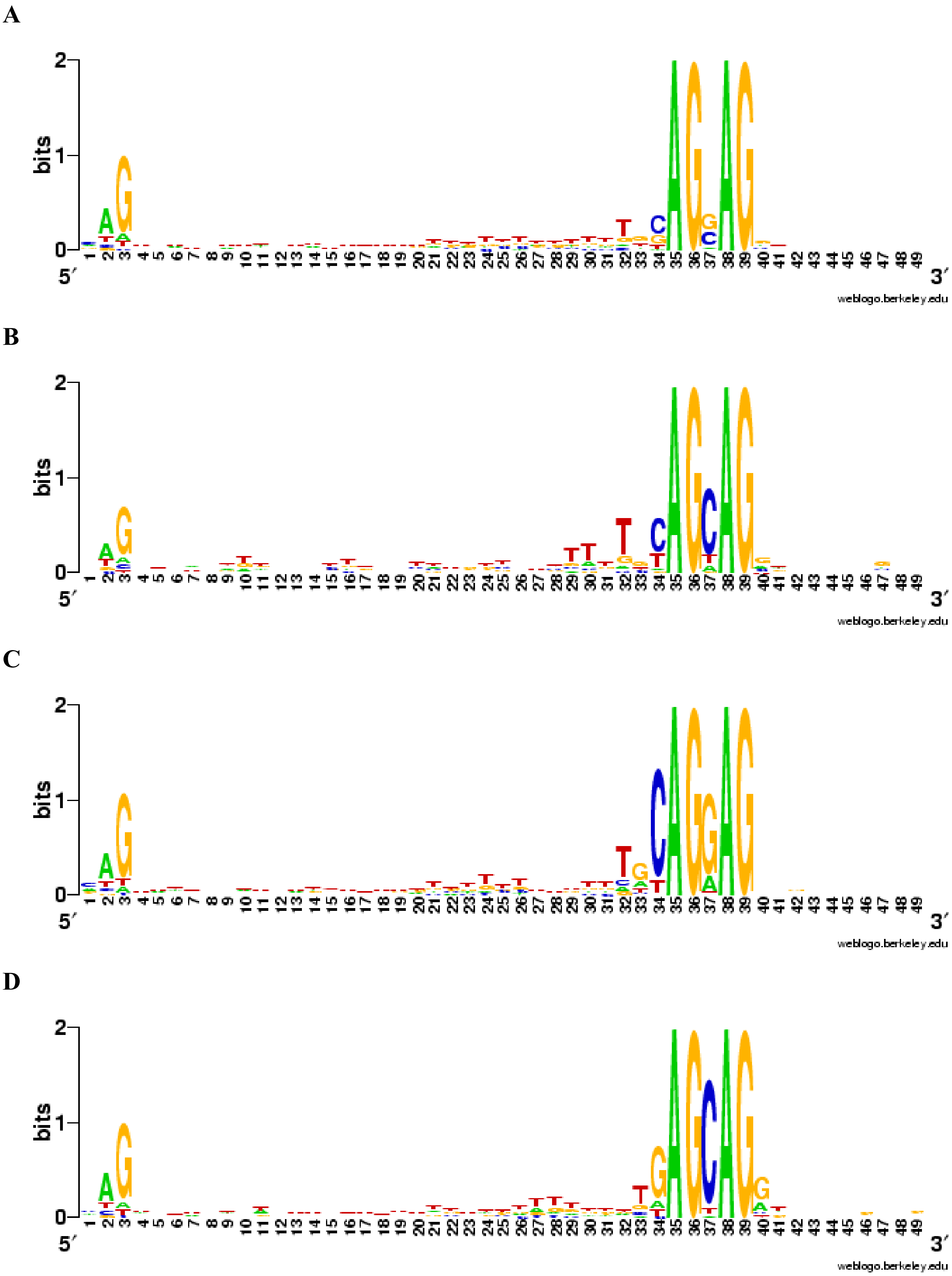


Figure 2. Sequence logos (http://weblogo.berkeley.edu/logo.cgi) of NAGNAG splice sites. The first three positions represent the last 3 nucleotides (nt) of the upstream exon, followed by the 30 nt upstream of the NAGNAG, the NAGNAG motif itself, and the 10 nt downstream of the NAGNAG (total 49 positions). A: all splice sites; B: EI sites, C: E sites; D: I sites.

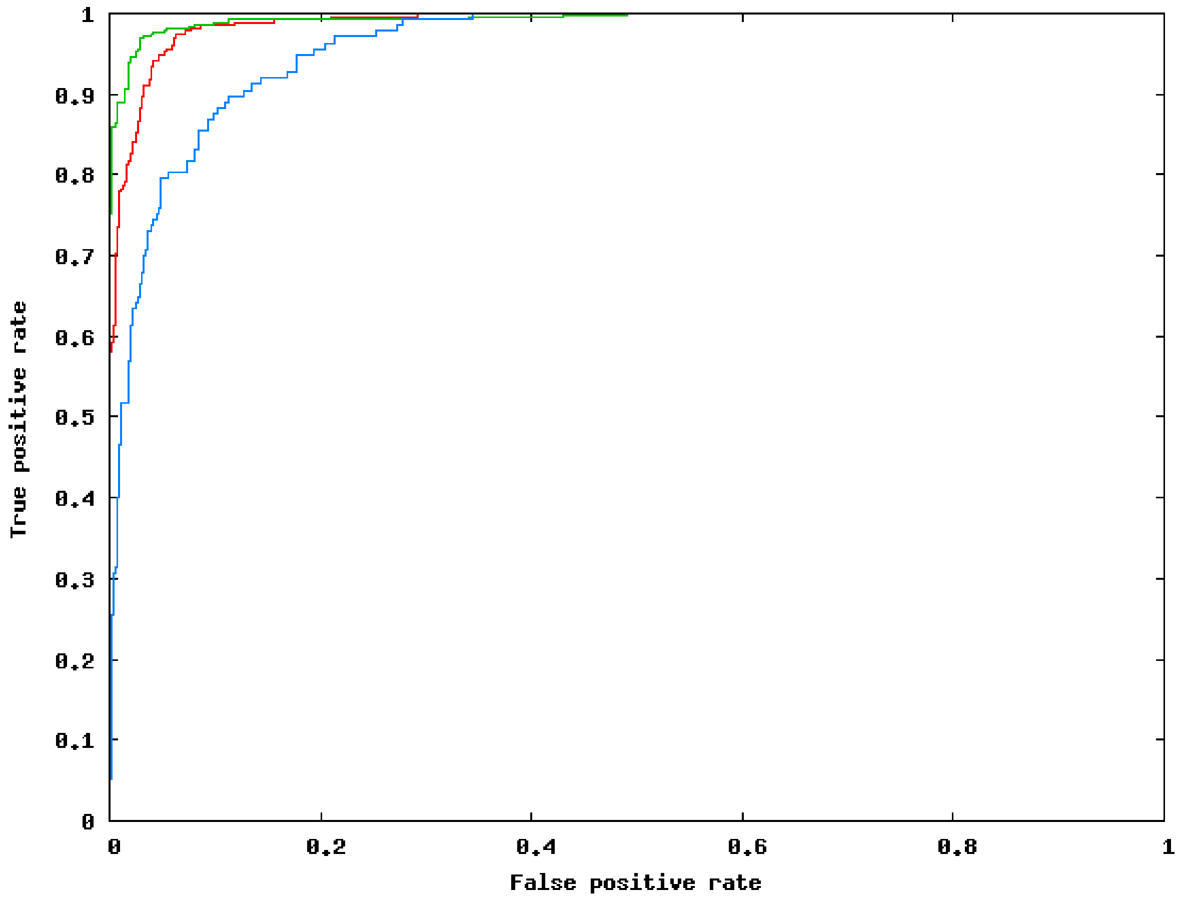


Figure 3. ROC plot

Depicting the *in silico* performance on the 3-class [I-class (red), E-class (green), and EI-class (blue)] classification problem. The EI-class, or AS, harder to predict (AUC = 0.96) than the two constitutive variants, E and I (AUC = 0.99 for both).

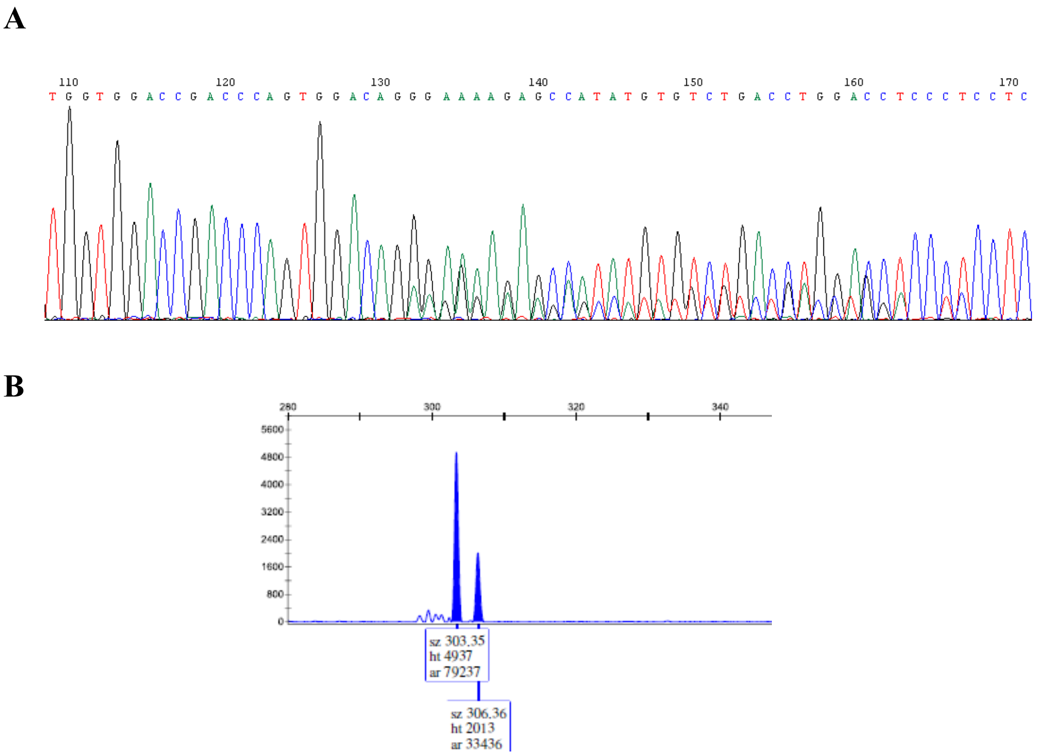


Figure 4. Example of the validation procedures employed.

A: Electropherogram of the direct sequencing of a cDNA PCR product. Starting with the NAGNAG AS site at position 132 the two alternatively spliced sequence signatures are overlaid.

B: FAM fluorescence intensity peaks of the two splice variants (length difference three nucleotides). The lower peak constitutes 40% of the area (ar) of the bigger one, i.e. an approximate ratio of 3:1.

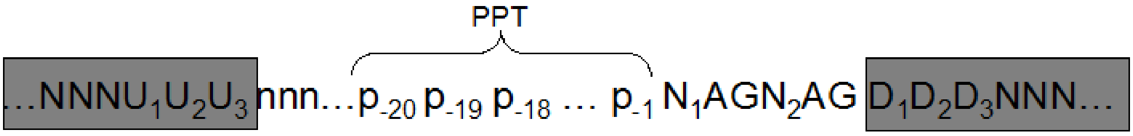

PPT
...NNNU1U2U3 nnn...p-20 p-19 p-18 ... p-1 N1AGN2AG D1D2D3NNN...

Figure 5. Nomenclature of features used in this study

Nomenclature of sequence features used to analyze NAGNAG splicing. The region used to derive all 31 features is shown, along with the names given to the positional features.

Michael Hiller
Karol Szafranksi
Rileen Sinha
Klaus Huse
Swetlana Nikolajewa
Philip Rosenstiel
Stefan Schreiber
Rolf Backofen
Matthias Platzer

Assessing the fraction of short-distance tandem splice sites under purifying selection

RNA 2008, 14(4): 616-629.

Alternative splice involving events involving short-distance tandem splice sites are frequent, but it is unknown how many are functionally important. We used phylogenetic conservation to address this question for tandems with a distance of 3–9 nucleotides. We showed that a statistical paradox (Simpson's paradox) can explain previous contradicting results on whether alternative or constitutive tandem motifs are more conserved between species. Taking biases into account, we found higher conservation of human alternative tandems and their intronic flanks in mouse, dog, and even chicken, zebrafish, and *Fugu* genomes. While the absolute number of conserved tandem motifs decreases with the evolutionary distance, the fraction under selection increases. Interestingly, a number of frameshifting tandems are under selection, suggesting a role in regulating mRNA and protein levels via nonsense-mediated decay (NMD). We propose that stochastic splice site selection, which is likely involved in a majority of tandem splice events, can be an advantageous mechanism, allowing constant splice variant ratios in situations where a deviation in this ratio is deleterious.

Assessing the fraction of short-distance tandem splice sites under purifying selection

Michael Hiller, Karol Szafranski, Rileen Sinha, Klaus Huse, Swetlana Nikolajewa, Philip Rosenstiel, Stefan Schreiber, Rolf Backofen and Matthias Platzer

RNA 2008 14: 616-629; originally published online Feb 11, 2008;
Access the most recent version at doi:10.1261/rna.883908

Assessing the fraction of short-distance tandem splice sites under purifying selection

MICHAEL HILLER,[1] KAROL SZAFRANSKI,[2] RILEEN SINHA,[2] KLAUS HUSE,[2] SWETLANA NIKOLAJEWA,[3] PHILIP ROSENSTIEL,[4] STEFAN SCHREIBER,[4] ROLF BACKOFEN,[1] and MATTHIAS PLATZER[2]

[1]Bioinformatics Group, Albert-Ludwigs-University Freiburg, 79110 Freiburg, Germany
[2]Genome Analysis, Leibniz Institute for Age Research, Fritz Lipmann Institute, 07745 Jena, Germany
[3]Department of Bioinformatics, Friedrich-Schiller-University Jena, 07743 Jena, Germany
[4]Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, 24105 Kiel, Germany

ABSTRACT

Many alternative splice events result in subtle mRNA changes, and most of them occur at short-distance tandem donor and acceptor sites. The splicing mechanism of such tandem sites likely involves the stochastic selection of either splice site. While tandem splice events are frequent, it is unknown how many are functionally important. Here, we use phylogenetic conservation to address this question, focusing on tandems with a distance of 3–9 nucleotides. We show that previous contradicting results on whether alternative or constitutive tandem motifs are more conserved between species can be explained by a statistical paradox (Simpson's paradox). Applying methods that take biases into account, we found higher conservation of alternative tandems in mouse, dog, and even chicken, zebrafish, and *Fugu* genomes. We estimated a lower bound for the number of alternative sites that are under purifying (negative) selection. While the absolute number of conserved tandem motifs decreases with the evolutionary distance, the fraction under selection increases. Interestingly, a number of frameshifting tandems are under selection, suggesting a role in regulating mRNA and protein levels via nonsense-mediated decay (NMD). An analysis of the intronic flanks shows that purifying selection also acts on the intronic sequence. We propose that stochastic splice site selection can be an advantageous mechanism that allows constant splice variant ratios in situations where a deviation in this ratio is deleterious.

Keywords: purifying selection; subtle alternative splicing; tandem splice site; comparative genome analysis; Simpson's paradox

INTRODUCTION

Alternative splicing is a widespread mechanism to produce transcript and protein diversity in animals and plants (Campbell et al. 2006; Tress et al. 2007). Detailed studies revealed many examples where the existence and regulation of alternative splice variants are crucial for cellular functions. For example, alternative splice variants have important roles in the nervous (Ule et al. 2005; Licatalosi and Darnell 2006) and immune (Lynch 2004) systems and during sex determination in *Drosophila* (Black 2003). Moreover, human and mouse splicing factor genes extensively produce nonfunctional splice forms, which provides a potential mechanism for autoregulating the protein level (Stoilov et al. 2004; Lareau et al. 2007; Ni et al. 2007). Misregulation of alternative splicing is a frequent cause of disease (Pagani and Baralle 2004), and the human *SFRS1* gene encoding the splicing factor ASF/SF2 was shown to be a proto-oncogene (Karni et al. 2007).

Despite these facts, the general extent of functional alternative splicing is unknown. Some splice forms such as the skipping of exon 12 of human *CFTR* were described to have no functional advantage (Raponi et al. 2007), and the tissue-specific inclusion of exon 8 of mouse *Psap* shows no phenotypic differences in a knockout mouse lacking this exon (Cohen et al. 2005). Furthermore, about one-third of the human alternative splice events lead to an early stop codon, thus yielding truncated proteins and/or subjecting the mRNA to the nonsense-mediated decay (NMD) pathway (Lewis et al. 2003). Apart from their potential to regulate the protein level by reducing the level of transcripts

Abbreviations: f_s, fraction of confirmed and conserved tandem splice sites estimated to be under purifying selection; CMH test, Cochran–Mantel–Haenszel test.

Reprint requests to: Michael Hiller, Bioinformatics Group, Albert-Ludwigs-University Freiburg, Georges-Koehler-Allee 106, 79110 Freiburg, Germany; e-mail: hiller@informatik.uni-freiburg.de; fax: 49 (761) 203-7462.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.883908.

encoding the full-length protein, their function is often not obvious.

To assess function, one usually considers sequence conservation or the conservation of an event as an important criterion that implies purifying (negative) selection because deviations confer a disadvantage to the organism. Indeed, conserved exon skipping events have a tendency to preserve the protein reading frame (Resch et al. 2004; Sorek et al. 2004; Yeo et al. 2005). These exons and their intronic flanks also exhibit an increased sequence conservation (Sorek and Ast 2003). Furthermore, tissue-specific exon skipping is associated with conserved exons and with reading frame preservation (Xing and Lee 2005). However, genome-wide studies found only a small percentage (~10%–20%) of exon skipping events to be conserved between human and mouse, with most alternative exons being either skipped in only one species or occurring only in one genome (Modrek and Lee 2003; Sorek and Ast 2003; Pan et al. 2004; Yeo et al. 2005). Thus, while alternative splicing is undoubtedly frequent, most of the splice events seem to have no functional role that is conserved in evolution.

Apart from exon skipping, numerous human and mouse alternative splice events occur at alternative donor and acceptor splice sites. The majority of these splice site pairs are in close proximity (Clark and Thanaraj 2002; Zavolan et al. 2003; Sugnet et al. 2004), thus leading to subtle mRNA changes. In this study, we analyze pairs of donor or acceptor sites that are 3–9 nucleotides (nt) apart (Δ3–Δ9 nt) and use the term "tandem sites" to denote these splice site pairs (Fig. 1). The most frequent of these subtle events is alternative splicing at NAGNAG acceptors (Zavolan et al. 2003; Hiller et al. 2004; Sugnet et al. 2004). At the donor site, Δ4 tandem splice sites are most prominent as dictated by the donor consensus sequence (Dou et al. 2006; Ermakova et al. 2007). For most tandem sites, it is likely that their underlying alternative splicing mechanism is based on a stochastic selection of either splice site, also called "noisy splicing" (Chern et al. 2006). A recent study showed that the region between the branch point and the acceptor has a strong influence on the splicing ratio of alternatively spliced NAGNAG sites (Tsai et al. 2007).

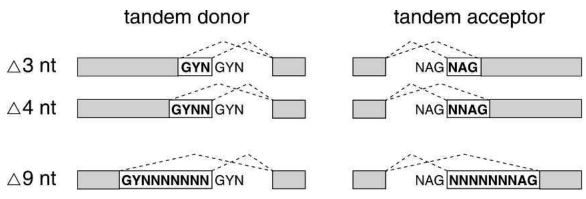


FIGURE 1. Schematic representation of the tandem sites analyzed in this study. (Boxes) Exons; (dashed lines) splice events; (boldface) the variable exonic parts; (GYNGYN and NAGNAG) tandem sites with a distance of 3 nt; (Δ4–Δ9) tandem donors and acceptors that are 4–9 nt apart, respectively; (N) A, C, G, or T; (Y) C or T.

Targeted experimental studies have revealed functional roles for tandem splice events. For example, conserved tandem acceptors in human and mouse transcription factor genes (NAGNAG acceptors in *PAX3* and *PAX7*, Δ6 acceptor in *IRF2*) result in protein isoforms that differ in the ability to activate transcription (Vogan et al. 1996; Koenig Merediz et al. 2000). Conserved Δ6 donors lead to protein variants of human *ALDH18A1* that are insensitive to ornithine inhibition (Hu et al. 1999) and produce protein isoforms of mouse *Fgfr1* that are unable to bind FRS2 and thus cannot activate the Ras/MAPK signaling pathway (Burgar et al. 2002). Furthermore, a splice event at a conserved Δ6 donor in human *EDA* tightly controls binding specificity by remodeling the properties of the receptor binding site, such that the longer protein binds only to the EDAR receptor, while the shorter variant binds only to the XEDAR receptor (Yan et al. 2000; Hymowitz et al. 2003). Another example is the Δ9 donor of human *WT1* exon 9 that leads to the insertion of three amino acids (KTS). Both splice forms have distinct transcriptional regulation properties, hetero- and homozygous mouse mutants lacking one of the two splice forms show severe defects in kidney development and function (Hammes et al. 2001), and a mutation in this donor motif leads to Frasier syndrome in humans (Barbaux et al. 1997).

While these individual studies demonstrate that several of these subtle splice events are functionally important, the general extent remains unknown. Moreover, there is a discussion whether tandem sites that are alternatively spliced are better conserved in evolution than those that are constitutively spliced (Hiller et al. 2006c) since conflicting results were published for NAGNAG acceptors (Hiller et al. 2004; Chern et al. 2006). As alternative and constitutive NAGNAG sites have different preferences for specific NAGNAG motifs (Hiller et al. 2004; Akerman and Mandel-Gutfreund 2006), we considered the possibility that the comparison of two heterogeneous groups caused a statistical paradox, which is often called Simpson's paradox. This paradox is frequently encountered in biomedical studies (Julious and Mullee 1994) and describes a situation in which a trend observed between two groups is reversed when the two groups are split into several subgroups (Simpson 1951). A well-known example of Simpson's paradox is described in Bickel et al. (1975) and refers to university admission data. In this case, the overall admission rates indicated a significant bias against female applications, while investigating all departments individually provided evidence for the opposite—a bias in favor of female applicants. As described in Bickel et al. (1975), the explanation of this apparently paradox is: "The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into."

Here, we show that previous conflicting conclusions for the evolutionary conservation of NAGNAG acceptors (Hiller et al. 2004; Chern et al. 2006) arose from Simpson's paradox caused by substantial conservation differences between specific NAGNAG motifs. Controlling for biases, we found that alternatively spliced NAGNAG acceptors are significantly more conserved than those that are constitutively spliced. We extended the analysis to human tandem donor and acceptor sites that are up to 9 nt apart and estimated a lower bound for the fraction of tandem sites being under purifying selection, and thus expected to have an evolutionarily advantageous phenotype.

RESULTS

Conservation of human NAGNAG acceptors differs between the NAGNAG motifs

First, we analyzed the sequence conservation of human NAGNAG acceptor motifs that are located within the protein coding sequence (CDS). Our data set consists of 1597 confirmed (at least one mRNA/EST indicates splicing at the upstream and at least one mRNA/EST splicing at the downstream acceptor) and 7452 unconfirmed (currently available mRNA/EST data indicate no alternative splicing) NAGNAG acceptors (Hiller et al. 2007). We tested the pairwise conservation of human NAGNAG acceptors over different evolutionary distances: rhesus (∼23 million years ago [mya] since split of the common ancestor), mouse (∼90 mya), dog (∼92 mya), chicken (∼310 mya), and zebrafish and *Fugu* (∼450 mya) (Ureta-Vidal et al. 2003). Although the close distance human–rhesus might limit the power to detect conservation differences, we include rhesus to cover a large spectrum of evolutionary distances.

We used very stringent criteria to define conservation between two NAGNAG tandems to increase the likelihood that an orthologous tandem site, which is considered to be conserved, has the same splicing pattern (alternative or constitutive splicing) in the other species. Previous observations suggest that the tandem splice site motif is the strongest factor determining the splicing pattern (Chern et al. 2006; Hiller et al. 2006a). For this reason, we considered a human NAGNAG as conserved in another species if the orthologous acceptor pattern is identical to the human NAGNAG motif, except for the first N, where we allow variation between C and T. We allow this C/T variation since pyrimidines are the most frequent nucleotides at the −3 position of standard acceptors (Abril et al. 2005) and are not expected to affect the splicing efficiency significantly.

We first performed a global analysis and compared the conservation of all confirmed and all unconfirmed human NAGNAG acceptors in each of the other species. We found that unconfirmed NAGNAG acceptors are more conserved than confirmed ones in the pairwise comparisons (Fig. 2, left parts; Table 1), as previously reported for human and mouse (Chern et al. 2006). The differences are significant in a Fisher's exact test (*P*-values <0.01 for all pairwise comparisons). For this and the following tests, we also computed a standard measurement in biostatistics, the odds ratio (OR). The interpretation of an OR is as follows: an OR > 1 indicates higher conservation for confirmed NAGNAG tandems, an OR < 1 indicates higher conservation for unconfirmed tandems, and an OR = 1 indicates no differences between confirmed and unconfirmed tandems. In the global test, we observed ORs < 1 (Table 1), indicating higher conservation for unconfirmed ones.

Next, we compared the conservation between confirmed and unconfirmed tandems for each of the 16 NAGNAG motifs individually. Strikingly, this motif-specific comparison revealed for 10 of the 16 motifs a higher conservation level for confirmed NAGNAG acceptors in mouse (Fig. 2C). Similarly, confirmed NAGNAG acceptors are more conserved for 10 motifs in rhesus and for 11 in dog and chicken (Fig. 2B,D,E). This apparently contradicts the results of the global analysis. As evident from Figure 2, motifs differ considerably in their overall conservation levels. For example, 51% of all CAGCAG but 70% of all CAGGAG motifs are conserved in mouse.

We hypothesized that these substantial differences in the conservation levels are caused by constraints on the acceptor splice site consensus YAG | G (| indicates the intron–exon boundary; Y = C or T). While a G at the 5′ exon end conforms with the acceptor consensus sequence, a C at this position leads to a weaker acceptor. Thus, CAGCAG acceptors without functional importance are more likely to accept mutations of the unfavored C at position +4 in this motif, while CAGGAG acceptors are less likely to allow mutations of the preferred G at +4. To further test this, we grouped NAGNAG acceptors according to the nucleotide at the second N position and determined the overall conservation. We found that NAGGAG (68.1% conserved) and NAGAAG (62.3%) tandems are generally more conserved than NAGCAG (49.5%) and NAGTAG (56.6%) tandems, in agreement with the preferred 5′-most exon nucleotides, which are G and A followed by C and T (Abril et al. 2005). Moreover, GAG at the 5′ exon end might also be more constrained than CAG, since GAG is more often a core of the splicing enhancer motif identified in Stadler et al. (2006) than CAG (14% versus 12%). Thus, we identified the individual NAGNAG motif as a confounding variable that considerably affects the conservation levels. In such a situation, a global calculation can lead to wrong conclusions.

Higher conservation for confirmed versus unconfirmed human NAGNAG acceptors in rhesus, mouse, dog, and chicken

An unbiased analysis of the conservation level has to correct for the influence of the confounding variable NAGNAG motif. To this end, we used the Cochran–Mantel–Haenszel (CMH) test, which is an extension of the χ^2 test and commonly used in such a situation. The

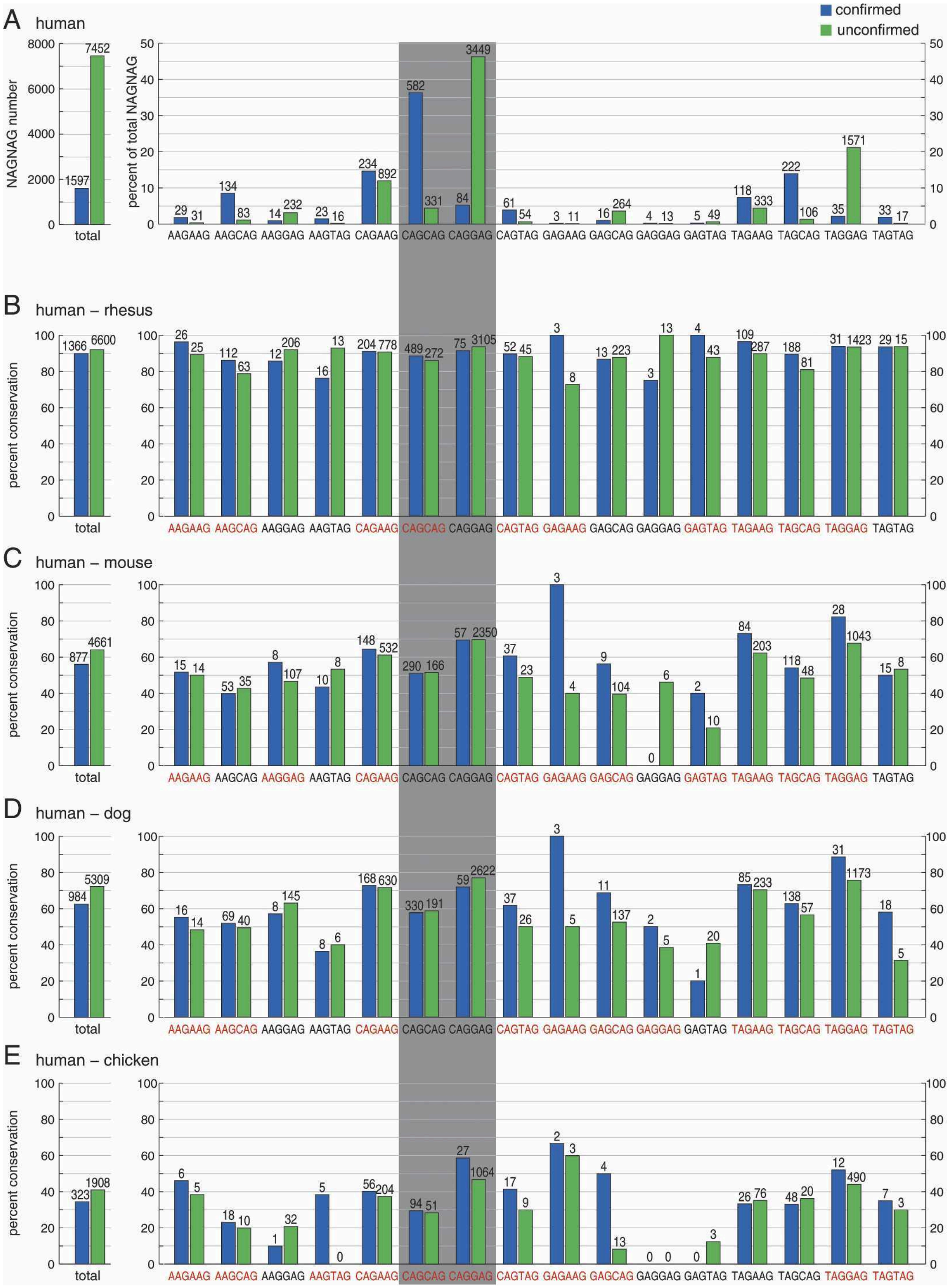


FIGURE 2. Overall and individual conservation of human confirmed and unconfirmed NAGNAG motifs. (*A*, *left* panel) Total number of confirmed and unconfirmed NAGNAG acceptors; (*right* panel) the fraction of individual motifs of the total number of confirmed and unconfirmed human NAGNAG acceptors. The numbers above the bars are absolute numbers of confirmed and unconfirmed NAGNAG acceptors. (*B–E*) The conservation of human NAGNAG acceptor motifs in (*B*) rhesus, (*C*) mouse, (*D*) dog, and (*E*) chicken was analyzed in a (*left* panel) global and (*right* panel) motif-specific comparison. As expected, the overall conservation drops with increased evolutionary distance from rhesus to chicken. A human NAGNAG acceptor is considered to be conserved if it is identical to the orthologous mouse acceptor motif except for an allowed variation between C and T at the first position. (Red) Motifs with a higher conservation for confirmed tandem acceptors. Note that all NAGNAG acceptors for which no pairwise alignment block with the respective species was found were discarded in the conservation analysis. The numbers above the bars are the absolute numbers of conserved NAGNAG sites.

TABLE 1. Pairwise NAGNAG conservation results for a global and a motif-specific analysis of human NAGNAG acceptors in six vertebrate species

	Global conservation			Motif-specific conservation by CMH test		
Species	Confirmed (%)	Unconfirmed (%)	Odds ratio[a]	Odds ratio[b]	Confidence interval[c]	*P*-value[d]
Rhesus	89.8	92.0	0.77	1.29	1.02–1.64	**0.031**
Mouse	56.0	64.1	0.71	1.16	1.00–1.34	**0.047**
Dog	62.6	72.3	0.64	1.10	0.95–1.28	0.205
Chicken	34.4	41.0	0.75	1.18	0.97–1.43	0.097
Zebrafish	17.0	33.4	0.41	0.77	0.59–1.02	0.065
Fugu	17.0	32.3	0.43	0.94	0.71–1.22	0.643

While the global analysis indicates a lower conservation for confirmed NAGNAG tandems (left part), the motif-specific analysis indicates the opposite (right part). *P*-values in bold are significant at the 0.05 level.

[a] An odds ratio (OR) >1 indicates higher conservation for confirmed, <1 higher conservation for unconfirmed NAGNAG tandems; OR is computed as $(n_{cc}/n_{cn})/(n_{uc}/n_{un})$, where n_{cc} = number confirmed and conserved; n_{cn} = confirmed and nonconserved; n_{uc} = unconfirmed and conserved; n_{un} = unconfirmed and nonconserved.

[b] OR computed by the CMH test and corrected for the influence of the NAGNAG motif.

[c] Confidence interval for the OR.

[d] *P*-value (computed by the CMH test) that the OR is unequal to 1.

CMH test estimates an OR that is corrected for the influence of the NAGNAG motif. As shown in the right part of Table 1, using the CMH test, we observed ORs > 1 for rhesus, mouse, dog, and chicken. This indicates a higher conservation for confirmed NAGNAG acceptors. However, in zebrafish and *Fugu*, confirmed NAGNAG tandems have a lower conservation even after correcting for the motif (Table 1), and this holds for the following analyses as well.

The contradictory results of the global and the motif-specific conservation analysis are an example of the above-described Simpson's paradox (Simpson 1951). Here, the paradox occurs since (1) the conservation level (Fig. 2B–E), and (2) the distribution of confirmed and unconfirmed NAGNAG acceptors (Fig. 2A) vary greatly among the different motifs. The most dramatic difference is caused by the weakly conserved CAGCAG motif (that makes up 36% of the confirmed but only 4% of the unconfirmed NAGNAG acceptors) and the strongly conserved motif CAGGAG (that makes up only 5% of the confirmed but 46% of the unconfirmed NAGNAG acceptors), shaded gray in Figure 2. Thus, confirmed NAGNAG acceptors are enriched in weakly conserved motifs, while unconfirmed ones are enriched in highly conserved motifs (analogous to the above example of Simpson's paradox). This bias causes the misleading result of a lower conservation of confirmed versus unconfirmed tandem acceptors in the global analysis. Moreover, this bias explains previous conflicting conclusions because the data set used in Hiller et al. (2004) ("intronic extra AGs") (see Supplementary Note in Hiller et al. 2004) contains virtually none of the strongly conserved NAGGAG motifs, while NAGGAG motifs make up a large fraction of all unconfirmed NAGNAG sites that were analyzed in Chern et al. (2006).

The unequal NAGNAG motif distribution was observed in previous studies (Hiller et al. 2004; Akerman and Mandel-Gutfreund 2006) that showed that >90% of the alternative NAGNAG acceptors have an HAGHAG motif (H = A, C, T), while those tandems having a GAG are rarely alternatively spliced (Hiller et al. 2006b). Furthermore, standard acceptors are mostly CAG or TAG, with AAG and especially GAG being rare. This reflects the binding affinity of the U2AF35 splicing factor (Wu et al. 1999). Thus, the splicing machinery may select either acceptor in an HAGHAG motif, resulting in alternative splicing (Chern et al. 2006).

Estimating the number of human NAGNAG acceptors that are under purifying selection

Higher conservation of confirmed NAGNAG tandems indicates that a certain fraction is under purifying selection, which prevents the destruction of the NAGNAG motif in the course of evolution. Since the CMH test does not estimate how many confirmed tandem acceptors are under selection, we developed two simulations to answer this question. We used unconfirmed NAGNAG tandems to estimate the expected or background conservation that reflects evolutionary constraints to preserve a functional acceptor and the coding sequence that overlaps the NAGNAG motif. The number of confirmed and conserved tandem acceptors that exceed the expected conserved number is considered to be subject to purifying selection, which preserves the alternative splice event. In the following, we use f_s for the fraction of confirmed and conserved tandem splice sites estimated to be under purifying selection.

Applying the first simulation (called the "balanced motif distribution"; see Materials and Methods) to the rhesus, mouse, dog, and chicken conservation data, we estimate that between 2.95% (rhesus) and 9.55% (chicken) of the confirmed and conserved NAGNAG acceptors are under purifying selection (Table 2). Furthermore, the *P*-values are significant at the 0.05 level for all four comparisons. To further support this estimation, we applied another simulation (called the "balanced OR"; see Materials and Methods) and found highly similar results (Table 2).

To extend the pairwise approach, we considered as a four-way conserved NAGNAG acceptor a human tandem that is conserved in rhesus, mouse, dog, and chicken (237 confirmed and 1360 unconfirmed four-way conserved sites) (Supplemental Table 1). NAGNAG sites that lack conservation in one or more species are considered as nonconserved in this test (1351 confirmed and 6101 unconfirmed sites). We found that confirmed NAGNAG acceptors have a significantly higher four-way conservation (CMH test: OR = 1.28, $P = 0.014$) than unconfirmed ones. The balanced motif distribution simulation estimates an f_s of 20.3% (OR = 1.31, $P < 0.0001$), indicating that 48 of the 237 four-way conserved tandems are under selection. Thus, four-way conserved NAGNAG acceptors have a stronger tendency to be under purifying selection.

As pointed out above, CAGCAG is the motif with the highest number of confirmed human tandem acceptors. Confirmed CAGCAG acceptors show a slightly higher conservation than unconfirmed CAGCAG sites in rhesus and chicken but not in mouse and dog (Fig. 2B–E). To further investigate the conservation of this motif, we considered four-way conserved CAGCAG sites and found that human confirmed CAGCAG acceptors have a 3% higher four-way conservation level than unconfirmed ones, suggesting that 17 CAGCAG sites are under selection.

Finally, we analyzed conservation of NAGNAG tandems located in the untranslated region (UTR). In contrast to NAGNAG tandems in the CDS, we found no indication that UTR tandems are under selection (data not shown).

Conservation of human NAGNAG alternative splicing in mouse

Confirmed NAGNAG acceptor motifs are likely to be under purifying selection because the alternative splice event provides an advantageous phenotype. Therefore, we considered conservation of the alternative splice event in mouse. Of the human confirmed NAGNAG acceptors that are conserved in mouse, we found that 59% of the orthologous mouse NAGNAG acceptors are alternatively spliced in mouse. This shows that conservation of the NAGNAG motif is associated with conservation of the splice event. In particular, confirmed NAGNAG sites that have no GAG acceptor have a high chance to be confirmed in mouse (Fig. 3), presumably because their splice variant ratio is often rather balanced so that few ESTs can be sufficient to detect alternative splicing in mouse. As the mouse transcript coverage is only 62% of the human coverage (~5 million mouse ESTs and mRNAs versus ~8 million for human), our finding that 59% of the alternative splice events are conserved is a lower bound.

Human tandem donors and acceptors with up to Δ9 nt under purifying selection

Next, we extended our conservation analysis to human tandem donors with Δ3–Δ9 nt and tandem acceptors with Δ4–Δ9 nt (Fig. 1) that are located within the CDS. As for NAGNAG acceptors, we found that constraints on the donor and acceptor consensus are one reason for the different conservation levels of individual tandem motifs (Materials and Methods). Furthermore, the motif distribution differs between confirmed and unconfirmed tandems, probably because some tandem motifs allow selection of either splice site by the spliceosome, while in other tandems the stronger splice site is used exclusively (Chern et al. 2006). To exclude potential biases, we used the balanced motif distribution simulation to assess f_s in the following.

In contrast to NAGNAG acceptors, confirmed GYNGYN donors (Hiller et al. 2006b) are not conserved significantly more than unconfirmed ones. Only the mouse and chicken comparisons indicate that a few confirmed GYNGYN tandems might be under selection (Fig. 4A, left). However, conserved tandem donors with larger splice site distances contain more sites under purifying selection.

TABLE 2. Pairwise estimation of f_s, the fraction of confirmed and conserved human NAGNAG acceptors under purifying selection

		Balanced motif distribution simulation				Balanced OR simulation	
Human versus	Number of confirmed and conserved tandems	Average OR[a]	*P*-value	f_s (%)	Number of tandems under selection	f_s (%)	Number of tandems under selection
Rhesus	1366	1.32	0.001	2.95	40	2.56	35
Mouse	877	1.14	0.022	5.47	48	5.82	51
Dog	984	1.11	0.027	3.63	36	3.25	32
Chicken	323	1.17	0.036	9.55	31	9.60	31

Note that the average ORs of the balanced motif distribution simulation are in good agreement with the estimations from the CMH test (see Table 1).

[a] Average of 1000 iterations.

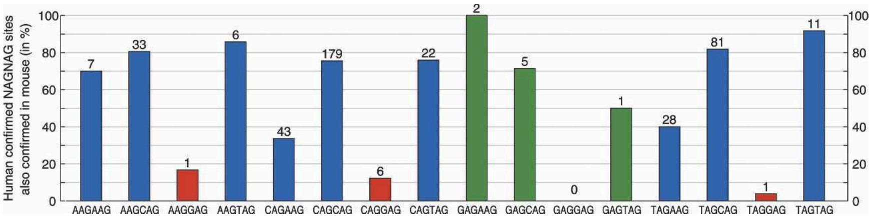


FIGURE 3. Conservation of alternative NAGNAG splice events in human and mouse. Each bar is the percentage of human confirmed NAGNAG acceptors that is also confirmed in mouse, split into the 16 NAGNAG patterns. Absolute numbers are given above the bars. Only those human NAGNAG sites that are conserved in mouse are considered. (Blue) Tandem acceptors with the pattern HAGHAG (H = A, C, T); (red) acceptors with the pattern HAGGAG; (green) acceptors with the pattern GAGHAG. Note that there is no human confirmed GAGGAG acceptor that is conserved, hence none can be confirmed in mouse.

We observed that f_s increases with the evolutionary distance and that frame-preserving donor sites are preferentially under selection for large evolutionary distances (Fig. 4A, right). Strikingly, for the human–*Fugu* comparison, f_s increases to 0.75 for Δ6 donors, indicating that three-quarters of the confirmed tandem donors that are conserved over ~450 mya are under purifying selection. Indeed, these cases include the functionally important tandem donors in human *EDA* (Yan et al. 2000) and *ALDH18A1* (Hu et al. 1999). Apart from frame-preserving sites, frameshifting tandem donors contain sites under selection, even in the human–fish comparison.

While the f_s estimations for individual Δ3–Δ9 acceptors are often lower than the respective donor classes and rarely significant, the absolute number under selection is mostly higher due to a larger number of confirmed tandem acceptors (Fig. 4A, left). In particular, NAGNAG sites contribute the biggest portion. As for tandem donors, f_s increases for larger evolutionary distances. In general, Δ3–Δ6 tandems contain more sites under selection than Δ7–Δ9 tandems. The human–rhesus comparison reveals selection for only a few tandem site classes, which is presumably due to the close evolutionary distance that leads to a high background conservation rate.

Assessing purifying selection for mouse tandem splice sites

Up to now, we have assessed f_s for human confirmed tandem sites by pairwise comparison with other species. Apart from human, only the mouse genome has a transcript coverage (~5 million ESTs) that allows us to create sufficiently large sets of confirmed tandem sites. In contrast to mouse, many human ESTs are sampled from tumor tissue, and this might affect the above conclusions. To provide an independent estimation, we used the balanced motif distribution simulation to estimate f_s for confirmed mouse tandem sites by analyzing the conservation in human. Noteworthy, a high fraction (70%) of the mouse confirmed and conserved NAGNAG sites is also alternatively spliced in human.

Consistently, the estimated number of mouse confirmed tandem sites under purifying selection is similar to the estimations for human confirmed sites (Fig. 4, cf. B and A). The mouse-based analysis estimates an even higher number of Δ4 donors, NAGNAG, and Δ9 acceptors to be under selection. It should be noted that mouse confirmed CAGCAG sites have a 3% higher conservation level than unconfirmed ones, suggesting that 13 confirmed mouse CAGCAG sites are under selection. This is in agreement with the estimation for four-way conserved human CAGCAG sites (see above).

Conservation of the intronic flanks

To provide further support that conservation of the tandem motif implicates conservation of the splicing pattern (alternative or constitutive splicing), we determined the conservation of the intronic flanking regions. Previous studies showed that exons, which are alternatively spliced in human and mouse, are flanked by highly conserved intronic regions (Sorek and Ast 2003; Yeo et al. 2005), and the same was observed for human and mouse confirmed GYNGYN and NAGNAG tandems (Akerman and Mandel-Gutfreund 2006; Hiller et al. 2006b). Thus, high intronic flank conservation is a hallmark of conserved alternative splice events. To abstract from pairwise conservation (often human–mouse), we used the PhastCons conservation scores, which are based on multiple genome sequences and a given phylogeny (Siepel et al. 2005).

Analyzing the average per-position conservation score for the 30-nt intronic flank of all tandems with Δ3–Δ9 nt, we found that confirmed and mouse-conserved human tandems have generally the highest intronic flank conservation, indicating that purifying selection also acts on the intronic context. In particular, Δ4, Δ6, Δ7, and Δ9 donors and Δ3 and Δ5 acceptors have significantly higher intron conservation (Supplemental Figs. 1,2), and this coincides with the tandem classes that have a significant fraction under purifying selection (Fig. 4A, mouse). These observations indicate that confirmed and conserved human tandem sites are associated with alternative splice events in other species.

DISCUSSION

Given the abundance of alternative splicing at tandem sites, it is of interest to find out what fraction of these events is biologically meaningful. Apart from experimental investigations (Condorelli et al. 1994; Vogan et al. 1996;

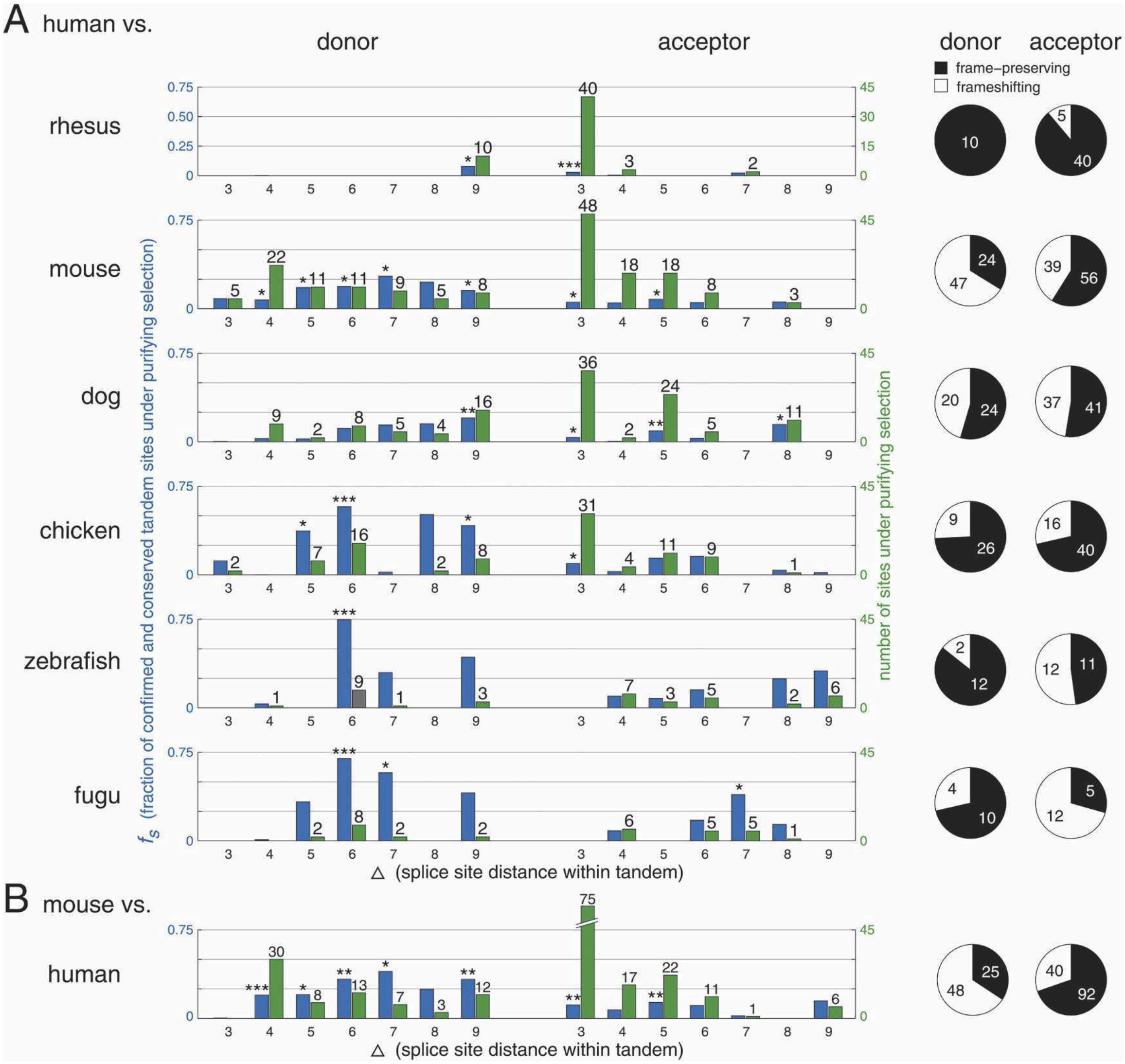


FIGURE 4. Tandem donor and acceptor sites with Δ3–Δ9 nt under purifying selection. (*A*) Analyzing the conservation of confirmed human tandem sites in six vertebrate species. (*Left* chart, blue bars) f_S; (green bars) the number of confirmed and conserved tandems under purifying selection (numbers >0 are given above the bars). (*Right* pie charts) The fraction of all frame-preserving and all frameshifting tandems that are under selection. (*B*) Analyzing the conservation of confirmed mouse tandem sites in human. *P*-values are determined by repeatedly testing the null hypothesis that confirmed tandems are conserved according to the motif-specific background conservation level (Materials and Methods). Significance is indicated as (***) $P < 0.01$; (**) $P < 0.01$; (*) $P < 0.05$.

Hu et al. 1999; Koenig Merediz et al. 2000; Yan et al. 2000; Joyce-Brady et al. 2001; Burgar et al. 2002; Tadokoro et al. 2005), another approach to address this question is to estimate the fraction of tandem sites under purifying selection. Here, we show that the sequence conservation level differs between tandem motifs due to constraints on the splice site consensus and possibly on splicing enhancer motifs as well as on the coding sequence. Together with differences in the tandem motif distribution, this bias seriously affects the conclusion whether confirmed tandems are more conserved than unconfirmed ones. Applying methods that control for this bias, we estimate the fraction of tandem sites under purifying selection.

Interestingly, we found that frame-preserving and frameshifting tandems are under selection. Frameshifting tandem splice events can have a functional role by creating truncated proteins as exemplified for a Δ4 acceptor in the last intron of the zebrafish *pou5f1* gene (Takeda et al. 1994) or by subjecting the mRNA to the NMD pathway. In agreement with this, at least 21% of the human–mouse conserved exon skipping events lead to an NMD-inducing transcript, suggesting a potential role in regulating the protein level (Baek and Green 2005). Furthermore, NMD-inducing exon skipping and intron retention events in splicing factor genes are likely to be important because these alternative regions overlap highly or even ultraconserved elements (Lareau et al. 2007; Ni et al. 2007). It is noteworthy that experimental studies also revealed functional differences for tandem sites that lack deep evolutionary conservation. For example, the CAGCAG acceptor of human *IGF1R* exon 14, which leads to changes in the signaling activity and the internalization rate of the receptor (Condorelli et al. 1994), is not conserved in mouse, rat, dog, and chicken. Thus, similarly to the predicted functional roles of species-specific alternative splice events occurring at conserved exons (Pan et al. 2005), species- or lineage-specific alternative splice events at tandem sites may have functional consequences.

It is important to note that our estimation of the fraction of confirmed tandem sites under selection (Fig. 4) is a lower bound. A major reason is that our set of unconfirmed tandems is likely to be contaminated with sites that are alternatively spliced but currently lack transcript confirmation. To provide a rough estimate of how many unconfirmed NAGNAG acceptors might be alternatively spliced, we determined how many of those have a local context of 3 nt upstream and downstream ($N_3NAGNAGN_3$) that is identical to a confirmed NAGNAG. As the local sequence context primarily determines if a NAGNAG is alternatively spliced (Chern et al. 2006), these unconfirmed NAGNAG sites are expected to allow alternative splicing. We found that 10.5% of the unconfirmed human NAGNAG acceptors have a local context identical to a confirmed tandem. Remarkably, this fraction increases to 26% for those unconfirmed human NAGNAG acceptors with a C or T at the N-positions (YAGYAG), and these unconfirmed sites have a fivefold lower EST coverage than the confirmed ones. Requiring the identity of only 2 nt upstream and downstream ($N_2NAGNAGN_2$), 72% of the unconfirmed YAGYAG sites have a confirmed counterpart. Thus, a substantial fraction of the unconfirmed NAGNAG acceptors is likely to be alternatively spliced, although this is not indicated by current transcript data. Therefore, the background conservation level computed from unconfirmed tandems is likely to be overestimated, and consequently the real fraction under selection is underestimated. In particular, frameshifting tandem splice sites are expected to contain many unconfirmed but alternatively spliced cases, since NMD removes the alternative transcripts (Baek and Green 2005; Chern et al. 2006; Lareau et al. 2007; Ni et al. 2007). If the down-regulation of the mRNA encoding the full-length protein has functional relevance, unconfirmed but alternatively spliced tandem sites are probably conserved in evolution, which, in turn, leads to an overestimated background conservation level.

Two confirmed NAGNAG acceptors are located in ultraconserved elements (defined as at least 200-nt-long regions that are identical between human, mouse, and rat) (Bejerano et al. 2004). The first is the CAGCAG in *PAX2* exon 2, which leads to a ProGly-to-Arg exchange immediately upstream of the Paired box domain. Interestingly, NAGNAG splice events within the Paired box domain in *PAX3* and *PAX7* affect DNA binding (Vogan et al. 1996). The second case is a CAGAAG in *CLK4* exon 4 that leads to the insertion/deletion of a Lys upstream of the protein kinase domain. These two NAGNAG acceptors are also identical between human and chicken. Both ultraconserved elements overlap a large region of the intron–exon boundary; thus it is unknown if purifying selection on the NAGNAG acceptor and its context was the driving force for these ultraconserved elements.

Although tissue- or cell-line-specific splicing has been observed at tandem acceptors (Koenig Merediz et al. 2000; Hiller et al. 2004; Xu et al. 2004; Tadokoro et al. 2005) and tandem donors (Hu et al. 1999; Yan et al. 2000), stochastic selection of either of the two splice sites likely explains alternative splicing at most tandems (Chern et al. 2006). Stochastic splice events are expected to yield similar splice variant ratios in different tissues, and this was observed in many cases (Vogan et al. 1996; Hammes et al. 2001; Burgar et al. 2002; Tadokoro et al. 2005; Hiller et al. 2006b). Noteworthy, stochastic splicing does not preclude functional importance of the alternative splice event (Hiller et al. 2006c). Especially in a situation where both protein isoforms are required ubiquitously, stochastic splice site selection based only on spliceosomal core components offers the advantage of producing the two variants nearly independent of other conditions that regulate alternative splicing. This is likely to be the case for the functionally relevant tandem sites in mouse *Fgfr1* and human *PAX3* and *PAX7* (see Introduction) that produce a constant ratio of the two splice variants (Vogan et al. 1996; Burgar et al. 2002). Another striking example is the Δ9 donor of human *WT1* (see Introduction). This tandem donor site as well as its flanking regions is perfectly conserved between vertebrates, and the two splice variants have distinct functional roles. The splice variant ratio is constant in human tissues and cell lines (Barbaux et al. 1997; Davies et al. 2000) as well as in mouse (Hammes et al. 2001) and in zebrafish (C. Englert, pers. comm.). A deviation in this ratio is highly deleterious and leads to pronounced phenotypes (Hammes et al. 2001). In this case, stochastic donor selection by the ubiquitously expressed U1 snRNP would be a probable mechanistic basis of the constant ratio. Similar to NAGNAG acceptors (Tsai et al. 2007), sequences in the intronic flank might be important for the ratio of the two donor sites, which would explain the high intronic conservation. Apart from tandem sites, a stochastic mechanism that controls splicing of 48 mutually exclusive exons in *Drosophila DSCAM* is essential for axon guidance and is conserved over 300 mya in the insect lineage (Graveley 2005).

While we provided quantitative evidence that a fraction of tandem sites is under purifying selection and thus functional, their identity remains unknown. We found that NAGNAG acceptors with a strong minor splice site are more conserved than those with a weak one, suggesting that the frequency of the alternative splice event might be important. Furthermore, deep conservation in several species such as four-way conserved tandems (Supplemental Table 1), conservation over large evolutionary distances (Supplemental Table 2), and high intronic flank conservation (Supplemental Figs. 1,2) might be reasonable criteria to select promising candidates for further experimental studies.

MATERIALS AND METHODS

Data sets

We downloaded from the UCSC Genome Browser (Kuhn et al. 2007) the human genome assembly (hg17, May 2004) as well as RefSeq annotation (refFlat.txt.gz, November 2006). We screened all splice sites for the presence of a tandem donor and acceptor Δ3–Δ9 motif. Donor sites without GT or GC and acceptors without AG intron termini were omitted. The RefSeq annotation of the open reading frame was used to decide if a tandem site affects the CDS. A tandem site was considered as confirmed if there is at least one EST/mRNA that matches the short and at least one EST/mRNA that matches the long transcript. For NAGNAG and GYNGYN tandems, we downloaded EST information from TassDB (Hiller et al. 2007). For Δ4–Δ9 tandem sites, we used BLAST against all ESTs and mRNAs to obtain confirmation for the putative alternative splice event. BLAST was done as described in Hiller et al. (2006b). The total size of the obtained confirmed and unconfirmed data sets is as follows: GYNGYN: 116 confirmed and 8031 unconfirmed; Δ4 donors: 595 and 97,539; Δ5: 161 and 27,254; Δ6: 161 and 40,262; Δ7: 89 and 33,329; Δ8: 63 and 31,501; Δ9: 160 and 34,793; NAGNAG acceptors: 1597 confirmed and 7452 unconfirmed; Δ4 acceptors: 603 and 8093; Δ5: 364 and 7912; Δ6: 266 and 11,754; Δ7: 118 and 12,917; Δ8: 100 and 11,338; Δ9: 156 and 14,040.

Conservation was detected by analyzing the genome-wide pairwise alignments downloaded from the UCSC Genome Browser (assemblies: human hg17, rhesus rheMac2, mouse mm7, dog canFam2, chicken galGal2, zebrafish danRer3, fugu fr1) using the genomic locus of the human tandem sites to select the respective alignment chain. This approach allows a highly accurate detection of true orthologous splice sites, since the alignment of the individual exons and their splice sites is embedded in the syntenic context of the UCSC whole-genome alignment. Furthermore, coding exons are a class of functional elements that can be reliably aligned between distant genomes (Thomas et al. 2003). Tandem sites, for which no alignment chain was found, were excluded from the pairwise analysis as it is not clear if the entire exon is missing in the other species, if the tandem site is contained in two different alignment chains, or if these cases are due to wrong alignments. It should be noted that considering these tandem sites as nonconserved leads to an even higher conservation difference in favor of confirmed sites.

PhastCons scores for alignments of 16 vertebrate genomes with the human hg17 assembly (phastCons17way) were downloaded from the UCSC Genome Browser.

Statistics

The odds ratio is defined as $(n_{cc}/n_{cn})/(n_{uc}/n_{un})$, where n_{cc} is the number of confirmed and conserved tandems, n_{cn} is the number of confirmed and nonconserved tandems, n_{uc} is the number of unconfirmed and conserved tandems, and n_{un} is the number of unconfirmed and nonconserved tandems. Statistical tests (Fisher's exact test, CMH test, Wilcoxon rank-sum test) were performed using the R software (http://www.r-project.org/).

Different filtering and conservation criteria for NAGNAG acceptors

Given two orthologous NAGNAG acceptor motifs, we define "conservation" as an identical NAGNAG motif allowing only a variation between C and T at the first position. We tested if the conservation results for NAGNAG tandems were affected by this definition of conservation. Higher conservation for confirmed NAGNAG acceptors was consistently found if we (1) consider NAGNAG conservation as the conservation of both AGs allowing both Ns to vary (CMH test OR: 1.26 for rhesus, 1.15 for mouse, 1.06 for dog, 1.16 for chicken) (Supplemental Fig. 3); (2) consider NAGNAG conservation as the identity of the entire hexamer; i.e., conservation of both AGs and both Ns (CMH test OR: 1.11 for rhesus, 1.1 for mouse, 1.01 for dog, 1.18 for chicken).

Higher conservation for confirmed NAGNAG tandems was also observed if we (1) exclude unconfirmed NAGNAG tandems from the analysis that have only single EST support and hence cannot be confirmed (CMH test OR: 1.24 for rhesus, 1.15 for mouse, 1.1 for dog, 1.18 for chicken); (2) exclude confirmed NAGNAG tandems where the minor acceptor is supported by only a single EST (CMH test OR: 1.35 for rhesus, 1.11 for mouse, 1.07 for dog, 1.25 for chicken). As confirmed NAGNAG acceptors have an approximately twofold higher EST coverage, we tested if the overall EST coverage affects our results. Splitting all confirmed and unconfirmed NAGNAG tandems into those with at most 10 and at least 10 ESTs, we found a higher conservation for confirmed NAGNAG sites in both groups except for dog (CMH test OR: 1.64 for at most 10 ESTs and 1.05 for at least 10 ESTs for rhesus, 1.23 and 1.04 for mouse, 1.08 and 0.97 for dog, 1.10 and 1.12 for chicken).

We also found higher conservation for confirmed NAGNAG acceptors, when we restrict the analysis only to those tandems that contain no GAG site (CMH test OR: 1.39 for rhesus, 1.12 for mouse, 1.1 for dog, 1.1 for chicken). Consistently, restricting the analysis only to those NAGNAG sites that have a C or T at both N positions, we also found higher conservation for confirmed ones (CMH test OR: 1.39 for rhesus, 1.07 for mouse, 1.11 for dog, 1.07 for chicken).

Balanced motif distribution simulation for NAGNAG acceptors

The basic idea for the balanced motif distribution simulation is that Simpson's paradox cannot occur if the distribution of the 16 motifs is equal between confirmed and unconfirmed NAGNAG tandems. To correct the unequal motif distribution, we did the following. For each NAGNAG motif, we constructed two lists containing the confirmed and unconfirmed tandems. From the list with the higher entry number, we randomly removed entries so that the entry number in this list equals the number in the other list. This procedure was repeated for all splice site motifs. Then, we combined all confirmed and unconfirmed lists, counted the total number of conserved confirmed and unconfirmed tandems, and determined the OR. Note that after correcting the unequal motif distribution (Supplemental Fig. 4), a global analysis provides a fair comparison; thus we can directly determine how many confirmed tandem acceptors are under purifying selection. The whole procedure was repeated 1000 times. Finally, we computed the average of the 1000 odds ratios and the difference between the average number of conserved confirmed and conserved unconfirmed tandems. This difference divided by the number of conserved and confirmed tandems is f_s. We also tested bootstrapping (allowing a single tandem site to be selected more than once in one iteration) and found virtually identical results (data not shown).

Additionally, we computed a *P*-value by repeatedly testing the null hypothesis that confirmed tandems are conserved according to the motif-specific background conservation level. To this end, we used the motif-specific percentage p of conserved unconfirmed NAGNAG acceptors as the background conservation level. Let n be the number of confirmed NAGNAG acceptors with a given motif. Then, we generated n random numbers and counted as m the number of cases which are $\leq p$. The interval $[0–p]$ represents the conserved part of the background, and the interval $(p–1]$ represents the nonconserved part. For example, the background conservation level in mouse for AAGAAG is 50% (Fig. 2C). Since there are 29 confirmed AAGAAG acceptors in our data set, we generated 29 random numbers and counted how many of those are ≤ 0.5. We repeated that for all motifs and determined the total sum of motif-specific m. The *P*-value is the fraction of 10,000 performed iterations where this sum is equal to or higher than the actual number of confirmed and conserved tandems. This *P*-value is independent of the CMH test.

Balanced OR simulation

The rationale for the balanced OR simulation is that the CMH test should estimate an OR of 1 if there is no difference in the conservation. Thus, we determined which fraction of the confirmed and conserved tandems has to be artificially considered as nonconserved to get an OR of 1; this fraction is the estimation for f_s. Specifically, for a given fraction f, we changed the conservation status of $f \cdot n$ randomly selected confirmed NAGNAG acceptors from conserved to nonconserved, where n is the total number of confirmed and conserved tandems. Then, we computed the OR using the CMH test. For a given f, this was repeated 1000 times, and we determined the average OR and the standard deviation. If $f = f_s$, we expect that the OR = 1. Starting from $f = 0$, we increased f to obtain average ORs well below 1. The highest f for which the average OR is still >1 is taken as an estimate of f_s (Supplemental Fig. 5).

The balanced OR simulation was only performed for NAGNAG acceptors as the number of motifs increases for GYNGYN and Δ4–Δ9 tandem sites, while the number of confirmed sites decreases. In a situation with many motifs mostly having a low case number, the CMH test cannot reliably estimate the OR.

Definition of conservation of two tandem sites

With increasing distance between the two acceptors of a confirmed tandem, the sequence between the two AGs has a tendency to contain pyrimidines (Dou et al. 2006), probably reflecting the requirement for a second polypyrimidine tract. Furthermore, the nucleotide downstream from the AGs, which is frequently a G for confirmed tandems, influences the splicing pattern (Dou et al. 2006). To account for these observations, we required identity of the +4 position (in the following, numbering starts at the first position in a Δ3–Δ9 acceptor or donor motif). All other positions between the first four and last three positions were required to be either a pyrimidine or a purine for Δ5–Δ9 motifs (for example, a CAGG<u>C</u>CAG is conserved to a TAGG<u>T</u>CAG but not to a TAGG<u>A</u>CAG). To fulfill these constraints, two tandem acceptors have to be highly similar; indeed, tandem acceptors are often identical between species as the part downstream from the first AG overlaps with protein-coding sequence.

Previously, we found that all GYNGYN donors that are confirmed in human and mouse are identical between both species and that the GTAGTT donor of *STAT3* exon 21 even yields virtually identical splice variant ratios in human and mouse (Hiller et al. 2006b). Therefore, we required identity of the first and last three positions for Δ3–Δ9 donors. Analyzing the nucleotide preferences for the positions between the two GYNs, we found a preference for a purine at position +4 for Δ4–Δ9 donors, at +5 for Δ5–Δ9 donors, and at the position upstream of the second GYN for Δ6–Δ9 donors, which is in agreement with the general donor consensus. To account for this, we required either a purine or an identical nucleotide at these three positions.

Balanced motif distribution simulation for Δ4–Δ9 tandem sites

For tandem sites that are more than 6 nt apart, each motif basically becomes unique; thus it is no longer practical to compare in this simulation the conservation between confirmed and unconfirmed sites with equal motifs. Therefore, we modified the balanced motif distribution simulation to compare confirmed tandems with identical or highly similar unconfirmed tandems. To this end, we constructed for each confirmed tandem motif two lists: the first list contains all confirmed tandems with this motif, and the second list contains all unconfirmed tandems that are either identical or highly similar to this motif. Taking similar unconfirmed tandems into account makes the second list contain at least as many entries as the first one, so that this list can be used to sample a subset of unconfirmed tandems. Random sampling of unconfirmed tandems was repeated 1000 times.

For Δ4–Δ6 donors, we sampled only from identical unconfirmed tandems. Δ7–Δ9 donors were considered as similar if Δ7 motifs are identical in positions +1 to +5 and +7 to +10; Δ8 motifs are identical in positions +1 to +5 and +8 to +11; and there is at most one mismatch at positions +6 and +7; Δ9 motifs are identical in positions +1 to +5 and +9 to +12; and there are at most two mismatches at position +6, +7, and +8. Δ4–Δ9 acceptor motifs were considered as similar if they fulfill the conservation definition given above.

The reason to use this simulation is that the conservation differs between the motifs and the motif distribution differs between confirmed and unconfirmed ones. For example, the balanced motif distribution simulation estimates that 51% of the confirmed and 47.1% of the unconfirmed Δ4 donors are conserved; a difference of 3.9%. However, the global conservation is only 41.3% for unconfirmed Δ4 donors; a much higher difference of 9.7%. This indicates that unconfirmed Δ4 tandems are enriched in weakly conserved motifs that do not occur among the confirmed ones; for example, all of the 2226 GTAAGCA donors are unconfirmed, and this motif has an exceptionally low conservation level of 29.6% (660 of 2226). As the balanced motif distribution simulation compares either identical or highly similar motifs, it gives a fair estimation of a lower bound for f_s.

Correlation between motif conservation and splice site consensus constraints

For NAGNAG acceptors, we found that constraints on the acceptor splice site consensus are one main reason for the motif-specific conservation differences. To further test this, we considered Δ4 acceptors. As most of these acceptors are predominantly spliced at the upstream acceptor, we focused on the +4 position, which is often the 5′ exon end. The conservation level is 62.2% for NAGGNAG sites, 61.5% for NAGANAG, 60.5% for NAGCNAG, and 59.4% for NAGTNAG. Thus, the order G > A > C > T exactly correlates with the preference of the +1 position in the acceptor consensus (Abril et al. 2005), even though the conservation differences are not as pronounced as observed for NAGNAG acceptors.

We also determined the overall conservation level of Δ4 donors with a GTNNGTN motif, focusing on the +4 position in the tandem motif. GTNAGTN donors have the highest overall conservation level with 45.4%, followed by GTNGGTN (40%), GTNTGTN (35.6%), and GTNCGTN (16.9%). Again, the order A > G > T > C correlates perfectly with the +4 position preference in the donor consensus (Abril et al. 2005). For donors with a GCNNGTN motif, the GTN donor is predominant in most cases; thus the +4 position in the tandem motif represents the −1 position in the donor consensus. At the −1 position, G is preferred over A, T, and C (Abril et al. 2005). Again, this order correlates with the conservation level: GCNGGTN, 55.6%; GCNAGTN, 54.8%; GCNTGTN, 46.2%; and GCNCGTN, 42.1%. Thus, constraints on the donor and acceptor consensus are likely to be a major reason for the observed differences in the overall conservation levels of tandem motifs.

SUPPLEMENTAL DATA

Supplemental material can be found at http://www.rnajournal.org.

ACKNOWLEDGMENTS

We thank Christoph Englert, Jürgen Schulte Mönting, and Anke Busch for helpful discussions; and three anonymous referees for useful comments and suggestions. This work was supported by grants from the German Ministry of Education and Research to S.S. (01GS0426) and M.P. (01GR0504, 0313652D) as well as from the Deutsche Forschungsgemeinschaft to M.P. (SFB604-02) and K.H. (Hu498/3-1).

Received October 19, 2007; accepted January 3, 2008.

REFERENCES

Abril, J.F., Castelo, R., and Guigo, R. 2005. Comparison of splice sites in mammals and chicken. *Genome Res.* **15:** 111–119.

Akerman, M. and Mandel-Gutfreund, Y. 2006. Alternative splicing regulation at tandem 3′ splice sites. *Nucleic Acids Res.* **34:** 23–31. doi: 10.1093/nar/gkj408.

Baek, D. and Green, P. 2005. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc. Natl. Acad. Sci.* **102:** 12813–12818.

Barbaux, S., Niaudet, P., Gubler, M.C., Grunfeld, J.P., Jaubert, F., Kuttenn, F., Fekete, C.N., Souleyreau-Therville, N., Thibaud, E., Fellous, M., et al. 1997. Donor splice-site mutations in WT1 are responsible for Frasier syndrome. *Nat. Genet.* **17:** 467–470.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. 2004. Ultraconserved elements in the human genome. *Science* **304:** 1321–1325.

Bickel, P.J., Hammel, E.A., and O'Connell, J.W. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* **187:** 398–404.

Black, D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72:** 291–336.

Burgar, H.R., Burns, H.D., Elsden, J.L., Lalioti, M.D., and Heath, J.K. 2002. Association of the signaling adaptor FRS2 with fibroblast growth factor receptor 1 (Fgfr1) is mediated by alternative splicing of the juxtamembrane domain. *J. Biol. Chem.* **277:** 4018–4023.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M., and Buell, C.R. 2006. Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics* **7:** 327. doi: 10.1186/1471-2164-7-327.

Chern, T.M., van Nimwegen, E., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Zavolan, M. 2006. A simple physical model predicts small exon length variations. *PLoS Genet.* **2:** e45. doi: 10.1371/journal.pgen.0020045.

Clark, F. and Thanaraj, T.A. 2002. Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.* **11:** 451–464.

Cohen, T., Auerbach, W., Ravid, L., Bodennec, J., Fein, A., Futerman, A.H., Joyner, A.L., and Horowitz, M. 2005. The exon 8-containing prosaposin gene splice variant is dispensable for mouse development, lysosomal function, and secretion. *Mol. Cell. Biol.* **25:** 2431–2440.

Condorelli, G., Bueno, R., and Smith, R.J. 1994. Two alternatively spliced forms of the human insulin-like growth factor I receptor have distinct biological activities and internalization kinetics. *J. Biol. Chem.* **269:** 8510–8516.

Davies, R.C., Bratt, E., and Hastie, N.D. 2000. Did nucleotides or amino acids drive evolutionary conservation of the WT1 ± KTS alternative splice? *Hum. Mol. Genet.* **9:** 1177–1183.

Dou, Y., Fox-Walsh, K.L., Baldi, P.F., and Hertel, K.J. 2006. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *RNA* **12:** 2047–2056.

Ermakova, E.O., Nurtdinov, R.N., and Gelfand, M.S. 2007. Overlapping alternative donor splice sites in the human genome. *J. Bioinform. Comput. Biol.* **5:** 991–1004.

Graveley, B.R. 2005. Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell* **123:** 65–73.

Hammes, A., Guo, J.K., Lutsch, G., Leheste, J.R., Landrock, D., Ziegler, U., Gubler, M.C., and Schedl, A. 2001. Two splice variants of the Wilms' tumor 1 gene have distinct functions during sex determination and nephron formation. *Cell* **106:** 319–329.

Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R., and Platzer, M. 2004. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat. Genet.* **36:** 1255–1257.

Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R., and Platzer, M. 2006a. Single-nucleotide polymorphisms in NAGNAG acceptors are highly predictive for variations of alternative splicing. *Am. J. Hum. Genet.* **78:** 291–302.

Hiller, M., Huse, K., Szafranski, K., Rosenstiel, P., Schreiber, S., Backofen, R., and Platzer, M. 2006b. Phylogenetically widespread alternative splicing at unusual GYNGYN donors. *Genome Biol.* **7:** R65.

Hiller, M., Szafranski, K., Backofen, R., and Platzer, M. 2006c. Alternative splicing at NAGNAG acceptors: Simply noise or noise and more? *PLoS Genet.* **2:** e207; author reply e208. doi: 10.1371/journal.pgen.0020207.

Hiller, M., Nikolajewa, S., Huse, K., Szafranski, K., Rosenstiel, P., Schuster, S., Backofen, R., and Platzer, M. 2007. TassDB: A database of alternative tandem splice sites. *Nucleic Acids Res.* **35:** D188–D192. doi: 10.1093/nar/gkl762.

Hu, C.A., Lin, W.W., Obie, C., and Valle, D. 1999. Molecular enzymology of mammalian Δ^1-pyrroline-5-carboxylate synthase. Alternative splice donor utilization generates isoforms with different sensitivity to ornithine inhibition. *J. Biol. Chem.* **274:** 6754–6762.

Hymowitz, S.G., Compaan, D.M., Yan, M., Wallweber, H.J., Dixit, V.M., Starovasnik, M.A., and de Vos, A.M. 2003. The crystal structures of EDA-A1 and EDA-A2: Splice variants with distinct receptor specificity. *Structure* **11:** 1513–1520.

Joyce-Brady, M., Jean, J.C., and Hughey, R.P. 2001. γ-Glutamyltransferase and its isoform mediate an endoplasmic reticulum stress response. *J. Biol. Chem.* **276:** 9468–9477.

Julious, S.A. and Mullee, M.A. 1994. Confounding and Simpson's paradox. *BMJ* **309:** 1480–1481.

Karni, R., de Stanchina, E., Lowe, S.W., Sinha, R., Mu, D., and Krainer, A.R. 2007. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat. Struct. Mol. Biol.* **14:** 185–193.

Koenig Merediz, S.A., Schmidt, M., Hoppe, G.J., Alfken, J., Meraro, D., Levi, B.Z., Neubauer, A., and Wittig, B. 2000. Cloning of an interferon regulatory factor 2 isoform with different regulatory ability. *Nucleic Acids Res.* **28:** 4219–4224. doi: 10.1093/nar/28.21.4219.

Kuhn, R.M., Karolchik, D., Zweig, A.S., Trumbower, H., Thomas, D.J., Thakkapallayil, A., Sugnet, C.W., Stanke, M., Smith, K.E., Siepel, A., et al. 2007. The UCSC Genome Browser Database: Update 2007. *Nucleic Acids Res.* **35:** D668–D673. doi: 10.1093/nar/gkl928.

Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., and Brenner, S.E. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446:** 926–929.

Lewis, B.P., Green, R.E., and Brenner, S.E. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci.* **100:** 189–192.

Licatalosi, D.D. and Darnell, R.B. 2006. Splicing regulation in neurologic disease. *Neuron* **52:** 93–101.

Lynch, K.W. 2004. Consequences of regulated pre-mRNA splicing in the immune system. *Nat. Rev. Immunol.* **4:** 931–940.

Modrek, B. and Lee, C.J. 2003. Alternative splicing in the human, mouse, and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* **34:** 177–180.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., and Ares Jr., M. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes & Dev.* **21:** 708–718.

Pagani, F. and Baralle, F.E. 2004. Genomic variants in exons and introns: Identifying the splicing spoilers. *Nat. Rev. Genet.* **5:** 389–396.

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T.R., Morris, Q.D., et al. 2004. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol. Cell* **16:** 929–941.

Pan, Q., Bakowski, M.A., Morris, Q., Zhang, W., Frey, B.J., Hughes, T.R., and Blencowe, B.J. 2005. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet.* **21:** 73–77.

Raponi, M., Baralle, F.E., and Pagani, F. 2007. Reduced splicing efficiency induced by synonymous substitutions may generate a substrate for natural selection of new splicing isoforms: The case of CFTR exon 12. *Nucleic Acids Res.* **35:** 606–613. doi: 10.1093/nar/gkl1087.

Resch, A., Xing, Y., Alekseyenko, A., Modrek, B., and Lee, C. 2004. Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res.* **32:** 1261–1269. doi: 10.1093/nar/gkh284.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.

Simpson, E.H. 1951. The interpretation of interaction in contingency tables. *J. R. Stat. Soc. [Ser A]* **13:** 238–241.

Sorek, R. and Ast, G. 2003. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13:** 1631–1637.

Sorek, R., Shamir, R., and Ast, G. 2004. How prevalent is functional alternative splicing in the human genome? *Trends Genet.* **20:** 68–71.

Stadler, M.B., Shomron, N., Yeo, G.W., Schneider, A., Xiao, X., and Burge, C.B. 2006. Inference of splicing regulatory activities by sequence neighborhood analysis. *PLoS Genet.* **2:** e191. doi: 10.1371/journal.pgen.0020191.

Stoilov, P., Daoud, R., Nayler, O., and Stamm, S. 2004. Human tra2-β1 autoregulates its protein concentration by influencing alternative splicing of its pre-mRNA. *Hum. Mol. Genet.* **13:** 509–524.

Sugnet, C.W., Kent, W.J., Ares Jr., M., and Haussler, D. 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.* **2004:** 66–77.

Tadokoro, K., Yamazaki-Inoue, M., Tachibana, M., Fujishiro, M., Nagao, K., Toyoda, M., Ozaki, M., Ono, M., Miki, N., Miyashita, T., et al. 2005. Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: The case of Gln in DRPLA affects subcellular localization of the products. *J. Hum. Genet.* **50:** 382–394.

Takeda, H., Matsuzaki, T., Oki, T., Miyagawa, T., and Amanuma, H. 1994. A novel POU domain gene, zebrafish pou2: Expression and roles of two alternatively spliced twin products in early development. *Genes & Dev.* **8:** 45–59.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.

Tress, M.L., Martelli, P.L., Frankish, A., Reeves, G.A., Wesselink, J.J., Yeats, C., Olason, P.L., Albrecht, M., Hegyi, H., Giorgetti, A., et al. 2007. The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl. Acad. Sci.* **104:** 5495–5500.

Tsai, K.W., Tarn, W.Y., and Lin, W.C. 2007. Wobble splicing reveals the role of the branch point sequence-to-NAGNAG region in 3′ tandem splice site selection. *Mol. Cell. Biol.* **27:** 5835–5848.

Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., et al. 2005. Nova regulates brain-specific splicing to shape the synapse. *Nat. Genet.* **37:** 844–852.

Ureta-Vidal, A., Ettwiller, L., and Birney, E. 2003. Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4:** 251–262.

Vogan, K.J., Underhill, D.A., and Gros, P. 1996. An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity. *Mol. Cell. Biol.* **16:** 6677–6686.

Wu, S., Romfo, C.M., Nilsen, T.W., and Green, M.R. 1999. Functional recognition of the 3′ splice site AG by the splicing factor U2AF35. *Nature* **402:** 832–835.

Xing, Y. and Lee, C.J. 2005. Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet.* **1:** e34. doi: 10.1371/journal.pgen.0010034.

Xu, Q., Belcastro, M.P., Villa, S.T., Dinkins, R.D., Clarke, S.G., and Downie, A.B. 2004. A second protein L-isoaspartyl methyltransferase gene in *Arabidopsis* produces two transcripts whose products are sequestered in the nucleus. *Plant Physiol.* **136:** 2652–2664.

Yan, M., Wang, L.C., Hymowitz, S.G., Schilbach, S., Lee, J., Goddard, A., de Vos, A.M., Gao, W.Q., and Dixit, V.M. 2000. Two-amino acid molecular switch in an epithelial morphogen that regulates binding to two distinct receptors. *Science* **290:** 523–527.

Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T., and Burge, C.B. 2005. Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl. Acad. Sci.* **102:** 2850–2855.

Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D.A., Hayashizaki, Y., and Gaasterland, T. 2003. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.* **13:** 1290–1300.

RILEEN SINHA
THORSTEN LENSER
NIELS JAHN
ULRIKE GAUSMANN
SWETLANA FRIEDEL
KAROL SZAFRANKSI
KLAUS HUSE
PHILIP ROSENSTIEL
JOCHEN HAMPE
STEFAN SCHUSTER
MICHAEL HILLER
ROLF BACKOFEN
MATTHIAS PLATZER

TassDB2 - A comprehensive database of subtle alternative splicing events

(under review at BMC Bioinformatics)

Subtle alternative splicing (AS) events involving tandem splice sites separated by a short (2-12 nucleotides) distance are frequent and evolutionarily widespread in eukaryotes, and contribute to the complexity of transcriptomes and proteomes. We have substantially revised and extended our database TassDB (Tandem Splice Site DataBase, version 1), which stores data about AS events at tandem splice sites separated by 3 nt in eight species. The currently available version 2 contains extensive information about tandem splice sites separated by 2-12 nt for the human and mouse transcriptomes, including data on the conservation of the tandem motifs in five vertebrates. TassDB2 offers a user-friendly interface to search for specific genes or for genes containing tandem splice sites with specific features. The results provide detailed information about the subtle tandem splice sites, as well as the possibility to download result datasets. Links are provided to the UCSC genome browser and other external resources.

TassDB2 - A comprehensive database of subtle alternative splicing events

Rileen Sinha[1, 2], Thorsten Lenser[3], Niels Jahn[2], Ulrike Gausmann[2], Swetlana Friedel[4], Karol Szafranski[2], Klaus Huse[2], Philip Rosenstiel[5], Jochen Hampe[6], Stefan Schuster[7], Michael Hiller[8], Rolf Backofen[1] and Matthias Platzer[2*]

Address: [1]Bioinformatics group, Albert-Ludwigs-University Freiburg, Georges-Koehler-Allee 106, 79110 Freiburg, Germany

[2] Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany

[3] Bio Systems Analysis Group, Friedrich Schiller University Jena, Ernst-Abbe-Platz 1–4, D-07743 Jena, Germany

[4] Leibniz Institute for Natural Product Research and Infection Biology, Hans-Knöll-Institute, Systems Biology/Bioinformatics, Beutenbergstrasse.11a, 07745 Jena, Germany

[5] Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Schittenhelmstrasse, 12, 24105 Kiel, Germany

[6] Department of General Internal Medicine, University Hospital Schleswig-Holstein, Campus Kiel, Schittenhelmstrasse, 12, 24105 Kiel, Germany

[7] Department of Bioinformatics, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany

[8] Department of Developmental Biology, Stanford University, Stanford, CA 94305, USA

Email : Rileen Sinha – rsinha@fli-leibniz.de; Thorsten Lenser - thorsten.lenser@minet.uni-jena.de; Niels Jahn - nielsj@fli-leibniz.de; Ulrike Gausmann – ugau@fli-leibniz.de; Swetlana Friedel – swetlana.friedel@hki-jena.de; Karol Szafranski – szafrans@fli-leibniz.de; Klaus Huse – khuse@fli-leibniz.de; Philip Rosenstiel – p.rosenstiel@mucosa.de; Stefan.Schuster - Stefan.Schu@uni-jena.de; Michael Hiller – hillerm@stanford.edu; Rolf Backofen - backofen@informatik.uni-freiburg.de; Matthias Platzer – mplatzer@fli-leibniz.de

Abstract

Background

Subtle alternative splicing events involving tandem splice sites separated by a short (2-12 nucleotides) distance are frequent and evolutionarily widespread in eukaryotes, and a major contributor to the complexity of transcriptomes and proteomes. However, these events have been either omitted altogether in databases on alternative splicing, or only the cases of experimentally confirmed alternative splicing have been reported. Thus, a database which covers all confirmed cases of subtle alternative splicing as well as the numerous putative tandem splice sites (which might be confirmed once more transcript data becomes available), and allows to search for tandem splice sites with specific features and download the results, is a valuable resource for targeted experimental studies and large-scale bioinformatics analyses of tandem splice sites. Towards this goal we recently set up TassDB (Tandem Splice Site DataBase, version 1), which stores data about alternative splicing events at tandem splice sites separated by 3 nt in eight species.

Description

We have substantially revised and extended TassDB. The currently available version 2 contains extensive information about tandem splice sites separated by 2-12 nt for the human and mouse transcriptomes including data on the conservation of the tandem motifs in five vertebrates. TassDB2 offers a user-friendly interface to search for specific genes or for genes containing tandem splice sites with specific features as well as the possibility to download result datasets. For example, users can search for cases of alternative splicing where the proportion of EST/mRNA evidence supporting the minor isoform exceeds a specific threshold, or where the difference in splice site scores is specified by the user. The predicted impact of each event on the protein is also reported, along with information about being a putative target for the nonsense-mediated decay (NMD) pathway. Links are provided to the UCSC genome browser and other external resources.

Conclusion

TassDB2, available via http://www.tassdb.info, provides comprehensive resources for researchers interested in both targeted experimental studies and large-scale bioinformatics analyses of short distance tandem splice sites.

Background

Alternative splicing (AS), a process which enables the production of multiple mRNA transcripts by the same gene via the variable inclusion of parts of the primary transcript, is very widespread in eukaryotes – almost all multi-exonic human genes are believed to undergo AS [1, 2]. Thus, AS is a major contributor to the complexity and diversity of eukaryotic transcriptomes and proteomes. The splice variants produced can either exhibit different properties (e.g. half-life, translational efficiency), be translated into different protein isoforms with potentially different functions, or can be degraded via pathways such as the nonsense-mediated decay (NMD) [3]. AS can often be specific to a tissue type or developmental stage, and the majority of human AS events are believed to be regulated in this sense [1]. The regulation of AS has been shown to play an important role in several developmental processes in various organisms, and defects in AS can lead to diseases [4].

Subtle AS, involving splice sites separated by a distance of 2-12 nt, is an important, evolutionarily widespread subclass of AS [5]. Such AS is called subtle because the resulting mRNA isoforms differ by only a few nucleotides. While alternative acceptors (AA) and alternative donors (AD) together constitute about a third of all AS events in humans, subtle AS events comprise about a third of AA and AD events – for example, subtle events constitute 1,586 (38%) out of 4,179 AA events and 774 (28%) out of 2,728 AD events in the "alt events" track of the UCSC genome browser [6] for a combined total of 34% (2,360/6,907). Another reason for treating these events separately is that the mechanisms behind such events are likely different from those involving splice sites separated by larger distances – for example, the emergence of a second polypyrimidine tract can be observed for alternative acceptors separated by 8 or more nucleotides, and events which result in a frame-preserving difference of transcript length are seen to be more common than frame-shifting ones, once we move beyond a difference of 12 nt [7, 8].

It is a matter of debate as to what fraction of subtle AS events are truly functional, as opposed to being a result of a noisy process in which the spliceosome stochastically selects between nearby competing alternatives [5, 9-11]. Consistent with estimations that a fraction of those subtle AS events is under purifying selection [12], there are several known cases where they result in functionally different protein isoforms or affect the translational efficiency when located in the untranslated regions (UTR) [5]. Moreover, subtle AS can also have a decidedly unsubtle effect in cases where a premature stop codon can be created, which is especially likely in cases where the splice sites are separated by a distance which is not a multiple of 3. Mutations that create frame-preserving tandem splice sites affecting the coding region are selected against [13] and in the case of *ABCA4* are associated with human disease [14]. In the following, we shall use the notation *Δx* to denote a subtle splice event involving sites separated by *x* nucleotides, so for example, the class *Δ3* shall be used to mean all GYNGYN and NAGNAG AS events (Y stands for C or T; N for A, C, G, or T), and so on.

TassDB1 (TAndem Splice Site DataBase, version 1), the first database devoted to subtle AS, provides large collections of *Δ3* donors and acceptors in eight species [15]. We have extended TassDB1 considerably, to create TassDB2, which provides a comprehensive collection of all human and mouse donors and acceptors in the *Δ2*- *Δ12* range. We note that while TassDB provided data on 8 species, TassDB2 only includes 2 species, human and mouse. This is because the transcriptome coverage by ESTs/mRNAs in the remaining species was insufficient for detection of a non-negligible number of AS events involving the larger distances in the *Δ2*- *Δ12* range. TassDB2 includes data on the conservation of the tandem motifs in five vertebrates (human, mouse, dog, chicken and zebrafish). Thus, TassDB2 provides comprehensive information on 22 event types, compared to 2 (NAGNAG and GYNGYN) in TassDB1. Thus TassDB2 is effectively a new database rather than just a simple extension. A user-friendly search interface features both a “quick search” mode, in which a user can search using gene symbol, as well as an “advanced search” mode, in which several different criteria can be specified by the user, and the possibility to download result datasets.

Database construction and content

Data

TassDB2 uses an annotation pipeline based on transcript-to-genome mappings taken from the UCSC genome browser [16]. We used the RefSeq annotation as well as the UCSC ‘knownGene’ set for human (build hg18) and mouse (build mm9). The exon–intron structure as well as the protein-coding sequence (CDS) annotation was as per the UCSC annotation. Alternatively spliced tandem splicing events were identified using BLAST against all ESTs and mRNAs from the respective species as described in [17, 18].

For each tandem splice site and the confirmed or putative AS event, TassDB2 contains the following data: the splice site motif, its genomic locus, its location in the transcript (5'/3'-UTR or CDS with intron phase 0/1/2), the (predicted) impact of the splice event on the protein, the sequences and length of the up-/downstream exon and the intron, and information about the ESTs/mRNAs that indicate usage (if any) of the splice sites. As the strength of the splice sites in a tandem often helps to distinguish between alternatively and non-alternatively spliced tandem motifs [7, 9, 18], we also computed splice site scores for both splice sites in each tandem [19].

TassDB2 holds splice site specific data as well as transcript-dependent data. Some features, such as the tandem motif (the two NAGs or GTNs, and the intervening sequence, if any – *Δ2* being a special case, with motifs NAGAG and GTGTN), the genomic locus and the splice site scores, are independent of transcript annotation. However, other features such as intron phase, protein impact, EST confirmation and predicted targeting by NMD depend on the CDS annotation and the exon–intron structure of the transcript. Targeting by NMD is predicted in the usual manner - we calculated the nucleotide distance between the stop position (corresponding to the given splice variant) and the position of the last exon-exon junction, and if this distance was greater than 50, targeting by the NMD pathway was predicted.

Database Design

The web-frontend to TassDB2 is created in HTML with PHP and JavaScript. The data is stored in a relational database, running under the MySQL database system. The data is primarily organized in the database tables *splicesite, spliceeventdata,* and *transcript* (Fig. 1).

The table *splicesite* contains sequence-dependent information such as the genomic locus, the splice site pattern with its sequence context, the splice site scores, and conserved tandem sequences (if available) in human/mouse, chicken, dog, and zebrafish. All transcript-dependent data is stored in table *spliceeventdata*: the transcripts which have the tandem site in their exon-intron structure, the annotated splice site, the number of ESTs for each (potential) tandem splice variant along with the two BLAST queries used to find the ESTs, the predicted protein impact, and the NMD prediction. The table *transcript* contains the information on the transcripts that is independent from the splice sites. The three main tables are linked through the *ss2transcript2sed* table.

Additionally, each splice site is linked to information on its gene (table *gene*), and its conservation in other species (table *splicesite_conservation*; species are human, mouse, dog, chicken, zebrafish, representing the major vertebrate clades). The splicing events are linked to their supporting ESTs in the table *est.* The user interface contains links giving a detailed description of each data field.

Summary statistics of human tandem splice sites in TassDB2 are given in Table 1.

Utility and Discussion

User interface – quick search and advanced search

We anticipate that the most frequent use of TassDB2 will be a search for tandem splice sites of a given gene. Therefore, TassDB2 provides a “quick search” interface where a user need only specify a gene symbol or a transcript accession number, and the entire information of both confirmed and unconfirmed tandem splicing events for this gene is displayed.

Often, however, users might be interested in information which requires a selection of tandem splice sites with specific features. To address this, TassDB2 also provides an “advanced search interface” (Fig. 2) where the search can be restricted using one or more of the following features: (i) Δ - the distance between the splice sites, (ii) frame-preserving or/and frame-shifting, (iii) number of ESTs/mRNAs that match both splice forms, (iv) “minor isoform ratio”, that is the fraction of ESTs/mRNA that support the minor isoform, (v) tandem site conservation in any or all of five organisms (human, mouse, dog, chicken, and zebrafish) (vi) splice site scores for the two splice sites, (vii) the difference in the splice site scores, and (ix) location in the UTR or CDS. Thus, it is easy to formulate queries such as: “Show all confirmed $\Delta 3$ events with a minor isoform ratio ≥ 0.4 ”, “Show all tandem splice sites where both splice forms are represented by at least two ESTs/mRNAs and the minor isoform ratio is

≥ 0.15” or “Show all confirmed frame-shifting tandem donors which are located in the CDS”. Additionally, the search can be restricted to certain genes.

User interface – reporting results

The result of the search consists of two parts: (i) a summary table listing the affected genes and their number of tandem splice sites of each type, and (ii) detailed tables containing information regarding the individual tandem splice sites. These detailed result tables also provide links to the ESTs/mRNAs for both splice forms as well as links to the UCSC genome browser. If the transcript specific data differ between transcripts, TassDB2 shows detailed result tables with more than two columns. Features that differ between transcripts are shown in black while those that are identical in all transcripts are shown in grey color.

Examples

Searching for all confirmed tandem splice sites in the gene *HHIP* (hedgehog interacting protein) in human leads to the result page shown in Fig. 3: *HHIP* has one confirmed *Δ4* tandem acceptor event, with the upstream and downstream acceptor supported by 30 and 34 ESTs/mRNAs, respectively. The event is predicted to lead to targeting by NMD according to one of the representative transcripts (uc003ijs.1, NM_022475), but not according to the other (uc003ijr.1).

While AS has now been established as a widespread phenomenon and a substantial contributor to the complexity of eukaryotic transcriptomes and proteomes, it is still a matter of great debate as to how many AS events are truly functional [3, 20, 21]. The literature regarding this question is the motivation behind providing the options for searching by splice site score difference and minor isoform ratio in TassDB2. It has been observed that comparable splice site strength is often indicative of both splice sites in a tandem being used, whereas a higher fraction of ESTs/mRNA supporting the minor isoform is a good test of whether the event is likely to be genuine AS events rather than just noise [22, 23]. As an example, searching for all confirmed tandem splicing events with a minor isoform ratio of≥

0.45 yields 300 results, and increasing the threshold of supporting ESTs/mRNAs to ≥ 10 for each variant yields 170 results.

The TassDB2 resource also includes the BayNAGNAG webserver (available at http://www.tassdb.info/baynagnag/), which uses Bayesian networks to predict the splicing outcome at NAGNAG tandem splice sites in an EST/mRNA independent way based on splice site features [24].

Conclusions

TassDB2 is a comprehensive resource for information regarding subtle AS. Users can easily search for individual genes, as well as by various criteria corresponding to different features of the tandem splice sites. Some of the criteria can be used to enrich for splicing events which are likely to have functional significance. The results can be downloaded for further exploration, and flat files have also been made available for those who wish to carry out their own large-scale bioinformatics studies. Thus TassDB2 should be a very useful resource for scientists interested in subtle AS.

Availability and requirements

TassDB2 is freely available for online use at http://www.tassdb.info

TassDB2 can be used via any standard internet browser.

Author's contributions

RS participated in database design and testing, found the alternative tandem splice sites, and drafted the manuscript. TL participated in database design and testing, and implemented several parts of the front-end and back-end of TassDB2. NJ improved the implementation of the database, and optimized it for faster searches. UG participated in database design and performed extensive testing. MH, SF and KS participated in database design and testing, and provided data and scripts. SF and UG took part in front-end improvement. KH participated in database design and testing. PR, JH, SS, MH, RB and MP designed and supervised the project, participated in database design and testing, and manuscript preparation. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by grants from the German Ministry of Education and Research (01GS08182, 01GS0809, 0313652), the Deutsche Forschungsgemeinschaft (SFB604-02, Hi 1423/2-1) and the Human Frontier Science Program (Fellowship LT000896/2009-L).

References

1. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.
2. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays**. *Science* 2003, **302**(5653):2141-2144.
3. Lareau LF, Green RE, Bhatnagar RS, Brenner SE: **The evolving roles of alternative splicing**. *Curr Opin Struct Biol* 2004, **14**:273 - 282.
4. Faustino NA, Cooper TA: **Pre-mRNA splicing and human disease**. *Genes Dev* 2003, **17**(4):419 - 437.
5. Hiller M, Platzer M: **Widespread and subtle: alternative splicing at short-distance tandem sites**. *Trends in Genetics* 2008, **24**(5):246-255.
6. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A *et al*: **The UCSC genome browser database: update 2007**. *Nucl Acids Res* 2007, **35**(suppl_1):D668-673.

7. Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ: **Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site**. *RNA* 2006, **12**(12):2047-2056.

8. 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* 2008, **9**(1):202.

9. Chern T-M, van Nimwegen E, Kai C, Kawai J, Carninci P, Hayashizaki Y, Zavolan M: **A Simple Physical Model Predicts Small Exon Length Variations**. *PLoS Genetics* 2006, **2**(4):e45.

10. Hiller M, Szafranski K, Backofen R, Platzer M: **Alternative Splicing at NAGNAG Acceptors: Simply Noise or Noise and More?** *PLoS Genetics* 2006, **2**(11):e207.

11. van Nimwegen E, Zavolan M: **Authors' Reply**. *PLoS Genetics* 2006, **2**(11):e208.

12. Hiller M, Szafranski K, Sinha R, Huse K, Nikolajewa S, Rosenstiel P, Schreiber S, Backofen R, Platzer M: **Assessing the fraction of short-distance tandem splice sites under purifying selection**. *RNA* 2008, **14**(4):616-629.

13. Hiller M, Szafranski K, Huse K, Backofen R, Platzer M: **Selection against tandem splice sites affecting structured protein regions**. *BMC Evolutionary Biology* 2008, **8**(1):89.

14. Maugeri A, van Driel MA, van de Pol DJR, Klevering BJ, van Haren FJJ, Tijmes N, Bergen AAB, Rohrschneider K, Blankenagel A, Pinckers AJLG *et al*: **The 2588G'C Mutation in the ABCR Gene Is a Mild Frequent Founder Mutation in the Western European Population and Allows the Classification of ABCR Mutations in Patients with Stargardt Disease**. 1999, **64**(4):1024-1035.

15. Hiller M, Nikolajewa S, Huse K, Szafranski K, Rosenstiel P, Schuster S, Backofen R, Platzer M: **TassDB: a database of alternative tandem splice sites**. *Nucl Acids Res* 2007, **35**(suppl_1):D188-192.

16. Karolchik D, Baertsch R, Diekhans R, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ: **The UCSC Genome Browser Database**. *Nucleic Acids Res* 2003, **31**:51 - 54.

17. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M: **Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity**. *Nat Genet* 2004, **36**(12):1255-1257.

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

19. Yeo G, Burge CB: **Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals**. *Journal of Computational Biology* 2004, **11**(2-3):377-394.

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

21. Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason Pl, Albrecht M, Hegyi H, Giorgetti A *et al*: **The implications of alternative splicing in the ENCODE protein complement**. *PNAS* 2007, **104**(13):5495-5500.

22. Szafranski K, Schindler S, Taudien S, Hiller M, Huse K, Jahn N, Schreiber S, Backofen R, Platzer M: **Violating the splicing rules: TG dinucleotides function as alternative 3' splice sites in U2-dependent introns**. *Genome Biology* 2007, **8**(8):R154.

23. Kan Z, States D, Gish W: **Selecting for Functional Alternative Splices in ESTs**. *Genome Res* 2002, **12**(12):1837-1845.

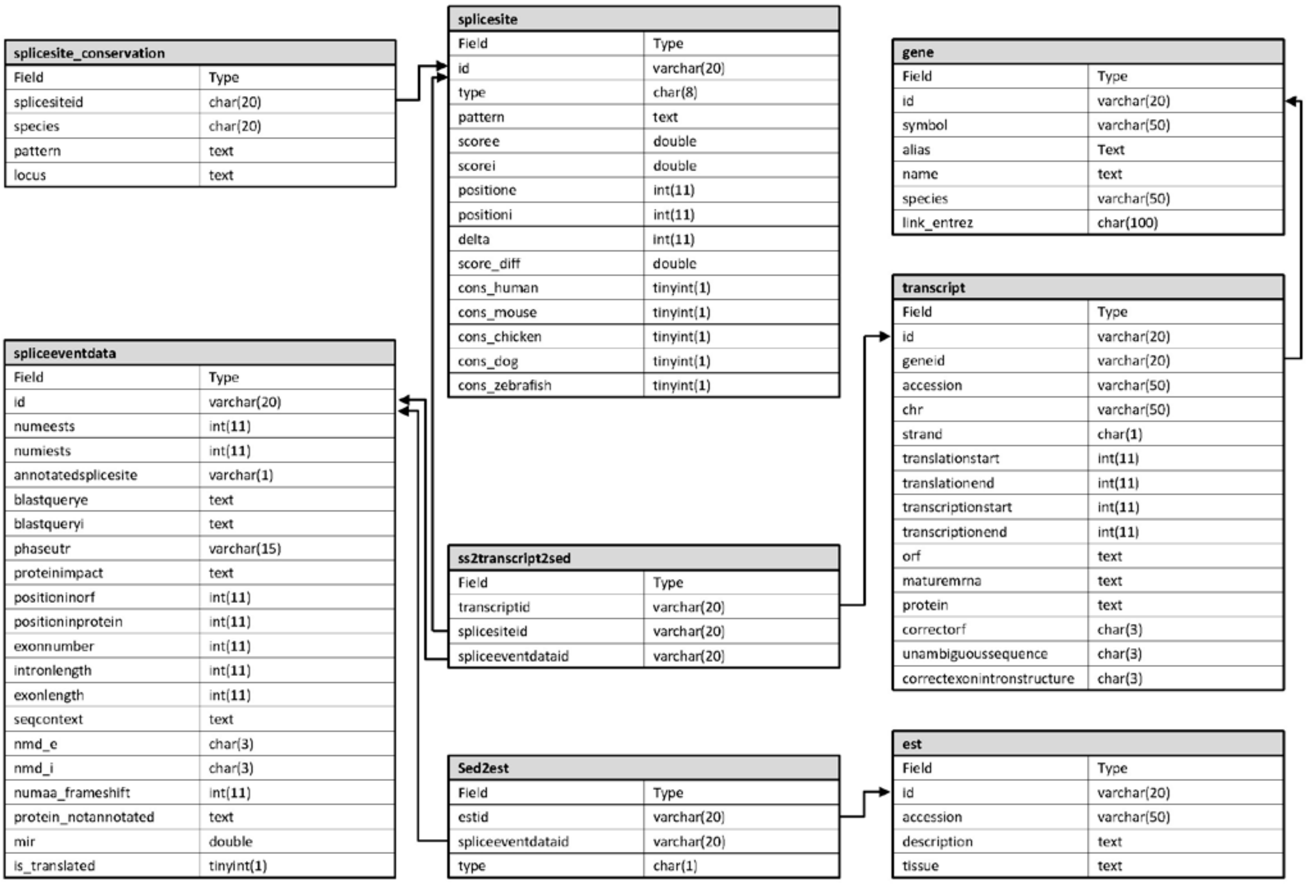
24. Sinha R, Nikolajewa S, Szafranski K, Hiller M, Jahn N, Huse K, Platzer M, Backofen R: **Accurate prediction of NAGNAG alternative splicing**. *Nucl Acids Res* 2009, **37**(11):3569-3579.

Table 1. Statistics of human tandem splice sites in TassDB2.

Delta	Donors # %				Acceptors # %			
	Tandem splice sites		Confirmed* alternative		Tandem splice sites		Confirmed alternative	
2	9,825	*1.9*	164	*1.7*	11,135	*6.7*	252	*2.3*
3	11,164	*2.2*	166	*1.5*	12,542	*7.5*	2,272	*18.1*
4	130,104	25.3	955	*0.7*	11,852	*7.1*	961	*8.1*
5	36,643	*7.1*	269	*0.7*	11,314	*6.8*	609	*5.4*
6	54,142	*10.5*	275	*0.5*	16,495	*9.9*	396	*2.4*
7	45,161	*8,8*	150	*0.3*	17,290	*10.4*	179	*1.0*
8	42,670	*8.3*	150	*0.4*	15,386	*9.2*	175	*1.1*
9	46,688	*9.1*	249	*0.5*	18,645	*11.2*	212	*1.1*
10	47,092	*9.1*	217	*0.5*	17,294	*10.4*	160	*0.9*
11	44,831	*8.7*	157	*0.4*	15,832	*9.5*	123	*0.8*
12	46,654	*9.1*	267	*0.6*	18,819	*11.3*	204	*1.1*
Total	514,974	*100.0*	3,019	*0.6*	166,604	*100.0*	5,543	*3.3*

* Tandem splice sites are considered confirmed if both splice forms have at least one supporting EST/mRNA.

Figure 1. The database scheme of TassDB2.



splicesite_conservation	
Field	Type
splicesiteid	char(20)
species	char(20)
pattern	text
locus	text

splicesite	
Field	Type
id	varchar(20)
type	char(8)
pattern	text
scoree	double
scorei	double
positione	int(11)
positioni	int(11)
delta	int(11)
score_diff	double
cons_human	tinyint(1)
cons_mouse	tinyint(1)
cons_chicken	tinyint(1)
cons_dog	tinyint(1)
cons_zebrafish	tinyint(1)

gene	
Field	Type
id	varchar(20)
symbol	varchar(50)
alias	Text
name	text
species	varchar(50)
link_entrez	char(100)

spliceeventdata	
Field	Type
id	varchar(20)
numeests	int(11)
numiests	int(11)
annotatedsplicesite	varchar(1)
blastquerye	text
blastqueryi	text
phaseutr	varchar(15)
proteinimpact	text
positioninorf	int(11)
positioninprotein	int(11)
exonnumber	int(11)
intronlength	int(11)
exonlength	int(11)
seqcontext	text
nmd_e	char(3)
nmd_i	char(3)
numaa_frameshift	int(11)
protein_notannotated	text
mir	double
is_translated	tinyint(1)

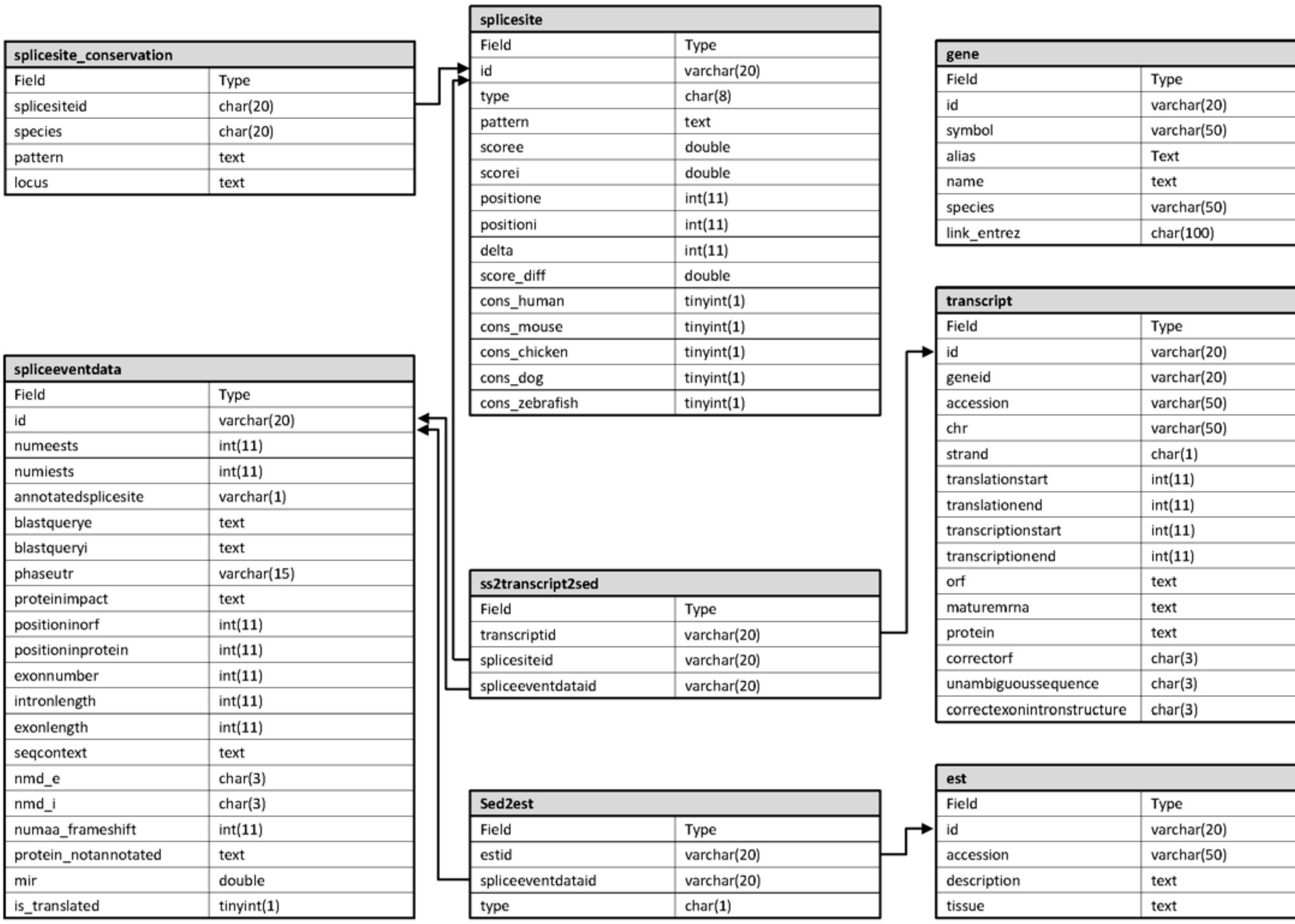
transcript	
Field	Type
id	varchar(20)
geneid	varchar(20)
accession	varchar(50)
chr	varchar(50)
strand	char(1)
translationstart	int(11)
translationend	int(11)
transcriptionstart	int(11)
transcriptionend	int(11)
orf	text
maturemrna	text
protein	text
correctorf	char(3)
unambiguoussequence	char(3)
correctexonintronstructure	char(3)

ss2transcript2sed	
Field	Type
transcriptid	varchar(20)
splicesiteid	varchar(20)
spliceeventdataid	varchar(20)

Sed2est	
Field	Type
estid	varchar(20)
spliceeventdataid	varchar(20)
type	char(1)

est	
Field	Type
id	varchar(20)
accession	varchar(50)
description	text
tissue	text

Figure 1. The database scheme of TassDB2.



splicesite_conservation

Field	Type
splicesiteid	char(20)
species	char(20)
pattern	text
locus	text

splicesite

Field	Type
id	varchar(20)
type	char(8)
pattern	text
scoree	double
scorei	double
positione	int(11)
positioni	int(11)
delta	int(11)
score_diff	double
cons_human	tinyint(1)
cons_mouse	tinyint(1)
cons_chicken	tinyint(1)
cons_dog	tinyint(1)
cons_zebrafish	tinyint(1)

gene

Field	Type
id	varchar(20)
symbol	varchar(50)
alias	Text
name	text
species	varchar(50)
link_entrez	char(100)

spliceeventdata

Field	Type
id	varchar(20)
numeests	int(11)
numiests	int(11)
annotatedsplicesite	varchar(1)
blastquerye	text
blastqueryi	text
phaseutr	varchar(15)
proteinimpact	text
positioninorf	int(11)
positioninprotein	int(11)
exonnumber	int(11)
intronlength	int(11)
exonlength	int(11)
seqcontext	text
nmd_e	char(3)
nmd_i	char(3)
numaa_frameshift	int(11)
protein_notannotated	text
mir	double
is_translated	tinyint(1)

transcript

Field	Type
id	varchar(20)
geneid	varchar(20)
accession	varchar(50)
chr	varchar(50)
strand	char(1)
translationstart	int(11)
translationend	int(11)
transcriptionstart	int(11)
transcriptionend	int(11)
orf	text
maturemrna	text
protein	text
correctorf	char(3)
unambiguoussequence	char(3)
correctexonintronstructure	char(3)

ss2transcript2sed

Field	Type
transcriptid	varchar(20)
splicesiteid	varchar(20)
spliceeventdataid	varchar(20)

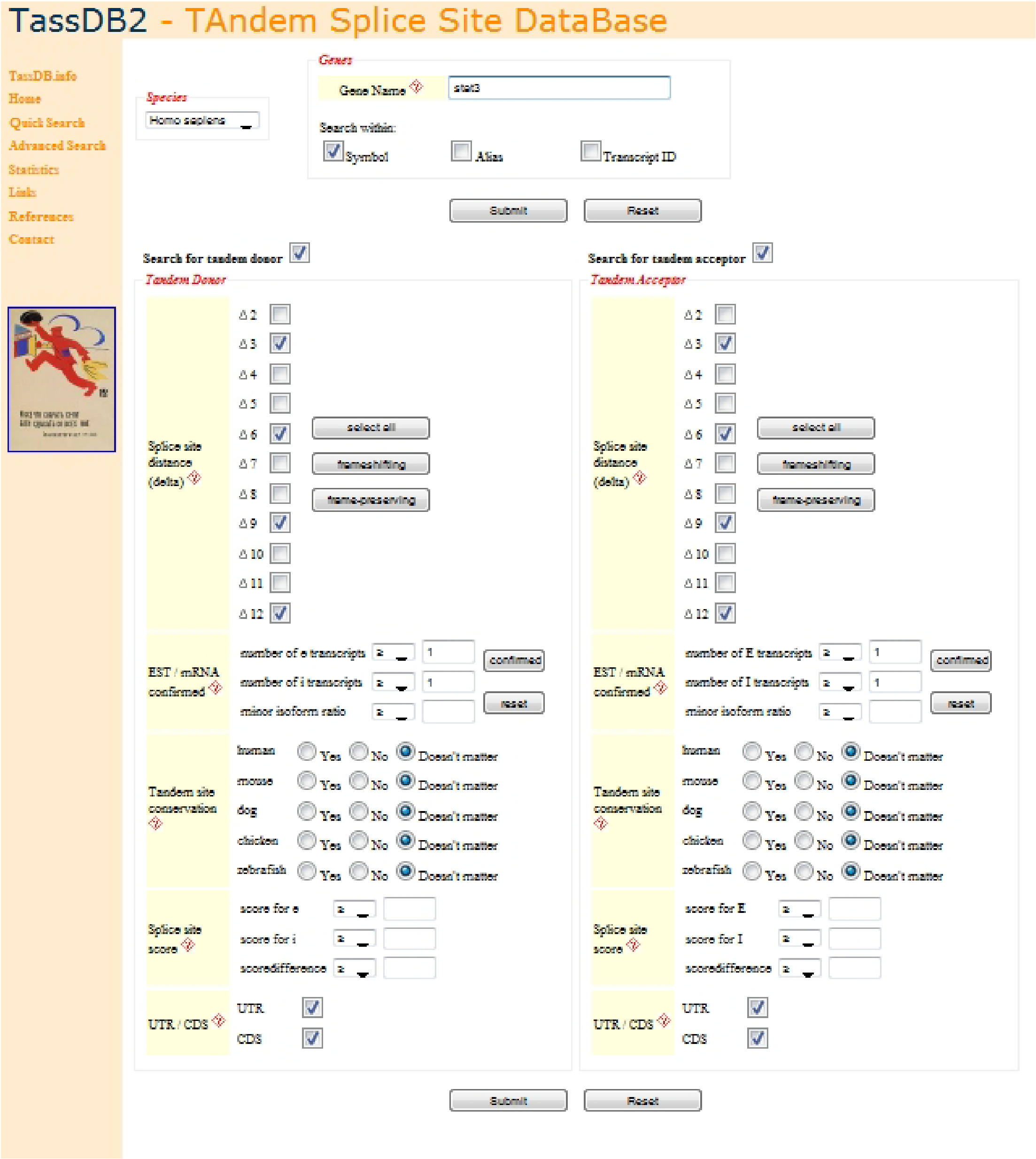
Sed2est

Field	Type
estid	varchar(20)
spliceeventdataid	varchar(20)
type	char(1)

est

Field	Type
id	varchar(20)
accession	varchar(50)
description	text
tissue	text

Figure 2. The advanced search interface of TassDB2.



Species

Homo sapiens

Genes

Gene Name: stat3

Search within: ☑ Symbol □ Alias □ Transcript ID

Search for tandem donor ☑

Tandem Donor

Splice site distance (delta): Δ2 □ Δ3 ☑ Δ4 □ Δ5 □ Δ6 ☑ Δ7 □ Δ8 □ Δ9 ☑ Δ10 □ Δ11 □ Δ12 ☑

EST / mRNA confirmed:
number of e transcripts ≥ 1
number of i transcripts ≥ 1
minor isoform ratio ≥

Tandem site conservation:
human ○ Yes ○ No ● Doesn't matter
mouse ○ Yes ○ No ● Doesn't matter
dog ○ Yes ○ No ● Doesn't matter
chicken ○ Yes ○ No ● Doesn't matter
zebrafish ○ Yes ○ No ● Doesn't matter

Splice site score:
score for e ≥
score for i ≥
scoredifference ≥

UTR / CDS: UTR ☑ CDS ☑

Search for tandem acceptor ☑

Tandem Acceptor

Splice site distance (delta): Δ2 □ Δ3 ☑ Δ4 □ Δ5 □ Δ6 ☑ Δ7 □ Δ8 □ Δ9 ☑ Δ10 □ Δ11 □ Δ12 ☑

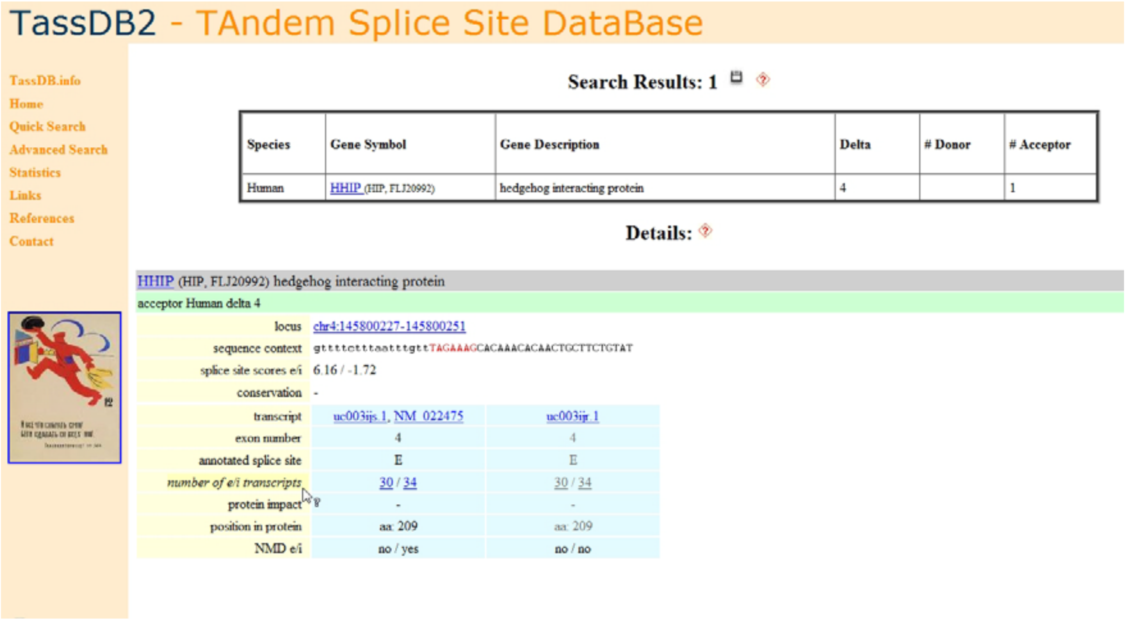
EST / mRNA confirmed:
number of E transcripts ≥ 1
number of I transcripts ≥ 1
minor isoform ratio ≥

Tandem site conservation:
human ○ Yes ○ No ● Doesn't matter
mouse ○ Yes ○ No ● Doesn't matter
dog ○ Yes ○ No ● Doesn't matter
chicken ○ Yes ○ No ● Doesn't matter
zebrafish ○ Yes ○ No ● Doesn't matter

Splice site score:
score for E ≥
score for I ≥
scoredifference ≥

UTR / CDS: UTR ☑ CDS ☑

Figure 3. Result page for all confirmed tandem splice sites in the gene *HHIP*.



TassDB2 - TAndem Splice Site DataBase

TassDB.info
Home
Quick Search
Advanced Search
Statistics
Links
References
Contact

Search Results: 1

Species	Gene Symbol	Gene Description	Delta	# Donor	# Acceptor
Human	HHIP (HIP, FLJ20992)	hedgehog interacting protein	4		1

Details:

HHIP (HIP, FLJ20992) hedgehog interacting protein

acceptor Human delta 4

locus	chr4:145800227-145800251	
sequence context	gttttctttaatttgttTAGAAAGCACAAACACAACTGCTTCTGTAT	
splice site scores e/i	6.16 / -1.72	
conservation	-	
transcript	uc003ijs.1, NM_022475	uc003ijr.1
exon number	4	4
annotated splice site	E	E
number of e/i transcripts	30 / 34	30 / 34
protein impact	-	-
position in protein	aa: 209	aa: 209
NMD e/i	no / yes	no / no

DISCUSSION

Discussion

Alternative Splicing (AS) has emerged as a key element of gene regulation in higher eukaryotes, and an important contributor to transcriptome and proteome diversity [19]. The focus of this thesis is on prediction of various alternative splicing events in animals and plants using Machine Learning and novel features, and studying the extent and conservation of subtle alternative splicing. Through my studies of alternative splicing, I have sought to understand the mechanisms involved in constitutive splicing, exon skipping, and subtle alternative splicing.

Improved identification of conserved cassette exons

The motivation for my first work [108] came partly from the fact that while large-scale detection of AS is often performed using EST data and is heavily dependent on the extent of coverage afforded by such data (greater the coverage, more are the AS events detected) [25, 36, 94], the amount of available EST data varies greatly across the organisms whose genomes have been sequenced to date [142, 143]. At the time that I started to work on conserved cassette exon prediction in 2006, there were only seven species with over a million ESTs in dbEST – this has now increased to 13 species (dbEST data from http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html, as of November 13, 2009), but the fact remains that many species have very low EST coverage. Moreover, even for a reasonably well characterized organism such as human, novel AS events are continually detected with the addition of more ESTs – for instance, while TassDB1 [121] had 1,945 cases of confirmed NAGNAG AS in human, TassDB2 (manuscript under review) has 2,272 such cases . Furthermore, the majority of human ESTs have been sampled from cancer tissue, which means that there might be AS events (possibly even dominant ones) in normal tissue which are as yet undetected [88]. The number of ESTs available for different tissues also differs substantially – for example, the collection of ESTs for humans in dbEST [143] has nearly eight times as many brain-derived ESTs as heart-derived ESTs [144]. Moreover, EST datasets are biased towards highly expressed variants, as the expression level of a transcript must be sufficiently high for it to be detected within the settings of a particular EST sequencing effort. Thus, minor isoforms can escape detection, even though they might be quite important in the cell [96, 145]. The single read nature of ESTs also makes them error prone [146] and difficult to interpret. AS can be condition/tissue/developmental stage-specific [59], which means that it shall go undetected until and unless ESTs are sampled under the appropriate conditions. All these reasons provide a strong motivation for other methods of detecting and predicting AS. Microarrays are another popular experimental method of large scale detection of AS, which have achieved a more uniform coverage of tissues than ESTs [147]. However, they have limitations stemming from hybridization related artifacts [148], design constraints that limit the detection of RNA splice patterns, and a limited ability to distinguish between closely related isoforms. Thus computational methods of AS prediction have much to contribute. Finally, computational predictions on species different from the species on which data the method was trained can give insights how conserved the underlying splicing mechanism between species is.

However, it must be asked how the advent of next-generation sequencing technologies impacts the situation. On the one hand, these technologies make it possible to sample transcriptomes at unprecedented depth, enabling the discovery of novel exons and AS events which were not detected using ESTs derived by Sanger technology alone [27, 28, 98, 99, 149-151]. On the other hand, shorter read length in comparison to Sanger sequencing means that events involving longer sequences are harder to detect using these approaches. This is especially true of (earlier versions of) technologies such as Illumina/Solexa and ABI/SOLiD, with read lengths of 25-36 nt in most studies reported to date. The situation is likely to change as these companies work to make longer read lengths possible, in fact it has already changed to some extent, with paired end reads as well read lengths of 50-100 bp becoming available in the latest versions of these platforms. Roche's 454 sequencing platform is less limited in this sense, with current read lengths of ~250-400 nt, but few AS studies have been published so far using this platform due to the higher costs/nt of this technology. But with over 200 cell types, several developmental stages, and many other physiological conditions under which AS can occur, it is still infeasible to cover all the specific combinations via deep sequencing approaches. It must also be pointed out that next-generation sequencing platforms are still too expensive for many laboratories and institutes, which means that scientists working in these places continue to use Sanger ESTs, mRNAs, full-length cDNAs, RT-PCR and microarrays to detect AS. Therefore, it is clear that computational methods of predicting AS continue to be useful.

A quite independent reason which makes computational prediction of AS useful is that it can help shed further light on the mechanisms involved in AS. As an example, in our work on conserved cassette exon prediction, we managed to show that the majority of exons which were labeled constitutive but assigned a probability ≥ 0.7 of being alternative by the BN were, in fact, alternative according to the latest transcript evidence. But what about the exons which were labeled alternative, yet received a low probability of being alternative by the BN? These exons were clearly more like constitutive exons in terms of the features we used for prediction, yet they were known to be alternative – why might that be so? By examining such cases in the genome browser, at least two clear patterns emerged:

(1) There were some conserved cassette exons which had predicted conserved secondary structure in their splice site neighborhoods. This suggested a mechanism in which exons which are otherwise like constitutive exons, get alternatively spliced due to secondary

structure which somehow interfere with the process of splicing. This view is supported by a recent paper which reports that conserved secondary structures promote alternative splicing [152].

(2) There are exons which resemble constitutive exons in most of the features, but have an extremely large or small value for some important feature(s) which explain why they are alternatively spliced. For instance, they might have an unusually long upstream intron (say > 20 kb long, which is longer than 95% of human internal introns [153]). It is known that the exon-intron architecture of a gene plays an important role in determining whether it shall be alternatively spliced, and long flanking introns make AS much likelier [38, 40]. If the orthologous exon also has an unusually long upstream intron, it may be alternatively spliced as well, giving rise to conserved AS which is due to architectural constraints rather than conserved regulators.

Thus, by studying where our computational method goes wrong, we pick up useful hints about alternative mechanisms of conserved AS.

Figure 4 shows the distribution of posterior probabilities of human-mouse conserved exons being alternative for constitutive as well as alternative exons. Ideally, the probabilities should be very low for constitutive exons, and very high for alternative exons. The distribution for the constitutive exons is not far from the ideal case, with a clear peak in the region 0-0.25. However, while the main peak for alternative exons is in the region 0.75-1, there is a smaller peak in 0.1-0.2 region, and a plateau in the 0.25-0.65 region. These curves explain why the specificity of our method is quite high, but the sensitivity can be improved further. They also suggest that there are potentially many more undiscovered conserved cassette exons in human and mouse.

It is also worth noting that the lessons learned from computational approaches to AS prediction can also be applied to study the wealth of data being produced by next-generation sequencing. An important recent study reported “switch-like” exons, which are alternatively spliced exons whose inclusion levels vary remarkably across tissues [27]. Such exons seem to share several properties with conserved cassette exons (also called alternative conserved exons, or ACEs, in [100] – we use the term "conserved cassette exon" as there are other kinds of alternative conserved exons, for example conserved alternative acceptors and donors, that we did not consider). Thus it is not unnatural to think of developing computational methods to identify and characterize “switch like” exons, and other kinds of alternative exons which are likely to be of functional importance.

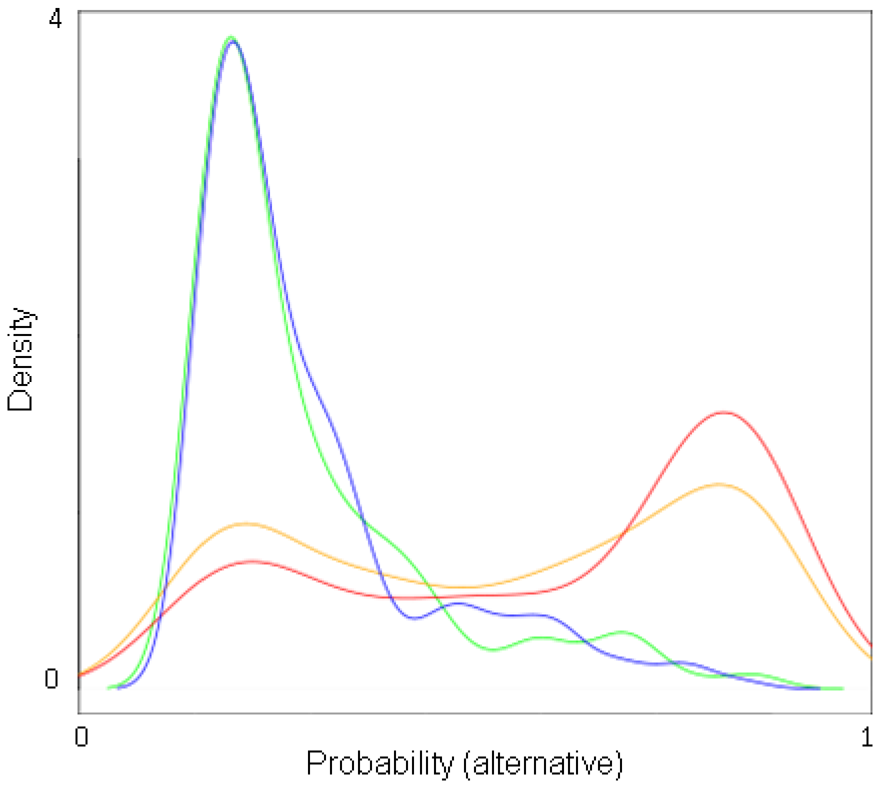


Figure 4. Posterior probabilities of being alternative for constitutive and alternative exons.
The distribution of posterior probabilities of being alternative, for constitutive exons (D1 – green, D2 - blue) and alternative exons (D1 – red, D2- orange) of the datasets D1 (from [102]) and D2 (from [100]) used in our study [108].

One of the most encouraging results in my work on predicting conserved cassette exons was that when I inspected the exons which were labeled as constitutive by ESTs available at the time of BN training but received a probability ≥ 0.7 of being alternative, the majority (20/29) of them turned out be alternative according to the latest transcript evidence. This further strengthened the claim that our classifier was able to distinguish constitutive exons from alternative exons with high accuracy. This also suggests that one could devise a stringent test for classifiers by deliberately mislabeling a few samples, and check whether the classifier still classifies them correctly.

Ideally, we should be able to predict AS without having to rely on conservation based features, that is, using only information which is in the given genome or transcriptome, and thus in principle available to the spliceosome. While we took a promising step in this direction by achieving $TP_{0.5}$ = 29% without using conservation, this is only about half the performance achieved when including conservation-based features. Furthermore, while conserved AS events are a very important subclass of events by virtue of being enriched in likely functional events, they are only a minority of AS events according to the current literature – up to 50% of all human AS exons may be human-specific, that is, not conserved in other species, and of the conserved exons, up to 60% of the AS exons may be alternative in a species-specific way [154]. Thus, prediction and in general further characterization of both these classes of AS exons is of great interest.

Accurate prediction of NAGNAG alternative splicing

Several studies in the recent years have shown that a substantial fraction of alternative acceptors (AA) and alternative donors (AD) are separated by a relatively small distance – these are instances of so-called "subtle alternative splicing" [35, 118-120]. NAGNAG AS, involving acceptors separated by 3 nt, is the most widespread subclass of subtle AS [35, 119, 120]. The true extent of subtle AS is still unknown, as many regions harboring tandem splice sites are not well covered by ESTs – for example, while 16% (1,815 of 10,740) of human NAGNAG acceptors are alternatively spliced according to the data present in TassDB1 [121], 40% (3,562) of the remaining NAGNAG acceptors have less than ten ESTs each. Thus a subset of these NAGNAGs, rather than being truly constitutive, may simply lack evidence of AS due to insufficient sampling of the transcriptome. An accurate predictive method can provide a meaningful estimate of how many alternative NAGNAG acceptors remain to be discovered. While previous work on predicting alternative 3' splicing had reported good results overall, the results for predicting NAGNAG AS were modest, in comparison to cases involving larger distances [122]. This is at odds with the notion that a simple model based on splice site strength suffices to explain NAGNAG and other short-distance tandem AS [123]. Under such a model, NAGNAG and other short distance AS is mainly caused by a spliceosome which "slips" occasionally, and is the outcome of stochastic selection between two nearby, competing alternative splice sites. This implies that subtle AS at many NAGNAG AAs is a noisy rather than a regulated process [123, 124] – nevertheless, this does not rule out the existence of a small subset of NAGNAG AAs where the noise is "cultivated" to provide stable, functionally important isoform ratios (reviewed in [136]).

. Motivated by these reasons, I used BNs and TassDB1 to develop an approach to accurately predicting the splicing outcome at NAGNAG acceptors. I treated the problem of predicting NAGNAG splicing outcome as a 3-class prediction problem, corresponding to the three possible outcomes (E, I, or EI), instead of two 2-class. problems (E versus EI and I versus EI), as was done earlier [122]. 42 features were derived from the NAGNAG and its surrounding region (positions -20 to +3), as well as the last 3 positions of the upstream exon (due to the possible impact on the protein context). These features included the splice site strengths of the two putative acceptors, the sequence composition of the region, and the lengths of the flanking exons and introns, as well as features capturing information about the pyrimidine content of the PPT [125]. To begin with, I worked with human data, and achieved a balanced sensitivity and specificity of ≥ 92%. This high *in-silico* performance gave us the confidence to perform experimental validation, with 63 NAGNAGs being investigated in one of the most extensive validations reported to date. Our predictions agreed with 80% (50/63) of the cases, with the agreement rising to 89% (25/28) for cases where AS was predicted with a probability ≥ 90%. We also showed that a higher predicted posterior probability corresponds to more accurate predictions, and that relaxing the threshold on the abundance of the minor isoform results in higher accuracy of AS prediction, as expected.

One of the reasons behind the success of our approach to NAGNAG AS prediction was the careful preparation of training and test datasets. Since it is known that EST data can contain artifacts [146], I took care to include only those NAGNAGs in the training set, whose categorization based on EST data was very likely to be correct. Specifically, to call a NAGNAG constitutive, we required a support of ≥ 10 ESTs for one of the splice sites in the tandem, and none for the other; while to label a NAGNAG as alternative, we required a support of ≥ 2 ESTs for each variant, and a minor abundance threshold of 10%, that i s, ≥ 10% of the ESTs covering the NAGNAG had to support the minor isoform. These stringent criteria decreased the chances of erroneous inclusion of “constitutive” exons which were in fact alternative but merely poorly covered by ESTs, as well as exons which appeared to be alternative due to a single, erroneous EST, and thus significantly reduced the noise in the dataset. We then kept all the remaining NAGNAGs in the training set. Apart from the performance of AUC = 0.94, 0.98 and 0.96 for the EI, E and I forms, respectively, we were also able to show that our criteria could help improve the performance on a dataset from the literature [122]. By comparison with TassDB1, it was evident that the main reason for the lower performance in NAGNAG AS prediction reported in [122] was the contamination of the set of constitutive NAGNAGs by several NAGNAGs which were in fact alternative (as documented in TassDB1), but lacked sufficient transcript coverage at the time the dataset was compiled. This improvement was independently verified by the use of next generation sequencing data from the Solexa platform described in [27].

One of the most interesting results in my work on NAGNAG AS prediction was the high accuracy achieved on the mouse, rat, dog and zebrafish NAGNAGs by a BN which had been trained using only human NAGNAG data. Since this classifier achieved results which were nearly identical to those achieved on human data itself, this was a strong hint that the mechanism behind NAGNAG splicing is conserved for 450 My among vertebrates. This also shows that our classifier should be able to predict NAGNAG splicing outcomes in vertebrate genomes that currently lack extensive transcript data. Since the same classifier also achieved a good performance on fruit fly and *C. elegans* – albeit with a drop as compared to prediction on vertebrate genomes – this suggests that the mechanism is conserved beyond vertebrates.

The most informative features in NAGNAG AS prediction were the two Ns in the NAGNAG motif, and the splice site scores. The scores for the upstream and downstream splice sites, and the upstream and downstream Ns can be substituted by each other with nearly identical results. The next important features were the nucleotides immediately neighboring the NAGNAG, while other features made only small contributions to the prediction performance. This shows that most of the information required for NAGNAG splicing prediction is encoded in the immediate splice site neighborhood. Moreover, simpler approaches like using only the two Ns in the NAGNAG motif, or only the splice site scores (computed by MAXENTSCAN [155]), or using a naïve Bayes classifier, worsened performance only slightly, indicating that the other features and the corresponding dependencies learned by the BN are weak in their discriminative power as well as in generalization to other datasets. All this further strengthens the view that the mechanism behind NAGNAG splicing is simple in nature. Taken together, our results show that BNs can produce highly reliable predictions of NAGNAG splicing outcomes, and that the mechanism behind NAGNAG splicing is simple in nature, and maintained during the course of evolution.

Characterization and prediction of NAGNAG alternative splicing in the moss *Physcomitrella patens*

In contrast to the remarkable rise in the number of studies of AS in animals in the last decade, the study of AS in plants is still in its initial stage. This also means that some of the lessons learned from analyzing AS in animals can be exploited when studying AS in plants [31]. It appears that AS in plants, while prevalent, is only about one-third to half as frequent as in animals [31] – while AS in vertebrates is estimated to occur in 50%-75% of genes according to EST and microarray data [26, 32, 34, 45, 79], the corresponding estimates for plants are 20%-35% [29, 30, 126]. The estimates using deep coverage of transcriptomes by the RNA-seq approach using next-generation sequencing technologies is > 90% for human [27] and 42%-56% for Arabidopsis [127]. While it is likely that the estimates of AS abundance in plants will keep rising with more comprehensive characterization of the transcriptomes of various tissues as well as under the diverse conditions under which plants respond to environmental stress, the frequency of AS in plants may nevertheless be less than that in animals. One possible reason for this might be the tendency of plants to undergo genome duplication and/or polyploidization [31]. Like AS, genome duplication (followed by divergence) is also a potential source of proteomic diversity, and some studies have observed an inverse correlation between gene family size and AS frequency in animals [156, 157]. This suggests that the higher rate of duplications in plants might be related to a lower rate of AS when compared to animals [31]. It must however be noted that AS seems to have a more drastic effect on protein sequence and structure than does genome duplication followed by divergence [158].

While there have been papers on the detection of AS in plants in the recent years [29, 30, 126, 127], there are has not been much work on computational prediction of AS in plants – to the best of our knowledge, our work on predicting NAGNAG AS in moss is one of the first such studies (manuscript submitted). The model moss *Physcomitrella patens* was the first bryophyte genome to be sequenced, and the distribution of AS events in it seemed similar to other plants studied so far [128]. Since it was known that NAGNAG AS is widespread in rice and *Arabidopsis Thaliana* [30], I sought to extend our approach to NAGNAG AS prediction to plants by using moss as a case study. Since the extent of NAGNAGs in moss had not been reported, this work involved both characterization as well prediction of NAGNAG splicing in moss. We found that NAGNAG AS is also widespread in moss – out of 9,427 NAGNAG acceptors in *P. patens*, 5,031 have Sanger EST coverage, with 295 (5.9%) being alternatively spliced (EI form), 2,695 (53.6%) constitutively spliced at the first (intron proximal) acceptor (E form), and 2,041 (40.5%) constitutively spliced at the second (intron distal) acceptor (I form). Thus, NAGNAG AS is common in *Physcomitrella*. Using a feature set similar to that used in my work on NAGNAG AS prediction in animals, an in-silico performance of AUC = 0.96, 0.99 and 0.99 was achieved for the EI, E and I forms, respectively. In the course of our work, 454 data were generated for the moss transcriptome, to supplement our characterization of NAGNAGs in moss using evidence for this next generation sequencing platform. Even though the 454 data covers only 75% (3,745/5,031) of the genomic NAGNAGs covered by ESTs, it helps us to detect 371 alternative NAGNAGs – 9.9% of the covered NAGNAGs, compared to 7.5% using ESTs. Furthermore, the 454 data provides support for 73 alternative NAGNAGs among gene models not covered by Sanger ESTs. Combining the results from ESTs and 454 data, *Physcomitrella patens* has 664 alternative NAGNAGs.

Since I trained the classifier using Sanger ESTs, I was able to use the 454 data for independent validation of the predictions of the classifier. When considering AS according to 454 reads alone, 64% (80/125) of the well-supported (≥ 2 reads per variant, ≥ 10% of the reads support the minor variant) cases of AS were predicted correctly, which increased to 79% (30/38) if we required ≥ 4 reads per variant while keeping the threshold of minor variant abundance at ≥ 10%. The corresponding numbers were 62% (41/66) and 75% (9/12) when I considered the well-supported additional NAGNAG AS events detected using the combined Sanger EST and 454 data. Altogether, our results show that NAGNAG AS is widespread in moss, and the mechanism behind NAGNAG AS in plants seems very similar to that in animals. The pervasiveness of NAGNAG AS suggests that it may be a general feature of splicing in animals and plants, and possibly in all intron containing organisms.

Assessing the conservation of tandem splice sites

While the abundance of subtle AS events is well accepted [118-120, 132], their functional impact has been the matter of debate [123, 124]. Briefly, the two opposing views can be summarized as “Subtle alternative splicing is a means of fine-tuning the transcriptome, and thereby the proteome”, versus “It’s not subtle, it’s noise – the spliceosome may ‘slip’ when confronted with competing alternatives in close vicinity of each other” –however, both views can be accommodated by considering that noise refers to the mechanism, whereas function (or the lack thereof) depends on how the species makes use of the AS event. The fourth work of my thesis [159] sought to address this question by studying the conservation of subtle splicing. It was known that NAGNAG splicing is biased in terms of the motif [118, 138]. In particular, while YAGYAG (Y = C or T) motifs are often alternative, GAGHAG and HAGGAG (H = A, C, or T) are only very rarely alternative. This is due to the fact that GAGs rarely function as acceptors, while YAGs are strong acceptors, and there is a distinct order of preference for the nucleotide preceding the AG in the acceptor which is C > T > A > G. At the genomic level, HAGGAGs are by far the most common NAGNAG motif, in keeping the 3’ splice site motif. Thus, any analysis which does not take the biased distribution of NAGNAG motifs into account is heavily influenced by the dominant HAGGAGs. To control for bias, we looked at the conservation of the 16 different NAGNAG motifs, and found that alternative NAGNAGs are in fact more conserved than constitutive NAGNAGs. This is an example of Simpson’s paradox [140], in which splitting two groups into subgroups reverses observed trends. By taking biases into account, we could arrive at an estimate of the fraction of subtle AS events under selection.

Interestingly, both frame-preserving as well as frame-shifting tandems are under selection, even though frame-shifting tandems could often have deleterious consequences, by introducing a PTC. One possible explanation is that frame-shifting tandems could play role in creating truncated proteins, another is that they could cause the transcripts to be degraded via the NMD pathway [82], thus regulating the protein level. The coupling of AS and NMD is very frequent, and the term RUST (Regulated Unproductive Splicing and Translation) [81] has been coined to refer to regulation of gene expression via this coupling. RUST has been found to occur in splicing factor genes, in highly as well as ultraconserved regions [160, 161] (ultraconserved regions are defined as at least 200-nt-long regions that are identical between human, mouse, and rat [162]), and at least 21% of conserved cassette exons also lead to an NMD-inducing transcript [163]. Our estimates for the fraction of tandem sites under selection are only a lower bound, because as shown in our work on predicting NAGNAG AS, sets of constitutive exons are often contaminated by as-yet-undiscovered alternative exons. Therefore, such undetected alternative exons can contribute to increasing the background conservation level for constitutive exons, thus reducing the estimated fraction of alternative exons under selection. Finally, while tandem sites under selection are very likely to be functional, it must be noted that functional differences have also been revealed by experiments and/or predicted by computational approaches for tandem and other AS events which are species-specific [154, 164]. Thus, conserved AS events are in turn themselves a lower bound on the number of functional AS events.

TassDB2 – a comprehensive resource on tandem splice sites

Subtle alternative splice events involving tandem splice sites separated by a short (2-12 nt) distance are frequent and evolutionarily widespread in eukaryotes, and a major contributor to the complexity of transcriptomes and proteomes [35, 118-120, 134]. However, these events have been either omitted altogether in databases on alternative splicing, or only the cases of confirmed alternative splicing have been reported. Thus, a database which covers all confirmed cases of subtle alternative splicing as well as the numerous putative tandem splice sites (which might be confirmed once more transcript data becomes available), and allows to search for tandem splice sites with specific features and download the results, can be a valuable resource for targeted experimental studies and large-scale bioinformatics analyses of tandem splice sites.

We have substantially revised and extended TassDB1 [121], which stores extensive data about alternative splice events at tandem splice sites separated by 3 nt, in eight species [121]. TassDB2 (manuscript under review) contains information about tandem splice sites separated by 2-12 nt for the human and mouse transcriptomes. Thus, TassDB2 provides comprehensive information on 22 event types, compared to 2 (NAGNAG and GYNGYN) in TassDB1. TassDB2 is therefore effectively a new database rather than just a simple extension. TassDB2 offers a user-friendly interface to search for specific genes or for genes containing tandem splice sites with specific features as well as the possibility to download search results and large datasets. For example, the users can search for cases of alternative splicing where the proportion of EST/mRNA evidence supporting the minor isoform exceeds a specific threshold, or where the difference in splice site scores is specified by the user. The predicted impact (if any) of each event on the protein sequence is also reported, along with information about being a putative target for the nonsense-mediated decay (NMD) pathway [82]. Links are provided to the UCSC genome browser [165] and other external resources.

TassDB2 is a comprehensive resource for information regarding subtle alternative splicing. Users can easily search for individual genes, as well as by various criteria corresponding to different features of the tandem splice sites. Some of the criteria can be used to enrich for splicing events which are likely to have functional significance. The results can be downloaded for further exploration, and flat files have also been made available for those who wish to carry out their own large-scale bioinformatics studies. Thus TassDB2 should be a very useful resource for scientists interested in subtle alternative splicing.

Outlook

In my doctoral research on the bioinformatics analysis of alternative splicing, I have mostly focused on prediction of AS using Machine Learning and features derived from splice-relevant regions, and subtle AS. The field of research on AS, which was already very active in 2006, has grown further and changed rapidly in the three-and-a-half years of my research. Perhaps the biggest change has come about due to the advent of next-generation sequencing technologies, since they readily generate vast quantities of transcript data under controlled conditions, thus removing several limitations associated with Sanger EST data. At the time I started my doctoral research in 2006, analysis of ESTs was the most common approach to large-scale detection of AS, and had several limitations in terms of coverage, data quality, knowledge about the conditions under which data was collected, and so on. The new sequencing technologies remove these limitations, and as read length gets longer, Sanger ESTs may soon become less relevant in cases where RNA-seq data with a comparable or even longer read length is available.

Thus, new questions are arising in the light of new data. It would be interesting to reassess many of the results regarding the various properties distinguishing AS exons from constitutive ones, as use of the new large-scale datasets should reduce contamination of datasets of constitutive exons by undiscovered alternative exons, which means that more meaningful comparisons can be performed. We are now moving into a phase where AS is being analyzed at a new level, where instead of trying to characterize the transcriptome of an organism, we have enough data to actually study the various transcriptomes a given organism has, corresponding to different tissues, developmental stages, environmental stresses, and so on.

The availability of ever increasing amounts of data means that the over-dependence on conservation as a filter for finding likely cases of functional AS is also decreasing. This is a welcome development, because while conservation as an indicator of function is the bedrock of comparative genomics, we cannot hope to understand the physiological differences between species without studying what is lineage-specific yet functional [166]. Since it is estimated that the majority of human alternative exons are not conserved in mouse, and further that the majority of exons which are conserved and alternative, are alternative in a species-specific manner [154], it is natural to expect that AS contributes to species-specific differences. Therefore it is critical to develop conservation-independent methods of assessing the functionality of AS. A start has been made recently - for instance, deep coverage of transcriptomes allows us to estimate the fractions of genes and exons which are expressed in a tissue or condition-biased manner, which is another way of inferring functionality. Furthermore, predictions and studies of motifs involved in AS can also take advantage of the recently available and rapidly accumulating data to focus on regulation under specific conditions, tissues and so forth. Recent studies have started providing “RNA binding maps” for important RNA-binding proteins and regulators of AS and other processes, thus opening doors to new areas of detailed knowledge about the regulation of transcription and splicing [55, 57].

Another exciting and emerging area concerns the coupling of various fundamental processes. For instance, transcription and splicing are often coupled, and the rate of transcription can have profound effects on splicing [167]. A recent study suggests that the coupling of transcription to AS might be a key feature of the DNA-damage response [168]. Very recent work shows that epigenetics and splicing may also be interrelated, as nucleosome occupancy seems to be a strong discriminator between introns and exons [169, 170]. We are in the midst of an exciting era in biological research and research in transcriptional, co- and post-transcriptional regulation in particular. The quest to crack the splicing code, a continuing endeavour, is also a part of cracking the regulatory code itself.

“I can’t be as confident about computer science as I can about biology. Biology easily has 500 years of exciting problems to work on. It’s at that level.” – Donald Knuth (Computer Literacy Booksops Interview, 1993)

BIBLIOGRAPHY

1. Gilbert W: **Why Genes in Pieces**. *Nature* 1978, **271**(5645):501-501.
2. Crick F: **Central Dogma of Molecular Biology**. *Nature* 1970, **227**(5258):561-&.
3. Berget SM, Moore C, Sharp PA: **Spliced segments at the 5 ' terminus of adenovirus 2 late mRNA** *Proc Natl Acad Sci USA* 1977, **74**(8):3171-3175.
4. Chow LT, Gelinas RE, Broker TR, Roberts RJ: **Amazing Sequence Arrangement at 5' Ends of Adenovirus-2 Messenger-Rna**. *Cell* 1977, **12**(1):1-8.
5. Lerner MR, Boyle JA, Mount SM, Wolin SL, Steitz JA: **Are Snrnps Involved in Splicing**. *Nature* 1980, **283**(5743):220-224.
6. Rappsilber J, Ryder U, Lamond AI, Mann M: **Large-scale proteomic analysis of the human splicesome**. *Genome Research* 2002, **12**(8):1231-1245.
7. Jurica MS, Moore MJ: **Pre-mRNA Splicing: Awash in a Sea of Proteins**. *Molecular Cell* 2003, **12**(1):5-14.
8. Wahl MC, Will CL, Luhrmann R: **The Spliceosome: Design Principles of a Dynamic RNP Machine**. *Cell* 2009, **136**(4):701-718.
9. Burge CB, Padgett RA, Sharp PA: **Evolutionary fates and origins of U12-type introns**. *Molecular Cell* 1998, **2**(6):773-785.
10. Patel AA, Steitz JA: **Splicing double: Insights from the second spliceosome**. *Nature Reviews Molecular Cell Biology* 2003, **4**(12):960-970.
11. Abril JF, Castelo R, GuigÃ³ R: **Comparison of splice sites in mammals and chicken**. *Genome Research* 2005, **15**(1):111-119.
12. Burset M, Seledtsov IA, Solovyev VV: **SpliceDB: database of canonical and non-canonical mammalian splice sites**. *Nucl Acids Res* 2001, **29**(1):255-259.
13. Zhang MQ: **Statistical features of human exons and their flanking regions**. *Hum Mol Genet* 1998, **7**(5):919-932.
14. Breathnach R, Chambon P: **Organization and Expression of Eukaryotic Split Genes-Coding for Proteins**. *Annu Rev Biochem* 1981, **50**:349-383.
15. Chen M, Manley JL: **Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches**. *Nat Rev Mol Cell Biol* 2009, **advance online publication**.
16. Cartegni L, Chew SL, Krainer AR: **Listening to silence and understanding nonsense: exonic mutations that affect splicing**. *Nat Rev Genet* 2002, **3**(4):285-298.
17. Dietrich RC, Peris MJ, Seyboldt AS, Padgett RA: **Role of the 3 ' splice site in U12-dependent intron splicing**. *Molecular and Cellular Biology* 2001, **21**(6):1942-1952.
18. Dietrich RC, Fuller JD, Padgett RA: **A mutational analysis of U12-dependent splice site dinucleotides**. *Rna-a Publication of the Rna Society* 2005, **11**(9):1430-1440.
19. Black DL: **Mechanisms of alternative pre-messenger RNA splicing**. *Annu Rev Biochem* 2003, **72**:291-336.
20. Zamore PD, Green MR: **Identification, Purification, and Biochemical-Characterization of U2 Small Nuclear Ribonucleoprotein Auxiliary Factor**. *Proceedings of the National Academy of Sciences of the United States of America* 1989, **86**(23):9243-9247.
21. Berglund JA, Chua K, Abovich N, Reed R, Rosbash M: **The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC**. *Cell* 1997, **89**(5):781-787.
22. Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L: **Two mRNAs can be produced from a single immunoglobulin [mu] gene by alternative RNA processing pathways**. *Cell* 1980, **20**(2):313-319.
23. Mironov AA, Fickett JW, Gelfand MS: **Frequent Alternative Splicing of Human Genes**. *Genome Res* 1999, **9**(12):1288-1293.

24. Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Bork P: **EST comparison indicates 38% of human mRNAs contain possible alternative splice forms**. *FEBS Letters* 2000, **474**(1):83-86.
25. Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes**. *Nucl Acids Res* 2001, **29**(13):2850-2859.
26. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays**. *Science* 2003, **302**(5653):2141-2144.
27. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.
28. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nat Genet* 2008, **40**(12):1413-1415.
29. Wang B-B, Brendel V: **Genomewide comparative analysis of alternative splicing in plants**. *PNAS* 2006, **103**(18):7175-7180.
30. Campbell M, Haas B, Hamilton J, Mount S, Buell CR: **Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis**. *BMC Genomics* 2006, **7**(1):327.
31. Barbazuk WB, Fu Y, McGinnis KM: **Genome-wide analyses of alternative splicing in plants: Opportunities and challenges**. *Genome Research* 2008, **18**(9):1381-1392.
32. Kim E, Magen A, Ast G: **Different levels of alternative splicing among eukaryotes**. *Nucl Acids Res* 2007, **35**(1):125-131.
33. Takeda J-i, Suzuki Y, Nakao M, Barrero RA, Koyanagi KO, Jin L, Motono C, Hata H, Isogai T, Nagai K *et al*: **Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56 419 completely sequenced and manually annotated full-length cDNAs**. *Nucl Acids Res* 2006, **34**(14):3917-3928.
34. Takeda J-i, Suzuki Y, Nakao M, Kuroda T, Sugano S, Gojobori T, Imanishi T: **H-DBAS: Alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational**. *Nucl Acids Res* 2007, **35**(suppl_1):D104-109.
35. Sugnet CW, Kent WJ, Jr. AM, Haussler D: **Transcriptome and Genome Conservation of Alternative Splicing Events in Humans and Mice**. *Pacific Symposium on Biocomputing* 2004, **9**:66-77.
36. Kan Z, States D, Gish W: **Selecting for Functional Alternative Splices in ESTs**. *Genome Res* 2002, **12**(12):1837-1845.
37. Schindler S, Szafranski K, Hiller M, Ali G, Palusa S, Backofen R, Platzer M, Reddy A: **Alternative splicing at NAGNAG acceptors in Arabidopsis thaliana SR and SR-related protein-coding genes**. *BMC Genomics* 2008, **9**(1):159.
38. Fox-Walsh KL, Dou Y, Lam BJ, Hung S-p, Baldi PF, Hertel KJ: **The architecture of pre-mRNAs affects mechanisms of splice-site pairing**. *PNAS* 2005, **102**(45):16176-16181.
39. Lander ES, Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al.: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
40. Sterner DA, Carlo T, Berget SM: **Architectural limits on split genes**. *PNAS* 1996, **93**(26):15081-15085.

41. Berget SM: **Exon recognition in vertebrate splicing**. *J Biol Chem* 1995, **270**:2411 - 2414.
42. Lim SR, Hertel KJ: **Commitment to splice site pairing coincides with a complex formation**. *Molecular Cell* 2004, **15**(3):477-483.
43. Kotlajich MV, Crabb TL, Hertel KJ: **Spliceosome Assembly Pathways for Different Types of Alternative Splicing Converge during Commitment to Splice Site Pairing in the A Complex**. *Molecular and Cellular Biology* 2009, **29**(4):1072-1082.
44. Holste D, Huo G, Tung V, Burge CB: **HOLLYWOOD: a comparative relational database of alternative splicing**. *Nucl Acids Res* 2006, **34**(suppl_1):D56-62.
45. Kim N, Alekseyenko AV, Roy M, Lee C: **The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species**. *Nucl Acids Res* 2007, **35**(suppl_1):D93-98.
46. 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* 2005, **33**(13):4276 - 4284.
47. Lee Y, Lee Y, Kim B, Shin Y, Nam S, Kim P, Kim N, Chung WH, Kim J, Lee S: **ECgene: an alternative splicing database update**. *Nucleic Acids Res* 2007, **35**(Database issue):D99 - 103.
48. Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, Tang Y, Barbosa-Morais NL, Thanaraj TA: **ASD: a bioinformatics resource on alternative splicing**. *Nucleic Acids Res* 2006, **34**(Database issue):D46 - 55.
49. Lim LP, Burge CB: **A computational analysis of sequence features involved in recognition of short introns**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(20):11193-11198.
50. Smith CWJ, Valcárcel J: **Alternative pre-mRNA splicing: the logic of combinatorial control**. *Trends in Biochemical Sciences* 2000, **25**(8):381-388.
51. Dreyfuss G, Kim VN, Kataoka N: **Messenger-RNA-binding proteins and the messages they carry**. *Nat Rev Mol Cell Biol* 2002, **3**(3):195-205.
52. Long JC, Caceres JF: **The SR protein family of splicing factors: master regulators of gene expression**. *Biochem J* 2009, **417**(1):15-27.
53. Graveley BR: **Sorting out the complexity of SR protein functions**. *RNA* 2000, **6**(09):1197-1211.
54. Tacke R, Manley JL: **Determinants of SR protein specificity**. *Current Opinion in Cell Biology* 1999, **11**(3):358-362.
55. Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, Gaasterland T, Blencowe BJ, Darnell RB: **An RNA map predicting Nova-dependent splicing regulation**. *Nature* 2006, **444**(7119):580-586.
56. Hui J, Hung L-H, Heiner M, Schreiner S, Neumuller N, Reither G, Haas SA, Bindereif A: **Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing**. *EMBO J* 2005, **24**(11):1988-1998.
57. Yeo GW, Coufal NG, Liang TY, Peng GE, Fu X-D, Gage FH: **An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells**. *Nat Struct Mol Biol* 2009, **16**(2):130-137.
58. Mauger DM, Lin C, Garcia-Blanco MA: **hnRNP H and hnRNP F Complex with Fox2 To Silence Fibroblast Growth Factor Receptor 2 Exon IIIc**. *Mol Cell Biol* 2008, **28**(17):5403-5419.
59. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H: **Function of alternative splicing**. *Gene* 2005, **344**:1 - 20.
60. Barbaux S, Niaudet P, Gubler MC, Grunfeld JP, Jaubert F, Kuttenn F, Fekete CN, SouleyreauTherville N, Thibaud E, Fellous M *et al*: **Donor splice-site mutations in WT1 are responsible for Frasier syndrome**. *Nature Genetics* 1997, **17**(4):467-470.

61. Buratti E, Baralle M, Baralle FE: **Defective splicing, disease and therapy: searching for master checkpoints in exon definition**. *Nucl Acids Res* 2006, **34**(12):3494-3510.

62. Burgar HR, Burns HD, Elsden JL, Lalioti MD, Heath JK: **Association of the signaling adaptor FRS2 with fibroblast growth factor receptor 1 (Fgfr1) is mediated by alternative splicing of the juxtamembrane domain**. *Journal of Biological Chemistry* 2002, **277**(6):4018-4023.

63. Hammes A, Guo JK, Lutsch G, Leheste JR, Landrock D, Ziegler U, Gubler MC, Schedl A: **Two splice variants of the Wilms' tumor 1 gene have distinct functions during sex determination and nephron formation**. *Cell* 2001, **106**(3):319-329.

64. Hu CA, Lin WW, Obie C, Valle D: **Molecular enzymology of mammalian Delta(1)-pyrroline-5-carboxylate synthase - Alternative splice donor utilization generates isoforms with different sensitivity to ornithine inhibition**. *Journal of Biological Chemistry* 1999, **274**(10):6754-6762.

65. Hymowitz SG, Compaan DM, Yan MH, Wallweber HJA, Dixit VM, Starovasnik MA, de Vos AM: **The crystal structures of EDA-A1 and EDA-A2: Splice variants with distinct receptor specificity**. *Structure* 2003, **11**(12):1513-1520.

66. Lareau LF, Green RE, Bhatnagar RS, Brenner SE: **The evolving roles of alternative splicing**. *Curr Opin Struct Biol* 2004, **14**:273 - 282.

67. Tsai K-W, Tseng H-C, Lin W-c: **Two wobble-splicing events affect ING4 protein subnuclear localization and degradation**. *Experimental Cell Research* 2008, **314**(17):3130-3141.

68. Vogan KJ, Underhill DA, Gros P: **An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity**. *Molecular and Cellular Biology* 1996, **16**(12):6677-6686.

69. Yan MH, Wang LC, Hymowitz SG, Schilbach S, Lee J, Goddard A, de Vos AM, Gao WQ, Dixit VM: **Two-amino acid molecular switch in an epithelial morphogen that regulates binding to two distinct receptors**. *Science* 2000, **290**(5491):523-527.

70. Tress M, Bodenmiller B, Aebersold R, Valencia A: **Proteomics studies confirm the presence of alternative protein isoforms on a large scale**. *Genome Biology* 2008, **9**(11):R162.

71. Resch A, Xing Y, Modrek B, Gorlick M, Riley R, Lee C: **Assessing the impact of alternative splicing on domain interactions in the human proteome**. *Journal of Proteome Research* 2004, **3**(1):76-83.

72. Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S: **Increase of functional diversity by alternative splicing**. *Trends Genet* 2003, **19**(3):124 - 128.

73. Xing Y, Xu Q, Lee C: **Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains**. *Febs Letters* 2003, **555**(3):572-578.

74. Cline MS, Shigeta R, Wheeler RL, Siani-Rose MA, Kulp D, Loraine AE: **The Effects of Alternative Splicing on Transmembrane Proteins in the Mouse Genome**. In: *Pacific Symposium on Biocomputing: 2004*; 2004: 17-28.

75. Ule J, Ule A, Spencer J, Williams A, Hu JS, Cline M, Wang H, Clark T, Fraser C, Ruggiu M *et al*: **Nova regulates brain-specific splicing to shape the synapse**. *Nature Genetics* 2005, **37**(8):844-852.

76. Wojtowicz WM, Flanagan JJ, Millard SS, Zipursky SL: **Alternative splicing of Drosophila Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding**. *Cell* 2004, **118**(5):619-633.

77. Lynch KW: **Consequences of regulated pre-mRNA splicing in the immune system**. *Nature Reviews Immunology* 2004, **4**(12):931-940.

78. Roberts AG, Redding SJ, Llewellyn DH: **An alternatively-spliced exon in the 5 '-UTR of human ALAS1 mRNA inhibits translation and renders it resistant to haem-mediated decay**. *Febs Letters* 2005, **579**(5):1061-1066.

79. Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason Pl, Albrecht M, Hegyi H, Giorgetti A *et al*: **The implications of alternative splicing in the ENCODE protein complement**. *PNAS* 2007, **104**(13):5495-5500.

80. Birzele F, Csaba G, Zimmer R: **Alternative splicing and protein structure evolution**. *Nucl Acids Res* 2008, **36**(2):550-558.

81. Lewis BP, Green RE, Brenner SE: **Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans**. *Proceedings of the National Academy of Sciences* 2003, **100**(1):189-192.

82. Maquat LE: **Nonsense-mediated mRNA decay in mammals**. *J Cell Sci* 2005, **118**(9):1773-1776.

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

84. Faustino NA, Cooper TA: **Pre-mRNA splicing and human disease**. *Genes Dev* 2003, **17**(4):419 - 437.

85. Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R: **Are splicing mutations the most frequent cause of hereditary disease?** *Febs Letters* 2005, **579**(9):1900-1903.

86. Lynch KW, Weiss A: **A CD45 polymorphism associated with multiple sclerosis disrupts an exonic splicing silencer**. *Journal of Biological Chemistry* 2001, **276**(26):24341-24347.

87. Kalnina Z, Zayakin P, Silina K, Line A: **Alterations of pre-mRNA splicing in cancer**. *Genes Chromosomes & Cancer* 2005, **42**(4):342-357.

88. Xu Q, Lee C: **Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences**. *Nucleic Acids Research* 2003, **31**(19):5635-5643.

89. Garcia-Blanco MA, Baraniak AP, Lasda EL: **Alternative splicing in disease and therapy**. *Nat Biotech* 2004, **22**(5):535-546.

90. Sazani P, Kole R: **Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing**. *Journal of Clinical Investigation* 2003, **112**(4):481-486.

91. Tazi J, Bakkour N, Stamm S: **Alternative splicing and disease**. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 2009, **1792**(1):14-26.

92. Marquis J, Meyer K, Angehrn L, Kampfer SS, Rothen-Rutishauser B, Schumperli D: **Spinal Muscular Atrophy: SMN2 Pre-mRNA Splicing Corrected by a U7 snRNA Derivative Carrying a Splicing Enhancer Sequence**. *Mol Ther* 2007, **15**(8):1479-1486.

93. Blencowe BJ: **Alternative Splicing: New Insights from Global Analyses**. *Cell* 2006, **126**(1):37-47.

94. Lee C, Wang Q: **Bioinformatics analysis of alternative splicing**. *Brief Bioinform* 2005, **6**(1):23-33.

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

96. Matos P, Collard JG, Jordan P: **Tumor-related Alternatively Spliced Rac1b Is Not Regulated by Rho-GDP Dissociation Inhibitors and Exhibits Selective Downstream Signaling**. *Journal of Biological Chemistry* 2003, **278**(50):50442-50448.

97. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank**. *Nucl Acids Res* 2008, **36**(suppl_1):D25-30.

98. Schuster SC: **Next-generation sequencing transforms today's biology**. *Nat Meth* 2008, **5**(1):16-18.

99. Mardis ER: **The impact of next-generation sequencing technology on genetics**. *Trends in Genetics* 2008, **24**(3):133-141.

100. Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB: **Identification and analysis of alternative splicing events conserved in human and mouse**. *PNAS* 2005, **102**(8):2850-2855.

101. Sorek R, Ast G: **Intronic Sequences Flanking Alternatively Spliced Exons Are Conserved Between Human and Mouse**. *Genome Res* 2003, **13**(7):1631-1637.

102. Dror G, Sorek R, Shamir R: **Accurate identification of alternatively spliced exons using support vector machine**. *Bioinformatics* 2005, **21**(7):897-901.

103. Raetsch G, Sonnenburg S, Scholkopf B: **RASE: recognition of alternatively spliced exons in C.elegans**. *Bioinformatics* 2005, **21**(suppl_1):i369-377.

104. Hiller M, Huse K, Platzer M, Backofen R: **Non-EST based prediction of exon skipping and intron retention events using Pfam information**. *Nucl Acids Res* 2005, **33**(17):5611-5621.

105. Leparc GG, Mitra RD: **Non-EST-based prediction of novel alternatively spliced cassette exons with cell signaling function in Caenorhabditis elegans and human**. *Nucl Acids Res* 2007, **35**(10):3192-3202.

106. Ohler U, Shomron N, Burge CB: **Recognition of Unknown Conserved Alternatively Spliced Exons**. *PLoS Computational Biology* 2005, **1**(2):e15.

107. Philipps DL, Park JW, Graveley BR: **A computational and experimental approach toward a priori identification of alternatively spliced exons**. *RNA* 2004, **10**(12):1838-1844.

108. Sinha R, Hiller M, Pudimat R, Gausmann U, Platzer M, Backofen R: **Improved identification of conserved cassette exons using Bayesian networks**. *BMC Bioinformatics* 2008, **9**(1):477.

109. Beaumont MA, Rannala B: **The Bayesian Revolution In Genetics**. *Nature Reviews Genetics* 2004, **5**(4):251-261.

110. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR: **Inference in Bayesian networks**. *Nat Biotech* 2006, **24**(1):51-53.

111. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR: **A Primer on Learning in Bayesian Networks for Computational Biology**. *PLoS Comput Biol* 2007, **3**(8):e129.

112. Pudimat R, Schukat-Talamazzini E-G, Backofen R: **A multiple-feature framework for modelling and predicting transcription factor binding sites**. *Bioinformatics* 2005, **21**(14):3082-3088.

113. Barash YKT, Friedman N, Elidan G: **Modeling Dependencies in Protein-DNA Binding Sites**. In: *The 7th International Conference on Research in Computational Molecular Biology (RECOMB): 2003; Berlin: 2003*; 2003: 28 - 37.

114. Cai D, Delcher A, Kao B, Kasif S: **Modeling splice sites with Bayes networks**. *Bioinformatics* 2000, **16**(2):152-158.

115. Chen T-M, Lu C-C, Li W-H: **Prediction of splice sites with dependency graphs and their expanded bayesian networks**. *Bioinformatics* 2005, **21**(4):471-482.

116. Deforche K, Silander T, Camacho R, Grossman Z, Soares MA, Van Laethem K, Kantor R, Moreau Y, Vandamme AM, on behalf of the non BW: **Analysis of HIV-1 pol sequences using Bayesian Networks: implications for drug resistance**. *Bioinformatics* 2006, **22**(24):2975-2979.

117. de la Grange P, Dutertre M, Correa M, Auboeuf D: **A new advance in alternative splicing databases: from catalogue to detailed analysis of regulation of expression and function of human alternative splicing variants**. *BMC Bioinformatics* 2007, **8**(1):180.

118. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M: **Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity**. *Nat Genet* 2004, **36**(12):1255-1257.
119. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Group RG, Members GSL, Hayashizaki Y, Gaasterland T: **Impact of Alternative Initiation, Splicing, and Termination on the Diversity of the mRNA Transcripts Encoded by the Mouse Transcriptome**. *Genome Res* 2003, **13**(6b):1290-1300.
120. Hiller M, Huse K, Szafranski K, Rosenstiel P, Schreiber S, Backofen R, Platzer M: **Phylogenetically widespread alternative splicing at unusual GYNGYN donors**. *Genome Biology* 2006, **7**(7):R65.
121. Hiller M, Nikolajewa S, Huse K, Szafranski K, Rosenstiel P, Schuster S, Backofen R, Platzer M: **TassDB: a database of alternative tandem splice sites**. *Nucl Acids Res* 2007, **35**(suppl_1):D188-192.
122. Akerman M, Mandel-Gutfreund Y: **Does distance matter? Variations in alternative 3' splicing regulation**. *Nucl Acids Res* 2007, **35**(16):5487-5498.
123. Chern T-M, van Nimwegen E, Kai C, Kawai J, Carninci P, Hayashizaki Y, Zavolan M: **A Simple Physical Model Predicts Small Exon Length Variations**. *PLoS Genetics* 2006, **2**(4):e45.
124. Hiller M, Szafranski K, Backofen R, Platzer M: **Alternative Splicing at NAGNAG Acceptors: Simply Noise or Noise and More?** *PLoS Genetics* 2006, **2**(11):e207.
125. Sinha R, Nikolajewa S, Szafranski K, Hiller M, Jahn N, Huse K, Platzer M, Backofen R: **Accurate prediction of NAGNAG alternative splicing**. *Nucl Acids Res* 2009, **37**(11):3569-3579.
126. Wang B-B, O'Toole M, Brendel V, Young N: **Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes**. *BMC Plant Biology* 2008, **8**(1):17.
127. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong W-K, Mockler TC: **Genome-wide mapping of alternative splicing in Arabidopsis thaliana**. *Genome Research* 2009:-.
128. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud P-F, Lindquist EA, Kamisugi Y *et al*: **The Physcomitrella Genome Reveals Evolutionary Insights into the Conquest of Land by Plants**. *Science* 2008, **319**(5859):64-69.
129. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD *et al*: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies**. *Nucl Acids Res* 2003, **31**(19):5654-5666.
130. Iida K, Shionyu M, Suso Y: **Alternative Splicing at NAGNAG Acceptor Sites Shares Common Properties in Land Plants and Mammals**. *Mol Biol Evol* 2008, **25**(4):709-718.
131. Clark F, Thanaraj TA: **Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human**. *Hum Mol Genet* 2002, **11**(4):451-464.
132. 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* 2008, **9**(1):202.
133. Ermakova EO, Nurtdinov RN, Gelfand MS: **Overlapping alternative donor splice sites in the human genome**. *Journal of Bioinformatics and Computational Biology* 2007:991–1004.

134. Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ: **Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site**. *RNA* 2006, **12**(12):2047-2056.

135. Tsai K-W, Tarn W-Y, Lin W-c: **Wobble Splicing Reveals the Role of the Branch Point Sequence-to-NAGNAG Region in 3' Tandem Splice Site Selection**. *Mol Cell Biol* 2007, **27**(16):5835-5848.

136. Hiller M, Platzer M: **Widespread and subtle: alternative splicing at short-distance tandem sites**. *Trends in Genetics* 2008, **24**(5):246-255.

137. Merediz SAK, Schmidt M, Hoppe GJ, Alfken J, Meraro D, Levi BZ, Neubauer A, Wittig B: **Cloning of an interferon regulatory factor 2 isoform with different regulatory ability**. *Nucleic Acids Research* 2000, **28**(21):4219-4224.

138. Akerman M, Mandel-Gutfreund Y: **Alternative splicing regulation at tandem 3' splice sites**. *Nucl Acids Res* 2006, **34**(1):23-31.

139. Julious SA, Mullee MA: **Confounding and Simpson's paradox**. *BMJ* 1994, **309**(6967):1480-1481.

140. Simpson EH: **The Interpretation of Interaction in Contingency Tables**. *J R Stat Soc Ser B-Stat Methodol* 1951, **13**(2):238-241.

141. Bickel PJ, Hammel EA, Oconnell JW: **Sex Bias in Graduate Admissions - Data from Berkeley**. *Science* 1975, **187**(4175):398-404.

142. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucl Acids Res* 2009, **37**(suppl_1):D26-31.

143. Boguski MS, Lowe TMJ, Tolstoshev CM: **dbEST - database for "expressed sequence tags"**. *Nat Genet* 1993, **4**(4):332-333.

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

145. Maugeri A, van Driel MA, van de Pol DJR, Klevering BJ, van Haren FJJ, Tijmes N, Bergen AAB, Rohrschneider K, Blankenagel A, Pinckers AJLG *et al*: **The 2588G'C Mutation in the ABCR Gene Is a Mild Frequent Founder Mutation in the Western European Population and Allows the Classification of ABCR Mutations in Patients with Stargardt Disease**. 1999, **64**(4):1024-1035.

146. Modrek B, Lee C: **A genomic view of alternative splicing**. *Nat Genet* 2002, **30**(1):13-19.

147. Sugnet CW, Srinivasan K, Clark TA, O'Brien G, Cline MS, Wang H, Williams A, Kulp D, Blume JE, Haussler D *et al*: **Unusual Intron Conservation near Tissue-Regulated Exons Found by Splicing Microarrays**. *PLoS Comput Biol* 2006, **2**(1):e4.

148. Eklund AC, Turner LR, Chen P, Jensen RV, deFeo G, Kopf-Sill AR, Szallasi Z: **Replacing cRNA targets with cDNA reduces microarray cross-hybridization**. *Nat Biotech* 2006, **24**(9):1071-1073.

149. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D *et al*: **A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome**. *Science* 2008, **321**(5891):956-960.

150. Li H, Lovci MT, Kwon Y-S, Rosenfeld MG, Fu X-D, Yeo GW: **Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model**. *Proceedings of the National Academy of Sciences* 2008, **105**(51):20179-20184.

151. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Meth* 2008, **5**(7):621-628.

152. Shepard PJ, Hertel KJ: **Conserved RNA secondary structures promote alternative splicing**. *RNA* 2008, **14**(8):1463-1469.

153. Collins L, Penny D: **Investigating the Intron Recognition Mechanism in Eukaryotes**. *Mol Biol Evol* 2006, **23**(5):901-910.

154. Pan Q, Bakowski MA, Morris Q, Zhang W, Frey BJ, Hughes TR, Blencowe BJ: **Alternative splicing of conserved exons is frequently species-specific in human and mouse**. *Trends in Genetics* 2005, **21**(2):73-77.

155. Yeo G, Burge CB: **Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals**. *Journal of Computational Biology* 2004, **11**(2-3):377-394.

156. Su Z, Wang J, Yu J, Huang X, Gu X: **Evolution of alternative splicing after gene duplication**. *Genome Res* 2006, **16**(2):182-189.

157. Kopelman NM, Lancet D, Yanai I: **Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms**. *Nature Genetics* 2005, **37**(6):588-589.

158. Talavera D, Vogel C, Orozco M, Teichmann SA, de la Cruz X: **The (In)dependence of Alternative Splicing and Gene Duplication**. *PLoS Comput Biol* 2007, **3**(3):e33.

159. Hiller M, Szafranski K, Sinha R, Huse K, Nikolajewa S, Rosenstiel P, Schreiber S, Backofen R, Platzer M: **Assessing the fraction of short-distance tandem splice sites under purifying selection**. *Rna* 2008, **14**(4):616-629.

160. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE: **Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements**. *Nature* 2007, **446**(7138):926-929.

161. Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, Ares M, Jr.: **Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay**. *Genes Dev* 2007, **21**(6):708-718.

162. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome**. *Science* 2004, **304**(5675):1321-1325.

163. Baek D, Green P: **Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing**. *PNAS* 2005, **102**(36):12813-12818.

164. Condorelli G, Bueno R, Smith RJ: **2 Alternatively Spliced Forms of the Human Insulin-Like Growth-Factor-I Receptor Have Distinct Biological-Activities and Internalization Kinetics**. *Journal of Biological Chemistry* 1994, **269**(11):8510-8516.

165. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A *et al*: **The UCSC genome browser database: update 2007**. *Nucl Acids Res* 2007, **35**(suppl_1):D668-673.

166. Ponting CP: **The functional repertoires of metazoan genomes**. *Nat Rev Genet* 2008, **9**(9):689-698.

167. Kornblihtt AR, de la MM, Fededa JP, Munoz MJ, Nogues G: **Multiple links between transcription and splicing**. *Rna* 2004, **10**:1489 - 1498.

168. Muñoz MJ, Santangelo MSP, Paronetto MP, de la Mata M, Pelisch F, Boireau S, Glover-Cutter K, Ben-Dov C, Blaustein M, Lozano JJ *et al*: **DNA Damage Regulates Alternative Splicing through Inhibition of RNA Polymerase II Elongation**. *Cell* 2009, **137**(4):708-720.

169. Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R: **Nucleosome positioning as a determinant of exon recognition**. *Nat Struct Mol Biol* 2009, **16**(9):996-1001.

170. Schwartz S, Meshorer E, Ast G: **Chromatin organization marks exon-intron structure**. *Nat Struct Mol Biol* 2009, **16**(9):990-995.

Acknowledgements

I would like to thank my supervisors Prof. Dr. Stefan Schuster, Prof. Dr. Rolf Backofen, and Dr. Matthias Platzer. I have learned a lot from them during the last three-and-a-half years, both about science as well as life, and many nuggets to remember, above all – “it always takes longer than you think” (Matthias).

I would like to thank my former and current lab mates, especially: Klaus Huse, Karol Szafranski, Ulrike Gausmann, Swetlana Friedel, Marius Felder, Marco Groth, Kathrin Reichwald, Stefanie Schindler, Stefanie Stepanow, Jeannette Kirschner, Andreas Petzold, Andrew Heidel, Oliver Müller, Stefan Taudien, Markus Schilhabel, Niels Jahn, Bernd Senf, Marcel Kramer, Gernot Glöckner, Hella Ludewig, Ivonne Heinze, Kathleen Seitz and Oliver Hoffman in Jena; and Michael Hiller, Rainer Pudimat, Martin Mann, Sebastian Will, Anke Busch, Sven Siebert, Michael Beckstette, Steffen Heyne, Andreas Richter, Monika Degen-Hellmuth, Stefan Jankowsky, Sita Lange, Matthias Möhl, Daniel Matizcka, Kousik Kundu and Oliver Krieg in Freiburg. They have all contributed to the great experience that my Ph.D has been, both by contributing to a great environment at work as well as by including me in various social events and ensuring that I never lead the life of a lonely outsider. I must especially thank Karol, Michael H, Klaus and Ulrike for lots of help with my research, and endlessly stimulating discussions; Klaus and Marius for the several football games at the stadium to which they treated me, and to all the enthusiastic fellow “Kicker” players for much pleasure at that particular table. Prof. Dr. Klemens Hertel was a guest of our group in Freiburg while on a sabbatical from University of California, Irvine, into whose deep reservoir of knowledge on splicing I could barely dip – it was an honour to share an office with him.
.
I cannot go without thanking my good friends, who have supported me and made life outside academics very enjoyable, thereby enhancing the enjoyment of research as well. Though there are more such friends than I could possibly thank, at the very least I must mention Siddhartha, Anu, Ashes, Diptendu, Arun, Naga, Abhishek, Ritesh, Rashmi, Anindita, Gregor, Leila, Radhika, Sada, Sachin and Monika .

I extend my gratitude to the Leibniz Graduate School on Ageing and Age-related diseases (LGSA) and its coordinator, the ever-helpful Dr. Claudia Müller, for all the support received, and all the people at Friedrich-Schiller-University involved in the dissertation procedure. I thank the SFB604 “Multifunctional Signalling Proteins” and its coordinator Prof. Dr. Reinhard Wetzker for my PhD funding via project B10, and all the support received during this period.

Last but exactly the opposite of least, I would like to thank my parents, sister and the rest of my family for always being there for me. I was born while my parents were doing their respective Ph.Ds - looks like I ended up being a chip of the old block(s), after all!

Declaration of Independent Assignment

I declare in accordance with the conferral of the degree of doctor from the School of Biology and Pharmacy of Friedrich Schiller University Jena that the submitted thesis was written only with the assistance and literature cited in the text.

People who assisted in the experiments, data analysis and writing of the manuscripts are listed as co-authors of the respective manuscripts. I was not assisted by a consultant for doctorate theses.

The thesis has not been previously submitted whether to the Friedrich-Schiller-University Jena or to any other University.

Jena, December 3rd, 2009 ………………
Rileen Sinha

Curriculum Vitae

Personal Details
Name: Rileen Sinha
Date of birth: 06.01.1973
Place of birth: Pittsburgh, U.S.A
Nationality: Indian
Address: Blücherstrasse 11, Freiburg im Breisgau
Email: sinha@informatik.uni-freiburg.de

Education:
2009-present: Scientific researcher, Department of Computer Science, University of Freiburg, Freiburg, Germany
2006-2008: Ph.D. student at Fritz-Lipmann Institute and Friedrich Schiller University, Jena
2005-2006: Postgraduate course in Bioinformatics at the Cologne University Bioinformatics Centre (CUBIC), Cologne, Germany
2002-2003: Master of Science, Advanced Computing, King's College London, London, United Kingdom.
1995-1999: Master of Engineering, Electrical Communication Engineering, Indian Institute of Science, Bangalore, India.
1992-1995: Bachelor of Science, Faculty of Science, Banaras Hindu University, Varanasi, India.

Publications:
Improved identification of conserved cassette exons using Bayesian networks
Rileen Sinha, Michael Hiller, Rainer Pudimat, Ulrike Gausmann, Matthias Platzer and Rolf Backofen
BMC Bioinformatics 2008, 9:477.

Accurate prediction of NAGNAG alternative splicing
Rileen Sinha*, Swetlana Nikolajewa*, Karol Szafranski, Michael Hiller, Niels Jahn, Klaus Huse, Matthias Platzer and Rolf Backofen
Nucleic Acids Research 2009, Vol. 37, No. 11 3569-3579.
* joint first authors

Identification and characterization of NAGNAG alternative splicing in the moss *Physcomitrella patens*
Rileen Sinha, Andreas D. Zimmer, Kathrin Bolte, Daniel Lang, Ralf Reski, Matthias Platzer, Stefan A. Rensing and Rolf Backofen
(manuscript submitted)

Assessing the fraction of short-distance tandem splice sites under purifying selection
Michael Hiller, Karol Szafranski, **Rileen Sinha**, Klaus Huse, Swetlana Nikolajewa, Philip Rosenstiel, Stefan Schreiber, Rolf Backofen, and Matthias Platzer
RNA 2008, 14(4): 616-629.

TassDB2 - A comprehensive database of subtle alternative splicing events
Rileen Sinha, Thorsten Lenser, Niels Jahn, Ulrike Gausmann, Swetlana Friedel, Karol Szafranski, Klaus Huse, Philip Rosenstiel, Jochen Hampe, Stefan Schuster, Michael Hiller, Rolf Backofen and Matthias Platzer
(under review at BMC Bioinformatics)

Posters:
1) "Effective RNA length and its applications" at the Alternative Splicing SIG Meeting, ISMB 2009, Stockholm Sweden.
2) "Accurate prediction of NAGNAG alternative splicing using Bayesian networks", ISMB 2008, Toronto, Canada.
3) "Improved prediction of conserved exon skipping using Bayesian networks", ISMB 2007, Vienna, Austria.
4) "Learning genetic networks – combined evaluation of transcription regulation information and gene expression experiments", GCB 2006, Tübingen, Germany.
5) "Fast alignment searching using indexing, filtering and vectorization, GCB 2004, Bielefeld, Germany.

Scientific Talks:
1) "Towards *ab-initio* Prediction of Alternative Splicing in Signalling Proteins". SFB604 Symposium, Jena, Germany, September 2007.
2) "Violating the splicing rules : TG dinucleotides function as alternative 3' splice sites in U2-dependent introns." (on behalf of K. Szafranski), Alternative Splicing SIG Meeting, ISMB 2007, Vienna, Austria, July 2007.
3) "A subtle event whose effects are not – frame shifting alternative splicing at short distance tandem acceptors". JCB Workshop, Jena, Germany, May 2007
4) "Towards *ab-initio* Prediction of Alternative Splicing in Signalling Proteins". SFB604 Workshop, Jena, Germany, March 2007
5) "Suffix trees in Bioinformatics", Jawaharlal Nehru University, Delhi, India, March 2005.