**DIVERSIDADE NA ARQUITETURA E EXPRESSÃO GÊNICA:
UMA ANÁLISE QUANTITATIVA DE *EXON SHUFFLING* E
SPLICING ALTERNATIVO**

Fabio Passetti
Dissertação de Mestrado

Orientador: Dr. Sandro J. de Souza

SÃO PAULO
22 de maio de 2002

“Diversidade na Arquitetura e Expressão Gênica: Uma Análise Quantitativa de *Exon Shuffling* e *Splicing* Alternativo”

FABIO PASSETTI

***Dissertação de Mestrado submetida ao Instituto de
Química da Universidade de São Paulo como parte dos
requisitos necessários à obtenção do grau de Mestre em
Ciências – Área: Bioquímica.***

Aprovado por:

Prof. Dr. SANDRO JOSÉ DE SOUZA
IQ – USP
(Orientador e Presidente)

Prof. Dr. SERGIO VERJOVSKI DE ALMEIDA
IQ – USP

Prof. Dr. PAULO MARINHO DE ANDRADE ZANOTTO
ICB – USP

SÃO PAULO
20 DE JUNHO 2002.

BIBLIOTECA
INSTITUTO DE QUÍMICA
Universidade de São Paulo
60.128

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA

**DIVERSIDADE NA ARQUITETURA E EXPRESSÃO GÊNICA:
UMA ANÁLISE QUANTITATIVA DE *EXON SHUFFLING* E
SPLICING ALTERNATIVO**

Fabio Passetti
Dissertação de Mestrado

Orientador: Dr. Sandro J. de Souza

SÃO PAULO
22 de maio de 2002

DEDALUS - Acervo - CQ

30100004820

Ficha Catalográfica

Elaborada pela Divisão de Biblioteca e
Documentação do Conjunto das Químicas da USP.

P287d Passetti, Fabio
Diversidade na arquitetura e expressão gênica : uma
análise quantitativa de exon shuffling e splicing alternativo
/ Fabio Passetti. -- São Paulo, 2002.
86p.

Dissertação (mestrado) – Instituto de Química da Universidade
de São Paulo. Departamento de Bioquímica.
Orientador : Souza, Sandro José de

1. Transcrição genética 2. Evolução gênica I. T. II.
Souza, Sandro José de, orientador.

574.873223 CDD

À minha família.

AGRADECIMENTOS

Inicialmente, ao Dr. Ricardo Renzo Brentani pela oportunidade de estar no Instituto Ludwig de Pesquisa sobre o Câncer. Agradeço também a todos os funcionários e estudantes deste instituto, pela ajuda em todas as horas.

À Pós-graduação do Instituto de Química da USP.

Ao Dr. Sandro J. de Souza, por me deixar participar de seu grupo de pesquisa e me aventurar numa nova linha de pesquisa na qual eu não tinha prévia experiência. Obrigado pela confiança.

Ao Paulo Sérgio Lopes de Oliveira, pela amizade e pelo apoio profissional.

Aos membros do Laboratório de Biologia Computacional, em especial ao Maarten Rudolph Leerkes e à Helena Brentani Samaia, pelo excelente convívio e amizade. Não podendo também me esquecer do constante auxílio técnico de Jorge Estéfano de Souza, Ítalo Migotto e Daniel Furtado Simão.

À Patrícia Sávio de Araújo Souza que me apoiou de forma incondicional ao longo do desenvolvimento desta dissertação com muito carinho e companheirismo.

Àqueles que me iniciaram e apoiaram na vida acadêmica: Carmem Sílvia Fontanetti Christofolletti, Maria Isabel Camargo Mathias, Edison Zefa, Marcelo Bento Soares, Vera da Costa Soares e Maria de Fátima Bonaldo.

À FAPESP, pelo apoio financeiro;
Muito obrigado.

Sumário

Lista de Figuras	2
Lista de Tabelas.....	3
Lista de Abreviaturas	4
I. Resumo	5
II. Introdução.....	6
III. Materiais e métodos.....	18
III.a. <i>Exon shuffling</i>	18
III.b. <i>Splicing</i> alternativo	23
IV. Resultados.....	30
IV.a. <i>Exon shuffling</i>	30
IV.b. <i>Splicing</i> alternativo	44
V. Discussão	53
V.a. <i>Exon shuffling</i>	53
V.a. <i>Splicing</i> alternativo.....	59
VI. Conclusões	69
VII. Referências bibliográficas.....	70
VIII. Anexos.....	79
VIII.a. <i>Curriculum vitae</i>	79
VIII.b. Trabalhos publicados.....	82

Lista de Figuras

Figura 1. Representação esquemática da estrutura da base de dados relacional exonDB.	20
Figura 2. Esquema de uma matriz binária hipotética.	25
Figura 3. Representação esquemática da criação das variantes de <i>splicing</i> a partir das matrizes binárias.	26
Figura 4. Representação esquemática da estrutura da base de dados relacional “AEU”	28
Figura 5. Frequência por tamanho de introns em ExInt.....	34
Figura 6. Distribuição dos introns ao longo das proteínas.....	35
Figura 7. Distribuição da frequência de exons de diferentes tamanhos..	37
Figura 8. Relação do número de seqüências nos clusters e taxa de UAE em mRNAs conhecidos.	48
Figura 9. Distribuição de eventos de uso alternativo de exons ao longo dos transcritos.	49
Figura 10. Número de variantes tecido-específicos por 1.000 cDNAs.....	50
Figura 11. Taxa de UAE nos clusters com crescente frequência de cDNAs tumorais (vermelho). Para cada dado encontrado em vermelho, a taxa de UAE para os clusters de mesma média de tamanho foi determinada (azul) e somente considerando seqüências normais.	52

Lista de Tabelas

Tabela 1. Descrição de exons e introns de ExInt.	32
Tabela 2. Distribuição das fases dos introns.....	33
Tabela 3. Associação das três fases dos introns em seis classes de simetria exônica.	39
Tabela 4. Associação das três fases dos introns em duas classes de exons: simétricos e assimétricos.....	40
Tabela 5. Introns preditos <i>in silico</i> confirmados por ESTs.....	42
Tabela 6. Estatística para eventos de UAE.....	47

Lista de Abreviaturas

AEU: *Alternative exon usage*; UAE.

cDNA: *Complementary deoxyribonucleic acid*; Ácido desoxirribonucleico complementar.

CDS: *Coding sequence*; região ou sequência codificante.

ESE: *Exon splicing enhancer*; *enhancer* de *splicing* de exon.

ESS: *Exon splicing silencer*; silenciador de *splicing* de exon.

EST: *Expressed Sequence Tags*; marcadores de sequências expressas.

ISE: *Intron splicing enhancer*; *enhancer* de *splicing* de intron

mRNA: *Messenger ribonucleic acid*; Ácido ribonucleico mensageiro.

PC: *Personal computer*; computador pessoal.

PVM: *Parallel virtual machine*; máquina virtual paralela.

UAE: Uso alternativo de exons.

UTR: *Untranslated region*; região não traduzida.

I. Resumo

A função e a arquitetura dos genes está começando a ser elucidada a partir do estudo de genomas completos tanto de procariotos como de eucariotos. Diversos estudos foram ultimamente realizados a respeito de *exon shuffling* do ponto de vista evolutivo, fenômeno relacionado à origem de novos genes através de recombinações de DNA mediadas por introns. Apesar de eventos de *exon shuffling* serem responsáveis pelo aumento da modularidade gênica, outros processos foram desenvolvidos ao longo da evolução para que houvesse o aumento da diversidade do proteoma sem a conseqüente expansão dos genomas, sendo *splicing* alternativo um dos mais freqüentes. Apresentamos nesta dissertação duas extensivas análises: 1) a análise de uma base de dados de genes eucarióticos contendo pelo menos um intron que apresentou excesso de introns de fase 0 e exons simétricos, dados que suportam *exon shuffling* como um importante mecanismo de evolução gênica. Avaliamos também a confiabilidade de introns preditos por programas de computador através de alinhamento de ESTs; e 2) a análise do uso alternativo de exons (UAE), um tipo de *splicing* alternativo, em transcritos humanos detectando que cerca de 51% dos genes humanos possuem mais de uma variante de *splicing* e que este tipo de processamento pós-transcricional parece ser mais freqüentemente encontrado em tecidos tumorais.

II. Introdução

Os bancos públicos de seqüências cresceram exponencialmente nos últimos anos em função da contínua adição de novas informações por meio dos projetos genoma e do seqüenciamento de *Expressed Sequence Tags* (ESTs). A automatização do seqüenciamento de proteína e DNA, conjuntamente com o desenvolvimento de diferentes abordagens para a obtenção de genomas completos e seus transcriptomas, fez da década de 1990 um período de prosperidade para a biologia molecular. Ao passo que estes bancos de dados cresceram ao longo dos anos, as suas análises tornaram-se cada vez menos práticas sem o auxílio computacional. Uma eficiente abordagem para a manipulação dessa enorme e crescente quantidade de dados, tornando-os mais acessíveis aos biólogos, pode ser realizada através da utilização de ferramentas de bioinformática. A associação da rapidez do processamento de dados oferecido pela computação ao conhecimento molecular dos biólogos mostra-se fundamental para que sejam realizadas análises de fenômenos biológicos.

O seqüenciamento dos genomas de *C. elegans* (Waterston *et al.*, 1997), *S. cerevisiae* (Mewes *et al.*, 1997), *D. melanogaster* (Adams *et al.*, 2000) e *H. sapiens* (Venter *et al.*, 2001; Lander *et al.*, 2001) tornou

acessíveis os primeiros genomas completos de eucariotos. Essas informações são imprescindíveis para que sejam estudadas as diversas formas através das quais um gene pode ser transcrito e quais as suas funções.

Alguns mecanismos de evolução molecular já estão devidamente comprovados pela sua importância no aumento da modularidade gênica, dentre os quais encontra-se *exon shuffling*. O conceito de *exon shuffling* foi proposto por Walter Gilbert, em 1978, como um processo de recombinação de exons mediado por seqüências intrônicas. A descoberta, um ano antes, da estrutura interrompida dos genes eucarióticos (revisto por Sharp, 1994) levou Gilbert a sugerir que a presença de introns (em média dez vezes maiores que os exons) agilizaria o processo de formação de novos genes. Assim, recombinações dentro dos introns levariam à formação de genes em mosaico, constituídos de exons encontrados em outros genes (Gilbert, 1978; 1987).

Alguns genes tiveram a sua origem relacionada a *exon shuffling*. O gene do fator de crescimento *EGF* e o gene para o receptor de *LDL* compartilham um conjunto de oito exons no homem e sua origem ocorreu através de *exon shuffling* (Südhof *et al.*, 1985). Outro gene originado por *exon shuffling* é o *Jingwei* em *Drosophila melanogaster*. Long & Langley (1993) demonstraram que sua origem se deu através de uma retroposição de um transcrito do gene *Adh* em outro cromossomo. Originariamente, o gene

de *Adh* está presente no cromossomo 3. Entretanto a incorporação do DNA complementar (cDNA) ocorreu num intron no cromossomo 2, formando assim um novo gene.

Os mecanismos moleculares que levam a *exon shuffling* estão sendo progressivamente elucidados (revisto por Long, 2001). Um deles é através de transcriptases reversas que produzem um DNA complementar (cDNA) a partir de um RNA mensageiro (mRNA) e, logo em seguida, ocorre a incorporação do mesmo em alguma parte do genoma. Se esta incorporação acontecer dentro de um intron e esta proteína for funcional, rapidamente foi criado um novo gene. Uma vez que esses processos decorrentes de transcriptases reversas são os responsáveis pela criação de pseudogenes, *exon shuffling* torna-se extremamente dependente da região genômica em que ocorrer a inserção. Outro mecanismo já descrito que pode promover *exon shuffling* é o de recombinação ilegítima. O gene αA -cristalino em hamster possui na sua arquitetura uma duplicação de um exon interno, originada através da ação da topoisomerase I (van Rijk *et al.*, 2000). Esta enzima reconhece preferencialmente um sítio de três nucleotídeos (CAT) e excisa a fita de DNA, promovendo assim recombinação entre regiões com pouca ou nenhuma identidade. Elementos repetitivos podem também aumentar a taxa de *exon shuffling* nos organismos. Por meio deste

mecanismo denominado transdução, partes de um gene ou um gene completo podem ser incorporados em outras partes do genoma após a sua expressão. No genoma humano, encontra-se em abundância um destes elementos, o LINE. Por corresponder entre 16% (Venter *et al.*, 2001) e 20% (Lander *et al.*, 2001) do genoma humano, presume-se que o seu papel ao longo da evolução foi de extrema importância para que os exons fossem continuamente embaralhados (Long, 2001).

A medida que mais análises foram sendo realizadas, o papel das fases dos introns que flanqueiam os exons envolvidos nesse tipo de recombinação começou a ser evidenciado (Gilbert & Glynias, 1993). A fase dos introns está associada à fase de leitura, podendo ser 0, 1 e 2. Introns de fase 0 interrompem a fase de leitura entre dois códons. Já introns de fase 1 estão presentes entre o primeiro e o segundo nucleotídeo do códon e os de fase 2 entre o segundo e terceiro nucleotídeos. A associação de dois introns adjacentes a um exon pode ser classificada em dois grupos: exons simétricos (0,0), (1,1) e (2,2); e exons assimétricos (0,1), (1,0), (1,2), (2,1), (0,2) e (2,0).

A fase dos introns e a simetria dos exons então começou a ser correlacionada a um método eficaz de comprovação de que eventos de *exon shuffling* são essenciais à evolução gênica, uma vez que a inserção de exons

simétricos não perturba a fase de leitura (Patthy, 1987). Fedorov e colaboradores (1992) já haviam descrito que introns de fase 0 são encontrados mais freqüentemente do que as duas outras fases. Diversos estudos de genes presentes tanto em procariotos quanto em eucariotos desvendaram que os introns que estão presentes interrompendo a estrutura gênica são mais freqüentes por serem de fase 0 e os exons por serem simétricos do tipo (1,1). Este fato corrobora que *exon shuffling* é uma das mais importantes vias para o aumento da modularidade gênica (Long *et al.*, 1995).

Nos últimos anos tem aumentado o número de proteínas descritas cujos genes são resultantes de eventos de *exon shuffling* (van Rijk, 1999; Wang *et al.*, 2000; Jones *et al.*, 2000). Atualmente, o conceito de *exon shuffling* é amplamente aceito como um dos mecanismos mais importantes na formação de novos genes (Patthy, 1999).

O estudo de proteínas *in vitro* corroborou dados de De Souza e colaboradores (1996; 1998) e Long e colaboradores (1995) sobre a correlação entre domínios protéicos, módulos funcionais, simetria de exons e *exon shuffling* (Kolkman & Stemmer, 2001). Long e colaboradores (1995) e Deutsch & Long (1999) estimaram que pelo menos 19% dos exons estejam relacionados a *exon shuffling*.

Os genomas estão sob a influência de eventos que auxiliam o aumento da diversidade gênica, como *exon shuffling* e duplicação gênica (Ohno, 1968). Mesmo assim, existem ainda perguntas pertinentes em relação a outros tipos de controles que a evolução utilizou para que houvesse o aumento da variedade do proteoma sem o acréscimo do número de genes, sendo *splicing* alternativo um deles. Estes processos não estão ainda completamente elucidados, principalmente pelo fato de desconhecermos como ocorre o reconhecimento de todos os sítios de *splicing*. Entretanto, *splicing* alternativo é um versátil mecanismo para a regulação da expressão gênica durante a transcrição, sendo que a partir de um gene, podem ser formadas diversas proteínas diferentes.

Distintas maneiras são encontradas nos organismos para que uma isoforma seja constituída: exons podem ser retirados, introns podem ser mantidos e sítios crípticos doadores e aceptores de *splicing* podem ser usados na concepção do transcrito (Modrek & Lee, 2002). Outras formas que podem levar também à criação de múltiplas variantes de um mesmo gene são as alterações no sítio de início de transcrição e no sítio de poliadenilação, e a edição de RNA.

A diversidade nos transcritos que *splicing* alternativo cria em eucariotos pode ser uma das respostas para proteomas tão ricos em

variabilidade (revisto por Chabot, 1996; e por Smith & Valcárcel, 2000). Em alguns casos extremos, como o gene *para* em *D. melanogaster*, o número de isoformas hipotéticas formadas por *splicing* alternativo pode ultrapassar um milhão quando associada com a edição de RNA (Graveley, 2001). Em casos em que somente *splicing* alternativo é reconhecido como o único responsável pela diversidade nos transcritos, como o do gene *dscam* também em *D. melanogaster*, podem ser produzidos quase 40.000 diferentes transcritos (Schmucker *et al.*, 2000). Um caso em que a produção de transcritos é mais restrita, mas que exemplifica bem a importância de *splicing* alternativo na regulação de processos celulares, é o da neurexina. Cerca de 1.000 transcritos distintos podem ser produzidos a partir de *splicing* e promotores alternativos. Neste caso, quando um transcrito não possui um exon específico, a neurexina interage com outras moléculas para a formação de sinapses (Missler & Südhof, 1998). Desta forma, correlacionaram-se os eventos de uso alternativo de exons e mudança de função protéica de um mesmo gene (Black, 2000; Graveley, 2001).

As diversas formas em que um transcrito pode ser processado têm sido continuamente descritas na literatura (revisto por Chabot, 1996; e por Smith & Valcárcel, 2000). Durante o *splicing* de mRNAs, introns são precisamente removidos e os exons que estavam originalmente flanqueando-

o são unidos. O *splicing* é realizado através de uma partícula multi-enzimática denominada spliceossomo (um grande complexo ribonucleoprotéico contendo cinco snRNAs e diversas proteínas), que é formado sobre o pré-mRNA. Existem três sítios nos introns dos pré-mRNAs que são de fundamental importância para que este processo ocorra corretamente: o 5' (doador), o 3' (aceptor) e a sequência consenso rica em pirimidinas (Thanaraj & Clark, 2001).

A primeira sub-unidade que surge acoplada ao pré-mRNA é a pequena ribonucleoproteína (snRNP) U1 no sítio 5' do intron a ser excisado. O fator auxiliar U2AF é formado a partir de duas sub-unidades: a 65 que aloja-se na região rica em pirimidinas e a 35 que reconhece o sítio 3' do intron. Uma vez que U1 e U2AF são as moléculas que realizam o reconhecimento dos sítios onde ocorre a excisão do pré-mRNA, elas podem sofrer a influência de diversos fatores tanto *cis* como *trans*.

Os sítos 5' de *splicing* de introns são geralmente bem conservados, enquanto os sinais de *splicing* em 3' demonstram maior variabilidade. O reconhecimento destes sítios pode ser influenciado por proteínas ricas em serina-arginina (SR), uma família gênica extremamente conservada nos metazoa (Tacke & Manley, 1999). Este tipo protéico possui dois domínios funcionais: o motivo de reconhecimento de RNA (RRM) e domínio RS para

interação proteína-proteína (Cartegni *et al.*, 2002). O motivo RRM tem como função o reconhecimento de sítios específicos dentro do pré-mRNA, auxiliando no funcionamento dos domínios RS como pontes de ligação entre as snRNPs no reconhecimento de sítios de *splicing* alternativos. Há ainda outro tipo protéico que atua no reconhecimento de sítios alternativo de *splicing*, apesar de não ser tão bem descrito como as proteínas SR: as ribonucleoproteínas heterogêneas nucleares (hnRNP). Estas moléculas possuem uma estrutura muito similar às proteínas SR: um motivo de ligação com o pré-mRNA e outro que realiza interações entre proteínas. Entretanto, estas duas partículas atuam diferencialmente tanto na região do pré-mRNA onde se acoplam como na forma que auxiliam no reconhecimento de sítios de *splicing*.

Existem algumas seqüências que são amplamente conservadas entre os genes eucarióticos e têm a sua nomenclatura relacionada à sua função: silenciador de *splicing* de exon (ESS), *enhancer* de *splicing* de intron (ISE) e *enhancer* de *splicing* de exon (ESE). Acredita-se que os sítios presentes em exons (ESS e ESE) sejam reconhecidos por proteínas SR, e que as seqüências que estão presentes em introns (ISE) possuam afinidade por hnRNPs. Este reconhecimento destas seqüências por fatores específicos pode ser o responsável por padrões de reconhecimento de sítios de *splicing*.

As proteínas SR tendem a reconhecer sítios de *splicing* mais próximos da região onde se acoplaram. Entretanto, as hnRNPs atuam de forma a pular um ou mais sítios de *splicing* próximos de seu sítio de reconhecimento no pré-mRNA e determinam aqueles mais distantes (revisto por Smith & Valcárcel, 2000).

A determinação de sítios alternativos de *splicing* muitas vezes é dada através da concentração dos fatores que atuam como bloqueadores ou estimuladores para o acoplamento do spliceossomo no pré-mRNA. Apesar de somente poucos reguladores terem sido descritos em mamíferos, eles podem ser necessários para a seleção de sítios de *splicing* mais restritivos. Reguladores de *splicing* podem ser ainda necessários em casos em que a sequência ou a posição do *enhancer* não sejam muito favoráveis para a formação do spliceossomo através das snRNPs. Assim, quanto mais forte for o sinal do sítio de *splicing*, mais facilmente ele será reconhecido pela maquinaria de *splicing* e fatores como as proteínas SR e hnRNPs atuarão desta forma.

Um dos mecanismos moleculares que é aceito para a produção de uso alternativo de exons (UAE) foi resumidamente descrito por Chabot (1996). Se tomarmos um exemplo hipotético de um exon que não é constitutivamente encontrado nos transcritos de seu gene por ser flanqueado por fracos sinais

de *splicing*, então proteínas SR específicas acopladas à *enhancers* presentes dentro do referido exon favoreceriam a escolha de um sítio alternativo 5' pela snRNP U1. Assim sendo, U2AF reconheceria o sítio 3' deste exon hipotético inserindo-o num mecanismo alternativo de escolha de exons. Portanto, moléculas regulatórias estão diretamente relacionadas com o grau de reconhecimento de todos os sítios de *splicing* de um transcrito pela maquinaria do spliceossomo, sendo que usualmente elas ocupam determinados sítios bloqueando o acesso da snRNP U1, fortalecendo assim a afinidade do spliceossomo por sítios alternativos (Cáceres *et al.*, 1997). Uma molécula bem descrita na literatura que aumenta a taxa de uso alternativo de exons é a hnRNP-I ou PTB (revisto por Smith & Valcárcel, 2000). Os autores sugerem que esta proteína poderia competir por sítios de ligação da U2AF e, mais especificamente para UAE, PTB estaria diretamente relacionada atuando em grandes concentrações em exons específicos, mantendo-os no mRNA maduro. Outra proteína hnRNP, a A1, controla o *splicing* alternativo de seus próprios transcritos por sítios de ligação da proteína no pré-mRNA otimizados (Blanchette & Chabot, 1999).

A manifestação de alguns casos patológicos foi relacionada à desregulação do balanço desses fatores de regulação do spliceossomo, como no caso de tumorigênese de mama (Stickeler *et al.*, 1999). As disfunções

causadas por esta estequiometria têm levado a obtenção de marcadores tumorais resultantes de variantes de *splicing* específicos de tumor (revisto por Caballero *et al.*, 2001).

Alguns estudos em larga escala referentes a *splicing* alternativo já foram realizados nos últimos anos, culminando na formação de bases de dados específicas (Mironov *et al.*, 1999; Croft *et al.*, 2000; Hide *et al.*, 2001; Modrek *et al.*, 2001). No genoma humano, estima-se que pelo menos 35% dos genes estejam sob *splicing* alternativo. Entretanto, como as informações são provenientes de ESTs e somente um número limitado de tecidos e estágios de desenvolvimento foram analisados, há suspeita que este número seja uma sub-estimativa (Black, 2000).

Recentemente, Brett e colaboradores (2002) realizaram um estudo no qual determinaram a taxa de *splicing* alternativo em sete espécies-modelo através da normalização de genes estudados e transcritos mapeados. Os autores concluem que a taxa de *splicing* alternativo não estaria sofrendo alterações significativas dentre os organismos estudados.

Levando-se em consideração as informações relatadas acima, o nosso objetivo é a abordagem em bioinformática de eventos de dois mecanismos conhecidos que atuam diferencialmente para o aumento da diversidade protéica: *exon shuffling* no genoma e *splicing* alternativo no transcriptoma.

III. Materiais e métodos

III.a. Exon shuffling

Primeiramente foi realizado o download dos arquivos de seqüências protéicas da base de dados ExInt (<http://intron.bic.nus.edu.sg/exint/newexint/exint.html>). Ambos os arquivos apresentam as seqüências no formato FASTA e foram confeccionados a partir da versão 122 do GenBank (<http://www.ncbi.nlm.nih.gov>), sendo subdivididos em entradas que contêm pelo menos um intron predito *in silico* ou experimentalmente definido. Para cada um dos dois arquivos provenientes de ExInt adotaram-se os seguintes critérios de restrição para a retirada de seqüências redundantes dos subconjuntos, remanescendo sempre a maior: identidade acima de 25% e sobreposição mínima de 70% em relação ao tamanho da menor seqüência. Escrevemos um programa (DBPURGE) em linguagem PERL para a determinação do conjunto não-redundante, utilizando-se de um cluster de 8 PCs através de uma versão do programa FASTA para PVM (<http://ftp.virginia.edu/pub/fasta/>) para a sua realização. Criamos então uma base de dados relacional denominada “exonDB” (Figura 1) em MySQL com três tabelas: Main, Exon e Intron. Os campos na tabela Main referem-se a: número de acesso de ExInt (id_seq); espécie da qual a

seqüência é originada (species); anotação da seqüência (seq_annot); número total de introns (intron_total); número de referência da seqüência no GenBank (nid); número de acesso ao GenBank (axn); número total de exons (exon_total); tamanho da seqüência em aminoácidos (total_aa); seqüência de aminoácidos (aa_seq); tipo de origem da seqüência (status); identificador das seqüências não-redundantes (nr); e proveniência nuclear ou organelar (origin). Já a tabela Exon é composta pelos seguintes campos: número de acesso de ExInt (id_seq); simetria dos introns flanqueando o referido exon (exon_sym); tamanho do exon na seqüência de aminoácidos (exon_size); número do exon a partir da extremidade amino terminal (exon_num); seqüência em aminoácidos do exon (exon_seq); fase dos introns que flanqueiam o referido exon (exon_phase); e o tamanho do exon na seqüência de nucleotídeos (exon_length). Por fim, a tabela Intron apresenta os seguintes campos: número de acesso de ExInt (id_seq); número do intron a partir da extremidade amino terminal (intron_num); início e final do intron na seqüência de mRNA (intron_map); fase do intron (phase); posição em que se encontraria o intron na seqüência de aminoácidos (position_aa); tamanho do intron na seqüência nucleotídica (intron_length); seqüência do intron em nucleotídeos (intron_seq); e confirmação dos introns preditos *in silico* através de alinhamento de ESTs (est_conf).

Figura 1. Representação esquemática da estrutura da base de dados relacional exonDB.

A análise de confiabilidade dos limites dos introns preditos por programas ocorreu da seguinte maneira: 1) Download das seqüências nucleotídicas dos genes contendo introns preditos de ExInt a partir do GenBank; 2) Download do dbEST (versão 123) também do GenBank; e 3) alinhamento usando o programa BLAST (Altschul *et al.*, 1997) de cada gene do conjunto previamente descrito com o conjunto de cDNAs referente à espécie do gene em questão. Após o alinhamento, foi escrito um programa na linguagem PERL para a confirmação dos limites dos introns preditos nos genes e aqueles encontrados a partir do alinhamento entre a seqüência nucleotídica do gene e os cDNAs, ressaltando que os limites de cada intron encontrado através do alinhamento com cDNAs foi sempre acrescido ou decrescido em 10 pares de bases para que erros de seqüenciamento fossem minimizados no resultado da análise.

Após a confecção desta base de dados relacional, escrevemos outro programa na linguagem PERL para que os dados fossem carregados nas referidas tabelas. Usando sintaxe apropriada para MySQL, foram respondidas as perguntas sobre: frequência dos exons e sua simetria; introns e suas fases; e distribuição dos introns ao longo das proteínas. Os dados estão armazenados nos computadores do Laboratório de Biologia

Computacional do Instituto Ludwig de Pesquisa Sobre o Câncer em São Paulo – SP.

III.b. Splicing alternativo

Através de um trabalho em colaboração com os grupos de pesquisa de P.S. de Oliveira (Universidade Santo Amaro), J.P. Kitajima (Universidade de Campinas) e R.L. Strausberg (National Cancer Institute – EUA), uma base de dados relacional que armazena as informações referentes ao alinhamento de todas as seqüências de cDNA humanas no genoma humano foi confeccionada. A partir desta base de dados, criamos uma secundária, específica para determinação de uso alternativo de exons no transcriptoma humano.

As seqüências provenientes do GenBank anotadas como genômicas e maiores que 10.000 bases foram consideradas como clones genômicos. O transcriptoma humano foi definido através da versão de março de 2001 do GenBank para as subdivisões dbEST e nr. Não foram utilizadas seqüências montadas no intuito de evitarmos artefatos criados durante a sua confecção. Preparamos ainda um conjunto o qual denominamos de *full-length* que correspondem a cDNAs que não sejam ESTs. A todas as seqüências aplicamos o pacote para mascaramento de elementos repetitivos e contaminantes denominado PFP da empresa Paracel (Pasadena, CA, EUA).

Após esta primeira etapa, utilizamos o programa BLAST (Altschul *et al.*, 1997) para que os pares de seqüências de cDNA e genômicas que

possuem identidade fossem encontrados. Determinados estes pares, utilizamos o programa Sim4 (Florea *et al.*, 1998) no intuito de estabelecermos os melhores limites de exons e introns (realizado gentilmente por C.V. Jongeneel e membros de seu laboratório – Ludwig Institute for Cancer Research – Lausanne Branch). Aquelas seqüências com identidade abaixo de 95% em 30 pares de bases quando alinhados aos clones genômicos foram excluídos (valores determinados empiricamente). Assim, foram usadas as coordenadas do mapeamento para agruparmos as seqüências que possuíssem pelo menos a sobreposição de 10 pares de bases de um exon (clusters). Na tentativa de ser retirada a redundância dos clones genômicos, utilizamos o conjunto de *Golden Path* (<http://genome.ucsc.edu>).

Desenvolvemos numa segunda etapa uma nova metodologia para a representação das variantes de *splicing* para cada cluster: a construção de matrizes binárias. Para cada seqüência mapeada foi criada uma matriz binária na qual cada coluna representa um exon e cada linha corresponde a uma seqüência. Uma dada célula da matriz é numerada com “1” se o respectivo exon estiver presente, de outra forma a célula é numerada com “0” (Figura 2). Para a construção das matrizes binárias foi escrito um programa em linguagem PERL (realizado gentilmente por P.S. de Oliveira – Universidade Santo Amaro).

Matriz de Uso de Exons

	1	2	3	4
A	1	1	1	1
B	1	0	1	1
C	1	0	0	1

Figura 2. Esquema de uma matriz binária hipotética. Cada linha representa uma seqüência e cada coluna um exon.

Outro programa foi escrito usando a linguagem PERL para a retirada dos clusters redundantes a partir do seguinte grau de restrição: uma seqüência que teve o seu mapeamento em dois ou mais clusters foi utilizada para que ambos os clusters fossem comparados. Aquele cluster de menor tamanho foi eliminado e, conseqüentemente, o outro foi considerado um cluster não-redundante. Uma vez que tínhamos como objetivo o estudo de *splicing* alternativo, utilizamos somente clusters que contivessem dois ou mais exons. É importante salientar que nestes clusters considerados não-

redundantes são encontrados os clusters em que os cDNAs *full-length* foram mapeados.

A partir de todas as seqüências mapeadas e as matrizes formadas, foram geradas as variantes únicas de cada cluster. Primeiramente, todas aquelas matrizes que não apresentavam o padrão $1[0]_n1$, onde “n” significa o número de “0” entre dois “1”, foram consideradas como pertencentes a uma única variante de *splicing*. (A e B na Figura 3) As remanescentes foram agrupadas de forma que o padrão de *splicing* fosse preservado dentro de cada variante formada, como pode ser observado em C e D na Figura 3.

	I	II	III
A	1111	1111	1
B	1001	1001	2
C	1010	1011	3
D	1011	1011	3

Figura 3. Representação esquemática da criação das variantes de *splicing* (coluna II) a partir da matriz binária do mapeamento dos exons (coluna I). As isoformas estão enumeradas na coluna III.

Após todas as informações necessárias terem sido adquiridas, carregamos um banco relacional em MySQL denominado “AEU” (Figura 4) com quatro tabelas: *Full_length*, *Full_length_compl*, *Clusters* e *Clusters_compl*. A tabela *Full_length* possui os seguintes campos: número de

acesso ao GenBank que possui maior comprimento em nucleotídeos em relação às demais clusterizadas (`full_length`); clone genômico (`genomic_id`); número do cluster (`cluster`); número de acesso ao GenBank da sequência que compartilhou ao menos parcialmente um exon com a sequência selecionada no campo “`full_length`” (`gene_id`); matriz binária contendo o uso alternativo de exons (`alt_splicing`); tecido proveniente (`tissue`); orientação da sequência em relação ao clone genômico (`strand`); e a matriz binária identificadora de variantes únicas (`variant`). Na tabela `Full_length_compl` encontram os seguintes campos: `full_length`; `genomic_id`; `gene_id`; tipo de metodologia empregada (`origin`); referência sobre a presença da referida sequência em tecidos tumoral ou normal (`tumor`); e o cromossomo no qual a sequência foi mapeada (`chr_id`). Já na tabela `Clusters`, os campos são: `cluster`; `genomic_id`; `gene_id`; `alt_splicing`; `tissue`; `strand`; `variant`; e se o cluster é redundante ou não (`nr`). Na tabela `Clusters_compl` foram inseridos os campos: `cluster`; `genomic_id`; `gene_id`; `origin`; `tumor`; e `chr_id`.

Figura 4. Representação esquemática da estrutura da base de dados relacional “AEU”.

A análise estatística que aplicamos para a obtenção do valor de qui-quadrado encontrado para a Figura 11 foi realizada através do programa EpiInfo6 (Centers For Disease Control & Prevention (CDC), USA, 1997, versão 6.04b), gentilmente assessorado por P.C. Maciag (Instituto Ludwig de Pesquisa sobre o Câncer – São Paulo Branch).

Os dados aqui apresentados foram obtidos através de programas em linguagem PERL ou foi utilizada sintaxe apropriada para MySQL. Os dados estão armazenados nos computadores do Laboratório de Biologia Computacional do Instituto Ludwig de Pesquisa Sobre o Câncer em São Paulo – SP.

IV. Resultados

IV.a. Exon shuffling

Análises em larga escala sobre seqüências gênicas que possuem introns vêm sendo feitas nos últimos anos e diversos bancos de dados relacionados a este assunto foram formados, como ExInt (Sakharkar *et al.*, 2000; Sakharkar *et al.*, 2002), EID (Saxonov *et al.*, 2000) e IDB/IEDB (Schisler & Palmer, 2000). O banco ExInt, por ter sido desenvolvido pelo nosso grupo foi o utilizado para as análises, entretanto outros bancos em menor escala já foram feitos para propósitos similares (Deutsch & Long, 1999).

A partir de ExInt, desenvolveu-se uma base de dados relacional em MySQL denominada “exonDB” que possibilitou que as análises fossem realizadas mais rapidamente.

A atual versão da base de dados (Sakharkar *et al.*, 2002) contém 120.573 entradas, sendo 87.429 contendo introns preditos *in silico* e 33.144 contendo introns definidos experimentalmente. Os introns encontrados *in silico* correspondem a estruturas gênicas preditas por programas de computador como GRAIL (Uberbacher & Mural, 1991) ou GENESCAN (Burge & Karlin, 1997). Outra importante subdivisão foi a definição da

localização celular dos genes em nuclear (119.069 entradas) ou organelar (1.504 entradas).

Utilizando uma adaptação da metodologia empregada por Long e colaboradores (1995) para a retirada de seqüências redundantes provenientes de genes homólogos e variantes de *splicing* de bases de dados, obtivemos um subconjunto de ExInt não-redundante. Dos arquivos originais de ExInt, restaram 21.079 entradas (17,5%), sendo que 1.716 (1,4%) contêm introns definidos experimentalmente e 19.363 (16,1%) contêm introns preditos (Sakharkar *et al.*, 2002). A completa descrição de entradas para cada espécie, de exons e de introns pode ser encontrada na Tabela 1 (Sakharkar *et al.*, 2002).

Tabela 1. Descrição de exons e introns de ExInt.

	Entradas no GenBank	Exons	Introns
ExInt total	94.615	518.169	525.870
ExInt não-redundante	15.271	113.457	128.065
<i>R. norvegicus</i>	835	4.889	7.191
<i>H. sapiens</i>	8.287	60.499	43.127
<i>M. musculus</i>	3.044	18.920	15.407
<i>D. melanogaster</i>	15.220	64.271	89.969
<i>C. elegans</i>	18.924	121.708	108.803
<i>A. thaliana</i>	25.216	158.629	127.386
<i>S. cerevisiae</i>	589	1.695	1.438

A correlação entre a fase dos introns, a estrutura gênica e os eventos de *exon shuffling* tem sido extensivamente demonstrada através de análises em larga escala (Long *et al.*, 1995; de Souza *et al.*, 1998; Deutsch & Long, 1999). Foi demonstrado por Long e colaboradores, em 1995, que para análises sobre fase de introns e simetria de exons a retirada de redundância das seqüências não é necessária, o que justificou a utilização do conjunto redundante de seqüências para as análises aqui descritas. Para um estudo mais detalhado do comportamento de exons e introns nos genes, dividimos a

análise estatística em sete espécies modelo: *Rattus norvegicus*, *Mus musculus*, *Caenorhabditis elegans*, *Homo sapiens*, *Arabidopsis thaliana*, *Drosophila melanogaster* e *Saccharomyces cerevisiae* (Tabela 2).

Tabela 2. Distribuição das fases dos introns

	0	1	2
ExInt total	257.713 (49%)	147.625 (28%)	120.532 (23%)
ExInt não-redundante	60.979 (48%)	35.438 (28%)	31.608 (24%)
<i>M. musculus</i>	6.703 (44%)	5.921 (38%)	2.783 (18%)
<i>C. elegans</i>	51.251 (47%)	28.553 (26%)	28.999 (27%)
<i>H. sapiens</i>	19.102 (44%)	15.423 (36%)	8.602 (20%)
<i>A. thaliana</i>	71.958 (56%)	28.178 (22%)	27.250 (22%)
<i>D. melanogaster</i>	38.101 (42%)	28.896 (32%)	22.972 (26%)
<i>R. norvegicus</i>	2.842 (39%)	2.365 (33%)	1.384 (28%)
<i>S. cerevisiae</i>	641 (45%)	428 (30%)	369 (25%)

A frequência do tamanho dos introns foi descrita primeiramente por Deutsch & Long (1999) como bimodal. Analisando esta frequência no banco de ExInt, também a encontramos para o conjunto completo de ExInt (Figura 5). Encontramos este mesmo padrão bimodal tanto para o conjunto completo de ExInt, como para o não-redundante.

Figura 5. Frequência por tamanho de introns em ExInt. Todos os introns de ExInt em preto, conjunto não-redundante em vermelho e introns experimentalmente definidos em amarelo.

A distribuição e frequência dos introns ao longo das proteínas (Figura 6) é relevante tanto para estudos de padrão de *splicing* alternativo como para o conhecimento de como se comportam os introns na maioria dos genes. Num estudo sobre *splicing* alternativo em transcritos humanos, Modrek e colaboradores (2001) encontraram a tendência destes eventos ocorrerem na metade C-terminal das proteínas. Assumindo isto, demonstramos que há uma maior incidência de introns na metade C-terminal das proteínas.

Figura 6. Distribuição dos introns ao longo das proteínas (sentido amino- para carboxi-terminal). Em preto fase 0, em vermelho fase 1 e em amarelo fase 2.

O número de exons (Figura 7) do conjunto não-redundante é de 150.353 (pico em 33 resíduos de aminoácidos) e para o conjunto redundante é de 518.169 (pico em 35 resíduos de aminoácidos). A média de exons por gene para o conjunto total de ExInt foi de 5,48, sendo que o gene com o maior número de exons possui 96, um gene definido *in silico* (número de acesso ao GenBank AL445795).

Figura 7. Distribuição da frequência de exons de diferentes tamanhos para os conjuntos redundante (preto) e não-redundante (vermelho).

Há uma grande discussão na literatura sobre a ausência ou presença de introns no organismo ancestral de eucariotos e procariotos. A linha *introns-late* sustenta que os introns foram inseridos ao longo da evolução dos genomas eucariotos (Palmer & Logsdon, 1991; Stoltzfus *et al.*, 1994; Logsdon *et al.*, 1995). Em contrapartida, outros autores defendem que os introns estavam presentes no genoma dos primeiros organismos (*introns-early*). Estando os introns presentes nestes organismos, eventos de *exon shuffling* poderiam ter acelerado a formação de novos genes (Gilbert, 1987; Gilbert & Glynias, 1993; Long *et al.*, 1995; de Souza *et al.*, 1996).

Exons simétricos, por serem inseridos em outros introns sem interromper a fase de leitura, são de extrema importância para a detecção de eventos de *exon shuffling* em grandes bases de dados. Realizamos então duas distintas análises estatísticas: a primeira leva em consideração as seis possíveis combinações entre as fases dos introns e a segunda distingue somente os exons simétricos dos assimétricos. Como pode ser observado na Tabela 3, os três tipos de exons simétricos possuem excesso acima do esperado e os exons (0,0) apresentaram-se mais abundantemente dentre todas as classes. Na Tabela 4 são encontrados os valores para as duas possibilidades de simetria de exons e os simétricos novamente apresentaram-se com excesso acima do esperado.

Tabela 3. Associação das três fases dos introns em seis classes de simetria exônica.

Teste das seis possibilidades (valores absolutos (% em relação ao esperado))						
Nº de exons Internos	(0,0)	(0,1) + (1,0)	(0,2) + (2,0)	(1,1)	(1,2) + (2,1)	(2,2)
	Nº total de exons			X²		
	P					
1	111.959 (12%)	97.398 (-14%)	92.348 (-1%)	37.923 (7%)	50.644 (-5%)	24.475 (11%)
				414.747	4.170	< 0,001

Tabela 4. Associação das três fases dos introns em duas classes de exons: simétricos e assimétricos.

Teste das duas possibilidades (valores absolutos (% em relação ao esperado))				
Nº de exons internos	(0,0) + (1,1) + (2,2)	(0,1) + (1,0) + (0,2) + (2,0) + (1,2) + (2,1)	X²	P
1	174.357 (13%)	240.390 (-8%)	4.812	<0,001

Um dos maiores anseios da bioinformática é poder prever confiavelmente a arquitetura gênica a partir de seqüências originárias do genoma. Alguns programas já foram criados com esta finalidade, embora sua precisão seja questionada por alguns estudos. Uma vez que há um crescente aumento no depósito de dados provenientes de projetos de seqüenciamento de ESTs, este tipo de informação pode ser utilizado confiavelmente para determinar os limites entre exons e introns. Assim sendo, para cada uma das sete espécies modelo que estudamos, separamos as seqüências de origem genômica e o seu respectivo transcrito e, através da comparação dos limites de exons e introns preditos *in silico* e os valores obtidos com os alinhamentos de seu transcriptoma, pudemos validar algumas das estruturas preditas anteriormente *in silico* (Tabela 5). Objetivamos assim a validação de introns preditos *in silico*, sendo que não foram avaliados introns não determinados por estes programas de predição da estrutura gênica.

Tabela 5. Introns preditos *in silico* confirmados por ESTs.

	Número de acessos ao GenBank com introns preditos	Número de acessos ao GenBank com introns preditos confirmados	Introns preditos	Número de ESTs	Introns preditos confirmados por ESTs
<i>R. norvegicus</i>	23	10	183	273.591	31 (17%)
<i>M. musculus</i>	137	73	1.704	1.296.332	389 (23%)
<i>C. elegans</i>	3.016	2.283	100.977	58.367	17.454 (17%)
<i>H. sapiens</i>	1.852	1.149	23.235	3.406.430	6.013 (26%)
<i>A. thaliana</i>	1.592	1.438	125.567	112.999	31.873 (25%)
<i>D. melanogaster</i>	703	542	52.639	116.099	10.278 (20%)
<i>S. cerevisiae</i>	317	38	1.024	11.159	38 (4%)

Esta análise da base de dados, incluindo as análises estatísticas, está aberta a toda comunidade científica através do endereço <http://intron.bic.nus.edu.sg/exint/newexint/exint.html> (Sakharkar *et al*, 2002).

Na página de ExInt, podem ser feitas buscas de seqüências nucleotídicas e de aminoácidos contra o próprio ExInt utilizando BLASTN, BLASTP e BLASTX. Para facilitar a pesquisa, é possível escolher separadamente cinco conjuntos de seqüências: 1) genes contendo introns preditos *in silico*; 2) genes contendo introns definidos experimentalmente; 3) genes de origem

organelar (mitocôndria e cloroplasto); 4) genes de origem nuclear e; 5) conjunto não-redundante. Ainda é possível fazer a busca no banco de dados relacional através de palavras-chave, número de acesso ao GenBank e espécie. Uma importante implementação no banco de dados relacional que também pode ser visualizada na mesma página de internet é a sequência dos exons em aminoácidos e a sequência do introns em nucleotídeos. Para os pesquisadores que se interessarem em ter tanto o banco de dados em formato FASTA e o conjunto não-redundante ou com os dados já formatados para bancos de dados MySQL, é possível fazer o download dos arquivos livremente.

IV.b. *Splicing alternativo*

O *splicing* de pré-mRNAs acontece através de um sistema multi-enzimático, o spliceossomo. Subunidades de pequenas ribonucleoproteínas nucleares (snRNP) e diversas outras proteínas são necessárias para que ocorra, de maneira coordenada, a agregação destas ao transcrito para formar o spliceossomo.

O término do projeto genoma humano revelou um surpreendentemente baixo número de genes (Venter *et al.*, 2001; Lander *et al.*, 2001). Mesmo que ao longo dos anos este número aumente, eventos de *splicing* alternativo serão ainda um dos prováveis mecanismos celulares responsáveis pela enorme diversidade encontrada no proteoma humano. Consensualmente estima-se que pelo menos 1/3 dos genes humanos sejam alvos deste tipo de mecanismo de aumento de variabilidade dos transcritos. O mecanismo completo de reconhecimento de todos os sítios de *splicing* ainda não é bem elucidado, mas a partir de bases de dados, que contêm as informações do mapeamento de ESTs no genoma humano ou nos transcritos de referência (*full-length*), pode-se inferir com um grau de confiabilidade alto a taxa de *splicing* alternativo no transcriptoma humano (Mironov *et al.*, 1999; Croft *et al.*, 2000; Hide *et al.*, 2001; Modrek *et al.*, 2001).

Um dos maiores problemas em estudos referentes à análise em larga escala de *splicing* alternativo é a representação gráfica das isoformas. Muitas vezes, as variantes de *splicing* somente são detectadas através do alinhamento múltiplo de um conjunto de seqüências. No entanto, estes alinhamentos são longos e sua visualização não é fácil. Desenvolvemos então uma nova metodologia de fácil identificação e interpretação de eventos de uso alternativo de exons (UAE) em transcritos humanos e pudemos formular algumas perguntas de interesse biológico através do uso de ferramentas de bioinformática. Desenvolvemos uma nova metodologia capaz de detectar UAE em transcritos através da construção de matrizes binárias, nas quais cada seqüência é representada por uma linha e um exon por uma coluna. Para um dado cluster, a presença de um exon numa seqüência é enumerada com “1”, de outra forma com “0”. As vantagens desta metodologia são: 1) sua interpretação é bastante simples; 2) podem ser utilizadas para análises estatísticas; 3) são facilmente manipuladas para clusterização; e 4) são facilmente trabalhadas em estudos em larga escala. Contudo, é necessário apontar que uma limitação desta metodologia está na capacidade de detecção de somente um dos tipos de *splicing* alternativo, o uso alternativo de exons.

A partir da base de dados relacional desenvolvida por nosso grupo denominada “AEU” (Materias e métodos), dicotomizamos a nossa análise em duas vertentes: nos 255.382 clusters não-redundantes e nos 17.999 clusters que continham os genes que consideramos *full-length*.

Analisando o conjunto de *full-length*, encontramos que 51% deles possuem mais de uma variante de *splicing*, totalizando 65.481 isoformas. Um dos maiores problemas na detecção de variantes de *splicing* está no fato de que alguns delas são somente representadas por um transcrito, muitas vezes consideradas decorrentes de erros que acontecem durante tanto a transcrição como durante o *splicing* de RNAs. Assim, excluindo estas isoformas, a taxa de UAE nestes *full-length* decai para 41%, acumulando 38.086 variantes.

Na análise que realizamos nos clusters não-redundantes, encontramos uma grande quantidade de clusters compostos somente por um exon. Ao excluirmos tais clusters, reduzimos o número de clusters não-redundantes de 255.382 para 36.688. Além de não se apresentarem redundantes, estes clusters ainda possuem todos os 17.999 clusters previamente separados como sendo do conjunto que contém os cDNA *full-length* mapeados. Assim, para este novo conjunto de clusters que determinamos, cerca de 28% deles estão produzindo mais que uma variante, somando 67.030 isoformas.

Novamente, excluindo-se aquelas variantes que possuem somente uma EST, a taxa de UAE decresce para 19%, remanescendo 43.500 variantes. Uma descrição completa das taxas de UAE e média de isoformas para cada análise pode ser encontrada na Tabela 6.

Tabela 6. Estatística para eventos de UAE.

		Total		Variantes com mais de uma sequência	
		UAE	Média de variantes	UAE	Média de variantes
cDNAs full-length	17.999	51%	3,6	41%	2,1
Clusters não-redundantes e com mais de um exon	36.688	28%	1,8	19%	1,1

Corroborando dados de Hide e colaboradores (2001) e Modrek e colaboradores (2001), a taxa de *splicing* é correlacionada com o número de seqüências que compõem um cluster. Constatamos que a partir de clusters com mais de 100 cDNAs a taxa de UAE manteve-se constante em cerca de 75% (Figura 8).

Figura 8. Relação do número de seqüências nos clusters e taxa de UAE em mRNAs conhecidos.

A distribuição de eventos de splicing alternativo ao longo dos transcritos possui uma grande importância biológica, conforme inicialmente proposto por Mironov e colaboradores (1999). Posteriormente, em 2001, Modrek e colaboradores também avaliaram a mesma questão, porém os dados são controversos. Assim, analisamos a distribuição da frequência de UAE ao longo dos transcritos e encontramos que eles são mais frequentemente encontradas na metade 5' dos transcritos. Uma vez que

Makalowski & Bogunski (1998) propõem que os 5,5% iniciais dos transcritos sejam referentes à 5' UTR e os 19,6% finais refiram-se ao 3' UTR, analisamos que, para o conjunto de cDNAs *full-length*, cerca de 14% e 13% dos eventos de *splicing* alternativo estariam acontecendo supostamente dentro da 5' e 3' UTR, respectivamente (Figura 9).

Figura 9. Distribuição de eventos de uso alternativo de exons ao longo dos transcritos. A maior fração foi observada dentro da região codificante.

Uma das grandes indagações que a biologia levanta em relação à *splicing* alternativo é a análise das variantes tecido-específicos e expressos

em diferentes fases do desenvolvimento. Em outros organismos, já foram caracterizados muitos casos de variantes expressos durante a diferenciação sexual, como o gene *dsx* em *Drosophila* (Nothiger & Steinmann-Zwicky, 1987). A análise dos cDNAs *full-length* auxiliou para que pudéssemos entender o padrão de expressão de isoformas tecido-específicas destes genes em 12 diferentes tecidos. As variantes tecido-específicas não se apresentaram distribuídas uniformemente, predominando a diversidade de isoformas em cérebro e testículo (Figura 10).

Figura 10. Número de variantes tecido-específicos por 1.000 cDNAs.

Uma vez que estudos recentes demonstraram diferenças no padrão do uso alternativo de exons entre os transcriptomas tumoral e normal (revisto por Caballero *et al.*, 2001), diferenciamos ambos os transcriptomas em nosso estudo.

Primeiramente, calculamos a taxa de UAE para os distintos transcriptomas. Para as 893.480 seqüências de cDNA provenientes de amostras de tecido normal, observamos um total de 36.859 isoformas. Por outro lado, para 484.516 seqüências de cDNA provenientes de amostras de tecido tumoral, encontramos 26.690 variantes de *splicing*. Ao passo que, em média, 24,2 seqüências normais são necessárias para encontrarmos uma nova isoforma, somente 17,3 seqüências tumorais são suficientes para localizarmos uma nova variante de *splicing* ($p < 0,001$).

Como pode ser observado na Figura 11, a percentagem de seqüências tumorais de um dado cluster e sua respectiva taxa de UAE aumentam de forma correspondente (linha vermelha). Entretanto, poderia estar ocorrendo uma variação estatisticamente significativa no tamanho do cluster para cada classe de porcentagem tumoral analisada. Assim, calculamos a média do tamanho do cluster de cada classe de porcentagem tumoral e a filtramos para obtermos, a partir daqueles clusters que contêm seqüências de origem tecidual normal, os clusters com tamanho $\pm 5\%$ da média calculada ($p < 1 \times$

10^{-6}) (linha em azul da Figura 11). Os valores obtidos através de clusters com mais de 60% de seqüências tumorais foram excluídos devido à pequena amostragem.

Figura 11. Taxa de UAE nos clusters com crescente frequência de cDNAs tumorais (vermelho). Para cada dado encontrado em vermelho, a taxa de UAE para os clusters de mesma média de tamanho foi determinada (azul) e somente considerando seqüências normais.

V. Discussão

V.a. Exon shuffling

A arquitetura gênica teve a sua nomenclatura inicialmente proposta por Walter Gilbert em 1978. Os termos exon e intron foram pela primeira vez definidos pelo o autor a partir das expressões (em inglês) *expressed region* e *intragenic region*, respectivamente. O autor ainda propõe que os introns, por serem cerca de dez vezes maiores que os exons, aumentariam a taxa de recombinação entre dois genes na ordem de 10^8 vezes (Gilbert, 1978; 1987). Este mecanismo de desenvolvimento de novos genes, é denominado por Gilbert de *exon shuffling*. Desde então, diversos estudos têm sido realizados comprovando que este processo de recombinação mediada por introns é um das mais eficientes vias de surgimento de novos genes (Patthy, 1987; Long *et al.*, 1995; de Souza *et al.*, 1996; de Souza *et al.*, 1998; Fedorov *et al.*, 2001).

A predição da estrutura gênica é alvo de diversas pesquisas em bioinformática e alguns programas são amplamente utilizados pela comunidade científica para este fim. Como demonstrado pela análise do cromossomo 22 humano (Dunham *et al.*, 1999), 30% dos exons deste

cromossomo não foram confirmados por métodos experimentais e 20% dos determinados experimentalmente não conseguiram ser preditos por programas. Rogic e colaboradores (2001) realizaram uma extensiva análise da eficiência de diversos programas de predição da arquitetura gênica e concluem que cerca de 70% dos exons podem ser detectados *in silico*. Assim, aplicamos uma extensiva análise em que conseguimos validar introns preditos por estes programas em aproximadamente 20% dos casos através de alinhamentos com ESTs (Tabela 5). Estes dados podem ser utilizados tanto por pesquisadores interessados na melhora dos atuais programas de predição da estrutura gênica, como por biólogos interessados na validação de introns em estudos de genes específicos. Entretanto, é necessário ressaltar que esta metodologia possui a limitação da utilização de transcritos seqüenciados e depositados em bases públicas de dados para a confirmação dos limites de exons e introns preditos *in silico*. Portanto, os introns preditos que não foram confirmados através do alinhamento não podem ter os seus limites de exons e introns desconsiderados, pelo simples fato de não serem observados. Neste contexto, deve ser levado em consideração que a cobertura do transcriptoma não foi ainda finalizada (Sakharkar *et al.*, 2002).

A frequência do tamanho dos introns, similar ao que também foi observado por Deutsch e Long (1999), apresentou-se de forma bimodal

(Figura 5). Pouco se sabe da biologia deste padrão para que os introns tenham entre 40 e 120 nucleotídeos, com uma nítida queda na frequência por volta de 70 nucleotídeos. Podemos inferir que de maneira geral este padrão não é influenciado por introns preditos *in silico*, uma vez que os introns experimentais apresentam-se de maneira semelhante.

A análise de bases de dados cada vez maiores tornou possível que a quantificação das fases dos introns numa dada população gênica seja uma eficaz maneira de se detectar *exon shuffling*, uma vez que estas não se apresentam randomicamente distribuídas. Se a distribuição dos introns nas três fases fosse randômica, seriam esperadas frequências de 1/3 para cada fase. Inicialmente, em 1992, Fedorov e colaboradores apresentaram os primeiros resultados que demonstraram o excesso de introns de fase 0. A importância evolutiva de se encontrarem mais introns de fase 0 está no fato deles não interromperem a fase de leitura dos exons que estão flanqueando-o, tornando mais fácil o embaralhamento dos exons que *exon shuffling* propõe. Três anos mais tarde, Long e colaboradores (1995) fizeram uma extensiva análise em genes que contenham introns presentes tanto em procariotos quanto em eucariotos e encontram um excesso significativo de introns de fase 0. Corroborando estes dados, realizamos uma análise com um conjunto de dados maior que Deustch & Long (1999), não sendo estudados

somente os genes considerados pelos autores para a sua pesquisa. Encontramos que introns de fase 0 estão presentes em ExInt com excesso de 32% acima do esperado ($p < 0,001$) (Sakharkar *et al.*, 2002). Além da avaliação da fase dos introns como meio de detecção de *exon shuffling* em bases de dados, a simetria dos exons também têm sido utilizada para tal (Long *et al.*, 1995; Deustch & Long, 1999). Assim, como pode ser observado nas tabelas 3 e 4, independente do teste estatístico aplicado, encontramos um excesso significativo de exons simétricos (13%) quando comparados aos assimétricos (-8%). Long e colaboradores (1995) determinaram que os exons simétricos do tipo (1,1) são os que possuem maior diferença estatística. Entretanto, os autores observaram que havia em sua base de dados quantitativamente mais exons simétricos do tipo (0,0). Encontramos também esta preferência quantitativa de exons (0,0) em nosso estudo, entretanto a classe de simetria exônica que apresentou-se com melhor significado estatístico foi novamente (0,0) e não mais (1,1), como observado anteriormente. Este significativo excesso de exons simétricos (0,0) pode ter auxiliado os genes durante a evolução para que eles pudessem ser inseridos em introns de mesma fase sem a perturbação da fase de leitura que a assimetria causaria. Nestes estudos para a análise da arquitetura gênica, Long e colaboradores (1995) e Deustch & Long (1999) também

encontraram maior frequência de exons entre 25 e 40 resíduos de aminoácidos. Encontramos fundamentalmente o mesmo padrão de curva que ambos os trabalhos demonstram quando analisamos a distribuição da frequência de tamanho dos exons em ExInt, inclusive o pico em 35 resíduos de aminoácidos (Figura 7).

Continuando com a análise dos introns encontrados em ExInt, o estudo da distribuição da posição dos introns nas proteínas pode ser uma fonte de recursos para que diversos processos biológicos sejam elucidados, sendo *splicing* alternativo um deles. Modrek e colaboradores (2001) realizaram um estudo em larga escala sobre *splicing* alternativo no transcriptoma humano e detectaram que há uma tendência para que estes eventos ocorram na metade C-terminal das proteínas. Uma vez que verificamos que os introns são mais frequentemente presentes na metade C-terminal das proteínas (Figura 6), poderíamos inferir que esta alta taxa de *splicing* alternativo seja decorrente justamente da alta frequência de introns na parte gênica que codifica esta região das proteínas.

Assim, através de uma extensiva análise quantitativa de ExInt, demonstramos que eventos de *exon shuffling* podem ser detectados em bases de dados de genes que contenham introns. Utilizamos-nos de duas premissas que podem favorecer eventos de *exon shuffling* e que poderiam ser

quantificadas em nossa base de dados: os excessos acima dos esperados de introns de fase 0 e de exons simétricos. Além de ser uma base de dados com grande potencial para estudos evolutivos, ExInt pode também ser utilizada para o estudo dos exons e dos introns. Demonstramos neste estudo somente uma de suas aplicações: a validação de introns preditos *in silico*. Outras subdivisões ainda não foram devidamente estudadas em ExInt como, por exemplo, diferença dos genes de origem organelar e nuclear. Assim, novas abordagens para o estudo qualitativo de eventos de *exon shuffling* são necessárias para que novos casos sejam descritos e os seus mecanismos moleculares sejam melhor compreendidos, sendo que ExInt possui todas as qualidades para que elas sejam realizadas.

V.a. *Splicing alternativo*

O descobrimento dos introns interrompendo os genes (Berget *et al.*, 1977; Chow *et al.*, 1977) levou a uma mudança no conceito de arquitetura gênica. Walter Gilbert, no mesmo artigo que realiza as primeiras colocações referentes a *exon shuffling* em 1978, apresenta a possibilidade dos exons serem utilizados de diferentes formas para viabilizar a formação de produtos distintos a partir de um mesmo gene. A este mecanismo denominaram *splicing* alternativo e muitos de seus mecanismos moleculares já são bem elucidados (revisto por Chabot, 1996; e por Smith & Valcárcel, 2000).

Através de uma análise na qual mapeamos os transcritos humanos no genoma humano, pudemos confirmar que a maioria dos genes humanos têm seus transcritos envolvidos em eventos de *splicing* alternativo. Obtivemos que, para os cDNAs *full-length* que estudamos, 51% deles apresentaram mais de uma variante de *splicing* (Tabela 6). Entretanto, como somente analisamos um tipo de *splicing* alternativo, o uso alternativo de exons, quando levados em consideração os outros possíveis, este número tenderá a aumentar. Lander e colaboradores (2001), por exemplo, constataram que após o mapeamento de genes conhecidos no cromossomo 22 a taxa de *splicing* alternativo chegou a 59%. Realizando uma análise por tamanho dos

clusters dos cDNA *full-length* escolhidos, encontramos que a partir de 100 cDNAs a taxa de UAE permanece constante em torno de 75% (Figura 8). Portanto, o aumento de seqüências depositadas nas bases públicas de dados tende a aumentar o número de isoformas conhecidas por gene, sendo de fundamental importância a continuidade de projetos de seqüenciamento de ESTs.

No entanto, os projetos de seqüenciamento de ESTs têm levado a um acúmulo de seqüências nas extremidades 5' e 3' dos transcritos. Dias-Neto e colaboradores (2000) analisaram esta distribuição e determinaram que nestas regiões a freqüência de seqüências é de 26% e 65%, respectivamente, remanescendo apenas uma pequena percentagem restrita à parte interna dos mesmos.

Num dos primeiros estudos referentes à *splicing* alternativo em bioinformática, Mironov e colaboradores (1999) procuraram determinar a freqüência de eventos de *splicing* alternativo ao longo dos transcritos. Os autores demonstram que cerca de 80% destes eventos estão restritos à 5' UTR e os 20% restantes inclusos tanto na região codificante como em 3' UTR. Posteriormente, Modrek e colaboradores (2001) apresentam dados sobre a sua análise destes eventos ao longo de transcritos humanos, ocorrendo preferencialmente na região codificante (74%), mais

especificamente na metade correspondente à C-terminal das proteínas. Os autores ainda apresentam uma separação para os eventos de *splicing* alternativo nas UTRs, sendo 22% em 5' UTR e 4% para a 3' UTR. Ao analisarmos a distribuição dos eventos de UAE ao longo dos transcritos, encontramos que a grande maioria dos eventos concentraram-se na região codificante, sendo predominante na metade 5' dos transcritos (Figura 9). Separando-se os eventos de UAE em 5' UTR, região codificante (CDS) e 3' UTR, obtivemos as frequências de 14%, 73%, 13%, respectivamente. Ao analisarmos somente a CDS dos cDNAs *full-length*, encontramos uma tendência para que eventos de UAE ocorram na região referente à sua metade inicial (45%) quando comparada à sua final (28%). Talvez os dados apresentados por Modrek e colaboradores (2001) tenham sido influenciados de duas maneiras: 1) pelo baixo número de genes estudados (apenas 50); e 2) que os genes escolhidos possuem uma maior representatividade de introns na região dos transcritos que codificam a metade C-terminal das proteínas, como demonstramos para um grande grupo de genes que contêm introns (Figura 6), levando ao acúmulo de eventos de *splicing* nesta região dos transcritos. Assim, por termos utilizado um conjunto maior de genes, presumimos que a maior distribuição de introns na metade C-terminal das proteínas não necessariamente influencia a distribuição de eventos de UAE

nos transcritos humanos. Assim, a nossa abordagem em 17.999 genes possibilitou que possíveis tendências como famílias gênicas com maior taxa de UAE numa dada região de seus genes fossem diluídas no resultado final. Portanto, análises em larga escala podem ser considerados bons indicadores de padrões biológicos e, trabalhando-se com quantidades de dados cada vez maiores, pode-se observar com uma confiabilidade cada vez maior o comportamento de eventos biológicos específicos.

Biologicamente, esta maior frequência de variação nos transcritos em sua metade 5' pode acarretar tanto na eficiência em que ocorre o acoplamento dos ribossomos para o início da tradução como no reconhecimento do primeiro limite de exon e intron. Já variações na metade 3' dos transcritos seriam responsáveis pela estabilidade do RNA e por mudança no sítio de poliadenilação (revisto por Goldstrohm *et al.*, 2001).

Para eventos de *splicing* alternativo tecido-específicos, uma variedade de fatores são atribuídos para o reconhecimento diferencial e estrito de alguns sítios de *splicing*. Reguladores tecido-específicos podem estabilizar o acoplamento das proteínas SR para que a comunicação destas com um sítio específico de *splicing* seja melhorada (revisto por Chabot, 1996). Encontramos que testículo apresenta maior diversidade de isoformas tecido-específicas por cDNA *full-length* estudado (Figura 10). Desde o seu

desenvolvimento até a espermatogênese, os genes expressos neste órgão podem ser regulados tanto durante a sua transcrição como durante o *splicing* dos introns, sendo plausíveis mecanismos alternativos ao spliceossomo (Gonzalez *et al.*, 1999; Wegner, 1999). A determinação sexual e o desenvolvimento embrionário em algumas espécies é extremamente dependente de fatores de transcrição regulados através de *splicing* alternativo, podendo também ser utilizados na sua regulação. Um caso muito bem descrito na literatura é o do gene *dsx* (Nothiger & Steinmann-Zwicky, 1987), o qual fatores tecido-específicos estabilizam o acoplamento de proteínas SR para que melhore o reconhecimento do sítio alternativo de *splicing*. No ser humano, a determinação sexual foi demonstrada ser coordenada pelo gene *SRY* tanto em humanos quanto em murinos (revisto por McLaren, 1990). A expressão deste gene que induz o desenvolvimento dos testículos e a subsequente diferenciação sexual masculina é decorrente dos hormônios produzidos por este órgão (Sinclair *et al.*, 1990), sendo um clássico exemplo de expressão tecido-específica. A proteína SRY possui um domínio de ligação ao DNA denominado HMG que tem 50% de similaridade com outra proteína: a SOX. Recentemente, comprovou-se que genes da família *SOX* estejam relacionados na regulação de *splicing* alternativo tecidos-específicos atuando como fatores de transcrição, incluindo o

testículo (Ohe *et al.*, 2002). Entretanto, é aceito que a expressão de genes *SOX* ocorra em algum período da vida de todas as células (revisto por Wegner, 1999). Outro órgão que também demonstrou alta taxa de *splicing* alternativo foi o cérebro. Uma vez que este é considerado o órgão de maior diversidade protéica, os dados aqui apresentados (Figura 10) levam-nos a inferir que esta diversidade possa ser resultante de uma expressão diferenciada de seus genes a partir de eventos de uso alternativo dos exons, como acontece com as neurexinas (Missler & Südhof, 1998). Diversos outros estudos vêm demonstrando que o cérebro é um dos órgãos de maior abundância de eventos de *splicing* alternativo (revisto por Grabowski & Black, 2001). Dois exemplos, um em humano e outro em *Drosophila*, têm revelado o poder de criação de variabilidade de *splicing* alternativo: caderinas (Wu & Maniatis, 1999) e *DSCAM* (Schmucker *et al.*, 2000), respectivamente.

Um dos grandes interesses da medicina contemporânea é a detecção de variantes de *splicing* específicos de tecidos tumorais, para que eles sirvam como marcadores tumorais. A importância de eventos de *splicing* em estados celulares tumorais pode ser exemplificada pelo estudo de Matter e colaboradores (2000). Os autores descrevem um complexo protéico relacionado à *splicing* controlado por vias oncogênicas. Como observado por

Smith & Valcárcel (2000), variações nas concentrações de fatores de *splicing*, hnRNPs e proteínas SR afetam a escolha de um sítio em diversos RNAs. Esta fina estequiometria pode definir um código celular para estabelecimento de escolhas de sítios de *splicing* específicos de um tecido, de um estágio embrionário ou mesmo durante a tumorigênese. Mudanças nos níveis de proteínas SR que ocorrem durante a tumorigênese são correlacionadas com o uso alternativo de exons no gene de *CD44*, que codifica uma molécula de adesão celular envolvida em estados metastáticos celulares (Stickeler *et al.*, 1999). Assim, nossa observação de que as variantes de *splicing* são mais freqüentemente encontradas em tecidos tumorais (Figura 11), pode ser explicada tanto por uma maior atividade celular encontrada nesses estados patogênicos como por uma regulação específica do spliceossomo que somente ocorre durante estes processos celulares. Entretanto, as altas taxas de *splicing* alternativo encontradas para ambos os transcriptomas sugerem que este tipo de evento que aumenta a variabilidade dos transcritos é essencial para a manutenção de um estado constitutivo celular, podendo também estar envolvida na sua desregulação e criação de um novo estado, o tumoral. Pela amostragem ser ainda reduzida para o transcriptoma tumoral, acreditamos que eventualmente o seu aumento não possibilite mais observarmos esta alta taxa de UAE que encontramos,

quando comparada ao transcriptoma normal. Os transcriptomas normal e tumoral provavelmente possuam grande similaridade de isoformas expressas, mesmo que somente nos níveis basais de transcrição, podendo ser encontradas menos freqüentemente variantes de *splicing* específicas de cada estágio celular. Assim, possivelmente, a grande diferença entre processos celulares considerados normal e tumoral deve estar na quantidade de fatores e reguladores temporais tanto da transcrição como envolvidos no processo de *splicing* de pré-mRNAs.

Uma vez que alguns estudos sobre *splicing* alternativo no transcriptoma humano foram realizados (Mironov *et al.*, 1999; Croft *et al.*, 2000; Hide *et al.*, 2001; Modrek *et al.*, 2001), alguns problemas inerentes a estes tipos de análises passam a ser detectados. Em artigo recentemente publicado, Modrek & Lee (2002) discutem estes problemas detalhadamente.

O nosso estudo poderia ter sido afetado por alguns fatores. Desenvolvemos uma metodologia que somente é capaz de detectar um dos tipos de *splicing* alternativo, o uso alternativo de exons. Assim, outras formas de *splicing* não representadas podem fazer variar os valores aqui demonstrados. Como discutido anteriormente, a distribuição das ESTs não ocorre uniformemente ao longo dos transcritos (Dias-Neto *et al.*, 2000). Isto pode acarretar na observação mais fácil de eventos de *splicing* alternativo

nas extremidades dos transcritos, sendo então essencial que técnicas que aumentam a amostragem de ESTs na sua porção mais interna sejam utilizadas. Assim, possivelmente, algumas isoformas que não haviam sido detectadas anteriormente, possam ser observadas e estudadas. Neste sentido, a determinação exata do início das UTRs e da região codificante dos transcritos pode ser um fator limitante em nosso estudo. Uma vez que não consideramos os limites inicial e final da CDS, o estudo minucioso através de genes que possuem anotação para as UTRs pode alterar os resultados aqui descritos. A diferença observada entre os clusters não-redundantes e os genes *full-length* estudados pode ser devida tanto ao grande número de clusters formados por somente uma sequência como pelos genes que possuem pelo menos um transcrito *full-length* serem mais representados nas bases de dados. Outro fator que pode ter influenciado os nossos resultados nas análises de isoformas é a anotação das sequências. Uma vez que utilizamos palavras-chave para a classificação das sequências em tumoral e normal, tanto o erro de anotação por parte de quem a depositou, como nosso, por não termos utilizado palavras-chave que possibilitassem a correta separação dos transcriptomas podem levar a mudanças nos dados aqui apresentados, sendo que a determinação tecidual das sequências que realizamos está sujeita a ser influenciada da mesma forma.

Assim, pesquisas em larga escala têm sido cada vez mais utilizadas para que sejam elucidadas tanto a arquitetura quanto a expressão gênica. A arquitetura dos genes eucarióticos já é bem conhecida e descrita na literatura, entretanto análises em bases de dados maiores podem ajudar a confirmar dados obtidos em estudos de menor escala realizados anteriormente. Entretanto, algumas discordâncias ainda são encontradas quando estudadas as diferentes maneiras que um gene pode ser expresso. Somente novos estudos e novas abordagens em bioinformática podem auxiliar na confirmação de padrões biológicos observados em transcriptomas e assim auxiliar os biólogos no entendimento de processos celulares que atuam no controle da expressão gênica.

VI. Conclusões

Este estudo abordou, em duas análises distintas, eventos que contribuem para o aumento da diversidade protéica, sendo *exon shuffling* em genomas e *splicing* alternativo em transcriptomas.

Encontramos os excessos de introns de fase zero e de exons simétricos na base de dados ExInt, que são considerados bons indicadores de eventos de *exon shuffling* em análises em larga escala.

Validamos um conjunto de introns preditos *in silico* de sete espécies modelo através de uma nova abordagem que utiliza o mapeamento de ESTs no seu genoma.

Identificamos que a distribuição dos introns ao longo das proteínas ocorre preferencialmente na metade correspondente à C-terminal. Entretanto, encontramos numa extensiva análise de uso alternativo de exons no transcriptoma humano que estes eventos ocorrem mais frequentemente na metade 5' dos transcritos.

Determinamos também que pelo menos metade dos genes humanos devem produzir mais de um transcrito através de uso alternativo de exons e que são mais facilmente encontradas isoformas no transcriptoma tumoral quando comparada ao normal.

VII. Referências bibliográficas

- Adams M.D., *et al.* The genome sequence of *Drosophila melanogaster*. *Science*, v.287, n.5461, p.2185-95, 2000.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, v.25, n.17, p.3389-402, 1997.
- Berget, S.M., Moore, C., Sharp, P.A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA*, v.74, n.8, p.3171-5, 1977.
- Black, D. L. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, v.103, n.3, p.367-70, 2000.
- Blanchette, M., Chabot, B. Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization. *EMBO J*, v.18, n.7, p.1939-52, 1999.
- Brett, D., Pospisil, H., Valcarcel, J., Reich, J., Bork, P. Alternative splicing and genome complexity. *Nat Genet.*, v.30, n.1, p. 29-30, 2002.
- Burge, C., Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, v.268, n.1, p.78-94, 1997.
- Caballero, O.L., de Souza, S.J., Brentani, R.R., Simpson, A.J. Alternative spliced transcripts as cancer markers. *Disease Markers*, v.17, n.2, p. 67-75, 2001.

- Cáceres, J.F. *et al.* Role of the modular domains of SR proteins in subnuclear localization and alternative splicing specificity. *J Cell Biol*, v.138, n.2, p.225-38, 1997.
- Cartegni, L., Chew, S.L., Krainer, A.R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet*, v.3, n.4, p.285-98, 2002.
- Chabot, B. Directing alternative splicing: cast and scenarios. *TIG*, v.12, n.11, p.472-8, 1996.
- Chow, L.T., Gelinas, R.E., Broker, T.R., Roberts, R.J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, v.12, n.1, p.1-8, 1977.
- Croft, L. *et al.* ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat Genet.*, v.24, n.4, p. 340-1, 2000.
- de Souza, S.J., Long, M., Schoenbach, L., Roy, S.W., Gilbert, W. Intron positions correlate with module boundaries in ancient proteins. *Proc. Natl. Acad. Sci. USA*, v.93, p.14632-14636, 1996.
- de Souza, S.J., Long, M., Klein, R.J., Roy, S., Lin, S., Gilbert, W. Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl. Acad. Sci. USA*, v.95, p.5094-5099, 1998.
- Deutsch M., Long M. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.*, v.27, n.15, p.3219-28, 1999.

- Dias-Neto, E. *et al.* Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc Natl Acad Sci U S A.*, v.97, n.7, p. 3491-6, 2000.
- Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature*, v.402, p.489-495, 1999.
- Fedorov, A., Suboch, G., Bujakov, M., Fedorova, L. Analysis of nonuniformity in intron phase distribution. *Nucleic Acids Res*, v.20, n.10, p.2553-7, 1992.
- Fedorov, A., Cao, X., Saxonov, S., de Souza, S.J., Roy, S.W., Gilbert, W. Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. *Proc Natl Acad Sci USA*, v.98, n.23, p.13177-82, 2001.
- Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, v.8, n.9, p. 967-74, 1998.
- Gilbert, W. Why genes in pieces? *Nature*, v.271, p.501, 1978.
- Gilbert, W. The exon theory of genes. *Cold Spring Harbor Symposia on Quantative Biology*, v.LII, p.901- 905, 1987.
- Gilbert, W., Glynias, M. On the ancient nature of introns. *Gene*, v.135, p. 137-44, 1993.
- Goldstrohm, A.C., Greenleaf, A.L., Garcia-Blanco, M.A. Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. *Gene*, v.277, n.1-2, p.31-47, 2001.

- Gonzalez, T.N., Sidrauski, C., Dorfler, S., Walter, P. Mechanism of non-spliceosomal mRNA splicing in the unfolded protein response pathway. *EMBO J*, v.18, n.11, p.3119-32, 1999.
- Grabowski, P.J., Black, D.L. Alternative RNA splicing in the nervous system. *Prog Neurobiol*, v.65, n.3, p.289-308, 2001.
- Graveley, B.R. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, v.17, n.2, p. 100-7, 2001.
- Hide, W.A. *et al.* The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res*, v.11, n.11, p. 1848-53, 2001.
- Jones, J.M. *et al.* The mouse neurological mutant flier expresses a novel hybrid gene derived by exon shuffling between *Gnb5* and *Myo5a*. *Hum. Mol. Gen.*, v.9, n.5, p. 821-8, 2000.
- Kolkman, J.A., Stemmer, P.C. Directed evolution of proteins by exon shuffling. *Nature Biotech.*, v.19, p. 423-8, 2001.
- Lander, E.S., *et al.* Initial sequencing and analysis of the human genome. *Nature*, v.409, n.6822, p.860-921, 2001.
- Logsdon, J. M. Jr. *et al.* Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc. Natl. Acad. Sci. USA*, v.92, n.18, p.8507-11, 1995.
- Long, M., Langley, C.H. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science*, v.260, n.5104, p. 91-5, 1993.

- Long, M., de Souza, S.J., Gilbert, W. Evolution of the intron-exon structure of eukariotic genes. *Curr Opin Genet Dev*, v.5, p.774-778, 1995.
- Long, M. Evolution of novel genes. *Curr Opin Genet Dev.*, v.11, n.6, p. 673-80, 2001.
- Makalowski, W., Boguski, M.S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. USA*, v.95, n.16, p.9407-12, 1998.
- Matter, N. *et al.* Heterogeneous ribonucleoprotein A1 is part of an exon-specific splice-silencing complex controlled by oncogenic signaling pathways. *Journal Biol. Chem.*, v.275, n.45, p. 35353-35360, 2000.
- McLaren, A. Sex determination. What makes a man a man? *Nature*, v.346, n.6281, p.216-7, 1990.
- Mewes, H.W., et al. Overview of the yeast genome. *Nature*, v.387, p. 7-65, 1997.
- Mironov, A.A., Fickett, J.W., Gelfand, M.S. Frequent alternative splicing of human genes. *Genome Res.*, v.9, n.12, p.1288-93, 1999.
- Missler M., Sudhof T.C. Neurexins: three genes and 1001 products. *Trends Genet.*, v.14, n.1, p.20-6, 1998.
- Modrek, B., Lee, C. A genomic view of alternative splicing. *Nat Genet.*, v.30, n.1, p. 13-9, 2002.

- Modrek, B., Resch, A., Grasso, C., Lee, C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, v.29, n.13, p. 2850-9, 2001.
- Nothiger R., Steinmann-Zwicky M. Genetics of sex determination: what can we learn from *Drosophila*? *Development.*, v.101, supp.17-24, 1987.
- Ohe, K., Lalli, E., Sassone-Corsi, P. A direct role of SRY and SOX proteins in pre-mRNA splicing. *Proc Natl Acad Sci USA*, v.99, n.3, p.1146-51, 2002.
- Ohno, S., Wolf, V., Atkin, N.B. Evolution from fish to mammals by gene duplication. *Hereditas*, v.59, n.1, p.169-187, 1968.
- Patthy, L. Intron-dependent evolution: preferred types of exons and introns. *FEBS Lett.*, v.214, n.1, p.1-7, 1987.
- Patthy, L. Genome evolution and the evolution of exon-shuffling – a review. *Gene*, v.238, p. 103-114, 1999.
- Palmer, J.D., Logsdon, J.M. The recent origin of introns. *Curr. Opin. Genet. Dev.*, v.1, n.4, p.470-7, 1991.
- Rogic, S., Mackworth, A.K., Ouellette, F.B. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.*, v.490, p. 685-90, 2001.
- Sakharkar, M., Long, M., Tan, T.W., de Souza, S.J. ExInt: an Exon/Intron database. *Nucleic Acids Res.*, v.28, n.1, p.191-192, 2000.

Sakharkar, M., **Passetti, F.**, de Souza, J.E., Long, M., de Souza, S.J. ExInt: an Exon Intron database. *Nucleic Acids Res.*, v.30, n.1, p. 191-4, 2002.

Saxonov, S., *et al.* EID: the Exon-Intron Database-an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.*, v.28, n.1, p.185-90, 2000.

Schisler, N., Palmer, J.D. The IDB and IEDB: intron sequence and evolution databases. *Nucleic Acids Res.*, v.28, n.1, p. 181-184, 2000.

Schmucker, D. *et al.* *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, v.101, n.6, p.671-84, 2000.

Sharp, P.A. Split genes and RNA splicing. *Cell*, v.77, p.805-815, 1994.

Sinclair, A.H., *et al.* A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature*, v.346, n.6281, p.240-4, 1990.

Smith, C.W., Valcárcel, J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci*, v.25, n.8, p. 381-8, 2000.

Stickeler, E., Kittrell, F., Medina, D., Berget, S.M. Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis. *Oncogene*, v.18, n.24, p.3574-82, 1999.

Stoltzfus, A. *et al.* Testing the exon theory of genes: the evidence from protein structure. *Science*, v.265, n.5169, p.202-7, 1994.

- Südhof, T.C., Russell, D.W., Goldstein, J.L., Brown, M.S., Sanchez-Pescador, R., Bell, G.I. Cassette of eight exons shared by genes for LDL receptor and EGF precursor. *Science*, v.228, n.4701, p.893-5, 1985.
- Tacke, R., Manley, J.L. Determinants of SR protein specificity. *Curr Opin Cell Biol*, v.11, n.3, p.358-62, 1999.
- Thanaraj, T.A., Clark, F. Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res.*, v.29, n.12, p.2581-93, 2001.
- van Rijk, A.A., de Jong W.W., Bloemendal, H. Exon shuffling mimicked in cell culture. *Proc Natl Acad Sci U S A*, v.96, n.14, p. 8074-9, 1999.
- van Rijk, A.F., van den Hurk, M.J., Renkema, W., Boelens, W.C., de Jong, W.W., Bloemendal, H. Characteristics of super alphaA-crystallin, a product of in vitro exon shuffling. *FEBS Lett*, v.480, p.79-83, 2000.
- Venter, J.C., *et al.* The sequence of the human genome. *Science*, v.291, n.5507, p.1304-51, 2001.
- Wang, W., Zhang, J., Alvarez, C., Llopart, A., Long, M. The origin of the *Jingwei* gene and the complex modular structure of its parental gene, *yellow emperor*, in *Drosophila melanogaster*. *Mol.Biol. Evol.*, v.17, n.9, p. 1294-1301, 2000.
- Waterston, R.H., Sulston, J.E., Coulson, A.R. The genome. In: *C. elegans II*, Cold Spring Harbor, NY, EUA: CSHL Press, 1997. p.23-45.
- Wegner, M. From head to toes: the multiple facets of Sox proteins. *Nucleic Acids Res*, v.27, n.6, p.1409-20, 1999.

Wu, Q., Maniatis, T. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell*, v.97, n.6, p.779-90 1999.

Uberbacher, E.C., Mural, R.J. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA*, v.88, n.24, p.11261-11265, 1991.