UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA

**Programa de Pós-Graduação em Ciências Biológicas
(Bioquímica)**

NOBORU JO SAKABE

ANÁLISE DE CARACTERÍSTICAS DAS SEQUÊNCIAS GENÔMICAS
RELACIONADAS A EVENTOS DE *SPLICING* ALTERNATIVO DO TIPO
RETENÇÃO DE INTRON NO TRANSCRIPTOMA HUMANO

versão eletrônica, sem errata
texto revisado e corrigido
pronto para impressão frente e verso

São Paulo

Data do Depósito na SPG:
30/11/2006

NOBORU JO SAKABE

ANÁLISE DE CARACTERÍSTICAS DAS SEQUÊNCIAS GENÔMICAS
RELACIONADAS A EVENTOS DE *SPLICING* ALTERNATIVO DO TIPO
RETENÇÃO DE INTRON NO TRANSCRIPTOMA HUMANO

Tese apresentada ao Instituto de
Química da Universidade de São Paulo para
obtenção do Título de Doutor em Ciências
Biológicas (Bioquímica)

Orientador(a): Prof(a). Dr(a). Sandro José de Souza

São Paulo
2006

*O Senhor tem direito à sua opinião
Mas eu confio na Minha*

Agradecimentos

Agradeço ao Sandro por me receber em seu laboratório, pelas oportunidades oferecidas, por ter permitido que eu desenvolvesse minhas próprias idéias e meu trabalho com liberdade. Isso estimulou a minha independência e maturação como pesquisador independente.

Agradeço à FAPESP pela bolsa concedida; à Ana Cláudia pela preciosa ajuda com os problemas do dia-a-dia e ao suporte técnico provido por André Zaiats e equipe.

Se uma pós-graduação é de alguma maneira formativa, agradeço aos meus pais por terem me dado toda a base dessa formação que agora simbolicamente termina. Filho de pais universitários, em verdade, a obtenção deste título é só uma consequência do que a minha família, desde os meus avós que imigraram para o Brasil, construiu. Embora muitas exceções existam, a carreira científica, principalmente no Brasil, é quase dependente ou pelo menos enormemente facilitada pelo apoio familiar com que sempre contei.

Sempre quis ter a oportunidade de agradecer formalmente à minha primeira orientadora ainda na graduação, a professora Shirley Schreier do departamento de Bioquímica do IQ-USP, que faz complicados cálculos no verso do papel e deduz equações físico-químicas no canto de um artigo. Agradeço a ela por ter tido um papel fundamental na minha formação científica, não só nos ensinamentos técnicos, mas principalmente nos filosóficos. Entre muitas coisas, ela me ensinou que nossa capacidade científica evolui exponencialmente. Isso não me poupou (e não poupa) das frustrações da vida científica, mas é certamente um ensinamento valioso que me ajudou a manter a perspectiva adequada das coisas (bom, quase sempre). Devo aqui acrescentar que a fase *lag* da minha evolução aconteceu em boa parte sob sua orientação. Nessa fase, aprendi com a professora Shirley que, às vezes, ir devagar é o melhor a se fazer. Herdei dela a devoção aos detalhes, o prazer de fazer “bem-feito” e do perfeccionismo, sempre me lembro dela quando estou corrigindo a última vírgula da décima versão de um texto ou corrigindo um chatíssimo *bug* em um *script*. Ela é também meu maior parâmetro de rigor e seriedade científica.

Agradeço também ao meu segundo orientador de iniciação científica, o professor Bayardo Baptista Torres, também do departamento de Bioquímica do IQ-USP, alguém a quem nós que já fomos membros do seu laboratório chamamos de *mentor*. Com ele

aprendi, entre muitas coisas, perdoem o horroroso lugar-comum, a aprender. Esse deve ter sido um dos ensinamentos mais valiosos que recebi na universidade. Por outro lado, a convivência com ele me levou, pelo menos por algum tempo, à crença totalmente incorreta de que todo acadêmico é um grande ser humano. Ele me ensinou também que não me faltava inteligência, muitas das aulas e palestras que tive que assistir eram péssimas mesmo. Aprendi que quando sentimos que a palestra deve ter sido excelente porque não entendemos nada, é porque na verdade o apresentador é que foi péssimo. Também aprendi nos cursos que dei com ele o prazer de partilhar o que sei, como ensinar é uma atividade que requer muita seriedade e comprometimento, acima de tudo boa-fé, mas que traz enormes recompensas. No professor Bayardo encontrei o reflexo de vários valores humanos e científicos. Por isso, acima de tudo, a convivência com este verdadeiro mentor foi importante para a minha formação como ser humano.

Por essas razões, agradeço a esses dois docentes que tiveram um grande papel na minha formação se dispondo a olhar para mim não só com o que diz respeito à produção.

O “título” de mentor foi dado pelo Guilherme, com quem convivi por alguns anos no laboratório do prof. Bayardo. A ele, mais velho do que eu, também agradeço pelas dicas sobre as artimanhas da vida científica, conselhos e palpites sempre úteis; além das gargalhadas que damos até hoje pelo Skype.

Agradeço aos meus colegas de laboratório Pedro, Robson, Jorge, Natanja e Maria, com quem partilhei as angústias, incertezas e alegrias, pães recheados, risotos e fondues e com quem, juntos, formamos um verdadeiro time. Em particular, agradeço ao Robson pelas frutíferas, e geralmente longuíssimas, discussões científicas e técnicas de alto nível e pelas infindáveis ajudas nos momentos mais cruciais, por exemplo às 9 da noite. Com o Pedro também tive grandes discussões científicas, sobre Ciência e a vida científica, sobre a vida em geral e as ações da bolsa de valores. Junto com o Robson, o Pedro e eu formamos por alguns anos o Clube do Açaí, em cujas reuniões tínhamos a oportunidade de falar mal de todo mundo e de discutir assuntos científicos (em partes iguais). Com o Jorge, embora ele não goste de açaí, também tive boas discussões técnicas, além das nossas conversas depois do expediente sobre evolução, vida científica, vida alienígena, vida de outrem, vida, formas alternativas de alimentação baseadas em itens de qualidade duvidosa, Ciência, a existência de Deus e os Simpsons. Quantas vezes partilhamos as intragáveis bolachas genéricas que ele comprava, cheias de gorduras trans? A Natanja, uma holandesa-israelense mais brasileira do que eu, com

quem troquei algumas receitas culinárias e *scripts*, foi sempre amiga, e ainda é, nas alegrias e nas tristezas. Até hoje partilhamos *herzenkronkels* e *gedachtengangs* (caraminholas e formas de pensar, respectivamente). E por fim, agradeço à Maria, por todo o companheirismo, paciência e ajuda técnica, intelectual e emocional durante todo o doutorado, principalmente no meu último artigo em que ela, mesmo estando fora do laboratório, me ajudou tanto, fosse interpretando as complicadas correlações de Spearman ou criticando o texto. O Pedro e a Maria também foram aqueles que revisaram esta tese, e a Maria ainda me ajudou a contrabandear artigos da Roma do nosso século para sua escrita. Sem a ajuda e o companheirismo destes colegas, tudo teria sido infinitamente mais difícil e certamente, muito menos engraçado.

Não poderia deixar de agradecer também a todos aqueles que me serviram de exemplo do que eu não quero ser jamais, do caminho a não se seguir. Ao longo desse caminho, não foram poucos. Embora seja importante se ter gurus e mentores, a existência desses exemplos é indubitavelmente preciosa.

Meu último agradecimento é para aqueles que desenvolvem *softwares* livres, como a linguagem de programação, o sistema de bancos de dados e o sistema operacional que utilizamos, o Linux. Seria impossível realizar uma fração sequer do que realizamos se usássemos Window\$. Agradeço sobretudo ao espírito de boa-fé desses desenvolvedores que não só distribuem seus códigos de forma aberta, mas principalmente produzem programas excelentes, de incomparável robustez, eficiência e qualidade. É o espírito que afronta a tirania e mostra que aqueles que acreditam que a busca pelo lucro pessoal é só o que importa e até mesmo que o capitalismo é somente obter o máximo de lucro estão tristemente errados. Como o espírito científico, esse espírito tem como objetivo maior não alguma forma de lucro ou de benefício pessoal escuso, mas realizar o melhor trabalho possível, da forma mais honesta, coletivamente, colaborativamente e sempre estar disposto a evoluir. É esse o *zeitgeist* que eu gostaria de ver triunfar. Foi com esse espírito que por quatro anos realizei meu doutorado.

Sobre regras e suas mudanças

Eu gosto de regras. Não hiperventilo se tenho que seguir formatos e instruções. Até prefiro que existam. Contanto que sejam razoáveis, é claro, e sejam feitas para melhorar a vida das pessoas.

Alguma vez comentei que era uma pena a falta de instruções formais impressas para escrita da tese.

De toda forma, na falta de regras impressas acessíveis (sim, a ampla divulgação e acessibilidade às regras faz parte do processo), interpelei a SPG em busca de instruções e formatações. Fui informado que não havia, que o formato era livre.

Embora goste de regras, também gosto de fazer as minhas, e assim escrevi e formatei a tese com a liberdade que me foi concedida.

Para minha surpresa, cinco dias antes do prazo final de depósito da tese, (sim, cinco dias) foram divulgadas regras formais impressas para confecção da tese. Se por um lado fiquei feliz por agora existirem as tais regras, fiquei preocupado pelo fato de terem sido emitidas cinco dias antes do prazo final. Normalmente, em situações como esta, dá-se um período de tolerância até o início da validade das regras. É razoável. É humano.

Entre imprimir a tese e tentar cumprir o máximo de formatações possíveis nos cinco dias que me restavam, entreguei minha tese com meu manuscrito contendo meus resultados em inglês, prática que aprendi de teses anteriores de outras pessoas. Diga-se, portanto aprovadas por reuniões da CPG anteriores.

Para minha surpresa (surpreendi-me muito nesses dias), minha tese foi recusada pela reunião da CPG. Eu não havia respeitado o formato. Tinha que traduzir o manuscrito. Sim, cinco dias depois de terem sido emitidas, as regras foram aplicadas.

Felizmente, a CPG concedeu-me uma forma de ter minha banca indicada antes da próxima reunião, que só aconteceria em março de 2007. Agradeço enormemente à CPG por dispor um tempo extra para a solução do meu problema.

Mas não posso deixar de pensar se havia mesmo um problema. Eu havia sido informado por meios oficiais (um funcionário da SPG) de que o formato era livre. Que problema existe em seguir regras informais endossadas por reuniões oficiais do próprio IQUSP? Que problema existe em não seguir regras quando elas foram divulgadas cinco dias antes? De que adiantam regras que ninguém conhece nem pratica? Além disso,

note-se ainda que a referida regra diz que "Em conformidade com a Resolução CoPGr nº 4678/99 da USP, as dissertações e teses devem ser redigidas em português." Minha tese estava escrita em português, com uma seção em inglês. Estritamente, a regra não diz que a tese *inteira* deve ser redigida em português.

Este acontecimento me lembrou um pouco um episódio dos Simpsons em que o Sr. Burns abre uma penitenciária privada e precisava de presos para enchê-la para ganhar dinheiro. Na falta de prisioneiros, ele passou, através da polícia, a buscar velhas leis para poder prender todo e qualquer cidadão, como por exemplo a proibição de chutar uma latinha mais de cinco vezes (o que leva o Homer à prisão). Ou seja, a aplicação de uma lei depende da realidade.

Não estou dizendo que há uma conspiração perpetrada pelo Sr. Burns, apenas que as regras não devem ser feitas para existirem por si, mas para melhorar a vida das pessoas, coletivamente. Mas se a regra não condiz mais com uma realidade moderna, que mude-se as regras e não a realidade. O fato de que muitos alunos têm podido transformar suas teses ou parte delas em artigos, e que isso afeta a escrita de suas teses não pode mais ser negligenciado. Se há um problema jurídico, que se reveja as leis e não nos tornemos reféns delas como vemos em tantas situações cotidianas do nosso país.

Vale ainda notar que ainda que não me custe tanto assim seguir a regra, levei 3 dias de trabalho intenso para traduzir e reformatar a tese. Somando ao tempo que levei para preparar a tese original, teria levado sete dias para entregar a tese. Ou seja, se eu tivesse seguido as regras divulgadas cinco dias antes do prazo, não teria conseguido ter entregue minha tese a tempo.

Embora possa parecer que minha principal intenção é discutir a regra, esse não é o ponto. Tenho ciência que mesmo discordando de regras, temos o dever de segui-las, se elas forem divulgadas e praticadas por todos igualmente. Mas esse é o ponto, não pedi um tratamento especial, ao contrário, pedi um tratamento igual baseado em um precedente aberto pela própria CPG. Os precedentes estão nas prateleiras da biblioteca da Química.

Não posso também deixar de achar que há uma certa ironia nesse episódio. Justo agora que o IQUSP fala em internacionalizar seu programa de pós-graduação, há um reforço da regra de que as teses devem ser escritas somente em português, uma língua, que como muitos sabem, não é a língua mais usada em ciência no mundo todo.

Curiosamente, as “novas regras” (não sei quais outras regras já existiam mas não eram praticadas) trazem ainda a abolição da escrita de uma tese final. Agora escreve-se apenas uma errata. Tenho sinceras dúvidas se isso não acarretará em uma menor qualidade das teses produzidas, já que apenas pequenos erros poderão ser corrigidos. Acho que os alunos ficarão muito felizes em não precisar mais corrigir a tese. Mas não me parece que essa “nova regra” tenha sido estabelecida para melhorar a vida dos alunos.

De toda forma, espero que a partir de agora, todos, absolutamente todos, sigam as regras, a começar pelas reuniões da CPG, sem exceções. Mas não só essas regras, mas todas as regras. Que agora sejam divulgadas e aplicadas.

Viva as regras.

Resumo

Sakabe, N.J., **Análise de características das sequências genômicas relacionadas a eventos de *splicing* alternativo do tipo retenção de intron no transcriptoma humano**. 2006. (169 pág.) Tese (Doutorado) - Programa de Pós-Graduação em Bioquímica. Instituto de Química, Universidade de São Paulo, São Paulo.

Os genes eucarióticos, em sua maioria, são divididos em exons e introns, requerendo processamento do RNAm para remover as sequências intrônicas e juntar os exons (*splicing*). As bordas exon/intron são definidas por sítios de *splice* que normalmente são reconhecidos com alta fidelidade, gerando os mesmos RNAs processados a cada vez. Apesar desse reconhecimento preciso, tem sido observada a junção de exons de maneiras alternativas (*splicing* alternativo), foco de muitos estudos recentes devido à sua importância em vários processos biológicos. Este processamento alternativo do RNAm pode ser principalmente de três tipos: *exclusão de exon*, no qual um exon pode ser incluído ou não no RNAm maduro; *uso alternativo de sítios de splice*, resultando em exons mais longos ou mais curtos e *retenção de intron*, o tipo menos estudado, no qual uma sequência intrônica é mantida no RNAm maduro. Um dos aspectos cruciais no entendimento de *splicing* alternativo é conhecer os mecanismos que levam à geração de diferentes transcritos. Coerente com a importância dos sítios de *splice* no *splicing* de RNAs, a retenção de intron parece ser causada por falha no reconhecimento daqueles que são sub-ótimos. Como os sítios de *splice* são reconhecidos aos pares ao se estabelecer uma ponte através de exons ou introns, dependendo de qual é mais curto, uma falha no reconhecimento de um exon ou de um intron leva a diferentes tipos de *splicing* alternativo (*exclusão de exon* ou *retenção de intron*, respectivamente). Desta forma, acredita-se que a ocorrência de retenção de intron esteja também associada a uma falha no reconhecimento de introns curtos. Embora estudos de introns retidos individuais tenham abordado estas questões, poucas análises sistemáticas de grandes quantidades de dados foram conduzidas sobre as características gerais que levam à retenção de intron. Para este fim, realizamos uma análise de bioinformática de sequências do genoma e transcriptoma (RNAm) humanos armazenadas em formato de computador. Para realizar as análises computacionais, desenvolvemos um sistema de anotação de *splicing* alternativo completo. Particionamos os eventos de retenção de intron identificados em sequências expressas pelo nosso sistema de anotação em dois grupos, com base na abundância relativa das duas isoformas (um grupo de eventos com <50% e outro com >50% de transcritos retendo o intron) e comparamos características relevantes. Verificamos que

uma maior frequência de retenção de intron em humano está associada a sítios de *splice* mais fracos, genes com introns mais curtos e maior nível de expressão gênica, e menor densidade de um conjunto de elementos inibitórios exônicos e do promotor de *splicing* intrônico GGG. Os dois grupos apresentaram eventos conservados em camundongo, nos quais os introns retidos também eram curtos e apresentavam sítios de *splice* mais fracos. Embora nossos resultados tenham confirmado que sítios de *splice* mais fracos estão associados à retenção de intron, eles mostram que uma fração não-desprezível dos eventos não pode ser explicada apenas por esta característica. Nossa análise sugere que elementos reguladores em *cis* provavelmente têm um papel na regulação da retenção de intron e também revelou características previamente desconhecidas que parecem influenciar sua ocorrência. Estes resultados salientam a importância de considerar o compromisso entre estas características na regulação da frequência relativa de retenção de intron.

Palavras-chave: *splicing* alternativo, retenção de intron, transcriptoma, bionformática, frequência relativa de isoformas

Abstract

Sakabe, N.J., **Analysis of genomic sequence features related to alternative splicing events (intron retention) in the human transcriptome**. 2006. (169 pg) PhD Thesis - Graduate Program in Biochemistry. Instituto de Química, Universidade de São Paulo, São Paulo.

Most eukaryotic genes are split in exons and introns, requiring mRNA processing to remove intervening sequences and join exons (splicing). Exon/intron borders are defined by splice sites that are normally recognized with high fidelity, yielding the same processed mRNA each time. Notwithstanding such precise recognition, alternative joining of exons has been observed (alternative splicing) and is the focus of many recent studies, due to its importance in several biological processes. This alternative mRNA processing can be mainly of three types: *exon skipping*, whereby an exon may be included or not in the mature mRNA; *alternative use of splice sites*, resulting in longer or shorter exons and *intron retention*, the least studied type whereby an intronic sequence is maintained in the mature mRNA. One of the key aspects in understanding alternative splicing is to know the mechanisms that lead to the generation of different transcripts. Coherent with the importance of splice sites in mRNA splicing, intron retention seems to be caused by failure in the recognition of those that are sub-optimal. As splice sites are recognized in pairs by bridging either exons or introns, depending on which is the shortest, failure to recognize an exon or an intron leads to different types of alternative splicing (exon skipping or intron retention, respectively). This way, the occurrence of intron retention is believed to be associated to failure in recognition of short introns also. Although studies on individual retained introns have addressed such issues, few systematic surveys of large amounts of data have been conducted on the general features leading to intron retention. To this end, we performed a bioinformatics analysis of human genome and transcriptome (mRNA) sequences stored in computer format. To perform the computational analyses we developed a complete alternative splicing annotation system. We partitioned intron retention events identified in expressed sequences by our annotation system in two groups based on the relative abundance of both isoforms (one group of events with <50% and another with >50% of transcripts retaining the intron) and compared relevant

features. We found that a higher frequency of intron retention in human is associated to weaker splice sites, genes with shorter intron lengths and higher expression level, and lower density of a set of exonic inhibitory elements and the intronic splicing enhancer GGG. Both groups of events presented conserved events in mouse, in which the retained introns were also short and presented weaker splice sites. Although our results confirmed that weaker splice sites are associated to intron retention, they showed that a non-negligible fraction of events can not be explained by this feature alone. Our analysis suggests that *cis*-regulatory elements are likely to play a crucial role in regulating intron retention and also revealed previously unknown features that seem to influence its occurrence. These results highlight the importance of considering the interplay among these features in the regulation of the relative frequency of intron retention.

Keywords: alternative splicing, intron retention, transcriptome, bioinformatics, relative isoform frequency

Abreviaturas, definições e convenções

Embora esta tese seja escrita em português, alguns termos técnicos empregados foram mantidos em seu original em inglês, como por exemplo *splicing*, siglas como ESE (*exonic splicing enhancer*), snRNP (*small nuclear ribonucleoproteins*) entre outros. Esta decisão foi tomada porque apesar de ser possível utilizar traduções destes termos, (i) há termos cujo significado está profundamente associado à sua forma inglesa (por exemplo, o termo sítio de *splice*) e são utilizados como tal no mundo todo. Outras vezes, os termos possuem tradução mas são quotidianamente empregados em inglês (principalmente as siglas), (ii) nem sempre há traduções oficiais na língua portuguesa, (iii) as traduções nem sempre são satisfatórias, (iv) portanto, tendem a confundir mais do que ajudar (já que na realidade a esmagadora maioria dos leitores de língua portuguesa utilizam quotidianamente os termos em inglês), (v) às vezes terminam por serem um desserviço ao leitor iniciante que terá dificuldades em associar os termos traduzidos aos correntemente utilizados na literatura internacional, (vi) a totalidade dos periódicos importantes são escritos na língua inglesa.

Em contrapartida, a utilização de estrangeirismos desnecessários foram fortemente evitados.

Nos casos (arbitrários) em que a tradução era intuitiva e não trazia grandes prejuízos ao texto, os termos traduzidos foram utilizados (ex.: *commitment complex* = complexo de engajamento). De toda forma, com exceção das siglas, os termos em inglês foram salientados em itálico.

Abaixo seguem abreviaturas, convenções e definições usadas neste trabalho:

a 3' de, na direção 3' de - downstream

a 5' de, na direção 5' de – upstream

cDNA – complementary DNA

cluster, cluster de cDNA – agrupamento de sequências expressas (cDNAs: ESTs e RNAs) idealmente representando a coleção de transcritos de um gene baseado em alinhamento ao genoma

complexo de engajamento – commitment complex

cross-linking – ligação covalente de duas moléculas por irradiação de UV

EJC – exon junction complex

ESE – exonic splicing enhancer

ESS – exonic splicing silencer

EST – expressed sequence tag

exon 1 – para facilitar as descrições, o exon 1 é aquele que considerando o *splicing* de um único intron entre dois exons, situa-se a 5' do intron removido (ver **FIGURA A**)

exon 2 – para facilitar as descrições, o exon 2 é aquele que considerando o *splicing* de um único intron entre dois exons, situa-se a 3' do intron removido (ver **FIGURA A**)

força do sítio de *splice* – embora os nucleotídeos dos sítios de *splice* sejam bastante conservados, a maioria das posições é bastante flexível quanto à sua composição nucleotídica. Dependendo de quanto elas desviam da composição ótima para ligação de componentes do spliceossomo, o sítio pode ser reconhecido de forma mais ou menos eficiente. Diz-se então que o sítio tem uma “força” maior ou menor, e diferentes medidas quantitativas existem para expressar esta “força”.

hnRNP – heterogeneous nuclear ribonucleoprotein

ISE – intronic splicing enhancer

kb – kilobases (mil nucleotídeos)

kDa – kilodaltons (mil daltons)

lnRNP – large nuclear ribonucleoprotein

NMD – nonsense mediated decay

nt – nucleotídeo(s)

RNA_m(s) – RNA mensageiro(s)

(domínio) RS – domínio rico em arginina e serina

RUST – regulated unproductive splicing and translation

SAGE – serial analysis of gene expression, técnica de análise de expressão gênica em larga escala baseada na geração de grandes quantidades de sequências (etiquetas) curtas da porção 3' de RNAs. Idealmente o número de etiquetas produzidas é proporcional ao nível expressão gênica.

sítio de *splice* – borda exon/intron, englobando os nucleotídeos -3 a +5 (5') e -20 a +2 (3') (embora as coordenadas possam mudar de acordo com diferentes autores), com o nucleotídeo 0 sendo o primeiro do intron.

sítio de *splice* 3' – borda exon/intron a 3' do intron, ou seja, refere-se à extremidade 5' do exon 2 (**FIGURA A**).

sítio de *splice* 5' – borda exon/intron a 5' do intron, ou seja, refere-se à extremidade 3' do exon 1 (**FIGURA A**).

(proteínas) **SR** – proteínas que possuem um domínio **RS**

snRNP – small nuclear ribonucleoprotein

splicing – processamento de RNAs no qual os introns são removidos do pré-RNAs, emendando os exons sequencialmente

FIGURA A. Nomenclatura dos sítios de *splice*.

Sobre a Introdução a esta tese

O estudo de *splicing* alternativo em larga-escala é uma área bastante nova que teve uma impressionante expansão nos últimos anos e uma vasta literatura acerca dos mais variados aspectos do fenômeno têm sido produzida. Na Introdução a esta tese, busquei ilustrar a diversidade desses aspectos e permitir que o leitor vislumbre a dimensão da importância e do impacto biológico do fenômeno. Além disso, revisei o mecanismo de *splicing* extensamente, com um grande número de detalhes para possibilitar uma compreensão minuciosa dos possíveis pontos de regulação do processo onde o *splicing* alternativo poderia ocorrer.

Este esforço resultou em uma introdução possivelmente detalhada em excesso para a maioria dos leitores. Assim, nem todas as partes da Introdução são absolutamente relevantes para a compreensão do estudo sobre o qual versa a tese e o leitor ou leitora pode optar por não ler a Introdução inteira. Procurei agrupar as partes mais detalhadas na seção 1.4 que é inteiramente opcional. Para uma compreensão adequada dos resultados apresentados nesta tese, recomenda-se ler inteiramente as seguintes seções:

- “1.1 *Splicing*”
- “1.2 Sinais básicos de *splicing* no pré-RNAs”
- “1.3 O spliceossomo, uma máquina extraordinária: o mecanismo canônico de *splicing*”
- “1.5 O mecanismo canônico geral de *splicing* revisto”
- “1.7 *Splicing* alternativo”, à exceção possivelmente das seções “1.7.2 Características do fenômeno de *splicing* alternativo” e “1.7.4 Outros estudos”
- “1.8 Bioinformática: uma nova maneira de abordar velhos e novos problemas”, esta seção é opcional mas traz uma perspectiva sobre o porquê de se realizar uma análise computacional em biologia
- “1.9 Característica das sequências de RNAs envolvidas na retenção de introns”

Índice

Introdução.....	1
1.1 <i>Splicing</i>	1
1.2 Sinais básicos de <i>splicing</i> no pré-RNA.....	6
1.3 O spliceossomo, uma máquina extraordinária: o mecanismo canônico de <i>splicing</i>	7
1.3.1 A máquina	7
1.3.2 Modelo canônico de montagem do spliceossomo	8
1.3.3 Seleção dos sítios de splice: definição de introns versus definição de exons	11
1.3.3.1 Os introns de mamíferos são tipicamente muito maiores que os exons.....	11
1.3.3.2 As interações através do exon devem ser convertidas para interações através do intron	18
1.3.3.3 Polaridade da definição de exon	19
1.4 O mecanismo de <i>splicing</i> em detalhes.....	20
1.4.1 Interações através do intron.....	20
1.4.1.1 O reconhecimento do sítio de <i>splice</i> 3' requer U2AF e outras proteínas auxiliares	21
1.4.1.2 As interações através do intron ocorrem bastante cedo (complexo de engajamento/E)	22
1.4.1.3 O complexo A também apresenta interações entre os sítios de <i>splice</i> 5' e 3'	25
1.4.2 O papel das proteínas SR nas interações através dos introns e na montagem do spliceossomo	25
1.4.3 As interações através dos exons também envolvem proteínas SR (modelo de definição de exon)	26
1.4.3.1 Comunicação entre sítios de <i>splice</i> através do exon.....	26
1.4.3.2 Proteínas SR poderiam formar pontes através do exon	28
1.4.3.3 As interações através do intron envolvendo proteínas SR podem também ocorrer através do exon.....	30
1.4.3.4 Junção dos exons e conversão das interações através do exon para através do intron.....	31
1.4.4 Detalhes do reconhecimento do sítio de <i>splice</i> 5'	33
1.4.4.1 Proteínas SR estão envolvidas no reconhecimento do sítio de <i>splice</i> 5'	33
1.4.4.2 Os nucleotídeos conservados GT não são necessários para o reconhecimento inicial por U1	35
1.4.4.3 O sítio de <i>splice</i> 5' deve ser reconhecido diversas vezes.....	36
1.4.5 Detalhes no reconhecimento do sítio de <i>splice</i> 3'	37
1.4.5.1 Estruturação do sítio de <i>splice</i> 3' por U2AF ⁶⁵	37
1.4.5.2 Proteínas SR e ESEs também participam do reconhecimento de sítios de <i>splice</i> 3'	38
1.4.6 Detalhes do mecanismo de atuação de proteínas SR: interação proteína-proteína versus interação com RNA.....	40
1.4.7 Mecanismo de ativação do spliceossomo: do complexo B para o complexo C	43

1.4.7.1	A ativação do spliceossomo requer diversos rearranjos das snRNPs	43
1.4.7.2	O papel da snRNP U5 no <i>splicing</i>	45
1.4.8	Alguns detalhes da catálise	47
1.4.8.1	A proteína PRP8 parece ter um importante papel na catálise	48
1.4.8.2	Um ou dois sítios catalíticos?	49
1.5	O mecanismo canônico geral de <i>splicing</i> revisto	50
1.6	Mecanismos de fidelidade do <i>splicing</i> e controle de qualidade	53
1.6.1	A hipótese do holospliceossomo e o modelo do supraspliceossomo	53
1.6.2	Controle de qualidade da produção de RNAm: non-sense mediated decay	56
1.7	<i>Splicing</i> alternativo	59
1.7.1	Alinhamento de cDNAs no genoma, identificação de <i>splicing</i> alternativo no transcriptoma	63
1.7.2	Características do fenômeno de <i>splicing</i> alternativo	66
1.7.2.1	Uma alta frequência de <i>splicing</i> alternativo em diversos organismos	67
1.7.2.2	Função do <i>splicing</i> alternativo	69
1.7.2.3	Relação de <i>splicing</i> alternativo com outros processos celulares	73
1.7.3	Mecanismos de regulação do <i>splicing</i> alternativo	76
1.7.3.1	O papel dos elementos <i>cis</i> e <i>trans</i> na regulação do <i>splicing</i> alternativo	76
1.7.3.2	Identificação de motivos de elementos reguladores de <i>splicing</i> alternativo em <i>cis</i>	81
1.7.3.3	Características dos exons e introns alternativos com implicações mecanísticas	87
1.7.3.4	Mecanismos de <i>splicing</i> alternativo diversos	90
1.7.4	Outros estudos de <i>splicing</i> alternativo	91
1.7.4.1	Outros métodos de identificação de <i>splicing</i> alternativo	91
1.7.4.2	<i>Splicing</i> alternativo em diferentes tecidos	92
1.7.4.3	<i>Splicing</i> alternativo e doenças	93
1.7.4.4	A conservação do <i>splicing</i> alternativo em diferentes organismos e evolução	94
1.8	Bioinformática: uma nova maneira de abordar velhos e novos problemas	99
1.9	Característica das sequências de RNAm envolvidas na retenção de introns	102
1.9.1	A busca por um código de <i>splicing</i>	105
2	Objetivos	107
3	Materiais e métodos	109
3.1	Dados primários utilizados	109
3.2	Anotação de <i>splicing</i> alternativo no transcriptoma humano	110
3.2.1	Anotação de eventos de retenção de intron e filtragem do conjunto de dados	113
3.3	Identificação de sítios de <i>splice</i>	113
3.3.1	Cálculo do conteúdo de informação de Shannon para avaliação da força dos sítios de <i>splice</i>	114

3.3.2	Cálculo da “pontuação de Shapiro & Senapathy” para avaliação da força dos sítios de splice	116
3.4	Medição do nível de expressão gênica	118
3.5	Densidade de motivos <i>reguladores</i> de <i>splicing</i> em <i>cis</i>	120
3.6	Conservação de retenção de intron em camundongo	121
4	Resultados	123
4.1	A força dos sítios de <i>splice</i> como causa da retenção de intron	123
4.2	A influência do comprimento dos exons flanqueantes e dos introns na retenção de intron.....	127
4.3	Genes com retenção de intron são expressos em mais tecidos e em maiores níveis.....	132
4.4	Distribuição e densidade de elementos reguladores de <i>splicing</i> em <i>cis</i>	138
4.5	Conservação da frequência de retenção de intron em camundongo	144
5	Discussão	147
6	Conclusão.....	153
7	Referências.....	155
8	Anexo A - Documentação da ferramenta ASC utilizada para mapeamento e agrupamento de cDNAs e anotação de <i>splicing</i> alternativo	171

Página propositalmente deixada em branco

Introdução

1.1 Splicing

Down where I darn with the milk-eyed mender

“Sadie” – Joanna Newsom

Genes eucarióticos, em sua maioria, apresentam-se divididos em exons e introns. Após a transcrição do gene em RNA mensageiro (RNAm), em realidade um pré-RNAm, ocorre um processo de remoção dos introns e emenda dos exons em uma molécula única (RNAm maduro) (revisado em Sharp (2005)). Este processamento é denominado pela palavra inglesa *splicing* que significa “unir, juntar (duas pontas de uma corda ou de uma fita)”. No caso dos genes que codificam proteínas, a molécula resultante composta somente por exons inclui de forma sequencial uma região não-traduzida a 5’, uma região que codifica a proteína e uma região não-traduzida 3’.

Quimicamente, este processamento do pré-RNAm consiste em duas quebras e duas religações de ligações fosfodiéster entre nucleotídeos dos exons e do intron, ou seja, duas transesterificações (**FIGURA 1**). A primeira transesterificação ocorre quando um grupo OH livre em 2’ de uma adenina localizada na região 3’ do intron a ser excisado promove um ataque nucleofílico ao fosfato da ligação fosfodiéster entre o último nucleotídeo do exon a 5’ (exon 1) e o primeiro do intron, rompendo a ligação exon 1/intron e formando-se um laço intrônico (**FIGURA 1a**). A separação do exon 1/intron deixa um grupo 3’-OH livre no exon 1 (**FIGURA 1b**) que pode atacar o fosfato da ligação fosfodiéster entre o último nucleotídeo do intron a ser excisado e o primeiro

nucleotídeo do exon 2, rompendo-a e ligando os dois exons (**FIGURA 1c**), constituindo a segunda transesterificação. O intron em forma de laço foi removido e é degradado.

FIGURA 1. Mecanismo geral de *splicing* nuclear e de grupo II.

(a) O grupo 2'-OH da adenina no ponto de ramificação na região 3' do intron ataca o fosfato da ligação entre o "exon 1" e a guanina em GT do intron na primeira transesterificação, formando um laço intrônico. (b) na segunda transesterificação, o último nucleotídeo a 3' do "exon 1" possui um grupo 3'-OH livre para atacar o fosfato da ligação fosfodiéster entre a guanina a 3' (AG) e o primeiro nucleotídeo do "exon 2". (c) o resultado é o rompimento da ligação intron-"exon 2" e a junção dos exons 1 e 2.

Nem todos os introns encontrados na natureza são sujeitos a *splicing* pelo mecanismo descrito acima. Na verdade, são conhecidos três tipos de introns: **introns de *splicing* nuclear**, introns de **grupo I** e de **grupo II**. Os introns dos grupos I e II são os chamados auto-catalíticos porque não necessitam de nenhum complexo enzimático especial para a excisão intrônica, ou seja, o *splicing* é uma propriedade intrínseca a esses RNAs (**FIGURA 2**). Já os ditos introns nucleares necessitam de uma maquinaria de remoção do intron, o spliceossomo, encontrada no núcleo das células eucarióticas.

O *splicing* dos introns de grupos II e do tipo nuclear ocorrem da maneira descrita acima (revisto em Gordon et al (2000), Butcher e Brow (2005)), a partir de um ataque nucleofílico de uma adenina intrônica. Já os introns do grupo I sofrem *splicing* através de um processo ligeiramente diferente no qual a primeira reação de transesterificação envolve um grupo 3'-OH livre de uma guanina exógena ao pré-RNA_m (cofator), portanto não formando um laço intrônico (**FIGURA 2**). Magnésio atua como cofator, quebrando a ligação exon 1/intron em uma reação de transesterificação. Em seguida, o *splicing* procede como para os grupos II e nuclear, o grupo 3'-OH livre do exon ataca a ligação intron/exon 2, liberando o intron. É interessante notar que não há gasto de ATP nestas reações.

FIGURA 2. Tipos de intron conforme o mecanismo de excisão do intron e junção dos exons.

Tanto no grupo II quanto no *splicing* nuclear o *splicing* inicia-se com um ataque nucleofílico de um grupo 2'-OH de uma adenina intrônica. Já no grupo I, uma guanina exógena é que realiza o primeiro passo do *splicing* com um ataque nucleofílico de um grupo funcional 3'-OH. Tanto no grupo II quanto no grupo I o *splicing* ocorre independentemente de uma maquinaria exógena, somente a formação de estruturas secundárias do intron são requeridas. Extraída de Weiner (1993)

Vários estudos têm mostrado que as reações de transesterificação que ocorrem nos introns de grupos I, II e nuclear são mecanisticamente idênticas. Tanto os introns de grupo I (Piccirilli et al (1993), Weinstein et al (1997)), do grupo II (Sontheimer et al (1999)) e nuclear (Sontheimer et al (1997)) fazem uso de um íon magnésio coordenado

pelo grupo abandonador (o oxianion 3' do exon 1) para estabilizar o surgimento de carga negativa no estado de transição durante a catálise, possivelmente nos dois passos (revisado em Gordon et al (2000), Butcher e Brow (2005), Valadkhan (2005)). Corroborando essas idéias, tanto em introns do grupo II (que formam uma estrutura secundária) como na snRNP U6 do spliceossomo (que forma o sítio catalítico, ver abaixo) foram encontrados possíveis sítios de ligação para estes íons metálicos (revisado em Butcher e Brow (2005)). Além disso, Gordon et al (2000) salientam outra similaridade entre o *splicing* de grupo II e nuclear, a mudança conformacional que é um passo limitante entre as duas catálises. Tais similaridades dos mecanismos de *splicing* dos diferentes grupos sugeriram que os diferentes grupos de introns evoluíram a partir de um ancestral comum e que RNA, e não proteínas, tem um importante papel catalítico. Esta idéia da ancestralidade comum é ainda bastante controversa, já que a possibilidade de uma convergência evolutiva ou “determinismo químico” não pode ser excluída (Weiner (1993)).

Entre os introns de *splicing* nuclear, a grande maioria é flanqueada por GT..AG e é processada pelo spliceossomo principal (*major spliceosome*). Uma pequena fração dos introns que sofrem *splicing* são flanqueados por AT..AC e são processados por outro spliceossomo (*minor spliceosome*). Nesta tese, somente são tratados os introns de *splicing* nuclear processados pelo spliceossomo principal. Para revisão sobre os diferentes tipos de spliceossomo, veja Burge et al (1999).

1.2 Sinais básicos de *splicing* no pré-RNAm

É notável que um gene sempre seja editado da mesma forma, indicando que de alguma forma a célula reconhece precisamente a borda entre exons e introns (mas veja **seção 1.7** abaixo). Atualmente, possuímos um conhecimento razoável acerca do processo de *splicing*. Por exemplo, sabemos que em vários organismos, senão em todos, a borda exon/intron é quase sempre marcada pela presença de nucleotídeos conservados: **GT** na extremidade 5' do intron e **AG** na extremidade 3' (~99.6% das bordas exon/intron de mamíferos, segundo dados recentes de Sheth et al (2006)). Outro sinal que parece ser importante para a demarcação da borda 3' é a existência de um **trato degenerado rico em pirimidinas (Y)** que se inicia um nucleotídeo antes (na direção 5') do sítio 3'-AG e se estende por até 20 nucleotídeos na direção 5' (**FIGURA 3**). Por fim, entre estes sinais básicos está o **sítio do ponto de ramificação** que abriga a adenina que promoverá o ataque nucleofílico à borda exon 1/intron na ponta 5' do intron (**FIGURA 3**) e que também é reconhecido pela maquinaria que efetua o *splicing*. Em mamíferos este sinal pode ser bastante degenerado e sua posição pode ser bem variável em relação à distância dos sítios de *splice* 5' e 3'.

FIGURA 3. Sinais básicos de *splicing* em introns nucleares.

O esquema mostra os sítios de *splice* 5' e 3', trato de polipirimidina (poli-Y) e a adenina no sítio do ponto de ramificação.

1.3 O spliceossomo, uma máquina extraordinária: o mecanismo canônico de *splicing*

If there was a better way to go then it would find me

I can't help it, the road just rolls out behind me

Be kind to me or treat me mean

I'll make the most of it, I'm an extraordinary machine

“Extraordinary machine”, Fiona Apple

1.3.1 A máquina

O núcleo do spliceossomo é constituído por 5 partículas (**U1, U2, U4, U5 e U6**) chamadas de pequenas ribonucleoproteínas nucleares (*small nuclear ribonucleoproteins*, **snRNPs**) que, conforme o modelo canônico atualmente aceito, se ligam em uma sequência ordenada de eventos ao RNAm, formando complexos de snRNPs distintos a cada momento. Cada snRNP U é formada por um ou dois **snRNAs** e várias (>3 e <20) **proteínas associadas**, como por exemplo as proteínas U1C, U1A, U1-70 kDa que se ligam a U1 (Rappsilber et al (2002), Zhou et al (2002)). Em adição a essas snRNPs, também se ligam fatores auxiliares importantes (não-snRNPs, como proteínas SR, ver por exemplo Fu (1995b)), constituindo o spliceossomo completo. Uma descrição detalhada da estrutura e outras características das snRNPs U e de alguns fatores auxiliares pode ser encontrada em Green (1991), Krämer (1996). Estudos

recentes têm mostrado que o spliceossomo deve ser a maquinaria celular mais complexa conhecida (Nilsen (2003)), pelo menos em termos de quantidade de proteínas diferentes associadas (291 proteínas distintas, Rappsilber et al (2002); ~145, Zhou et al (2002)). É interessante notar que nestes estudos não só proteínas associadas a *splicing* e processamento de RNA foram encontradas, mas também associadas a transcrição, exportação de RNAm e até tradução (Rappsilber et al (2002), ver **seção 1.7.2.3**).

1.3.2 Modelo canônico de montagem do spliceossomo

Conforme o modelo corrente, o spliceossomo é uma partícula montada passo a passo sobre o pré-RNAm a ser processado (**FIGURA 4**, para extensas revisões ver Moore et al (1993), Burge et al (1999), Brow (2002)). A primeira snRNP a se ligar ao RNAm é **U1** que reconhece o sítio de *splice* 5'-GT através do estabelecimento de pontes de hidrogênio com cerca de 4-7 nucleotídeos nesse sítio (Seraphin e Rosbash (1989), revisado em Reed (2000b)). O reconhecimento inicial do sítio de *splice* 3' (sítio do ponto de ramificação + trato de polipirimidina + AG) se dá pela ligação de um fator auxiliar, **U2AF** (Ruskin et al (1988), Merendino et al (1999), Wu et al (1999), Zorio e Blumenthal (1999)). Em seguida, **U2** liga-se ao sítio do ponto de ramificação de forma instável (Das et al (2000)). Esta etapa, onde os sítios de *splice* 5' e 3' são reconhecidos, é conhecida como a formação do **complexo de engajamento** (*commitment complex*) em levedura (Seraphin e Rosbash (1989), revisado em Rosbash e Séraphin (1991)) ou **complexo E** (*early*) em metazoários (Michaud e Reed (1991), Jamison et al (1992), Michaud e Reed (1993)). Para que a cisão da borda exon 5'/intron, a remoção do intron e a ligação dos exons ocorram, é necessário haver a aproximação dos nucleófilos e eletrófilos envolvidos. Há evidências de que já no complexo E haja interações entre U1

e U2AF, promovidas por fatores de *splicing* conhecidos como proteínas SR (revisado em Fu (1995a), Graveley (2000)), aproximando os dois exons a serem unidos (revisado em Rosbash e Séraphin (1991), Michaud e Reed (1993)).

A ligação estável de **U2** ao RNAm leva à formação do **complexo A** que consome ATP para permitir rearranjos de interações entre os componentes (revisado em Burge et al (1999)). As snRNPs U1 e U2 também são ligadas através de proteínas neste complexo (Fu e Maniatis (1992)).

O próximo passo do mecanismo de *splicing* é a ligação de três snRNPs (**U4/U6.U5**) ao complexo U1/U2, formando o **complexo B** (Konarska e Sharp (1987)). Uma vez que este pentâmero tenha se montado, U1 se desliga do sítio de *splice* 5', dando lugar a uma interação deste com U6 (Konforti et al (1993)), o pareamento entre U4 e U6 é desfeito e um pareamento entre U6 e U2 é estabelecido (revisado em Murray e Jarrell (1999)). O complexo resultante é o **complexo C** composto por U2, U5 e U6, onde a catálise das transesterificações ocorre.

FIGURA 4. Mecanismo básico de *splicing* e os complexos *spliceossomais* formados.

Extraída de Graveley (2000).

Aparentemente, a snRNP U6 é o núcleo catalítico do spliceossomo, embora um envolvimento de U2 e proteínas associadas não seja excluído (veja **seção 1.4.8**). Várias evidências suportam esta idéia: U6 é a snRNP mais conservada de todas (revisado em Weiner (1993)), o extenso pareamento entre U6 e U4 deve ser desfeito logo antes do primeiro passo catalítico, mutações em dois motivos conservados de U6 bloqueiam o *splicing* e U6 faz contato com o sítio de *splice* 5'. Um outro argumento muito

interessante é a observação de um intron no snRNA de U6 de *S. pombe*, sendo que nenhum outro snRNA possui introns. Este intron pode ser remanescente de uma excisão incompleta realizada por U6. Um evento de retroposição associado pode ter fixado este RNA alterado ao genoma (revisado em Green (1991)).

Todos os eventos acima permitem a execução do primeiro passo catalítico, no qual a primeira transesterificação, ou seja, o rompimento da ligação fosfato exon 1/intron seguida da ligação da ponta 5' livre do intron ao 2'-OH livre da adenina do sítio do ponto de ramificação, ocorre. O segundo passo catalítico ocorre já no spliceossomo ativado composto de U2, U6 e possivelmente de U5.

1.3.3 Seleção dos sítios de splice: definição de introns versus definição de exons

1.3.3.1 Os introns de mamíferos são tipicamente muito maiores que os exons

Embora o modelo canônico de *splicing* explique precisamente como pode ser realizada a junção de exons pelo reconhecimento de sítios de *splice* **através de um intron curto**, fato que o caracteriza como mecanismo de **definição de intron**, ele não é adequado para explicar o processamento de pré-RNAs de organismos que possuam introns longos (alguns Kb). Nesse caso, o estabelecimento de interações entre os sítios de *splice* 5' e 3' deve ser bastante problemático. Embora quase todos os eucariotos tenham seus genes divididos em exons e introns, há uma vasta diversidade na forma de organização destes segmentos. Por exemplo, levedura possui pouquíssimos introns sendo que muitos dos genes possuem um único intron, fugu possui introns bem mais

curtos que seus correspondentes em mamíferos, enquanto *Drosophila* e *Caenorhabditis* possuem exons mais longos que os introns, sendo que estes são curtos (~100 nt). Os introns de mamíferos em geral podem ter até alguns milhares de bases. Assim, é bem provável que existam diferentes mecanismos de *splicing* conforme a estrutura exon/intron.

O modelo canônico de *splicing* apresentado acima foi amplamente baseado em estudos realizados em levedura ou em mini-genes artificiais metazoários contendo apenas um intron curto e sítios de *splice* fortes. Robberson et al (1990) observaram que a ligação de U2 ao sítio de *splice* 3' requeria a existência de um sítio de *splice* 5' (flanco 3' do exon) do mesmo exon (veja **FIGURA A** para orientação dos sítios de *splice*), sugerindo que de alguma forma U1 participava na interação de U2. Em seguida, utilizando um mini-gene contendo apenas um exon, os autores observaram que a ligação de U1 no sítio de *splice* 5' requeria ligação de U2 a um sítio de *splice* 3' do mesmo exon. Esses resultados indicaram que deve haver comunicação entre U2 e U1 **através do exon**, já que a existência de uma estrutura sítio de *splice* 3' + exon + sítio de *splice* 5' nessa polaridade era crucial para a montagem do spliceossomo e *splicing*. Outro fator que interferiu no *splicing* foi o **tamanho do exon**, ao se alongá-lo acima de 300 nt não ocorria *splicing*, provavelmente por afetar a interação entre U1 e U2. Com base nesses resultados, os autores propuseram que o reconhecimento dos sítios de *splice* se dá através do exon, denominando o mecanismo de **definição de exon (FIGURA 5)**.

FIGURA 5. Modelo de definição de exons proposto por Robberson et al (1990).

As caixas cinza representam exons e as linhas horizontais introns. As esferas representam U1 (1) e U2 (2). O reconhecimento do sítio de *splice* 5' se daria através do exon, após reconhecimento inicial do sítio de *splice* 3'. As setas curvadas indicam interações entre U1 e U2, primeiro através do exon e depois através do intron. Extraído de Robberson et al (1990).

A **FIGURA 6** ilustra as interações através do intron e através do exon simultaneamente para comparação das estruturas formadas nos diferentes complexos *spliceossomais*.

FIGURA 6. Reconhecimento de sítios de *splice* através do intron e através do exon.

À esquerda o mecanismo de definição de introns e à direita a definição de exon.

Extraído de Matlin et al (2005).

De acordo com o modelo da **FIGURA 5**, U2 iniciaria a definição do exon (interno) ao se ligar ao sítio de *splice* 3' (flanco 5' do exon). Através de um mecanismo de varredura (setas) o sítio de *splice* 5' no flanco 3' do exon dentro de uma distância menor que 300 nt seria procurado.

A interação entre U2 e U1 através do exon poderia se dar através de uma **ponte de interações proteína-proteína (FIGURA 7)**. Tal ponte poderia ser formada por **proteínas SR** que possuem **domínios RS** para ligação **proteína-proteína** e têm um

importante papel no processo de *splicing*. Essas proteínas estão envolvidas no reconhecimento de sítios de *splice*, estabilização de interações de componentes do spliceossomo e na comunicação entre componentes do spliceossomo ligados aos sítios de *splice* (ver **seção 1.4**). Embora muitos dos dados dessas interações tenham sido originalmente interpretados no âmbito da comunicação através do intron, as mesmas proteínas poderiam estar envolvidas em interações através do exon (ver **seção 1.4.3.3** para detalhes). Uma evidência de interação através do exon vem do trabalho de Hoffman e Grabowski (1992), no qual utilizando-se *cross-linking* de snRNPs ao RNA, observou-se que a ligação de U1 coincidia com a seleção do sítio de *splice* 3' do mesmo exon por U2AF de acordo com o esperado para definição de exon. A adição de U1 e U2AF purificados não foi suficiente para reconstituir *splicing*, indicando que a ligação entre U1 e U2AF requeria algum fator intermediário. Assim, esses resultados sugeriram que havia uma interação através do exon ocorrendo mediada por fatores desconhecidos (veja também a **seção 1.4.3**).

FIGURA 7. Modelo de interações através do exon, ligando os sítios de *splice* e envolvendo ESEs.

Sub-sequências curtas localizadas em exons denominadas *exonic splicing enhancers* (ESEs, veja o texto) podem servir de sítio de ligação a fatores de *splicing* como

proteínas SR (pSR). Essas proteínas, por sua vez, podem interagir entre si e com componentes basais do spliceossomo, formando uma ponte protéica que define o exon.

Essas proteínas se ligariam a elementos especiais localizados nos exons (**FIGURA 7**) denominados promotores de *splicing* exônicos (*exonic splicing enhancer*, **ESE**, para revisão ver Cartegni et al (2002), Matlin et al (2005)) fazendo uma ponte protéica de contatos através do exon. De fato, Tanaka et al (1994) observaram que o sítio de *splice* 5', um ESE e o sítio de *splice* 3' atuavam de forma compensatória na seleção de um exon de IgM e propuseram que o reconhecimento do exon se daria por uma soma destes sinais. Xu et al (1993) ao estudarem um ESE no gene de troponina T cardíaca, observaram que sua presença só afetava o reconhecimento do sítio de *splice* 3' (no flanco 5' do exon com o ESE), sugerindo que o ESE auxiliava no reconhecimento deste sítio diretamente. No entanto, como esse sítio de *splice* era forte e supostamente não precisaria ser reforçado por um ESE, uma explicação alternativa para o modo de atuação deste elemento promotor de *splicing* no exon poderia ser que ele participa do estabelecimento de uma ponte protéica de definição através do exon.

Deve-se notar que embora o modelo vise particularmente os exons internos flanqueados por sítios de *splice* 5' e 3', o primeiro e último exons também poderiam ser processados através de um mecanismo de definição de exon modificado, no qual as interações com os sítios de *splice* seriam substituídas por interações com proteínas depositadas nas extremidades do pré-RNAm (ver Zheng (2004)).

Um aspecto interessante do modelo é que ele oferece várias predições, que de fato são observadas (revisado em Berget (1995)):

- A definição de exons implica que, assim como para introns muito longos, distâncias grandes entre os sítios de *splice* através de exons seriam problemáticas. De fato, observou-se que a inclusão de exons acima de 500 nt é extremamente deficiente (Sternier et al (1996)). No entanto, quando os exons longos eram flanqueados por introns curtos, a inclusão do exon se tornava mais eficiente, provavelmente porque os sítios de *splice* eram reconhecidos através do intron. Ou seja, o modelo de definição de exon introduziu a idéia que os sítios de *splice* são reconhecidos através do segmento mais curto disponível seja um exon ou um intron. De acordo com essa idéia, Fox-Walsh et al (2005) utilizaram estudos de cinética de ligação de componentes do spliceossomo ao RNAm para determinar o comprimento máximo em que ocorre definição de intron ou de exon. Em mini-genes com introns acima de 200 nt, notou-se uma mudança na cinética de ligação que indicava uma mudança na forma de definição dos sítios de *splice*, confirmando que o comprimento do intron é crucial para determinar o modo de reconhecimento. Além disso, os autores notaram que um exon flanqueado por sítios de *splice* sub-ótimos era frequentemente não reconhecido quando flanqueado por introns longos. Sua inclusão era resgatada quando os introns eram encurtados, de acordo com o observado por Sternier et al (1996).
- em mamíferos, mutação de sítios de *splice* 5' resultaria em exclusão do exon e não retenção de intron, de forma que o tipo de *splicing* alternativo mais observado em vertebrados seria a inclusão alternativa de exon, o que de fato é observado (ver adiante a **seção 1.7**)
- dado o limite de comprimento do exon para sua perfeita definição, a maioria dos exons de metazoários seria <300 nt, também observado

1.3.3.2 As interações através do exon devem ser convertidas para interações através do intron

Embora segundo o modelo de definição de exons os sítios de *splice* sejam reconhecidos através dos exons, as reações de transesterificação devem ocorrer através dos introns (**FIGURA 1**). Dessa forma, em algum momento as interações através do exon devem ser convertidas em interações através do intron (**FIGURA 5**). Não se sabe ainda exatamente o momento em que isto ocorre, mas provavelmente só deve ocorrer após a entrada de U2 no pré-spliceossomo, ou seja, da formação do complexo A em diante, quando os dois sítios de *splice* 5' e 3' já devem estar estavelmente ligados a U1 e U2.

Neste ponto, é interessante comentar o trabalho de Lim e Hertel (2004) que, utilizando um ensaio de armadilha cinética que é capaz de bloquear o *splicing* de pré-RNAs com diferentes complexos pré-spliceossomo montados, mostraram que até a formação do complexo A é possível alterar a escolha do sítio de *splice*. A partir da formação deste complexo, quando U2 é “travado” no pré-RNA a escolha parece ser irreversível. Em outras palavras, o **engajamento do spliceossomo** com um dado par de sítios de *splice* ocorre no complexo A. Os autores sugerem que haja dois passos de engajamento do *splicing*, um que ocorre na **formação do complexo de engajamento/E** que engaja o pré-RNA na via geral de *splicing* e outro que **engaja pares específicos de sítios de *splice***. Esta idéia é condizente com a ocorrência da transição de interações através do exon para interações através do intron após a formação do complexo A, já que este evento deve envolver o rearranjo das interações, possivelmente também daquelas com o pré-RNA.

No entanto, estudos de *trans splicing*, ou seja, em que ocorre junção de exons localizados em moléculas de pré-RNA distintas (em *trans*) indicam que o engajamento dos sítios de *splice* pode ocorrer mais adiante, na formação do complexo B

(Chiara e Reed (1995)). Nesse estudo, os autores observaram que mesmo após a montagem do complexo A no sítio de *splice* 3' de um exon em uma molécula de pré-RNAm ainda era possível haver “escolha” de um sítio de *splice* 5' remoto (em outra molécula de pré-RNAm), mostrando que mesmo após a ligação de U2 ainda é possível selecionar o sítio de *splice* 5'.

1.3.3.3 Polaridade da definição de exon

Deve-se notar que embora definição de exon na polaridade 5'-3' tenha sido observada em alguns trabalhos (por exemplo Robberson et al (1990), Zeng e Berget (2000)) e de certa forma seja uma parte importante do modelo de definição de exon, pelo menos na sua formulação original (Robberson et al (1990), Berget (1995) e **FIGURA 5**), outras observações de reconhecimento do exon no sentido 3'-5' também existem (Hoffman e Grabowski (1992), Millhouse e Manley (2005)). Assim, embora o modelo de definição de exon seja suportado por diversas evidências, vários detalhes ainda devem ser esclarecidos.

1.4 O mecanismo de *splicing* em detalhes

1.4.1 Interações através do intron

Um ponto crucial no mecanismo canônico de *splicing* descrito acima é a comunicação entre U1 e U2 para que ocorra justaposição dos sítios de *splice* 5' e 3'. Vários estudos mostraram evidências indiretas de que de fato pode haver um contato entre U1 e U2 (e U2AF) logo nos complexos pré-spliceossomo iniciais (para revisão, ver Rosbash e Séraphin (1991)). Por exemplo, mesmo na ausência de um sítio de *splice* 5' foi observada ligação da snRNP U1, provavelmente a algum outro componente do spliceossomo (Zillmann et al (1987)). Michaud e Reed (1993) mostraram que complexos E podem ser montados independentemente em sítios de *splice* 5' e 3', sendo denominados E5' e E3', mas que a montagem do complexo E era mais eficiente quando os sítios de *splice* 5' e 3' estavam juntamente presentes. Estes resultados sugeriram que há uma interação física entre os sítios de *splice* para que ocorra uma colaboração na montagem dos complexos E5' e E3'. Além disso, observou-se que a ligação de U2AF no sítio de *splice* 3' é favorecida pela presença do sítio de *splice* 5' (Côté et al (1995)). Uma das comprovações mais diretas dessas interações veio mais recentemente; Kent e MacMillan (2002) monitoraram diversos contatos entre os componentes do spliceossomo durante a formação do complexo E e mostraram que em estágios bastante precoces da montagem do spliceossomo há inúmeros contatos ($< 20 \text{ \AA}$) entre os sítios de *splice* 5' e 3' intermediados por diversos componentes do spliceossomo.

1.4.1.1 O reconhecimento do sítio de *splice* 3' requer U2AF e outras proteínas auxiliares

Atualmente, diversos detalhes das interações entre os fatores de *splicing* que reconhecem os sítios de *splice* 5' e 3' são conhecidos. Para compreender essas interações é importante fornecer alguns detalhes sobre o mecanismo de *splicing* descrito acima.

Em realidade, o reconhecimento do sítio de *splice* 3', mais precisamente do sítio do ponto de ramificação, não envolve apenas U2AF, tanto em levedura quanto em metazoários, mas uma intrincada rede de fatores auxiliares. Em **metazoários**, U2AF é um heterodímero (**U2AF⁶⁵** + **U2AF³⁵**), sendo que, segundo atualmente aceito, a sub-unidade de 65 kD de U2AF liga-se ao trato de polipirimidina formando o **complexo E**. Em introns com trato de poli-pirimidina curtos, a sub-unidade de 35 kD liga-se ao sítio de *splice* 3'-AG (introns de *splicing* dependente de AG, Merendino et al (1999), Wu et al (1999), Zorio e Blumenthal (1999)). Em **levedura**, é conhecida uma proteína homóloga à sub-unidade U2AF⁶⁵ de metazoários, a **MUD2** que possui um papel bastante similar (Abovich et al (1994)). Ainda no complexo de engajamento/E, após o reconhecimento do trato de polipirimidina por U2AF⁶⁵/MUD2, outro fator não-snRNP, **SF1 em metazoários** e **BBP em levedura**, também reconhece o sítio do ponto de ramificação, interagindo com U2AF⁶⁵/MUD2 (Selenko et al (2003) e referências citadas em Reed (2000b)).

Embora o mecanismo de reconhecimento do sítio de *splice* 3' em leveduras e em metazoários seja muito parecido, há uma diferença funcional muito importante. Mutações no sítio do ponto de ramificação de leveduras podem mesmo inviabilizar o *splicing*, mas em mamíferos, pontos de ramificação crípticos são ativados. Coerentemente, o sítio do ponto de ramificação em levedura é extremamente

conservado, enquanto em metazoários é bastante degenerado (Green (1991)). Outra diferença é a aparente falta de um correspondente de U2AF³⁵ em levedura, indicando que há diferenças no mecanismo de reconhecimento do sítio 3' em relação aos metazoários. No entanto, há evidências que mostram que U2AF³⁵ não é um fator de *splicing* essencial (Zamore e Green (1991)). Isso deve ser principalmente verdadeiro para introns com trato de polipirimidina forte uma vez que U2AF³⁵ é importante para reconhecer o sítio 3'-AG e que estes nucleotídeos são principalmente importantes nos introns com trato de polipirimidina fraco (Merendino et al (1999), Wu et al (1999), Zorio e Blumenthal (1999)).

A **transição do complexo E para A** envolve uma substituição da interação entre SF1/BBP e o pré-RNA por uma **ligação estável de U2** ao sítio do ponto de ramificação (**FIGURA 4**). Este passo envolve a participação de ATPases específicas (Burge et al (1999)). Além do demonstrado papel de domínios RS (tanto de U2AF⁶⁵ como de outras proteínas SR) de recrutar U2 (revisado em Will e Lührmann (1997)), esta transição parece envolver uma ATPase (UAP56, Fleckner et al (1997)). Além de UAP56, Gozani et al (1998) mostraram que uma proteína associada a U2, SAP 155 interage diretamente com U2AF⁶⁵ e com o RNA, no sítio do ponto de ramificação e deve ser muito importante para o recrutamento de U2 (ver revisão de Staley e Guthrie (1998) sobre o papel das ATPases no rearranjo do spliceossomo).

1.4.1.2 As interações através do intron ocorrem bastante cedo (complexo de engajamento/E)

Os componentes do spliceossomo mencionados acima que se ligam ao pré-RNA no complexo E/engajamento promovem interações entre os sítios de *splice* 5' e 3' já nessa etapa. Wu e Maniatis (1993) mostraram que a interação entre os sítios de

splice 5' e 3' se dá através da ligação de **proteínas SR SC35** e **ASF/SF2** à proteína U1-70K, componente da snRNP U1, posicionada no sítio de *splice* 5' e **U2AF**³⁵ posicionado no sítio do ponto de ramificação (**FIGURAS 8 e 9**, revisado em Reed (2000b)). Esta interação é tão forte que foi até mesmo observada entre sítios de *splice* 5' e 3' em *trans*, ou seja, entre sítios de *splice* localizados em moléculas de RNA distintas (Bruzik e Maniatis (1995)). Outra interação demonstrada entre os sítios de *splice* 5' e 3' ainda no complexo E/complexo de engajamento se dá através de **PRP40 em levedura** (Abovich e Rosbash (1997)) e possivelmente de **FBP11 em metazoários** (**FIGURA 8**, revisado em Reed (2000b)).

FIGURA 8. Interações através do intron no complexo de engajamento em *S. cerevisiae* e no complexo E de metazoários.

(a) *Complexo de engajamento*: proteínas associadas à snRNP U1 e proteínas ligadas ao cap (CBC20 e CBC80) interagem com o pré-RNA na região contendo o sítio de *splice* 5'. A *branchpoint binding protein* (BBP) reconhece o sítio do ponto de ramificação (BPS). Interações proteína-proteína entre BBP e MUD2, e entre BBP e a proteína da snRNP U1, PRP40, podem funcionar como uma ponte entre as pontas 5' e 3' do intron (Abovich e Rosbash (1997)).

(b) *Complexo E*: metazoários contêm homólogos de algumas das proteínas associadas à snRNP U1 sugerindo que interações similares entre RNA-proteína podem ocorrer no

sítio de *splice* 5' no complexo E. Em metazoários, mBBP (SF1), U2AF⁶⁵ e U2AF³⁵ reconhecem o BPS, o trato de polipirimidina (py tract), e o AG no sítio de *splice* 3', respectivamente. A conexão entre os sítios de *splice* 5' e 3' no complexo E pode envolver as mesmas interações do complexo de engajamento (PRP40/FBP11). Proteínas SR também podem atuar como uma ponte através do intron no complexo E (Wu e Maniatis (1993)). Extraída de Reed (2000b).

FIGURA 9. Interações entre os sítios de *splice* 5' e 3' no complexo E.

Os sítios de *splice* 5' e 3' são diretamente ligados por pareamento de bases a U1 e U2AF, respectivamente. Diversos fatores não-snRNPs como SF1, FBP11 e SC35 promovem interações através do intron já no complexo E. Extraída de Kent e MacMillan (2002).

1.4.1.3 O complexo A também apresenta interações entre os sítios de *splice 5' e 3'*

Embora estas interações iniciais possivelmente contribuam para aproximar as extremidades do intron, elas não são permanentes. Na formação do complexo A, SF1/BBP se desligam do sítio do ponto de ramificação e dão lugar a uma interação direta entre U2 e o sítio do ponto de ramificação com o auxílio de uma ATPase contendo o motivo DExH/D (revisado em Schwer (2001)). Ao mostrarem que a ligação de U2 ao RNA requeria a presença de U1, Barabino et al (1990) sugeriram que esta interação poderia se dar diretamente ou através de um fator desconhecido. De fato, a **proteína SR SC35** promove uma interação entre **U1** e **U2**, como demonstraram Fu e Maniatis (1992). Também foi observado que em *S. pombe* uma interação mais estável se dá no complexo A com a participação do fator não-snRNP **PRP5** que parece ligar U1 e U2 diretamente (Xu et al (2004)). Estas interações também foram observadas em células metazoárias HeLa. Assim, o papel de PRP5 seria o de promover a ligação estável de U2 ao RNA e ao mesmo tempo servir como ponte de ligação dos sítios de *splice 5' e 3'*.

1.4.2 O papel das proteínas SR nas interações através dos *introns e na montagem do spliceossomo*

A interação descrita acima entre U1 (através de sua proteína constituinte U1-70 kDa) e U2AF⁶⁵ que ocorre já no complexo E/complexo de engajamento envolve dois fatores de *splicing* protéicos, **SC35** (*spliceosomal component 35*) e **SF2/ASF** (*splicing factor 2/alternative splicing factor*) (Wu e Maniatis (1993)). Estas **proteínas auxiliares** fazem parte da **família SR** e possuem papéis importantes no *splicing* de **metazoários**

(Fu (1995a), Graveley (2000), Hertel e Graveley (2005)), não tendo sido descritas em leveduras até hoje. O nome SR provém do fato que todas as proteínas da família possuem um domínio de ligação a RNA rico em serinas e argininas (**domínio RS**) que é encontrado em um grande número de proteínas diferentes associadas ao processamento de RNA. Em geral, estas proteínas exercem uma ação estimulatória no *splicing* ao recrutar snRNPs e ancorá-los ao RNA, auxiliando no reconhecimento de sítios de *splice* e ao promover a interação entre componentes do spliceossomo como por exemplo no tri-snRNP U4/U6.U5. Uma revisão sobre os diversos papéis e características das proteínas SR pode ser encontrado em Fu (1995a), Graveley (2000).

Como estas proteínas estão envolvidas em vários pontos da montagem do spliceossomo, elas possuem um importante papel na regulação do processo de *splicing*. Este assunto é tratado em detalhes posteriormente na **seção 1.7**.

1.4.3 As interações através dos exons também envolvem proteínas SR (modelo de definição de exon)

1.4.3.1 Comunicação entre sítios de *splice* através do exon

O exon E4 do gene de preprotachiquinina de rato provê um exemplo de reconhecimento de sítios de *splice* através do exon já que sua inclusão é diretamente aumentada pelo fortalecimento do sítio de *splice* 5' (Hoffman e Grabowski (1992)). Utilizando este modelo, Hoffman e Grabowski (1992) observaram que a seleção do sítio de *splice* 3' do exon E4 era aumentada pela ligação de U2AF⁶⁵, e facilitada pela ligação de U1 ao sítio de *splice* 5' (a presença de um sítio de *splice* 5' era necessária e a porção 5' terminal de U1 era requerida para se observar inclusão do exon). Neste trabalho,

embora os autores tenham demonstrado uma comunicação entre componentes do spliceossomo nos sítios de *splice* 3' e 5' do mesmo exon, eles não identificaram diretamente as interações entre U2AF⁶⁵ e U1. Algumas possibilidades foram levantadas como um contato direto entre proteínas componentes da snRNP U1 e U2AF⁶⁵ ou através de uma ponte de outras proteínas através do exon. Wu e Maniatis (1993) sugeriram que a proteína SR SC35 poderia interagir com a proteína U1-70K componente de U1 no sítio de *splice* 5' e interagir com U2AF³⁵ no sítio de *splice* 3', na extremidade 5' do mesmo exon. O modelo de interações proposto por Hoffman e Grabowski (1992) é apresentando na **FIGURA 10**.

FIGURA 10. Modelo de interações através do exon proposto por Hoffman e Grabowski.

“A” ligação de U1 ao sítio de *splice* 5' no flanco 3' do exon, “B” interação de U1 e U2AF⁶⁵ através do exon. O sinal “?” indica que a comunicação entre U1 e U2AF⁶⁵ pode ser direta ou através de outras proteínas não identificadas no trabalho, “C” interação de U2AF⁶⁵ com o trato de polipirimidina a 5' do exon. Wu e Maniatis (1993) sugeriram que “?” poderia ser SC35. Extraída de Hoffman e Grabowski (1992).

1.4.3.2 Proteínas SR poderiam formar pontes através do exon

Nas situações descritas acima, as proteínas SR interagem com outras proteínas através de seus domínios SR. No entanto, essas proteínas também possuem um outro domínio que é responsável por **ligação ao RNA** (domínios **RRM**, *RNA recognition motif*). Através destes domínios RRM, essas proteínas podem ligar-se a sub-sequências curtas do RNAm. Essas curtas sequências, tipicamente localizadas em exons de metazoários são **sítios extremamente degenerados** que ao ligar proteínas SR promovem o reconhecimento dos sítios de *splice* sendo denominadas promotoras de *splicing* exônicas (*exonic splicing enhancer*, **ESE**, para revisão ver Cartegni et al (2002), Matlin et al (2005)). A base bioquímica para esse reconhecimento de sequências tão degeneradas provém do fato que arginina pode formar pontes de hidrogênio com diferentes sequências de DNA, reconhecendo virtualmente qualquer motivo (revisado em Hertel e Graveley (2005)). Um modelo de como ESEs poderiam estar envolvidos na formação de uma ponte protéica através do exon foi apresentado na **FIGURA 6**.

Embora a idéia de que o reconhecimento através dos exons envolva a formação de uma ponte protéica entre os sítios de *splice* 3' e 5' seja bastante aceita, evidências diretas ainda são escassas. Uma evidência indireta da ligação provém do trabalho de Lam e Hertel (2002) que observaram que proteínas SR ligadas a ESEs no exon favorecem o reconhecimento simultâneo de dois sítios de *splice* fracos flanqueando um mesmo exon (**FIGURA 11**).

FIGURA 11. Comunicação entre os sítios de *splice* 3' e 5' através do exon proposta por Lam e Hertel (2002).

Tendo observado um recrutamento simultâneo de componentes do spliceossomo por ESEs, Lam e Hertel (2002) propuseram que deve haver uma interação através do exon, ligando os dois sítios de *splice*. Essa interação deve envolver a ligação de proteínas SR através de seus domínios RRM a motivos curtos no exon (ESE). As proteínas SR utilizariam seus domínios RS para estabelecer uma ponte protéica conectando os sítios de *splice*. Extraída de Lam e Hertel (2002).

Uma outra confirmação experimental indireta do papel das proteínas SR na definição de exons foi fornecida por Boukis et al (2004), mostrando que a ligação de U1 ao sítio de *splice* 5' pode ser independente da presença de proteínas SR, mas que a promoção da ligação de U2 por U1 depende destes fatores. Ou seja, deve existir uma interação física entre U2 no sítio de *splice* 3' e U1 no sítio de *splice* 5' flanqueando o mesmo exon, sendo que uma ponte de proteínas SR deve realizar tal conexão. Este resultado é similar ao obtido por Hoffman e Grabowski (1992) onde a adição de U1 e U2AF⁶⁵ purificados não foram capazes de ativar o *splicing*, provavelmente pela falta de um fator que ligasse os dois componentes (a ligação de U1 e de U2AF⁶⁵ ao RNA eram facilitadas reciprocamente). Similarmente, Izquierdo et al (2005) mostraram que o exon

6 do gene *fas* também tem a ligação de U2 ao sítio de *splice* 3' facilitada pela ligação de U1 ao sítio de *splice* 5'. Os autores mostraram que a ligação de uma proteína inibitória (PTB) a uma sequência no exon pode abolir essa cooperação entre U1 e U2, possivelmente por abolir a formação da ponte protéica.

1.4.3.3 As interações através do intron envolvendo proteínas SR podem também ocorrer através do exon

Cabe salientar que a interação de U1 e U2 através de PRP5 observada por Xu et al (2004) descrita na seção “Interações através do intron”, que pode ter um papel na justaposição dos sítios de *splice* 5' e 3' foi observada em células de levedura e em células metazoárias HeLa (Xu et al (2004)). Os autores levantaram a possibilidade de que as interações entre U1 e U2 através de PRP5, no caso da definição de exon, ocorreriam através do exon e não através do intron, embora não haja evidências diretas para suportar esta idéia.

É importante salientar que as interações através do intron descritas anteriormente também poderiam ocorrer através do exon (ver também Reed (2000b)). Por exemplo, embora o trabalho de Wu e Maniatis (1993) sobre o papel de SC35 e ASF/SF2 na comunicação dos sítios de *splice* tenha sido realizado em *Drosophila* e seja sempre citado no contexto de interações através do intron, os próprios autores propuseram que esta comunicação poderia se dar através do exon, para explicar os resultados de Robberson et al (1990) e Hoffman e Grabowski (1992) descritos nas seções **1.3.3.1 e 1.4.3.1.**

1.4.3.4 Junção dos exons e conversão das interações através do exon para através do intron

Um ponto importante no modelo de reconhecimento de sítios de *splice* através dos exons é que em algum momento deve ocorrer uma conversão das interações através do exon para através do intron, já que os sítios de *splice* são unidos através destes segmentos de DNA, de forma a possibilitar a união dos exons. Stark et al (1998) mostraram através de ensaios de cromatografia de afinidade e de microscopia de imageamento por espectroscopia que proteínas SR que ligam a ESEs são capazes de promover a aproximação de dois exons, deixando um laço intrônico exposto.

Zeng e Berget (2000) sugeriram que o domínio C-terminal da RNA polimerase II (CTD, *C-Terminal Domain*) poderia ter um papel na aproximação de dois exons adjacentes. O acoplamento da transcrição e do *splicing* já é bastante conhecido (ver referências citadas em Zeng e Berget (2000), revisão de Maniatis e Reed (2002) e a **seção 1.7.2.3**), embora nem todos os detalhes de como esse acoplamento se dá sejam conhecidos. Zeng e Berget (2000) observaram ativação do *splicing* na presença de CTD somente quando havia um exon interno completo (flanqueado pelos dois sítios de *splice*), de acordo com o modelo de definição de exon. Os autores propuseram que o papel do CTD na definição de exons seria de segurar exons já reconhecidos para trazê-lo perto dos próximos exons, na polaridade 5'-3' (**FIGURA 12**).

FIGURA 12. Possível papel do domínio CTD da polimerase de RNA II na junção de exons durante o *splicing*.

No primeiro passo, o CTD media a associação de fatores de *splicing* no exon recém transcrito, auxiliando o reconhecimento do exon. No passo dois, a polimerase move-se para o próximo exon, trazendo também o exon anterior para perto do próximo exon.

Extraída de Zeng e Berget (2000).

1.4.4 Detalhes do reconhecimento do sítio de splice 5'

1.4.4.1 Proteínas SR estão envolvidas no reconhecimento do sítio de *splice 5'*

Fatores não-snRNP podem ter um papel bastante importante no reconhecimento do sítio de *splice 5'*. Por exemplo, Fu (1993) demonstrou que a proteína SR **SC35 purificada** era capaz de **promover engajamento do *splicing*** por si própria. Zuo e Manley (1994) verificaram que **ASF/SF2 se liga ao sítio de *splice 5'*** de maneira sequência-específica, de tal forma que mutações no sítio levaram a uma redução de até 100 vezes da ligação.

Estas observações levantaram a idéia de que o papel das proteínas SR no reconhecimento do sítio de *splice 5'* seria de auxiliar no recrutamento de U1 ao RNA. De fato, a **interação inicial de U1 com o sítio de *splice 5'*** parece requerer **estabilização por ASF/SF2** (Eperon et al (1993), Kohtz et al (1994), Staknis e Reed (1994)), aumentando sua afinidade a todos os sítios de *splice 5'* disponíveis (Eperon et al (1993)). Note-se que esta estabilização pode não ser obrigatória para todos os introns, como demonstrado por Boukis et al (2004), por exemplo.

Jamison et al (1995) observaram que, reciprocamente, U1 também estabiliza a ligação de ASF/SF2 ao pré-RNA e por isso propuseram que as proteínas SR e U1 se estabilizam mutuamente. De qualquer forma, a **ligação de ASF/SF2 antes de U1** aumenta consideravelmente a **cinética do *splicing*** (Kohtz et al (1994), Jamison et al (1995)). Várias outras proteínas SR podem ter um papel no reconhecimento do sítio de *splice 5'*. Zahler e Roth (1995) mostraram que, como esperado, diferentes proteínas SR têm diferentes especificidades para ligação em sítios de *splice 5'* alternativos. No caso, SRp40 auxiliaria U1 a reconhecer um sítio distal e SRp30b um sítio proximal em

relação ao sítio de *splice* 3'. Esta observação é interessante porque implica que a presença de **diferentes proteínas SR** resultará em **diferentes formas de *splicing***.

Com base nessas observações, acredita-se que na verdade, a **ligação de ASF/SF2** seja o primeiro passo para a formação do complexo E, **anterior à ligação de U1** (Kohtz et al (1994), Jamison et al (1995)) o que explicaria o engajamento de *splicing* promovido por um membro desta família de proteínas observada por Fu (1993), Zuo e Manley (1994).

Em adição ao reconhecimento direto do sítio de *splice* 5' por proteínas SR e à estabilização promovida por elas, Tarn e Steitz (1995) sugeriram que estas proteínas podem auxiliar na escolha de sítios de *splice* 5' autênticos em um segundo passo mais indireto, através de interação com o sítio de *splice* 3'. Barabino et al (1990) observaram que a ligação de U2 ao sítio do ponto de ramificação requeria a ligação de U1, através de um contato direto entre as duas snRNPs ou de um fator até então desconhecido. Tarn e Steitz (1995) mostraram que proteínas SR, além de auxiliarem a ligação de U1 e promoverem a comunicação entre U2 e U1, podem estabilizar a ligação de U2 ao sítio de *splice* 3' e até substituir U1 neste papel, oferecendo uma identidade ao fator desconhecido de Barabino et al (1990). O complexo formado por U2 e proteínas SR pode posteriormente se comunicar com o sítio de *splice* 5' e assim modular a escolha de sítios de *splice* 5' em um passo distinto ao reconhecimento inicial de U1 descrito no parágrafo acima. Note-se que essas observações são plenamente compatíveis com o modelo de definição de exon.

Interessantemente, uma outra proteína SR que também parece poder modular a ligação de U1 ao RNA é U2AF⁶⁵. Embora pelo modelo canônico o papel desta proteína seja de reconhecer o sítio de *splice* 3', Förch et al (2003) mostraram que em sítios de

splice 5' fracos com sequências ricas em uridina próximas, U2AF⁶⁵ pode auxiliar na ligação de U1.

Por fim, o papel mais surpreendente das **proteínas SR** vem da observação que elas podem até mesmo **substituir U1** completamente, como mostraram estudos de complementação de proteínas SR a extratos depletados de U1 (Crispino et al (1994), Tarn e Steitz (1994)). A explicação para esta curiosa observação seria que as proteínas SR auxiliariam U6 a reconhecer o sítio de *splice* 5' em complexos *spliceossomos* tardios, na realidade estabilizando a ligação de U2 ao sítio de *splice* 3' e permitindo assim o estabelecimento das interações U6-U2 e U6-sítio de *splice* 5' (Tarn e Steitz (1994), Tarn e Steitz (1995)). Outra possibilidade é que as proteínas SR estariam ajudando a recrutar U4/U6.U5 (Roscigno e Garcia-Blanco (1995)) e indiretamente auxiliando no reconhecimento tardio do sítio de *splice* 5' (este sítio deve ser reconhecido diversas vezes ao longo da montagem do *spliceossomo*, ver **seção 1.4.4.3**). É possível ainda que o excesso de proteínas SR seja suficiente para promover estabilização mútua da ligação ao sítio de *splice* 5' e assim tornar o reconhecimento do sítio de *splice* 5' independente de U1 (Jamison et al (1995)).

Embora aparentemente uma proteína SR possa ser suficiente para promover *splicing* por si, o papel de U1 parece ser importante no reconhecimento de um sítio de *splice* 5' autêntico, em detrimento de outros sítios disponíveis, ou seja, poderia ser um componente importante da **fidelidade** do processo (Tarn e Steitz (1994)).

1.4.4.2 Os nucleotídeos conservados GT não são necessários para o reconhecimento inicial por U1

A altíssima conservação dos nucleotídeos GT nas posições 0 e -1 do sítio de *splice* 5' sugere que estas bases devem ser imprescindíveis para o reconhecimento deste

sítio. Vários estudos mostram que pelo menos no passo de ligação de U1, os nucleotídeos conservados do sítio de *splice* 5' não são necessários. Por exemplo, Aebi et al (1987) fizeram todas as seis mutações possíveis do sítio de *splice* 5'-GT de um intron de β -globina e observaram que nenhum mutante impediu a execução da primeira clivagem, mostrando que há um outro motivo pelo qual estes nucleotídeos são conservados que não somente para o reconhecimento por U1.

1.4.4.3 O sítio de *splice* 5' deve ser reconhecido diversas vezes

A observação acima pode estar relacionada ao fato de que além do reconhecimento por U1 para formação do complexo E, o sítio de *splice* 5' é reconhecido diversas outras vezes por outros componentes do spliceossomo. Na verdade, há evidências que o sítio é inicialmente reconhecido pela proteína **U1C** componente da snRNP U1 (Du e Rosbash (2002)). Através do uso de oligonucleotídeos complementares ao sítio de *splice* 5' para inibir a ligação de snRNPs e de experimentos de *cross-linking* de snRNPs ao RNA, Konforti et al (1993) mostraram que é necessário haver desligamento da snRNP U1 do sítio de *splice* 5' dando lugar à ligação de U4/U6 para a ocorrência de *splicing*. Neste complexo (B), o sítio de *splice* 5' é diretamente reconhecido pela proteína **PRP8** associada a U5 que se liga a 5'-GT (Maroney et al (2000)) e a própria snRNP **U5** **contata o sítio de *splice* 5'** (Wyatt et al (1992)). Um último contato é realizado por **U6** (veja referências em Staley (1999)) através da ação de uma ATPase (PRP28) contendo o motivo DExD/H (Staley (1999) e revisado em Schwer (2001)).

Em realidade, o papel de U1 no *splicing* de mamíferos ainda é discutido. Vários estudos demonstraram que o evento de clivagem é independente do engajamento do RNA na via de *splicing* (ligação de U1). Coerentemente, outras snRNPs além de U1

parecem ter um papel na seleção do ponto de clivagem do sítio de *splice* 5'. Kandels-Lewis e Séraphin (1993) construíram mutantes do snRNA de **U6** e observaram que alterações do pareamento de bases com o sítio de *splice* 5' resultava em mudanças na **escolha do sítio de clivagem**. Mutações em **U2** (Luukkonen e Séraphin (1998)) e em **U5** (Newman e Norman (1991)) tiveram um impacto considerável na **determinação da posição da primeira clivagem**. Estes dados são coerentes com a observação de Lim e Hertel (2004) que o comprometimento com os sítios de *splice* só se dá no complexo A; a ligação de U1 ao sítio de *splice* 5' pode ser alterada após formação do complexo E e não deve ser crucial para a escolha do sítio de corte. Além disso, foi mostrado que em extratos de células metazoárias **depletados de U1**, mas enriquecidos com a **proteína SR SC35**, mutações no RNA que aumentam a complementaridade com **U6** facilitam o *splicing* (Crispino e Sharp (1995)). Todas estas observações em conjunto sugerem que a snRNP **U1 não é essencial para *splicing*** ou que, pelo menos em mamíferos, deve haver uma **via independente desta snRNP**.

1.4.5 Detalhes no reconhecimento do sítio de splice 3'

1.4.5.1 Estruturação do sítio de *splice* 3' por U2AF⁶⁵

Embora o papel de U2AF⁶⁵ seja plenamente reconhecido no mecanismo de *splicing*, as razões estruturais desse papel não foram imediatamente conhecidas. Recentemente, Kent et al (2003) analisaram dados de raio-X de complexos proteína-RNA para entender o papel de U2AF⁶⁵ na estruturação do sítio de *splice* 3' (**FIGURA 13**). A ligação de U2AF⁶⁵ ao sítio de *splice* 3' do intron justapõe o sítio do ponto de ramificação (UGCUGAC) ao sítio de *splice* (3'-GAC-5') promovendo um contato entre o domínio RS de U2AF⁶⁵ e o RNA, como posteriormente detalhado por Shen e Green

(2004) através de estudos de *cross-linking* de proteínas SR/RNA (revisado em Hertel e Graveley (2005) e ver a **seção 1.4.6**). Dessa forma, esse fator auxiliar deve promover contatos e arranjar o sítio de *splice* 3' de forma a permitir o estabelecimento de interações posteriores.

FIGURA 13. Estruturação do sítio de *splice* 3'.

A ligação de U2AF⁶⁵ ao trato de polipirimidina (PPT) do intron justapõe o sítio do ponto de ramificação (UGCUGAC) ao sítio de *splice* (GAC) promovendo um contato entre o domínio RS de U2AF⁶⁵ e o RNA. RRM: *RNA binding motif*. Extraída de Kent et al (2003).

1.4.5.2 Proteínas SR e ESEs também participam do reconhecimento de sítios de *splice* 3'

Xu et al (1993) estudaram um ESE no gene de troponina T cardíaca e observaram que a presença desse elemento em *cis* promotor de *splicing* só afetava o reconhecimento do sítio de *splice* 3'. Não se pode, no entanto, descartar uma explicação alternativa para essa observação. Como esse sítio de *splice* era forte e supostamente não precisaria ser reforçado por um ESE, uma explicação alternativa para o modo de atuação de um elemento promotor de *splicing* no exon poderia ser que ele participa do estabelecimento de uma ponte protéica de definição através do exon.

Wang et al (1995) observaram um aumento da ligação de U2AF a um sítio de *splice* 3' fraco somente na presença de um ESE e de um complexo de ligação ao ESE (complexo promotor de *splicing*), contendo a proteína SR SC35. Outros estudos mostraram que U2AF³⁵ era indispensável para ligação de U2AF e concluíram que o papel do ESE era de recrutar U2AF³⁵ e U2AF⁶⁵, sendo que U2AF³⁵ seria uma ponte entre o complexo no promotor de *splicing* e U2AF⁶⁵ (Zuo e Maniatis (1996), Graveley et al (2001)). Essas evidências sugeriram que proteínas SR ligadas a ESEs fariam contato proteína-proteína com U2AF⁶⁵ e aumentariam sua ligação ao RNA, promovendo assim o *splicing* (FIGURA 14).

FIGURA 14. Modelo de atuação da proteína SR TRA no reconhecimento do sítio de *splice* 3' do gene *dsx* de *Drosophila*.

As proteínas SR TRA e U2AF³⁵ fariam contato através de seus domínios RS.

Entretanto, nem sempre uma dependência de ESEs para a ligação de U2AF no trato de pirimidina foi observada (Li e Blencowe (1999)). Kan e Green (1999) também não observaram tal dependência no gene de IgM que também possui um ESE essencial para *splicing in vitro*. Nesse estudo, os autores propuseram que o papel do ESE não era de ligar uma proteína SR para auxiliar a ligação de U2AF no trato de polipirimidina, mas de inibir o efeito de um silenciador.

1.4.6 Detalhes do mecanismo de atuação de proteínas SR: interação proteína-proteína versus interação com RNA

O modelo proposto na **FIGURA 14** acima propõe que o complexo promotor de *splicing* ligado ao ESE faria contatos com U2AF³⁵ através de interações proteína-proteína utilizando os domínios RS. Entretanto, um outro modo de atuação das proteínas SR não pode ser excluído: interações proteína-RNA.

Uma forte evidência que colocou em dúvida o mecanismo de interação proteína-proteína foi a observação que proteínas U2AF³⁵ sem o domínio RS, que interagiria com a proteína SR ligada ao ESE, são capazes de realizar *splicing* no gene *dsx* de *Drosophila* (revisado em Matlin et al (2005), Hertel e Graveley (2005)).

Trabalhos recentes mostraram que o domínio RS deve mesmo estar envolvido diretamente na ligação ao RNA (Shen et al (2004), Shen e Green (2006), revisado em Hertel e Graveley (2005)). Através de experimentos de *cross-linking* sítio-específico, os autores dos estudos mostraram que há contatos do **domínio RS de U2AF⁶⁵** diretamente com o **sítio do ponto de ramificação no complexo E**. Mesmo com a saída de U2AF⁶⁵ para a formação do complexo A, o RNA continua a ser contatado por um domínio RS, mas desta vez de uma **outra proteína SR** ligada a um **ESE** localizado no exon 2. Estes resultados explicariam porque o domínio RS de U2AF³⁵, que supostamente contataria o domínio RS da proteína SR ligada ao ESE, não é necessário para que o *splicing* prossiga. A proteína SR ligada ao ESE, ao contatar diretamente o sítio do ponto de ramificação com seu domínio RS seria suficiente.

Foi ainda demonstrado que durante o curso da montagem do spliceossomo, uma **outra proteína SR** utilizando seu domínio RS para contatar o **sítio de splice 5'**, possivelmente para **auxiliar a ligação de U6** é requerida (em acordo com as

observações de Roscigno e Garcia-Blanco (1995) que proteínas SR podem recrutar U4/U5/U6). O processo descrito é ilustrado na **FIGURA 15**.

FIGURA 15. Interações sequenciais entre domínios RS e o pré-RNA.

No **complexo E**, o domínio RS de U2AF⁶⁵ contata o sítio do ponto de ramificação. No **complexo A**, o contato realizado pelo domínio RS de U2AF⁶⁵ é substituído no sítio do ponto de ramificação pelo domínio SR de uma proteína SR ligada a um ESE no exon 2. Finalmente, em complexos *spliceossomais* tardios, o domínio RS de uma segunda proteína SR não ligada ao ESE contata o pré-RNA no sítio de *splice* 5' em um

momento em que o sítio de *splice* 5' também é contatado por U6. Extraída de Shen e Green (2004).

Mais recentemente, Shen e Green (2006) utilizando os mesmos experimentos de *cross-linking* entre RNA e snRNPs utilizados nos trabalhos mencionados acima, mostraram que um **domínio RS** fosforilado **promove pareamento de bases** da snRNP U2 com o sítio de ponto de ramificação. Além disso, eles observaram que uma melhoria do pareamento de bases entre U1 e o sítio de *splice* 5' torna a presença do domínio RS dispensável. Juntamente com a observação que o domínio RS estudado contata seletivamente RNA em dupla fita (snRNA+RNAm), estes resultados indicam que o papel deste domínio na promoção do *splicing* seria de **melhorar o pareamento de snRNPs com o RNA** e não de estabilizar interações proteína-proteína.

Estes resultados podem ainda explicar o fato que um excesso de proteínas SR pode compensar a ausência da snRNP U1 (Crispino et al (1994), Tarn e Steitz (1994)), como discutido acima na **seção 1.4.4.1**. Segundo os autores, a presença de domínios RS em abundância favoreceria a ligação de U2 ao ponto de ramificação, mesmo na ausência das interações com U1.

O envolvimento de domínios RS tanto na interação com proteínas quanto com RNA é teoricamente possível. Sabe-se que o nível de fosforilação de proteínas SR é crucial para o seu papel no *splicing* e a fosforilação dos domínios SR é uma forma de controle importante sobre o papel das proteínas SR no *splicing*. O grau de fosforilação parece interferir na localização sub-celular e na capacidade de realizar interações com outras proteínas (revisto em Fu (1995a), Graveley (2000)). Serinas desfosforiladas poderiam interagir com RNA, enquanto quando fosforiladas teriam carga negativa que

possibilitaria interagir com as argininas de outros domínios RS (Hertel e Graveley (2005)).

1.4.7 Mecanismo de ativação do spliceossomo: do complexo B para o complexo C

1.4.7.1 A ativação do spliceossomo requer diversos rearranjos das snRNPs

O evento que marca a formação do complexo B é a ligação da tri-snRNP U4/U6.U5 ao complexo A (U1 e U2). Neste ponto, diversos rearranjos entre as ligações entre snRNPs e RNA ocorrem com a participação de ATPases específicas para cada passo (revisado em Staley e Guthrie (1998), Schwer (2001)). A **FIGURA 16** retrata seis rearranjos do complexo B que ativam o spliceossomo para o primeiro passo catalítico.

FIGURA 16. A montagem do spliceossomo catalítico requer inúmeros rearranjos.

A linha fina representa o pré-RNAm e as linhas grossas snRNPs.

(a) troca de U1 por U6

(b) troca de BBP por U2

(c) rearranjo intramolecular de U2

(d) desemparelhamento da haste II U4/U6, formação de uma haste/laço 3' em U6

(e) desemparelhamento da haste I U4/U6 e da haste/laço em 5' de U2, formação da hélice I U2/U6

(f) desemparelhamento da haste/laço U2 5', formação da hélice II U2/U6

U5 não é mostrada para facilitar a visualização.

Extraída de Staley e Guthrie (1998)

O primeiro passo da ativação do *spliceossomo* envolve o desligamento de U1 do sítio de *splice* 5' e emparelhamento de U6 (**FIGURA 16a**). Um aspecto interessante da ativação do *spliceossomo* foi revelado pelo trabalho de Staley (1999), estendendo os resultados de Konforti et al (1993). Eles mostraram que a habilidade de U6 de se ligar ao sítio de *splice* 5' é crucial para a execução do *splicing*, de tal forma que é necessário haver um balanço entre a força de ligação de U1 e U6 para que a transição do *spliceossomo* inativo para o ativo se dê com sucesso. Ou seja, se a ligação entre U1 e o sítio de *splice* 5' for muito forte, o *splicing* não pode prosseguir, já que U6 não pode se ligar ao RNA (revisado por Murray e Jarrell (1999)). Neste trabalho os autores ainda reportaram a existência de uma proteína, **PRP28**, que poderia ser responsável pela troca de U1 por U6, talvez atuando no desemparelhamento de U1 com o sítio de *splice*.

Uma outra proteína associada a desemparelhamento de snRNPs é **PRP8**, atuando no desligamento de U4 de U6 mostrado nas **FIGURAS 16d e 16e** (Kuhn et al (1999), revisado em Murray e Jarrell (1999)). Como U4 não participa da catálise, acredita-se que o papel desta snRNP seja de sequestrar o domínio catalítico de U6 até pouco antes do momento da catálise.

O último rearranjo ocorre com o pareamento de U6 e U2, formando o núcleo catalítico do *spliceossomo*. O complexo C está formado e o primeiro passo catalítico é efetuado. Alguns detalhes sobre a catálise são tratados mais abaixo.

1.4.7.2 O papel da snRNP U5 no *splicing*

Como descrito anteriormente, U5 liga-se ao sítio de *splice* 5' antes do primeiro passo catalítico (Wyatt et al (1992)), podendo até mesmo participar na determinação da posição de clivagem desse sítio (Newman e Norman (1991)). No entanto, em diferentes trabalhos os esquemas do mecanismo de *splicing* retratam a saída de U5 do

spliceossomo (**FIGURA 17**, à esquerda) ou sua permanência durante a catálise (**FIGURA 17**, à direita).

FIGURA 17. O papel da snRNP U5 no complexo C ainda não é totalmente claro.

À esquerda, no esquema apresentado por Hertel e Graveley (2005), U5 se desligaria totalmente do spliceossomo, deixando apenas U2 e U6 para realizar a catálise. Já no esquema à direita de Graveley (2000), U5 se manteria ligado ao RNA e a U2 e U6.

Estudos em levedura sugeriram que esta snRNP teria um importante papel no segundo passo catalítico do *splicing*, segurando e alinhando os exons para que sejam ligados (revisado em Newman (1997)). São observadas interações do laço I, que é conservado em diferentes organismos, com o exon 1 tanto antes quanto depois do primeiro passo catalítico. No entanto, interações com o exon 2 só são detectáveis após a

clivagem da ligação exon 5'/intron, sugerindo que **U5 alinha os dois exons** para efetuar sua junção no segundo passo. Coerentemente com um papel de U5 no alinhamento dos exons para execução do segundo passo catalítico, mutações no laço I comprometem a execução do segundo passo, mas não do primeiro (O'Keefe et al (1996)).

Assim, muitos modelos de *splicing* incluem uma ligação de U5 na parte exônica do sítio de *splice* 5' assim que U4/U6.U5 contata o RNA e U6 emparelha com U2. Em seguida, U5 se mudaria para a parte intrônica, dando lugar para o pareamento de U6 com o sítio de *splice* 5' enquanto U4 se solta de U6. O spliceossomo está pronto para realizar a primeira catálise. Assim que ela ocorre, U5 seria translocado para o sítio de *splice* 3', segurando e alinhando os dois exons para U2/U6 efetuarem a segunda catálise.

Entretanto, estudos posteriores mostraram que em mamíferos o *splicing* pode prosseguir mesmo sem a presença do laço I conservado de U5 (Ségault et al (1999)). Uma explicação para esta observação é que proteínas associadas a U5 realizariam a tarefa realizada pelo laço I de U5 de levedura. De fato, somente 6 nucleotídeos do laço I são utilizados na interação com os exons, o que pode ser muito pouco para segurá-los em posição, de forma que proteínas associadas poderiam ter um importante papel auxiliar na ligação aos exons, como por exemplo **PRP8**, uma proteína associada a U5 (Ségault et al (1999), ver mais sobre PRP8 na seção abaixo).

1.4.8 Alguns detalhes da catálise

Como descrito acima, o complexo *spliceossomal* C é composto pelas snRNPs U2, U5 e U6 e várias proteínas auxiliares, formando o spliceossomo maduro que é responsável pelos dois passos catalíticos. Há fortes indícios de que o sítio catalítico seja

formado por snRNPs, mais precisamente por U2/U6, sendo que U6 deve conter os nucleotídeos cruciais para a catálise. A observação de que U6 é a snRNP mais conservada de todas, além das evidências apresentadas acima na **seção 1.3.2** juntamente com observações de que componentes do spliceossomo desproteinizados eram capazes de realizar uma fraca atividade catalítica imediatamente sugeriram que os componentes catalíticos do spliceossomo envolvidos em *splicing* de introns do grupo nuclear seriam principalmente constituídos de RNA e não de proteínas (revisado em Green (1991), Sontheimer (2001), Butcher e Brow (2005)).

Estudos de mutagênese indicaram que uma caixa ACAGAGA invariante e a tríade AGC em U6 são cruciais para catálise (revisado em Valadkhan (2005)). Além disso, há evidências de que haja um sítio de ligação para o íon metálico divalente essencial para a catálise em uma alça da snRNP U6, que se estrutura de forma semelhante a introns do grupo II (revisado em Butcher e Brow (2005)).

1.4.8.1 A proteína PRP8 parece ter um importante papel na catálise

Além de U6, sabe-se que uma proteína bastante importante para a execução da catálise, principalmente da segunda reação, é PRP8, uma enorme proteína associada a U5. Uma extensa revisão sobre PRP8 pode ser encontrada em Grainger e Beggs (2005). Somente alguns pontos mais relevantes à catálise são levantados aqui. Esta proteína faz contato com os sítios de *splice* 5' e 3', com o sítio do ponto de ramificação e com U5 e U6. Até hoje, esta é a única proteína conhecida que contata todos esses elementos envolvidos na catálise. Além disso, PRP8 é uma das proteínas conhecidas mais conservadas, com 61% de identidade, desde levedura a humano, de forma análoga a U6, a snRNP mais conservada, fatos que suportam a idéia de que U6 e PRP8 estão relacionadas de alguma forma à catálise das reações de *splicing*.

Vários estudos permitiram construir um modelo de atuação de PRP8. Esta proteína contataria o sítio de *splice* 5' antes das reações de transesterificação. Após o primeiro passo, PRP8, auxiliada por outras proteínas seguraria o exon 1 recém clivado no spliceossomo. Em seguida, ocorrem alterações conformacionais no sítio de *splice* 3' para posicionar os exons para junção. Uma helicase, PRP16, se ligaria ao sítio de *splice* 3' hidrolisando ATP e se desligaria, permitindo uma ligação mais forte de PRP8 à região. Assim, o exon 1 seria trazido próximo ao sítio de *splice* 3' (e portanto do exon 2 a 3') para a execução do segundo passo. Devido à sua importância no *splicing*, PRP8 é tida como um cofator na reação. Note-se que embora as evidências apontem para U6 como contendo o sítio catalítico, não se exclui a possibilidade de uma participação mais direta de PRP8 na catálise.

Outro entre os vários papéis desta proteína parece ser regular a troca de U1 por U6 no sítio de *splice* 5' durante o rearranjo conformacional do complexo B para C. PRP8 atuaria também como um “fecho” que impede o desligamento de U4/U6 no complexo C até que todos os componentes estejam ligados. Ou seja, esta proteína está envolvida em diversos momentos do processo de *splicing* e é crucial para sua execução.

1.4.8.2 Um ou dois sítios catalíticos?

Um dos pontos ainda controversos em relação à catálise do *splicing* é se as duas reações de transesterificação ocorrem em um ou dois sítios catalíticos do spliceossomo. Para os introns de grupo I, há evidências que suportam a idéia de que os dois passos catalíticos ocorrem em um só sítio catalítico: as duas reações envolvem ataque nucleofílico de um oxigênio a 3' e ambas as transesterificações são reações S_N2 com a mesma orientação relativa dos reagentes como se fossem reações direta e inversa (Moore e Sharp (1993)). Em relação ao *splicing* nuclear, entretanto, Moore e Sharp

(1993) argumentaram que dificilmente os sítios catalíticos poderiam ser os mesmos já que uma reação envolve um ataque nucleofílico de um grupo 2'-OH e a outra um 3'-OH, a orientação relativa dos reagentes é diferente e a segunda reação não parece ser o reverso da primeira. O rearranjo do spliceossomo que é imprescindível para a ocorrência da segunda catálise seria um passo de conversão de um sítio ativo para o outro.

Os autores propuseram que o primeiro passo seria homólogo ao primeiro passo dos introns de grupo II (reconhecimento de uma guanina ligada ao fosfato reativo através de um oxigênio em 5', ataque nucleofílico de um oxigênio em 2' no intron ao fosfato reativo, mesma orientação relativa dos reagentes) e o segundo homólogo aos de grupo I (reconhecimento de uma guanina ligada ao fosfato reativo em 3', ataque nucleofílico de um oxigênio a 3' do exon 1 ao fosfato ligado ao exon 2, mesma orientação relativa dos reagentes). No entanto, para Gordon et al (2000) o fato dos mecanismos de reação serem tão similares entre os dois passos, requerendo magnésio para facilitar a saída do grupo abandonador, seria evidência de que as reações ocorrem em um mesmo sítio ativo.

1.5 O mecanismo canônico geral de *splicing* revisto

Com base nos detalhes acerca do processo de *splicing* geral expostos acima, uma descrição resumida mais completa dos diversos eventos que ocorrem para a montagem do spliceossomo segundo o modelo canônico (definição de intron) é fornecida abaixo. Deve-se lembrar, entretanto, que ainda não é conhecida a ordem exata dos eventos, há exceções, possíveis vias alternativas e há diferenças entre mamíferos e leveduras (os principais modelos). Assim, a intenção do modelo abaixo é apenas servir como uma

revisão geral dos diversos eventos relacionados ao processo de *splicing*, de forma a resumir os principais detalhes apresentados acima.

1. formação do complexo de engajamento (revisado em Rosbash e Séraphin (1991))/complexo E (Michaud e Reed (1991), Jamison et al (1992), Michaud e Reed (1993))

- ligação de U1C (componente da snRNP U1, Du e Rosbash (2002)) ao sítio de *splice* 5'
- ligação de proteínas SR ASF/SF2 ao sítio de *splice* 5' (Zuo e Manley (1994))
- ligação de U1 ao sítio de *splice* 5' CAG|GTRAGT através do estabelecimento de pontes de hidrogênio com cerca de 4-7 nucleotídeos nesse sítio (Seraphin e Rosbash (1989), revisado em Reed (2000b)), estabilizado pela ligação a ASF/SF2 (Eperon et al (1993), Kohtz et al (1994), Staknis e Reed (1994))) e estabilizando a interação desta proteína SR (Jamison et al (1995))
- ligação de U2AF⁶⁵ (Ruskin et al (1988))/MUD2 (Abovich et al (1994)) ao sítio do ponto de ramificação. Reconhecimento do sítio 3'-AG por U2AF³⁵, principalmente em introns com trato de poli-pirimidina fraco (Merendino et al (1999), Wu et al (1999), Zorio e Blumenthal (1999))
- ligação de SF1/BBP a U2AF⁶⁵/MUD2 no sítio do ponto de ramificação (Selenko et al (2003) e referências citadas em Reed (2000b))
- ligação de U1-70K (U1) a U2AF³⁵ através de SC35 e SF2/ASF (Wu e Maniatis (1993)), ligação de U1 a U2AF⁶⁵/MUD2 através de contatos entre BBP-PRP40 (levedura, Abovich e Rosbash (1997))/SF1-FBP11 (metazoários, revisado em Reed (2000b)) a U1-70 K

- ligação da snRNP U2 ao sítio do ponto de ramificação de forma instável (Das et al (2000))

2. formação do complexo A (pré-spliceossomo)

- substituição da ligação de SF1/BBP ao sítio do ponto de ramificação pela ligação da snRNP U2 de forma estável (revisado em Will e Lührmann (1997)), com o auxílio de proteínas auxiliares (Fleckner et al (1997), Staley e Guthrie (1998), Gozani et al (1998)), esta etapa consome ATP
- as snRNPs U1 e U2 são ligadas através de SC35 (Fu e Maniatis (1992))
- PRP5 também parece ligar U1 e U2 (Xu et al (2004))

3. formação do complexo B (spliceossomo)

3.1 complexo B1

- ligação da tri-snRNPs (U4/U6.U5) ao complexo U1/U2 (Konarska e Sharp (1987))
- U5 se liga ao sítio de *splice* 5' (Wyatt et al (1992), revisado em Newman (1997))
- U1 se desliga do sítio de *splice* 5', dando lugar a uma interação com U6 (Konforti et al (1993), Staley (1999))
- rearranjos internos de U2 (revisado em Staley e Guthrie (1998))

3.2 complexo B2 ou B (spliceossomo ativado)*

- pareamento entre U4 e U6 é desfeito (revisado em Staley e Guthrie (1998))
- pareamento entre U6 e U2 é estabelecido (revisado em Staley e Guthrie (1998))

4. formação do complexo C e catálise

- a catálise é realizada por U2/U6, mais provavelmente por U6 com o possível auxílio de proteínas associadas (revisado em Butcher e Brow (2005))

- U5 (revisado em Newman (1997)) e a proteína associada PRP8 (Ségault et al (1999), Grainger e Beggs (2005)) seguram e alinham os exons para junção no segundo passo catalítico

1.6 Mecanismos de fidelidade do *splicing* e controle de qualidade

Como descrito acima, a montagem do spliceossomo, de acordo com o modelo passo-a-passo, envolve o reconhecimento de vários sinais do pré-RNA múltiplas vezes (ex.: reconhecimento do sítio de *splice* 3' por U2AF/MUD2 e depois por U2). Acredita-se que a fidelidade do complexo processo de *splicing* seja justamente garantida pela necessidade dessas múltiplas interações (Reed (2000a)).

É interessante ressaltar que somente 4% dos sítios de *splice* 5' naturais são totalmente complementares ao sítio de ligação em U1 (Shapiro e Senapathy (1987a)), indicando que a existência de sinais fracos é compensada por múltiplos reconhecimentos ao longo da montagem do spliceossomo, que somados resultam no reconhecimento correto da estrutura exon/intron a ser processada (Reed (2000a)).

1.6.1 A hipótese do holospliceossomo e o modelo do supraspliceossomo

Há estudos recentes que propõem que o pentâmero de snRNPs não seja montado passo-a-passo ao ligar-se aos exons e introns, mas que este existe como um complexo em solução que liga-se inteiro ao pré-RNA (Stevens et al (2002)). Várias observações

pontuais anteriores suportam indiretamente esta observação. Por exemplo, Konarska e Sharp (1988), observaram complexos formados por várias snRNPs na ausência de RNAm em extratos humanos. Há evidências que mostram que U2 já se encontra no complexo E (Das et al (2000)), a ligação prematura de U1 e U2 neste complexo (Lam e Hertel (2002)) e a comunicação entre U1 e U4/U6.U5 também neste complexo (Maroney et al (2000)). Outros exemplos podem ser encontrados em (Nilsen (2002)) e em (Brow (2002)). Neste cenário, um dos possíveis papéis de proteínas SR na montagem do spliceossomo seria o de agregadoras do complexo ainda em solução, como sugerido por Lam e Hertel (2002) para explicar suas observações.

Stevens et al (2002), usando condições salinas moderadas, isolaram em levedura um penta-snRNP estável de 45 S (o spliceossomo inteiro sedimenta na faixa de 60 S em gradientes de sacarose). O penta-snRNP continha também diversas proteínas auxiliares e embora não fosse imediatamente funcional, quando um extrato contendo fatores de *splicing* foi adicionado, se tornou competente.

Este estudo pôs em questão as tradicionais observações de uma montagem do spliceossomo passo-a-passo através da detecção de diferentes complexos em géis de eletroforese. Primeiramente, o estudo de Stevens et al (2002) empregou uma concentração salina bastante mais moderada que as empregadas nos estudos tradicionais, próxima às condições da célula. Ou seja, é possível que ao empregar condições salinas extremas em estudos de eletroforese, somente as interações mais fortes se mantenham, resultando nos complexos individuais frequentemente observados. Segundo, alguns dos estudos que mostraram discordâncias do modelo passo-a-passo mencionados acima utilizaram técnicas de *cross-linking* ao invés de eletroforese. Terceiro, a observação da formação de novas interações que são a base do modelo passo-a-passo podem ser na realidade apenas decorrentes de rearranjos de interações *no*

contexto do penta-snRNP ao invés da montagem do complexo. Assim, a hipótese do holospliceossomo considera que o penta-snRNP já formado em solução sofre uma cascata de *ativação alostérica* ao contatar o pré-RNA que envolve o estabelecimento de novas interações entre as snRNPs U, os diversos fatores protéicos e o pré-RNA (Brow (2002)).

A idéia tradicional de que o spliceossomo se monta passo-a-passo sobre cada par de junções exon/intron é ainda mais abalada frente a uma linha de evidências que defende que o spliceossomo é na verdade uma sub-unidade de uma lnRNP (*large nuclear RNP*) de 200 S composta de 4 spliceossomos (Wassarman e Steitz (1993), Sperling et al (1997), Müller et al (1998), Azubel et al (2006)), compondo um supraspliceossomo. É importante notar que foi demonstrado que os spliceossomos nativos purificados a partir de lnRNPs isoladas de células HeLa eram funcionais (Azubel et al (2004), Azubel et al (2006)). Interessantemente, é proposto que a lnRNP contenha mais um domínio composto de proteínas auxiliares de processamento da extremidade 5' (proteínas de ligação ao *cap*), poliadenilação e enzimas de edição A-para-I, ou seja, a lnRNP seria a maquinaria completa de processamento de pré-RNAs (Raitskin et al (2002)).

Os autores destes estudos propõem que cada pré-RNA seria processado por uma lnRNP, independente da quantidade de introns (Spann et al (1989), Azubel et al (2006)), sendo que até 4 introns podem ser processados ao mesmo tempo, mas não obrigatoriamente na ordem que se encontram no pré-RNA. Além disso, o próprio pré-RNA seria responsável pela manutenção da estrutura quaternária da lnRNP (Azubel et al (2006)). No modelo do supraspliceossomo (Azubel et al (2006)), é proposto que o pré-RNA seria passado através do complexo de 4 spliceossomos, processando até 4 introns por vez, mas não necessariamente de forma simultânea. O modelo também é

compatível com a inclusão alternativa de exons, bastando formar um laço intron+exon+intron, excluindo o exon do processamento. Devido à proximidade dos 4 spliceossomos no supraspliceossomo, pode haver comunicação entre os diferentes introns que estão sendo processados, o que poderia ajudar a explicar porque observa-se em alguns casos que o *splicing* de um dado intron é dependente do *splicing* de introns adjacentes (Azubel et al (2006)).

Por fim, os autores do modelo propõem que o intron é reconhecido como uma entidade praticamente de uma vez só, ao invés de depender do reconhecimento de várias partes conspícuas, o que aumentaria a fidelidade do processo. Embora tanto no modelo de montagem do spliceossomo passo-a-passo como no modelo do holospliceossomo haja o reconhecimento de múltiplos sinais fracos, a interpretação de como a fidelidade é garantida difere nas duas visões. Enquanto no primeiro caso o fato do reconhecimento ser feito múltiplas vezes ao longo do tempo garantiria múltiplas verificações da validade dos sinais, garantindo a fidelidade do processo, no modelo do holospliceossomo, o reconhecimento destes sinais múltiplos somados de uma só vez é que garantiria a fidelidade. Embora contraditórias, as duas explicações parecem razoáveis e uma resolução da aparente discrepância só deverá ocorrer nos próximos anos.

1.6.2 Controle de qualidade da produção de RNAm: non-sense mediated decay

O processamento do RNAm está associado a diversos outros processos celulares, entre eles a própria transcrição (revisado em Maniatis e Reed (2002)). Durante este processo, após a excisão de um intron e a junção de dois exons, o domínio C-terminal da RNA polimerase II associa-se a diversos fatores celulares, entre eles um complexo

protéico que é depositado a ~20 nt a 5' das bordas exon/exon recém ligadas (**FIGURA 18**). Esses fatores formam um complexo denominado **EJC** (*exon junction complex*) que se mantém ligado ao RNAm maduro e promove sua exportação para o citoplasma até o momento da tradução. No **primeiro ciclo de tradução**, o ribossomo ao encontrar um EJC remove-o do RNAm e os múltiplos ciclos de tradução seguintes prosseguem normalmente. Entretanto, verificou-se que raramente a região 3' não-traduzida do RNAm (3' UTR de UnTranslated Region) de mamíferos é formada por mais de um exon, de forma que o códon de terminação sempre se localiza após o último EJC (revisado recentemente em Cartegni et al (2002), Lewis et al (2003), Maquat (2005)). Quando existe um códon de terminação antes do último EJC (**códon de terminação prematuro**), o ribossomo abandona o RNAm ao encontrá-lo e deixa todos os EJC a 3' ainda ligados ao RNAm. Nesta situação, um mecanismo chamado de *non-sense mediated decay* (**NMD**) é acionado. Este mecanismo envolve fatores protéicos que se ligam ao complexo de pós-terminação depositado pelo ribossomo na região do códon de parada prematuro e contatam o EJC a 3' que foi abandonado. Quando ocorre interação entre o complexo de pós-terminação no códon de parada prematuro e o EJC abandonado, um mecanismo de **degradação do RNAm** é ativado. Ou seja, o NMD é um mecanismo de reconhecimento de mensagens incorretas que carregam um códon de terminação prematuro e acionam a degradação destas.

Embora indireta, esta também é considerada uma forma de garantir que somente os RNAs corretos serão traduzidos. Ou seja, eventos de *splicing* errôneos que resultem na alteração do quadro de leitura aberta, causando a ocorrência de um códon de parada prematuro serão eliminados posteriormente (Cartegni et al (2002)).

FIGURA 18. Mecanismo de *non-sense mediated decay*.

A inclusão de um códon de parada prematuro resulta no abandono de EJCs que ao serem reconhecidos pelo complexo de pós-terminação, desencadeiam a degradação do RNAm (ver texto).

1.7 *Splicing* alternativo

I noticed that my opponent is always on the go

And, won't go slow so as not to focus and I notice

He'll hitch a ride with any guide as long as they go fast from whence he came

But he's no good at being uncomfortable so he can't stop staying exactly the same

“Extraordinary machine”, Fiona Apple

Apesar da espantosa fidelidade do processo de *splicing*, vários tipos alternativos de *splicing* de um pré-RNAm transcrito de um dado gene são observados, desde variações ínfimas entre as isoformas produzidas ao se utilizar sítios de *splice* alternativos localizados a alguns nucleotídeos de distância até a perda de vários exons. Por muito tempo acreditou-se que o fenômeno de *splicing* alternativo fosse bastante raro. No entanto, com o advento da genômica e da geração de milhões de sequências expressas completas e parciais, foi observada uma extraordinária variabilidade do transcriptoma de vários organismos (cerca de 80% dos genes humanos apresentam *splicing* alternativo segundo Johnson et al (2003), ver revisão de Lee e Wang (2005)). De acordo com Stamm et al (2005), de 1985 para 1998, o número de artigos descrevendo *splicing* alternativo aumentou de 16 para 1073 e a partir de então, cerca de 1000 artigos por ano tratam das várias facetas do *splicing* alternativo.

Um bordão da área genômica é que, frente ao baixo número de genes encontrados no genoma humano em relação a organismos como *Drosophila*, uma explicação para a origem do suposto maior número de proteínas expressas para gerar

uma maior “complexidade” em mamíferos seria a produção de diversas isoformas de proteínas de um mesmo gene através de *splicing* alternativo. Outro aspecto interessante do *splicing* alternativo é que ele coloca em cheque um dos pontos do Dogma Central da Biologia “um gene, uma enzima”. Frente a este tipo de variabilidade, a frase poderia ser modificada para “um gene, várias enzimas - talvez” (na realidade, frente à crescente constatação da importância de RNAs não-codificadores (revisado em Eddy (2001)), a frase deveria ser reformulada de forma ainda mais radical para “um gene, vários transcritos - talvez”).

Independente de explicar ou não uma suposta maior diversidade protéica em mamíferos, sabe-se que o *splicing* alternativo está envolvido em vários processos fisiológicos complexos como por exemplo, determinação de sexo em *Drosophila* (revisado em Black (2003)), apoptose (Boise et al (1993)), reconhecimento de som (revisado em Black (1998)) ou ainda em processos de adaptação fisiológica. Por exemplo, alterações nutricionais como a mudança na quantidade de ácidos graxos poli-insaturados disponíveis alteram o padrão de produção de variantes de *splicing* de glicose-6-fosfato desidrogenase (para mais exemplos, ver Stamm (2002)).

Basicamente, existem três tipos de *splicing* alternativo:

i- **inclusão alternativa de exon** (também denominado “*exon skipping*”, “*cassette exon*”), no qual um ou mais exons podem ser excluídos juntamente com os introns flanqueantes, resultando em transcritos alternativos mais curtos (**FIGURA 19**).

ii- **sítio de *splice* alternativo** (5’ ou 3’), engloba os eventos de *splicing* alternativo que envolvem junções exon/intron diferentes daquelas “normalmente” utilizadas, resultando em um encurtamento ou alongamento do exon (**FIGURA 19**).

iii- **retenção de intron** (uso alternativo de intron), envolve os casos em que um intron é mantido no transcrito final. Para estes casos, observa-se duas populações de RNAm: uma que inclui um trecho de DNA genômico flanqueado por “exons” (exon + intron + exon) nos RNAs maduros e outra população na qual o mesmo trecho de DNA genômico é excisado como um intron qualquer, sendo também flanqueado por exons (exon + exon, **FIGURA 19**).

FIGURA 19. Tipos de *splicing* alternativo.

(a) inclusão alternativa de exon, o exon 2 pode ser mantido ou excluído do RNAm maduro (b) sítio de *splice* alternativo. O exon 2 pode ser delimitado por dois sítios de *splice* 5' diferentes e o exon 4 por sítios de *splice* 3' alternativos, (c) retenção do intron entre os exons 2 e 3.

Note-se que há diferentes esquemas de classificação de eventos de *splicing* alternativo. Por exemplo, a base de dados ASD (Thanaraj et al (2004)) classifica (ii) como “isoforma do exon” e “isoforma do intron”, quando ocorre encurtamento do exon e alongamento do exon (encurtamento do intron), respectivamente. Há também quem

classifique o uso mutuamente exclusivo de exons (**FIGURA 20**) como um outro tipo de *splicing* alternativo (ex.: Black (2003), Matlin et al (2005)), mas na minha opinião, este é, formalmente, apenas um caso específico da inclusão alternativa de exons, só que envolvendo dois transcritos que nunca contêm dois dados exons ao mesmo tempo. Igualmente, há quem classifique o uso alternativo do primeiro e último exons das regiões não-traduzidas 5' ou 3', respectivamente, como outros dois tipos de *splicing*, mas formalmente também são casos de inclusão alternativa de exons, embora não seja possível usar o mesmo procedimento de bioinformática para identificá-los (não possuem um dos exons flanqueantes, tornando sua detecção correta impossível).

FIGURA 20. Uso mutuamente exclusivo de exons.

A nomenclatura usada em (ii) é bastante genérica, contemplando tanto o caso da forma alternativa alongar ou encurtar um exon, sem tender a descrever somente uma das formas, como é o caso de “*exon skip*”. No caso de “retenção de intron”, talvez o mais correto fosse denotar este tipo de evento como “uso alternativo de intron”, mas por enquanto a primeira denominação é bem aceita. Por enquanto, cada autor define sua terminologia. É possível que futuramente haja nomenclaturas “oficiais” que sejam mais neutras, já que mais recentemente é que começou a ser aceito que é importante considerar a frequência em que diferentes isoformas são expressas (Clark e Thanaraj (2002), Modrek e Lee (2003), Resch et al (2004a), Baek e Green (2005), Xing e Lee (2005), para citar alguns trabalhos). É possível ainda que a classificação seja desdobrada

para denominar separadamente os tipos de isoforma com base na frequência em que ocorrem (inclusão alternativa de exon/exclusão alternativa de exon, por exemplo).

Em adição à nomenclatura empregada para definir os diferentes tipos de *splicing* alternativo, um outro termo frequentemente utilizado é “**exon constitutivo**” que define exons que são sempre incluídos da mesma forma no RNAm maduro. Entretanto, a definição do termo é flexível, variando um pouco com o contexto. Pode-se considerar um exon como “constitutivo” para inclusão alternativa de exons, aquele que é sempre incluído, ainda que apresente uma forma mais longa e outra mais curta (uso de sítios de *splice* alternativos), por exemplo.

Como mencionado acima, diversos trabalhos de *splicing* alternativo foram realizados nos últimos anos e uma vasta literatura sobre os mais variados aspectos do fenômenos foi produzida, abrangendo desde a frequência do fenômeno em diversos organismos, seu possível impacto funcional, até formas de regulação e evolução, descritos nas seções a seguir. Muitos dos trabalhos foram baseados em análises em larga-escala de dados digitais e dessa forma, na seção abaixo alguns conceitos e termos utilizados nos estudos são brevemente definidos.

1.7.1 Alinhamento de cDNAs no genoma, identificação de *splicing* alternativo no transcriptoma

Assim como o **genoma** é definido como o conjunto total de genes de um organismo, o **transcriptoma** é definido como o conjunto total de transcritos gerados a partir de um genoma. O avanço dos métodos de sequenciamento permitiu a produção de milhares a milhões de **sequências expressas (cDNAs)**. Por causa da tecnologia utilizada, essas sequências podem ser parciais (*Expressed Sequence Tags*, **ESTs**) ou

completas (denominadas **RNAs**). Em estudos de *splicing* alternativo, as sequências dos transcritos armazenadas em formato digital têm sido rotineiramente utilizadas como fonte de evidências do fenômeno (revisto por Lee e Wang (2005)), como relatado a seguir.

Com o sequenciamento do genoma humano (Lander et al (2001)) e de camundongo (Consortium et al (2002)) entre outros, tornou-se possível **alinhar as sequências expressas ao genoma** e identificar de que gene cada transcrito provém. Mais importantemente, o alinhamento permite a identificação direta de exons (trechos de cDNA alinhados ao genoma) e introns (apresentam-se como vãos (*gaps*) no alinhamento) (**FIGURA 21**). Programas específicos permitem a identificação precisa dos sítios de *splice* (por exemplo SIM4, Florea et al (1998); est2genome, Rice et al (2000)). Com base na sobreposição desses alinhamentos é realizado então um **agrupamento de cDNAs** indireto gene a gene, compondo **grupos ou clusters de cDNA**. Esse agrupamento indireto se dá através da comparação das coordenadas dos alinhamentos das sequências expressas, utilizando-se critérios de sobreposição mínima para considerar duas sequências como parte de um mesmo grupo. Ou seja, um grupo ou *cluster* de cDNA é a coleção conhecida de transcritos parciais e integrais de um dado gene. Por essa razão, a palavra *cluster* é coloquialmente usada de forma intercambiável com gene, embora formalmente as palavras sejam conceitualmente diferentes.

Dentro de cada grupo de transcritos de um gene é então possível **identificar eventos de *splicing* alternativo** buscando-se sequências que apresentem certas diferenças (**FIGURA 21**). Por exemplo, um evento de inclusão alternativa de exons pode ser detectado ao se encontrar um par de cDNAs alinhados em que haja ausência de um exon em uma sequência e a presença do mesmo em outra (**FIGURA 21**). De forma similar, um evento de retenção de introns é evidenciado pela presença de um segmento

em um cDNA entre dois exons que em outra sequência esteja ausente (**FIGURA 21**). Por isso, os cDNAs ou transcritos conhecidos de um gene são também denominados **evidências de *splicing* alternativo**.

FIGURA 21. Utilização de dados de sequências expressas para determinação da estrutura exon/intron de genes e identificação de eventos de *splicing* alternativo através de Bioinformática.

No esquema são representados dois *clusters* de cDNA diferentes alinhados ao genoma, compostos por sequências de RNAm e de ESTs. Os eventos de *splicing* alternativo são identificados através da comparação das coordenadas dos alinhamentos dos exons.

Como a detecção de eventos de *splicing* alternativo é baseada em evidências de ESTs e RNAs, a existência ou não destas sequências em *clusters* de cDNA é um dos fatores limitantes em análises do transcriptoma. A quantidade de cDNAs existentes para um dado gene (*cluster*) é denominada **tamanho do cluster**, e de forma geral, a disponibilidade de cDNAs para um dado gene ou genoma é denominada **cobertura**. A

definição não é precisa e por vezes é confundida com tamanho do *cluster*, ou até referente a uma região do *cluster*. De qualquer maneira, tanto tamanho do *cluster* quanto cobertura referem-se à disponibilidade de dados de sequências expressas para um dado genoma. Como o tamanho do *cluster* representa, em princípio, o número de transcritos produzidos por um gene, há uma certa relação deste valor com o **nível de expressão gênica**, pelo menos de forma grosseira e qualitativa.

Todas as análises mencionadas até aqui podem ser realizadas em qualquer organismo que tenha dados digitais disponíveis. Com base na idéia que características conservadas entre organismos estão sob pressão seletiva e são portanto indicativos de função biológica, a comparação de características biológicas entre organismos é extremamente útil e recebe o nome de “**Genômica Comparativa**”.

1.7.2 Características do fenômeno de *splicing* alternativo

Como mencionado acima, nos últimos anos o número de estudos sobre *splicing* alternativo cresceu enormemente. Estudos sobre diversos aspectos do fenômeno foram conduzidos, principalmente aqueles em larga-escala que buscam encontrar padrões e mecanismos gerais. Nas seções seguintes, tentei abordar os aspectos mais importantes que têm sido estudados nos últimos anos e ilustrar as mais diferentes áreas estudadas e abordagens utilizadas citando alguns artigos como exemplo, principalmente aqueles que realizaram análises em larga-escala. Note-se que devido ao enorme volume de artigos existentes, somente alguns foram citados, não significando necessariamente que são os únicos sobre determinado assunto. Nem todas as sub-seções abaixo constituem uma revisão aprofundada de cada área já que em muitos casos cada tópico é por si uma área de pesquisa bastante extensa. Outras fontes de informação sobre o fenômeno são as

várias revisões de caráter geral existentes, por exemplo Black (2003), Zheng (2004), Matlin et al (2005), Blencowe (2006).

1.7.2.1 Uma alta frequência de *splicing* alternativo em diversos organismos

- utilizando montagens de sequências expressas curtas (ESTs), um estudo pioneiro mostrou que 35% dos genes humanos apresentam *splicing* alternativo (Mironov et al (1999)). Um outro trabalho produzido nesta linha analisou eventos de *splicing* alternativo em genes anotados do cromossomo 22 empregando anotação manual para obter um conjunto de dados de alta confiabilidade. Também foi mostrada uma alta taxa de *splicing* alternativo (30-40%, Hide et al (2001)).
- vários outros estudos similares se seguiram e confirmaram a alta taxa de *splicing* alternativo em genes humanos (40-60% dos genes, revisado em Modrek e Lee (2002)). Estes valores eram sabidamente subestimados devido à escassez de dados para muitos dos genes, já que os estudos dependiam da existência de ESTs para identificar eventos de *splicing* alternativo (cobertura do *cluster*). Assim, com o aumento da quantidade de sequências expressas disponíveis, o número de genes que sofrem *splicing* alternativo também aumentou.
- O estudo de Kan et al (2002) utilizou grafos para estudar a frequência de *splicing* alternativo, utilizando as frequências das isoformas para obter variantes com maior confiabilidade (ver revisão do uso de grafos em estudos deste tipo em Lee e Wang (2005)). Requerendo um alto nível de confiabilidade, os autores estimaram que 17-28% dos genes apresentavam *splicing* alternativo e 5% dos genes apresentaram retenção de introns.

- mais recentemente, uma estimativa da frequência de *splicing* alternativo foi realizada sem empregar ESTs, utilizando microarrays de junções exon/exon (74% dos genes humanos apresentaram *splicing* alternativo (Johnson et al (2003))).
- estudos similares aos realizados em humano foram realizados em camundongo (Zavolan et al (2003)). Também foi mostrada uma alta frequência do fenômeno neste organismo (estimado em até 60% dos genes).
- análises em *Arabidopsis thaliana* estimaram que, com a cobertura de ESTs disponível, ~20% dos genes deste organismo apresentam *splicing* alternativo de forma que o fenômeno deve ser frequente também em plantas (Ner-Gaon et al (2004), Ner-Gaon e Fluhr (2006)).
- é interessante notar que Brett et al (2002), usando ESTs de diferentes organismos, mostraram que não há diferenças na quantidade de isoformas produzidas por invertebrados e mamíferos, sugerindo que o *splicing* alternativo não é responsável pela complexidade de mamíferos.

Alguns dos trabalhos citados acima revelaram que dentre os três tipos básicos de *splicing* alternativo existentes, o **mais frequente** em humano e em outros vertebrados é a **inclusão alternativa de exons** (revisado em Modrek et al (2001)). A predominância deste tipo de *splicing* alternativo é uma predição do modelo de reconhecimento de exons, já que se um sítio de *splice* fraco não for reconhecido, o exon que ele flanqueia será considerado parte de um intron e será excisado (Berget (1995)). Já para organismos em que o reconhecimento de introns deve ser responsável pela maior parte dos casos de *splicing*, a predição seria de uma alta frequência de **retenção de introns**. Este é o caso, por exemplo, de *Arabidopsis thaliana* que possui introns curtos (40% dos eventos, Ner-Gaon et al (2004)). Este tipo de *splicing* alternativo é o mais difícil de ser distinguido de

transcritos não-completamente processados, ou seja, de artefatos. No caso de *Arabidopsis thaliana*, mostrou-se que transcritos que retêm introns estão associados a ribossomos e devem, portanto, ser produtos funcionais (Ner-Gaon et al (2004)).

A alta frequência de *splicing* alternativo verificada não implica que todos os eventos observados são funcionais. Uma idéia bastante aceita é que eventos conservados em diferentes organismos devem estar sob pressão seletiva e portanto serem funcionais. Embora Nurtdinov et al (2003), Sorek et al (2004a) tenham observado uma baixa conservação de eventos de *splicing* alternativo entre humano e camundongo (~25% dos eventos), estes estudos não consideraram a baixa cobertura de ESTs que afeta diretamente a detecção de eventos de *splicing* alternativo. Quando a cobertura é considerada, através de uma extrapolação matemática estimou-se que pelo menos 61% de eventos são conservados entre os dois organismos (Thanaraj et al (2003), ver também **seção 1.7.4.4**).

1.7.2.2 Função do *splicing* alternativo

Pelo menos duas funções podem ser atribuídas ao *splicing* alternativo:

- *alteração da estrutura das proteínas codificadas* (inserção ou deleção de partes da proteína).
- *regulação dos níveis de transcritos produzidos* (através da inserção de códons de parada prematuros que seriam reconhecidos pelo mecanismo de NMD mencionado acima na **seção 1.6.2** e na **seção 1.7.2.3** abaixo, ou através da alteração da estabilidade do RNAm).

A inserção ou deleção de partes da proteína por eventos de *splicing* alternativo pode acarretar em mudanças na sua estrutura, alterando propriedades de ligação, atividade enzimática, localização celular, inserção na membrana, estabilidade, carga e

modificações pós-traducionais. Além disso, a simples introdução de códons de parada na região codificadora pode resultar em proteínas mais curtas, com funções diferentes das isoformas completas. Uma revisão de casos individuais já conhecidos de alterações funcionais causadas por *splicing* alternativo pode ser encontrada em Stamm et al (2005).

Vários dos estudos em larga-escala mencionados no tópico acima sobre a frequência de *splicing* alternativo analisaram o possível impacto do fenômeno no potencial codificador dos transcritos, ou seja, a localização dos eventos em relação à região codificadora. Observou-se que uma grande parte dos eventos (principalmente inclusão alternativa de exon) afetaria a proteína codificada (92% dos eventos (Hide et al (2001)), 74% dos eventos, (Modrek et al (2001))), 74% dos eventos em camundongo (Zavolan et al (2003))).

Dessa forma, é razoável imaginar que *splicing* alternativo poderia ser um mecanismo importante de geração de variabilidade protéica. Especula-se que tantas modificações poderiam alterar o repertório de proteínas expressas a tal ponto de expandir o proteoma de um dado organismo. Essa hipótese é discutida em diversos trabalhos, como por exemplo na revisão de Maniatis e Tasic (2002).

No entanto, o impacto do fenômeno de *splicing* alternativo poderia afetar desigualmente as proteínas, ocorrendo preferencialmente em grupos funcionais específicos onde a variabilidade é crucial. As análises realizadas sobre este tema se beneficiaram do advento do banco de anotações funcionais Gene Ontology (GO, Ashburner et al (2000)) que permite classificar funcionalmente milhares de genes automaticamente com alta confiabilidade. Por exemplo, um estudo pioneiro foi o de Modrek et al (2001) que observou uma alta taxa de *splicing* alternativo em receptores com funções no sistema imune e no sistema nervoso. Em camundongo, Zavolan et al (2003) não verificou o mesmo viés que Modrek et al (2001), mas encontrou que

enzimas em geral parecem estar menos sujeitas a *splicing* alternativo. Johnson et al (2003) que analisaram *splicing* alternativo utilizando microarrays encontraram que as funções mais sujeitas a *splicing* alternativo seriam comunicação celular e regulação enzimática. Um estudo realizado por nós não detectou diferenças nos padrões de *splicing* alternativo em genes associados à matriz extracelular (Sakabe et al (2004)). Por hipótese, o grupo de proteínas codificadas por estes genes (envolvidas em comunicação celular, migração e adesão celular, funções em que a diversidade da interação proteína-proteína deve ser útil) poderia apresentar um padrão de inclusão alternativa de exon diferenciado (maior número de isoformas, mais eventos envolvendo domínios de interação proteína-proteína, eventos mais longos), o que não foi observado.

Outros estudos refinaram a hipótese acima, buscando verificar se *splicing* alternativo afeta domínios específicos, como por exemplo relacionados a funções específicas onde a interação combinatória seria útil (interação proteína-proteína). Embora isso tenha sido demonstrado para casos específicos (por exemplo Nurminsky et al (1998), Sudo e Hidaka (1998), Lim et al (1999)), ainda não é claro se este é um mecanismo geral e os dados ainda são controversos. Por exemplo:

- Kriventseva et al (2003) mostraram haver uma maior ocorrência de *splicing* alternativo envolvendo domínios inteiros ou afetando resíduos do sítio catalítico do que esperado por acaso, sugerindo que *splicing* alternativo é efetivamente usado para regular a função de proteínas.
- Resultado similar foi encontrado para estruturas terciárias do banco de estruturas SCOP (Murzin et al (1995)) no estudo de Homma et al (2004).
- Um outro estudo utilizando somente RNAs curados (RefSeq, Pruitt et al (2005)) mostrou haver uma preferência de remoção de domínios envolvidos em

comunicação celular, sinalização e apoptose acima do esperado (Liu e Altman (2003)).

- Uma maior ocorrência de *splicing* alternativo em proteínas associadas a certas funções biológicas também foi mostrada por Yura et al (2006) (transdução de sinal e transcrição/tradução, afetando domínios de ligação proteína-proteína, proteína-DNA/RNA ou proteína-ligante/cofator). Neste trabalho foi também utilizado um conjunto de RNAs supostamente de maior confiabilidade, provenientes do projeto H-invitational japonês (Imanishi et al (2004)). Note-se entretanto a discrepância das funções encontradas neste estudo e naquele de Liu e Altman (2003) mencionado acima.
- Domínios de interação proteína-DNA também apareceram como preferencialmente afetados por *splicing* alternativo de forma tecido-específica, indicando que as isoformas regulam a transcrição diferencialmente, em um estudo em fatores de transcrição de camundongo (Taneri et al (2004)).
- Entretanto, em outro estudo, embora tenha sido observado que a maioria dos domínios afetados por *splicing* alternativo são associados a interações proteína-proteína, a fração correspondia simplesmente à maior ocorrência destes domínios em proteínas humanas (Resch et al (2004b)).
- Outro estudo que contradiz a hipótese de uma regulação da função de proteínas através da remoção de domínios inteiros analisou um conjunto pequeno, mas bastante curado, de proteínas envolvidas em interação e não obteve indícios de haver preferência na remoção de domínios de ligação proteína-proteína (Offman et al (2004)). Neste estudo, os autores levantam a hipótese de que é possível que os eventos de *splicing* alternativo mesmo ocorrendo em outras partes da proteína não diretamente associadas à interação proteína-proteína podem modificar a estrutura a

tal ponto de afetar a ligação a outras proteínas. Deve-se levar em conta que embora de alta confiabilidade, o conjunto analisado era bastante reduzido, o que pode ter afetado as conclusões.

- Embora tenha sido encontrado que exons alternativos codifiquem domínios inteiros, a fração encontrada foi <4%, indicando que a remoção de domínios não deve ser um papel tão importante do fenômeno de *splicing* alternativo. Além disso, a fração de exons alternativos e constitutivos codificando domínios foi a mesma (Magen e Ast (2005)).
- Outras discrepâncias entre diferentes estudos também merece atenção. Enquanto alguns mostram que há uma preferência da ocorrência de *splicing* alternativo na superfície da proteína, o que afetaria menos sua estabilidade (75% dos casos, segundo um estudo que modelou variantes de *splicing* por *threading*, Wang et al (2005b)), outros mostraram o contrário (35% dos casos de *splicing* alternativo não afetariam a estabilidade da proteína codificada (Yura et al (2006))). Tanto para este tópico como para o discutido acima, algumas explicações podem ser cogitadas para as discrepâncias observadas. Os diferentes estudos empregaram diferentes conjuntos de dados com critérios de qualidade distintos o que pode ter influenciado as conclusões. Além disso, os conjuntos de dados resultantes possuem tamanhos bastante diferentes, podendo não representar a mesma população de proteínas. De qualquer maneira, estas discrepâncias mostram que uma compreensão do impacto funcional do *splicing* alternativo de maneira geral ainda não foi atingida.

1.7.2.3 Relação de *splicing* alternativo com outros processos celulares

Inúmeras evidências suportam a idéia de que todos os processos envolvidos na expressão gênica em eucariotos, desde a iniciação da transcrição, passando pelo

encapamento da extremidade 5', transcrição, *splicing*, poliadenilação da extremidade 3' do RNAm, NMD, até sua exportação para o citoplasma não ocorrem separadamente, mas fazem parte de uma rede que acopla todos os processos. Mecanicamente, a principal razão para esse acoplamento é o fato que o domínio C-terminal (CTD) da polimerase de RNA II se associa ordenadamente aos diversos fatores que controlam esses processos (para revisão ver Maniatis e Reed (2002), Reed e Hurt (2002)).

Em relação ao *splicing*, há evidências que indicam que o fenômeno ocorre junto com a transcrição, embora não compulsoriamente. Por exemplo, embora o intron 3 do gene “Balbiani ring 1” (*br1*) seja excisado cotranscricionalmente, somente cerca de 10% dos transcritos têm o intron 4 excisado dessa forma. No restante, a excisão se dá pós-transcricionalmente (Baurén e Wieslander (1994)).

Splicing alternativo está fortemente acoplado à **iniciação da transcrição** (escolha alternativa de promotores) e à **taxa de transcrição** (revisado em Kornblihtt et al (2004)). Ou seja, observa-se que diferentes isoformas são geradas dependendo do promotor utilizado para transcrever o gene ou da taxa de transcrição.

Basicamente, duas idéias foram oferecidas para explicar o acoplamento da escolha de um promotor em particular com modificação do padrão de *splicing*. A primeira é que como a escolha do promotor pode modular a associação de **diferentes conjuntos de proteínas SR** (Cramer et al (1999)) e estas estão envolvidas no reconhecimento de exons e introns, diferentes isoformas poderiam ser geradas (ver **seção 1.7.3.1** abaixo). A segunda envolve o acoplamento da taxa de transcrição ao *splicing* alternativo. Observa-se que pausas na transcrição ou diminuição da taxa de elongação alteram o padrão de *splicing* de uma molécula, provavelmente por aumentar o **tempo para reconhecimento** de exons, introns e elementos reguladores em *cis* (Roberts et al (1998)). Neste trabalho, os autores mostraram que um retardo da

transcrição leva a um aumento da inclusão do exon 3 no gene de alfa-tropomiosina. O mecanismo parece ser que devido a uma demora na transcrição do intron localizado após o exon 3 onde há o inibidor de *splicing* intrônico DRE (*downstream regulatory element*) a inibição da inclusão do exon 3 não ocorre. Observou-se também que diferentes taxas de transcrição podem possibilitar a formação de diferentes **estruturas secundárias do pré-RNA** e estas, ao expor diferentes partes da molécula, podem afetar a seleção de sítios de *splice* (Eperon et al (1988)).

Uma outra etapa da transcrição que pode afetar o padrão de *splicing* alternativo é a **alteração da estrutura da cromatina** no início da transcrição, provavelmente por afetar o tempo disponível para reconhecimento de sítios de *splice*. Ao se utilizar um inibidor da acetilação de histonas, observou-se a não-inclusão do exon EDI de fibronectina (revisado em Kornblihtt et al (2004)).

Como comentado acima, outro processo celular associado ao *splicing* é o **NMD**, que tem por finalidade servir como controle de qualidade dos RNAs transcritos. Como o *splicing* alternativo pode facilmente gerar moléculas com **códons de parada prematuros**, acredita-se que uma das funções deste fenômeno, associado à **degradação de transcritos por NMD**, seria justamente de regular os **níveis de expressão gênica** ao desviar a produção de mensagens que seriam traduzidas (o mecanismo foi denominado **RUST**, *regulated unproductive splicing and translation*, “*splicing* e tradução improdutivos regulados”, Lewis et al (2003)). Neste trabalho, foi mostrado que cerca de um terço dos transcritos produzidos por *splicing* alternativo podem estar sujeitos a degradação por NMD. No entanto, há evidências de que embora sujeitos a degradação, alguns RNA alternativos escapam da via e geram proteínas funcionais (Dreumont et al (2005)).

1.7.3 Mecanismos de regulação do *splicing* alternativo

1.7.3.1 O papel dos elementos *cis* e *trans* na regulação do *splicing* alternativo

As sequências ESEs que auxiliam no reconhecimento dos sítios de *splice* ou na formação de pontes protéicas para definição de exons não são as únicas sequências em *cis* que atuam no processo de *splicing*. Em realidade, há quatro classes desses elementos, agrupados de acordo com a sua localização (no intron ou no exon) e função (promovendo ou inibindo o *splicing*), embora tal separação seja mais formal que funcional. Os ESEs já mencionados acima promovem o *splicing*, juntamente com os ISEs (*intronic splicing enhancers*). Já os ESSs (*exonic splicing silencers*) e ISSs (*intronic splicing silencers*) seriam supressores do *splicing* (revisado em Cartegni et al (2002), Matlin et al (2005)). Um exemplo de como a classificação é mais formal que funcional vem de um conjunto de sequências inicialmente identificadas como ESSs que atuam como ISEs quando localizados em introns (Wang et al (2004a)).

As quatro classes de elementos em *cis* citadas acima possuem um importante papel na regulação do *splicing* e na geração de formas de RNAm alternativas, já que sua presença ou ausência pode levar à remoção de exons inteiros, ao reconhecimento de sítios de *splice* alternativos ou até à retenção de introns no RNAm maduro. Pseudo-exons parecem não ser reconhecidos como exons devido a diferenças na composição de elementos reguladores em *cis*, tanto de ESEs quanto de ESSs (Sironi et al (2004)). O papel desses elementos em *cis* é tão importante que mutações envolvendo nucleotídeos cruciais ao reconhecimento das bordas exon/intron podem resultar em doenças (revisado em Cartegni et al (2002)).

Os fatores *trans* que se ligam a estes elementos em *cis* incluem **proteínas SR** (tipicamente promotoras do *splicing*, revisto em Fu (1995b)) e **hnRNPs** (*heterogeneous nuclear ribonucleoproteins*, tipicamente inibitórias) que atuam de forma combinatória para reconhecer exons e introns (Graveley (2000)). Com base nessa idéia, acredita-se que transcritos que sofrem *splicing* alternativo estariam associados a conjuntos específicos de proteínas reguladoras, hipótese recentemente confirmada por Mabon e Misteli (2005). Uma vez que as concentrações e mais importante, as **razões entre proteínas SR e hnRNPs** variam entre tecidos ou tipos celulares (Hanamura et al (1998) e exemplos na revisão de Stamm (2002)), tal ação combinatória chega a gerar padrões de *splicing* específicos a tecidos (ver **seção 1.7.4.2**). A disponibilidade de proteínas SR pode também ser influenciada em resposta a estímulos, como por exemplo na mudança do padrão de *splicing* do gene de glicose-6-fosfato desidrogenase frente a mudanças na quantidade de ácidos graxos poli-insaturados (revisto Stamm (2002)).

Muitos desses silenciadores e promotores de *splicing* constitutivo ou alternativo atuam auxiliando ou impedindo a formação dos complexos E e A. A **FIGURA 22**, extraída de Matlin et al (2005) mostra algumas das formas de atuação destes fatores *trans* interagindo com elementos em *cis*. ISEs e ESEs podem auxiliar na ancoragem de elementos dos complexos pré-spliceossomo, estabilizando seus contatos com o RNA, por exemplo. O modelo mais citado desta forma de regulação é a ligação de U2AF³⁵ ao sítio de *splice* 3' do gene *dsx* de *Drosophila* auxiliada pela proteína SR TRA (**FIGURA 14**). A ligação de TRA a U2AF³⁵ promoveria a ligação de U2AF⁶⁵ ao sítio de polipirimidina do intron a 5' do exon 4 (**FIGURA 22a**). No entanto, outras formas de atuação desta proteína SR que não interagindo diretamente com U2AF³⁵ não são excluídas (veja **seção 1.4.6**).

FIGURA 22. Mecanismos de promoção e silenciamento de *splicing* alternativo.

- (a) Modelos de função de ESEs. Proteínas SR ligadas a um ESE poderiam ativar o *splicing* ao recrutar U2AF⁶⁵ a sítios de polipirimidina fracos através da interação de seus domínios RS com U2AF³⁵ (1), ao interagir com o co-ativador SRm160 (2) ou ao contatar o sítio do ponto de ramificação no RNA (3, ver **seção 1.4.6**).
- (b) Um sítio de *splice* 5' fraco no transcrito FAS é fortalecido por ligação de TIA1 a um ISE a 3'. TIA1 promove cooperativamente a interação de U1 com o pré-RNAm
- (c) A repressão do sítio de *splice* 3' não-sexo-específico de *tra* envolve a interação de SXL com um ISS localizado dentro do trato de polipirimidina e a prevenção da

ligação de U2AF. Isto leva à seleção do sítio de *splice* 3' específico de fêmeas, localizado alguns nucleotídeos a 3' (esquema à esquerda). Em machos (esquema à direita) não ocorre expressão de SXL, e portanto o sítio mais a 5' é usado.

- (d) A inclusão do exon 3 do gene *tat* de HIV1 é determinada pela razão de certas hnRNPs/proteínas SR no núcleo. A multimerização propagativa de hnRNPA1 (inibidor) a partir de um ESS de alta afinidade (esquema à esquerda) é bloqueada estericamente pela interação da proteína SR SF2/ASF com um ESE a 5' (esquema à direita). Nesse caso, a função do ESE requer os domínios de ligação a RNA mas não os RS de interação proteína-proteína.
- (e) A regulação do *splicing* do exon N1 no transcrito *src* fornece um exemplo de controle combinatório por cooperação e antagonismo entre vários fatores que atuam positiva e negativamente. Em células não-neuronais (à esquerda), N1 é excluído, enquanto em neurônios (à direita), é incluído no RNA maduro. Exons constitutivos são mostrados como caixas bege, enquanto exons alternativos são mostrados como caixas azuis. O exon N1 à esquerda é apresentado com SF2/ASF ligado para indicar que o mecanismo não envolve repressão por propagação de múltímeros de proteínas repressoras. KSRP: *KH-type splicing regulatory protein*; nPTB: *neural polypyrimidine tract binding protein* é uma PTB que não apresenta papel repressivo, somente expressa em neurônios.

Extraída de Matlin et al (2005).

Um outro modelo de proteína SR reforçando o reconhecimento de um sítio de *splice*, ainda não totalmente comprovado, é para o reconhecimento do sítio de *splice* 5' do gene *fruitless* de *Drosophila*. O mesmo ESE observado próximo ao sítio de *splice* 3' de *dsx* foi observado próximo ao sítio de *splice* 5' de um exon de *fruitless*. Com base em

experimentos que sugerem que TRA também influencia o reconhecimento do sítio de *splice* 5' de *fruitless*, postula-se que este seja reconhecido com o estabelecimento de uma ligação de TRA ao exon, estabelecendo uma ponte com a proteína U1-70K, componente da snRNP U1 (revisto em Graveley (2000)).

ISEs teriam um papel similar ao de ESEs, ou seja, são sequências próximas aos sítios de *splice* às quais se ligam fatores de *splicing* que auxiliam na ligação de snRNPs U ou outros elementos do spliceossomo. Um exemplo de ISE auxiliando na ligação ao sítio de *splice* 5', de forma análoga a TRA, é TIA1. Este fator de *splicing*, embora seja estimulatório, é uma hnRNP. Esta proteína parece ligar-se ao intron e promover a ligação de U1 ao sítio de *splice* 5', ao estabilizar sua ligação ao RNA (**FIGURA 22b**).

O silenciamento ou repressão do *splicing* por ligação de fatores *trans* a elementos *cis* envolve, em geral, hnRNPs. Vários mecanismos para sua atuação são conhecidos ou postulados. Por exemplo, a proteína PTB (*polypyrimidine tract binding*) ao ligar-se ao trato de polipirimidina impede a ligação de U2AF, resultando na não-formação do complexo E ou ainda ligando-se a um ESS e inibindo a seleção de um sítio de *splice* 3' sub-ótimo (revisado em Wagner e Garcia-Blanco (2001)). Outro exemplo é a proteína SXL, expressa somente em fêmeas de *Drosophila* que inibe um dos sítios de *splice* 3' do gene *tra*, para promover *splicing* de outro sítio de *splice* 3' localizado mais a 3' (**FIGURA 22c**). Um outro mecanismo seria a ligação de fatores repressores a vários elementos em *cis* do mesmo tipo, criando uma “zona de silenciamento” (revisado em Wagner e Garcia-Blanco (2001), Matlin et al (2005)). Nessa “zona de silenciamento” ficam encobertos sítios de *splice*, exons inteiros ou ESEs/ISEs. Essa zona pode ser criada ao cobrir-se um segmento de RNA com proteínas repressoras (**FIGURA 22d**) ou através da formação de um laço contendo o segmento (**FIGURA 22e**). A **FIGURA 22d** mostra ainda um caso de promoção do *splicing* indireto, ou seja, a proteína SR não

promove a ligação de snRNPs ou de U2AF, mas impede o estabelecimento da “zona de silenciamento”.

Por fim, na **FIGURA 22e** é apresentado um mecanismo de controle combinatório onde há envolvimento de PTB, um parálogo de PTB que não inibe *splicing* (nPTB), hnRNPs e possivelmente proteínas SR. O mecanismo é extremamente complexo e ainda não é plenamente compreendido (Matlin et al (2005)).

Outros exemplos de regulação de *splicing* alternativo por interação de fatores de *splicing* com elemento em *cis* podem ser encontrados em Black (2003), Matlin et al (2005).

1.7.3.2 Identificação de motivos de elementos reguladores de *splicing* alternativo em *cis*

Dado o importante papel de elementos reguladores em *cis* na regulação do *splicing* alternativo, vários estudos foram realizados em busca de novos elementos. Recentemente, valendo-se da disponibilidade de grandes quantidades de dados de sequência que provê poder estatístico, diversos trabalhos de bioinformática, algumas vezes associados a técnicas experimentais, têm sido realizados para buscar novos elementos reguladores de *splicing* (revisado em Ladd e Cooper (2002)). O número de trabalhos existentes é extremamente grande e seria impossível listar todos. Abaixo, são citados diferentes artigos para ilustrar algumas das estratégias utilizadas, com destaque àqueles com implicação nos resultados apresentados nesta tese. Uma extensa listagem de elementos reguladores de *splicing* conhecidos pode ser encontrada em Ladd e Cooper (2002).

Estudos experimentais:

- o promotor de *splicing* intrônico (ISE) GGG foi descoberto em introns vertebrados reconhecidos por definição de introns ao invés de definição de exon (McCullough e Berget (1997)). Quando elementos GGG no intron eram mutados, transcritos retendo o intron eram produzidos. Os autores verificaram que a proximidade do ISE aos sítios de *splice* era importante para o funcionamento do elemento regulador. Em trabalho posterior, os mesmos pesquisadores analisaram o modo de funcionamento deste ISE, verificando que a snRNP U1 liga-se a GGG através de pareamento de bases. Esta ligação deve acumular snRNPs próximas ao sítio de *splice* 5', facilitando o reconhecimento da borda exon/intron e por consequência, aumentando a eficiência do *splicing* desses introns (McCullough e Berget (2000)).
- Matrizes de pontuação para motivos de ESEs reconhecidos pelas proteínas SR SRp40, SRp55, SF2/ASF (Liu et al (1998)) e SC35 (Liu et al (2000)) foram construídas a partir da técnica experimental denominada SELEX (*selective evolution of ligands by exponential enrichment*). O método baseia-se na substituição de um ESE conhecido (e imprescindível para *splicing* de um dado exon) por dezenas de motivos aleatórios em um exon de um mini-gene que é submetido a *splicing*. Os produtos são purificados em gel e os transcritos processados com sucesso, provavelmente por conterem um motivo com atividade de ESE, são selecionados. Os ESEs são então substituídos pelos motivos dos transcritos selecionados novamente e esse procedimento é realizado algumas vezes para selecionar os motivos com maior afinidade por proteínas SR. No caso dos estudos aqui referidos, somente uma das proteínas SR em questão é fornecida no passo de *splicing* do mini-gene, de forma que somente ESEs para aquela proteína serão selecionados (ver sumário do procedimento SELEX na **FIGURA 23**). Um problema desta estratégia é

que nem sempre os sítios de maior afinidade são aqueles com maior atividade. Isso pode ser devido ao fato do SELEX ser realizado em condições bastante diferentes daquelas do *splicing* in vivo. De toda forma, diversas sequências resultam desse procedimento e um processo computacional é realizado para se agrupar motivos similares, permitindo produzir matrizes de pontuação para cada um dos sítios de ligação das proteínas SR. Essas matrizes podem ser utilizadas para encontrar tais sítios em uma sequência qualquer (**FIGURA 24**). Cada posição de uma sub-sequência (nesse caso de 5-7 nt) é pontuada separadamente, atribuindo-se diferentes pesos para cada nucleotídeo com base na matriz. Quando a soma dessa pontuação ultrapassa um certo valor, admite-se que aquela sub-sequência corresponde a um sítio de ligação putativo.

- 72 novos ESSs (*exonic splicing silencers*) foram identificados utilizando-se um método denominado FAS (*fluorescent-activated splicing*, Wang et al (2004a)). Nesse método, um mini-gene contendo um exon que é sempre incluído sofre inserção de oligonucleotídeos aleatórios. Quando o exon não é incluído, uma proteína funcional é expressa, sendo que ela emite fluorescência detectável. A presença de oligonucleotídeos que contiverem ESSs causa exclusão do exon do RNAm maduro e pode ser reconhecida pela emissão de fluorescência. Em trabalho posterior, o mesmo grupo analisou funcionalmente os ESSs obtidos em diferentes tipos de *splicing* alternativo (Wang et al (2006)). No primeiro trabalho já havia sido observado que pseudo-exons, exons usados alternativamente e exons que fazem uso de sítio de *splice* alternativo são enriquecidos em ESSs, em comparação a exons constitutivos. Desta vez, ao analisar casos de uso de sítio de *splice* alternativo, verificaram experimentalmente que os ESSs colocados entre dois sítios de *splice* podem inibir a seleção do sítio mais próximo do intron. Além disso, uma parte

desses ESSs (FAS-hex3 classe 2) apresentaram capacidade de aumentar *splicing* de um intron alternativo (retido em alguns transcritos). Ou seja esses ESSs denominados classe 2 podem contrabalancear a retenção de um intron, promovendo sua excisão.

FIGURA 23. O método de SELEX funcional.

(a) Um ESE natural essencial é substituído por oligonucleotídeos aleatórios. (b, c) O mini-gene é submetido a *splicing*. (d) Os produtos são purificados e aqueles que foram processados com sucesso, e possivelmente contêm um ESE funcional, são recolhidos. (e) O ESE putativo que eles contêm é reinserido no mini-gene e um novo ciclo se inicia.

Extraída de Matlin et al (2005).

		ATGTTCAACAAGTGCCTAAGTGTCTTA			
	→				
	→				
	→				
	→				
		1	2	3	4
matriz de pontuação	A	-1.8	+2.2	-1.5	+0.1
	T	+1.9	+1.2	-0.5	+1.1
	C	-1.5	-2.2	-0.7	+0.5
	G	+0.8	+0.2	+1.1	-1.1
		pontuação (ATGT) = +1.6			
		pontuação (TTCA) = +2.5			

FIGURA 24. Uso de uma matriz de pontuação para buscar sítios de ligação de proteínas SR em uma sequência qualquer.

Dada uma sequência qualquer, para cada sub-sequência de tamanho n (setas), cada posição l é pontuada de acordo com a matriz. Cada nucleotídeo recebe uma pontuação (negativa ou positiva). A soma das pontuações de cada posição é comparada a um valor de referência, acima do qual considera-se que a sub-sequência é suficientemente similar à matriz de pontuação e assim a um sítio de ligação da proteína SR considerada. No exemplo, dado um valor mínimo de pontuação de 2.0, a sub-sequência ATGT não seria um motivo válido.

Estudos computacionais:

- Elementos reguladores putativos foram identificados comparando-se exons de genes contendo e não-contendo introns e buscando 5-6mêros super-representados, sob a hipótese que aqueles exons de genes que contêm introns devem estar enriquecidos em ESEs importantes para *splicing* (Fedorov et al (2001)).
- Fairbrother et al (2002) criaram uma estratégia denominada RESCUE (*relative enhancer and silencer classification by unanimous enrichment*) para identificar novos ESEs em humano. Foram selecionados 238 hexâmeros super-representados em regiões próximas às bordas de exons com sítios de *splice* fracos. Os hexâmeros foram agrupados por similaridade e um representante de cada grupo foi validado

experimentalmente. Interessantemente, mostrou-se que há uma relação inversa da distribuição em exons da frequência de SNPs (*single nucleotide polymorphism*) e estes ESEs (Fairbrother et al (2004a)). De acordo com a hipótese que esses ESEs auxiliam a ancoragem de componentes do spliceossomo aos sítios de *splice*, há uma maior frequência desses ESEs próximos às bordas dos exons. Já os SNPs que podem potencialmente mutar ESEs e causar danos ao *splicing* de um dado exon, são encontrados em menor frequência próximos às bordas exon/intron, indicando que há pressão seletiva para a preservação de ESEs.

- Pseudo-exons, ou seja, sequências internas a introns de genes extensivamente anotados foram selecionadas por sua semelhança com exons (flanqueados por sequências similares a “sítios de *splice*” fortes e presença de motivos de ESEs) e nelas foi realizada uma busca por hexâmeros super-representados. Foi encontrado um silenciador potente confirmado experimentalmente (Sironi et al (2004)).
- Uma outra estratégia de identificação de ESEs e ESSs foi comparar exons de regiões não-codificantes (para evitar o viés introduzido por códons) com pseudo-exons (regiões genômicas que não sofrem *splicing*) e regiões 5' não traduzidas de genes que não sofrem *splicing* (Zhang e Chasin (2004)). Octâmeros super (ESEs) e sub-representados (ESSs) foram selecionados, gerando 2069 ESEs denominados PESEs e 974 ESSs denominados PESSs.

Com a disponibilidade de tantos novos elementos reguladores, alguns trabalhos buscaram analisar a distribuição destes elementos em exons constitutivos e alternativos e em introns. Uma vez que nem todos os motivos encontrados em um exon devem ser ESEs, seria razoável esperar que essas sequências apresentem então uma densidade maior destes elementos, o contrário para introns, o que de fato foi observado (Wu et al

(2005), Wang et al (2005a)). Estes dois trabalhos analisaram a distribuição dos SELEX-ESEs e a relação da densidade de ESEs com sítios de *splice* fortes ou fracos, mostrando que exons alternativos possuem uma ligeira depleção destes elementos reguladores, mas que a relação é bastante complexa.

A densidade de RESCUE-ESEs, PESE, PESS e FAS-ESSs também foi analisada em exons constitutivos, alternativos e introns retidos (Zheng et al (2005)) em eventos de *splicing* alternativo detectados por microarray. O estudo revelou diferenças entre exons alternativos e constitutivos, sendo que introns retidos apresentaram densidades similares aos introns não-retidos.

Outro trabalho nessa linha comparou a distribuição dos RESCUE-ESEs em humano, camundongo e fuga (Yeo et al (2004b)). Enquanto não foi observada diferença na distribuição de ESEs, coerente com a conservação de proteínas SR nesses organismos, somente mamíferos apresentaram o ISE GGG. Em contrapartida, fuga parece possuir outros ISE formados por curtas repetições de AC e GT. A diferença entre ISEs nos diferentes organismos também é coerente com a heterogeneidade da estrutura e abundância de hnRNPs nos diferentes organismos.

1.7.3.3 Características dos exons e introns alternativos com implicações mecánísticas

Há certas características das próprias sequências de RNAm que podem levar ao *splicing* alternativo. Por exemplo, como os componentes do spliceossomo devem se ligar aos sinais básicos para *splicing* (sítios de *splice* 5' e 3', trato de pirimidina e o sítio do ponto de ramificação) a composição nucleotídica destes é fundamental à inclusão constitutiva de exons no RNAm maduro. Vários estudos mostraram que **sítios de *splice* mais fracos** (menos conservados) estão associados à ocorrência de *splicing* alternativo

em humano e em outros organismos (Stamm et al (1994), Stamm et al (2000), Zavolan et al (2003), Zheng et al (2005) para citar alguns exemplos). Como esperado, a frequência de exclusão de um exon (Clark e Thanaraj (2002), Zheng et al (2005)) ou retenção de um intron (Zheng et al (2005)) também varia com a força do sítio de *splice*. Quanto mais forte o sítio, menor a frequência de exclusão ou de retenção.

Uma outra característica que está associada a *splicing* alternativo do tipo inclusão alternativa de exon é o comprimento destes. Em geral, **exons incluídos alternativamente são mais curtos** (Stamm et al (1994), Stamm et al (2000), Zavolan et al (2003), Zheng et al (2005) para citar alguns exemplos). Acredita-se que a proximidade demasiada dos sítios de *splice* resulte em impedimento estérico entre as snRNPs, comprometendo o reconhecimento do exon.

Ainda em relação ao comprimento dos exons, o modelo de definição de exon implica em um limite para reconhecimento de **exons longos** (maiores que 300 nt, Berget (1995)). Distâncias muito grandes entre os sítios de *splice* flanqueando um exon seriam problemáticas para a formação de uma ponte protéica ligando as duas extremidades. De fato, observou-se que a inclusão de exons acima de 500 nt é deficiente (Sternier et al (1996), Wang et al (2005a)). Neste trabalho, também foi mostrado que quando os exons longos eram flanqueados por introns curtos, a inclusão do exon era mais eficiente, provavelmente porque os sítios de *splice* eram reconhecidos através do intron. Posteriormente, observou-se uma maior eficiência de inclusão de um exon quando o mecanismo de definição de intron em relação à definição de exon e que, coerentemente, em *Drosophila* e em humano a probabilidade de um exon ser usado alternativamente é maior se ele for flanqueado por **introns longos** (maiores que 250 nt, Fox-Walsh et al (2005)).

A maior eficiência de *splicing* através de introns curtos pode ser explicada pela observação que introns curtos em *Drosophila melanogaster* e *Caenorhabditis elegans*, organismos em que a definição de intron é o mecanismo predominante, possuem praticamente toda a informação necessária para dirigirem seu *splicing* (Lim e Burge (2001)). Já em humano, os introns curtos dependem da presença de elementos reguladores e de outros fatores além dos sinais de *splicing* na sequência.

Em relação à **retenção de introns**, embora tenha sido observado por nós (Galante et al (2004)) e em outros estudos (Sugnet et al (2004), Zheng et al (2005)) que este tipo de *splicing* alternativo frequentemente envolve introns curtos, isto pode ser simplesmente devido a um artefato amostral. É possível que tal enriquecimento da amostra em introns curtos seja devido ao fato que as ESTs que são usadas para evidenciar os introns retidos são sequências tipicamente curtas. É também possível que haja um viés experimental para a clonagem de introns retidos mais curtos.

A presença e distribuição diferencial de elementos reguladores em *cis* em exons alternativos também é uma característica da sequência importante para a ocorrência de *splicing* alternativo. De forma geral, exons alternativos parecem apresentar uma **menor densidade de ESEs e maior densidade de ESSs** (Stamm et al (2000), Zavolan et al (2003), Wang et al (2004b), Zheng et al (2005), Wang et al (2005a) para citar alguns exemplos). A observação que trechos intrônicos próximos às bordas exon/intron de exons alternativos conservados são mais conservados em camundongo que introns comuns sugere que estas regiões contêm elementos reguladores (Sorek e Ast (2003), Sugnet et al (2004)).

A relação de características de sequência como comprimento do exon, conservação intrônica e força do sítio de *splice* parece explicar cerca de 25% da

variabilidade da frequência relativa da inclusão alternativa de exons em humano (Baek e Green (2005)).

Outras características de exons alternativos relacionadas a aspectos evolutivos e não necessariamente mecânicos podem ser encontradas na **seção 1.7.4.4** abaixo.

1.7.3.4 Mecanismos de *splicing* alternativo diversos

- A retenção de intron do gene *msl2* e do elemento P de *Drosophila* é causada por ligação de fatores *trans* reguladores específicos ao pré-RNA. No caso do *msl2*, a proteína SXL liga-se aos sítios de *splice* 5' e 3' que flanqueiam o intron inibindo sua excisão (revisado em Black (2003)). A retenção do intron 3 do elemento P envolve também fatores *trans* que se ligam a sequências internas ao exon, sequestrando U1 para um sítio de *splice* 5' não-funcional (**FIGURA 25**, revisado em Black (2003)).

FIGURA 25. Regulação da retenção do intron 3 do elemento P de *Drosophila*.

O exon 3 possui sítios de ligação para fatores *trans* (PSI e hrp48) e um sítio de *splice* 5' não-funcional. Quando estes fatores *trans* estão presentes, a ligação de U1 ao sítio de

splice 5' não-funcional ocorre e de alguma forma impede a excisão do intron 3, resultando em sua retenção no RNAm maduro. Embaixo, na ausência dos fatores *trans* reguladores, U1 liga-se ao sítio de *splice* 5' funcional e o *splicing* procede normalmente (revisado em Black (2003)). Extraída de Black (2003).

- Graveley (2005) estudou o gene *dscam* de *Drosophila* que apresenta exons removidos de forma mutuamente exclusiva. Graveley verificou que há elementos reguladores no intron competidores, selecionando um ou outro exon. O mecanismo envolve ainda um fator *trans* e a formação de estrutura secundária do pré-RNAm.
- foi mostrado que a ordem de remoção de exons pode levar à remoção de vários exons ao mesmo tempo, sugerindo um mecanismo para a ocorrência de eventos de uso alternativo de múltiplos exons (Takahara et al (2002)).
- a retenção/excisão de um intron do gene de trombopoietina, além de ser influenciada pela fraqueza dos sítios de *splice* e do comprimento do intron, é dependente do *splicing* de introns a 5' (Romano et al (2001)).

1.7.4 Outros estudos de splicing alternativo

1.7.4.1 Outros métodos de identificação de *splicing* alternativo

Dadas as limitações dos dados de cDNA, embora o uso de ESTs e RNAs como evidências das ocorrências de *splicing* alternativo seja o mais comum para detecção dos eventos, outras alternativas foram também usadas.

Oligonucleotídeos correspondentes a junções exon/exon foram utilizados como sondas em experimentos de microarray onde eram hibridizados RNAs extraídos de

células (por exemplo Johnson et al (2003), revisto em Matlin et al (2005), Blencowe (2006)).

Métodos puramente computacionais para identificação de eventos de *splicing* alternativo também têm sido desenvolvidos. Por exemplo, características de exons alternativos conservados foram usadas em *clusters* com baixa cobertura de ESTs para identificar novos exons utilizados alternativamente em humano e camundongo (Sorek et al (2004b)). Um método modificado independente da existência simultânea de ESTs em dois organismos comparados foi empregado em genomas de *D. melanogaster* e *D. pseudoobscura* para identificar novos exons utilizados alternativamente (Philipps et al (2004)). Estratégia similar foi usada em humano, rato e camundongo identificando também introns retidos (Chen et al (2006)). Em uma outra estratégia que não fez uso de características de eventos conservados gerou-se isoformas putativas computacionalmente e verificou-se aquelas que aumentavam a pontuação do alinhamento contra domínios Pfam (Bateman et al (2004)) acima de um certo valor (Hiller et al (2005)). Supostamente, o aumento da pontuação deve ser causado pela inclusão de uma sequência que faz parte do domínio Pfam. Além deste trabalho, outras estratégias *ab initio* foram utilizadas como por exemplo (Ohler et al (2005), Yeo et al (2005)).

1.7.4.2 *Splicing* alternativo em diferentes tecidos

Dado que os diferentes tecidos expressam diferentes conjuntos de genes, é razoável imaginar que diferentes conjuntos de fatores de *splicing* como proteínas SR e hnRNPs sejam expressos, o que é de fato observado (Hanamura et al (1998) e exemplos na revisão de Stamm (2002)). Isso resultaria em uma regulação tecido-específica de *splicing* alternativo, gerando isoformas encontradas exclusivamente em certos tecidos.

Ainda dentro da hipótese que *splicing* alternativo poderia gerar diversidade especificamente para determinados grupos de proteínas (envolvidas em comunicação celular, por exemplo), também é possível que tal diversidade seja concentrada em tecidos onde uma especialização das proteínas seja importante. De acordo com essa idéia, o estudo de Modrek et al (2001) mostrou que cérebro era o tecido que apresentava maior frequência de *splicing* alternativo, entretanto sem levar em conta a cobertura de ESTs de cada tecido. Em trabalho posterior, ao se levar em conta a cobertura, os tecidos com maior número de genes apresentando *splicing* alternativo foram cérebro, testículo e fígado (Yeo et al (2004a)). Os tecidos com menos variabilidade foram músculo, útero, mama, estômago e pâncreas. Em cérebro, os autores ainda identificaram motivos super-representados em exons de genes sujeitos a *splicing* alternativo como candidatos a ESEs, como também realizado por Brudno et al (2001). Esses motivos podem ser elementos reguladores tecido-específico utilizados para a regulação do *splicing* alternativo.

Como alguns tecidos apresentaram consideravelmente mais genes com *splicing* alternativo, é bem possível que parte das isoformas produzidas sejam específicas ao tecido. Uma busca em larga-escala foi realizada para identificar esses casos (Xu et al (2002)), encontrando tais formas mais frequentemente em cérebro, olho-retina, pulmão, fígado, pâncreas, placenta, ovário, útero, testículo, músculo, linfa e pele.

1.7.4.3 *Splicing* alternativo e doenças

A alta frequência de *splicing* alternativo em metazoários não tem apenas efeitos no aumento da diversidade protéica, mas também pode estar relacionada a doenças. Um evento de *splicing* alternativo não-regulado pode ser causado por um descontrole da expressão de elementos reguladores *trans*, pela mutação de sítios de *splice* ou de

elementos reguladores em *cis* (ESEs, por exemplo). Note-se que mesmo mutações sinônimas podem ter efeitos profundos no *splicing* ao danificar elementos reguladores em *cis* importantes ou sítios de *splice*. Além de causar danos diretos à estrutura da proteína, a inclusão ou exclusão indevida de um exon ou até sua extensão/encurtamento, bem como a retenção de um intron, podem causar a inclusão de um códon de parada prematuro, por exemplo. Nesse caso, uma proteína truncada seria expressa ou o transcrito seria degradado por NMD. A relação de mutações em elementos reguladores em *cis*, doenças e NMD foi revista por Cartegni et al (2002). Para outras revisões de *splicing* alternativo envolvido em doenças ver (Maquat (2001), Cartegni et al (2002), Nissim-Rafinia e Kerem (2002), Blencowe (2006)).

Dessa forma, é possível que no futuro, a detecção de diferentes isoformas de *splicing* alternativo seja utilizada rotineiramente como método diagnóstico/prognóstico de doenças, como já realizado com microarrays para melhorar a detecção de tumor de próstata (revisto em Blencowe (2006)).

1.7.4.4 A conservação do *splicing* alternativo em diferentes organismos e evolução

I seem to you to seek a new disaster every day

You deem me due to clean my view and be at peace and lay

I mean to prove I mean to move in my own way

And say I've been getting along for long before you came into the play

“Extraordinary machine”, Fiona Apple

Um dos aspectos do fenômeno de *splicing* alternativo que tem recebido bastante atenção nos últimos anos é a sua evolução. Em sua revisão, Ast (2004) salienta que pelo menos dois modelos para a evolução de *splicing* alternativo poderiam ser propostos. Ambos baseiam-se na existência de genes ancestrais com introns flanqueados por sítios de *splicing* fortes que eram reconhecidos constitutivamente e que passaram a apresentar características tais que possibilitam seu reconhecimento alternativo (sítio de *splice* fracos (Stamm et al (1994)), menor comprimento (Sorek et al (2004a)), divisibilidade por 3 para manter o quadro de leitura aberta (Resch et al (2004a), Sorek et al (2004a))).

O primeiro modelo propõe que mutações nas sequências genômicas causaram enfraquecimento dos sítios de *splice* de tal forma que alguns exons passaram a ser reconhecidos alternativamente. O segundo propõe que o surgimento de proteínas SR e de outros fatores *trans* para auxiliar o reconhecimento de exons constitutivos permitiu que parte desses tivessem sítios de *splice* mais fracos, tornando-se alternativos. Ambos os modelos se baseiam na idéia de que a célula passou a produzir transcritos novos, “experimentais”, sem comprometer a expressão do gene original, idéia proposta por Gilbert (1978).

É interessante notar que a ocorrência de *splicing* alternativo varia entre diferentes organismos. Por exemplo, leveduras em geral apresentam pouquíssimo, ou não apresentam, *splicing* alternativo. Os poucos casos conhecidos referem-se a retenção de intron, e nunca inclusão alternativa de exons. Isso se deve, muito provavelmente, ao fato de que nesses organismos o mecanismo de reconhecimento de introns prevalece. Além disso, ao se comparar os sítios de *splice* de leveduras com o de mamíferos notam-se algumas diferenças importantes. Basicamente, nota-se que o sítio de *splice* 5’ de *S. cerevisiae* é mais “rígido” que o de mamíferos, sendo que os nucleotídeos intrônicos são

muito conservados. De *S. pombe* para mamíferos parece ter ocorrido uma flexibilização do sítio de *splice*, possivelmente associado à ocorrência de *splicing* alternativo. Em mamíferos os nucleotídeos do intron são menos conservados, compensados por uma maior conservação dos nucleotídeos exônicos (revisado em Ast (2004)).

Embora seja tentador tomar a ausência de *splicing* alternativo em leveduras como uma evidência da direcionalidade da evolução do fenômeno, já que este é abundante em mamíferos, é possível que esses organismos tenham sofrido um processo de perda de variabilidade por razões de “economia”. Ou seja, por serem organismos de crescimento rápido onde erros são menos tolerados, a ocorrência de *splicing* alternativo poderia ser selecionada negativamente, já que é também um processo que pode resultar em proteínas não-funcionais (Ast (2004)).

Vários outros aspectos da evolução de *splicing* alternativo foram estudados (revisado em Lareau et al (2004)), como por exemplo:

- Como mencionado acima, embora alguns estudos tenham observado uma baixa taxa de conservação de eventos de *splicing* alternativo entre humano e camundongo (~25%, Nurtdinov et al (2003), Sorek et al (2004a)), estes estudos não consideraram a baixa cobertura de ESTs que afeta diretamente a detecção de eventos de *splicing* alternativo. Quando a cobertura é considerada, foi encontrado que pelo menos 61% de eventos são conservados entre os dois organismos (Thanaraj et al (2003)).
- Os trechos intrônicos próximos às bordas exon/intron de exons alternativos conservados entre humano e camundongo são mais conservados que introns comuns, possivelmente por conterem elementos reguladores (Sorek e Ast (2003), Sugnet et al (2004)).
- O grau de conservação de *splicing* alternativo entre humano e camundongo não é o mesmo para toda a população de eventos. Na realidade, ela varia com a frequência

em que os exons são incluídos nos transcritos (Modrek e Lee (2003)). Os exons alternativos que são incluídos na maior parte dos transcritos humanos (forma majoritária) também são alternativos em 98% dos transcritos de camundongo; sendo que esta é a mesma fração para exons constitutivos. Já para os exons alternativos incluídos na minoria dos transcritos, apenas um quarto são conservados. Ou seja, estes últimos parecem ser provenientes de um processo recente de perda/ganho. A observação de que exons constitutivos e alternativos majoritários não foram ganhos nem perdidos recentemente sugere que os minoritários apresentam uma taxa de evolução mais acelerada, suportando a idéia de que novas isoformas alternativas geradas estariam livres de pressão seletiva e serviriam como um laboratório de novas proteínas.

- Deve-se notar, no entanto, que a extrapolação feita por Thanaraj et al (2003) para considerar a falta de cobertura de ESTs indicou que 60% das formas minoritárias seriam conservadas.
- No estudo de Sorek et al (2004a), foi mostrado que entre os exons conservados entre humano e camundongo há uma maior frequência de exons com comprimento divisível por 3, mais curtos, sem códons de parada e sem elementos repetitivos. Preservação do quadro de leitura também foi uma característica dos exons alternativos conservados entre camundongo e humano detectados *ab initio* (Yeo et al (2005)) e de exons alternativos incluídos na minoria das ESTs (Resch et al (2004a)).
- De acordo com as observações apresentadas acima, análises evolutivas de exons alternativos conservados revelaram que os exons incluídos minoritariamente estão sob menor pressão seletiva negativa que exons constitutivos (Xing e Lee (2005)). Intrigantemente, os exons alternativos que apresentam taxa de inclusão mais baixa

sofrem maior pressão para preservação do quadro de leitura e para evitar mutações sinônimas. Os autores sugerem que isso ocorra para preservar elementos reguladores como ESEs. De fato, há exemplos experimentalmente conhecidos de mutações sinônimas que alteraram o padrão de *splicing* (Cartegni et al (2002)).

- Em 48 de 73 casos de *splicing* alternativo estudados com base em dados de proteínas, a forma ancestral (encontrada em procariotos) era a mais longa, indicando que a forma derivada era resultante de exclusão alternativa de exon (Kondrashov e Koonin (2001)). Supostamente, o ganho de novas partes codificadoras é mais difícil que a perda. Lareau et al (2004) salientam que, no entanto, a adição de um exon não-codificante que induza RUST (degradação por NMD) poderia ocorrer com maior facilidade.
- Um dos mecanismos de geração de exons alternativos pode ser a duplicação de exons. Por exemplo, para 10% dos eventos de *splicing* alternativo foi encontrada evidência de origem por duplicação em tandem (Kondrashov e Koonin (2001)). Outro estudo mostrou que de 2438 exons duplicados (de 10% dos genes de humano, *Drosophila* e *C. elegans*), a maioria estava envolvida em *splicing* alternativo mutuamente exclusivo (Letunic et al (2002)).
- Uma fração (5.2%) dos exons alternativos humanos também parece ter surgido a partir de elementos repetitivos Alu (Sorek et al (2002)). Alguns estudos acerca de mecanismos de exonização de Alu foram realizados. Por exemplo, a exonização do exon denominado AluYa5 do gene que codifica a enzima conversora de angiotensina envolveu mutações para criação de sítios de *splice* 3' e 5' e de um ESE (Lei et al (2005)).
- Wang e Brendel (2006) compararam eventos de *splicing* alternativo entre *Arabidopsis* e arroz, mostrando que retenção de introns é o evento mais comum nos

dois organismos, até ~30% dos eventos de *splicing* alternativo são conservados, 40% dos genes apresentam *splicing* alternativo nos dois organismos e que 36-43% dos eventos seriam sujeitos a NMD.

1.8 Bioinformática: uma nova maneira de abordar velhos e novos problemas

Como mostrado acima, muitos dos trabalhos sobre *splicing* alternativo realizados nos últimos anos foram exclusivamente análises de Bioinformática (por exemplo, Lim e Burge (2001), Modrek et al (2001), Kan et al (2002), Kriventseva et al (2003), Lewis et al (2003), Zavolan et al (2003), Baek e Green (2005), Wang e Brendel (2006), para citar alguns).

A Bioinformática é mais que apenas um conjunto de métodos, técnicas e ferramentas provenientes da interação de biólogos com cientistas da computação, ou da utilização de programas de computador para resolver problemas biológicos. A Bioinformática trouxe uma nova maneira de estudar problemas biológicos ao se considerar que grandes coleções de dados podem ser analisadas como um todo, tal qual um problema estatístico. É verdade que algumas áreas da Biologia já utilizavam análises matemáticas, estatísticas ou computacionais. Uma área pioneira é sem dúvida a de análises filogenéticas, em que o emprego de algoritmos de computador são essenciais para a análise de dados e construção de árvores filogenéticas (que em verdade é uma aplicação particular de técnicas de agrupamento largamente usadas em Ciência da Computação, como por exemplo, *neighbor-joining*). Assim, na área de Biologia Molecular, fora das análises filogenéticas, a conversão de dados biológicos em formato

de computador, a padronização dos dados, a coleção destes em bancos e a postura de encarar estes dados como retratos de fenômenos é relativamente nova, embora alguns estudos pioneiros sejam antigos (na área de *splicing* por exemplo, Shapiro e Senapathy (1987b), Stephens e Schneider (1992)).

Epistemologicamente, as prerrogativas gerais da Bioinformática, principalmente aquelas usadas nesta tese, voltada à análise de sequências, podem ser resumidas como se segue:

- i- existem padrões biológicos presentes em um número razoável de casos que podem ser descobertos ao se estudar uma coleção razoável destes casos
- ii- os padrões observados podem ser causados ou serem a causa de fenômenos biológicos relacionados aos objetos de estudo
- iii- uma história evolutiva comum pode originar tais padrões (subentendendo-se que um mecanismo comum também, mas este deve ter sido selecionado e portanto implica em história evolutiva comum)
- iv- a identificação e estudo destes padrões pode revelar novos conhecimentos acerca de fenômenos biológicos relacionados ao objeto de estudo
- v- a correlação dos padrões com outras características do fenômeno estudado é um meio de realizar o item iv

De certa forma, a Bioinformática vale-se então do estudo de amplas coleções de registros de fenômenos biológicos em formato de computador, e da busca por sinais ou padrões emergentes que estejam relacionados ao fenômeno estudado. Uma boa análise de Bioinformática envolve necessariamente um amplo conhecimento biológico da área estudada, dos métodos disponíveis e também de criatividade para executar as análises.

Na minha opinião, é principalmente na elaboração das análises baseada em conhecimento biológico, e em parte em sua implementação correta, que a Bioinformática se apresenta como mais que um conjunto de técnicas. Por isso é possível fazer da Bioinformática uma prática de ciência baseada em hipótese sem restringir-se à criação de ferramentas e métodos ou simples catalogações de um dado fenômeno. Em outras palavras, a Bioinformática permite realizar uma pesquisa embasada teoricamente e profundamente baseada em hipóteses, mas que visa estudar fenômenos de forma abrangente, uma vez que se ocupa em descobrir padrões gerais, sem deixar de se preocupar com detalhes, como por exemplo um mecanismo molecular complexo. Deve-se notar que este é um objetivo não menos importante desta tese.

Embora a Bioinformática possa ser usada para realizar experimentos *in silico* e assim testar hipóteses ao tratar grandes coleções de dados simultaneamente, seu papel como geradora de novos panoramas e *insights* é fundamental para a proposição de hipóteses que possam ser tratadas experimentalmente em trabalhos posteriores. Por exemplo, estudos que poderiam ser considerados pioneiros na área de Bioinformática utilizaram argumentos estatísticos ou puramente baseados em análises de sequência para gerar *insights* importantíssimos em áreas fortemente experimentais. Na área de *splicing*, poderia-se citar como exemplos: Lerner et al (1980), Rogers e Wall (1980), baseando-se em comparação de sequências, predisseram que U1 faria pareamento de bases com o sítio de *splice* 5', o que de fato foi constatado. Outro exemplo é o suporte à hipótese de definição de exons proveniente do fato que a grande maioria dos exons metazoários são mais curtos que os introns flanqueantes (revisto em Berget (1995)). A hipótese certamente não teria desbancado a definição de introns em mamíferos sem o argumento estatístico. Outro exemplo é a observação de que há uma relação inversamente proporcional da força do trato de polipirimidina e a conservação de AG

que gerou o *insight* do papel de U2AF³⁵ no reconhecimento do sítio AG (Reed (1989)). Estudos do conteúdo de informação dos sítios de *splice* aumentaram a compreensão de como esses sinais são reconhecidos e trouxeram novas idéias de como eles podem ter evoluído (Stephens e Schneider (1992)). Ou seja, embora a Biologia seja uma ciência experimental, a geração de *insights* é crucial e essa é uma das principais contribuições que a Bioinformática pode fazer.

1.9 Característica das sequências de RNAm envolvidas na retenção de introns

A esmagadora maioria dos estudos descritos acima versaram sobre basicamente um tipo de *splicing* alternativo, a inclusão alternativa de exons. O tipo de *splicing* alternativo menos estudado é, sem dúvida, a retenção de introns (RI). Este tipo de evento é o mais raro em mamíferos e muitas vezes é excluído das análises pela incerteza de que os transcritos que contêm um intron não são simplesmente RNAm parcialmente processados. Só mais recentemente as análises têm incluído ou até se concentrado sobre esse tipo de evento (por exemplo, Zheng et al (2005), Ohler et al (2005), Hiller et al (2005), Chen et al (2006)).

A RI é inesperada uma vez que afeta o transporte do RNAm ao citoplasma (Nott et al (2003)) e pode inserir códons de parada prematuros no transcrito maduro que seria então degradado por NMD (**seção 1.6.2**). Na verdade, diminuir os níveis de RNAs pode ser uma das funções do *splicing* alternativo como já discutido acima. Este, entretanto, pode não ser o destino de todos os RNAs com um intron retido. Por exemplo, mostrou-se que uma isoforma do gene *tgif2* de camundongo que retém um

intron na região codificadora possui atividade biológica (Melhuish e Wotton (2006)). Além disso, mostramos previamente que introns retidos no transcriptoma humano possuem sinais de serem biologicamente funcionais, como conteúdo de GC e uso de códons similares àqueles de exons codificantes e a capacidade de codificar domínios protéicos (Galante et al (2004)).

Em relação à natureza mecanística da RI, a explicação mais intuitiva é que estes introns são flanqueados por sítios de *splice* fracos que ocasionalmente não são reconhecidos corretamente. Stamm et al (2000) observaram uma menor conservação dos sítios de *splice* flanqueando introns retidos. Estudos com genes humanos mostraram que o fortalecimento de sítios de *splice* sub-ótimos flanqueando introns retidos causaram um aumento nos seus níveis de remoção (Romano et al (2001)) ou aboliram a retenção completamente (Dirksen et al (1995)). Por outro lado, outros estudos mostraram que enfraquecer sítios de *splice* ótimos levam à retenção de intron (gene de troponina I de frango, Sterner e Berget (1993), α -globina 2 humana, McCullough e Berget (1997), gene *zeste* de *Drosophila melanogaster*, Talerico e Berget (1994), gene *cdc2* de *S. pombe*, Romfo et al (2000)).

Interessantemente, todos os introns envolvidos nestes estudos eram curtos (< 274 nt). Esta observação pode ter uma razão mecanística de acordo com os modelos atuais de reconhecimento de sítios de *splice*. No modelo de definição de exon (**seção 1.3.3**), o não-reconhecimento de um par de sítios de *splice* fracos levaria à não-inclusão do exon e não a RI. Entretanto, quando os introns são curtos prevalece a definição de intron (**seção 1.3.3**) e o não-reconhecimento de sítios de *splice* levaria à retenção do intron, mesmo em vertebrados.

Além da variação de sítios de *splice*, alguns estudos mostraram que a expressão das isoformas que retêm e que excisam o intron varia em diferentes tecidos (Popielarz et

al (1995), Galante et al (2004), Ledee et al (2004)) indicando que fatores celulares também estão envolvidos no controle da retenção de intron. De fato, o reconhecimento dos sítios de *splice* depende de diversos fatores como por exemplo elementos reguladores localizados na sequência de RNAm (**seção 1.7.3**) e na concentração de fatores protéicos (**seções 1.4.4.1 e 1.7.3**). Foi mostrado que a presença de elementos reguladores específicos pode de fato contrabalancear a retenção de introns individuais (Dirksen et al (1995), Lejeune et al (2001)).

Embora estudos sobre introns retidos individuais (por exemplo, Sterner e Berget (1993), Dirksen et al (1994), Popielarz et al (1995), Carlo et al (1996), Romano et al (2001), Ledee et al (2004), Michael et al (2005), Melhuish e Wotton (2006)) e algumas poucas análises em larga-escala tenham sido realizadas (Stamm et al (2000), Zheng et al (2005)), poucas análises sistemáticas sobre características gerais das sequências de RNAm envolvidas no mecanismo de retenção foram conduzidas.

Para entender como características da sequência estão associadas ou controlam a RI em humano de maneira geral, conduzimos uma análise de Bioinformática utilizando dados de genoma e de transcriptoma (ESTs e RNAs). Os eventos de retenção de intron foram identificados em cDNAs alinhados ao genoma e particionados em dois grupos, de acordo com a abundância relativa de ambas as isoformas: um grupo contendo os casos em que a forma que retém o intron era minoria e um segundo grupo em que eram a maioria. Características da sequência como conservação dos sítios de *splice*, comprimento dos introns e outras foram comparadas entre os dois grupos sob a hipótese que os fatores que apresentassem contrastes poderiam ser responsáveis pelas diferentes frequências relativas de retenção.

Além dos resultados terem confirmado que sítios de *splice* fracos estão associados à retenção de intron, como previamente conhecido, eles mostraram que essa

relação pode explicar, em parte, também as diferentes frequências de retenção. No entanto, uma fração não desprezível dos eventos não pode ser explicada por essa característica apenas. Nossa análise revelou que a densidade de um conjunto específico de ESSs (FAS-hex3 classe 2, Wang et al (2004a)) e do promotor de *splicing* intrônico GGG devem ser cruciais para gerar diferentes frequências relativas de retenção. Outras características relacionadas a RI previamente desconhecidas como genes com um maior nível de expressão gênica e introns em geral mais curtos também foram reveladas na nossa análise. Parte destas características foram analisadas em eventos conservados em camundongo, suportando as observações em humano.

A impossibilidade de explicar todos os casos de retenção de introns pelas características de sequência analisadas indica que outros fatores mais complexos, como por exemplo formação de estrutura secundária no RNAm, devem estar envolvidos em uma fração dos eventos. Este estudo salienta a importância de se considerar o compromisso entre essas características na regulação da frequência relativa de retenção de intron e não somente algumas poucas características como a força dos sítios de *splice*. O trabalho reforça também a importância de se analisar os eventos divididos pela frequência relativa em que as isoformas ocorrem.

1.9.1 A busca por um código de splicing

Como exposto acima, o *splicing* alternativo parece ser um fenômeno de enorme importância biológica, tanto por causa da enorme variabilidade protéica que ele permite alcançar, como pelo seu papel em diversos processos fisiológicos normais e em doenças. Assim, o entendimento de como e quando se dá a geração das diferentes isoformas é de grande interesse.

Mais recentemente, tem amadurecido a idéia de se compilar todo o conhecimento mecanístico dos diferentes tipos de *splicing* alternativos e determinar o conjunto de regras que os regem, o hipotético “código de *splicing*” (ver por exemplo Wang et al (2004b), Matlin et al (2005), Blencowe (2006)). Hipoteticamente, esse código permitiria prever as isoformas produzidas por um gene apenas com base em dados sobre a sequência, por exemplo. Também é de interesse poder prever em que tecido, sob que condições uma dada isoforma é expressa e como uma mutação poderia alterar o padrão de *splicing* de um dado gene.

A maior parte dos métodos *ab initio* de detecção de eventos de *splicing* alternativo tem utilizado características de conservação evolutiva como preditores. Nesses métodos, diversas características de exons alternativos conservados são fornecidas a um reconhecedor automático, que é utilizado para identificar casos semelhantes ainda não conhecidos. Por exemplo, para detectar exons alternativos conservados entre camundongo e humano utilizou-se como características o comprimento dos exons, comprimento dos introns flanqueantes, força dos sítios de *splice*, identidade dos exons, identidade dos introns (Yeo et al (2005)). HMMs (*hidden Markov models*, modelos de Markov ocultos) treinados para reconhecer eventos conservados entre camundongo e humano pela identidade da sequência também foram utilizados (Ohler et al (2005)). Alguns outros preditores de primeira-geração foram construídos (Lim e Burge (2001), Wang et al (2004b)), utilizando características como pontuação para sítios de *splice* putativos e para elementos reguladores de *splicing* putativos, por exemplo. Estudos como o realizado nesta tese poderiam, eventualmente, ser úteis para o avanço de tais preditores.

2 *Objetivos*

I – Estudar as características das sequências de RNAm humanas que pudessem estar envolvidas no mecanismo de retenção de intron, com atenção aos elementos reguladores *em cis*.

II – Produzir um sistema de anotação de *splicing* alternativo que possibilitasse o uso inteligente dos dados, a reutilização de código e assim, análises mais rápidas e eficientes.

III – Realizar um estudo de bioinformática que envolvesse um estilo de ciência baseada em hipóteses, sem recair no modelo ordinário de catalogação de um fenômeno, gerando *insights* mecanísticos sobre um problema relevante à área de *splicing* alternativo.

Página propositalmente deixada em branco

3 *Materiais e métodos*

My method is uncertain

It's a mess but it's working

And maybe if you try it out

you won't like it when you're crying out

“Please please please” - Fiona Apple

3.1 Dados primários utilizados

Para detecção dos eventos de retenção de intron foram utilizadas basicamente a sequência genômica humana e sequências expressas disponíveis (representando o transcriptoma humano). Tanto sequências expressas parciais conhecidas como *Expressed Sequence Tags* (ESTs) e completas (RNAm) (genoma montagem 35 e cDNAs de maio de 2004 obtidos da página do Genome Browser da Universidade da Califórnia em Santa Cruz, Karolchik et al (2003)) foram utilizadas para construir um mapa de transcritos no genoma humano. Brevemente, as sequências eram alinhadas no genoma humano com SIM4 (Florea et al (1998)), sendo que somente o melhor alinhamento (aquele com maior identidade percentual com o genoma humano e com a maior extensão alinhada) para cada sequência foi mantido, e agrupadas com base na sobreposição das suas coordenadas genômicas. Para serem incluídas no mesmo *cluster*, duas sequências tinham que apresentar ao menos uma borda exon/intron similar (± 3 nt) ou uma sobreposição de pelo menos 15 nt se as sequências se sobrepunham somente nas extremidades.

Assim, cada agrupamento representa idealmente os transcritos de um gene individual (**FIGURA 21**). O processo de agrupamento de cDNAs foi implementado por

Pedro AF Galante e está documentado na ferramenta apresentada a seguir (seção 5-B “Clustering cDNAs” da documentação da ferramenta ASC). Estes agrupamentos de cDNA constituem o dado inicial das análises.

3.2 Anotação de *splicing* alternativo no transcriptoma humano

Os bancos de dados que armazenam basicamente as sequências de transcritos ou genomas, ou qualquer tipo de informação biológica pura são conhecidos como **bancos primários**. Sobre estes dados é necessário então construir **bancos secundários** que identifiquem evidências de fenômenos de interesse, como explicado acima, e as armazenem de forma inteligente, dando poder analítico sobre os dados primários. Ou seja, bancos secundários guardam, em geral, **anotações** ou **conhecimento** sobre eventos biológicos. Atualmente existem diversos bancos de *splicing* alternativo, como por exemplo o ASDB (Gelfand et al (1999)), ASD (Thanaraj et al (2004)), o ASAP (Lee et al (2003)), e outros. Muitos destes bancos são relativamente recentes e não existiam na época do início do trabalho aqui apresentado, sendo que muitos mantinham somente eventos confirmados experimentalmente. Além do mais, a elaboração de um banco de dados de *splicing* alternativo requer modelagem do fenômeno. Por exemplo, o estabelecimento de critérios de identificação de um dado evento de *splicing* alternativo é uma etapa de tal modelagem. Realizar a modelagem a partir dos dados brutos permite um maior controle sobre a anotação e evidentemente gera um conhecimento inestimável ao modelador.

Dessa forma, para este trabalho, em conjunto com Pedro A. F. Galante, desenvolvi uma ferramenta para anotação de eventos de *splicing* alternativo detectados

no transcriptoma humano. Esta ferramenta, denominada ASC (Automatic Annotation of Alternative *Splicing* based on cDNA *Clusters*, disponível em <http://www.compbio.ludwig.org.br/asc>) tem como dados primários uma representação do transcriptoma humano, baseada no mapeamento e agrupamento de sequências de cDNA em relação ao genoma humano. A ferramenta ASC realiza uma anotação de *splicing* alternativo para cada agrupamento de cDNAs, baseando-se em todos os cDNAs presentes (em realidade somente os que apresentam pelo menos um intron) e gera uma base de dados MySQL secundária onde os eventos de *splicing* alternativo estão anotados e prontos para consulta.

A ferramenta possui uma vasta documentação que detalha os diversos aspectos e é apresentada no **Anexo A** (também disponível em <http://www.compbio.ludwig.org.br/asc>). Somente alguns pontos de maior relevância são levantados nesta seção.

- 1- mapeamento e agrupamento de cDNAs no genoma humano – as coordenadas dos cDNAs indexadas pelo genoma são comparadas umas contra as outras. Se duas sequências apresentarem ao menos uma borda exon/intron comum são consideradas provenientes do mesmo gene. Se elas não apresentarem nenhuma borda em comum, ainda assim podem ser agrupadas se tiverem sobreposição mínima *das extremidades* de N (padrão = 15) nucleotídeos.
- 2- a ferramenta de anotação implementa uma modelagem dos eventos de *splicing* alternativo que visa agrupá-los em torno dos exons de um dado gene. Os eventos anotados foram: inclusão alternativa de exon (a presença/ausência de um dado exon), retenção de intron (a presença/ausência de uma sequência intrônica), uso de sítios de *splice* alternativos (o encurtamento/alongamento de exons).

- 3- a base de dados produzida foi concebida para ser consultada de forma orientada a objeto, ou seja, o usuário possui à sua disposição uma interface contendo objetos como Exon, Transcrito, *Cluster* de cDNA, etc. para realizar suas consultas. Por ser uma base de dados secundários, todo processo de, por exemplo, identificar todas as sequências que possuem um dado exon já foi realizada e as informações podem ser acessadas facilmente pelo usuário. A escrita de programas complexos para consulta da anotação é simplificada ao não requerer que o usuário i- conheça todos os detalhes da base de dados, já que oferece uma interface de nível mais alto, ii- escreva métodos para consultas básicas, iii- reescreva códigos similares ou iguais para tarefas de rotina.
- 4- a interface permite uma consulta a dados como por exemplo: obter todos os eventos de *splicing* alternativo de um certo tipo, o número de cDNAs que representam um dado evento de *splicing* alternativo, bem como os números de acesso destes cDNAs, se um dado exon está envolvido em retenção de intron, e muitos outros. Para outros exemplos e maiores detalhes, ver a documentação da interface em http://www.compbio.ludwig.org.br/asc/api_doc/basic
- 5- a ferramenta permite gerar diferentes anotações, conforme os parâmetros fornecidos pelo usuário. Por exemplo, utilizar somente RNAs para anotar eventos, ou requerer um número mínimo de cDNAs para anotar a existência de um exon.
- 6- RNAs são usados para determinar a orientação do agrupamento de cDNAs na fita de DNA.

3.2.1 Anotação de eventos de retenção de intron e filtragem do conjunto de dados

Somente sequências com evidência de terem sido processadas pelo spliceossomo (apresentando vãos no alinhamento ao genoma) foram usadas para anotação de *splicing* alternativo. Retenção de intron, inclusão alternativa de exon e exons constitutivos foram identificados em *clusters* de cDNAs por comparação de coordenadas genômicas com nossa ferramenta ASC. Um evento de RI foi anotado para qualquer ocorrência de pelo menos um cDNA (ESTs e RNAs) definindo um intron através da existência de seus exons flanqueantes e pelo menos um outro cDNA contendo pelo menos partes (> 1 nt) dos exons flanqueantes e o intron.

Para obter um conjunto de eventos de RI de alta confiança, os conjuntos iniciais de 6370 eventos de baixa e 1838 de alta-FRI foram filtrados para 1515 e 250 eventos, respectivamente (ver definição de FRI na pág. 123). Somente aqueles eventos nos quais ambas as formas retendo e não-retendo o intron eram confirmadas por pelo menos dois cDNAs cada (no caso de ESTs, de diferentes bibliotecas) foram aceitos. Também, introns retidos com bordas de *splice* alternativas foram descartados. A única exceção aceita foi quando a borda mais frequente era confirmada por >10 cDNAs e havia apenas uma borda alternativa evidenciada por um único cDNA. Somente a borda mais frequente foi usada nas análises. Introns retidos mais curtos que 50 nt foram descartados.

3.3 Identificação de sítios de *splice*

Para cada evento de RI do conjuntos de dados filtrado, uma sequência representativa foi selecionada para guiar a extração de sítios de *splice* do genoma. Para

evitar bordas de *splice* alinhadas erroneamente, somente bordas exon/intron identificadas por SIM4 que não apresentavam vãos ou desalinhamentos em uma janela de 20 nt no exon a partir de ambas as bordas foram aceitas. Somente alinhamentos com o sentido dos sítios de *splice* coerente com a fita de alinhamento de RNAs no *cluster* de cDNA foram aceitos. O conteúdo de informação de bordas de *splice* foi calculado usando o programa rseq do pacote DELILA (Schneider et al (1984)) para as posições -3 a +6 para o sítio de *splice* 5' e para -20 a +3 para o sítio de *splice* 3'. Logos dos sítios foram gerados usando WebLogo (Crooks et al (2004)).

3.3.1 Cálculo do conteúdo de informação de Shannon para avaliação da força dos sítios de splice

O conteúdo de informação é um valor expresso em bits que revela quanta informação uma dada posição em uma sequência de DNA possui (Stephens e Schneider (1992)). A idéia é que haveria um “reconhecedor” biológico, uma máquina, que procuraria por informações sobre onde se ligar ao DNA. Se uma dada posição l sempre é ocupada por G, esta é uma informação inequívoca para o reconhecedor. Se uma outra posição contiver A, T, C ou G com iguais probabilidades, essa posição não é informativa.

O conteúdo de informação é expresso em unidades de informação, bits. De maneira mais informal, 1 bit corresponde à quantidade de escolhas necessárias para se descobrir uma informação, ou ainda o número de perguntas (ou respostas) necessárias. Imagine-se que o reconhecedor pode fazer indagações a uma posição nucleotídica. Primeiramente, ele indaga àquela posição se ela é ocupada por uma purina. A base pode responder binariamente, “sim” ou “não”. Digamos que ela responda “sim”, de forma que ela só pode ser uma citosina ou uma guanina. O reconhecedor então indaga se

aquela posição contém uma citosina. Sendo “sim” ou “não”, o reconhecedor obtém a informação exata de qual nucleotídeo se encontra naquela posição. Assim, 1 bit é a quantidade de informação necessária e suficiente para escolher entre 2 possibilidades. No caso de nucleotídeos, o máximo de informação necessária e suficiente para se informar a ocupação de uma posição é portanto 2 bits.

Desse racional, deriva-se a seguinte fórmula para cálculo da incerteza H de uma dada posição:

$$H(l) = - \sum_{b \in \{a, t, c, g\}} f(b, l) \log_2 f(b, l)$$

onde l é cada posição do sítio considerado, $f(b, l)$ é a frequência de cada base b na posição l .

Pode-se verificar que caso um mesmo nucleotídeo seja sempre observado em uma dada posição, a incerteza será 0. Se qualquer dos 4 nucleotídeos forem possíveis em igual frequência, a incerteza é de 100%, já que a posição não é informativa.

O conteúdo de informação R da posição é então:

$$R_{sequence}(l) = 2 - H(l)$$

O conteúdo de informação de um sítio qualquer é:

$$R_{sequence} = \sum_l R_{sequence}(l)$$

Assim, o conteúdo de informação é calculado com base na frequência das bases de um sítio qualquer de sequências alinhadas. Quanto maior seu valor, maior a conservação de um sítio qualquer, e mais informação contém.

3.3.2 Cálculo da “pontuação de Shapiro & Senapathy” para avaliação da força dos sítios de splice

Esta medida de força (pontuação S&S, Shapiro e Senapathy (1987b)), ou mais apropriadamente, de conservação dos sítios de *splice* é baseada também na frequência das bases de um sítio qualquer. No entanto, ela foi criada para pontuar sítios individuais com base em uma referência canônica de conservação. Para este fim, Shapiro e Senapathy criaram um conjunto de sítios de *splice* 5' (posições -3 a +5 somente contendo GT nas posições 0 e +1, respectivamente) e 3' (posições -14 a +1 somente contendo AG nas posições -1 e 0, respectivamente) flanqueando exons humanos e construíram um modelo que pontua cada característica dos sítios de *splice* com diferentes pesos.

Fórmula para a pontuação do sítio de *splice* 5':

$$\text{Pontuação S\&S} = 100 (t - \text{mint}) / (\text{maxt} - \text{mint})$$

onde

t: soma das frequências no conjunto de referência correspondentes a cada nucleotídeo das posições -3 a +5 na sequência teste

mint : a soma de todas as menores frequências em cada uma das posições do conjunto de referência (pior sítio possível)

maxt : a soma de todas as maiores frequências em cada uma das posições do conjunto de referência (melhor sítio possível)

Fórmula para a pontuação do sítio de *splice* 3':

$$\text{Pontuação S\&S} = 100 \left((t1-l1)/(h1-l1) + (t2-l2)/(h2-l2) \right) / 2$$

onde

t1: soma das 8 frequências mais altas de 10 frequências em posições de pirimidina

t2: soma das 4 frequências em posições CAGG (-2 a +1)

l1: soma das 8 frequências mais baixas das 10 mais baixas em posições de pirimidina (pior sítio possível)

l2: soma das 4 frequências mais baixas em posições CAGG (-2 a +1) (pior sítio possível)

h1: soma das 8 frequências mais altas das 10 mais altas em posições de pirimidina (melhor sítio possível)

h2: soma das 4 frequências mais altas em posições CAGG (-2 a +1) (melhor sítio possível)

Para uma dada sequência de bases, verifica-se a frequência em que cada base é encontrada na referência canônica, na posição correspondente. A frequência de cada posição é somada para gerar um valor observado para a sequência inteira (valores t). A pontuação final é simplesmente a razão dos valores observados (subtraído do mínimo possível: soma das frequências mais baixas) e do máximo possível (subtraído do mínimo possível: soma das frequências mais baixas). A pontuação máxima por sítio de *splice* é 100. Assim, a pontuação pode ser entendida como uma porcentagem de

conservação de um dado sítio em relação ao sítio mais conservado possível (referência canônica), acima de um limite mínimo.

3.4 Medição do nível de expressão gênica

Embora os cDNAs em um *cluster* representem os transcritos de um gene, a quantificação da expressão gênica não é precisa quando realizada contando-se a quantidade de cDNAs conhecidos para um dado gene (tamanho do *cluster*), já que esta técnica amostra somente os genes mais expressos e muitas das bibliotecas de ESTs são normalizadas, ou seja, têm os transcritos mais abundantes removidos. Além disso, as ESTs existentes para um gene provêm dos mais variados tecidos e condições patológicas. Assim, deve-se utilizar técnicas de quantificação do nível de expressão gênica mais sensíveis, como por exemplo a técnica de “*Serial Analysis of Gene Expression*” (SAGE, Velculescu et al (1995)) na qual são produzidas mini-ESTs de até 14 nt. Devido ao seu pequeno tamanho e à forma como são geradas, tais “mini-ESTs” conhecidas como SAGE “tags” (etiqueta gênica) representam a quantidade de transcritos produzidos por um gene de maneira mais fiel e podem ser facilmente manipuladas *in silico*. Assim, para um dado gene conhecido, determina-se a etiqueta que este geraria e contabiliza-se nas bibliotecas de SAGE quantas vezes ela aparece, resultando em uma estimativa da expressão gênica.

O número de tecidos no qual um dado gene é expresso (amplitude) e os níveis de expressão foram determinados através de contagens de etiquetas de SAGE. Bibliotecas de SAGE-curto (10 nt) foram obtidas da página do SAGE Genie (<http://cgap.nci.nih.gov/SAGE>).

Níveis de expressão: para um RNAm com cauda de poli(A) (pelo menos 5 As consecutivos no fim 3' da sequência) de cada *cluster* de cDNA, o sítio NlaIII mais a 3' foi buscado e 10 nt a 3' do sítio foram extraídos como uma etiqueta de SAGE bona-fide produzida por um dado gene (Boon et al (2002)) A etiqueta foi então procurada em todas as bibliotecas de SAGE e sua frequência contada. Etiquetas redundantes (mesma etiqueta para pelo menos dois genes diferentes) foram excluídas. As frequências das etiquetas foram normalizadas pelo tamanho da biblioteca dividindo a frequência da etiqueta pelo número total de etiquetas na biblioteca multiplicado por um fator arbitrário de 200.000. Contagens normalizadas maiores que 500 foram desconsideradas, representando cerca de 0.3 % das contagens.

Comparação dos níveis de expressão gênica entre conjuntos de genes: o nível de expressão global de um grupo de genes para um dado tecido foi tomado como todos os níveis de expressão de todas as bibliotecas disponíveis para o tecido. Para comparar níveis de expressão entre um dado grupo experimental (genes de manutenção celular, baixa ou alta-FRI e todos os outros genes, um procedimento de amostragem de dados semelhante a *bootstrap* foi realizado. Dez mil subconjuntos aleatórios do tamanho do grupo experimental foram gerados a partir do conjunto de níveis de expressão de todos os dados e tiveram seu nível de expressão médio comparado com aquele do conjunto experimental. Se a expressão aleatória fosse diferente de (maior ou menor, dependendo do tecido), o contador era aumentado em 1. Níveis de expressão foram considerados diferentes quando no máximo 5% dos subconjuntos aleatórios tinham nível de expressão maior que (ou menor, dependendo do tecido) ou igual àquele do conjunto experimental.

3.5 Densidade de motivos *reguladores de splicing* em *cis*

Elementos *reguladores* em *cis* putativos e testados experimentalmente obtidos da literatura foram buscados em sequências de exon, de introns retidos e não retidos. Quatro matrizes de pontuação para SF2/ASF, SC35, SRp40 e SRp55 (SELEX-ESEs) e suas pontuações mínimas padrão (Cartegni et al (2003)) foram usadas como amostras de ESEs. Além disso, 238 ESEs putativos de Fairbrother et al (2002) foram buscados juntos como uma categoria única (RESCUE-ESEs). O ESE GAA (Louie et al (2003)), ESSs de Wang et al (2006) (FAS-hex3) e GGG, um ISE conhecido (McCullough e Berget (1997)) também foram buscados. Para normalizar o comprimento, somente os primeiros e últimos 200 nt foram varridos, avançando 1 nt por vez. Para sequências mais curtas que 400 nt, a sequência inteira foi varrida. Todas as ocorrências foram então contadas e somadas nas suas respectivas categorias e divididas pelo número total de nucleotídeos varridos para gerar densidades de motivos (motivos/nt). A avaliação da significância estatística de densidades diferentes em cada conjunto de sequências foi realizada com um procedimento similar a *bootstrap*, como o usado para medições de expressão gênica.

3.6 Conservação de retenção de intron em camundongo

Identidade de introns e exons entre humano/camundongo: cDNAs humanos sem o intron retido foram alinhados com Blast, (Altschul et al (1997), E-valor = 10^{-10}) à sequência genômica de camundongo (montagem 33) obtida do UCSC (Maio, 2004, Karolchik et al (2003)) e posições de intron homólogas foram determinadas. Somente introns que apresentaram bordas (± 10 nt) e comprimentos ($\pm 15\%$) similares em ambos os organismos foram aceitas. Também, o alinhamento tinha que apresentar coordenadas de cDNA contínuas, i.e., os dois exons definindo o intron de camundongo podiam ter um vão (*gap*) máximo de 20 nt no cDNA, para garantir que o intron não era simplesmente o produto de um longo vão no alinhamento. Quando múltiplos alinhamentos existiam, aquele com o menor E-valor foi preferido. Introns encontrados em camundongo foram posteriormente alinhados globalmente com *needle* (Rice et al (2000)) usando parâmetros padrão e tiveram a média de suas identidades percentuais calculadas. O mesmo procedimento foi usado para determinar a identidade de introns não-retidos. A identidade dos exons foi obtida a partir do alinhamento de exons flanqueantes humanos a cDNAs de camundongo.

Conservação dos eventos de RI: blocos de cDNA humanos correspondentes a segmentos “exon + intron retido + exon” foram alinhados a ESTs e RNAs de camundongo. Alinhamentos correspondendo a coordenadas do intron retido mais partes dos exons flanqueantes (pelo menos -20 nt a 5' e +20 nt a 3') foram aceitos como evidência de retenção de intron em camundongo (somente alinhamentos com evidência

de ter sofrido *splicing* foram usadas). Alinhamentos correspondendo a exon + exon nos quais as bordas exon/intron eram similares por ± 10 nt foram tomadas como evidência do intron ser excisado em um dado cDNA de camundongo. O cDNA também tinha que ser contínuo (vão < 20 nt). Sítios de *splice* foram identificados ao alinhar cDNAs de camundongo sem o intron retido ao genoma com SIM4, usando o mesmo procedimento aplicado para eventos de RI humanos.

4 Resultados

O número total de *clusters* ou genes disponíveis em nosso conjunto de dados com pelo menos um RNAm era de 31074, sendo que 18622 continham mais de 2 cDNAs e podiam portanto apresentar alguma forma de *splicing* alternativo. A forma mais abundante era do tipo inclusão alternativa de exon, observada em 14721 genes.

4.1 A força dos sítios de *splice* como causa da retenção de intron

Para entender como a força dos sítios de *splice* controla a RI, analisamos introns retidos identificados em *clusters* de cDNA humanos (um *cluster* idealmente representa um gene e seus transcritos) particionados em duas categorias, de acordo com sua frequência relativa de isoformas (FRI = número de cDNAs com o intron retido/número de cDNAs definindo o intron). O grupo de baixa-FRI continha eventos de RI onde os cDNAs retendo introns eram a forma minoritária (FRI < 1, que corresponde a < 50 % dos cDNAs com RI) e o grupo de alta-FRI continha aqueles eventos onde a RI era a forma majoritária (FRI > 1, > 50% de RI). Como queríamos contrastar características dos grupos de eventos de RI com diferentes FRI, eventos com igual número de evidências definindo e retendo intron não foram consideradas (3% dos casos filtrados).

Os eventos de RI detectados foram filtrados para se obter um conjunto de maior qualidade (ver Métodos). Alguns introns retidos flanqueados por sítios de *splice* não-canônicos foram encontrados no conjunto filtrado (1 no grupo de baixa e 46 no de alta-FRI). Como eles não eram introns G(T/C)..AG e em vários casos diferentes sítios de *splice* não-canônicos podiam ser obtidos por ajustes no alinhamento, eles foram

excluídos do conjunto de dados. O número final de eventos de RI em cada categoria analisada é apresentado na **TABELA 1**.

TABELA 1. Partição de eventos de retenção de introns no transcriptoma humano de acordo com suas frequências relativas de isoforma (FRI).

grupo de FRI	número de eventos	número de <i>clusters</i> de cDNA*	número de <i>clusters</i> c/ > 1 evento da mesma FRI**	número de eventos/ <i>cluster</i>
baixa (< 50% retenção)	1515	1114	282	1.4
alta (> 50% retenção)	250	244	6	1.0

* 53 *clusters* apresentaram tanto eventos de baixa como de alta-FRI (≥ 2 introns retidos diferentes no mesmo gene)

** ≥ 2 introns retidos distintos no mesmo gene, veja Discussão

Para cada grupo, os sítios de *splice* 5' e 3' tiveram seus conteúdos de informação calculados (**FIGURA 26**). A figura mostra que os sítios de *splice* do grupo de baixa-FRI (**FIGURA 26A**) se parece com aqueles que flanqueiam exons constitutivos (**FIGURA 26D**) mas o trato de polipirimidina é mais fraco em comparação ao de exons constitutivos. No grupo de alta-FRI, os sítios de *splice* são menos conservados e apresentam um trato de polipirimidina ainda mais fraco, indicando que a alta frequência de RI é devida a sítios de *splice* mais fracos (**FIGURAS 26B e 26C**). Stamm et al (2000) também observaram que introns retidos apresentam um maior desvio do trato de polipirimidina em comparação a outros tipos de *splicing* alternativo. Também é interessante notar que introns curtos de *Drosophila* não apresentam trato de polipirimidina fortes (Mount et al (1992)) e que sua ausência favorece o modo de definição de intron (Talerico e Berget (1994)).

FIGURA 26 – Logos para sítios de *splice* 5' e 3' de introns retidos, flanqueando exons constitutivos e incluídos alternativamente.

Os logos foram obtidos para os sítios de *splice* flanqueando: (A) Introns retidos em baixa-FRI (1515 sítios), (B) Introns retidos em alta-FRI (250 sítios), (C) Introns retidos em alta-FRI flanqueados por GT..AG (240 sítios), (D) Exons constitutivos (aqueles que não apresentam qualquer evidência de envolvimento em *splicing* alternativo, 90185 sítios), (E) Exons incluídos alternativamente com FRI < 1 (31414 sítios), (F) Exons incluídos alternativamente com FRI > 1 (9567 sítios). A FRI para inclusão alternativa de exon foi calculada de forma análoga à retenção de intron (RIF = número de cDNAs excluindo o exon/cDNAs incluindo o exon).

Reforçando a idéia que sítios de *splice* fracos são uma das causas de RI, há uma fraca correlação negativa entre a força do sítio de *splice* e a frequência de retenção, quando medidas usando o esquema de pontuação de Shapiro e Senapathy (pontuação S&S, Shapiro e Senapathy (1987)) e a FRI em cada grupo (baixa-FRI: $\rho = -0.08$, $P = 0.001$, alta-FRI: $\rho = -0.22$, $P = 0.0008$, **FIGURA 27**, painéis A e B). Os baixos coeficientes de correlação observados sugerem que a FRI é influenciada por vários outros fatores além da força do sítio de *splice*. Além disso, a FRI como calculada aqui é

apenas uma aproximação grosseira, já que não leva em conta a variabilidade de tecidos, normalização de bibliotecas e possível viés de clonagem.

FIGURA 27 – Associação da força do sítio de *splice*, nível de expressão e FRI.

Painéis (A) e (B): Frequência de retenção em função de pontuações Shapiro e Senapathy (doador + aceitador). (C) e (D): Frequência de retenção em função de níveis de expressão gênica (1306 eventos de baixa-FRI e 203 de alta-FRI para os quais havia dados de SAGE). Expressões médias por gene foram usadas. Os gráficos internos mostram os dados transformados em logaritmos e as linhas são as regressões lineares por mínimos quadrados. Coeficientes de correlação: ρ =ordem de rank de Spearman e R =momento linear de Pearson (sobre o dado logarítmico). Note que a regressão linear deve ser tomada como um indicador da tendência e não como modelagem formal dos dados. Uma análise de reamostragem de 1000 conjuntos aleatórios a partir do maior conjunto (baixa-FRI), sendo que cada um dos 1000 conjuntos tinha o tamanho do menor conjunto (alta-FRI), suportou a significância das correlações (Pearson e Spearman) do maior conjunto.

Outra forma de ver a dependência da FRI da força do sítio de *splice* é comparar as frequências dos sítios de *splice* com diferentes pontuações S&S nos dois grupos de FRI (TABELA 2). Como esperado, há uma alta porcentagem de introns de baixa-FRI com altas pontuações S&S (46.8% vs. 34.4% no grupo de alta-FRI e 69.3% de introns

flanqueando exons constitutivos têm pontuação > 160). Entretanto, cerca de metade dos introns retidos em ambos os grupos de FRI apresentaram pontuações S&S na faixa de 140-160 (29.3% dos introns flanqueando exons constitutivos), resultando em chances iguais de serem retidos em baixa ou alta-FRI. Esta observação indica que outros fatores importantes também têm um papel na determinação da frequência de retenção de intron.

TABELA 2. Frequência de introns retidos com sítios de *splice* com diferentes pontuações de Shapiro e Senapathy.

Pontuação S&S (5' ss + 3' ss)	ao redor de exons constitutivos	% de sítios de <i>splice</i>		
		baixa-FRI	alta-FRI	razão baixa/alta-FRI
80-100	0	0	0.4	0
100-119	0.1	0.3	3.6	0.1
120-139	1.4	4.1	8.4	0.5
140-159	29.3	48.8	53.2	0.9
160-179	61.1	43.8	33.2	1.3
180-200	8.2	3.0	1.2	2.5

4.2 A influência do comprimento dos exons flanqueantes e dos introns na retenção de intron

Baseado no estudo de retenção de um intron de 116 nt flanqueado por exons de 30 e de 7 nt (Sternier e Berget (1993)) e observações em introns de *Drosophila*, Talerico e Berget (1994) propuseram que em vertebrados, a retenção ocorreria quando os introns e os exons flanqueantes são pequenos. Nos conjuntos de dados analisados aqui, os dois exons flanqueando o intron retido têm aproximadamente o tamanho médio de exons

internos humanos (~135 nt), com pouquíssimos casos de exons muito curtos (**TABELA 3 e FIGURAS 28A e 28B**). Também, exemplos individuais de RI da literatura envolvem exons com comprimentos regulares (**TABELA 4**). Portanto, a retenção de intron não parece ser relacionada a micro-exons.

TABELA 3. Comprimentos médios dos introns retidos e não-retidos, exons flanqueantes, unidades “exon + intron + exon” e exons internos.

	baixa-FRI	alta-FRI
	média ± d.p. (N)	média ± d.p. (N)
exon a 5' do intron retido ¹	136 ± 131 (1193)	136 ± 99 (98)
exon a 3' do intron retido ¹	127 ± 108 (1193)	138 ± 91 (98)
outros exons não flanqueando um intron retido	134 ± 119 (33,405)	135 ± 134 (7385)
exon+intron retido+exon ²	496 ± 305 (1193)	476 ± 232 (98)
exon+intron não-retido+exon ^{2,3}	2022 ± 10997 (13,486)	2987 ± 10,675 (2275)
intron retido	259 ± 290 (1515)	219 ± 209 (250)
intron retido (EST) ⁴	170 ± 112 (1114)	173 ± 103 (214)
intron retido (RNAm) ⁵	506 ± 447 (401)	504 ± 396 (35)
introns não-retidos ⁶	2005 ± 11,797 (15,710)	3207 ± 12,160 (2758)
introns não-retidos de genes com dados de expressão (SAGE)	2051 ± 12,400 (13,701)	3207 ± 12,136 (2281)

¹ o número de eventos é menor que do conjunto de dados inteiro porque apenas exons flanqueantes pareados que possuíam todas as bordas exon/intron definidas foram usadas para garantir que os exons haviam sido completamente sequenciados. Resultados similares foram obtidos projetando coordenadas exon + exon em unidades “exon + intron retido + exon” com ambas as bordas evidenciadas por introns.

² somente blocos “exon + intron + exon” que tinham bordas exon/intron definidas foram usados

³ os comprimentos médios não são maiores que aquele dos introns apenas porque os primeiros exons, que são tipicamente maiores (Hong et al (2006)), não foram considerados (somente exons internos foram usados)

⁴ somente eventos reportados por ESTs retendo o intron

⁵ somente eventos reportados por RNAs retendo e definindo o intron

⁶ os comprimentos dos introns não-retidos foram calculados para introns definidos por um RNAm (aqueles com mais exons, para aumentar o tamanho do conjunto de dados) por *cluster* (gene)

FIGURA 28 – Comprimentos dos exons flanqueantes, introns retidos e unidades “exon + intron retido + exon”.

Painel (A) Distribuição de comprimentos de exons a 5' flanqueando o intron retido. (B) Distribuição de comprimentos de exons a 3' flanqueando o intron retido. Tamanho dos conjuntos de dados em (A) e (B): 1193 exons com ambas as bordas definidas por introns do grupo de baixa-FRI e 98 de alta-FRI. (C) Distribuição de comprimentos de introns retidos. (D) Distribuição de comprimentos de introns retidos + exons flanqueantes. Não há diferenças estatisticamente significativas.

TABELA 4. Comprimentos de introns retidos e exons flanqueantes da literatura.

exon a 5'	intron retido	exon a 3'	exon + intron + exon	gene	referência
99	118	2219	2436	tgif 2	Melhuish e Wotton (2006)
81	116	469*	666	trombopoietina	Romano et al (2001)
162	274	303*	379	hormônio de crescimento bovino	Dirksen et al (1995)
30	116 ^m	7	153	troponina de frango I	Sterner e Berget (1993)
177	875	75	1127	fator de <i>splicing</i> 9G8	Popielarz et al (1995)
129*	148	132	409	quemoquina CCL27	Ledee et al (2004)
132	205	174*	511	quemoquina CCL27	Ledee et al (2004)
205	129 ^m	238	472	α-globin 2	Carlo et al (1996)
290	118	137	545	calicreína	Michael et al (2005)

notas: exceto onde marcado, os genes estudados são de humano. *indica um exon terminal, ^m o intron retido só foi observado quando o sítio de *splice* 5' era mutado

Em organismos como *Drosophila* (Talerico e Berget (1994)) e *S. pombe* (Romfo et al (2000)), introns mais curtos que os exons são reconhecidos ao invés dos exons e o enfraquecimento dos sítios de *splice* frequentemente resulta em retenção de intron. Em vertebrados, RI frequentemente envolve introns curtos (comprimento médio ~100-200 nt, Stamm et al (2000), Galante et al (2004), Sugnet et al (2004), Zheng et al (2006) e exemplos individuais na **TABELA 4**). Portanto, o comprimento dos introns é provavelmente uma característica importante do mecanismo de RI, em concordância com o modelo de definição de intron. De fato, a **FIGURA 28C** mostra um viés para introns retidos curtos em ambos os grupos de FRI. Entretanto, é difícil avaliar se esta tendência é real usando dados de cDNA, já que ESTs enviesam os dados para introns curtos. Embora Zheng et al (2005) tenham encontrado uma diferença no comprimento de introns retidos em diferentes frequências, a falta de tal diferença em nossos dados pode ser porque ao requerer que cada evento seja representado por pelo menos 2 cDNAs, os eventos identificados foram aqueles com introns curtos em ambos os grupos de FRI que tinham uma alta chance de serem sequenciados múltiplas vezes. O estudo de Ohler et al (2005) não fez uso de ESTs para identificar eventos de RI, empregando somente o grau de conservação da sequência genômica em humano e camundongo como um preditor. Ainda assim, o comprimento médio de 21 introns retidos identificados era curto ($187 \text{ nt} \pm 156$).

Embora este viés para retenção de introns curtos possa também ser um resultado de seleção para inclusão de sequências curtas, outra evidência indireta que suporta uma possível relação mecânica de introns curtos e retenção é a alta frequência de *introns não-retidos* mais curtos em genes com RI. O comprimento médio de introns em nosso conjunto de dados é de 5580 nt (d.p. 17836, 120,949 introns), próximo a observações

anteriores (5025 nt, Eisenberg e Levanon (2003)). Genes que apresentam baixa-FRI possuem introns mais curtos (2005 ± 11797), próximo ao comprimento de genes de manutenção celular (*housekeeping genes*, 2573), de acordo com Eisenberg e Levanon (2003). Introns não-retidos do grupo de eventos de alta-FRI também são mais curtos (3207 ± 12160 , a distribuição de comprimentos é apresentada na **TABELA 5**). Em concordância com esta observação, o comprimento médio de introns não-retidos dos genes com retenção de intron identificados por Ohler et al (2005) sem dados de EST é 3513 ± 8337 nt (166 introns, removendo um gene que possui um comprimento médio de introns anormalmente alto (LACE1, ~19,000 nt) dos 21). É possível que este comprimento de introns mais curto em geral aumente a chance de um gene sofrer RI.

TABELA 5. Distribuição de comprimentos de introns não-retidos nos conjuntos de dados analisados.

Em geral, genes com retenção de intron apresentam uma maior frequência de introns mais curtos.

comprimento	todos os genes (%)	baixa-FRI (%)	alta-FRI (%)	Ohler et al (2005)* (%)
0-100	5.1	14.1	9.5	13.3
100-200	7.2	16.6	12.3	7.8
200-300	4.9	9.7	7.9	3.6
300-400	3.9	6.2	5.7	3.6
400-500	3.6	4.7	4.1	4.8
500-600	3.3	3.8	4.1	2.4
600-700	3.0	3.4	3.1	1.8
700-800	2.7	2.8	2.4	3.6
800-900	2.6	2.4	2.9	1.8
900-1000	2.5	2.3	2.4	1.8
>1000	61.2	34.1	45.5	55.4

*dados de eventos de retenção de intron conservados preditos, Tabela 2 de Ohler et al (2005). Um gene, LACE1 foi excluído por ser o único com uma média de comprimento de intron anormalmente alta (~19,000 nt)

A última característica relacionada a introns curtos analisada foi o comprimento de unidades “exon + intron retido + exon”. Independentemente de um viés do

comprimento para retenção, o fato que muitas destas unidades são mais curtas que 400 nt (**FIGURA 28D**), apenas um pouco acima do limite de comprimento para definição de exon de 300 nt proposto por Berget (1995), sugere que esta observação possa ter uma razão mecânica pelo menos para parte dos eventos como discutido depois.

4.3 Genes com retenção de intron são expressos em mais tecidos e em maiores níveis

Os genes de manutenção celular são definidos como aqueles que são altamente expressos na maioria, senão em todos os tecidos e apresentam uma estrutura compacta (menos introns e mais curtos, Eisenberg e Levanon (2003) e referências citadas). A **FIGURA 29** mostra que existe um viés para genes retendo intron de serem expressos em muitos tecidos em ambos os grupos de FRI, indicando um leve enriquecimento em genes de manutenção celular ou genes amplamente expressos. Além disso, genes de manutenção celular de Eisenberg e Levanon (2003) usados aqui como um controle positivo, apresentaram uma expressão maior que todos os genes considerados juntos em todos os tecidos disponíveis (**TABELA 6**). Genes com RI apresentaram níveis de expressão por tecido menores que de genes de manutenção celular, mas mais altos que todos os genes juntos em 25/31 (baixa-FRI) e 14/31 (alta-FRI) tecidos (**TABELAS 7 e 8**).

A análise de expressão gênica apresentada nesta seção não é uma análise extensiva e aprofundada. Existem métodos bastante sofisticados para comparar diferenças de expressão que levam em conta a variabilidade intra e inter bibliotecas de maneira mais precisa. Além disso, uma análise mais refinada deveria levar em conta o fato que muitos genes são co-expressos. A análise *bootstrap* deveria então sortear

grupos de genes, ao invés de genes individuais. A intenção desta análise foi basicamente verificar tendências de expressão gênica, sendo que o controle positivo representado pelos genes de manutenção celular apresentaram um perfil de expressão gênica compatível com o esperado.

FIGURA 29 – Amplitude de expressão (número de tecidos nos quais um gene é expresso) de genes que retêm introns.

A amplitude foi determinada pela contagem de etiquetas de SAGE para genes com retenção de intron (961 genes de baixa-FRI, 199 de alta-FRI), todos os genes na base de dados usada nesta análise (15953) e 255 de 575 genes de manutenção celular (determinado por Eisenberg e Levanon (2003) com dados de microarray). Genes com retenção de intron são mais frequentemente expressos em um maior número de tecidos em comparação a todos os genes.

TABELA 6. Comparação de níveis de expressão global (unidades arbitrárias) estimadas por SAGE de genes de manutenção celular e todos os genes, por tecido.

A coluna “n” mostra quantas medidas de expressão foram contadas, de forma que cada célula representa o número de genes X o número de bibliotecas no qual eles estão expressos. Comparações por re-amostragem que geraram diferenças em $< 5\%$ dos conjuntos aleatórios ($P < 0.05$) estão marcadas em negrito (diferença estatisticamente significativa, 31/31 tecidos).

Tecido	genes de manutenção celular		todos os genes		P-valor
	n	nível de expressão (u.a.) média $\pm$ d.p.	n	nível de expressão (u.a.) média $\pm$ d.p.	
medula	726	40.2 $\pm$ 67.6	13305	14.6 $\pm$ 31.5	$< 10^{-4}$
osso	160	43.0 $\pm$ 67.2	3370	15.8 $\pm$ 36.5	$< 10^{-4}$
cérebro	13463	34.0 $\pm$ 56.4	332771	12.0 $\pm$ 26.9	$< 10^{-4}$
cartilagem	1853	35.3 $\pm$ 58.3	38533	13.6 $\pm$ 32.4	$< 10^{-4}$
cerebelo	4807	34.4 $\pm$ 58.8	110861	12.5 $\pm$ 27.3	$< 10^{-4}$
cólon	2313	40.8 $\pm$ 70.4	46196	13.6 $\pm$ 30.0	$< 10^{-4}$
olho	222	32.4 $\pm$ 62.5	5395	9.7 $\pm$ 26.9	$< 10^{-4}$
trato gastrointestinal	167	23.1 $\pm$ 36.7	3081	11.4 $\pm$ 23.5	$< 10^{-4}$
coração	191	38.2 $\pm$ 66.3	4476	13.2 $\pm$ 34.5	$< 10^{-4}$
rim	751	40.6 $\pm$ 62.9	14503	15.0 $\pm$ 29.8	$< 10^{-4}$
fígado	793	30.3 $\pm$ 52.8	16638	13.1 $\pm$ 29.6	$< 10^{-4}$
pulmão	1374	37.7 $\pm$ 63.0	29168	14.3 $\pm$ 31.6	$< 10^{-4}$
nódulo linfático	206	24.6 $\pm$ 53.7	6340	9.2 $\pm$ 21.8	$< 10^{-4}$
glândula mamária	8567	34.1 $\pm$ 55.6	181819	14.4 $\pm$ 31.0	$< 10^{-4}$
músculo	348	32.2 $\pm$ 54.7	5883	15.0 $\pm$ 35.6	$< 10^{-4}$
outro	198	25.9 $\pm$ 59.7	4779	10.0 $\pm$ 26.8	$< 10^{-4}$
ovário	1583	42.8 $\pm$ 64.6	31301	16.5 $\pm$ 32.1	$< 10^{-4}$
pâncreas	1547	42.1 $\pm$ 67.7	27007	18.9 $\pm$ 35.2	$< 10^{-4}$
peritônio	387	34.5 $\pm$ 64.8	7428	15.9 $\pm$ 33.4	$< 10^{-4}$
placenta	384	28.5 $\pm$ 57.5	9916	10.4 $\pm$ 26.7	$< 10^{-4}$
próstata	2802	41.2 $\pm$ 66.7	55032	15.2 $\pm$ 32.9	$< 10^{-4}$
retina	634	25.2 $\pm$ 46.9	16620	10.2 $\pm$ 24.1	$< 10^{-4}$
pele	410	38.1 $\pm$ 64.7	7118	19.6 $\pm$ 37.7	$< 10^{-4}$
medula espinhal	197	32.7 $\pm$ 48.5	4791	13.6 $\pm$ 27.2	$< 10^{-4}$
célula tronco	2555	39.0 $\pm$ 67.8	81368	8.7 $\pm$ 25.5	$< 10^{-4}$
estômago	1837	34.2 $\pm$ 57.8	41658	13.4 $\pm$ 29.1	$< 10^{-4}$
tiróide	603	30.7 $\pm$ 61.3	17463	9.7 $\pm$ 25.8	$< 10^{-4}$
tecido não-caracterizado	353	41.9 $\pm$ 60.6	6578	19.1 $\pm$ 36.6	$< 10^{-4}$
útero	131	35.0 $\pm$ 58.9	1695	22.1 $\pm$ 37.0	0.0022
vascular	1187	36.8 $\pm$ 66.2	26377	13.2 $\pm$ 31.6	$< 10^{-4}$
linfócitos	2265	38.3 $\pm$ 66.3	46703	14.2 $\pm$ 31.2	$< 10^{-4}$

TABELA 7. Comparação de níveis de expressão global (unidades arbitrárias) estimadas por SAGE de genes de baixa-FRI e todos os genes, por tecido.

A coluna “n” mostra quantas medidas de expressão foram contadas, de forma que cada célula representa o número de genes X o número de bibliotecas no qual eles estão expressos. Comparações por re-amostragem que geraram diferenças em $< 5\%$ dos conjuntos aleatórios ($P < 0.05$) estão marcadas em negrito (diferença estatisticamente significativa, 25/31 tecidos).

Tecido	n	baixa-FRI	n	todos os genes	P-valor
		nível de expressão (u.a.) média $\pm$ d.p.		nível de expressão (u.a.) média $\pm$ d.p.	
medula	1463	17.9 $\pm$ 38.3	13305	14.6 $\pm$ 31.5	0.0007
osso	413	19.0 $\pm$ 36.9	3370	15.8 $\pm$ 36.5	0.0608
cérebro	30432	14.5 $\pm$ 33.1	332771	12.0 $\pm$ 26.9	$< 10^{-4}$
cartilagem	3871	15.9 $\pm$ 35.4	38533	13.6 $\pm$ 32.4	0.0002
cerebelo	10303	15.6 $\pm$ 35.5	110861	12.5 $\pm$ 27.3	$< 10^{-4}$
cólon	5057	16.3 $\pm$ 35.6	46196	13.6 $\pm$ 30.0	$< 10^{-4}$
olho	516	12.1 $\pm$ 31.9	5395	9.7 $\pm$ 26.9	0.0401
trato gastrointestinal	303	13.4 $\pm$ 28.7	3081	11.4 $\pm$ 23.5	0.111
coração	412	15.2 $\pm$ 33.7	4476	13.2 $\pm$ 34.5	0.1441
rim	1496	17.9 $\pm$ 35.3	14503	15.0 $\pm$ 29.8	0.0011
fígado	1545	15.7 $\pm$ 35.9	16638	13.1 $\pm$ 29.6	0.0014
pulmão	2987	17.0 $\pm$ 37.2	29168	14.3 $\pm$ 31.6	$< 10^{-4}$
nódulo linfático	551	11.5 $\pm$ 27.2	6340	9.2 $\pm$ 21.8	0.0154
glândula mamária	18323	17.7 $\pm$ 36.9	181819	14.4 $\pm$ 31.0	$< 10^{-4}$
músculo	635	14.9 $\pm$ 35.5	5883	15.0 $\pm$ 35.6	0.4899
outro	490	12.4 $\pm$ 25.5	4779	10.0 $\pm$ 26.8	0.0427
ovário	3217	19.4 $\pm$ 38.1	31301	16.5 $\pm$ 32.1	$< 10^{-4}$
pâncreas	3047	22.5 $\pm$ 42.0	27007	18.9 $\pm$ 35.2	$< 10^{-4}$
peritônio	756	20.8 $\pm$ 42.2	7428	15.9 $\pm$ 33.4	0.0005
placenta	817	11.8 $\pm$ 24.2	9916	10.4 $\pm$ 26.7	0.079
próstata	5818	18.1 $\pm$ 37.9	55032	15.2 $\pm$ 32.9	$< 10^{-4}$
retina	1474	13.5 $\pm$ 28.7	16620	10.2 $\pm$ 24.1	$< 10^{-4}$
pele	730	23.7 $\pm$ 47.6	7118	19.6 $\pm$ 37.7	0.0077
medula espinhal	440	14.6 $\pm$ 25.5	4791	13.6 $\pm$ 27.2	0.2378
célula tronco	7102	11.2 $\pm$ 32.6	81368	8.7 $\pm$ 25.5	$< 10^{-4}$
estômago	3858	16.5 $\pm$ 37.5	41658	13.4 $\pm$ 29.1	$< 10^{-4}$
tiróide	1481	11.2 $\pm$ 28.5	17463	9.7 $\pm$ 25.8	0.03
tecido não-caracterizado	658	25.1 $\pm$ 55.3	6578	19.1 $\pm$ 36.6	0.0011
útero	192	31.7 $\pm$ 52.6	1695	22.1 $\pm$ 37.0	0.0043
vascular	2547	15.4 $\pm$ 35.3	26377	13.2 $\pm$ 31.6	0.0012
linfócitos	4970	18.3 $\pm$ 39.6	46703	14.2 $\pm$ 31.2	$< 10^{-4}$

TABELA 8. Comparação de níveis de expressão global (unidades arbitrárias) estimadas por SAGE de genes de alta-FRI e todos os genes, por tecido.

A coluna “n” mostra quantas medidas de expressão foram contadas, de forma que cada célula representa o número de genes X o número de bibliotecas no qual eles estão expressos. Comparações por re-amostragem que geraram diferenças em $< 5\%$ dos conjuntos aleatórios ($P < 0.05$) estão marcadas em negrito (diferença estatisticamente significativa, 14/31 tecidos).

Tecido	n	alta-FRI	n	todos os genes	P-valor
		nível de expressão (u.a.) média $\pm$ d.p.		nível de expressão (u.a.) média $\pm$ d.p.	
medula	300	20.1 $\pm$ 37.3	13305	14.6 $\pm$ 31.5	0.0058
osso	55	22.4 $\pm$ 34.5	3370	15.8 $\pm$ 36.5	0.1039
cérebro	6188	13.3 $\pm$ 29.7	332771	12.0 $\pm$ 26.9	0.0004
cartilagem	762	13.6 $\pm$ 25.9	38533	13.6 $\pm$ 32.4	0.507
cerebelo	2153	16.7 $\pm$ 36.1	110861	12.5 $\pm$ 27.3	$< 10^{-4}$
cólon	923	16.4 $\pm$ 33.4	46196	13.6 $\pm$ 30.0	0.0053
olho	109	10.1 $\pm$ 17.7	5395	9.7 $\pm$ 26.9	0.3627
trato gastrointestinal	67	9.8 $\pm$ 10.6	3081	11.4 $\pm$ 23.5	0.3263
coração	81	12.1 $\pm$ 18.4	4476	13.2 $\pm$ 34.5	0.4383
rim	298	18.4 $\pm$ 33.1	14503	15.0 $\pm$ 29.8	0.0342
fígado	330	14.0 $\pm$ 28.6	16638	13.1 $\pm$ 29.6	0.275
pulmão	538	15.9 $\pm$ 32.8	29168	14.3 $\pm$ 31.6	0.1352
nódulo linfático	107	10.8 $\pm$ 19.9	6340	9.2 $\pm$ 21.8	0.1976
glândula mamária	3561	18.1 $\pm$ 40.5	181819	14.4 $\pm$ 31.0	$< 10^{-4}$
músculo	118	10.7 $\pm$ 15.1	5883	15.0 $\pm$ 35.6	0.0756
outro	97	11.9 $\pm$ 18.0	4779	10.0 $\pm$ 26.8	0.2227
ovário	644	20.4 $\pm$ 41.8	31301	16.5 $\pm$ 32.1	0.0036
pâncreas	549	21.3 $\pm$ 38.4	27007	18.9 $\pm$ 35.2	0.0655
peritônio	145	20.7 $\pm$ 46.1	7428	15.9 $\pm$ 33.4	0.0585
placenta	197	10.4 $\pm$ 16.7	9916	10.4 $\pm$ 26.7	0.4504
próstata	1094	18.5 $\pm$ 42.5	55032	15.2 $\pm$ 32.9	0.0014
retina	308	13.1 $\pm$ 34.9	16620	10.2 $\pm$ 24.1	0.0327
pele	153	18.9 $\pm$ 33.4	7118	19.6 $\pm$ 37.7	0.4424
medula espinhal	89	13.0 $\pm$ 26.3	4791	13.6 $\pm$ 27.2	0.4786
célula tronco	1453	15.1 $\pm$ 38.7	81368	8.7 $\pm$ 25.5	$< 10^{-4}$
estômago	803	13.9 $\pm$ 23.3	41658	13.4 $\pm$ 29.1	0.2995
tiróide	306	12.4 $\pm$ 33.9	17463	9.7 $\pm$ 25.8	0.0466
tecido não-caracterizado	133	25.4 $\pm$ 41.4	6578	19.1 $\pm$ 36.6	0.0391
útero	34	30.0 $\pm$ 69.0	1695	22.1 $\pm$ 37.0	0.1286
vascular	495	16.4 $\pm$ 36.2	26377	13.2 $\pm$ 31.6	0.0251
linfócitos	929	16.0 $\pm$ 32.5	46703	14.2 $\pm$ 31.2	0.047

Também observamos que há uma correlação entre FRI e níveis de expressão gênica (**FIGURA 28**, painéis C e D, baixa-FRI: $\rho = -0.31$, $P < 10^{-4}$, alta-FRI: $\rho = 0.16$, $P = 0.01$). Note que a correlação é negativa para o grupo de baixa-FRI e positiva para o de alta-FRI. Uma explicação é que os níveis mais altos de expressão aumentam a quantidade de todos os transcritos, tanto da forma minoritária quanto a majoritária, mas sob estas circunstâncias, formas muito raras também aparecem. Formas raras no grupo de baixa-FRI seriam aquelas com sítios de *splice* fortes que raramente retêm um intron, resultando no surgimento de eventos com baixa razão de retenção/não-retenção somente com expressão mais alta (correlação negativa). O inverso seria verdadeiro para o grupo de alta-FRI, no qual formas raras seriam aquelas com sítios de *splice* muito fracos que somente raramente removem o intron, e somente com expressão mais alta (correlação positiva).

Tal associação de expressão mais alta e FRI pode ocorrer ou porque uma expressão mais alta aumenta a taxa de erro da maquinaria de *splicing*, especialmente porque parece que *splicing* ocorre co-transcricionalmente (**seção 1.7.2.3**), ou simplesmente por permitir a detecção de RI (ou remoção de intron, no caso do grupo de alta-FRI), já que mais transcritos são produzidos. Esta última explicação é atraente porque verificamos que *clusters* que apresentam retenção de intron possuem mais cDNAs que todos os outros genes (por exemplo, ~17% dos *clusters* com retenção de intron e 6% de todos os outros *clusters*, possuem >400 cDNAs, teste χ^2 , 1 g.l., $P < 10^{-10}$, veja distribuição do tamanho de *clusters* na **FIGURA 30**).

FIGURA 30. Distribuição do número de cDNAs em *clusters* de baixa-FRI, alta-FRI e todos os *clusters*.

(A) *Clusters* com retenção de intron são maiores. (B) A diferença não é devida a uma escassez de *clusters* pequenos com RI, mas também existe mesmo considerando apenas *clusters* com >100 cDNAs.

4.4 Distribuição e densidade de elementos reguladores de *splicing* em *cis*

Elementos reguladores em *cis* (ESE, ESS, ISE e ISS), têm um importante papel no *splicing* constitutivo e alternativo, atuando como sítios de ligação para fatores em *trans* que promovem (ESE, ISE) ou inibem (ESS, ISS) o *splicing* (seção 1.7.3). Por exemplo, um ESE GAA é crucial para contrabalançar retenção do intron D do gene de hormônio de crescimento bovino (Dirksen et al (1995)) e do intron 3 do gene do fator de *splicing* 9G8 como mostrado por Lejeune et al (2001).

Como nem todos os motivos correspondendo a elementos reguladores em *cis* encontrados em exons ou introns são funcionais, uma hipótese é que eles estão presentes em densidades mais altas (motivos/nt) nas sequências em que eles atuam (Wu et al

(2005)). Recentemente, Wang et al (2006) mostraram que um conjunto particular de ESSs que promovem *splicing* de introns retidos está enriquecido nessas sequências.

Para avaliar se as diferentes FRI estão associadas a diferentes densidades de elementos reguladores em *cis*, varremos introns retidos e não-retidos, bem como exons dos mesmos genes, em busca de motivos reguladores conhecidos (TABELA 9). As densidades médias de SELEX-ESEs em exons e de FAS-hex3 ESSs em introns retidos e não-retidos mostradas na TABELA 9 estão na faixa observada em trabalhos anteriores (Wu et al (2005) e Wang et al (2006), respectivamente).

TABELA 9. Densidades de motivos reguladores em *cis* putativos em exons, introns retidos e não-retidos de genes dos grupos de baixa e alta-FRI.

	motivo regulador	introns			exons	
		retidos	não-retidos	flanqueando a 5'	flanqueando a 3'	todos os outros
Baixa	SF2/ASF	<i>0.0415⁺ⁱ</i>	<i>0.0365</i>	0.0458	0.0451	0.0430
	SC35	<i>0.0457⁺ⁱ</i>	<i>0.0413</i>	0.0445	0.0450	0.0425
	SRp40	0.0388	<i>0.0384</i>	0.0399	0.0395	0.0394
	SRp55	<i>0.0219</i>	<i>0.0220</i>	0.0272	0.0264	0.0261
	RESCUE-ESEs	<i>0.0442⁻ⁱ</i>	<i>0.0541</i>	0.0837	0.0850	0.0951
	GAA (ESE)	<i>0.0115⁻ⁱ</i>	<i>0.0142</i>	0.0191	0.0197	0.0219
	GGG (ISE)	0.0397⁺ⁱ	0.0333	0.0185	0.0190	0.0165
	FAS-ESS hex-3 classe 1	0.0047⁻ⁱ	0.0069	0.0010	0.0009	0.0011
	FAS-ESS hex-3 classe 2	0.0294⁺ⁱ	0.0276	0.0110	0.0115	0.0101
Alta	SF2/ASF	<i>0.0383⁺ⁱ</i>	<i>0.0345</i>	0.0445	0.0458	0.0423
	SC35	0.0409	<i>0.0387</i>	0.0447	0.0412	0.0428
	SRp40	<i>0.0394⁺ⁱ</i>	<i>0.0371</i>	0.0404	0.0397	0.0408
	SRp55	0.0231	<i>0.0219</i>	0.0256	0.0241	0.0250
	RESCUE-ESEs	<i>0.0548</i>	<i>0.0591</i>	0.0952	0.0971	0.0974
	GAA (ESE)	<i>0.0137⁻ⁱ</i>	<i>0.0156</i>	0.0227	0.0206	0.0224
	GGG (ISE)	0.0264⁻ⁱ	0.0300	0.0164	0.0173	0.0165
	FAS-ESS hex-3 classe 1	0.0043⁻ⁱ	0.0074	0.0017	0.0007	0.0013
	FAS-ESS hex-3 classe 2	0.0208⁻ⁱ	0.0253	0.0095	0.0106	0.0098

notas: as densidades médias em *itálico* são menores que em exons ($P \leq 0.05$), densidades médias em **negrito** são maiores que em exon ($P \leq 0.05$). O sinal ⁺ⁱ indica que a densidade do motivo em introns retidos é significativamente maior ($P \leq 0.05$) que em introns não-retidos, ⁻ⁱ indica que é menor.

Em comparação a exons, introns não-retidos são pobres em ESEs, coerentemente com um estudo prévio (Wu et al (2005)), e enriquecidos em ESSs e com o ISE GGG (**TABELA 9**).

Introns retidos também são diferentes de exons, mas com densidades de SELEX-ESEs um pouco mais altas, intermediárias entre exons e introns não-retidos. Esta tendência não foi observada para RESCUE-ESEs provavelmente porque esses motivos foram identificados selecionando hexâmeros especificamente super-representados em exons (Fairbrother et al (2002)). Ou seja, esses ESEs são provavelmente enviesados para a composição nucleotídica de exons e portanto sua presença é de alguma forma dependente do tipo de sequência sendo considerada, em comparação com SELEX-ESEs que seriam mais neutros, variando mais.

As diferenças mais marcantes entre os grupos de baixa e alta-FRI foram aquelas relacionadas à densidade de GGG e ESSs. Introns retidos em baixa-FRI apresentaram uma densidade de GGG um pouco maior que introns não-retidos, enquanto os introns retidos em alta-FRI apresentaram a tendência oposta (**TABELA 9 e FIGURA 31A**). Em relação a ESSs FAS-hex3, Wang et al (2006) mostraram que os ESSs classe 1 são na verdade sub-representados em introns retidos em relação a introns não-retidos, o que foi também observado em nossos dados para ambos os grupos de FRI. Os FAS-ESSs de classe 2, entretanto, seguiram a mesma tendência de GGG, apresentando densidades mais altas no grupo de baixa-FRI do que em introns não-retidos e densidades mais baixas no grupo de alta-FRI (**TABELA 9 e FIGURA 31B**). Esta diferença entre as duas classes de ESS é provavelmente relacionada ao fato que os de classe 2 são ricos em G (5/26 ESSs de classe 1 e 28/46 ESSs de classe 2 possuem pelo menos um triplete de GGG).

FIGURA 31 – Densidades de GGG/FAS-ESS classe 2 em introns retidos e não-retidos e variação com a força de sítio de *splice*.

Painéis (A) e (B): Distribuição da frequência de introns retidos e não-retidos com densidades variáveis de GGG (A) e de FAS-ESSs classe 2 (B). O grupo de baixa-FRI apresenta um viés para introns com alta densidade de GGG, enquanto o grupo de alta-FRI apresenta a tendência oposta. Painéis (C) e (D): Densidades médias de GGG (C) e de FAS-ESSs classe 2 (D) em introns não-retidos e retidos flanqueados por sítios de *splice* com diferentes pontuações S&S. Apenas classes de pontuações S&S com mais de 5 sequências foram usadas.

Como RI resulta em um longo exon único, que é a isoforma mais comum no grupo de alta-FRI, avaliamos a possibilidade que tal distribuição diferencial de motivos entre exons flanqueantes e o intron retido fosse na verdade intrínseca a exons longos divididos em dois segmentos flanqueantes e um segmento central (pseudo-intron retido), e não a introns retidos. Analisando exons de comprimento entre 300-600 nt e >600 nt, observou-se que essas sequências não apresentam o padrão observado para

unidades “exon + intron + exon” (TABELAS 10 e 11). O segmento central que corresponde ao pseudo-intron retido possui uma densidade de elementos reguladores muito similar à de exons, garantindo que as diferenças observadas em introns retidos são intrínsecas a essas sequências.

TABELA 10. Densidade de elementos reguladores em *cis* em 3207 exons longos (> 300 nt and < 600 nt) divididos em três segmentos.

Elemento regulador	Segmento 5'	Segmento central	Segmento 3'
SF2/ASF	0.0450	0.0461 ³	0.0443
SC35	0.0440	0.0447 ³	0.0431
SRp40	0.0424	0.0413 ⁵	0.0417
SRp55	0.0250	0.0250	0.0257
RESCUE-ESEs	0.0924	0.0870 ^{5,3}	0.0950
GAA (ESE)	0.0206	0.0191 ^{5,3}	0.0212
GGG (ISE)	0.0173	0.0200 ^{5,3}	0.0175
FAZ-ESS hex-3 classe 1	0.0015	0.0016	0.0015
FAZ-ESS hex-3 classe 2	0.0107	0.0116	0.0104

⁵ $P \leq 0.05$ em relação ao segmento 5'

³ $P \leq 0.05$ em relação ao segmento 3'

TABELA 11. Densidade de elementos reguladores em *cis* em 2570 exons longos (> 600 nt) divididos em três segmentos.

Elemento regulador	Segmento 5'	Segmento central	Segmento 3'
SF2/ASF	0.0427	0.0424	0.0425
SC35	0.0427	0.0427	0.0420
SRp40	0.0428	0.0421	0.0423
SRp55	0.0250	0.0245	0.0244
RESCUE-ESEs	0.0968	0.0890 ^{5,3}	0.0929 ⁵
GAA (ESE)	0.0219	0.0201 ^{5,3}	0.0210 ⁵
GGG (ISE)	0.0175	0.0197 ⁵	0.0187 ⁵
FAS-ESS hex-3 classe 1	0.0028	0.0034 ⁵	0.0032 ⁵
FAS-ESS hex-3 classe 2	0.0118	0.0135 ⁵	0.0130 ⁵

⁵ $P \leq 0.05$ em relação ao segmento 5'

³ $P \leq 0.05$ em relação ao segmento 3'

As densidades de ESS/GGG observadas poderiam sugerir que elementos reguladores em *cis* possuem um papel na regulação diferencial de RI nos dois grupos de FRI. Uma densidade mais alta de ESS/GGG em introns retidos poderia aumentar a eficiência do *splicing* do intron (assim, FRIs mais baixas) enquanto que para introns de alta-FRI ocorreria o inverso. A base para este argumento é que a mutação de diferentes

tripletes de GGG no segundo intron de α -globina 2 resulta em diferentes níveis de retenção de intron (McCullough e Berget (1997)) e que a inserção de ESSs no intron 3 do gene de NKIRAS2, que é frequentemente retido, aumentou os níveis da forma sem o intron (Wang et al (2006)). Isto explicaria porque introns com pontuação S&S na faixa de 140-160 podem ser retidos tanto em baixa quanto em alta FRI (veja a distribuição de pontuações S&S na **TABELA 2** e a densidade de GGG/ESS em função de pontuações S&S nas **FIGURAS 31C e 31D**).

A interpretação de densidades de SF2/ASF e SC35 ligeiramente maiores observadas em introns retidos não é simples. Uma possibilidade é que estes ESEs poderiam ajudar a definir unidades “exon + intron retido + exon” curtas como um exon único (veja Discussão), promovendo a retenção. A outra possibilidade é uma ação no reconhecimento de sítios de *splice*. Embora ESEs sejam exônicos por definição, foi mostrado que a ligação de proteínas SR a introns pode tanto aumentar como diminuir *splicing* em diferentes estudos (ver discussão em Ibrahim et al (2005), Wu et al (2005)). Se estes ESEs aumentam o *splicing* do intron nos eventos analisados aqui, eles poderiam contrabalançar a retenção do intron. Ambos os casos poderiam ocorrer em diferentes eventos de retenção de intron, mas o fato que as densidades no grupo de baixa-FRI para SF2/ASF e SC35 são maiores que em alta-FRI poderia indicar que seus papel é de contrabalancear a retenção de intron. Entretanto, os sinais são sutis demais para permitir a obtenção de conclusões definitivas.

Poderia-se argumentar que tanto a densidade mais alta de SELEX-ESEs e mais baixa de ESS/GGG no conjunto de alta-FRI poderia ser devido a seleção em sequências que são ocasionalmente exônicas, mas esta possibilidade é posta em dúvida frente aos dados de RESCUE-ESE e de FAS-ESS classe 1, cujas densidades são similares à de introns não-retidos.

4.5 Conservação da frequência de retenção de intron em camundongo

A conservação evolutiva de certas características é frequentemente tida como um sinal de função biológica sob a suposição que elas estão sob seleção negativa. O estudo de eventos de RI conservados poderia suportar as observações feitas com dados humanos e portanto ajudar a entender o mecanismo de retenção de intron. Como o critério que usamos para selecionar eventos de RI de alta confiança pode ter filtrado eventos verdadeiros, nesta análise usamos todos os eventos de RI inicialmente detectados para vasculhar o genoma e transcriptoma de camundongo. Com exceção do número de eventos conservados, todos os outros dados (identidades e porcentagens) foram os mesmos quando buscamos somente eventos humanos filtrados.

A **TABELA 12** (linha 1) mostra que introns retidos humanos do grupo de alta-FRI eram quase tão conservados quanto seus exons flanqueantes ($81\% \pm 9$ vs. $91\% \pm 3$) enquanto introns retidos em baixa-FRI em humano apresentaram uma identidade média correspondente àquela de introns não-retidos homólogos ($59\% \pm 10$ vs. $60\% \pm 10$).

TABELA 12. Conservação de eventos de retenção de introns humanos em camundongo.

	baixa-FRI	alta-FRI
1. % de identidade média (introns retidos humanos versus genoma de camundongo) ¹	59% ± 10 (1582 introns)	81% ± 9 (408 introns)
2. % de identidade média (exons)	91% ± 3	91% ± 3
3. número de eventos apresentando somente a forma sem o intron retido	1252/1297 ³ (97%)	16/201 ³ (8%)
4. número de eventos apresentando somente a forma contendo o intron retido	16/1297 ³ (1%)	177/201 ³ (88%)
5. número de eventos conservados ²	29/1297 ³ (2%)	8/201 ³ (4%)
6. número de eventos conservados com FRI conservada	22/29	7/8
7. % de identidade média de introns retidos conservados com FRI conservada	86% ± 10 (22 introns retidos)	91% ± 9 (7 introns retidos)
8. % de identidade média de introns retidos não-conservados ²	59% ± 9 (sem retenção em camundongo; 1252 introns)	82% ± 9 (sem a forma sem-retenção em camundongo; 177 introns)
9. comprimento médio de introns retidos conservados	152 ± 83 (22 introns)	142 ± 87 (7 introns)
10. conteúdo de informação dos sítios de <i>splice</i> ⁴	7.2 bits ± 0.2/9.7 bits ± 0.4 (22 introns, todos GT..AG)	5.4 bits ± 1.3/7.4 bits ± 1.3 (7 introns, todos GT..AG)

¹ somente introns com comprimentos e bordas similares em humano e camundongo foram considerados

² conservado em camundongo se o evento apresentava ≥ 1 cDNA com o intron retido e ≥ 1 sem. Não-conservado: somente a forma majoritária foi encontrada em cDNAs de camundongo

³ o número total de eventos (1297 e 201) corresponde somente àqueles eventos da linha 1 com pelo menos um cDNA de camundongo alinhado à região genômica correspondendo ao evento

⁴ com exceção do alto conteúdo de informação do sítio de *splice* 3' do grupo de baixa-FRI, os valores são similares aos observados para humano

A maioria dos eventos de RI possuíam apenas a forma majoritária em camundongo (TABELA 12, linhas 3 e 4), sendo considerados não-conservados. Os eventos conservados (formas minoritária e majoritária também encontradas em camundongo) totalizaram 29 no grupo de baixa-FRI e 8 no de alta-FRI. Entre esses eventos, 22/29 no grupo de baixa-FRI e 7/8 no grupo de alta-FRI apresentaram FRIs

conservadas. A pequena porcentagem de eventos conservados é provavelmente devida à falta de dados de cDNA cobrindo a região homóloga, e não falta de conservação.

De acordo com nossa suposição, 8/22 e 2/7 dos eventos humanos de baixa e alta-FRI, respectivamente, haviam sido filtrados do conjunto. Isto indica que nosso filtro foi rigoroso, removendo possíveis eventos verdadeiros, e que considerar todos os eventos na análise de camundongo foi adequado.

Interessantemente, introns retidos conservados apresentaram alta identidade, independente do grupo de FRI (TABELA 12, linha 7), próxima daquela de exons em geral, enquanto os eventos de RI humanos do grupo de baixa-FRI que apresentaram somente a forma majoritária em camundongo apresentaram a mesma identidade percentual de introns não-retidos (TABELA 12, linha 8).

Como observado em humano há uma diferença no conteúdo de informação de sítios de *splice*, sendo que aqueles do grupo de baixa-FRI apresentam maior conservação (TABELA 12, última linha). Independentemente da frequência de retenção, os introns retidos em camundongo eram consideravelmente mais curtos que a média de introns não-retidos. Não são fornecidos dados para a distribuição e densidade de elementos reguladores em *cis* para eventos conservados em camundongo devido ao tamanho reduzido do conjunto de dados. Vale a pena notar, entretanto, que a tendência da densidade de GGG/ESS em eventos conservados era consistente com aquela de humanos, *i.e.*, notavelmente maior em introns retidos em baixa-FRI.

5 Discussão

O principal problema de analisar RI é garantir que os eventos são ocorrências autênticas de *splicing* alternativo e não parte de mensagens parcialmente processadas. Entretanto, o fato que encontramos associações mecanisticamente coerentes da força do sítio de *splice* e densidades diferenciadas de motivos de elementos reguladores em *cis* com os dois grupos de FRI (corroborado pelos eventos conservados) indica que o nosso conjunto de dados contém uma grande quantidade de eventos de RI autênticos.

Em concordância com observações experimentais anteriores de introns retidos vertebrados (Sternier e Berget (1993), Dirksen et al (1995), Lejeune et al (2001), Romano et al (2001)) e com análises bioinformáticas (Stamm et al (2000), Zheng et al (2005)), verificamos que RI está associada a sítios de *splice* mais fracos. Estendemos esta observação geral mostrando que uma diminuição da força do sítio de *splice* leva a maiores frequências relativas de retenção. Entretanto, uma fração não-desprezível dos eventos possuem sítios de *splice* fortes. Estendendo observações anteriores que um conjunto específico de ESSs era capaz de inibir a retenção de intron (Ohler et al (2005)), mostramos que há uma associação inversa da densidade destes elementos reguladores (e também do ISE GGG) com a frequência de RI. O fato que uma fração considerável de introns retidos em baixa-FRI apresenta sítios de *splice* fortes e uma alta densidade de GGG/ESS indica que outros fatores mais complexos do que os estudados aqui também estão envolvidos na regulação deste tipo de *splicing* alternativo, como por exemplo a concentração relativa de fatores em *trans*, a formação de estruturas secundárias na molécula de RNA ou a dependência de *splicing* de introns adjacentes para a retenção de um intron específico, como observado para o gene de trombopoietina (Romano et al (2001)).

Embora dados de cDNA como os aqui apresentados sejam possivelmente enviesados para conter introns retidos curtos, outras observações não diretamente baseadas em dados de cDNA suportam a idéia que tais introns são curtos (Stamm et al (2000), Ohler et al (2005)). Junto com a nossa observação que os genes com RI apresentam introns mais curtos em geral, estes dados sugerem que isto possa ter razões mecánísticas para pelo menos parte dos eventos. Uma possibilidade atraente é que esses introns retidos curtos sejam parte de unidades maiores (até cerca de 300 nt, o limite para definição de exon de Berget (1995)) que são por vezes reconhecidas como exons longos únicos e por vezes como estruturas divididas em “exon + intron + exon”. De acordo com esta idéia, os exons flanqueantes parecem ter o mesmo comprimento de exons em geral. Além disso, os introns retidos (ou segmentos intra-exônicos) contêm características intrônicas necessárias para serem ocasionalmente removidos, tais como sítios de *splice* crípticos (fracos) e promotores de *splicing* como tripletes GGG que promovem a remoção de intron através da definição de intron (McCullough e Berget (1997)) e ESSs. De acordo com nossos dados, neste modelo, a frequência relativa de retenção seria regulada pela força dos sítios de *splice* e pela densidade de GGG/ESS.

Baseado neste cenário, propomos um mecanismo (A na **FIGURA 32**) como parte de um modelo bipartido para explicar como introns são retidos e removidos. O Modelo A também foi proposto por Sterner e Berget (1993) ao analisarem *in vivo* a retenção em uma sequência “exon + intron + exon” de troponina I de galinha, menor que 300 nt com o sítio de *splice* 5’ mutado. Mais tarde, Talerico e Berget (1994) sugeriram que este seria o mecanismo geral de retenção de intron em vertebrados.

FIGURA 32 – Modelo bipartido de retenção de intron.

(A) “Reconhecimento de sítio de *splice* em dois passos” (um modelo misto de reconhecimento de exon e intron) para introns curtos em unidades “exon + intron + exon” curtas (< 400 nt, próximo do limite de reconhecimento de exon). No passo 1, todos os exons são definidos, incluindo a unidade “exon + intron + exon” em preto e cinza. O passo 2 é alternativo, se a definição de intron ocorre na unidade “exon + intron + exon”, o intron será removido. A não-execução deste passo levará à retenção de intron.

(B) Modelo para introns longos, tipicamente com sítios de *splice* de alta pontuação em unidades “exon + intron + exon” longas. Passo 1: os exons são definidos, incluindo aqueles flanqueando o intron a ser retido (exons em cinza). No passo 2, os introns flanqueantes são removidos, juntando os exons que flanqueiam o intron retido aos exons externos. O passo 3 é alternativo, se os exons flanqueantes são unidos, o intron é removido; do contrário, a não-execução deste passo por falha hipotética na transição dos complexos E para A ou A para B levariam à retenção de intron.

O mecanismo no modelo A poderia estar relacionado ao fato que o manejo de exons longos é complicado para os spliceossomos vertebrados (Stern et al (1996)) e é tentador especular que o modelo poderia ser estendido para unidades “exon + intron retido + exon” maiores que 300-400 nt. Foi observado que um exon expandido de 566 nt não podia ser corretamente processado, mas sofria *splicing* interno, criando três novos exons pequenos de 63, 90 e 73 nt e dois introns mais curtos que 90 nt cada. Ao expandir para 894 nt, dois exons de 63 e 73 nucleotídeos e um intron foram criados (Stern et al (1996)). Estas observações suportam a idéia proposta no modelo A que introns retidos usariam sítios de *splice* críticos internos que seriam reconhecidos durante a definição de exon de um exon longo.

Entretanto, também verificamos que existem casos em que os introns retidos são muito mais longos que aqueles cuja definição é conhecida e não caberiam nas unidades “exon + intron + exon” relativamente curtas mencionadas acima. Para estes casos, propomos um modelo alternativo especulativo (mecanismo B na **FIGURA 32** que não envolve definição de intron. Este mecanismo implica que os introns flanqueando o intron retido seriam removidos primeiro, *i.e.*, a ordem de remoção dos introns pode ditar o padrão de *splicing* resultante, o que já foi observado no caso de inclusão alternativa de exon (Takahara et al (2002)). Ocasionalmente, a transição do complexo E para A ou mesmo de A para B poderia sofrer algum tipo de interrupção devido à fraqueza dos sítios de *splice*, impedindo a ocorrência de passos posteriores e retendo o intron.

A observação experimental da existência de diferentes complexos *spliceossomais* ocorre porque há gargalos cinéticos na montagem do spliceossomo vulneráveis a interrupções. Ou seja, praticamente todas as transições entre complexos, seriam em princípio, pontos potenciais para regulação de *splicing* alternativo (Matlin et al (2005)).

O modelo de retenção de introns longos (modelo B) na **FIGURA 32** baseia-se na idéia que a seleção dos sítios de *splice* só se dá na formação do complexo A (Lim e Hertel (2004)) ou mais tardiamente no complexo B (Chiara e Reed (1995)). A idéia é que os exons que flanqueiam o intron retido são definidos para que o *splicing* dos introns externos proceda normalmente. Assim, os sítios de *splice* flanqueando o intron retido devem ter pelo menos o complexo E montado com sucesso para engajar os exons na via de *splicing*, mas isso não implica que os sítios de *splice* já estejam engajados, ou seja, a excisão do intron “retido” não é compulsória. Ou seja, como o engajamento dos sítios de *splice* exatos só se dá na formação do complexo A ou ainda mais tardiamente no complexo B, é possível que mesmo após a excisão dos introns externos os sítios de *splice* do intron retido não sejam selecionados, interrompendo sua excisão. Isso talvez ocorra por problemas causados durante os rearranjos dos complexos devido à menor conservação dos sítios de *splice* ou pela menor densidade de ISEs que auxiliariam no reconhecimento dos sítios de *splice*. A alta ou baixa frequência com que essa interrupção ocorre deve ser influenciada pela força do sítio bem como por outros fatores como presença de proteínas SR/hnRNPs, estrutura secundária do RNA e outros.

Deve-se notar que como apresentado na **Introdução** desta tese, ainda se está muito longe de conhecer todos os detalhes da regulação do *splicing* e da geração de formas alternativas a partir de alterações na montagem do spliceossomo. Dessa forma, o mecanismo proposto é bastante especulativo. Muitos estudos ainda serão necessários para esclarecer se a retenção de introns longos realmente se dá pelo modelo sugerido e a forma que isto ocorre.

Os níveis de expressão gênica não seriam uma parte intrínseca do modelo proposto acima, mas podem contribuir para até mesmo abolir a retenção de intron em tecidos onde a expressão seja baixa. A ocorrência de mais de um evento por gene,

especialmente no grupo de baixa-FRI pode estar relacionada ao fato que estes genes apresentam em geral introns mais curtos, aumentando a probabilidade que um intron seja retido através do modelo A. Com expressão gênica mais alta, a retenção de introns com sítios de *splice* moderadamente fortes seria facilitada, ainda que em baixa frequência. Para os introns serem retidos em alta frequência, os sítios de *splice* teriam que ser consideravelmente mais fracos, o que devido à seleção são encontrados na minoria dos introns, e portanto a chance de ocorrer múltiplos eventos no mesmo gene é provavelmente menor.

6 Conclusão

Se você construir, ele virá
do filme “Campo dos sonhos”

O trabalho apresentado consistiu de uma análise puramente bioinformática sobre o fenômeno de *splicing* alternativo, um fenômeno de grande interesse biológico, atualmente. Embora o estudo tenha um caráter exploratório, na medida em que é uma busca por associações de características com a ocorrência de retenção de intron, a realização de cada análise envolveu a construção de hipóteses específicas para determinar que características seriam analisadas, com base no que se conhece sobre *splicing* alternativo. Um exemplo é a escolha dos elementos reguladores em *cis* a serem analisados ou a decisão de se estudar a expressão gênica. Além disso, as análises foram sempre dirigidas por hipóteses sobre possíveis mecanismos de retenção de intron.

Para a execução deste trabalho foi construída uma ferramenta computacional de anotação de *splicing* alternativo no transcriptoma humano representado por sequências expressas armazenadas em bancos de dados. A ferramenta permitiu a execução de várias análises complexas e atingiu o objetivo inicial de reutilização de código e uso inteligente dos dados primários através da construção de um banco secundário.

Com esta ferramenta, o trabalho de pesquisa realizado mostrou que características das sequências de RNA (sítios de *splice* mais fracos, densidade de elementos reguladores de *splicing* particulares) e dos genes (maior expressão, introns mais curtos) estão associadas a retenção de intron e devem ter alguma relação causal com o fenômeno. Embora outros fatores complexos desconhecidos não-relacionados diretamente à sequência de RNA parecem influenciar grandemente a ocorrência do

fenômeno, as características de sequência reveladas podem ter valor preditivo da ocorrência dos eventos. Além disso, a descoberta de novas características associadas à retenção de intron e a forma que elas podem atuar não só para gerar, mas para regular a frequência em que o fenômeno ocorre, contribui para o entendimento do fenômeno de *splicing* alternativo.

7 Referências

1. Abovich N, Liao XC, Rosbash M (1994) The yeast MUD2 protein: an interaction with PRP11 defines a bridge between commitment complexes and U2 snRNP addition. *Genes Dev* 8: 843-854.
2. Abovich N, Rosbash M (1997) Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. *Cell* 89: 403-412.
3. Aebi M, Hornig H, Weissmann C (1987) 5' cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5' splice region, not by the conserved 5' GU. *Cell* 50: 237-246.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25-29.
6. Ast G (2004) How did alternative splicing evolve? *Nat Rev Genet* 5: 773-782.
7. Azubel M, Habib N, Sperling R, Sperling J (2006) Native spliceosomes assemble with pre-mRNA to form supraspliceosomes. *J Mol Biol* 356: 955-966.
8. Azubel M, Wolf SG, Sperling J, Sperling R (2004) Three-dimensional structure of the native spliceosome by cryo-electron microscopy. *Mol Cell* 15: 833-839.
9. Baek D, Green P (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc.Natl.Acad.Sci.U.S.A* 102: 12813-12818.
10. Barabino SM, Blencowe BJ, Ryder U, Sproat BS, Lamond AI (1990) Targeted snRNP depletion reveals an additional role for mammalian U1 snRNP in spliceosome assembly. *Cell* 63: 293-302.
11. Bateman A, Coin L, Durbin R, Finn RD, Hollich V et al (2004) The Pfam protein families database. *Nucleic Acids Res* 32: D138-D141.
12. Baurén G, Wieslander L (1994) Splicing of Balbiani ring 1 gene pre-mRNA occurs simultaneously with transcription. *Cell* 76: 183-192.
13. Berget SM (1995) Exon recognition in vertebrate splicing. *J.Biol.Chem.* 270: 2411-2414.
14. Black DL (1998) Splicing in the inner ear: a familiar tune, but what are the instruments? *Neuron* 20: 165-168.
15. Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72: 291-336.
16. Blencowe BJ (2006) Alternative splicing: new insights from global analyses. *Cell* 126: 37-47.
17. Boise LH, González-García M, Postema CE, Ding L, Lindsten T et al (1993) bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell* 74: 597-608.
18. Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J et al (2002) An anatomy of normal and malignant gene expression. *Proc.Natl.Acad.Sci.U.S.A* 99: 11287-11292.

19. Boukis LA, Liu N, Furuyama S, Bruzik JP (2004) Ser/Arg-rich protein-mediated communication between U1 and U2 small nuclear ribonucleoprotein particles. *J Biol Chem* 279: 29647-29653.
20. Brett D, Pospisil H, Valcárcel J, Reich J, Bork P (2002) Alternative splicing and genome complexity. *Nat Genet* 30: 29-30.
21. Brow DA (2002) Allosteric cascade of spliceosome activation. *Annu Rev Genet* 36: 333-360.
22. Brudno M, Gelfand MS, Spengler S, Zorn M, Dubchak I et al (2001) Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucleic Acids Res* 29: 2338-2348.
23. Bruzik JP, Maniatis T (1995) Enhancer-dependent interaction between 5' and 3' splice sites in trans. *Proc Natl Acad Sci U S A* 92: 7056-7059.
24. Burge CB, Tuschli T, Sharp PA (1999) Splicing of precursors to mRNAs by the spliceosomes. 2nd Ed.: 525-560.
25. Butcher SE, Brow DA (2005) Towards understanding the catalytic core structure of the spliceosome. *Biochem Soc Trans* 33: 447-449.
26. Carlo T, Sterner DA, Berget SM (1996) An intron splicing enhancer containing a G-rich repeat facilitates inclusion of a vertebrate micro-exon. *RNA*. 2: 342-353.
27. Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat.Rev.Genet.* 3: 285-298.
28. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* 31: 3568-3571.
29. Chen FC, Chen CJ, Ho JY, Chuang TJ (2006) Identification and evolutionary analysis of novel exons and alternative splicing events using cross-species EST-to-genome comparisons in human, mouse and rat. *BMC Bioinformatics* 7: 136.
30. Chiara MD, Reed R (1995) A two-step mechanism for 5' and 3' splice-site pairing. *Nature* 375: 510-513.
31. Clark F, Thanaraj TA (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum Mol Genet* 11: 451-464.
32. Consortium MGS, Waterston RH, Lindblad-Toh K, Birney E, Rogers J et al (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520-562.
33. Côté J, Beaudoin J, Tacke R, Chabot B (1995) The U1 small nuclear ribonucleoprotein/5' splice site interaction affects U2AF65 binding to the downstream 3' splice site. *J Biol Chem* 270: 4031-4036.
34. Cramer P, Cáceres JF, Cazalla D, Kadener S, Muro AF et al (1999) Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. *Mol Cell* 4: 251-258.
35. Crispino JD, Blencowe BJ, Sharp PA (1994) Complementation by SR proteins of pre-mRNA splicing reactions depleted of U1 snRNP. *Science* 265: 1866-1869.
36. Crispino JD, Sharp PA (1995) A U6 snRNA:pre-mRNA interaction can be rate-limiting for U1-independent splicing. *Genes Dev* 9: 2314-2323.

37. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res.* 14: 1188-1190.
38. Das R, Zhou Z, Reed R (2000) Functional association of U2 snRNP with the ATP-independent spliceosomal complex E. *Mol Cell* 5: 779-787.
39. Dirksen WP, Hampson RK, Sun Q, Rottman FM (1994) A purine-rich exon sequence enhances alternative splicing of bovine growth hormone pre-mRNA. *J.Biol.Chem.* 269: 6431-6436.
40. Dirksen WP, Sun Q, Rottman FM (1995) Multiple splicing signals control alternative intron retention of bovine growth hormone pre-mRNA. *J.Biol.Chem.* 270: 5346-5352.
41. Dreumont N, Maresca A, Boisclair-Lachance JF, Bergeron A, Tanguay RM (2005) A minor alternative transcript of the fumarylacetoacetate hydrolase gene produces a protein despite being likely subjected to nonsense-mediated mRNA decay. *BMC Mol Biol* 6: 1.
42. Du H, Rosbash M (2002) The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing. *Nature* 419: 86-90.
43. Eddy SR (2001) Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* 2: 919-929.
44. Eisenberg E, Levanon EY (2003) Human housekeeping genes are compact. *Trends Genet.* 19: 362-365.
45. Eperon IC, Ireland DC, Smith RA, Mayeda A, Krainer AR (1993) Pathways for selection of 5' splice sites by U1 snRNPs and SF2/ASF. *EMBO J* 12: 3607-3617.
46. Eperon LP, Graham IR, Griffiths AD, Eperon IC (1988) Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? *Cell* 54: 393-401.
47. Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. *Science* 297: 1007-1013.
48. Fairbrother WG, Holste D, Burge CB, Sharp PA (2004) Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol* 2: E268.
49. Fedorov A, Saxonov S, Fedorova L, Daizadeh I (2001) Comparison of intron-containing and intron-lacking human genes elucidates putative exonic splicing enhancers. *Nucleic Acids Res* 29: 1464-1469.
50. Fleckner J, Zhang M, Valcárcel J, Green MR (1997) U2AF65 recruits a novel human DEAD box protein required for the U2 snRNP-branchpoint interaction. *Genes Dev* 11: 1864-1872.
51. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* 8: 967-974.
52. Förch P, Merendino L, Martínez C, Valcárcel J (2003) U2 small nuclear ribonucleoprotein particle (snRNP) auxiliary factor of 65 kDa, U2AF65, can promote U1 snRNP recruitment to 5' splice sites. *Biochem J* 372: 235-240.
53. Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF et al (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci U S A* 102: 16176-16181.
54. Fu XD (1993) Specific commitment of different pre-mRNAs to splicing by single SR proteins. *Nature* 365: 82-85.
55. Fu XD (1995b) The superfamily of arginine/serine-rich splicing factors. *RNA* 1: 663-680.

56. Fu XD (1995a) The superfamily of arginine/serine-rich splicing factors. *RNA*. 1: 663-680.
57. Fu XD, Maniatis T (1992) The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3' splice site. *Proc Natl Acad Sci U S A* 89: 1725-1729.
58. Galante PA, Sakabe NJ, Kirschbaum-Slager N, de Souza SJ (2004) Detection and evaluation of intron retention events in the human transcriptome. *RNA*. 10: 757-765.
59. Gelfand MS, Dubchak I, Dralyuk I, Zorn M (1999) ASDB: database of alternatively spliced genes. *Nucleic Acids Res* 27: 301-302.
60. Gilbert W (1978) Why genes in pieces? *Nature* 271: 501.
61. Gordon PM, Sontheimer EJ, Piccirilli JA (2000) Metal ion catalysis during the exon-ligation step of nuclear pre-mRNA splicing: extending the parallels between the spliceosome and group II introns. *RNA* 6: 199-205.
62. Gozani O, Potashkin J, Reed R (1998) A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site. *Mol Cell Biol* 18: 4752-4760.
63. Grainger RJ, Beggs JD (2005) Prp8 protein: at the heart of the spliceosome. *RNA* 11: 533-557.
64. Graveley BR (2000) Sorting out the complexity of SR protein functions. *RNA* 6: 1197-1211.
65. Graveley BR, Hertel KJ, Maniatis T (2001) The role of U2AF35 and U2AF65 in enhancer-dependent splicing. *RNA* 7: 806-818.
66. Graveley BR (2005) Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell* 123: 65-73.
67. Green MR (1991) Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. *Annu Rev Cell Biol* 7: 559-599.
68. Hanamura A, Cáceres JF, Mayeda A, Franza BR, Krainer AR (1998) Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. *RNA* 4: 430-444.
69. Hertel KJ, Graveley BR (2005) RS domains contact the pre-mRNA throughout spliceosome assembly. *Trends Biochem Sci* 30: 115-118.
70. Hide WA, Babenko VN, van Heusden PA, Seoighe C, Kelso JF (2001) The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res* 11: 1848-1853.
71. Hiller M, Huse K, Platzer M, Backofen R (2005) Non-EST based prediction of exon skipping and intron retention events using Pfam information. *Nucleic Acids Res* 33: 5611-5621.
72. Hoffman BE, Grabowski PJ (1992) U1 snRNP targets an essential splicing factor, U2AF65, to the 3' splice site by a network of interactions spanning the exon. *Genes Dev* 6: 2554-2568.
73. Homma K, Kikuno RF, Nagase T, Ohara O, Nishikawa K (2004) Alternative splice variants encoding unstable protein domains exist in the human brain. *J Mol Biol* 343: 1207-1220.
74. Hong X, Scofield DG, Lynch M (2006) Intron Size, Abundance, and Distribution within Untranslated Regions of Genes. *Mol.Biol.Evol.* 23: 2392-2404.

75. Ibrahim eC, Schaal TD, Hertel KJ, Reed R, Maniatis T (2005) Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proc.Natl.Acad.Sci.U.S.A* 102: 5002-5007.
76. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S et al (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol* 2: e162.
77. Izquierdo JM, Majós N, Bonnal S, Martínez C, Castelo R et al (2005) Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol Cell* 19: 475-484.
78. Jamison SF, Crow A, Garcia-Blanco MA (1992) The spliceosome assembly pathway in mammalian extracts. *Mol Cell Biol* 12: 4279-4287.
79. Jamison SF, Pasman Z, Wang J, Will C, Lührmann R et al (1995) U1 snRNP-ASF/SF2 interaction and 5' splice site recognition: characterization of required elements. *Nucleic Acids Res* 23: 3260-3267.
80. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM et al (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302: 2141-2144.
81. Kan JL, Green MR (1999) Pre-mRNA splicing of IgM exons M1 and M2 is directed by a juxtaposed splicing enhancer and inhibitor. *Genes Dev* 13: 462-471.
82. Kan Z, States D, Gish W (2002) Selecting for functional alternative splices in ESTs. *Genome Res* 12: 1837-1845.
83. Kandels-Lewis S, Séraphin B (1993) Involvement of U6 snRNA in 5' splice site selection. *Science* 262: 2035-2039.
84. Karolchik D, Baertsch R, Diekhans R, Furey TS, Hinrichs A et al (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.* 31: 51-54.
85. Kent OA, MacMillan AM (2002) Early organization of pre-mRNA during spliceosome assembly. *Nat Struct Biol* 9: 576-581.
86. Kent OA, Reayi A, Foong L, Chilibeck KA, MacMillan AM (2003) Structuring of the 3' splice site by U2AF65. *J Biol Chem* 278: 50572-50577.
87. Kohtz JD, Jamison SF, Will CL, Zuo P, Lührmann R et al (1994) Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. *Nature* 368: 119-124.
88. Konarska MM, Sharp PA (1987) Interactions between small nuclear ribonucleoprotein particles in formation of spliceosomes. *Cell* 49: 763-774.
89. Konarska MM, Sharp PA (1988) Association of U2, U4, U5, and U6 small nuclear ribonucleoproteins in a spliceosome-type complex in absence of precursor RNA. *Proc Natl Acad Sci U S A* 85: 5459-5462.
90. Kondrashov FA, Koonin EV (2001) Origin of alternative splicing by tandem exon duplication. *Hum Mol Genet* 10: 2661-2669.
91. Konforti BB, Koziolkiewicz MJ, Konarska MM (1993) Disruption of base pairing between the 5' splice site and the 5' end of U1 snRNA is required for spliceosome assembly. *Cell* 75: 863-873.
92. Kornblihtt AR, de la MM, Fededa JP, Munoz MJ, Nogues G (2004) Multiple links between transcription and splicing. *RNA*. 10: 1489-1498.

93. Krämer A (1996) The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annu Rev Biochem* 65: 367-409.
94. Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P et al (2003) Increase of functional diversity by alternative splicing. *Trends Genet* 19: 124-128.
95. Kuhn AN, Li Z, Brow DA (1999) Splicing factor Prp8 governs U4/U6 RNA unwinding during activation of the spliceosome. *Mol Cell* 3: 65-75.
96. Ladd AN, Cooper TA (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol* 3: reviews0008.
97. Lam BJ, Hertel KJ (2002) A general role for splicing enhancers in exon definition. *RNA* 8: 1233-1241.
98. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
99. Lareau LF, Green RE, Bhatnagar RS, Brenner SE (2004) The evolving roles of alternative splicing. *Curr Opin Struct Biol* 14: 273-282.
100. Ledee DR, Chen J, Tonelli LH, Takase H, Gery I et al (2004) Differential expression of splice variants of chemokine CCL27 mRNA in lens, cornea, and retina of the normal mouse eye. *Mol Vis* 10: 663-667.
101. Lee C, Atanelov L, Modrek B, Xing Y (2003) ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res* 31: 101-105.
102. Lee C, Wang Q (2005) Bioinformatics analysis of alternative splicing. *Brief Bioinform* 6: 23-33.
103. Lei H, Day INM, Vorechovský I (2005) Exonization of AluYa5 in the human ACE gene requires mutations in both 3' and 5' splice sites and is facilitated by a conserved splicing enhancer. *Nucleic Acids Res* 33: 3897-3906.
104. Lejeune F, Cavaloc Y, Stevenin J (2001) Alternative splicing of intron 3 of the serine/arginine-rich protein 9G8 gene. Identification of flanking exonic splicing enhancers and involvement of 9G8 as a trans-acting factor. *J Biol Chem* 276: 7850-7858.
105. Lerner MR, Boyle JA, Mount SM, Wolin SL, Steitz JA (1980) Are snRNPs involved in splicing? *Nature* 283: 220-224.
106. Letunic I, Copley RR, Bork P (2002) Common exon duplication in animals and its role in alternative splicing. *Hum Mol Genet* 11: 1561-1567.
107. Lewis BP, Green RE, Brenner SE (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* 100: 189-192.
108. Li Y, Blencowe BJ (1999) Distinct factor requirements for exonic splicing enhancer function and binding of U2AF to the polypyrimidine tract. *J Biol Chem* 274: 35074-35079.
109. Lim LP, Burge CB (2001) A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* 98: 11193-11198.
110. Lim S, Naisbitt S, Yoon J, Hwang JI, Suh PG et al (1999) Characterization of the Shank family of synaptic proteins. Multiple genes, alternative splicing, and differential expression in brain and development. *J Biol Chem* 274: 29510-29518.
111. Lim SR, Hertel KJ (2004) Commitment to splice site pairing coincides with A complex formation. *Mol Cell* 15: 477-483.

112. Liu HX, Chew SL, Cartegni L, Zhang MQ, Krainer AR (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol Cell Biol* 20: 1063-1071.
113. Liu HX, Zhang M, Krainer AR (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev* 12: 1998-2012.
114. Liu S, Altman RB (2003) Large scale study of protein domain distribution in the context of alternative splicing. *Nucleic Acids Res* 31: 4828-4835.
115. Louie E, Ott J, Majewski J (2003) Nucleotide frequency variation across human genes. *Genome Res.* 13: 2594-2601.
116. Luukkonen BG, Séraphin B (1998) A role for U2/U6 helix Ib in 5' splice site selection. *RNA* 4: 915-927.
117. Mabon SA, Misteli T (2005) Differential recruitment of pre-mRNA splicing factors to alternatively spliced transcripts in vivo. *PLoS Biol* 3: e374.
118. Magen A, Ast G (2005) The importance of being divisible by three in alternative splicing. *Nucleic Acids Res* 33: 5574-5582.
119. Maniatis T, Reed R (2002) An extensive network of coupling among gene expression machines. *Nature* 416: 499-506.
120. Maniatis T, Tasic B (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418: 236-243.
121. Maquat LE (2001) The power of point mutations. *Nat Genet* 27: 5-6.
122. Maquat LE (2005) Nonsense-mediated mRNA decay in mammals. *J Cell Sci* 118: 1773-1776.
123. Maroney PA, Romfo CM, Nilsen TW (2000) Functional recognition of 5' splice site by U4/U6.U5 tri-snRNP defines a novel ATP-dependent step in early spliceosome assembly. *Mol Cell* 6: 317-328.
124. Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: towards a cellular code. *Nat.Rev.Mol.Cell Biol.* 6: 386-398.
125. McCullough AJ, Berget SM (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol.Cell Biol.* 17: 4562-4571.
126. McCullough AJ, Berget SM (2000) An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5' splice sites. *Mol Cell Biol* 20: 9225-9235.
127. Melhuish TA, Wotton D (2006) The Tgif2 gene contains a retained intron within the coding sequence. *BMC.Mol.Biol.* 7: 2.
128. Merendino L, Guth S, Bilbao D, Martínez C, Valcárcel J (1999) Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG. *Nature* 402: 838-841.
129. Michael IP, Kurlender L, Memari N, Yousef GM, Du D et al (2005) Intron retention: a common splicing event within the human kallikrein gene family. *Clin.Chem.* 51: 506-515.
130. Michaud S, Reed R (1991) An ATP-independent complex commits pre-mRNA to the mammalian spliceosome assembly pathway. *Genes Dev* 5: 2534-2546.

131. Michaud S, Reed R (1993) A functional association between the 5' and 3' splice site is established in the earliest prespliceosome complex (E) in mammals. *Genes Dev* 7: 1008-1020.
132. Millhouse S, Manley JL (2005) The C-terminal domain of RNA polymerase II functions as a phosphorylation-dependent splicing activator in a heterologous protein. *Mol Cell Biol* 25: 533-544.
133. Mironov AA, Fickett JW, Gelfand MS (1999) Frequent alternative splicing of human genes. *Genome Res* 9: 1288-1293.
134. Modrek B, Resch A, Grasso C, Lee C (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* 29: 2850-2859.
135. Modrek B, Lee C (2002) A genomic view of alternative splicing. *Nat Genet* 30: 13-19.
136. Modrek B, Lee CJ (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.
137. Moore MJ, Query CC, Sharp PA (1993) Splicing of precursors to mRNA by the spliceosome. 1st Ed.: 303-357.
138. Moore MJ, Sharp PA (1993) Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing. *Nature* 365: 364-368.
139. Mount SM, Burks C, Hertz G, Stormo GD, White O et al (1992) Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res.* 20: 4255-4262.
140. Müller S, Wolpensinger B, Angenitzki M, Engel A, Sperling J (1998) A supraspliceosome model for large nuclear ribonucleoprotein particles based on mass determinations by scanning transmission electron microscopy. *J Mol Biol* 283: 383-394.
141. Murray HL, Jarrell KA (1999) Flipping the switch to an active spliceosome. *Cell* 96: 599-602.
142. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247: 536-540.
143. Ner-Gaon H, Fluhr R (2006) Whole-Genome Microarray in *Arabidopsis* Facilitates Global Analysis of Retained Introns. *DNA Res* 13: 111-121.
144. Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, Ophir R et al (2004) Intron retention is a major phenomenon in alternative splicing in *Arabidopsis*. *Plant J* 39: 877-885.
145. Newman A, Norman C (1991) Mutations in yeast U5 snRNA alter the specificity of 5' splice-site cleavage. *Cell* 65: 115-123.
146. Newman AJ (1997) The role of U5 snRNP in pre-mRNA splicing. *EMBO J* 16: 5797-5800.
147. Nilsen TW (2003) The spliceosome: the most complex macromolecular machine in the cell? *Bioessays* 25: 1147-1149.
148. Nilsen TW (2002) The spliceosome: no assembly required? *Mol Cell* 9: 8-9.
149. Nissim-Rafinia M, Kerem B (2002) Splicing regulation as a potential genetic modifier. *Trends Genet* 18: 123-127.

150. Nott A, Meislin SH, Moore MJ (2003) A quantitative analysis of intron effects on mammalian gene expression. *RNA* 9: 607-617.
151. Nurminsky DI, Nurminskaya MV, Benevolenskaya EV, Shevelyov YY, Hartl DL et al (1998) Cytoplasmic dynein intermediate-chain isoforms with different targeting properties created by tissue-specific alternative splicing. *Mol Cell Biol* 18: 6816-6825.
152. Nurtdinov RN, Artamonova II, Mironov AA, Gelfand MS (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum Mol Genet* 12: 1313-1320.
153. O'Keefe RT, Norman C, Newman AJ (1996) The invariant U5 snRNA loop 1 sequence is dispensable for the first catalytic step of pre-mRNA splicing in yeast. *Cell* 86: 679-689.
154. Offman MN, Nurtdinov RN, Gelfand MS, Frishman D (2004) No statistical support for correlation between the positions of protein interaction sites and alternatively spliced regions. *BMC Bioinformatics* 5: 41.
155. Ohler U, Shorom N, Burge CB (2005) Recognition of unknown conserved alternatively spliced exons. *PLoS Comput.Biol.* 1: 113-122.
156. Philipps DL, Park JW, Graveley BR (2004) A computational and experimental approach toward a priori identification of alternatively spliced exons. *RNA* 10: 1838-1844.
157. Piccirilli JA, Vyle JS, Caruthers MH, Cech TR (1993) Metal ion catalysis in the Tetrahymena ribozyme reaction. *Nature* 361: 85-88.
158. Popielarz M, Cavaloc Y, Mattei MG, Gattoni R, Stevenin J (1995) The gene encoding human splicing factor 9G8. Structure, chromosomal localization, and expression of alternatively processed transcripts. *J.Biol.Chem.* 270: 17830-17835.
159. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 33: D501-D504.
160. Raitskin O, Angenitzki M, Sperling J, Sperling R (2002) Large nuclear RNP particles--the nuclear pre-mRNA processing machine. *J Struct Biol* 140: 123-130.
161. Rappsilber J, Ryder U, Lamond AI, Mann M (2002) Large-scale proteomic analysis of the human spliceosome. *Genome Res* 12: 1231-1245.
162. Reed R (1989) The organization of 3' splice-site sequences in mammalian introns. *Genes Dev* 3: 2113-2123.
163. Reed R (2000b) Mechanisms of fidelity in pre-mRNA splicing. *Curr Opin Cell Biol* 12: 340-345.
164. Reed R (2000a) Mechanisms of fidelity in pre-mRNA splicing. *Curr Opin Cell Biol* 12 %6: 340-345.
165. Reed R, Hurt E (2002) A conserved mRNA export machinery coupled to pre-mRNA splicing. *Cell* 108: 523-531.
166. Resch A, Xing Y, Alekseyenko A, Modrek B, Lee C (2004a) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res* 32: 1261-1269.
167. Resch A, Xing Y, Modrek B, Gorlick M, Riley R et al (2004b) Assessing the impact of alternative splicing on domain interactions in the human proteome. *J Proteome Res* 3: 76-83.

168. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16: 276-277.
169. Robberson BL, Cote GJ, Berget SM (1990) Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol.Cell Biol.* 10: 84-94.
170. Roberts GC, Gooding C, Mak HY, Proudfoot NJ, Smith CW (1998) Co-transcriptional commitment to alternative splice site selection. *Nucleic Acids Res* 26: 5568-5572.
171. Rogers J, Wall R (1980) A mechanism for RNA splicing. *Proc Natl Acad Sci U S A* 77: 1877-1879.
172. Romano M, Marcucci R, Baralle FE (2001) Splicing of constitutive upstream introns is essential for the recognition of intra-exonic suboptimal splice sites in the thrombopoietin gene. *Nucleic Acids Res.* 29: 886-894.
173. Romfo CM, Alvarez CJ, van Heeckeren WJ, Webb CJ, Wise JA (2000) Evidence for splice site pairing via intron definition in *Schizosaccharomyces pombe*. *Mol.Cell Biol.* 20: 7955-7970.
174. Rosbash M, Séraphin B (1991) Who's on first? The U1 snRNP-5' splice site interaction and splicing. *Trends Biochem Sci* 16: 187-190.
175. Roscigno RF, Garcia-Blanco MA (1995) SR proteins escort the U4/U6.U5 tri-snRNP to the spliceosome. *RNA* 1: 692-706.
176. Ruskin B, Zamore PD, Green MR (1988) A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. *Cell* 52: 207-219.
177. Sakabe NJ, Vibranovski MD, de Souza SJ (2004) A bioinformatics analysis of alternative exon usage in human genes coding for extracellular matrix proteins. *Genet Mol Res* 3: 532-544.
178. Schneider TD, Stormo GD, Yarus MA, Gold L (1984) Delila system tools. *Nucleic Acids Res.* 12: 129-140.
179. Schwer B (2001) A new twist on RNA helicases: DEXH/D box proteins as RNAPases. *Nat Struct Biol* 8: 113-116.
180. Ségault V, Will CL, Polycarpou-Schwarz M, Mattaj IW, Branlant C et al (1999) Conserved loop I of U5 small nuclear RNA is dispensable for both catalytic steps of pre-mRNA splicing in HeLa nuclear extracts. *Mol Cell Biol* 19: 2782-2790.
181. Selenko P, Gregorovic G, Sprangers R, Stier G, Rhani Z et al (2003) Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. *Mol Cell* 11: 965-976.
182. Seraphin B, Rosbash M (1989) Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. *Cell* 59: 349-358.
183. Shapiro MB, Senapathy P (1987c) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* 15: 7155-7174.
184. Shapiro MB, Senapathy P (1987a) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res* 15: 7155-7174.

185. Shapiro MB, Senapathy P (1987b) RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* 15: 7155-7174.
186. Sharp PA (2005) The discovery of split genes and RNA splicing. *Trends Biochem Sci* 30: 279-281.
187. Shen H, Green MR (2004) A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. *Mol Cell* 16: 363-373.
188. Shen H, Green MR (2006) RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans. *Genes Dev* 20: 1755-1765.
189. Shen H, Kan JLC, Green MR (2004) Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Mol Cell* 13: 367-376.
190. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR et al (2006) Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res* 34: 3955-3967.
191. Sironi M, Menozzi G, Riva L, Cagliani R, Comi GP et al (2004) Silencer elements as possible inhibitors of pseudoexon splicing. *Nucleic Acids Res* 32: 1783-1791.
192. Sontheimer EJ (2001) The spliceosome shows its metal. *Nat Struct Biol* 8: 11-13.
193. Sontheimer EJ, Gordon PM, Piccirilli JA (1999) Metal ion catalysis during group II intron self-splicing: parallels with the spliceosome. *Genes Dev* 13: 1729-1741.
194. Sontheimer EJ, Sun S, Piccirilli JA (1997) Metal ion catalysis during splicing of premessenger RNA. *Nature* 388: 801-805.
195. Sorek R, Ast G (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res* 13: 1631-1637.
196. Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. *Genome Res* 12: 1060-1067.
197. Sorek R, Shamir R, Ast G (2004a) How prevalent is functional alternative splicing in the human genome? *Trends Genet* 20: 68-71.
198. Sorek R, Shemesh R, Cohen Y, Basechess O, Ast G et al (2004b) A non-EST-based method for exon-skipping prediction. *Genome Res* 14: 1617-1623.
199. Spann P, Feinerman M, Sperling J, Sperling R (1989) Isolation and visualization of large compact ribonucleoprotein particles of specific nuclear RNAs. *Proc Natl Acad Sci U S A* 86: 466-470.
200. Sperling R, Koster AJ, Melamed-Bessudo C, Rubinstein A, Angenitzki M et al (1997) Three-dimensional image reconstruction of large nuclear RNP (InRNP) particles by automated electron tomography. *J Mol Biol* 267: 570-583.
201. Staknis D, Reed R (1994) SR proteins promote the first specific recognition of Pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. *Mol Cell Biol* 14: 7670-7682.
202. Staley JP (1999) An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p. *Mol Cell* 3: 55-64.
203. Staley JP, Guthrie C (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* 92: 315-326.

204. Stamm S, Zhang MQ, Marr TG, Helfman DM (1994) A sequence compilation and comparison of exons that are alternatively spliced in neurons. *Nucleic Acids Res* 22: 1515-1526.
205. Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O et al (2000) An alternative-exon database and its statistical analysis. *DNA Cell Biol.* 19: 739-756.
206. Stamm S (2002) Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. *Hum Mol Genet* 11: 2409-2416.
207. Stamm S, Ben Ari S, Rafalska I, Tang Y, Zhang Z et al (2005) Function of alternative splicing. *Gene* 344: 1-20.
208. Stark JM, Bazett-Jones DP, Herfort M, Roth MB (1998) SR proteins are sufficient for exon bridging across an intron. *Proc Natl Acad Sci U S A* 95: 2163-2168.
209. Stephens RM, Schneider TD (1992) Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J Mol Biol* 228: 1124-1136.
210. Sterner DA, Berget SM (1993) In vivo recognition of a vertebrate mini-exon as an exon-intron-exon unit. *Mol.Cell Biol.* 13: 2677-2687.
211. Sterner DA, Carlo T, Berget SM (1996) Architectural limits on split genes. *Proc.Natl.Acad.Sci.U.S.A* 93: 15081-15085.
212. Stevens SW, Ryan DE, Ge HY, Moore RE, Young MK et al (2002) Composition and functional characterization of the yeast spliceosomal penta-snRNP. *Mol Cell* 9: 31-44.
213. Sudo T, Hidaka H (1998) Regulation of calyculin (S100A6) binding by alternative splicing in the N-terminal regulatory domain of annexin XI isoforms. *J Biol Chem* 273: 6351-6357.
214. Sugnet CW, Kent WJ, Ares MJr, Haussler D (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac.Symp.Biocomput.* 66-77.
215. Takahara K, Schwarze U, Imamura Y, Hoffman GG, Toriello H et al (2002) Order of intron removal influences multiple splice outcomes, including a two-exon skip, in a COL5A1 acceptor-site mutation that results in abnormal pro-alpha1(V) N-propeptides and Ehlers-Danlos syndrome type I. *Am.J.Hum.Genet.* 71: 451-465.
216. Talerico M, Berget SM (1994) Intron definition in splicing of small *Drosophila* introns. *Mol.Cell Biol.* 14: 3434-3445.
217. Tanaka K, Watakabe A, Shimura Y (1994) Polypurine sequences within a downstream exon function as a splicing enhancer. *Mol Cell Biol* 14: 1347-1354.
218. Taneri B, Snyder B, Novoradovsky A, Gaasterland T (2004) Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. *Genome Biol* 5: R75.
219. Tarn WY, Steitz JA (1994) SR proteins can compensate for the loss of U1 snRNP functions in vitro. *Genes Dev* 8: 2704-2717.
220. Tarn WY, Steitz JA (1995) Modulation of 5' splice site choice in pre-messenger RNA by two distinct steps. *Proc Natl Acad Sci U S A* 92: 2504-2508.
221. Thanaraj TA, Clark F, Muilu J (2003) Conservation of human alternative splice events in mouse. *Nucleic Acids Res* 31: 2544-2552.
222. Thanaraj TA, Stamm S, Clark F, Riethoven JJ, Texier VL et al (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res* 32: D64-D69.

223. Valadkhan S (2005) snRNAs as the catalysts of pre-mRNA splicing. *Curr Opin Chem Biol* 9: 603-608.
224. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. *Science* 270: 484-487.
225. Wagner EJ, Garcia-Blanco MA (2001) Polypyrimidine tract binding protein antagonizes exon definition. *Mol Cell Biol* 21: 3281-3288.
226. Wang BB, Brendel V (2006) Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci U S A* 103: 7175-7180.
227. Wang J, Smith PJ, Krainer AR, Zhang MQ (2005a) Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Res.* 33: 5053-5062.
228. Wang P, Yan B, Guo JT, Hicks C, Xu Y (2005b) Structural genomics analysis of alternative splicing and application to isoform structure modeling. *Proc Natl Acad Sci U S A* 102: 18920-18925.
229. Wang Z, Hoffmann HM, Grabowski PJ (1995) Intrinsic U2AF binding is modulated by exon enhancer signals in parallel with changes in splicing activity. *RNA* 1: 21-35.
230. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M et al (2004a) Systematic identification and analysis of exonic splicing silencers. *Cell* 119: 831-845.
231. Wang Z, Xiao X, Nostrand EV, Burge CB (2006) General and specific functions of exonic splicing silencers in splicing control. *Mol.Cell* 23: 61-70.
232. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M et al (2004b) Systematic identification and analysis of exonic splicing silencers. *Cell* 119: 831-845.
233. Wassarman DA, Steitz JA (1993) A base-pairing interaction between U2 and U6 small nuclear RNAs occurs in > 150S complexes in HeLa cell extracts: implications for the spliceosome assembly pathway. *Proc Natl Acad Sci U S A* 90: 7139-7143.
234. Weiner AM (1993) mRNA splicing and autocatalytic introns: distant cousins or the products of chemical determinism? *Cell* 72: 161-164.
235. Weinstein LB, Jones BC, Cosstick R, Cech TR (1997) A second catalytic metal ion in group I ribozyme. *Nature* 388: 805-808.
236. Will CL, Lührmann R (1997) Protein functions in pre-mRNA splicing. *Curr Opin Cell Biol* 9: 320-328.
237. Wu JY, Maniatis T (1993) Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* 75: 1061-1070.
238. Wu S, Romfo CM, Nilsen TW, Green MR (1999) Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature* 402: 832-835.
239. Wu Y, Zhang Y, Zhang J (2005) Distribution of exonic splicing enhancer elements in human genes. *Genomics* 86: 329-336.
240. Wyatt JR, Sontheimer EJ, Steitz JA (1992) Site-specific cross-linking of mammalian U5 snRNP to the 5' splice site before the first step of pre-mRNA splicing. *Genes Dev* 6: 2542-2553.
241. Xing Y, Lee C (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc Natl Acad Sci U S A* 102: 13526-13531.

242. Xu Q, Modrek B, Lee C (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res* 30: 3754-3766.
243. Xu R, Teng J, Cooper TA (1993) The cardiac troponin T alternative exon contains a novel purine-rich positive splicing element. *Mol Cell Biol* 13: 3660-3674.
244. Xu YZ, Newnham CM, Kameoka S, Huang T, Konarska MM et al (2004) Prp5 bridges U1 and U2 snRNPs and enables stable U2 snRNP association with intron RNA. *EMBO J* 23: 376-385.
245. Yeo G, Holste D, Kreiman G, Burge CB (2004a) Variation in alternative splicing across human tissues. *Genome Biol* 5: R74.
246. Yeo G, Hoon S, Venkatesh B, Burge CB (2004b) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A* 101: 15700-15705.
247. Yeo GW, Nostrand EV, Holste D, Poggio T, Burge CB (2005) Identification and analysis of alternative splicing events conserved in human and mouse. *Proc Natl Acad Sci U S A* 102: 2850-2855.
248. Yura K, Shionyu M, Hagino K, Hijikata A, Hirashima Y et al (2006) Alternative splicing in human transcriptome: Functional and structural influence on proteins. *Gene* 380: 63-71.
249. Zahler AM, Roth MB (1995) Distinct functions of SR proteins in recruitment of U1 small nuclear ribonucleoprotein to alternative 5' splice sites. *Proc Natl Acad Sci U S A* 92: 2642-2646.
250. Zamore PD, Green MR (1991) Biochemical characterization of U2 snRNP auxiliary factor: an essential pre-mRNA splicing factor with a novel intranuclear distribution. *EMBO J* 10: 207-214.
251. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA et al (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.
252. Zeng C, Berget SM (2000) Participation of the C-terminal domain of RNA polymerase II in exon definition during pre-mRNA splicing. *Mol Cell Biol* 20: 8290-8301.
253. Zhang XHF, Chasin LA (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* 18: 1241-1250.
254. Zheng CL, Fu XD, Gribskov M (2005) Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA*. 11: 1777-1787.
255. Zheng CL, Fu XD, Gribskov M (2006) Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA*. 11: 1777-1787.
256. Zheng ZM (2004) Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *J Biomed Sci* 11: 278-294.
257. Zhou Z, Licklider LJ, Gygi SP, Reed R (2002) Comprehensive proteomic analysis of the human spliceosome. *Nature* 419: 182-185.
258. Zillmann M, Rose SD, Berget SM (1987) U1 small nuclear ribonucleoproteins are required early during spliceosome assembly. *Mol Cell Biol* 7: 2877-2883.

259. Zorio DA, Blumenthal T (1999) Both subunits of U2AF recognize the 3' splice site in *Caenorhabditis elegans*. *Nature* 402: 835-838.
260. Zuo P, Maniatis T (1996) The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. *Genes Dev* 10: 1356-1368.
261. Zuo P, Manley JL (1994) The human splicing factor ASF/SF2 can specifically recognize pre-mRNA 5' splice sites. *Proc Natl Acad Sci U S A* 91: 3363-3367.

Página propositalmente deixada em branco

8 Anexo A - Documentação da ferramenta ASC utilizada para mapeamento e agrupamento de cDNAs e anotação de splicing alternativo

disponível em <http://www.compbio.ludwig.org.br/asc>

ASC documentation

version 0.01

Table of Contents

1.SYNOPSIS.....	3
2.AUTHORS.....	3
3.WHAT THE TOOLKIT DOES.....	4
4.INSTALLATION AND REQUIREMENTS.....	6
a. ASC (cDNA clustering, alternative splicing annotation tools and API).....	6
Requirements.....	6
Installation.....	6
b. ASC genome browser (optional resource).....	7
Requirements.....	7
Apache installation.....	7
Gbrowse installation.....	7
5.USING THE TOOLKIT.....	9
WHAT THE ASC TOOL KIT CONTAINS.....	9
Database creation scripts.....	9
Database access.....	10
BUILDING AN ASC DATABASE.....	10
A-Preparing input to ASC: formatting cDNA alignment coordinates (UCSC, BLAST, SIM4)....	12
ALTERNATIVE 1: Using data from the UCSC.....	13
>>Input: downloading UCSC data.....	13
>>Running: making UCSC data ready for use.....	14
Parameters in the configuration file required by ucsc2cct.pl, userBlast2cct.pl,	
userSim42cct.pl.....	14
>>Handling the output.....	15
ALTERNATIVE 2: Addition of in-house cDNAs or cDNAs not included in UCSC (BLAST or	
SIM4 user alignments).....	16
>>Input: aligning your sequences.....	16
>>Running: making BLAST or SIM4 data ready for use.....	18
>>Handling the output.....	18
B-Clustering cDNAs (cct.pl).....	19
Overview of the cDNA clustering procedure.....	19
1-Genome-based procedure (default).....	19
2-UniGene-based procedure.....	19
Setting parameters for different clustering methods.....	20

Parameters used in the configuration file required by cct.pl.....	20
ALTERNATIVE 1: Genome-based procedure (default).....	22
>>Input.....	22
>>Running: genome-based clustering.....	22
ALTERNATIVE 2: Using UniGene clusters.....	23
>>Input: downloading UniGene cluster info.....	23
>>Running: UniGene-based clustering.....	23
>>Handling the output: for both genome and UniGene-based clustering.....	24
C-Annotation of alternative splicing on genome-based cDNA Clusters (asc.pl).....	25
Overview of the alternative splicing annotation process.....	25
>>Input.....	26
>> Running asc.pl.....	26
>>Handling the output from asc.pl.....	27
Building ASC databases with different "quality" criteria.....	28
Parameters in the ASC configuration file (set1.conf).....	29
D-ACCESSING THE DATABASE I: Application Programming Interface.....	31
Initializing the API.....	31
Accessing cluster information.....	32
Accessing exon objects.....	32
Accessing alternative splicing events.....	33
ASC interface (API) documentation and reference.....	33
Sample script.....	33
Tips on improving API performance.....	36
Error checking in parameters passed to the methods.....	36
Querying only the cDNA clusters database.....	37
A very simple automatic documentation generator.....	38
Definition tags.....	39
E-ACCESSING THE DATABASE II: ASC browser.....	41
Installing ASC browser and its modified Gbrowse engine.....	41
Generating GFF's from cDNA clusters and alternative splicing annotation Javascript files.....	42
gbrowse_gff.pl parameters.....	43
Handling the output from gbrowse_gff.pl.....	45
A- handling gbrowse.dump.....	45
B- handling the .gff file.....	45
Setting up Gbrowse to display ASC clusters.....	45
C- handling the adaptor.myascdb directory.....	46
ASC browser/Gbrowse adaptors.....	46
Uninstalling ASC browser adaptors.....	47
6.TERMINOLOGY AND DEFINITIONS USED IN ASC.....	50
Terms.....	50
Biological (working) definitions.....	50
ASC API definitions.....	50
7.MAKING ALTERNATIVE SPLICING ANNOTATION RUN FASTER.....	51

1.SYNOPSIS

The ASC tool kit contains programs designed to simplify building, maintaining and querying a cDNA cluster based alternative splicing database. Primary data are cDNA alignment coordinates that are clustered based on the genome or any reference sequence. For such cDNA clusters, exons and alternative splicing events are represented in an object oriented fashion to facilitate handling and stored in a MySQL database. The user is allowed to use selected input sequences to annotate alternative splicing. The tool kit also contains a (object oriented) Perl API and a graphical viewer. UCSC's BLAT alignments can be used as the source of cDNA alignments to a genome.

The toolkit is implemented in Perl and is freely available at <http://www.compbio.ludwig.org.br/asc>

Written in November/2006 by Noboru Jo Sakabe in collaboration with Pedro AF Galante.

2.AUTHORS

Noboru Jo Sakabe [1,2], Pedro Alexandre Favoreto Galante [1,2], Sandro José de Souza [1]

[1] Ludwig Institute for Cancer Research, Sao Paulo Branch, Sao Paulo, Brazil

[2] PhD Program, Department of Biochemistry, Institute of Chemistry, University of Sao Paulo, Brazil

for questions on cDNA clustering: pgalante @ ludwig.org.br

for questions on alternative splicing annotation: noboru @ iq.usp.br

3.WHAT THE TOOLKIT DOES

Large-scale alternative splicing analyses involve the development of many programs to deal with great amounts of data. The data is typically a bunch of coordinates that must be compared, tested, summed, etc. in order to retrieve information. Many times, a "simplified version" of all such information would be enough to perform a given analysis. Some other times, one may need to know the specific coordinates of a given set of cDNA alignments and all the gory details.

Due to the great amount of information that one has to handle and the complex nature of the data, it is sometimes simpler to start writing a program from scratch to perform a similar task that a previous program did than modify a complex script. To avoid unnecessary code writing, minimizing the risk of introducing bugs in rewriting code and simplifying all tasks of building an alternative splicing database, we created a tool kit that aims at making the work simpler. After creating the database, the user can rely on simple methods to query the database whether sequence X is alternatively spliced, for example.

The ASC toolkit also allows the user to input only desired sequences or to exclude a few of a cDNA cluster, so to perform customized analyses.

An object-oriented solution

A way to simplify handling all such information, without losing its complexity, is to annotate not only cDNA coordinates but also alternative splicing *in relation to a higher monolithic entity* in a cDNA cluster, the exon. Such monolithic entity would be analogous to objects in Object Oriented Programming, a field in Computer Science (see Sun's JAVA handbook at <http://java.sun.com/docs/books/tutorial/java/concepts/index.html> for some concepts in Object Oriented Programming). Like the JAVA handbook explains, an object has states and behaviors. For example, a dog's name is a state, barking is a behavior. In the case of exons, the start and end coordinates would be states of the object. Whether it is alternatively or constitutively spliced would be a behavior, for example.

When one aligns cDNAs to a genome, one will obtain the coordinates of the "blocks" of the cDNAs that were aligned to a given genomic region, reporting an exon. For ASC, an exon object is an entity that groups all cDNA alignment "blocks" that report the same exon. It is a way to make things simpler, instead of handling tenths of exons with varying borders all the time, the "exon object" contains all information together, making it easier to handle the information as well as to query it. However, the complicated nature of alternative splicing makes an exon not so monolithic as we would like it to be. For example, an iron bar is an object with specific start and end coordinates in space. An exon is not; although in sequence it is essentially the same exon even if it is one nucleotide (or even more) shorter or longer, alternative splice sites break the ideally monolithic characterization of an exon. Even though, we can accept that an exon object represents a transcribed genomic region that also keeps track of all cDNA alignment "blocks" that report alternative borders.

Therefore, by calling an exon an exon object, all variability is encapsulated together and the fuzzy definition of an exon becomes more exact, and therefore easier to handle. It is easier now to add all alternative splicing information to a single place in the genome and verify if an exon (object) is skipped, if it presents alternative splice sites or if it is involved in intron retention with its neighbors.

Sometimes, when it is unavoidable to break the encapsulation (when pairwise evidences individually reports an event that does not exist considering all other sequences), ASC reports pairwise evidences.

Some conventions: intron retention

To put into practice the idea of "exon object", some conventions need to be made. In the case of intron

retention, we decided to create two separate exon objects when there is evidence of the existence of an intron in between. The "new" joining exon does not exist in our database as such, only the intron retention event. It would be weird to say that a gene presents 5 exons when one of them is in fact the composition of 2 of them plus an intron. Therefore, although the "exon object" itself does not exist, the information is present in the database and redundancy is avoided, *i.e.*, there is no overlap among "exon objects".

Object-oriented interface

It is important to note that the database itself is not object-oriented, in the sense that we do not store objects there. The object-oriented aspect of the database comes from the organization of the information around the exon object entity. To access the database in object-oriented fashion, one should use the Perl interface available in the tool kit. The interface will only accept object-oriented queries.

Many databases in one

ASC provides a means of building different alternative splicing databases using different stringency criteria like the minimum number of cDNAs that must exist to report an exon or an intron, restriction of use of EST sequences to build objects, etc. Therefore, over the same raw data, many different alternative splicing databases may be built.

This also means that the user can perform the annotation over any set of sequences, for example, a fresh set that one just finished sequencing last week.

Graphical viewer

The tool kit was designed to allow large-scale surveys to be conducted, and therefore one should use the Perl interface for that. However, a graphical viewer (using the Generic Genome Browser) is also available.

4.INSTALLATION AND REQUIREMENTS

a. ASC (cDNA clustering, alternative splicing annotation tools and API)

Requirements

- a computer running Linux
- Perl
- DBI/DBD
- MySQL up and running

All scripts run in ordinary PCs. ASC was developed on a Pentium 1.7 MHz, 528 Mb RAM and on an AMD64, 1.8 MHz, 1 Gb RAM, both running Linux.

Installation

Enter the directory of your ASC distribution asc-x.xx and do:

```
perl Makefile.PL
make
make install
make test
```

If anything in the installation fails, but you do have the requirements, all you have to do is:

- ◆ copy the entire directory ASC in asc-x.xx/ to your Perl modules' path, typically something like
 - /usr/local/lib/perl/5.8.7
 - /usr/local/share/perl/5.8.7
 - /usr/lib/perl5 /usr/share/perl5
 - /usr/lib/perl/5.8
 - /usr/share/perl/5.8
 - /usr/local/lib/site_perl
 - /usr/local/lib/perl/5.8.4
 - /usr/local/share/perl/5.8.4
- ◆ copy all .pl scripts in asc-x.xx/ to /usr/bin or /usr/local/bin or run ASC scripts by giving the full qualified path like ~/asc-0.01/asc.pl or /home/john/programs/asc-0.01/asc.pl for example

KEEP THE DIRECTORY asc-0.01 IN AN ACCESSIBLE LOCATION. It is required for installation of Gbrowse adaptors

b. ASC genome browser (optional resource)

Requirements

- a working installation of Gbrowse
- PHP >= 4.3.8
- Apache (should already be running for Gbrowse)
- CGI enabled (should already be running for Gbrowse)
- ability to connect MySQL databases from the Gbrowse host

Apache installation

Under Ubuntu Breezy, I installed the following packages:

```
apache  
libapache-mod-php4  
php4-mysql  
php4-cgi
```

I had to add the following line to my modules.conf at /etc/apache in order to run PHP in Apache:

```
LoadModule php4_module /usr/lib/apache/1.3/libphp4.so
```

Gbrowse installation

To install the ASC browser, you must have a server with Apache, PHP and MySQL running (above).

If you have, then install Gbrowse. These links might come handy:

1. [Getting BioPerl \(install before installing Gbrowse\)](http://bioperl.org/wiki/Getting_BioPerl) [http://bioperl.org/wiki/Getting_BioPerl]
2. [Installing Bioperl for Unix](http://bioperl.org/wiki/Installing_Bioperl_for_Unix) [http://bioperl.org/wiki/Installing_Bioperl_for_Unix]
3. [Getting Gbrowse](http://www.gmod.org/ggb/index.shtml) [http://www.gmod.org/ggb/index.shtml]
4. [Installing Gbrowse](http://www.gmod.org/?q=node/72) [http://www.gmod.org/?q=node/72]

Under Ubuntu Breezy I had no problems installing Gbrowse. Almost all missing Perl modules were installed through apt-get. My Apache was installed on /var/www and /var/www/cgi-bin is a symbolic link to /usr/lib/cgi-bin so I installed Gbrowse by doing:

```
perl Makefile.PL CONF=/etc HTDOCS=/var/www/ CGIBIN=/var/www/cgi-bin  
NONROOT=1  
make  
make install
```

After the installation is successful, enter your asc-x.xx directory and copy the directory *ascbrowser* to your www directory, which is typically /var/www (you must be root):

```
cp -r yourpath/asc-0.01/ascbrowser /var/www/  
chmod 777 /var/www/ascbrowser/gbrowse.html/tmp
```

IMPORTANT: in case your Gbrowse engine is located in a different host than the one where you installed ASC, copy the directory asc-x.xx/ASC to your Perl path in the Gbrowse host (or simply install ASC as explained in a.).

After you installed the ASC/Gbrowse system, you will need to install the adaptors that will connect ASC data and Gbrowse (see details later).

For now, use the test adaptor with test data (for Gbrowse 1.64, configured as above in /var/www, /var/www/cgi-bin, /etc/gbrowse.conf):

```
cp yourpath/asc-0.01/sample_data/gbrowse_test/asc.test.conf
/etc/gbrowse.conf
cp yourpath/asc-0.01/sample_data/gbrowse_test/gbrowse.asc.test
/var/www/cgi-bin
chmod 777 /var/www/cgi-bin/gbrowse.asc.test
cp -r yourpath/asc-0.01/sample_data/gbrowse_test/adaptor.test
/var/www/ascbrowser
```

That's all, if you point your web browser to yourserver/ascbrowser you will get the ASC browser interface. Select the set name “test” in the drop down menu. In the next screen you can view test cDNA clusters named “2454” and “2644”.

5.USING THE TOOLKIT

WHAT THE ASC TOOL KIT CONTAINS

Database creation scripts

<i>program</i>	<i>description</i>	<i>input</i>	<i>output</i>
ucsc2cct.pl	converts UCSC's BLAT alignments to the format used by the tool kit	UCSC alignments	MySQL database dump
userBlast2cct.pl userSim42cct.pl	converts BLAST alignments to the format used by the toolkit converts SIM4 alignments to the format used by the tool kit	Alignment result	MySQL database dump
ug2cct.pl	UniGene cluster information tool. Includes the UniGene cluster number for cDNA sequence in UCSC data (This is necessary to make clusters based on the UniGene database, see below).	UniGene cluster information	MySQL database dump (update)
cct.pl	cDNA clustering tool. Clustering based on cDNA alignments to the genome or on UniGene clusters	MySQL database	MySQL database dump (update)
cct-finish.pl	updates MySQL database above with additional information	MySQL database	MySQL database dump (update)
asc.pl	annotates alternative splicing on cDNA clusters generated by <i>cct.pl</i>	MySQL database	MySQL database dump
generate_bogus_asc.pl	produces an empty ASC database to allow the user to use the API only to query the cDNA alignment database	cluster ID list	MySQL database dump
gbrowse_gff.pl	generates data to be visualized with Gbrowse, an adaptor that connects ASC data and Gbrowse and javascript files that will display alternative splicing information and the GFF files that Gbrowse will display	MySQL database	MySQL database dump + GFF files
docparser.pl	produces HTML documentation for Perl modules, provided the tags required were used (see below)	Perl module	HTML
splitlist.pl	splits any given list in N pieces. Use this to run asc.pl faster by calling it multiple times	text file (a list)	multiple lists
gbrowse_gff_pseudoparallel.pl	runs gbrowse_gff.pl in parallel		

Database access

<i>utility</i>	<i>description</i>
ASC::DB_API	this Perl module is an object oriented interface (API) that contains several "methods" to access alternative splicing annotation
ASC browser	allows the user to view cDNA clusters as well as alternative splicing annotations on a web browser. Uses the Generic Genome Browser

BUILDING AN ASC DATABASE

Building a fresh alternative splicing database from scratch (cDNA alignment coordinates) with ASC involves 3 steps:

A- preparing input: formatting cDNA alignment coordinates and loading them to a MySQL database

B- cDNA clustering

C- alternative splicing annotation

Each of them is thoroughly described below. All steps are divided in 3 main sections:

- **Input:** describes what is required to run the current step (downloading data, previously built MySQL db, etc)
- **Running:** describes the commands and configuration files required to run a given step
- **Handling the output:** describes how to finish the current step (loading a MySQL db, etc)

The scheme below displays the process in detail, highlighting all the different steps performed. Use this figure as a reference for the steps to be performed.

For the examples below, the MySQL user is noboru, host=docontra. Please note that depending on how your MySQL host is configured, there might be differences on the way you can upload dumps. Here we assume your MySQL db asks for user password and accepts loading dumps directly from the command line.

See also the quickstart tutorial for a summary of the procedures below and approximate times that each task took during a test.

A-Preparing input to ASC: formatting cDNA alignment coordinates (UCSC, BLAST, SIM4)

The initial data for ASC are cDNA alignments to a genome (represented by the coordinate of the alignments). Currently, ASC accepts cDNA alignments from 3 different sources:

<i>alternative</i>	<i>source</i>	<i>comment</i>
1	UCSC	pre-computed cDNA-genome alignment data provided by UCSC can be used to generate a database of evidences of alternative splicing. Format the data to be used by ASC using <i>ucsc2cct.pl</i> . <i>Note that using this input data compromises detection of alternative splice sites, as BLAT does not precisely identify exon/intron borders as SIM4</i>
2	BLAST	you can align your own cDNA sequences using BLAST and use the parser <i>userBlast2gbcc.pl</i> to put alignments in ASC readable format. <i>Note that using this input data compromises detection of alternative splice sites, as BLAST does not precisely identify exon/intron borders as SIM4</i>
	SIM4	alternatively, use SIM4 alignments. Format the data to be used by ASC using <i>userSim42cct.pl</i>

In this step, the main task is to format raw data to ASC format. Other important tasks are:

- filtering data according to user parameters (length aligned, % identity)
- selecting only the best alignment per cDNA sequence to be retained
- dealing with problems in BLAT alignments (*ucsc2cct.pl* only)
 1. removing **gaps** when they occur with exactly the same size in both the genome and cDNA sequences (see below)
 2. removing **sequences** with discontinuous cDNA coordinates (± 3 nt) between alignment blocks (see below)
 3. removing **sequences** with very short (misaligned) blocks (see below))

ALTERNATIVE 1: Using data from the UCSC

>>Input: downloading UCSC data

Part A: cDNA-genome alignment data

The UCSC maps GenBank cDNA sequences to the genome of many organisms, for example, Human, Mouse, Chimpanzee, Rat and Dog (see the complete list at: <http://hgdownload.cse.ucsc.edu/downloads.html>).

1 - Go to the URL (<http://hgdownload.cse.ucsc.edu/downloads.html>), select an organism (for example: [Rat](#)) and click on [Annotation database](#). Then, download the files all_est.txt.gz and all_mrna.txt.gz. These files contain the cDNA (mRNAs and ESTs) alignment to the genome (cDNA mapping).

Alternatively, you can use wget:

```
wget http://hgdownload.cse.ucsc.edu/goldenPath/rn4/database/all_mrna.txt.gz
wget http://hgdownload.cse.ucsc.edu/goldenPath/rn4/database/all_est.txt.gz
```

2- Uncompress the data to go to the next step:

```
gunzip all_est.txt.gz all_mrna.txt.gz
```

Part B: CDS data

In addition to alignment coordinates, coding region coordinates can be stored in ASC databases.

1- Go to the URL (<http://hgdownload.cse.ucsc.edu/downloads.html>), select [Rat](#) and click on [Annotation database](#). Download the files cds.txt.gz and gbCdnaInfo.txt.gz. Alternatively, you can use wget:

```
wget http://hgdownload.cse.ucsc.edu/goldenPath/rn4/database/cds.txt.gz
wget http://hgdownload.cse.ucsc.edu/goldenPath/rn4/database/gbCdnaInfo.txt.gz
```

2- Uncompress the data

```
gunzip cds.txt.gz gbCdnaInfo.txt.gz
```

3- Parse CDS data

```
getCDS.pl -i cds.txt gbCdnaInfo.txt
```

A MySQL dump file named CDSallmRNAs.dump is generated (details on where to load this file in the section Handling the output).

>>Running: making UCSC data ready for use

3 – Next, you have to parse the UCSC alignments. The *ucsc2cct.pl* parser reads configuration parameters from a file.

a. Copy and paste the contents below to a file, for example named *cdna_align.conf*:

<i>cdna_align.conf</i> file contents:
identity=93; totalAlignment=0.8; minExonLength=8;

Parameters in the configuration file required by *ucsc2cct.pl*, *userBlast2cct.pl*, *userSim42cct.pl*

<i>parameter</i>	<i>values</i>	<i>description</i>	<i>default value</i>
Database connection		<i>the absence of these parameters cause the program to halt</i>	
identity	[floating point between 0 and 100]	minimum % identity in genomic alignment that a cDNA block must have to be used by ASC	93
totalAlignment	[floating point between 0 and 1]	minimum % of the extension of a cDNA sequence that must be aligned to the genome	0.8
minExonLength	[integer]	minimum size of an exon. Sometimes BLAT creates mini-exons due to misalignments in exon borders (see example below). As this is problematic for alternative splicing annotation, ASC excludes alignments with exons shorter than minExonLength	10

b. Parse files using the configuration file above:

```
ucsc2cct.pl -i all_est.txt -t est -c cdna_align.conf  
ucsc2cct.pl -i all_mrna.txt -t mrna -c cdna_align.conf
```

This program will parse UCSC alignments to ASC readable format and generate a MySQL dump named *[filename]_ucsc2cct.dump*, where *[filename]* is *all_est.txt* or *all_mrna.txt* in the example above.

About the parameter `minExonLength` and gaps in the alignments

According to the BLAT alignment available on-line in November/2006 at the UCSC webpage, the alignment of the cDNA **CO568725** consists of 3 blocks, one of which composed by 5 nt. Note, however, that there is a gap in both the genome and in the cDNA. This very short alignment block is not likely to be a real mini-exon, but rather due to misalignment.

```
(...)  
0000551 taggtccagatgctgcccacggccatgggggaacattgtattctgctgct 0000600  
<<<<<<< ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| <<<<<<<  
2288734 taggtccagatgctgcccacggccatgggggaacggtgtattctgctgct 2288685  
  
0000601 gtgtcctgtgatagcccaagaggtaaa.ctgcgtaacatt.gtccac 0000645 -> discontinuous by 24 nt  
<<<<<<< ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| <<<<<<<  
2288684 gtgtcctgtgatagcctaggaggtaaatctgcgtagcattgtccac 2288638 -> discontinuous by 28 nt
```

```
0000669 ggagac 0000674 <- a very short alignment block, due to misalignment caused by BLAT  
<<<<<<< ||||| <<<<<<<  
2288610 ggagac 2288605
```

```
0000676 gtgttcattgtttctct 0000692  
<<<<<<< ||||| <<<<<<<  
2288601 gtgttcattgtttctct 2288585
```

In this case, the cDNA coordinates are not continuous (gap of 24 nt) and ASC can do a few things about it.

The gap in the alignment is not exactly the same in both sequences (24 x 28 nt) and therefore, the alignment is not fixed. In this case, the sequence has to be removed because of the gap.

Even if it was not, the presence of an alignment block (exon) shorter than `minExonLength`, likely to be an error caused by BLAT, would cause its exclusion.

>>Handling the output

4- Create a MySQL database to store cDNA alignments (use the following commands or the ones you prefer to suit your MySQL configuration on passwords, permissions, host, etc):

```
mysql -u root -p -e 'create database mycdna_align_map'  
mysql -u root -p mycdna_align_map -e 'grant all privileges on  
mycdna_align_map.* to noboru'
```

5- Load the alignment dumps to the MySQL database:

```
mysql -u noboru -p mycdna_align_map < all_mrna.txt_ucsc2cct.dump  
mysql -u noboru -p mycdna_align_map < all_est.txt_ucsc2cct.dump
```

6- Load the CDS dump to the cDNA-genome alignment database

```
mysql -u noboru -p mycdna_align_map < CDSallmRNAs.dump
```

ALTERNATIVE 2: Addition of in-house cDNAs or cDNAs not included in UCSC (BLAST or SIM4 user alignments)

The user can also use his/her own sequence alignments (BLAST or SIM4) instead of UCSC outputs. User's alignments can also be ADDED to an existing UCSC database, or any alignment database created with ASC.

>>Input: aligning your sequences

Align your sequences with *blastall* or SIM4 and set the default to:

- **blastall** default output (full alignments) (parsed by ASC's *userBlast2cct.pl*)
- complete **SIM4** alignments (SIM4 parameters A=3 or A=4) (parsed by ASC's *userSim42cct.pl*)

Align ESTs and mRNAs separately, as ASC needs to know which is which.

blastall output format accepted by *userBlast2cct.pl*

```
BLASTN 2.2.1 [Jul-12-2001]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

Query= NM_001025
      (3665 letters)

Database: chr6.fa
      1 sequences; 170,899,992 total letters

Searching done

Sequences producing significant alignments:

                                Score      E
                                (bits)    Value
chr6                                2048    0.0

>chr6
      Length = 170899992

Score = 2048 bits (1033), Expect = 0.0
Identities = 1089/1117 (97%)
Strand = Plus / Plus

Query: 2166      aaattaacagtgcctaatgttattggtgtcttcactggatgtatttgactgctgtggactt 2225
                |||
Sbjct: 43860713 aaattaacagtgcctaatgttattggtgtcttcactggatgtatttgactgctgtggactt 43860772

Query: 2226      gagttgggaggggaatgttccactcagatcctgacaggaagaggaggagatgagagac 2285
                |||
Sbjct: 43860773 gagttgggaggggaatgttccactcagatcctgacaggaagaggaggagatgagagac 43860832

Query: 2286      tctggcatgatcnnnnnnnngtccacttggtggggccagggtcctctcccctgccagg 2345
                |||
Sbjct: 43860833 tctggcatgatctttttttgtccacttggtggggccagggtcctctcccctgccagg 43860892
```

SIM4 output format accepted by *userSim42cct.pl* (A=3 or A=4)

```
seq1 = seq_example1.fasta, 3665 bp
seq2 = chr6.fa (chr6), 170899992 bp

>NM_001025366
>chr6

1-1097 (43845931-43847027) 92% ->
1098-1149 (43850056-43850107) 100% ->
1150-1346 (43853184-43853380) 100% ->
1347-1423 (43854175-43854251) 100% ->
1424-1453 (43854604-43854633) 100% ->
1454-1576 (43856447-43856569) 100% ->
1577-1708 (43857671-43857802) 100% ->
2157-3655 (43860704-43862202) 95%

      0      .      :      .      :      .      :      .      :
      1 GGCTTGGGGCAGCCGGGTAGCTCGGAGGTCGTGGCGCTGGGGGCTAGCAC
        |||||||||||||||||||||||||||||||||||||||||||||||
43845931 GGCTTGGGGCAGCCGGGTAGCTCGGAGGTCGTGGCGCTGGGGGCTAGCAC

      50      .      :      .      :      .      :      .      :
      51 CAGCGCTCTGTCGGGAGGCGCAGCGGTTAGGTGGACCGGTCAGCGGACTC
        |||||||||||||||||||||||||||||||||||||||||||||||
43845981 CAGCGCTCTGTCGGGAGGCGCAGCGGTTAGGTGGACCGGTCAGCGGACTC

     100      .      :      .      :      .      :      .      :
     101 ACCGGCCAGGGCGCTCGGTGCTGGAATTTGATATTTCATTGATCCGGGTTT
        |||||||||||||||||||||||||||||||||||||||||||||||
43846031 ACCGGCCAGGGCGCTCGGTGCTGGAATTTGATATTTCATTGATCCGGGTTT
```

ASC reads the name of the sequences directly from the BLAST or SIM4 outputs. Therefore, the user must provide the definitive (and unique) names of the sequences in a standard way for ASC.

The header of FASTA files used for BLAST or SIM4 must be in this format:

```
>cDNA_name
sequences...
```

example of cDNA FASTA:

```
>NM_001025366
GGCTTGGGGCAGCCGGGTAGCTCGGAGGTCGTGGCGCTGGGGGCTAGCACCAGCGCTCTGTCGGGAGGCG
CAGCGGTTAGGTGGACCGGTCAGCGGACTCACCGGCCAGGGCGCTCGGTGCTGGAATTTGATATTTCATTG
ATCCGGGTTTTATCCCTCTCTTTTCTTTAAACATTTTTTTTAAACTGTATTGTTTCTCGTTTAAAT

>NM_0011254
TTTTCTTAAACATTTTTTTTAAACTGTATTGTTTCTCGTTTTTAATATCCGGGTTTTATCCCTCTTCTT
TTTAATATCCGGGTTTTATTATTTTGTCTTGCCATTCCCACTTGAATCGGGCCGACGGCTTGGGCGTAG
```

example of genomic FASTA:

```
>chr1
taaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
cctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
taaccctaaccctaaccctaaccctaaccctaaccctaaccctaacc
cctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
ccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
ccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
```

ATTENTION: This is important because the names of the sequences are required to build ASC databases.

>>Running: making BLAST or SIM4 data ready for use

The procedure is basically the same as for UCSC data. Parameters for filtering alignments are defined in the same way.

1-You only have to use another parser and provide BLAST or SIM4 alignments instead of UCSC files:

```
userBlast2cct.pl -i myalignments.blast -t mrna -c cdna_align.conf
```

or

```
userSim42cct.pl -i myalignments.sim4 -t mrna -c cdna_align.conf
```

>>Handling the output

2- Create a MySQL database to store cDNA alignments (**skip this step if you are adding alignments to a previously built database**):

```
mysql -u root -p -e 'create database mycdna_align_map'
mysql -u root -p mycdna_align_map -e 'grant all privileges on
mycdna_align_map.* to noboru'
```

3- Load the dumps to the MySQL database:

```
mysql -u noboru -p mycdna_align_map < myalignments.sim42cct.dump
```

or

```
mysql -u noboru -p mycdna_align_map < myalignments.blast2cct.dump
```

CDS data for user alignments

To add your own CDS data, create a MySQL dump containing the sequence accession number, the length of the CDS and the start and end coordinates:

```
insert into cds set seqacc="AB011407", len="353", start="1", end="354";
```

Load the dump to your alignment database:

```
mysql -u noboru -p mycdna_align_map < myCDSdata.dump
```

B-Clustering cDNAs (cct.pl)

Overview of the cDNA clustering procedure

The goal of cDNA clustering is to find subsets of sequences that are in the same genomic region (locus), in such a way that a **cluster** contains a **group of cDNA sequences** that were ideally transcribed from the same gene.

ASC (*cct.pl*) performs 2 types of clustering:

- i. genome-based, “from scratch”: based on the alignment of each transcript to the genome, *cct.pl* finds overlaps in the genome level and groups transcripts from the same gene
- ii. using third-party information (UniGene), *cct.pl* may assign UniGene clusters to the cDNA GenBank accession numbers present in the database

1-Genome-based procedure (default)

Given cDNA alignment to the genome coordinates, *cct.pl*:

1. finds overlapping cDNAs (N=1) that span a continuous genomic region (pre-clusters)
2. for each pre-cluster, *cct.pl* scans all cDNAs, comparing genomic alignment coordinates to identify groups of cDNAs that
 - i-share at least one exon-intron boundary or
 - ii-overlap by >N (default=15) in the extremities

See figure in the next section for illustration.

2-UniGene-based procedure

UniGene (<http://www.ncbi.nlm.nih.gov/UniGene/>) is the most used cDNA cluster information resource. The UniGene clustering process is both transcript (comparison within transcripts only) and genome-based and each cluster contains information on protein similarity, gene annotation, gene expression and gene homology. However, UniGene only has public sequences from GenBank.

cct.pl allows to:

- assign UniGene clusters to sequences in UCSC alignment data or to any (public) sequence with a GenBank accession number present in UniGene.
- in addition, in the case you are using **non-public** sequences, you can also use *cct.pl* to group them to UniGene clusters. This is possible because *cct.pl* will perform the genome-based clustering above to **add** these sequences to existing UniGene clusters.

Setting parameters for different clustering methods

The same script *cct.pl* is used to perform the 2 different clustering procedures above. To choose between the 2 types, one has simply to set a flag in a configuration file. This flag, MySQL database access information and all the clustering parameters required by *cct.pl* are presented below.

Parameters used in the configuration file required by cct.pl

<i>parameter</i>	<i>values</i>	<i>description</i>	<i>default value</i>
<i>Database connection</i>		<i>the absence of these parameters cause the program to halt</i>	
aligndbname	[string]	name of MySQL database containing cDNA alignments to reference sequence. Without this argument, <i>cct.pl</i> halts	-
aligndbhost	[string]	name of MySQL server. Without this argument, <i>cct.pl</i> halts	-
aligndbuser	[string]	name of MySQL user. Without this argument, <i>cct.pl</i> halts	-
aligndbpasswd	[string]	user password to access MySQL database, if necessary	-
<i>Clustering parameters</i>			
based_on_unigene	Y/N	Y=clustering process based on UniGene cluster N=cluster exclusively on overlap and similar exon/intron boundary	N
use_seq_out_unigene	Y/N	only works when based_on_unigene=Y: Y=use sequences not present in UniGene database. If you are using your own alignment and not using GenBank accession numbers, set this to Y N=non-public sequences are not clustered at all	N
overlap	[integer]	minimum number of nucleotides from two different transcripts that must overlap if there are no common splice borders. <i>This parameter only influences UniGene-based clustering if use_seq_out_unigene=Y</i>	15
border_slip	[integer]	number of nucleotides, plus or minus, by which two borders can differ and still be considered similar. <i>This parameter only influences UniGene-based clustering if use_seq_out_unigene=Y</i>	3
use_similar_boundary	Y/N	if Y, transcripts are only considered to belong to the same cluster if they share at least one splice border (differing by $\pm$ border_slip nt). <i>This parameter only influences UniGene-based clustering if use_seq_out_unigene=Y</i>	Y

You can see a graphical representation of the parameters used in *ucsc2cct.pl* and *cct.pl* in the scheme below:

Genome-based clustering

Setting `use_similar_boundary=N` will cluster cDNAs only by their overlap disregarding the exon/intron structure of transcripts.

Clustering with `use_similar_boundary=Y` is considerably more precise. Different overlapping genes can share exonic regions (for example cis-antisense genes), but rarely exon-intron boundaries. cDNAs from unprocessed mRNAs or genomic contaminant sequences also can share the exonic region with the cDNA prototype, but not the exon-intron structure. The criterion for considering two borders similar is a difference up to `border_slip` nt.

UniGene-based clustering (based_on_unigene=Y)

This clustering method was designed to cluster public sequences with an accession number in GenBank.

However, you can **add** non-public sequences or any sequence without a valid GenBank accession to existing UniGene clusters if you set `use_seq_out_unigene=Y`.

`cct.pl` will use the parameters `use_similar_boundary`, `border_slip` and `overlap` of the genome-base clustering method to find sequences in existing UniGene clusters that overlap with the non-public. *In order for this to work, you have to have public sequences clustered by UniGene. Note that it is impossible to build UniGene clusters with non-public sequences only.*

Therefore, the UniGene-based clustering with `use_seq_out_unigene=Y` allows you to **ADD non-public** sequences to **clusters of public sequences**. If you are **exclusively** using sequences without a GenBank accession number, you cannot use the UniGene-based clustering.

ALTERNATIVE 1: Genome-based procedure (default)

>>Input

The *cct.pl* script uses the MySQL database created in section “A-Preparing input to ASC” as input.

>>Running: genome-based clustering

a. Copy and paste the contents below to a file. In this example we will name it *clustering_param.conf*.

clustering_param.conf file contents:

```
# -----  
# 1. configure MySQL db where cDNA alignments are stored  
# -----  
aligndbhost=docontra  
aligndbname=mycdna_align_map  
aligndbuser=noboru  
aligndbpasswd=1259yzk  
# -----  
# 2. clustering parameters  
# -----  
based_on_unigene=N;  
use_seq_out_unigene=N;  
overlap=15;  
use_similar_boundary=Y;  
border_slip=3;
```

b. Parse files giving the configuration file above as a parameter:

```
cct.pl -c clustering_param.conf
```

This procedure will cluster cDNAs based on their alignment coordinates (from the MySQL database mycdna_align_map) and will group sequences that overlap by at least 15 nt in the extremities and share at least one exon/intron border.

See how to handle the output below.

ALTERNATIVE 2: Using UniGene clusters

Instead of clustering cDNAs from scratch, you can cluster them using pre-computed UniGene clusters (parameter: `based_on_unigene=Y` in the *clustering_param.conf* file). In this case you will need to download and parse data from UniGene:

>>Input: downloading UniGene cluster info

1. Go to the URL <http://www.ncbi.nlm.nih.gov/UniGene>, click on **Download UniGene** (left side on the web page), choose an organism (*Rattus norvegicus*) and download the file *.data.gz (for example, Rn.data.gz). Alternatively, you can use wget:

```
wget ftp://ftp.ncbi.nih.gov/repository/UniGene/Rattus_norvegicus/Rn.data.gz
```

2. Next, you have to gunzip, parse and upload the UniGene data to the MySQL database. Wait for complete loading of the MySQL database above before loading UGclusters.dump.

```
gunzip *.data.gz
ug2cct.pl -i Rn.data
mysql -u noboru -p mycdna_align_map < UGclusters.dump
```

>>Running: UniGene-based clustering

a. Set the parameter `based_on_unigene=Y` in the *clustering_param.conf* file:

NOTE: If you are using sequences which accession number are not in GenBank, probably non-public cDNA sequences, you must set `use_seq_out_unigene=Y` to cluster these sequences (see explanation above on how UniGene-based clustering for non-public sequences work).

clustering_param.conf file contents:

```
# -----
# 1. configure MySQL db where cDNA alignments are stored
# -----
aligndbhost=docontra
aligndbname=mycdna_align_map
aligndbuser=noboru
aligndbpasswd=1259yzk
# -----
# 2. clustering parameters
# -----
based_on_unigene=Y;
use_seq_out_unigene=N;
#overlap=15;
#use_similar_boundary=Y;
#border_slip=3;
```

b. Run *cct.pl*:

```
cct.pl -c clustering_param.conf
```

>>Handling the output: for both genome and UniGene-based clustering

Regardless of the clustering method you used, load the clustering data to your MySQL database:

```
mysql -u noboru -p mycdna_align_map < clustered.dump
```

This will assign cluster IDs to all transcripts that could be grouped.

The last procedure in generating the final cDNA cluster database is to run the script *cct_finish.pl* that will perform some final tasks on the cDNA clusters loaded in the MySQL database:

```
cct_finish.pl -c clustering_param.conf
```

Upload the dump:

```
mysql -u noboru -p mycdna_align_map < cct_finish.dump
```

C-Annotation of alternative splicing on genome-based cDNA Clusters (*asc.pl*)

Overview of the alternative splicing annotation process

asc.pl groups all cDNA alignment "blocks" that represent the same exon in an "object" (see scheme below) and annotates variations in the use of these Exon objects as alternative splicing events.

Scheme of objects and some concepts in ASC

¹ all exon/intron borders are annotated to one SpliceBorder object per Exon object border, here only 2 per object are circled.

Alternative splice sites are also recorded in SpliceBorder objects.

² all evidences are annotated to one ExonSkipEvent, here only 2 are circled

³ all evidences are annotated to one IRevent, here only 2 are circled

cDNA blocks 0-4 are "bound" to Exon object 5

The computational procedure involves ordering all cDNA blocks in a given cluster by their starting genomic coordinate (start_gnm in the scheme) and then looking if the current cDNA alignment block represents a new exon. In case it does, a new “Exon object” is created to group all similar blocks. Other objects as SpliceBorder, IRevent (intron retention), ExonSkipEvent, Transcript, Block and Cluster are also created to handle different data.

Alternative splicing evidenced by cDNA is then annotated to these objects. For example, intron retention will be annotated to both Exons 2 and 3 in the scheme above.

In the scheme, there is one mRNA in the cluster that is used to establish the strand of the gene. As it is in the minus strand, all Exons are numbered in the minus direction.

Details on the algorithm of exon object creation can be found in part II of this documentation.

>>Input

asc.pl uses the MySQL database produced by *cct.pl* and *cct_finish.pl* (cDNA clusters) as input and outputs another MySQL dump consisting of a number of tables with alternative splicing annotation.

>> Running *asc.pl*

To annotate alternative splicing with *asc.pl*, two input files are required:

a. The first must contain identifiers for the cDNA clusters (those in the alignment database) to which alternative splicing information will be built:

clusters.list file contents:

```
Rn.10  
Rn.1000  
Rn.10000
```

This file may be obtained directly through your freshly created database of clustered cDNA alignments “mycdna_align_map”.

```
mysql -u noboru -p mycdna_align_map -e "select cluster from clusters into  
outfile '/tmp/clusters.list'"  
cp /tmp/clusters.list .
```

b. the second file required is a configuration file. We will use a very simple one but there are several parameters that can be changed (see below). Save the contents below to a file.

set1.conf file contents:

```
aligndbname=mycdna_align_map;  
aligndbuser=noboru;  
aligndbpasswd=1259yzk;  
aligndbhost=docontra;
```

c. now to run *asc.pl* do:

```
asc.pl -i clusters.list -c set1.conf > asc.dump
```

Alternatively, one can annotate alternative splicing only to a subset of sequences of a cluster:

```
asc.pl -i clusters.list -c set1.conf -s input_seqs.list > asc.dump
```

The third argument above is a list of sequence accession numbers that should be either included or excluded from a given cluster. For example, the user wants to include only a bunch of sequences, not the entire cluster or wants to exclude two sequences. The file should be written as:

input_seqs.list file contents:

```
Hs.2134 NM_005266 +
Hs.2134 AB4005266 +
Hs.2134 AB6005276 +
Hs.0031 BC5006879 -
Hs.0031 AB4546848 -
```

In this case, for cluster Hs.2134, only sequences annotated with + will be considered. For cluster Hs.0031 all sequences will be considered, with the exception of both marked with “-”.

Therefore, the + token marks the sequences of the cluster to be included (all the others are excluded) and the - token those to be excluded (all the rest of the cluster is included).

Clusters with neither specification will have all their sequences automatically included.

Since excluding sequences implies accepting all others, it is not possible to both include and exclude sequences of the same cluster to build a data set. In case there are both specifications for the same cluster, only the inclusions (+) are considered. In the example below, only the two first sequences will be included in the cluster.

```
Hs.2134 NM_005266 +
Hs.2134 AB4005266 +
Hs.2134 AB6005276 -
```

>>Handling the output from *asc.pl*

The main output of *asc.pl* is a MySQL dump file.

Create another MySQL database for the alternative splicing annotation:

```
mysql -u root -p -e 'create database myascdb'
mysql -u root -p myascdb -e 'grant all privileges on myascdb.* to noboru'
```

Load the dump with:

```
mysql -u noboru -p myascdb < asc.dump
```

Building ASC databases with different "quality" criteria

The ASC tool kit allows one to build different alternative splicing annotations for the same cluster (see parameters **Selecting input sequences** and **Selecting exon objects** in the *asc.pl* configuration file). The usefulness of this feature is to allow the user to have different databases with varying parameters. For example, in one moment the user may want to perform a complete annotation, with all available cDNAs. Later, it might be interesting to only keep those exons that are confirmed by at least 2 cDNAs. Another user may need an annotation made using only mRNAs.

A table named `parameters` in the MySQL database containing all parameters is also created so the user can keep control of the differences among ASC DBs for the same cluster.

To build different databases, with different parameters, one should only modify the configuration file and re-run *asc.pl*.

For example, setting `input_useonlymrna=true` will discard all other sequences and of course they will not be used in any further step. All mRNAs are used, spliced or not. mRNA sequences are in general more trusted than regular ESTs and therefore this parameter should be used to produce more accurate, although less comprehensive, databases.

Unspliced cDNAs are automatically discarded, but setting the parameter `input_usenotspliced=true` will allow them to participate.

A "dirtier", but more complete, database would set `exonobj_useonlymrna=false`, allowing any EST to report exons and alternative splicing. One could make more stringent by setting `cdnas`, `spliced` and `upspliced` to more than 1, for example.

Exons with no cDNA bearing introns, and therefore with no precise borders are common, even if internal, because sometimes they are the first exon of a cDNA sequence. Therefore, setting `exonobj_splicedborders` to a value > 0 may be too stringent, but it is left as an option to the user.

Parameters in the ASC configuration file (set1.conf)

Changing the parameters bellow will affect exon object creation and alternative splicing annotation. Each parameter must be specified in one single line. Lines starting with '#' are ignored and treated as comments. Also, everything after ';' is ignored.

<i>parameter</i>	<i>values</i>	<i>description</i>	<i>default value</i>
<i>Database connection</i>		<i>the absence of these parameters cause the program to halt</i>	
aligndbname	[string]	name of MySQL database containing cDNA alignments to reference sequence generated by <i>cct.pl</i> . Without this argument, <i>asc.pl</i> halts	-
aligndbhost	[string]	name of MySQL server. Without this argument, <i>asc.pl</i> halts	-
aligndbuser	[string]	name of MySQL user. Without this argument, <i>asc.pl</i> halts	-
aligndbpasswd	[string]	user password to access MySQL database, if necessary	-
All the parameters below are optional			
<i>ASC database</i>			
comment	[string]	any comment, may be void	-
owner	[string]	name of the owner of the database, may be void	-
<i>Selecting input sequences</i>			
input_usenotspliced	true/false	if false, no unspliced cDNA from the cDNA alignment MySQL table (cdna_blocks) is considered	false
input_useonlymrna	true/false	if true, only mRNA sequences from the MySQL table cdna_blocks are used in all steps (building exon objects and annotating alternative splicing).	false
<i>Building exon objects</i>			
exonobj_join	[integer]	if two exon objects without an intron in between are at a distance < exonobj_join they are joined into one.	25
exonobj_useonlymrna	true/false	only blocks from mRNA sequences are considered for building exon objects, all the other sequences are used for alternative splicing annotation	false
exonobj_usenotspliced	true/false	if true, even unspliced fragments are allowed to create exon objects. Note that this may potentially include many unreal exons.	false

...continues in the next page

continued

<i>parameter</i>	<i>values</i>	<i>description</i>	<i>default value</i>
<i>Selecting exon objects</i>		<i>exon objects can be selected, allowing the user to keep only those with "better" evidence of being real. See Section "Building ASC databases with different 'quality' criteria"</i>	
cdnas	[integer]	minimum number of cDNAs (spliced or not) that must be bound to an exon object	1
spliced	[integer]	minimum number of spliced cDNAs that must be bound to an exon object	0
exonobj_splicedborders	[integer]	minimum number of cDNA fragments with an intron upstream or downstream of the exon object. Note that it is not required that <i>the same</i> fragment presents both introns, upstream and downstream.	0
exonobj_allowextrem	true/false	when exonobj_splicedborder > 0, exon objects in the extremities are discarded because they will never present up or downstream introns. To avoid this, set this parameter to true	false
defintron	[integer]	minimum number of introns that an I.3 object (exon object involved in intron retention) must bear to remain as an exon object, i.e. the number of intron evidences to consider before splitting an exon in two	1
<i>Alternative splicing detection</i>			
block_overlap	[integer]	when a block partially spans an intron, it can still be considered an intron retention if it is away by block_overlap nucleotides. If block_overlap = 0, then even if the block ends at, for example, 1000 and the next exon starts at 1001 it will not be considered an intron retention, but rather an alternative splice border of the previous exon	10
allow_IR_nonspliced	true/false	if true, allow even blocks that do not have both up and downstream introns at the same time to be evidence of intron retention	false

Notes: when using `input_usenotspliced=true`, you might get the following warning:

[WARNING]: AA069671 159 is not bound!

This might be because although unspliced sequences were used, they were not allowed to create exon objects.

Therefore, as some of them may lie in an intron (and likely to be genomic contaminants), they will not be bound to any existing exon object.

D-ACCESSING THE DATABASE I: Application Programming Interface

This tool kit in its present version contains two Perl modules that are APIs to the MySQL databases (cDNA alignment and alternative splicing annotation).

We tried to provide from very simple to more complex methods that may be useful for querying the database.

- **ASC::DB_API::Basic** contains many methods useful for querying alternative splicing events, features of the transcripts and cDNA blocks, etc. For example, it is used by *gbrowse_gff.pl* (one of the programs contained in the tool kit) to generate GFF files.
- **ASC::DB_API::Dumper** is a module with methods that makes one MySQL select only, fetching entire clusters, so data access is much faster when the intention is to iterate the entire cluster. It is also used by *gbrowse_gff.pl*. In fact, **ASC::DB_API::Basic** uses several of these methods.

The **ASC::DB_API::Basic** API consists of one single Perl module, divided in 8 other inner modules (objects): **Block**, **Cluster**, **Database**, **Exon**, **ExonSkipEvent**, **IRevent**, **SpliceBorder** and **Transcript**.

The use is simple, one must instantiate the object with `new( )` and use it:

```
use ASC::DB_API::Basic;  
  
$e=Exon->new("Rn.10");  
print $e->start(2);  
$ir=IRevent->new("Rn.10");  
print $ir->location(3);
```

All constructors are called `new( )` and require a cluster ID. The only exception is `Database->new( )` that does not require any parameter.

We opted to require the user to pass the exon object number as an argument instead of instantiating different objects for each exon or event for 2 main reasons: it is only necessary to call `new( )` once and it is easier to keep track of which object you are accessing.

Initializing the API

As the API queries the MySQL database, it is necessary to give it the data to access the database. Soon after you have loaded the API, initialize it:

```
use ASC::DB_API::Basic;  
ASC::DB_API::Basic->set_alignDB("mycdna_align_map", "yourhost", "yourname",  
"yourpassword");  
ASC::DB_API::Basic->set_ascDB("myascdb", "yourhost", "yourname", "yourpassword");
```

Accessing cluster information

Use the Cluster object to start your analyses. By querying this object you are able to find out how many Exon objects there are for this “gene”, whether it is alternatively spliced, what are the alternative splicing events, etc. Use the cluster number to query this object.

Accessing exon objects

In ASC, all exon objects are given a number that identifies them in the cluster. This number is called lnumber or lnum in ASC and is ordered by strand. That is, for plus clusters, the lnum increases with genomic coordinates. For clusters in the minus strand, lnum decreases with increasing genomic coordinate.

The first exon has lnum=1, increasing sequentially until the last in the cluster.

Use the method `Cluster->getLnumbers()` in the Basic API to retrieve the list of exon objects' lnum for a cluster. The lnum is required for most methods of object Exon.

Query this object to find information as the length of the consensus exon, the number of cDNAs that represent it, their accession numbers, etc.

Use the method `Exon->getBlocksBound()` to retrieve IDs for all cDNA blocks that represent a given exon object (are bound to in ASC lingo).

Remember that Exon objects group all information of all cDNA blocks bound. Therefore, information as the length may not correspond to a real observed transcript. For example, if a cDNA block started at 120 and finished at 200 and another started at 100 and ended at 198, the Exon object coordinates will be 100 to 200 and the length=100. However, no single cDNA block was observed with such length. In general, the Exon object is useful for pre-selecting data, but detailed “real-life” data should always come from individual selected transcripts.

Let's say I want to extract the sequence of exons shorter than 200 nt:

- for each cluster, fetch all Exon objects
- ask for the cDNA blocks “bound”
- for each cDNA block, ask for its length
- retain those <200 and store their start_gnm, end_gnm coordinates to extract the sequence from the genome, or better, start_cdna, end_cdna to extract sequence from cDNA FASTA

Accessing alternative splicing events

In order to access data on a specific alternative splicing event, events were given a code that mnemonically represent them:

alternative splicing event	object	code	comment
exon skipping	ExonSkipEvent	n or n,m	meaning that exon <i>n</i> is skipped or exons from <i>n</i> until <i>m</i> are skipped together
intron retention	IRevent	n,n+1	means that the intron between exons <i>n</i> and <i>n+1</i> is retained
alternative splice site	SpliceBorder	an integer	alternative splice sites are treated as exon borders and therefore are coded directly as splice border coordinates

For example, to access the number of sequences that retain intron between exon objects 2 and 3 of cluster Rn.10, you have to do:

```
$ir=IRevent->new("Rn.10");  
@array=$ir->getSeqsRetaining("2,3");
```

There are methods in the API to obtain event codes.

ASC interface (API) documentation and reference

Documentation for each method in the API is available at the ASC website and in your ASC package (asc-x.xx/api_doc/index.html).

Sample script

You can test the API by running the script below. Copy and paste the code to a file name something like test_script.pl. You can also find this script at asc-x.xx/sample_data/sample_script.pl.

Configure the lines that initialize the API with your data:

```
ASC::DB_API::Basic->set_alignDB("mycdna_align_map", "yourhost", "yourname",  
"yourpassword");  
ASC::DB_API::Basic->set_ascDB("myascdb", "yourhost", "yourname", "yourpassword");
```

```
#!/usr/bin/perl -w

use ASC::DB_API::Basic;
use ASC::DB_API::Dumper;

#sample script that uses the ASC::DB_API::Basic interface
#this script fetches information about all exons in the database. The user can
choose to pick skipped or non-skipped exons.
#input: 1-a list with cluster IDs, 2-noas or aeu

open(IN, "<$ARGV[0]") or die "no clusters list\n";
$type=$ARGV[1]; if(!$type){die "no type of exons [noas] [aeu]\n";}

ASC::DB_API::Basic->set_alignDB("mycdna_align_map", "docontra", "noboru",
"1259yzk");
ASC::DB_API::Basic->set_ascDB("myascdb", "docontra", "noboru", "1259yzk");

print "This script gets, for each cluster, all skipped (aeu) or constitutive
(noas) exons and, for each exon, prints its length, counts the number of cDNA
blocks bound (that are evidence of the exon object). For the exons not involved
in intron retention, the script chooses 1 representative cDNA block. For this
cDNA block, if it has both upstream and downstream introns, i.e. defined
exon/intron borders, prints the start and end coordinates\n\n\n";

$counter=0;
$cluster_previous="";
while(<IN){
    chomp;
    $cluster=$_;
    print "\n:: cluster:$cluster ", $counter++, "\n";

    # avoids instantiating the same cluster twice.
    if($cluster ne $cluster_previous){
        $e      = Exon->new($cluster);
        $tr      = Transcript->new($cluster);
        # making ASC more efficient, reusing instantiated objects
        $clusterobj = Cluster->new($cluster, $tr);
        $block      = Block->new($cluster, $tr);
        $ir          = IRevent->new($cluster, $block);

        # only accepting oriented clusters (w/ mRNA)
        $strand = $clusterobj->strand($cluster);
        if($strand eq "?"){print "excluding this cluster, it has no mRNAs\n";
next;}

        if($type eq "noas"){ @exons=$clusterobj->getConstitutiveSkipExons();}
        else{                @exons=$clusterobj->getAltSkipExons();        }
        if(!@exons){        print "this cluster has no $type exons\n"; next }
    }
    $cluster_previous=$cluster;

    # scans all exons of the cDNA cluster
    for($i=0; $i<@exons; $i++){
        $lnum=$exons[$i];
        # excludes all exons that are involved in any intron retention event
        if(!$ir->isEvent($lnum, ($lnum+1)) ){
            #choose a cDNA block to get exon length from, given it has an
            upstream and a downstream intron, i.e. a complete exon

```

```

        @blocks=$e->getBlocksBound($lnum);

        # picking additional info about the exon object (how many complete
and partial cDNAs represent this object)
        $numeid=$#blocks+1; # the number of blocks bound to the exon
        $len1 = $e->length($lnum);
        print "Exon $lnum of cluster $cluster is represented by:$numeid
cDNAs and has a length of $len1 nt\n";
        # scan all cDNA blocks that represent this exon object to chose 1
        for($j=0; $j<@blocks; $j++){
            # picking block ID
            $blid=$blocks[$j];
            # further verifying if this specific block is involved in
intron retention
            if(!$block->retainsIntron($blid)){
                # if the cDNA block has a downstream intron and
                $down=$block->hasIntronDown($blid);
                if($down){
                    # an upstream intron
                    $up=$block->hasIntronUp($blid);
                    if($up){
                        # let's get some information
                        $start = $block->start_gnm($blid);
                        $end = $block->end_gnm($blid);
                        $seqacc = $block->blockid2acc($blid);
                        $n_introns=$tr->n_introns($seqacc);
                        print " The blockID:$blid was selected and has
start coord:$start and end coord:$end.\n";
                        print " It belongs to sequence:$seqacc that has
$n_introns intron(s).\n";
                        # we only want 1 representative sequence per exon
object
                        last;
                    }
                }
            }
        }
    }
}
else{
    print " This exon object is involved in intron retention, I'll
discard it\n";
}
}
}

```

To run it, you should have already installed ASC::DB_API and have the ASC database that we built above ready. Run it doing:

```
perl yourpath/asc-x.xx/sample_data/sample_script.pl clusters.list aeu >
test.out
```

clusters.list is the list that you used in section C and “aeu” is a parameter to tell the script to fetch information only for exons with alternative exon usage. You can also try “noas” to do the same for non-skipped exons.

Tips on improving API performance

Granted, the ASC API is not fast, it is actually very slow. The greatest strength of the ASC API is to allow the generation of abstract programs that make data retrieval easier. There are a few things that may considerably improve the performance of your scripts.

Tip 1: As it is not fast, it is a good practice to write scripts to retrieve high-level data. You run it once and store the retrieved data in files. Later, you can parse these files for specific information, which is much faster.

Tip 2: To avoid extensive MySQL database queries (and make it even slower), some objects load entire MySQL tables to memory. For example, the Block object loads the table `cdna_blocks` of a given cluster. It also loads table `transcripts`.

The Transcript object itself also loads the MySQL table `transcripts` to memory. If you instantiate both Transcript and Block objects of the same cluster, they will load the MySQL table twice to memory. To save time, memory and improve general performance, you can:

1-instantiate a Transcript object

2-give it as an (optional) argument to a Block object (of the same cluster, of course), so it will not reload the table `transcripts` to memory.

```
$tr=Transcript->new($cluster);  
$bl=Block->new($cluster, $tr);  
$ir=IRevent->new($cluster, $bl);
```

The API documentation tells when you can give an already instantiated object as argument and which object.

Tip 3: A good practice is to avoid reloading objects of a cluster twice. In most occasions, your input data for a given script will include cluster IDs. It is always a good idea to sort them before running your script, making different entries for the same cluster appear one after the other. This way, you can instantiate each cluster only once:

```
# only instantiate new objects if the cluster is new  
if($cluster ne $cluster_previous){  
    print STDERR "instantiating objects for new cluster.";  
    $tr=Transcript->new($cluster);  
    $bl=Block->new($cluster, $tr);  
    $ir=IRevent->new($cluster, $bl);  
}  
else{  
    print STDERR "reusing objects.";  
}  
$cluster_previous=$cluster;
```

Error checking in parameters passed to the methods

Unfortunately, few checkings are done. The methods check if the parameters required were passed, but, in

general, not if they are valid parameters. In this regard, *lnum*, *event* and *seqacc* can never be 0 or "" (empty string). However, if you give *lnum*=11 for a cluster with *lnum*<6, no error checking will be made, but errors will occur (Perl warnings may spread everywhere...).

The best way to avoid these kind of errors is to always fetch values from Cluster methods. For example, in the *lnum* example, if you ask for

```
$c=Cluster->new("acluster");  
@lnum=$c->getLnumbers();
```

and iterate @lnum, the API will only work with valid *lnum*.

Other errors may be more difficult to avoid, but they should be rare.

Querying only the cDNA clusters database

Although the ASC tool was designed to generate an alternative splicing database based on cDNA clusters, you may have a particular reason to use only the cDNA clusters database. The API was designed to only work in the presence of both databases, the alternative annotation and cDNA database. All objects verify if the requested cluster ID exists in both databases, and if one of them doesn't exist returns 0.

In order to bypass this requirement of the API and use only the cDNA cluster database, it is necessary to “trick” the API. For this, you may load some “bogus” information in the alternative splicing database, so the API will verify that the cDNA cluster is actually there and instantiate the object. Of course you will only be able to use the methods that request alignment data.

Use the following script to generate an ASC db that will contain enough information to allow objects to be instantiated:

```
generate_bogus_asc.pl clusters.list > bogusasc.dump
```

Load the dump generated as a regular ASC db.

```
mysql bogus_asc < bogusasc.dump
```


A very simple automatic documentation generator

If you add or modify a method in the API, you can run the script *docparser.pl* to regenerate the HTML documentation. Go to the directory where the documentation will be and do:

```
docparser.pl -f path/ASC/DB_API/Basic.pm -n "API doc"
```

The script will produce all the HTML and CSS files in the current directory.

Note that *docparser.pl* is a humble documentation generator (far from Javadoc or anything similar). It was designed to produce fast and simple automatic annotation, no more than that. It doesn't extensively check correctness of the documentation provided, the user is asked to take full responsibility over the correctness of the fields provided.

"Private" methods declared as *sub private_name* are not documented.

The user should provide, right above the sub declaration, the following structure:

field	description
#args:=	brief description of arguments comma separated (ex.: cluster number, seqacc). The parser automatically grabs arguments given to the subs by looking for the token @_. Therefore, you must ALWAYS read all your parameters via (my \$self, my \$arg1, my \$arg2)=@_ The variable \$self contains the object of the method called. If your method does not receive arguments, it is not necessary to use @_
#desc:=	description of the method. You can use any valid HTML in here. You can also use definition tags (see below), assuming they were defined previously. The tag will be substituted by #deflink as defined (see below). You can add as many tags as you want.
#qdesc:=	quick description that will appear in the method's reference table
#ex:=	example code, will appear in Courier font
#return:=	return values (integer, array of integers, etc)
#returnd:=	more extensive description of return values
}	it is necessary to end your subs with "}" in the beginning of the line, it is used to end documentation for a given method. No other "}" in the middle of the code should be in the beginning of a line. sub test(){ (my \$self, my \$arg1, my \$arg2)=@_ code... } <-- end sub with this token in the beginning of the line

The "}" in the beginning of a line closing a sub is taken by the parser as the end of the documentation, so a typical (and working...) documentation should be like this.

```
#args:=cluster number
#desc:=Exon object constructor
#qdesc:=Exon object constructor
#ex:=$exon=Exon->new(564);<br>$exon->length();
#return:=Exon object
sub new(){
  (my $self, my $arg1)=@_;
  code...
} #<-- end of documentation
```

Definition tags

Sometimes, to explain a method it is necessary to explain other concepts. In order to avoid writing the same concept again and again, it is possible to write a definition of a given concept or idea and just link it to the documentation of the method.

The way to use it is to declare a definition and link it by a tag in a description.

A tag definition looks like this:

```
#defname:=Cluster strand determination
#defcode:=strand
#deflink:=(See <a href="definitions.html#strand" target="_blank">cluster
strand</a>)
#defd:=The strand where a cluster lies is defined on the basis of the
orientation of mRNAs present in that cluster. In principle, all mRNAs
should have the same orientation. If they don't have, the majority of the
orientations win and define whether the cluster is in the plus or minus
strand. If no mRNAs are available, the cluster is left as plus. The rule
for deciding whether a cluster is plus or minus can be modified by
changing the function verifyStrand() at the the main .pl file.
#defend <- token indicates end of description
```

and it is linked by a text by its #defcode that must be written as **::#defcode::strand::** anywhere in the documentation text.

```
#desc:=returns the ending coordinate of the given exon object, but considering if
the cluster is in the plus or minus strand ::#defcode::strand::. Coordinates are
genomic or reference sequence's. For plus clusters, the start_gnm is always lower
than end_gnm. For minus clusters, the start_gnm is actually the end_gnm coordinate
and therefore start_strand is always <b>greater</b> than end_strand.
```

becomes

description	returns the ending coordinate of the given exon object, but considering if the cluster is in the plus or minus strand (See cluster strand). Coordinates are genomic or reference sequence's. For plus clusters, the start_gnm is always lower than end_gnm. For minus clusters, the start_gnm is actually the end_gnm coordinate and therefore start_strand is always greater than end_strand.
--------------------	--

The **::#defcode::strand::** was substituted by the text "See [cluster strand](#)" as specified by #deflink and a hyperlink to the description was produced (click on the link to see it working).

Definitions work similarly for method documentation, it consists of specific fields that will later be shown as a fancy table. This table will be linked to a tag in your HTML text.

It is advised to declare all definitions in the beginning of the Perl file that you are documenting, so it will soon be loaded and any #defcode tag will be promptly substituted by its #deflink.

The user should specify the following fields:

<i>field</i>	<i>description</i>
#defname:=	definition title (required)
#defcode:=	a simple expression used to link a tag to the definition (required). To use this #defcode, your documentation text must contain the expression ::#defcode::code:: where code is a #defcode (see the strand example above and other examples in ASC/DB_API/Basic.pm)
#deflink:=	the text that will appear in all places where your defcode is put (required)
#defd:=	textual description of the method. You can use any valid HTML here. You can also use definition tags, assuming they were defining previously. The tag will be substituted by #deflink as defined previously.
#defex:=	example to illustrate a definition, any HTML is valid (optional)
#defend=	just a label for the docparser.pl to know that the definition ended (required)

Note that no verification is made to check whether the user correctly entered all the fields, it is the responsibility of the user to provide the correct fields. The penalty for wrong specification is probably only the non-inclusion of a description.

You can learn more about writing your own methods and documentation by hacking a bit the ASC/DB_API/Basic.pm module.

E-ACCESSING THE DATABASE II: ASC browser

The **ASC browser** is a graphical viewer that displays cDNA sequences grouped by clusters and information about exon objects. It is actually just an interface to the Generic Genome Browser (**Gbrowse**, <http://www.gmod.org/ggb/index.shtml>) that finds information for the cluster (displayed as a Javascript pop up) and calls Gbrowse to show the sequences.

Note that you don't need to use ASC browser as your only genome browser option if you don't want to!! You can simply use *gbrowse_gff.pl* to generate your GFFs and configure Gbrowse as you normally would. It is just that ASC browser provides a way to visualize alternative splicing annotation in your web browser and has a very simple interface to query your data.

Installing ASC browser and its modified Gbrowse engine

In order to ASC browser work, it requires:

- a working installation of Gbrowse (with MySQL)
- PHP >= 4.3.8
- Apache (should already be running for Gbrowse)
- CGI enabled (should already be running for Gbrowse)
- ability to connect MySQL databases with ASC DBs

You may find information on how to install a Gbrowse MySQL database at <http://www.gmod.org/ggb/tutorial/tutorial.html> (see Setting up a MySQL database in the referred document).

After you have installed Gbrowse, to install ASC browser you should simply do (must be root):

```
cd asc-x.xx
cp -r ascbrowser /var/www/
chmod 777 /var/www/ascbrowser/gbrowse.html/tmp
```

If you point your browser to *yourhost/ascbrowser* you should see this.

However, it won't be working because there is no data to be displayed, which takes us to the next step.

Generating GFF's from cDNA clusters and alternative splicing annotation Javascript files

Gbrowse displays data formatted in a specific way, the GFF format. The ASC toolkit contains a .gff file generator. You must provide a list of clusters of which you want the GFFs and a configuration file (see parameters below). Generate GFFs by using the command below.

```
gbrowse_gff.pl -i clusters.list -c gbrowsegff.conf > clusters.list.gff
```

gbrowsegff.conf file contents:

```
# configure MySQL where clustered cDNA alignments are stored
# -----
aligndbhost=docontra;
aligndbname=mycdna_align_map;
aligndbuser=noboru;
aligndbpasswd=1259yzk;

# configure MySQL where ASC alternative splicing info is stored
# -----
ascdbhost=docontra;
ascdbname=myascdb;
ascdbuser=noboru;
ascdbpasswd=1259yzk;

# configure MySQL where GFFs are stored
# -----
gffdbhost=docontra
gffdbname=mygff
gffdbuser=noboru
gffdbpasswd=1259yzk

#other conf
ascdb_setname=myascdb;
asc_path=/home/noboru/asc;
```

If you generated an ASC annotation for selected sequences (with *asc.pl*), instead of the entire cluster, you can give the same list as an argument to *gbrowse_gff.pl* so it will only generate GFF files for the selected sequences, but this is not obligatory.

```
gbrowse_gff.pl -i clusters.list -c gbrowsegff.conf -s input_seqs.list >
clusters.list.gff
```

***gbrowse_gff.pl* parameters**

<i>parameter</i>	<i>values</i>	<i>description</i>	<i>default value</i>
<i>Database connection (data sources)</i>		<i>the absence of these parameters cause the program to stop</i>	
aligndbhost	[string]	name of host with MySQL database containing cDNA alignments to genome or reference sequence	-
aligndbname	[string]	name of MySQL server	-
aligndbuser	[string]	name of MySQL user	-
aligndbpasswd	[string]	user password to access MySQL database	-
ascdbhost	[string]	name of host with MySQL database containing ASC annotation	-
ascdbname	[string]	name of MySQL server	-
ascdbuser	[string]	name of MySQL user	-
ascdbpasswd	[string]	user password to access MySQL database	-
<i>GFF MySQL database (Gbrowse will read data from)</i>		<i>the absence of these parameters cause the program to stop</i>	
gffdbhost	[string]	host name of the MySQL server storing GFFs	-
gffdbname	[string]	name of the MySQL database where the GFF will be	
gffdbuser	[string]	user name for access to MySQL database with GFF	-
gffdbpasswd	[string]	password name for access to MySQL database with GFF	-
<i>other important parameters</i>		<i>the absence of these parameters cause the program to stop</i>	
asc_path	[string]	fully qualified path of your asc-0.01 directory. <i>gbrowse_gff.pl</i> copies data from the directory asc-0.01	
ascdb_setname	[string]	a name for the viewer (spaces are removed)	

continues...

continued

<i>Optional parameters</i>			
http_path	[string]	only necessary if you install ASC browser in other path than /var/www	/var/www
ascbrowser_path	[string]	only necessary if you install ASC browser in other path than ascbrowser	ascbrowser
full_cgi_path	[string]	only necessary if your cgi-bin directory is not in /var/www	/var/www/cgi-bin
cgi_path	[string]	only necessary if you install ASC browser in other path than /cgi-bin, for example, "/cgi-bin/ascscrip". Notice that the fully qualified path is actually /var/www/cgi-bin set in full_cgi_path, but here only what is after /var/www matters	/cgi-bin
gbrowse_script_path	[string]	when installing Gbrowse, a test database is automatically installed. The script <i>gbrowse</i> is put under your /cgi-bin directory. You need to tell <i>gbrowse_gff.pl</i> where this script is	/var/www/cgi-bin/gbrowse
conf_dir	[string]	directory where your Gbrowse configuration files lay. It can be any directory, preferably not under /var/www	/etc/gbrowse.conf
<i>Data input</i>			
nounspliced	[integer]	1=do not display non-spliced sequences, 0=to display them (separate track)	1
<i>Output</i>			
show_max_evid	[integer]	To save space, each link of evidence information will display a maximum of show_max_evid sequences)	10
<i>Gbrowse</i>			
margin	[integer]	sets the amount of "blank space" to show before and after a cluster displayed by Gbrowse	1000
maxseq	[integer]	Gbrowse fails to handle large clusters, either getting too slow or crashing. To avoid this, large clusters are split into chunks of <i>maxseq</i> sequences.	1000
gbmaxlen	[integer]	this is the maximum length of a feature that Gbrowse displays. Clusters longer than this will have to be viewed by pieces	1 000 000

Handling the output from *gbrowse_gff.pl*

gbrowse_gff.pl outputs:

- **A- a MySQL dump file:** *[ascdb_setname].gbrowse.dump* contains data that must be loaded into the same MySQL database where your ASC DB lies.
- **B- to standard output:** (saved as *clusters.list.gff*) GFF file to be (loaded and) displayed by Gbrowse
- **C- a directory:** *adaptor.[ascdb_setname]* is a directory that connects ASC annotations and Gbrowse. This directory must be moved or copied under */var/www/ascbrowser* (see section below).

A- handling *gbrowse.dump*

Large clusters cannot be handled by Gbrowse. This way, we split them into smaller chunks of *maxseq* (you can configure this in the *gbrowse_gff.pl* configuration file) sequences. A MySQL table relating the original cluster and its split parts is also created so ASC browser can handle them.

>>Load *gbrowse.dump* to the same database where your source ASC data is located:

```
mysql -u noboru -p myascdb < myascdb.gbrowse.dump
```

B- handling the .gff file

In order to load the .gff file, you need to set a Gbrowse database.

Setting up Gbrowse to display ASC clusters

By default, Gbrowse will read GFFs from a MySQL database in your Gbrowse host.

Create a MySQL database, for example named *mygff*. Your Gbrowse installation has a tool for uploading GFF files to a MySQL database.

```
mysql -u root -p -e 'create database mygff'
mysql -u root -p mygff -e 'grant all privileges on mygff to noboru'
```

>>Upload your .gff file

```
bp_load_gff.pl -c -d mygff clusters.list.gff
```

If you are loading the GFFs to a MySQL server in another host, do:

```
bp_load_gff.pl -c --dsn
dbi:mysql:database=mygff:host=anotherhost:user=john:password=abc
clusters.list.gff
```

You can configure Gbrowse to read GFF from the disk. For this, change the configuration file named *asc.[ascdb_setname].conf* that is located inside your Gbrowse configuration directory (*/etc/gbrowse.conf*). See <http://www.gmod.org/nondrupal/tutorial/tutorial.html> for further information.

C- handling the adaptor.myascdb directory

The last output is a directory that has to be put under `/var/www/ascbrowser`. See below.

ASC browser/Gbrowse adapters

Up to now you have 1-installed Gbrowse, 2-generated ASC annotations, 3-generated GFFs to be displayed, 4-loaded GFFs to Gbrowse. There is one last step to makes ASC browser fully functional, installing an "adaptor" between Gbrowse and ASC databases:

1- we made some modifications in the Gbrowse *gbrowse* script so it would display a Javascript file with alternative splicing annotation. For each ASC database, a fresh *gbrowse* script is generated, together with all the necessary Gbrowse stuff, configuration files, etc.

2- the information on split clusters that you loaded after generating the GFF files also must be accessed by the ASC browser.

To deal with all this, for each user set of alignment and ASC annotation data, *gbrowse_gff.pl* creates an adaptor that contains all MySQL configurations. Therefore, the adaptor is the piece of software that manages all this interaction between Gbrowse and ASC databases.

ASC browser is therefore a PHP script that looks for adaptors (directories under `/ascbrowser`). These adaptors are also PHP scripts (generated by *gbrowse_gff.pl*) that are configured to display alternative splicing annotation from specific user sets and databases. Finally, Gbrowse displays clusters graphically in a web browser.

The adaptor is automatically generated when *gbrowse_gff.pl* is executed. All you have to do is to install it in the correct location (that you gave as parameters).

>> Install the adaptor

```
source install.myascdb.sh
```

gbrowse_gff.pl allows you to automatically install the adaptor in the place you configured (your *http_path*) normally */var/www/ascbrowser*. You may set different paths by changing the configuration file *gbrowsegff.conf*

```
gbrowse_gff.pl -i clusters.list -c gbrowsegff.conf -install >
clusters.list.gff
```

If for any reason you need to regenerate the adaptor (lost it or changed a database, for example) you don't need to generate all the GFF file (which takes a lot of time). You can give to *gbrowse_gff.pl* the argument *-adaptor*.

```
gbrowse_gff.pl -c gbrowsegff.conf -adaptor -install
```

The configuration file is read and the adaptor generated. You can substitute the older adaptor in */var/www/ascbrowser* and everything will work under this new configuration.

Uninstalling ASC browser adaptors

Everytime you run *gbrowse_gff.pl*, it will generate a shell script named *uninstall.setname* that you can run to uninstall the adaptor named “*ascdb_setname*”. The shell script will remove the adaptor.*[ascdb_setname]* directory, the *gbrowse.asc.[ascdb_setname]*, *asc.[ascdb_setname].conf* and *ascbrcnf.[ascdb_setname]* files.

NOTES:

1- In case you are having connection problems, verify your configuration file for the correct paths, passwords and MySQL database parameters. If you like to see blood, you can check configurations by yourself (assuming you installed Gbrowse at */var/www*, with config files at */etc/gbrowse.conf* and ASC browser as */var/www/ascbrowser*):

a. In file */etc/gbrowse.conf/asc.[ascdb_setname].conf*, verify whether your MySQL database containing your GFFs are correctly configured:

```
[GENERAL]
description = ASC browser - map5
db_adaptor = Bio::DB::GFF
db_args = -adaptor dbi::mysql
db_args      = -adaptor dbi::mysql
              -dsn dbi:mysql:database=mygff;host=docontra
user         = noboru
pass         = 1259yzk
```

b. In file */var/www/cgi-bin/gbrowse.asc.[ascdb_setname]*, verify whether your adaptor path is correctly configured:

```
#####
# Non-modperl users should change this variable if needed to point
# to the directory in which the configuration files are stored.
#
$CONF_DIR = '/var/www/ascbrowser/adaptor.[ascdb_setname]';
#
#####
```

c. Verify if *gbrowse.asc.[ascdb_setname]* is in the place where *loadcluster.php* expects it to be:

```
$gbrowse_script = "/cgi-bin/gbrowse.asc.[ascdb_setname]";
```

d. Check if

```
/var/www/ascbrowser/gbrowse.html/tmp  
/var/www/ascbrowser/adaptor.sample/js  
/var/www/cgi-bin/gbrowse.asc.setname
```

are with permissions 777.

2- If you want to modify Gbrowse or ASC browser installation paths you MUST modify the last part of the *gbrowse_gff.pl* configuration file to inform the new paths. Otherwise, no modifications should be made.

3-Two files control the access to the genome browser: *index.php* and *loadcluster.php*, each adaptor has both files configured differently.

The first parses the input (i.e. a cluster number, a cluster number and start/end coordinates or a sequence accession number) and asks *loadcluster.php* to display the cluster. If a sequence accession number is entered, this script will display information on the alignment of this sequence and tell in which cluster it is located. The user can then ask Gbrowse to display the sequence or the entire cluster.

This latter script will check if the cluster spans > N nt (default 1 000 000 nt). If it does, it will display the exon objects and ask the user to select a sub-region. This is because there is a parameter in Gbrowse to limit the length of the feature to be displayed. Note that it can be changed in the .conf file (section "Gbrowse details").

4-The alternative splicing tool tips are Javascript files downloaded by the adaptor for each cluster. Therefore, if you ask directly Gbrowse to display a cluster, through its original search box, it will not be able to display alternative splicing information. To do this you have to ask for clusters by the ASC browser interface.

General scheme of how ASC DBs are integrated to Gbrowse through ASC browser.

Once you made Step 1, the data will be permanently added to your MySQL data, together with your ASC data.

6.TERMINOLOGY AND DEFINITIONS USED IN ASC

Terms

_cdna	indicates the feature is relative to the cDNA coordinate system
_gnm	indicates the feature is relative to the genome coordinate system
bind, block bound to exon object	when a block overlaps an exon object, the block is considered "bound"
Gbrowse	The Generic Genome Browser (http://www.gmod.org/ggb/index.shtml)
ASC browser	The PHP interface brought by ASC that connects Gbrowse engine to ASC data

Biological (working) definitions

Alternative splice site	The existence of at least two splice borders in the same border of an exon object
Alternative splice site event	Each alternative splice border
block, cDNA block	a block of alignment of a cDNA to the genome (or any reference sequence with introns) obtained by alignment programs as BLAST, SIM4, BLAT. A transcript "exon"
cluster	a set of cDNAs that span the same genomic region, representing transcripts of a same gene
event	the occurrence of a given detectable difference among transcripts
Exon object	A data object that stores information on the evidences for this transcribed region, such as position, number, alternative splicing events and other
Exon skip event	Absence of one or more exons at a time from a transcript
Intron retention	Presence of an intron plus parts of two bordering exons in a transcript. An intron must have been defined by at least one other transcript, that is, another transcript must have only the two bordering exons.
Intron retention event	Occurrence of a defined/retained intron in different transcripts
Splice border	A block border, as defined by the alignment, that is accompanied by an up or downstream exon, thus defining the intron and therefore the border

ASC API definitions

You can find definitions of terms and concepts used in the implementation of ASC to annotate alternative splicing in the web page or at asc-x.xx/api_doc/basic/definitions.html.

7.MAKING ALTERNATIVE SPLICING ANNOTATION RUN FASTER

The easiest (and maybe the only...) way to make ASC faster is to make it run in parallel. As the input is a single list of clusters, the simplest parallelization possible is to split the input clusters in different computers. Often, ASC requires the full processor, so it is a good idea to split the input in more than one computer. The script *splitlist.pl* will, as the name says, split any list given in N pieces.

```
splitlist.pl -i clusters_list.in -n 5
```

The command above will split *clusters_list.in* in 5, naming them *clusters_list.in.1* to *clusters_list.5*. Now you can start *asc.pl* with each different split list in different computers.

The script *gbrowse_gff.pl* will also benefit from splitting input into many computers. However, *gbrowse_gff.pl* rarely uses the full capacity of the processor, so simply splitting the input in the same computer is already a good idea.

For making this simple parallelization, you can use *gbrowse_gff_pseudoparallel.pl* which will split the input list and fire many separate processes in the same computer. When I tested it splitting the input in 3 batches, it actually reduced processing time by almost 3-fold.

```
gbrowse_gff_pseudoparallel.pl -i clusters_list.in -c gbrowse.conf -n 3 -o test.out
```

The command above will split *clusters_list.in* in 3 batches, naming them *clusters_list.in.1*, *clusters_list.2* and *clusters_list.3* and start a process for each. The output GFF files will be named *test.out.1*, *test.out.2* and *test.out.3*.

You can check how many clusters were processed by doing something like:

```
awk '{print $1}' test.out.* | sort -n | uniq | wc
```

ASC documentation, part II (implementation)

version 0.01

Table of Contents

1.ABOUT THIS DOCUMENT.....	2
2.TERMINOLOGY AND DEFINITIONS USED IN ASC.....	3
Terms.....	3
Biological (working) definitions.....	3
ASC API definitions.....	4
3.ALGORITHM AND IMPLEMENTATION.....	5
Implementation.....	5
Perl modules.....	5
Database buildup procedure.....	5
Overview.....	5
Preprocessing.....	6
Phase I: creating exon objects.....	6
Algorithm for creating exon objects.....	8
Updating exon objects borders	9
The use of unspliced cDNA blocks is limited.....	9
Phase II: reassigning blocks to objects.....	10
Phase III: selecting exon objects.....	10
Phase IV: binding blocks to exon objects.....	11
Phase V: analyzing alternative splicing in the cluster.....	11
Cluster orientation (plus/minus strand).....	12
Exon object numbering systems.....	12
CDS annotation.....	12
Computational details.....	13
Data structures generated.....	13
4.ILLUSTRATION OF THE ALTERNATIVE SPLICING ANNOTATION PROCEDURE IN A SAMPLE CLUSTER.....	14

1.ABOUT THIS DOCUMENT

The ASC tool kit contains programs designed to simplify building, maintaining and querying a cDNA cluster based alternative splicing database. Primary data are cDNA alignment coordinates that are clustered based on the genome or any reference sequence. For such cDNA clusters, exons and alternative splicing events are represented in an object oriented fashion to facilitate handling and stored in a MySQL database. The user is allowed to use selected input sequences to annotate alternative splicing. The tool kit also contains a (object oriented) Perl API and a graphical viewer. UCSC's BLAT alignments can be used as the source of cDNA alignments to a genome.

This second part of the documentation describes in details how ASC builds exon objects and annotates alternative splicing to them.

The toolkit is implemented in Perl and is freely available at <http://www.compbio.ludwig.org.br/asc>

Written in November/2006 by Noboru Jo Sakabe.

2.TERMINOLOGY AND DEFINITIONS USED IN ASC

Terms

_cdna	indicates the feature is relative to the cDNA coordinate system
_gnm	indicates the feature is relative to the genome coordinate system
bind, block bound to exon object	when a block overlaps an exon object, the block is considered "bound"
Gbrowse	The Generic Genome Browser (http://www.gmod.org/ggb/index.shtml)
ASC browser	The PHP interface brought by ASC that connects Gbrowse engine to ASC data

Biological (working) definitions

Alternative splice site	The existence of at least two splice borders in the same border of an exon object
Alternative splice site event	Each alternative splice border
block, cDNA block	a block of alignment of a cDNA to the genome (or any reference sequence with introns) obtained by alignment programs as BLAST, SIM4, BLAT. A transcript "exon"
cluster	a set of cDNAs that span the same genomic region, representing transcripts of a same gene
event	the occurrence of a given detectable difference among transcripts
Exon object	A data object that stores information on the evidences for this transcribed region, such as position, number, alternative splicing events and other
Exon skip event	Absence of one or more exons at a time from a transcript
Intron retention	Presence of an intron plus parts of two bordering exons in a transcript. An intron must have been defined by at least one other transcript, that is, another transcript must have only the two bordering exons.
Intron retention event	Occurrence of a defined/retained intron in different transcripts
Splice border	A block border, as defined by the alignment, that is accompanied by an up or downstream exon, thus defining the intron and therefore the border

See below a graphical representation of objects present in the API and a few concepts.

Scheme of objects and some concepts in ASC

¹ exon/intron borders are annotated to one SpliceBorder object per Exon object border, here only 2 per object are circled

² all evidences are annotated to one ExonSkipEvent, here only 2 are circled

³ all evidences are annotated to one IRevent, here only 2 are circled

3.ALGORITHM AND IMPLEMENTATION

Implementation

The alternative splicing annotation system was implemented in Perl, consisting of a main script (asc.pl) and 6 modules. It requires the DBI/DBD module for database connection. Global variables were avoided, but there are still some to speed up. The main file contains only function calls to the modules.

Perl modules

AlternativeSplicing.pm	each type of alternative splicing (alternative exon usage, intron retention and alternative splice sites) is treated independently by an individual function.
BuildExonObjects.pm	exon object creation phases are separated in different functions in this module
Input.pm	fetches blocks, stores in array and do other preprocessing routines
PrintLog.pm	all logging functions are grouped here, including the functions that log exon and event objects, opening and closing of log file, etc
ReadParameters.pm	parameter file is read and parsed by this module
WriteMySQLDump.pm	produces dump file for MySQL loading

Database buildup procedure

Overview

The annotation for alternative splicing is generated in 5 "phases":

Preprocessing: retrieving cDNA blocks that will be used and storing them in an array for fast access

Phase I: creates exon objects - given the blocks aligned to the genome, decides if new exon objects should be created

Phase II: assign blocks to exon objects - one block may overlap 1 or more exon objects

Phase III: select exon objects according to user parameters – in the previous phase, cDNA blocks that represent a given exon object are counted. The user may only accept exon objects that were confirmed by at least 2 blocks, for example.

Phase IV: definitely binds blocks to exon objects - given the accepted exon objects, this phase will re-assign blocks to exon objects. It also annotates when a block is assigned to more than 1 exon object, which is important for intron retention analysis. Other annotations are performed, for example, if a given exon object is located in a CDS of a given mRNA.

Phase V: analyzes blocks and exon objects for alternative splicing - exon skipping, intron retention, alternative

splice sites.

Preprocessing

The first task of asc.pl is to retrieve the blocks that will be used for alternative splicing annotation from the alignment database (MySQL table `cdna_blocks`) to speed up processing.

During this phase, information as start and end coordinates of the alignment, % identity, etc is loaded into data structures (`@blocks`, `@mrna_strand`). Other information will be added to `@blocks` during buildup procedure, as for example, if it retains intron, to which exon objects it is bound and so on.

Implemented in Input.pm.

Phase I: creating exon objects

The first phase consists in identifying when a new exon should be created, that is, on the basis of the alignment of blocks of cDNAs (transcript exons) decide whether blocks belong together and thus represent the same exon or should be considered another exon downstream.

To accomplish this task, cDNA blocks are ordered by starting coordinate of the alignment to the genome and scanned in this order. Therefore, in the 3 first phases, no alignment strand is considered and everything in increasing genomic coordinates is "downstream".

There are 5 rules for creating an exon object:

I.1 - first object created, all variables = 0 (object 0). This object is discarded at the end of the building procedure. Needed for computational purposes.

I.2 – there is no overlap of the current block with the current object (the starting coordinate of the block is higher than the maximum ending coordinate of the previous object)

I.3 - the block overlaps the current exon object (object N), is spliced upstream and the intron starts inside exon object N. That is, there are at least 2 exons inside the current object, thus characterizing an intron retention event (Figure 1)

Figure 1 - Exons are represented as bars and introns as black lines. I.3 exon object creation requires a block with an upstream intron (and exon).

When an exon object is created by rule I.3, some "corrections" must be made in the previous object because it should in fact be considered a composition of 2 objects. One important correction is to fix the maximum end coordinate of the previous object, the value should be the end of the first block (Figure 2).

After correction, the previous I.2 exon object will vanish and blocks representing such event will be annotated as exon objects I.2 and I.3.

The next 2 rules are used very rarely, only when there are sequences that

- 1- present blocks in between exons N and N+1
- 2- the intron in between N and N+1 is (possibly) retained
- 3- the sequence does not have any block overlapping exon N, in such a way that its blocks between N and N+1 are "loosen"

I.4 block is "loosen", between exons N and N+1, and presents an intron "upstream" (Figure 3). This rule can only be used after creation of an I.3 exon object.

I.5 - same as above, but block presents no intron upstream, therefore depending on the parameter `exonobj_overlap` it could be bound or not to the "upstream" exon object. If it is too far, an exon object will have to be created (Figures 4 and 5).

Algorithm for creating exon objects

```

rule I.1: create OBJECT 0
for all cDNA blocks{
    rule I.2: if block_start > OBJECT_i_end, create new OBJECT_i+1
    rule I.3: if block is inside borders of OBJECT_i and has an upstream
intron inside borders of OBJECT_i, create new OBJECT_i+1. When creating an
I.3, look for "loosen" blocks
        update maximum end of OBJECT_i as the start of intron of block
    rule I.5: if block is inside a retained intron and has NO
upstream intron and is too far from previous exon object
    rule I.4: if block is inside a retained intron and has upstream intron
        if block_end > OBJECT_i_maxend: update maximum end of OBJECT
        if block_start < OBJECT_i_minstart: update minimum start of OBJECT
}

```

In this phase, all objects are numbered from 1 to n (global numbers), in increasing genomic coordinates.

Updating exon objects borders

The algorithm displayed above omits details on how borders are updated. Actually, both the start or end of the objects are updated for minimum and maximum, respectively, using all blocks until an up or downstream spliced block is found. After this, only spliced blocks are allowed to modify the minimum and maximum, since there is more precise evidence of the exon object borders (Figure 6).

Finally, a minimum start that overlaps a previous exon object cannot update the minimum. This case will be treated as an intron retention (Figure 7).

The use of unspliced cDNA blocks is limited

If the user allowed completely unspliced cDNAs to create exon objects (parameter `exonbj_usenotspliced=true`), they will only be used to create I.2 objects, but not I.3. This happens because I.3 objects are evidences that the last object created contained a retained intron. Therefore, the cDNA block that creates an I.3 object must necessarily bear an upstream exon (and intron) and consequently it is a spliced cDNA.

It may happen that internal exons are only evidenced by cDNA blocks with no upstream/downstream (in relation to the genome) intron. For example, the first and last exons are necessarily not defined by upstream/downstream introns. The default behavior is to allow such exon objects to exist. If the user does not want to trust these exons, he/she may require a minimum number of upstream/downstream (in relation to the genome) introns (parameter `exonobj_splicedborders > 0`).

If an I.2 object is created after another I.2 object, all blocks that could have bound to the previous object were tested and therefore the maximum end coordinate of the previous object was updated. However, when an I.3 object is created, the previous exon object created is no longer examined and there is no guarantee that the

maximum ending coordinate is the maximum possible. In other words, as in PHASE I blocks are ordered by their starting coordinates, once an object is created, all the previous objects are not looked in this phase anymore. To fix this, in PHASE II, all blocks are compared again to the exon object to which it is bound and may update the maximum ending coordinate, if they are spliced downstream.

If an exon object has an I.2 object downstream, it may have the contribution of unspliced cDNAs to its maximum ending coordinate. Otherwise, if an I.3 object is just downstream, it limited the I.2 object to its intron start.

The table below reports whether unspliced cDNAs contribute or not to the maximum ending coordinate of exon objects created by different rules. Note that for alternative splice sites only those blocks that bear an upstream or downstream intron, depending on the splice site, are considered. I.4 and I.5 objects are treated as I.2.

	I.2 (with an I.2 downstream)	I.2 modified by I.3	I.3 (with an I.2 downstream)
start	contributes	contributes	never contributes
maxend	may contribute if exonobj_usenotspliced=true	never contributes	may contribute if exonobj_usenotspliced=true
comment	this exon has free start and ends	this exon object has a free start but is limited by the intron that allowed the creation of the I.3 exon or by any intron upstream the I.3 object	the start of this object depends on the existence of an intron upstream.

Phase II: reassigning blocks to objects

In the second phase, all blocks are compared to all objects created in the previous phase. This process is necessary because, as PHASE I is left to right (increasing start coordinates), when an exon object is created by rule I.3 (reporting a previous object with intron retention) the block that caused the object creation cannot be assigned to the previous object too. Therefore, this is done in PHASE II. The second function of this phase is to count the number of cDNAs that overlap the object, the number of blocks with spliced borders, etc that will be used as selection criteria in Phase III.

Phase III: selecting exon objects

Accepted exons are those that fulfill parameters given by the user, as:

- minimum number of cDNAs representing an exon
- the minimum number of spliced cDNAs that an object has to have
- minimum number of blocks defining intron to create I.3 objects

If I.3, I.4 or I.5 objects are rejected, they must be bound to the exon object “upstream”. All information on this exon object must then be considered as part of the exon object to which it was joined. There are some gory details: as there are transcripts with an "intron" in the middle, one can not considered its internal borders for the splice borders of the exon objects, so they must be identified and treated specially.

Rejected I.2 objects are simply deleted.

Phase IV: binding blocks to exon objects

As exon objects present in Phase I are not necessarily present after filtering in Phase III, it is necessary to reassign blocks to the exon objects. For example, if an exon object $N+1$ created due to the existence of 1 single cDNA defining an intron is filtered out, block B will not bind anymore to N AND to $N+1$, but only to N .

Two annotations are possible:

i-the block considered only overlaps one object

ii-the block overlaps more than one object and in this case is assigned to all the objects it overlaps

The comparison is made with the minimum start and maximum ending coordinates of the blocks belonging to the object, (see PHASE I).

Phase V: analyzing alternative splicing in the cluster

Exon skipping

Exon skipping is detected by simply verifying all sequences, whether they are bound to exon object n , skip $n+1$ to $n+m$ and are also bound to $n+m+1$. In this case, they are considered evidence of exon object skipping and are registered as such. Information is attributed to all exon objects involved. Finally, an exon skipping event is created and coded as $(n+1, n+m)$, that is, the first and last skipped exons separated by comma.

Intron retention

Intron retention has 2 rounds of identification, one to identify sequences that retain an intron and to generate such event objects and a second round to identify sequences that defined that intron. All blocks are scanned and tested whether they bind to > 1 exon object. In such case, the exon objects are updated and an intron retention event is registered, coded similarly to exon skipping, but with a different meaning: $(n, n+1)$ meaning that intron between exon objects n and $n+1$ was retained.

To identify non-retaining blocks, the block must bind to an object n and the next block under the same accession number must bind to $n+1$, defining an intron.

By default, only retained introns bordered by introns, up and downstream, are considered, this decreases the probability of the retained intron being in fact a non-spliced transcript. This behavior can be changed when building the ASC DB.

Alternative splice sites

Alternative splice sites are identified by first putting all N coordinates of an exon in a hash. The keys are the coordinates. If the exon object presented > 1 coordinate, events must be created.

Implemented in AlternativeSplicing.pm.

Cluster orientation (plus/minus strand)

The orientation of a cDNA cluster is given only by its mRNA molecules present. If no mRNA belongs to the cluster, no orientation assignment is made. For those that present mRNAs, a simple "majority wins" rule is used to assign orientation.

Exon objects are always represented in relation to the genome, in increasing coordinates. When a cluster is in the minus strand, the exon objects *lnumbers* are reversed.

Exon object numbering systems

Three numbering systems exist and they are important in different occasions. The first numbering system is the genomic, which always increases from 1 to n and only exists during the entire database building procedure. The two next numbering systems are valid for database access.

To keep the concept of object consistent throughout the database, we provide two numberings for exon objects, one "global" to that cluster (*gnumber*) and another "local" that only has a meaning for the tables generated under a particular set of parameters given by the user (*lnumber*). The global numbering system enumerates exon objects for comparing different databases ***that were built upon the same input, but with different parameters for selecting valid exon objects.*** This way, if an exon object is filtered out in one database it will have its *gnumber* missing. For different inputs, the global numbering system may not work, since some cDNA blocks that would create certain exon objects may not be present at all.

For example, the user created 3 different ASC DBs for cluster number 88620 using the same input. For parameter set I he obtained 10 exon objects (A to J), for set II 8 (A, D to J) and for set III 5 (A, G to J). The twin numbering system will mark exon object J as number 10 in all sets ("global"), even if it is the 8th exon object in II and the 5th exon object in III. The "local" numbering will mark exon objects in set I from 1 to 10, in set II from 1 to 8 and in set III from 1 to 5, thus making exon object J "local" numbers 10, 8 and 5, respectively.

This way, it is possible to compare the same exon object across different databases.

Numbering minus clusters

The local numbering system is reversed for minus clusters and enumerates consecutively only those exon objects that are present in the final database.

For minus clusters, the *gnumber* is in reverse order of the *lnumber*.

Implemented in BuildExonObjects.pm.

CDS annotation

The goal of CDS annotation in ASC is to link exon coordinates to coding region coordinates. Therefore, CDS coordinates are searched within mapped exons and annotated to ASC.

Only exon objects that are represented by mRNA can be annotated as overlapping a CDS or not. Therefore, an exon object upstream the 5'UTR of a given mRNA that does not bind the mRNA will not be considered as located at its 5'UTR and no annotation will be provided.

Details in conversion of CDS coordinates from cDNA to genomic can be found in the method [Sequence->getCDS_gnm](#).

It may happen that the exon that contains the starting or ending exon coordinate of the CDS of a given mRNA was not aligned to the genome. In this case, the CDS information is useless and the corresponding converted genomic coordinates are left 0 in the MySQL table. Gbrowse will display such sequences as mRNAs without CDS. The CDS coordinates are still available through `Sequence->getCDS_gb ( )`.

Implemented in BuildExonObjects.pm.

Computational details

The database is accessed only once per cluster. Also to increase speed, all "main" data structures as @blocks and %transcripts were passed as globals, instead as function parameters. (see Wall, Christiansen and Schwartz Programming Perl (2nd Ed., pg. 543))

Data structures generated

During build-up procedure, some arrays and hashes are used to store data and objects:

- **%transcripts** - transcript information as number of blocks, type of sequence, etc
- **@mrna_strand** - this array saves polarities of mRNA alignments. In principle, as mRNAs are 5' -> 3', all mRNAs in a cluster should have the same orientation. However, as this may not be true, all polarities are saved and based on the number of + and -, the system decides if a cluster is + or -. The criterion of decision is the higher number wins. One may change the criterion by modifying the function `main->verifyStrand ( )`
- **@exonObjects** - stores hashes that represent exon objects.
- **@blocks** - this array is basically a copy of the MySQL table where cDNA alignment coordinates are stored, but it also contains some additional information.

4.ILLUSTRATION OF THE ALTERNATIVE SPLICING ANNOTATION PROCEDURE IN A SAMPLE CLUSTER

This cluster was artificially built so it contains many exquisite events and works as an exemplification of what ASC does in different situations. You can build it (almost) from scratch, so it's a nice way to learn how to use ASC. For this tutorial, I will assume you have a MySQL database located in the same computer you are running Gbrowse.

STEP 1: Create a MySQL database called mycdna_align (change noboru for your user name)

```
mysql -u root -p -e 'create database mycdna_align_map'
mysql -u root -p mycdna_align_map -e 'grant all privileges on
mycdna_align_map to noboru'
```

STEP 2: Load the MySQL dump with cDNA alignment coordinates (already clustered)

```
mysql -u noboru -p mycdna_align_map < yourpath/asc-
0.01/sample_data/sample_cluster/sample_cluster.align.dump
```

STEP 3: Copy and paste the above configuration to a file named sample.asc.config and fill in the blanks with your info.

```
# -----
# 1. configure MySQL db where clustered cDNA alignments are stored (data
source to ASC)
# -----
aligndbhost=
aligndbname=mycdna_align_map
aligndbuser=
aligndbpasswd=
```

STEP 4: Run alternative splicing annotation

```
echo hs.sample > sample_cluster.in
asc.pl -i sample_cluster.in -c sample.asc.config > sample_cluster.asc.dump
```

STEP 6: Create another database, where you are going to store ASC annotations (may even be in another host):

```
mysql -u root -p -e 'create database myascdb'
mysql -u root -p myascdb -e 'grant all privileges on myascdb.* to noboru'
```

STEP 7: Load alternative splicing annotation:

```
mysql -u noboru -p myascdb < sample_cluster.asc.dump
```

Alternative splicing annotation is stored and can be accessed through the Perl API. Now we will generate a viewer for this data, assuming you have successfully installed Gbrowse, Apache, PHP, etc in your machine.

STEP 8: Copy and paste the above configuration to a file named sample.gb.config and fill in the blanks with your info.

```
# -----
# 1. configure MySQL where clustered cDNA alignments are stored
# -----
aligndbhost=
aligndbname=mycdna_align_map
aligndbuser=
aligndbpasswd=

# -----
# 2. configure MySQL where ASC alternative splicing info is stored
# -----
ascdbhost=
ascdbname=myascdb;
ascdbuser=
ascdbpasswd=

# -----
# 3. adaptor configuration
# -----
ascdb_setname=sample;
asc_path=; # where is your asc-0.01 directory located? ex.
/home/john/myprograms

# where are you going to store your GFFs generated by gbrowse_gff.pl?
gffdbhost=
gffdbname=mygff
gffdbuser=
gffdbpasswd=
```

STEP 9: Now generate the Gbrowse files to view clusters and annotations. By running *gbrowse_gff.pl* with -install, *gbrowse_gff.pl* will automatically install the adaptor in your selected http_path and full_cgi_path.

```
gbrowse_gff.pl -i sample_cluster.in -c sample.gb.config -install >
sample_cluster.gff
```

STEP 10: Load dump to MySQL database (in the same database of alternative splicing annotation)

```
mysql -u noboru -p myascdb < sample.gbrowse.dump
```

STEP 11: Load GFF to MySQL database (another database)

```
mysql -u root -p -e 'create database mygff'
mysql -u root -p -e 'grant all privileges on mygff.* to noboru@localhost'
bp_load_gff.pl -c -d mygff sample_cluster.gff
```

Now you can access the alternative splicing annotation data by Gbrowse. Point your browser to *yourhost/ascbrowser*. Select set “sample” and look for cluster **hs.sample**.

The sample cluster contains some artificial cases that illustrate how alternative splicing annotation with ASC works. Notice that the examples are quite dramatic to better illustrate what ASC does:

Evidences for joining exon objects – the last block of sequence j and the first of sequence k could define exon objects 12 and 13, respectively. However, they are very close ($n < \text{exonobj_join}=25$) and are non-spliced, possibly representing the same exon. Therefore, they were joined, giving only one object (exon object 12).

The same happened with object 15, this time joining 3 distinct blocks that were very close ($n < 25$ nt). Note that as the block from sequence n was spliced, the other two blocks from m and o were considered as alternative splice sites.

When an intron is evidenced, even if in between two blocks that are close ($n < \text{exonobj_join}$), there is no joining (see exon objects 13 and 14 and sequence m).

Another case of not joining evidences is shown by exon objects 17 and 18. As sequence q and r present defined introns, they cannot be considered alternative splice sites and exon object 17 and 18 must exist alone.

Exon object 16, however, was formed by joining evidences n and p, yielding alternative splice sites because there was no evidence of an intron downstream n and o.

By clicking on the exon object 16, one can see the following information:

3' splice site(s):
2 seq(s) at 32000
1 seq(s) at 33020

Evidences for alternative splice sites must be spliced - even though exon object 4 was represented by 5 blocks, only 4 were considered for analyzing 5' alternative splice sites, because only these had a defined 5' border, *i.e.*,

they presented an exon "downstream".

Spliced intron retention evidences - as in the case of alternative splice sites, intron retention could also be an artifact of an unrealistically extended sequence. To avoid this, the default behavior is to only consider as intron retention those blocks that bear introns on both sides. For example, the intron between exon objects 2 and 3 is not considered retained by sequence f, since it is not spliced on both sides and could therefore be a genomic contamination. However, the user is left with the option of setting parameter `allow_IR_nonspliced=true`, and thus consider such evidences.

Groups of evidences of intron retention - the intron defined by exon objects 2 and 3 is retained by 4 spliced blocks and defined by 2 sequences. However, due to the alternative splice site that shortens/elongates exon object 2, evidence of intron retention exists in 2 groups (when grouped by the intron defining sequence): one containing sequences b,s/a,g (introns in b,s are only retained by a,g) and another with c/d,e (intron in c is retained by d,e in addition to a,g in the first group).

This information is provided by clicking on exon object 2 or 3.

Evidence groups for IR 2,3 in exon 2 of hs.sample

evidence group ID	intron defining evidences	intron retaining evidences
1	s:6-14 b:7-15 c:9-16	a:8 g:11
2	c:9-16	e:13 d:12

Pairwise evidences of alternative splice sites – in the case above, a block that shows intron retention can also be considered an alternative splice site. That is, when it does not overlap all blocks bound to the up or downstream exon object. One of such blocks might bear an alternative splice border. Exon objects 2 and 3 are examples of this.

In such case, a pairwise comparison would consider the intron retaining blocks of sequences d and e an alternative splice site in relation to sequences b and s, and therefore it is also annotated as such (exon object 3).

3' splice site(s):

1 seq(s) at 4100 | only pairwise: s:6 b:7

1 seq(s) at 4250 | only pairwise: s:6 b:7

2 seq(s) at 5900

1 seq(s) at 6000

Partial intron retention - when a block spans *almost* an entire intron, it can still be considered an intron retention event by setting parameter `block_overlap > 0`. If `allow_IR_nonspliced` is set to true partially retaining blocks will be considered as retaining introns. However, if `allow_IR_nonspliced` is set to false, this will only be effective in weird cases where the partial intron retention occurs in conjunction with an alternative splice site of the "downstream" exon as in sequences b and f.

In the sample cluster, setting `block_overlap = 0` will cause exon object 7 to be extended until 15,999 (transcript d), since this block was not considered to overlap the next object (8) that starts at 16,000. Setting `block_overlap > 1` will cause the block overlap the next exon object and therefore it is not an alternative border of 7 anymore.

`block_overlap = 0`

`block_overlap > 1`

Reciprocal retentions and conceptual exon objects - exon objects 5, 6, 7 and 8 show examples of cases where there are reciprocal intron retentions. Sequences a and b reciprocally retain introns 5,6 and 6,7. The sequence in between the two defined introns is therefore conceptually considered an exon (exon object 6), even though no sequence has explicitly presented it. This is denominated a "conceptual exon".

Likewise, exon object 7 was created because of evidence provided by c and e.

In principle, even a conceptual exon object of length 1 could exist, if there is evidence for it. However, we suppose such events are inexistent or very rare, but a formal solution for such cases is provided in the present modeling, as there is no better conceptual way to deal with them.

Conceptual retentions - the intron defined by exon objects 9 and 10 does not have the complications described above. However, when comparing sequence i (sequence i has no upstream defined border) to the annotated exon objects, an intron retention event is detected but not according to pairwise alignment (at least when using `block_overlap=0`). In this situation, due to the existence of evidence for the 2 exons in separate (9 and 10) the intron retention event conceptually exists and is noted as such.

For these cases, no intron defining sequence will be provided.

+ intron retention:

1 seq(s) joining exons 9,10

No seq defines intron 9,10 (*conceptual retention*)

Other situations where a conceptual retention are recorded are these:

Súmula curricular

Nome: Noboru Jo Sakabe
Local e data de nascimento: São Paulo, 19 de abril de 1977

formação acadêmica

2002 - 2006

Doutoramento direto em Bioquímica e Biologia Molecular no Departamento de Bioquímica do Instituto de Química da USP/Laboratório de Biologia Computacional do Instituto Ludwig de Pesquisa sobre o Câncer

1996 - 2002

Bacharelado em Biologia no Instituto de Biociências da USP

cursos de bioinformática

“Ideas avanzadas em filogenia” - Instituto de Estudios Avanzados - IDEA, Caracas, Venezuela - Maio, 13-19, 2004

“Bioinformatics International Course ICGEB/LNNC” - Laboratório Nacional de Computação Científica, Petrópolis, RJ, Brasil - Maio, 10 a 21, 2003

bolsas recebidas

doutorado direto

Agosto 2002 – Janeiro 2007

Instituição: FAPESP

Projeto: Estudo computacional de elementos reguladores de splicing alternativo em larga escala e identificação de novos elementos candidatos (FAPESP 02/05289-4).

Supervisor: Dr. Sandro José de Souza - Laboratório de Biologia Computacional, Instituto Ludwig de Pesquisa sobre o Câncer

iniciação científica

Março/2001-Agosto/2002

Instituição: Universidade de São Paulo

Projeto: Desenvolvimento do software "Estrutura de proteína".

Supervisor: Dr. Bayardo B. Torres - Grupo de Educação em Bioquímica, IQ-USP

Agosto/2000-Julho/2001

Instituição: FAPESP

Projeto: Desenvolvimento de experimentos de bioquímica de baixo custo para cursos de graduação (FAPESP 00/05830-1).

Supervisor: Dr. Bayardo B. Torres - Grupo de Educação em Bioquímica, IQ-USP

Março 1997-Março/2000

Instituição: FAPESP

Projeto: Anestésicos locais: síntese, propriedades estruturais e físico-químicas e interações com membranas modelo e biológicas (FAPESP 98/00348-5).

Supervisora: Dr. Shirley Schreier - Laboratório de Biologia Estrutural, IQ-USP

prêmios

Melhor artigo sobre "A influência do Portal Periódicos CAPES na pós-graduação brasileira" em concurso nacional promovido pela CAPES, categoria aluno de doutorado. Trabalho: "Análise da cobertura do Portal Periódicos da CAPES baseada na disponibilidade de artigos citados em publicações internacionais de alto nível", autor: Noboru Jo Sakabe - Agosto, 2006, Brasil.

Prêmio do programa MCT-PAPED pelo software "Estrutura de proteínas", versão 1.3, autores: Noboru Jo Sakabe, Guilherme A. Marson, Bayardo B. Torres – Novembro, 2005, Brasil

Menção honrosa da Associação Brasileira de Bioinformática e Biologia Computacional pelo pôster "Analysis of low and high frequency intron retention events and the role of cis-regulatory splicing elements", autores: Sakabe, N.J. e de Souza, S.J., apresentado na 1st Conference of the Brazilian Association for Bioinformatics and Computational Biology - Outubro, 4 a 7, 2005 – Brasil

Bolsa para viagem da International Society for Computational Biology para participar do congresso Intelligent Systems on Molecular Biology 2005 em Detroit, EUA – Julho, 2005

Bolsa para viagem da United Nations University – BIOLAC para participar do "Practical course on phylogeny" - Instituto de Estudios Avanzados, Caracas, Venezuela - Maio, 13 to 19, 2004

softwares

ASC: Automatic Annotation of Alternative Splicing on cDNA Clusters

Autores: Sakabe, N.J., Galante, P.F., de Souza, S.J. Disponível em: <http://www.compbio.ludwig.org.br/asc>

Tutorial on-line sobre estrutura de proteínas

Autores: Sakabe, N.J., Marson, G.A., Torres, B.B. Disponível em: <http://www.iq.usp.br/wwwdocentes/bayardo/proteina>

publicações

1. *Sakabe, N.J.* and de Souza, S.J. (2007) Sequence features responsible for intron retention in human. BMC Genomics 8:59.
2. Vibranovski, M.D.*, *Sakabe, N.J.**, de Souza S.J. (2006) A possible role of exon-shuffling in the evolution of signal peptides of human proteins. FEBS lett., 580(6):1621-1624. (*contribuição igual)
3. Vibranovski, M.D., *Sakabe, N.J.*, de Oliveira, R.S., de Souza S.J. (2005) Signs of ancient and modern exon-shuffling are correlated to the distribution of ancient and modern domains along proteins. J. Mol. Evol. 61(3):341-50
4. *Sakabe, N.J.*, Vibranovski, M.D., de Souza S.J. (2005) A bioinformatics analysis of alternative exon usage in human genes coding for extracellular matrix proteins. Genet. Mol. Res. 3 (4):532-544
5. Galante P.F.A., *Sakabe N.J.*, Kirschbaum-Slager N., and de Souza S.J. (2004) Detection and evaluation of intron retention events in the human transcriptome RNA: 10(5):757-765.
6. *Sakabe N.J.*, de Souza J.E., Galante P.A., de Oliveira P.S., Passetti F., Brentani H., Osorio E.C., Zaiats A.C., Leerkes M.R., Kitajima J.P., Brentani R.R., Strausberg R.L., Simpson A.J., de Souza S.J. (2003) ORESTES are enriched in rare exon usage variants affecting the encoded proteins. C R Biol. (10-11):979-85.

capítulo de livro

N. J. Sakabe & B.B. Torres (2005) "Trabalhando com enzimas" (p103-117) in Fisiologia Prática. R. Curi, J. Procópio, L.C. Fernandes, Organizadores. Ed. Manole, Barueri, SP. ISBN: 85-204-1621-7.