UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA

**O PAPEL DO FENÔMENO DE "EXON-SHUFFLING" ANTIGO E MODERNO NA
EVOLUÇÃO DE PROTEÍNAS**

Tese de Doutorado

Maria Dulcetti Vibranovski

ORIENTADOR:

Prof. Dr. Sandro José de Souza

SÃO PAULO

2005

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA

**O PAPEL DO FENÔMENO DE "EXON-SHUFFLING" ANTIGO E MODERNO NA
EVOLUÇÃO DE PROTEÍNAS**

Tese de Doutorado

Maria Dulcetti Vibranovski

ORIENTADOR:

Prof. Dr. Sandro José de Souza

SÃO PAULO

15 de Dezembro de 2005

"Não há nada que você possa fazer que não possa ser feito"

John Lennon e Paul McCartney

Aos meus pais, Vilma e Jitman

À minha irmã, Joana

AGRADECIMENTOS

Aos meus pais e à minha irmã, Joana. Não tenho nenhuma dúvida da importância de todo o apoio que eles me deram durante o desenvolvimento da minha tese. Sou profundamente grata pela união incondicional que eles estabeleceram para me apoiar nesta fase da minha vida.

Ao Sandro, meu orientador, por me dar a oportunidade trabalhar sob sua orientação e com um assunto tão interessante como evolução dos introns e do fenômeno de *exon-shuffling*.

Ao Noboru Jo Sakabe, por uma experiência pessoal e profissional em comum prazerosa, interessante e muitas vezes desafiante.

À Natanja S. Kirschbaum-Slager, companheira dos momentos bons e ruins. Pelos conselhos, desabafos, papos, correções de Inglês, discussões científicas e risadas. Por ter se tornado a minha melhor amiga durante estes anos que passei em São Paulo.

Ao Prof. Bayardo Batista, primeiro, por ele existir. Segundo por ter me recebido de braços abertos em seu laboratório e ter me deixado usufruir um pouquinho da sua existência.

Ao Noboru Jo Sakabe e ao Robson Franscisco de Souza pelas excelentes correções e discussões na redação da tese. Obrigada pela paciência e compreensão.

Ao Noboru Jo Sakabe, à Natanja S. Kirschbaum-Slager, ao Pedro A.F. Galante, ao Robson F. Souza e ao Jorge E.S. de Souza pela convivência no laboratório e pessoal, pelos incríveis momentos de descontração.

À Ana Claudia Pereira pela convivência, amizade e excelente ajuda administrativa.

À Helena Brentani pelo apoio incrível, pela grande amizade, pelos conselhos e por me ensinar a usar as minhas mãos.

Ao meu amigo Carlos Eduardo Schrago, pela troca de experiências.

À minha amiga Suzana Vaz, que por mais distante que estivesse, sempre esteve muito presente.

Ao pessoal que trabalha ou já trabalhou no laboratório de biologia computacional: Maarten, Elisson Osório, André Zaiats, Milton, Patrícia, Arthur Ramos, Arthur Fabri e Rodrigo de Oliveira.

Ao Raphael Bessa Parmigiani e à Dra. Anamaria Camargo pelo trabalho desenvolvido em conjunto.

Ao Rodrigo Soares de Oliveira pelo trabalho em conjunto.

Aos funcionários da informática pelo auxílio técnico.

Ao pessoal do Ludwig, pelos momentos de convivência.

À Dona Conceição pelo saboroso café diário.

Ao pessoal do laboratório de didática em bioquímica pela receptividade, em especial para Guilherme, Remo, André, Paulo e Silvia.

Aos monitores, professores e alunos da disciplina Bioquímica Experimental para Farmácia pela experiência e pela paciência.

À Comissão de Aperfeiçoamento de Pessoal de Nível Superior (**CAPES**) e à Fundação de Amparo à Pesquisa do Estado de São Paulo (**FAPESP**) pelas bolsas concedidas.

Aos funcionários da pós-graduação do Departamento de Bioquímica da USP pelo atendimento impecável.

A USP, por ser uma Universidade incrível.

Às meninas do futebol da biologia da USP.

Ao Noboru, à Sanai, à Sumire, ao Jun e à Iara pela convivência e por me acolherem.

À Natanja e à Helena pelos almoços das bruxas

À cidade de São Paulo que pode não ser uma Cidade Maravilhosa, mas que me ensinou a ver beleza naquilo que é feio. Mostrou-me que o bonito está muito mais naquilo que você sente do que naquilo que você realmente vê.

À rua 25 de março

Às mais de cem viagens de ônibus São Paulo - Rio que tive que enfrentar

À feira da Liberdade

Ao pastel de feira de todos os Sábados

Aos passeios na Av. Paulista

Ao concreto e à "grana que ergue e destrói coisas belas"

Enfim, a todos, muito agradecida

LISTA DE ABREVIATURAS

Å = Angstroms

BAC = Cromossomo Artificial de Bactéria

BLAST = Ferramenta de busca de alinhamento local

cDNA = DNA complementar

DNA = Ácido desoxirribonucléico

ExInt = Banco de dados de éxons e íntrons

Éxon 0-0 = éxon flanqueado por íntrons de fase zero

Éxon 1-1 = éxon flanqueado por íntrons de fase um

GO = Ontologia Gênica (traduzido do inglês *Gene ontology*)

HMM = Modelo de Markov oculto

L1 = Elemento repetitivo LINE 1

LR = Regiões de ligação entre módulos

mRNA = RNA mensageiro

rRNA = RNA ribossomal

NCBI = Centro Nacional para Informação de Biotecnologia

PAC = Cromossomo artificial derivado do bacteriófago P1

RAC = Regiões ancestrais conservadas

RNA = Ácido ribonucléico

UTR = região não traduzida

RESUMO

A partir da descoberta dos íntrons, muitas questões sobre sua origem vêm sendo discutidas como: porque eles existem em eucariotos e não são encontrados em procariotos, quando e como eles se originaram. Basicamente duas hipóteses existem para explicar a origem dos íntrons: “introns-early” e “introns-late”. A primeira hipótese sugere que íntrons e éxons já existiam nos primeiros genes e os íntrons foram perdidos posteriormente na linhagem de bactérias. A hipótese oponente, *introns-late*, assume que os íntrons foram adicionados posteriormente durante a evolução, somente em eucariotos. Introns de genes diferentes podem sofrer recombinação durante a divisão celular e assim formar novos genes. Este processo permite e aumenta a frequência de troca completa de éxons e, conseqüentemente, aumenta a probabilidade de novos genes funcionais serem formados. Este fenômeno é chamado de “exon-shuffling” e é um mecanismo importante em relação a origem de muitas proteínas novas em eucariotos. Entretanto, o papel do *exon-shuffling* na criação das proteínas no ancestral comum dos procariotos e eucariotos é o ponto de discordância entre as hipóteses *introns-early* e *introns-late* porque este mecanismo depende da presença dos introns no progenoto. Excesso de éxons simétricos é considerado uma evidência de *exon-shuffling*, já que a troca de éxons flanqueados por íntrons da mesma fase não muda o quadro de leitura do gene receptor. Nesta tese, apresentamos dois estudos relacionados ao papel do fenômeno de *exon-shuffling* na evolução das proteínas. No primeiro estudo, observamos que existe uma correlação significativa entre unidades simétricas de *shuffling* e a idade de domínios protéicos. Domínios antigos, presentes em procariotos e eucariotos, são mais freqüentemente

flanqueados por íntrons de fase zero e são preferencialmente localizados nas partes centrais das proteínas. Domínios modernos são mais freqüentemente flanqueados por íntrons de fase um e estão presentes predominantemente nas extremidades das proteínas. Propomos um modelo no qual o *shuffling* de domínios antigos flanqueados por íntrons de fase zero deve ter sido importante durante a criação das partes centrais das proteínas no ancestral comum de eucariotos e procariotos. *Shuffling* de domínios modernos, predominantemente flanqueados por íntrons de fase um, deve ter sido importante para a origem das extremidades das proteínas durante a evolução de eucariotos. O segundo estudo trata do possível papel do *exon-shuffling* na evolução de peptídeos sinal em proteínas humanas. Recentemente, foi mostrado que existe uma predominância de íntrons de fase um próximo ao sítio de clivagem de peptídeos sinal em genes humanos [Tordai, H., Patthy, L. (2004) FEBS lett. 575:109-111]. Os autores sugeriram que tal distribuição é causada pela inserção de íntrons em sítios de inserção preferencial AGG. Apresentamos evidências de que o sinal observado não é tão forte como inicialmente mostrado e que não existe excesso desproporcional de AGG que daria suporte a inserção em sítios preferenciais. Como estas proteínas evoluíram através de *exon-shuffling*, levantamos a possibilidade de que este fenômeno possa também ter sido amplamente responsável pelo excesso de íntrons de fase um. Acreditamos que os dados presentes nestes dois estudos representam uma contribuição importante para o campo de estudo de evolução de íntrons e do fenômeno de *exon-shuffling* porque estes apresentam dados importantes e originais acerca do papel do "exon-shuffling" antigo e moderno durante a evolução das proteínas.

ABSTRACT

Since the discovery of introns, many questions about their origin have been raised such as: why they exist in eukaryotic organisms and not in prokaryotes, when and how did they originate. Mainly, there are two hypotheses explaining the origin of introns: "introns-early" and "introns-late". The first hypothesis suggests that introns and exons already existed in the first genes and were lost later in the bacteria lineage. The opposing hypothesis, *introns-late*, assumes that introns were inserted late in evolution, in eukaryotic organisms only. Introns from different genes may suffer recombination during cell division and this way form new genes. This process allows and increases the frequency of exchanging *complete* exons and, consequently, increases the probability of forming new functional genes. This phenomenon is called "exon-shuffling" and is an important mechanism accounting for the origin of many new proteins in eukaryotes. However, the role of *exon-shuffling* in the creation of proteins in the ancestor of prokaryotes and eukaryotes is the point of disagreement between the hypotheses "introns-early" and "introns-late" because this mechanism depends of the presence of introns in the progenote. Excess of symmetric exons is thought to represent evidence for *exon-shuffling* since the exchange of exons flanked by introns of the same phase does not disrupt the reading frame of the host gene. In this thesis, we present two studies concerning the role of the *exon-shuffling* phenomenon in protein evolution. In the first study, we found that there is a significant correlation between symmetric units of shuffling and the age of protein domains. Ancient domains, present in both

prokaryotes and eukaryotes, are more frequently bounded by phase zero introns and their distribution is biased towards the central part of proteins. Modern domains are more frequently bounded by phase one introns and are present predominantly at the ends of proteins. We propose a model in which shuffling of ancient domains mainly flanked by phase zero introns would have been important during the creation of the central part of proteins in the ancestor of eukaryotes and prokaryotes. Shuffling of modern domains, predominantly flanked by phase one introns, would have accounted for the origin of the extremities of proteins during eukaryotic evolution. The second study accounts the possible role of *exon-shuffling* in the acquisition of signal peptides in human proteins. It was recently shown that there is a predominance of phase one introns near the cleavage site of signal peptides of human genes [Tordai, H., Patthy, L. (2004) FEBS lett. 575:109-111]. The authors suggested that it was due to intron insertion at AGG proto-splice sites. We present evidence that the signal observed is not as strong as initially shown and that there is no disproportional excess of AGG that would support insertion at proto-splice sites. As these proteins evolved by *exon-shuffling*, we raise the possibility that this phenomenon might also be largely responsible for such excess of phase one introns. We believe the data present in these two studies represent an important contribution to the field of introns and *exon-shuffling* evolution due to their important and original data concerning the role of ancient and modern exon-shuffling during the evolution of the proteins.

ÍNDICE

CAPÍTULO 1: INTRODUÇÃO GERAL	1
1.1. Íntrons	2
1.1.1. <i>Íntrons de splicing nuclear</i>	3
1.1.2. <i>Íntrons do grupo I</i>	5
1.1.3. <i>Íntrons do grupo II</i>	5
1.2. O fenômeno de exon-shuffling	7
1.3. Origem dos íntrons	9
1.4. Características dos íntrons	12
1.4.1. <i>Similaridade de seqüências</i>	12
1.4.2. <i>Posição e fase dos íntrons</i>	13
1.5. Características dos éxons	15
1.6. Análises e testes já realizados às características dos íntrons e dos éxons	19
1.6.1. <i>Análises filogenéticas de posições dos íntrons</i>	19
1.6.2. <i>Regiões ancestrais conservadas</i>	20
1.6.2.1. <i>Excesso de íntrons de fase zero</i>	22
1.6.2.2. <i>Excesso de éxons simétricos</i>	23
1.7. Exon-shuffling e unidades protéicas independentes	24
1.7.1. <i>Exon-shuffling e módulos protéicos</i>	25
1.7.2. <i>Exon-shuffling e domínios protéicos</i>	29
1.7.3. <i>Exon-shuffling em eucariotos e domínios protéicos flanqueados por íntrons de fase um</i>	30

1.8. Teoria sintética de evolução dos íntrons	34
1.9. Sítio de “proto-splice”	35
1.10. Resumo dos estudos realizados sobre a origem dos íntrons	41
1.11. Evolução de íntrons e do fenômeno de exon-shuffling	44
1.11.1. Distribuição de eventos de shuffling ao longo das proteínas	45
1.11.2. Eventos de shuffling em proteínas que contêm peptídeo sinal	45
1.12. Objetivo geral	51
1.12.1. Objetivos específicos	51
1.13. Apresentação da tese	52
 CAPÍTULO 2: MATERIAIS E MÉTODOS	 53
 2.1. Materiais	 54
2.1.1. ExInt: banco de dados de exons e intron	54
2.1.1.1. ExInt não redundante	55
2.1.1.2. ExInt experimental	56
2.1.2. SWISS-PROT	57
2.1.3. Pfam	58
2.2. Métodos	59
2.2.1. Programas de alinhamentos de seqüências	59
2.2.2. Método de construção de agrupamentos de seqüências do ExInt por níveis de identidade em alinhamento produzidos pelo programa FASTA	60

2.2.3. <i>Método de construção de agrupamentos de seqüências do SWISS-PROT por identidade e cobertura em alinhamento produzidos pelo programa BLAST</i>	62
2.2.4. <i>Simulações</i>	62
2.2.4.1. <i>Simulação da distribuição de éxons simétricos 0-0 e 1-1 ao longo das proteínas contendo domínios antigos e modernos</i>	63
2.2.4.2. <i>Simulação da freqüência de domínios antigos e modernos correspondentes a éxons 0-0 e 1-1.</i>	64
2.2.5. <i>Determinação da representação de termos de Ontologia Gênica (Gene Ontology) em genes humanos que codificam proteínas com e sem peptídeo sinal</i>	65
 CAPÍTULO 3: SINAIS DE EXON-SHUFFLING ANTIGO E MODERNO ESTÃO CORRELACIONADOS COM A DISTRIBUIÇÃO DE DOMÍNIOS ANTIGOS E MODERNOS AO LONGO DAS PROTEÍNAS	 67
 3.1. Artigo intitulado: <i>Signs of ancient and modern exon-shuffling are correlated to the distribution of ancient and modern domains along proteins</i>	 70
3.2. Análises de distribuição de domínios e excesso de éxons simétricos ao longo das proteínas provenientes de diferentes subconjuntos do ExInt	71
3.2.1. <i>Subconjuntos não redundantes do ExInt</i>	71
3.2.2. <i>Subconjunto experimental do ExInt</i>	73

CAPÍTULO 4: O POSSÍVEL PAPEL DO FENÔMENO DE <i>EXON-SHUFFLING</i>	
NA EVOLUÇÃO DE PROTEÍNAS HUMANAS QUE CONTÊM PEPTÍDEO SINAL	75
4.1. Manuscrito intitulado: <i>A possible role of exon-shuffling in the evolution of signal peptides of human proteins</i>	78
4.2. Material suplementar do manuscrito - <i>A possible role of exon-shuffling in the evolution of signal peptides of human proteins</i>	79
CAPÍTULO 5: DISCUSSÃO	80
5.1. Correlação do sinal de <i>exon-shuffling</i> antigo e moderno com a distribuição de domínios antigos e modernos ao longo das proteínas	82
5.1.1. <i>Distribuição espacial de domínios antigos e modernos ao longo das proteínas</i>	84
5.1.2. <i>Correspondência entre núcleos funcionais e domínios</i>	85
5.1.3. <i>Seleção de shuffling de exons 1-1</i>	86
5.2. O papel do fenômeno de <i>exon-shuffling</i> em proteínas humanas contendo peptídeo sinal	87
5.2.1. <i>Semi-exon-shuffling</i>	91
5.2.2. <i>Domínios e seqüências independentes pequenas em proteínas multi-domínios</i>	93
5.3. Conclusões	94

GLOSSÁRIO	95
BIBLIOGRAFIA	98
ANEXOS	112
1. Artigo de revisão intitulado <i>Insights on the origin and evolution of introns.</i>	113
2. <i>Curriculum Vitae</i>	114

CAPÍTULO 1: INTRODUÇÃO GERAL

1.1. Ítrons

De acordo com o dogma central da biologia molecular, proposto por Francis Crick (1970), a partir de cada gene (que codifica aminoácidos) é produzida uma cadeia polipeptídica em um fluxo unidirecional da informação (Figura 1.1). Algumas descobertas posteriores como duplicação de vírus de RNA, transcrição reversa representam exceções ao dogma proposto. No entanto, o fluxo é unidirecional em grande parte dos genes e ocorre através de dois processos seqüenciais: a transcrição e tradução (Figura 1.1) (Lewin 2000b). Basicamente, os genes são primeiro transcritos em uma molécula de RNA mensageiro (mRNA) e depois traduzidos em uma cadeia polipeptídica pelos ribossomos, onde cada três nucleotídeos (um códon) codificam um aminoácido (Lewin 2000b). Cabe ressaltar que alguns genes não seguem o dogma estabelecido porque não são traduzidos em uma seqüência de aminoácidos.

Figura 1.1: Esquema representativo da produção de uma cadeia polipeptídica a partir de um gene.

Além disso, este processo de transcrição e tradução não é exatamente igual em todos os organismos vivos. Diferentemente dos procariotos, no qual a maioria dos genes são constituídos por uma extensão de DNA colinear à proteína,

os genes de eucariotos são constituídos por seqüências de DNA de duas categorias distintas: os éxons e os íntrons (Berget et al. 2000; Chow et al. 1977; Gilbert 1978; Lewin 2000c). Em eucariotos, cada gene é primeiramente transcrito em uma molécula de RNA primário que é posteriormente processada em um mRNA maduro (Lewin 2000c). Os íntrons são removidos para formação da molécula de RNA mensageiro maduro em um processo denominado “splicing” (Sharp 1994). Desta maneira, os éxons são definidos como seqüências do gene presentes no mRNA maduro, que normalmente codificam aminoácidos, enquanto os íntrons são seqüências intervenientes não-codificantes, presentes somente no transcrito primário (Sharp 1994). Os íntrons podem ser classificados em diferentes grupos, de acordo com sua função e estrutura (Lewin 2000c).

1.1.1. *Íntrons de splicing nuclear*

Estes íntrons são encontrados na maioria dos eucariotos e são caracterizados principalmente por serem dependentes de um complexo ribonucleoprotéico denominado *spliceosome* para o processamento do *splicing* (ver revisão em (Sharp 1994)). Normalmente, os íntrons são removidos dos precursores de mRNA através de um mecanismo que envolve duas reações de transesterificação (Figura 1.2) (Lewin 2000c). A primeira consiste no ataque nucleofílico de um grupamento 2'-OH de uma adenosina (localizada internamente ao íntron) à ligação fosfodiéster de uma guanina invariante na extremidade 5' do íntron. Esta adenosina está localizada dentro de uma seqüência consenso chamada de ponto de ramificação próxima a extremidade 3' do próprio íntron. Esta reação forma uma ligação fosfodiéster 2'-5', resultando em uma estrutura

conhecida como “laço”. A segunda reação envolve o ataque nucleofílico da extremidade 3' do éxon anterior ao íntron. O éxon anterior apresenta, após a primeira reação, um grupamento OH livre, que ataca a ligação fosfodiéster da extremidade 3' do íntron. Desta maneira, as duas reações permitem a união dos dois éxons e a liberação do íntron intermediário (Figura 1.2) (Lewin 2000c).

Figura 1.2: *Splicing* nuclear ocorre através de duas reações de transesterificação, onde a hidroxila ataca a ligação fosfodiéster. Adaptada da Figura 22.7 de Lewin (2000c).

1.1.2. Íntrons do Grupo I

Estes íntrons são encontrados em genes que codificam rRNA no núcleo de eucariotos inferiores, em genes mitocondriais de fungo, em genes do fago T4 e em eubactérias (ver revisão em (Long & de Souza 1998)). Estes íntrons têm a capacidade de promover reações de *auto-splicing*. Sua atividade catalítica é consequência de sua estrutura secundária característica e da ação de uma guanossina livre, que promove o ataque nucleofílico à extremidade 5' do íntron que será excisado (Figura 1.3) (Lewin 2000a; Lewin 2000c).

Figura 1.3: Reações de transesterificação que ocorrem durante o *splicing* de íntrons do grupo I. Adaptada da Figura 22.15 de Lewin (2000c) .

1.1.3 Íntrons do Grupo II

Este íntrons são encontrados em genes mitocondriais de fungo, genes mitocondriais e de cloroplasto de vegetais superiores e genes de eubactérias (ver revisão em (Long & de Souza 1998)). Recentemente, foram encontrados íntrons

do grupo II no genoma da arqueobactéria *Methanosarcina acetivorans* (Dai & Zimmerly 2003). O mecanismo de *splicing* destes íntrons compartilha algumas similaridades com o dos íntrons de *splicing* nuclear, o que sugere que ambos possam ter um ancestral em comum (Rogers 1989). A principal semelhança é a formação de uma estrutura em forma de laço resultado da ligação fosfodiéster 2'-5' entre a extremidade 5' do íntron e uma adenosina no meio do íntron. No entanto, os íntrons do grupo II têm a capacidade de promover o *auto-splicing*. Durante o processo, uma estrutura secundária é formada, na qual dois domínios são justapostos de maneira que a primeira transesterificação possa ocorrer (Figura 1.4) (Lewin 2000a; Lewin 2000c).

Figura 1.4: Reações de transesterificação que ocorrem em *splicing* de íntrons do grupo II. Adaptada da Figura 22.15 de Lewin (2000c).

1.2. O fenômeno de *Exon-shuffling*

Seqüências de genes não homólogos podem sofrer recombinação durante a divisão celular, gerando novas seqüências, que podem ou não ser funcionais (ver mecanismos adiante) (Gilbert 1978; Graur & Li 2000b). A recombinação entre íntrons de genes não homólogos permite e aumenta a freqüência da troca de éxons inteiros e, conseqüentemente, aumenta a probabilidade de se formarem proteínas novas funcionais (Gilbert 1978; Gilbert 1987). Este fenômeno foi denominado “exon-shuffling” por Gilbert (Gilbert 1978; Gilbert 1987) (Figura 1.5). A existência de proteínas codificadas por genes que contêm regiões encontradas em genes não homólogos, chamadas de mosaicos, indica que o fenômeno de *exon-shuffling* ocorreu durante a história evolutiva destes genes (Graur and Li 2000b).

Figura 1.5: Esquema representativo de um evento de *exon-shuffling*. Através da recombinação intrônica entre os genes 1 e 2, o segundo éxon do gene 1 é inserido na seqüência do gene 2 formando um novo gene. Caixas representam éxons e linhas representam íntrons.

Vários estudos identificaram proteínas nas quais o fenômeno de *exon-shuffling* teve grande importância durante sua criação, como proteases fibrolíticas,

proteínas da coagulação sanguínea, proteínas das cascatas complementares da coagulação sanguínea e proteínas de matriz extracelular (Patthy 1999; Patthy 2003). Estes exemplos demonstram claramente como a combinação de éxons provenientes de genes diferentes, representando subunidades estruturais funcionais, podem gerar, a partir da recombinação não homóloga de íntrons, genes novos com novas funções (Patthy 1999; Patthy 2003). Além disso, estes exemplos mostram que íntrons são capazes de auxiliar e possivelmente acelerar a formação destas novas combinações (Gilbert 1978; Patthy 2003). Estes estudos também demonstraram que *shuffling* de éxons 1-1 foram importantes na formação de proteínas de organismos multicelulares durante a expansão de metazoários no período Cambriano (descrição mais detalhada na seção 1.7.3) (Patthy 1996; Patthy 1999; Patthy 2003).

São conhecidos dois mecanismos diferentes que podem viabilizar a troca de éxons: a recombinação ilegítima e a recombinação mediada por elemento repetitivo LINE-1 (L1) (Gilbert 1978; Gilbert 1987; Moran et al. 1999; van Rijk et al. 1999; van Rijk et al. 2000). O primeiro mecanismo consiste na recombinação entre seqüências de DNA não homólogas. Bloemendal e colaboradores (van Rijk et al. 1999; van Rijk et al. 2000) demonstraram que a duplicação intragênica de éxons de um gene chamado *alphaA-crystallin* resultou na formação de uma nova estrutura gênica através da recombinação ilegítima entre dois sítios CCCAT localizados em dois íntrons diferentes. Eles demonstraram que a seqüência CAT é preferencialmente cortada pela enzima topoisomerase 1, o que pode levar à recombinação ilegítima dos dois sítios CCCAT. O segundo mecanismo descrito como potencialmente eficiente para recombinação em genes de mamíferos envolve a presença do elemento de retrotransposição L1, que é transcrito

reversamente e pode se duplicar, inserindo novas cópias dentro do genoma (Moran et al. 1999). Este retroelemento possui um sinal fraco de término de transcrição, o que muitas vezes resulta na transcrição de uma parte do genoma localizada após a região 3' do elemento (Moran et al. 1999). Dependendo da posição do elemento L1, éxons subseqüentes podem ser transcritos com o elemento e recombinados com éxons do gene receptor (Moran et al. 1999).

1.3. Origem dos íntrons

Desde a descoberta dos íntrons (Berget et al. 2000; Chow et al. 1977), sua origem e função vêm sendo discutida. As perguntas principais são: qual a função dos íntrons, quando e como eles surgiram e porque íntrons de *splicing* nuclear (íntrons mais comumente encontrados) existem em eucariotos e não são encontrados em procariotos. Basicamente, existem duas hipóteses para explicar a origem dos íntrons: “íntrons-early” e “íntrons-late” (Figura 1.6) (Cavalier-Smith 1991; Doolittle 1978; Doolittle & Stoltzfus 1993; Gilbert 1978; Gilbert 1987; Palmer & Logsdon 1991). A primeira hipótese assume que os íntrons já estavam presentes nos genes do progenoto (ancestral comum de procariotos e eucariotos) (Doolittle 1978; Gilbert 1978; Gilbert 1987). Os íntrons teriam sido perdidos nos procariotos por um processo de redução do tamanho do genoma resultado da pressão seletiva em favor da replicação rápida do DNA (Doolittle 1978). Já a hipótese *íntrons-late* assume que os íntrons foram adicionados posteriormente durante a evolução, após a divergência entre procariotos e eucariotos (Cavalier-Smith 1991; Doolittle & Stoltzfus 1993; Palmer & Logsdon 1991; Stoltzfus et al. 1994).

Como o fenômeno de *exon-shuffling* (a troca de éxons inteiros entre genes não relacionados) ocorre mais frequentemente na presença de íntrons, as duas hipóteses também discordam sobre o período da evolução dos organismos em que este fenômeno ocorreu (Figura 1.6). A hipótese *introns-early* assume que os íntrons e os éxons formavam os primeiros genes e que este tipo de organização estrutural permitiria e facilitaria a formação de novos genes através do fenômeno de *exon-shuffling* (Blake 1978; Doolittle 1978; Gilbert 1978; Gilbert 1987). Desta maneira, o fenômeno de *exon-shuffling* teria um papel importante na criação das novas proteínas no início da evolução dos organismos (Doolittle 1978; Gilbert 1987). Já a hipótese *introns-late* argumenta que *exon-shuffling* é um fenômeno recente, importante para a variabilidade protéica somente em eucariotos (Doolittle & Stoltzfus 1993; Patthy 1996; Patthy 1999).

Figura 1.6: Esquema representativo dos passos evolutivos assumidos pelas duas hipóteses sobre a origem dos íntrons e do fenômeno de *exon-shuffling*.

O fato da existência de íntrons possibilitar e aumentar de frequência a recombinação entre éxons e com isso acelerar a evolução através da criação de novas proteínas não significa que os íntrons foram selecionados para isso

(Gilbert et al. 1997). Tal argumento não é compatível com as teorias de evolução por Seleção Natural já que a presença dos íntrons não pode ser explicada por vantagem adaptativa selecionada posteriormente à sua criação (Gilbert et al. 1997). A idéia de que íntrons favorecem a criação de novas proteínas através de *exon-shuffling*, tanto no progenoto quanto durante a evolução de eucariotos, está baseada na utilização de tais elementos antes da sua remoção por pressão seletiva negativa (Gilbert et al. 1997). Apesar dos íntrons não estarem normalmente sobre pressão seletiva como as seqüências codificantes, a redução do tamanho dos genomas pode ter sido selecionada em favor da rápida replicação do DNA em procariotos e outros organismos de rápido ciclo de vida, resultando na perda dos íntrons (Doolittle 1978).

Se os íntrons já existiam no progenoto, estes não deveriam ser similares aos íntrons conhecidos como nucleares e dependentes do *spliceosome*, já que os genes codificadores de moléculas deste complexo não são observados em eubactérias e arqueobactérias (Cavalier-Smith 1991). Os defensores da hipótese *introns-early* argumentam que os íntrons mais prováveis de terem existido no progenoto seriam similares aos íntrons capazes de exercer a função de *auto-splicing* como os íntrons do grupo I e II (Gilbert 1987; Rogers 1989). No entanto, muitos defensores da hipótese *introns-late* argumentam que não seria possível a ocorrência de *exon-shuffling* através da recombinação de tais íntrons, já que estes possuem seqüências internas específicas e que a perda de tais seqüências durante o processo de troca de éxons impediria o exercício da função de *auto-splicing* (Patthy 1987). No entanto, observou-se alguns casos de *shuffling* de éxons no bacteriófago T4 através da recombinação de íntrons que sofrem *auto-splicing*. Esta troca de éxons produziu um gene quimérico contendo um intron

híbrido ainda capaz de exercer a atividade de *auto-splicing* (ver revisão em (Long & de Souza 1998)). Tal resultado contraria o argumento dos defensores da hipótese *introns-late* porque mostra que a recombinação de introns do grupo I não necessariamente compromete sua atividade de *auto-splicing*.

As hipóteses *introns-early* e *introns-late* estão baseadas em propriedades dos íntrons e dos éxons observados na natureza. Entretanto, quais são as características que podem ser analisadas para testar as duas hipóteses? A seguir, estão descritas algumas dessas características e quais os resultados que cada uma das duas hipóteses espera observar em relação a elas.

1.4. Características dos íntrons

1.4.1. Similaridade de seqüências

A comparação entre as seqüências de introns de genes ortólogos (homólogos de espécies diferentes) poderia ser uma característica útil para determinar quais são os íntrons antigos (presentes em linhagens distantes evolutivamente) e os íntrons novos (inseridos somente em grupos evolutivamente próximos). Entretanto, como os íntrons apresentam uma alta taxa de mutação (aproximadamente 4×10^{-9} substituições por sítio por ano (Graur & Li 2000a)), a similaridade entre seqüências de espécies distantes não é possível de ser detectada.

1.4.2. Posição e fase de íntrons

Por outro lado, a posição dos íntrons, isto é, a sua localização em relação a seqüência de aminoácidos da proteína codificada pelo gene em questão, pode ser conservada entre espécies se os íntrons já existiam no ancestral em comum a elas (Fedorov et al. 2002; Patthy 1987; Rogozin et al. 2003). Se os íntrons já estavam presentes no progenoto como argumentado pelos defensores da hipótese *introns-early*, suas posições em um gene devem estar conservadas entre diferentes espécies (Figura 1.7). No entanto, se os íntrons foram inseridos aleatoriamente em seqüências eucarióticas durante a evolução, como assume a hipótese *introns-late*, não seria esperado encontrar posições correspondentes em genes ortólogos de organismos distantes.

Figura 1.7: Exemplo hipotético de íntrons inseridos recentemente e íntrons antigos. À esquerda, a árvore filogenética das espécies e à direita, o mapa de posição dos íntrons antigos e recentes no gene A.

Além da posição dos íntrons, a fase dos íntrons também pode ser conservada em espécies evolutivamente distantes (Fedorov et al. 2002; Long et al. 1995). A fase do íntron é definida pela posição em que este está localizado dentro dos códons (Figura 1.8) (Sharp 1981). Íntrons de fase zero estão localizados entre códons; íntrons de fase um estão localizados entre o primeiro e segundo nucleotídeo; e íntrons de fase dois entre o segundo e o terceiro nucleotídeo (Figura 1.8) (Sharp 1981).

Figura 1.8: Tipos de íntrons diferentes de acordo com suas fases. Íntrons de fase zero estão localizados entre dois códons; íntrons de fase um estão posicionados entre o primeiro e o segundo nucleotídeo; íntrons de fase dois estão localizados entre o segundo e o terceiro nucleotídeo. Éxons estão representados por retângulos e íntrons por linhas.

1.5. Características dos éxons

Outra característica que pode ser analisada é a simetria dos éxons, ou seja, a correlação entre fases dos íntrons adjacentes a um mesmo éxon (Patthy 1987). Éxons simétricos são aqueles flanqueados por íntrons de mesma fase, e.g., 0-0, 1-1 ou 2-2. Já éxons não simétricos são aqueles flanqueados por íntrons de fases diferentes (Figura 1.9).

Figura 1.9: Éxons simétricos e não simétricos. Retângulos estão representados por éxons e íntrons por linhas. Os números correspondem às fases dos íntrons que flanqueiam os éxons (representados por retângulos escuros).

Esta característica está altamente relacionada com o fenômeno do *exon-shuffling* (Patthy 1987). Éxons podem ser trocados entre genes, mas isso não significa que será produzida uma proteína funcional. A troca de éxons flanqueados por íntrons de fases diferentes (não simétricos) tem menor probabilidade de gerar um gene funcional devido à mudança ocasionada no quadro de leitura (Patthy 1987).

A Figura 1.10 mostra as proteínas produzidas a partir de *shuffling* de éxons simétricos e não simétricos através de três exemplos diferentes (Patthy 1987). A inserção do éxon simétrico (0-0) do gene 1 dentro do gene 2 através da recombinação de íntrons de mesma fase (zero) pode produzir uma nova proteína (Figura 1.10A). Esta é codificada por um gene contendo todos os éxons do gene 2, mais a mesma seqüência do éxon inserido proveniente do gene 1 que irá codificar os mesmos aminoácidos que eram codificados no gene 1.

Se ocorresse a inserção de um éxon simétrico, mas flanqueado por íntrons de fase um (Figura 1.10B), a nova proteína formada também conteria o éxon proveniente do gene 1 (mesma composição nucleotídica), porém, este éxon codificaria aminoácidos em um quadro de leitura diferente daquele codificado pelo gene 1.

Por fim, se o éxon inserido fosse não simétrico (Figura 1.10C), a proteína resultante do processo de *exon-shuffling* não apresentaria somente modificação do quadro de leitura do éxon inserido. Todo o quadro de leitura dos éxon subseqüentes ao éxon inserido seria modificado.

Assim, espera-se que a probabilidade de formar uma proteína funcional através de *exon-shuffling* seja bem mais alta para éxons simétricos do que para éxons não simétricos. Isto ocorre porque os éxons simétricos sempre contêm nucleotídeos em um número múltiplo de três e sua inserção nunca muda o quadro de leitura dos éxons subseqüentes (Patthy 1987).

Figura 1.10: Esquema representativo de genes gerados através de *exon-shuffling* entre o gene 1 e o gene 2. Um dos éxons do gene 1 é inserido no gene 2 através da recombinação intrônica, configurando um evento de *exon-shuffling*. O esquema mostra os aminoácidos codificados por cada novo gene gerado (A, B e C). A troca de éxons não simétricos muda completamente o quadro de leitura do éxon trocado e do restante da sequência do gene receptor. Figura e legenda reproduzidas da Figura 1 de Vibrantovski & de Souza (2004).

Segundo a hipótese *introns-early*, éxons simétricos devem ser encontrados em excesso mesmo em genes antigos (aqueles comuns a procariotos e eucariotos), já que o fenômeno de *exon-shuffling* já ocorria no progenoto, pois os íntrons já estavam presentes em seus genes (de Souza et al. 1998; Doolittle 1978; Gilbert 1978; Gilbert 1987; Long et al. 1995). Seus defensores também argumentam que a troca de éxons foi responsável por grande parte do aumento da variabilidade gênica naquele período (Doolittle 1978; Gilbert 1987).

Contrariamente, a hipótese *introns-late* não prediz nenhum tipo de excesso de éxons simétricos em genes antigos porque acredita que os íntrons foram

inseridos aleatoriamente em seqüências exônicas e não correspondem a unidades independentes de *shuffling* (Cavalier-Smith 1991; Palmer & Logsdon 1991).

Além disso, as duas hipóteses têm predições diferentes quanto à freqüência das fases dos introns em genes antigos. Se um tipo de éxon simétrico sofrer *shuffling* inicialmente mais freqüentemente que os outros tipos de éxons, espera-se que os íntrons da mesma fase que flanqueiam o éxon trocado aumentem de freqüência na população. Isto porque o *shuffling* de éxons simétricos (através da recombinação entre íntrons de mesma fase - Figura 1.10A) tem maior probabilidade de produzir uma proteína funcional. Desta forma, a hipótese *introns-early* espera observar freqüência desigual de fase de introns em genes antigos (de Souza et al. 1998; Long et al. 1995). Já a versão mais simples da hipótese *introns-late* prevê a distribuição igual das três fases dos íntrons em genes antigos devido à inserção aleatória destes em seqüências contínuas durante a evolução de eucariotos (Cavalier-Smith 1991; Long et al. 1995; Palmer & Logsdon 1991).

Cabe ressaltar que, posteriormente, a hipótese *introns-late* incorporou a possibilidade de inserção preferencial de íntrons para explicar diferença de freqüência de fase de íntrons observada na natureza. Esta nova argumentação, as observações que a suportam e a discussão sobre elas estão descritas na seção sobre sítios de “proto splice” (seção 1.9) devido à ordem cronológica do debate que está sendo apresentada durante esta tese.

1.6. Análises e testes já realizados relacionados às características dos íntrons e dos éxons

1.6.1. Análises filogenéticas de posições de íntrons

Muitos trabalhos identificaram genes contendo íntrons conservados na mesma posição em várias espécies diferentes (Fedorov et al. 2002; Gilbert et al. 1986; Rogozin et al. 2003). Ao mesmo tempo, foi relatado que outras posições intrônicas são únicas em uma espécie ou em um gênero em particular (Kwiatowski et al. 1995; Logsdon et al. 1995; Rogozin et al. 2003). Por exemplo, em genes que codificam a enzima trifosfato isomerase observou-se que muitas posições de íntrons são conservadas entre vários organismos (Gilbert et al. 1986). No entanto, análises filogenéticas demonstraram que outros íntrons foram inseridos nestes genes em posições diferentes ao longo da evolução de eucariotos (Kwiatowski et al. 1995; Logsdon et al. 1995).

Sabe-se que os íntrons são capazes de mudar de posição ao longo da evolução, deslocando-se até cerca de 15 nucleotídeos por um processo chamado *intron sliding* (Graur & Li 2000b). A partir das evidências em favor do *intron sliding*, os defensores da hipótese *introns-early* questionaram se a correspondência da posição de íntrons de genes ortólogos seria uma característica confiável para testar as duas hipóteses. Segundo estes defensores, íntrons em posições próximas em genes ortólogos poderiam representar tanto íntrons com origem comum que sofreram *sliding* quanto íntrons inseridos em momentos diferentes da evolução destes organismos.

Apesar destas críticas, Stoltzfus e colaboradores (1997) mostraram que a conservação de posição de íntrons em genes ortólogos pode ser utilizada para determinação de íntrons antigos e recentes. Eles demonstraram através de análises filogenéticas que o processo de mudança de posição dos íntrons não é suficiente para explicar a diferença de posição de 205 íntrons entre espécies diferentes para cinco proteínas estudadas. Tal estudo, somado a outros estudos posteriores, dão suporte à idéia de que muitos íntrons observados atualmente foram inseridos ao longo da evolução (Coghlan & Wolfe 2004; Dibb & Newman 1989; Kwiatowski et al. 1995; Logsdon et al. 1995; Qiu et al. 2004).

1.6.2. Análises das regiões ancestrais conservadas

É importante ressaltar que as duas hipóteses sobre a origem dos íntrons concordam que eles existem desde o início da evolução de eucariotos e que *exon-shuffling* é um fenômeno importante na formação de proteínas durante este período, chamadas de proteínas modernas.

O ponto de discordância entre as duas hipóteses é a presença de íntrons e do fenômeno de *exon-shuffling* no progenoto. Desta maneira, as duas hipóteses discordam quanto a algumas características dos éxons e íntrons presentes em genes antigos que são aqueles que já existiam antes da divergência entre procariotos e eucariotos.

Regiões de genes eucarióticos homólogas a seqüências procarióticas representam as regiões existentes nos genes atuais que existiam no progenoto e são denominadas regiões ancestrais conservadas (RACs) (Green et al. 1993). As RACs normalmente correspondem a domínios específicos (ou motivos) presentes

em várias proteínas, mas podem ainda compreender a maior parte ou toda a sequência de uma proteína altamente conservada (Green et al. 1993).

Já as outras regiões (que não correspondem as RACs) podem ter sido adicionadas nos genes ao longo da evolução individual das linhagens de eucariotos e procariotos através de *exon-shuffling*, duplicação ou outros processos.

As discordâncias entre as duas hipóteses recaem sobre as análises de características dos éxons e íntrons nas regiões antigas conservadas (RACs) (Long et al. 1995). Segundo a hipótese *introns-late*, os íntrons e o fenômeno de *exon-shuffling* não existiam no progenoto (Cavalier-Smith 1991; Doolittle & Stoltzfus 1993; Palmer & Logsdon 1991). Logo, todos os íntrons localizados nestas regiões conservadas teriam sido inseridos ao longo da evolução de eucariotos. Estes íntrons não poderiam estar envolvidos em eventos de *exon-shuffling* nos eucariotos porque estas regiões são colineares entre procariotos e eucariotos, ou seja, não apresentam deleções ou inserções consideráveis (Long et al. 1995). Se os íntrons foram inseridos aleatoriamente em sequências contínuas como argumenta os defensores da hipótese *introns-late*, não é esperado observar excesso de nenhum tipo de fase de íntron ou excesso de éxons simétricos nas RACs, pois estas características são evidências e sinais de eventos de *exon-shuffling* (Long et al. 1995).

Já para hipótese *introns-early*, sinais de *exon-shuffling* como excesso de éxons simétricos e abundância de um tipo de íntron nas RACs seria uma evidência da existência dos íntrons e do fenômeno de *exon-shuffling* antes da divergência de procariotos e eucariotos (Long et al. 1995).

1.6.2.1. Excesso de íntrons de fase zero

Nos últimos quinze anos, muitos trabalhos analisaram a frequência das fases dos íntrons nos genes inteiros, observando que existe excesso de íntrons de fase zero (de Souza et al. 1998; Fedorov et al. 1992; Long et al. 1995; Sakharkar et al. 2000; Sakharkar et al. 2002). Íntrons de fase zero representam 50% dos íntrons existentes na natureza, enquanto que os íntrons de fase um e fase dois representam 30 e 20%, respectivamente (Long et al. 1995). Este excesso poderia ser sinal de que o fenômeno de *exon-shuffling* teria ocorrido ao longo da evolução de eucariotos, já que os genes contêm regiões que podem ter sido adicionadas nos genes ao longo da evolução individual das linhagens de eucariotos e procariotos através de *exon-shuffling* (ver explicação na seção 1.6.2).

Long e colaboradores (1995) foram os primeiros a analisar tais características em regiões ancestrais conservadas. Eles observaram que nessas regiões, os íntrons de fase zero são os mais abundantes (50%), seguidos de íntrons de fase um e de íntrons de fase dois. Como as regiões ancestrais conservadas não sofreram *exon-shuffling* após a divergência de procariotos / eucariotos e se considerarmos que os íntrons são inseridos aleatoriamente, este resultado sugere que a diferença entre as frequências dos íntrons é produto de *shuffling* de éxons no progenoto. Principalmente de *shuffling* de éxons flanqueados por íntrons de fase zero que é o tipo de íntron observado em excesso (de Souza et al. 1998; Long et al. 1995).

1.6.2.2. Excesso de éxons simétricos

Em relação à distribuição de éxons simétricos, observou-se uma depleção de éxons não simétricos e um excesso de éxons simétricos quando genes inteiros foram analisados (de Souza et al. 1998; Long et al. 1995; Sakharkar et al. 2000; Sakharkar et al. 2002). Este excesso significativo é observado em todos os três tipos de éxons simétricos: 10% de excesso em éxons 0-0; 23% de excesso em éxons 1-1 e 12% de excesso em éxons 2-2 (de Souza et al. 1998; Long et al. 1995). Este excesso observado em todos os genes mostra que *exon-shuffling* é um fenômeno que deve ter ocorrido ao longo da evolução dos organismos, mas não necessariamente mostra que o fenômeno ocorreu no progenoto.

Long e colaboradores (1995) observaram também que estes excessos são encontrados nas regiões ancestrais conservadas, mostrando que este resultado deve ser produto de *exon-shuffling* ocorrido no progenoto.

Long e colaboradores (1995) mostraram que não somente éxons individuais simétricos (Figura 1.11A) são observados em excesso, mas conjuntos de éxons flanqueados por íntrons de mesma fase (Figura 1.11B) são encontrados em excesso em todos os genes e dentro das regiões ancestrais conservadas. Tal resultado está de acordo com o esperado pelo processo de *exon-shuffling*, já que as unidades de *shuffling* não necessariamente devem ser éxons individuais. Conjuntos de éxons podem também sofrer eventos de *shuffling* (Long et al. 1995).

As primeiras observações sobre o excesso de éxons simétricos em RACs representaram um importante ponto a favor da hipótese *introns-early* e juntamente com os dados de excesso de íntrons de fase zero sugeriram que uma

parte deste tipo de íntrons (fase zero) observados nos genes atuais são ancestrais (de Souza et al. 1998; Long et al. 1995).

Figura 1.11: Exemplo hipotético de éxon simétrico (A) e conjunto de éxons simétricos (B). Retângulos estão representados por éxons e íntrons por linhas. Os números correspondem às fases dos íntrons. Em vermelho estão representadas as fases dos íntrons que flanqueiam éxons ou conjuntos de éxons simétricos (retângulos escuros).

1.7. *Exon-shuffling* e unidades protéicas independentes

Outro ponto bastante discutido no debate sobre a origem dos íntrons é a correlação entre éxons e unidades protéicas independentes. Blake (1978) foi o primeiro a mencionar que se novas combinações de éxons são capazes de produzir novas proteínas, éxons devem corresponder a unidades estruturais independentes, ou seja, a domínios, combinações de domínios ou outras unidades estruturais. A hipótese *introns-early* acredita que éxons de genes antigos correspondem a unidades protéicas independentes (Blake 1978; Doolittle 1978; Gilbert 1987). A troca dos éxons que codificam estas unidades teriam formado proteínas novas funcionais no progenoto. Para a hipótese *introns-late*, éxons não necessariamente correspondem a unidades protéicas em seqüências

antigas já que os íntrons teriam sido inseridos em seqüências exônicas contínuas (Cavalier-Smith 1991; Doolittle & Stoltzfus 1993; Palmer & Logsdon 1991).

1.7.1 Exon-shuffling e módulos protéicos

Módulo protéico é considerado um tipo de unidade protéica independente (Graur & Li 2000b). Módulos são unidades definidas por um esqueleto estrutural protéico que cabe dentro de uma esfera de um determinado diâmetro (Go 1981; Graur & Li 2000b). Go (1981) observou correspondência entre éxons do gene de hemoglobina e módulos de 28Å (Angstroms) de diâmetro provenientes da estrutura terciária da proteína codificada por este gene (Figura 1.12). Após este trabalho, muitos outros foram publicados mostrando a correspondência entre unidades compactas estruturais (módulos) e éxons usando outras proteínas como: lisozima (Go 1983), triosefosfato isomerase (Gilbert et al. 1986), gliceraldeído fosfato desidrogenase (Long et al. 1996) e outras. Estes trabalhos também se basearam na determinação dos módulos protéicos pelo método de plotagem de Go (Go 1981), onde primeiramente é estipulado um valor máximo de distância entre carbonos alfa da cadeia polipeptídica que configura um módulo. Através da plotagem gráfica destas distâncias para cada cadeia, são definidos os maiores módulos protéicos em que as distâncias entre os carbonos alfa são menores do que o valor estipulado. Da mesma forma, são definidas as regiões de ligação entre os módulos. A correlação entre módulos e éxons é feita através da observação de íntrons nas regiões de ligação (de Souza et al. 1996).

Figura 1.12: Gráfico de Go (Go 1981) para hemoglobina. As regiões escuras representam pares de carbono alfa que estão separados por 28Å ou mais de distância. Cinco módulos foram identificados pelos triângulos largos ao longo da diagonal, cujos tamanhos são limitados pelas regiões escuras. As regiões de ligação (LR) são definidas pela interseção desses triângulos. Figura e legenda reproduzida de de Souza e colaboradores (1996).

A identificação de vários íntrons que se posicionavam nas regiões de borda dos módulos preditos constituiu o conjunto de dados que suportava a predição da hipótese *introns-early*. Concomitantemente, houve um grande número de descrições de íntrons de vários organismos localizados em posições que não correspondiam a estas bordas. Foram analisadas 62 posições de íntrons em quatro proteínas diferentes (álcool desidrogenase, globina, piruvato quinase e triosefosfato isomerase) e não foram observadas evidências de correlação entre muitas posições de íntrons e bordas de estrutura secundária, módulos ou domínios globulares (Stoltzfus et al. 1994).

Nenhuma destas evidências, até então, constituía um conjunto de dados com tamanho amostral suficientemente grande para dar suporte estatístico

definitivo a nenhuma das duas hipóteses. Não existia mais espaço no debate para exemplos em genes individuais de correlação entre éxons e unidades estruturais protéicas independentes. A discussão requeria uma análise em larga escala com suporte estatístico significativo.

Em 1996, com o aumento da quantidade de genomas seqüenciados, de seqüências de DNA disponíveis na Internet e de estruturas resolvidas no PDB (Protein Data Bank) (Westbrook et al. 2003), demonstrou-se correlação significativa entre módulos protéicos e éxons em um conjunto de proteínas antigas (de Souza et al. 1996). Neste trabalho, os autores definiram os módulos protéicos pelo método de plotagem de Go (1981) e testaram a relação entre as regiões de ligação entre módulos e o posicionamento dos íntrons de um conjunto de proteínas de eucariotos homólogas a seqüências gênicas procarióticas (de Souza et al. 1996). Quando utilizaram *a priori* módulos de 28 Å de distância, definidos anteriormente nos estudos com hemoglobina, encontram preferência significativa de posicionamento de íntrons nas regiões de ligação. Posteriormente, variando o tamanho máximo dessas distâncias, encontraram a mesma tendência significativa em outros dois tamanhos de módulos (de Souza et al. 1996). Dois trabalhos similares foram realizados com tamanhos amostrais maiores nos anos subseqüentes (de Souza et al. 1998; Fedorov et al. 2001). No primeiro trabalho, os autores, ao subdividirem a análise de posicionamento de íntrons em grupos diferentes de acordo com suas fases, observaram correlação somente entre íntrons de fase zero e regiões de ligação de módulos de proteínas ancestrais. Nenhuma correlação significativa entre as outras duas fases de introns e as regiões de ligação foi encontrada (de Souza et al. 1998). Este resultado sugere que os íntrons de fase zero estavam envolvidos no fenômeno de *exon-shuffling*

em proteínas antigas porque mostra evidências de correspondência entre exons 0-0 e unidades independentes de *shuffling* ancestrais (os módulos em proteínas antigas).

O segundo trabalho mostra que nenhuma correlação entre os íntrons de fase zero e as regiões de ligação entre módulos é encontrada no conjunto de proteínas que não possuem seqüências procarióticas correspondentes (chamadas de proteínas não-ancestrais) (Fedorov et al. 2001).

Estas observações quanto à correspondência de íntrons de fase zero e as regiões de ligação entre módulos antigos sugerem que os íntrons ancestrais (existentes no progenoto) correspondem a uma parte dos íntrons de fase zero observados nos genes atuais. Além disso, o *shuffling* de exons 0-0 não deve ter contribuído tanto na formação das proteínas modernas como deve ter contribuído na formação de proteínas antigas (Fedorov et al. 2001).

A partir destes resultados, os defensores da hipótese *introns-late* argumentaram que a correlação entre módulos estruturais e posicionamento dos íntrons pode ser resultado do processo pelo qual os íntrons foram inseridos (Endo et al. 2002). O argumento é baseado na idéia de que o processo de inserção de íntron poderia ser mutagênico. Isto é, poderia modificar a seqüência de aminoácidos produzida pelo gene no local da inserção. Desta maneira, seria esperado observar íntrons localizados preferencialmente em regiões pouco conservadas dos genes. Os defensores da hipótese *introns-late* argumentavam que as regiões de ligação entre os módulos são regiões menos conservadas do que as regiões dos centros dos mesmos em termos de composição de aminoácidos. Sendo assim, a inserção de íntrons nas regiões de ligação seria menos deletéria aos genes. Entretanto, Endo e colaboradores (2002) ao analisar

um grupo de proteínas antigas, mostraram que não há depleção de íntrons de fase zero nas regiões altamente conservadas, nem excesso nas regiões pouco conservadas. Desta maneira, não há suporte para o argumento de que íntrons localizados preferencialmente em regiões pouco conservadas seriam a base da correlação entre íntrons e regiões de ligação de módulos protéicos (Endo et al. 2002).

1.7.2. Exon-shuffling e domínios protéicos

Domínios protéicos são regiões bem definidas dentro de uma proteína que possuem uma função específica, como por exemplo, ligar-se a um substrato, ou constituir uma unidade estrutural independente, cujo enovelamento não depende de outras regiões da proteína da qual faz parte (Graur & Li 2000b). Domínios protéicos podem também ser definidos como as menores subunidades das proteínas que apresentam história evolutiva independente (e.g. PFam (Bateman et al. 2004)). Note porém que a maioria dos segmentos com história evolutiva autônoma são capazes de assumir sua estrutura e função independentemente de outros domínios, o que ilustra a forma como essas diferentes definições são relacionadas e complementares (Ponting & Russell 2002).

Muitos trabalhos mostraram a correspondência entre domínios e éxons em proteínas e em seus genes codificantes (Kaessmann et al. 2002; Patthy 1999; Patthy 2003). Em uma grande variedade de espécies eucarióticas, há excesso de domínios que apresentam bordas localizadas próximas as bordas de éxons, isto é, as bordas de domínios correspondem às bordas de éxons em um número de vezes maior do que o esperado ao acaso (Liu & Grigoriev 2004). Além disso,

estudos mostraram que domínios correspondentes a éxons são mais numerosos e mais amplamente distribuídos do que os não correspondentes (Liu et al. 2005). Estes domínios também co-ocorrem com outros domínios formando as chamadas proteínas mosaico (Liu et al. 2005; Tordai et al. 2005). Esses dados mostram que existe uma correlação significativa entre domínios e éxons e sugerem que estes devem corresponder às unidades de *shuffling* durante a criação das proteínas.

A identificação de casos de *shuffling* de éxons correspondentes a domínios se baseia em dois procedimentos i) identificação de dois domínios homólogos presentes em duas proteínas não homólogas; ii) a demonstração de que estes módulos foram transpostos através de recombinação mediada por íntrons (Patthy 1987). Em relação à fase dos íntrons que flanqueiam os domínios correspondentes a éxons, muitos estudos foram realizados. Estudos individuais identificaram muitos casos de *shuffling* de domínios correspondentes a éxons simétricos (ver explicação na seção 1.5) (Kaessmann et al. 2002; Liu et al. 2005; Liu & Grigoriev 2004; Patthy 1996; Patthy 1999; Tordai et al. 2005).

1.7.3 Exon-shuffling em eucariotos e domínios protéicos flanqueados por íntrons de fase um

Durantes alguns anos, Patthy identificou casos de *exon-shuffling* de módulos em diversas proteínas de matriz extracelular, na maioria das proteínas de regulação da coagulação sanguínea, fibronectinas e proteínas da ativação complementar, proteínas de adesão celular, receptores com atividade de tirosina quinases e receptores com atividade de tirosina fosfatase (ver revisão em (Patthy 1999; Patthy 2003)).

A maior parte dos módulos usados na construção destas proteínas são codificados por éxons 1-1 (Banyai & Patthy 2004; Patthy 1996; Patthy 1999; Patthy 2003). De fato, análises em larga escala mostraram que domínios correspondentes a éxons apresentam excesso de íntrons da mesma fase nas regiões flanqueadoras, principalmente de íntrons de fase um (Kaessmann et al. 2002; Patthy 1996; Patthy 2003; Tordai et al. 2005). Segundo as regras de compatibilidade de fase de íntrons para a ocorrência de *shuffling*, esta observação implica que a inserção destes módulos requer um íntron de fase um para mediar a recombinação entre genes (Patthy 1987). As predições sobre fase de íntron e simetria de éxons para tais eventos de *exon-shuffling* é a proliferação de íntrons de fase um e éxons 1-1 nestes genes. De fato, a classe de éxons simétricos observados em maior excesso em todos os genes é a de éxons flanqueados por íntrons de fase um (de Souza et al. 1998; Long et al. 1995; Sakharkar et al. 2002). Além de íntrons de fase um serem observados em maior frequência que íntrons de fase dois no geral, os primeiros são bastante abundantes nas proteínas estudadas por Patthy (Long et al. 1995; Patthy 2003; Tordai & Patthy 2004).

Proteínas compostas por domínios de classe 1-1 encontradas em vertebrados também foram observadas em esponjas, hidrozóia, moluscos, nematodos, artrópodos e equinodermas (Patthy 1999). Por outro lado, não existe praticamente nenhuma evidência da presença de tais proteínas em fungos e em plantas (Patthy 1999). A distribuição filogenética das proteínas contendo domínios 1-1 que evoluíram através de *shuffling* de éxons 1-1 sugere que tal fenômeno tenha se tornado importante para construção de proteínas durante a radiação dos metazoários (Patthy 1996; Patthy 1999). Patthy (1996, 1999, 2003)

sugere em seus trabalhos que a evolução acelerada de diversas espécies do filo dos metazoários no período rápido do Cambriano tenha tido contribuição significativa do fenômeno de *exon-shuffling*.

Desta maneira, *exon-shuffling* moderno é conhecido como evento de troca de éxons 1-1 ocorrido durante a evolução de metazoários na formação de novas proteínas importantes principalmente para formação de organismos multicelulares.

Kaessmann e colaboradores (2002) realizaram um elegante trabalho analisando a relação de éxons simétricos e domínios protéicos. Domínios podem ser codificados por um éxon ou por um conjunto de éxons (Figura 1.13A e 1.13B). Da mesma maneira, um éxon pode codificar um conjunto de domínios (Figura 1.13C). Considerando somente os casos de domínios que são codificados por vários éxons em genes humanos (Figura 1.13B), Kaessmann e colaboradores (2002) estudaram as fases dos íntrons que bordeiam os domínios e as fases dos íntrons internos aos mesmos. Eles observaram que os íntrons localizados nas bordas dos domínios apresentam excesso de combinações simétricas de fases (i.e., 0-0, 1-1, 2-2), enquanto que os éxons internos não apresentam nenhum excesso de simetria.

Figura 1.13: Ilustração de classes de domínio. (A) Domínio codificado por um éxon; (B) domínio codificado por um conjunto éxons; (C) conjunto de domínios codificado por apenas um éxon. Adaptado de Kaessmann e colaboradores (2002).

No mesmo trabalho, os autores observaram excesso de domínios correspondentes a éxons simétricos 1-1 e depleção de domínios correspondentes a éxons 0-0 quando analisaram somente domínios modernos (exclusivos de metazoários) (Kaessmann et al. 2002). Por outro lado, ao analisarem domínios antigos (presentes em metazoários e procariotos), o padrão contrário foi observado, ou seja, depleção de domínios correspondentes a éxons 1-1 e pequeno excesso de domínios correspondentes a éxons 0-0. Este dado é

coerente com outros trabalhos citados acima e com a hipótese de que a maioria dos íntrons antigos são de fase zero e estiveram envolvidos com o fenômeno de *exon-shuffling* em regiões antigas conservadas. É compatível também com a noção de que os íntrons de fase um e os eventos de *shuffling* de éxons 1-1 estão mais relacionados com eventos recentes (de Souza et al. 1996; de Souza et al. 1998; Long et al. 1995; Patthy 1999; Patthy 2003).

Todos esses resultados mostram a importância do fenômeno de *exon-shuffling* na formação de novos genes ao longo da evolução e da correspondência destas unidades de *shuffling* a unidades funcionais protéicas como os domínios. Mas, a contribuição mais importante ao debate em torno das hipóteses *introns-early* e *introns-late* foi a análise éxons ou conjuntos de éxons simétricos correspondentes a domínios modernos e antigos porque confirmou através de análises quantitativas que os éxons 1-1 estão envolvidos em fenômeno de *shuffling* recente, enquanto que éxons 0-0 provavelmente estiveram envolvidos em eventos antigos (Kaessmann et al. 2002).

1.8. Teoria sintética de evolução dos íntrons

Os estudos citados acima motivaram a introdução de uma teoria sintética de evolução dos íntrons (de Souza et al. 1998; de Souza 2003; Roy 2003). Esta teoria tem a intenção de conciliar diferentes aspectos das duas hipóteses originais, *introns-early* e *introns-late* (de Souza et al. 1998; de Souza 2003; Roy 2003). Ela postula que alguns dos íntrons de fase zero encontrados nos genes atuais são antigos, isto é, estavam presentes no progenoto, enquanto que os

demais (cerca de 65% de todos os íntrons) foram inseridos durante a evolução de eucariotos (de Souza et al. 1998).

O surgimento de uma teoria que combina as duas hipóteses para a origem dos íntrons, antiga e recente, é resultado do conjunto das observações acumuladas ao longo dos últimos anos. Por um lado, a distribuição filogenética de muitos íntrons fornece evidências irrefutáveis de que grande parte dos íntrons foi inserida ao longo da linhagem dos eucariotos (Dibb & Newman 1989; Qiu et al. 2004; Rogozin et al. 2003; Stoltzfus et al. 1997). Por outro lado, argumentos e dados como o excesso de éxons simétricos em regiões ancestrais conservadas e o excesso de íntrons de fase zero nas regiões de ligação entre módulos de genes antigos dão suporte à idéia de antiguidade do fenômeno de *exon-shuffling*, indicando que a hipótese de que pelo menos alguns íntrons de fase zero já existiam no progenoto, não pode ser descartada (de Souza et al. 1996; de Souza et al. 1998; Fedorov et al. 2001; Fedorov et al. 2003; Long et al. 1995).

1.9. Sítio de "proto-splice"

Por alguns anos, este debate sobre a origem dos íntrons parecia ter chegado a um consenso através da teoria sintética de evolução dos íntrons (de Souza et al. 1998; de Souza 2003; Roy 2003). No entanto, o conflito retomou a sua força após a publicação de alguns trabalhos relacionados à hipótese da existência de sítios de inserção preferencial de íntrons (Qiu et al. 2004; Sadusky et al. 2004; Sverdlov et al. 2004b).

Um dos pontos questionados pela hipótese *introns-late* em relação a distribuição de fase de íntrons é o suposto padrão não aleatório de inserção dos

íntrons ao longo da evolução. O argumento é baseado na possível existência de sítios de “proto-splice”, que são seqüências consenso (C/AAG|G/A), que seriam alvos para a inserção dos íntrons (“|” representa a posição de inserção) (Dibb 1991; Dibb & Newman 1989; Wolf et al. 2000). Dibb e Newman (1989) foram os primeiros a propor a existência de tal sítio em um estudo com íntrons presentes em genes de *Tubulina* e de *Actina* em diversas espécies. Eles identificaram mais de 60 íntrons que provavelmente foram ganhos durante a evolução e observaram uma seqüência consenso flanqueadora que chamaram de sítio de *proto-splice* (Dibb & Newman 1989). Esta seqüência consenso, na verdade, é muito similar ao consenso das bases nos éxons que flanqueiam todos os íntrons (Figura 1.14) (Stephens & Schneider 1992). Os autores sugerem que estas seqüências são realmente sítios de *proto-splice* porque estas bases flanqueadoras estão conservadas nas mesmas posições em outros membros dessa família gênica que não contêm íntrons nos sítios em questão (Dibb & Newman 1989). Embora exista a possibilidade destes sítios serem resquícios de perda de íntrons do ancestral, os autores argumentam que pela análise filogenética dos íntrons estudados é muito mais parcimonioso que estes íntrons tenham sido ganhos (Dibb & Newman 1989).

Figura 1.14: Representação gráfica desenvolvida por Tom Schneider e Mike Stephens do alinhamento múltiplo de nucleotídeos contidos nas bordas éxon-éxon de genes humanos (LOGO) (Schneider & Stephens 1990; Stephens & Schneider 1992). As posições negativas correspondem aos nucleotídeos do éxon 5' e as posições 0 e positivas correspondem aos do éxon 3' (em relação ao íntron).

Em relação ao mecanismo de inserção, acredita-se que o íntron excisado durante o processo de *splicing* ficaria associado aos componentes da maquinaria de *splicing* e seria inserido em um sítio de *proto-splice* de uma molécula de RNA diferente (Sharp 1985). A sequência intrônica seria adicionada à sequência genômica propriamente dita através da transcrição reversa e recombinação homóloga da molécula de RNA contendo íntron, que substituiria a cópia original (Sharp 1985). Tal mecanismo permitiria a inserção de um íntron em uma região exônica contendo os sinais necessários para a reação de *splicing*.

A possibilidade de íntrons serem inseridos não aleatoriamente em sequências pré-definidas (sítios de *proto-splice*) questiona a observação da frequência desigual de fase de íntrons (maior frequência de íntrons de fase zero). Esta distribuição poderia ser resultado de uma inserção não aleatória de íntrons (Qiu et al. 2004; Sadusky et al. 2004; Wolf et al. 2000). Long e colaboradores

(1998, 2000) demonstraram através de duas análises independentes que a distribuição de sítios de *proto-splice* observados atualmente não pode explicar a distribuição desigual de íntrons de fases diferentes, nem o excesso de éxons simétricos encontradas na natureza. A primeira análise consistiu em uma busca por sítios de *proto-splice* potenciais, localizados a cada dois códons em seis espécies modelo. A partir destes dados foi calculada qual seria a distribuição de fase de íntrons se todos os sítios de *proto-splice* encontrados tivessem sido usados para inserção de íntrons (Long et al. 1998). A segunda análise utilizou todos os sítios de *proto-splice* encontrados nas regiões codificadoras para identificar pseudo-éxons, isto é, éxons virtuais delimitados por dois sítios de *proto-splice* consecutivos (Long & Rosenberg 2000). Desta maneira, eles computaram a correlação entre as fases dos íntrons flanqueadores dos pseudo-éxons para calcular a porcentagem do excesso de pseudo-éxons simétricos e comparar com a distribuição observada do excesso de éxons simétricos. Ambas análises não observaram correspondência entre as distribuições virtuais (baseadas em sítios de *proto-splice*) de fase de íntrons e de excesso de éxons simétricos com estas mesmas distribuições encontradas na natureza (Long et al. 1998; Long & Rosenberg 2000). Esses resultados não corroboram os argumentos dos defensores da hipótese *introns-late*, que defendiam que as características dos íntrons atuais são resultado da inserção não aleatória de íntrons.

Durante alguns anos, estes trabalhos sobre sítios de *proto-splice* juntamente com as análises de correspondência entre íntrons, éxons, domínios e módulos, argumentaram em favor da teoria sintética da origem dos íntrons, citada anteriormente (de Souza et al. 1998; de Souza 2003; Roy 2003). Alguns estudos mais recentes, porém, reavivaram a discussão sobre a existência de sítios de

proto-splice (Coghlan & Wolfe 2004; Qiu et al. 2004; Sadusky et al. 2004; Sverdlov et al. 2004b). O primeiro destes trabalhos verificou que ao mutar seqüências C/AAG|G/A flanqueadoras de íntrons inseridos do gene *Actina* humano, sítios crípticos de *splicing* passavam a ser utilizados (Sadusky et al. 2004). Estes novos sítios correspondem exatamente a regiões flanqueadoras de íntrons de *Actina* existentes em outras espécies. Os autores sugerem que esta correspondência indica que estes sítios crípticos são alvos preferenciais de inserção (Sadusky et al. 2004).

Dois outros trabalhos mostraram que íntrons classificados filogeneticamente como íntrons inseridos apresentam distribuição desigual de fase (íntrons de fase zero são mais abundantes) e são flanqueados por seqüências consenso AG|G (Coghlan & Wolfe 2004; Qiu et al. 2004). No estudo específico de íntrons inseridos no gênero *Caenorhabditis*, a freqüência de bases AG|G flanqueando os íntrons é significativamente maior do que a encontrada nos íntrons conservados entre o gênero e outras espécies (Coghlan & Wolfe 2004).

Outro trabalho testou as hipóteses das seqüências consensos flanqueadoras dos éxons serem resquícios de sítios de *proto-splice* ou de terem evoluído convergentemente por seleção em favor da eficiência do *splicing* (Sverdlov et al. 2004b). Para tal, os autores examinaram as seqüências nucleotídicas flanqueadoras de íntrons localizados em regiões codificantes altamente conservadas em eucariotos. Teoricamente, estas seqüências evoluem sob um regime de forte seleção purificadora, não admitindo substituições não sinônimas. Espera-se, portanto, que estas regiões não estejam sujeitas a seleção em favor de uma maior eficiência de *splicing*. Mesmo assim, estas seqüências

apresentam seqüências consenso (A/C)AG|G flanqueando seus éxons (Sverdlov et al. 2004b).

Todos estes trabalhos descritos questionam dois pontos importantes em relação aos íntrons que devem ser analisados cuidadosamente. A primeira questão apontada é que estas observações corroboram a hipótese de existência de sítios preferenciais de inserção de íntron. No entanto, é importante ressaltar que nenhum dos estudos distingue entre as duas seguintes hipóteses: os íntrons foram inseridos em sítios de *proto-splice* ou os íntrons foram inseridos aleatoriamente e foram fixados somente nas regiões que contêm seqüências exônicas flanqueadoras “AG|G”. No caso do estudo entre as espécies do gênero *Caenorhabditis* (*C. elegans* e *C. briggsae*) (Coghlan & Wolfe 2004), o tempo evolutivo estimado de divergência entre elas (80 milhões de anos (Stein et al. 2003)) é suficiente para a fixação ou perda de qualquer *locus* polimórfico por deriva gênica (Hartl & Clark 1997). Desta maneira, íntrons sobre qualquer força seletiva mesmo que fraca para perda de íntrons em regiões que não fossem flanqueadas por “AG|G” teriam sido perdidos em muito menos do que 80 milhões de anos.

A segunda questão apontada pelos defensores da hipótese *introns-late* é que a observação de uma frequência maior de íntrons de fase zero no conjunto de íntrons inseridos questiona a natureza ancestral dos íntrons, já que esta é uma das características que suporta a hipótese *introns-early* (Qiu et al. 2004; Wolf et al. 2000). No entanto, a hipótese *introns-early* é sustentada por uma série de outras observações como excesso de éxons simétricos em regiões ancestrais conservadas, correspondências de íntrons de fase zero e regiões de ligação entre módulos protéicos em proteínas antigas e o excesso de éxons flanqueados

por íntrons de fase zero correspondentes a domínios antigos (de Souza et al. 1996; de Souza et al. 1998; Fedorov et al. 2001; Fedorov et al. 2003; Kaessmann et al. 2002; Long et al. 1995; Roy et al. 1999; Roy et al. 2002).

Parece que a questão de íntrons inseridos e sítios de *proto-splice* ainda não está totalmente resolvida e a solução para tal questão seria o estudo de íntrons inseridos entre espécies próximas já que muitos genomas de espécies do mesmo gênero estão sendo disponibilizados.

1.10. Resumo dos estudos realizados sobre a origem dos íntrons

Muitos estudos foram realizados para testar as hipóteses sobre a origem dos íntrons. Nesta seção estão organizadas as características (Tabela 1.1), as predições (Tabela 1.2) e as evidências (Tabela 1.3) características de cada uma das três hipóteses/teorias: *introns-early*, *introns-late* e teoria sintética da evolução dos íntrons.

Tabela 1.1: Características dos íntrons e do fenômeno de *exon-shuffling* segundo as três hipóteses/teorias estudadas.

	<i>Introns-early</i>	<i>Introns-late</i>	Teoria Sintética
Íntrons	São ancestrais, já existiam no progenoto	Inseridos somente em eucariotos	A maior parte foi inserida em eucariotos, mas alguns íntrons de fase zero já existiam no progenoto
Exon-shuffling	Existia no progenoto	Somente em eucariotos	Ocorreu no progenoto e em eucariotos

Tabela 1.2: Predições assumidas sobre a evolução dos íntrons e do fenômeno de *exon-shuffling* segundo as três hipóteses/teorias estudadas.

	<i>Introns-early</i>	<i>Introns-late</i>	Teoria Sintética
Distribuição das fases dos íntrons em RACs	Frequência desigual de fases	Frequência igual de fases (inserção aleatória)	Frequência desigual de fases
Distribuição de éxons em RACs	Excesso de éxons simétricos	Sem excesso de éxons simétricos	Excesso de éxons simétricos
Correspondência entre éxons as unidades protéicas independentes	Observada em proteínas antigas	Observada somente em proteínas modernas	Observada em proteínas antigas e modernas
Distribuição filogenética de posição de íntrons	Todas posições conservadas em várias espécies	Nenhuma posição conservada em todas as espécies	Algumas posições conservadas, mas a maioria não conservada em todas as espécies
Inserção de íntrons	Sem sítios de inserção	Sítios preferenciais de inserção (<i>proto-splice</i>)	Inserção aleatória de íntrons e seleção negativa de íntrons não flanqueados por seqüências específicas
Distribuição das fases dos íntrons inseridos	Sem íntrons inseridos	Frequência desigual de fases (inserção não aleatória)	Frequência igual de fases (inserção aleatória) / frequência desigual de fases (seleção negativa)

Tabela 1.3: Predições assumidas sobre a evolução dos íntrons e do fenômeno de *exon-shuffling* segundo as três hipóteses/teorias estudadas.

Em vermelho estão realçados quais pressupostos foram evidenciados.

	<i>Introns-early</i>	<i>Introns-late</i>	Teoria Sintética
Distribuição das fases dos íntrons em RACs	Frequência desigual de fases	Frequência igual de fases (inserção aleatória)	Frequência desigual de fases
Distribuição de éxons em RACs	Excesso de éxons simétricos	Sem excesso de éxons simétricos	Excesso de éxons simétricos
Correspondência entre éxons as unidades protéicas independentes	Observada em proteínas antigas	Observada em proteínas modernas	Observada em módulos e domínios 0-0 em proteínas antigas e domínios 1-1 em proteínas modernas
Distribuição filogenética de posição de íntrons	Todas posições conservadas em várias espécies	Nenhuma posição conservada em todas as espécies	Algumas posições conservadas, mas a maioria não conservada em todas as espécies
Inserção de íntrons	Sem sítios de inserção	Sítios preferenciais de inserção (<i>proto-splice</i>)	Inserção aleatória de íntrons e seleção negativa de íntrons não flanqueados por seqüências específicas
Distribuição das fases dos íntrons inseridos	Sem íntrons inseridos	Frequência desigual de fases (inserção não aleatória)	Frequência igual de fases (inserção aleatória) / frequência desigual de fases (seleção negativa)

1.11 *Evolução de íntrons e do fenômeno de exon-shuffling*

A teoria sintética de evolução dos íntrons concilia as duas hipóteses oponentes sobre a origem dos íntrons (de Souza et al. 1998; de Souza 2003; Roy 2003). Esta teoria combina características das duas hipótese e assume que a maioria dos íntrons foi inserida durante a evolução de eucariotos, mas que alguns íntrons de fase zero são ancestrais (de Souza et al. 1998; de Souza 2003; Roy 2003). Algumas questões ainda estão em aberto, como a hipótese de inserção não aleatória dos íntrons em sítios de *proto-splice*. Os defensores da hipótese *introns-late* ainda discutem sobre a existência dos íntrons no ancestral comum a procariotos e eucariotos.

No entanto, parece improvável que estes estudos alterem profundamente a hipótese sintética atual. O conflito remanescente entre as correntes que enfatizam a hipótese *introns-early* (de Souza et al. 1998; Roy et al. 2002; Roy 2003) e a que enfatiza os elementos da teoria *introns-late* (Coghlan & Wolfe 2004; Qiu et al. 2004; Sadusky et al. 2004; Stoltzfus 2004) se limita a buscar indícios definitivos sobre a presença ou ausência de íntrons no progenoto. Essa discussão tem poucas chances de produzir elementos novos, dada a dificuldade de encontrar evidências que não possam ser suportadas por versões mais ou menos sofisticadas de ambas as hipóteses e a impossibilidade de observação direta do genoma do progenoto.

Os estudos envolvendo a evolução dos íntrons e do fenômeno de *exon-shuffling* não se limitam apenas ao debate entre as hipóteses *introns-early* e *introns-late*. Durante muitos anos, esses estudos levaram à descoberta de diversas características e padrões dos genes, éxons, íntrons, domínios, módulos

e proteínas. Tais estudos foram fundamentais para a compreensão da evolução do fenômeno de *exon-shuffling* e da importância do fenômeno na formação de novos genes. Todavia, ainda existem problemas biológicos interessantes relacionados à evolução dos íntrons e ao fenômeno de *exon-shuffling*, alguns exemplos estão descritos a seguir.

1.11.1. *Distribuição de eventos de shuffling ao longo das proteínas*

Uma das contribuições importantes ao estudo de *exon-shuffling* foi a descoberta de que domínios antigos são enriquecidos com éxons 0-0 e domínios modernos são enriquecidos com éxons 1-1 (Kaessmann et al. 2002; Patthy 1999; Patthy 2003; Tordai et al. 2005). Estes estudos sugerem que dois tipos diferentes de *shuffling* de éxons correspondentes a domínios ocorreram em períodos diferentes da evolução. No entanto, não se tem nenhuma informação sobre a distribuição destes dois tipos de *shuffling* ao longo das proteínas. Isto é, se cada um destes tipos de *shuffling* ocorreram preferencialmente em locais diferentes das proteínas. A partir deste tipo de informação poderemos entender melhor como do fenômeno do *exon-shuffling* contribuiu para formação das proteínas atuais, além de complementar, confirmar ou corrigir a teoria sintética da evolução dos íntrons com estas informações.

1.11.2. *Eventos de shuffling em proteínas que contêm peptídeos sinais*

Recentemente, Tordai e Patthy (2004) compararam a distribuição de íntrons em genes humanos que codificam proteínas com e sem peptídeo sinal.

Peptídeo sinal é um tipo de seqüência encontrada em muitas proteínas extracelulares e transmembranares (formadas durante a evolução de metazoários), além de proteínas de bactérias (von Heijne 1990). Esta seqüência atua como sinal intrínseco da proteína que governa o seu transporte e localização dentro da célula (von Heijne 1990). O tipo mais comum de sinal são os peptídeos de 13 a 36 resíduos localizados na porção N-terminal de algumas proteínas. Em eucariotos, todas as proteínas secretadas pelo retículo endoplasmático e muitas proteínas transmembranares são sintetizadas com um peptídeo sinal que direciona a passagem pela membrana do retículo (Voet & Voet 1995a). Já em bactéria, proteínas que atravessam a membrana celular também possuem peptídeo sinal. Estes tipos de peptídeos são reconhecidos após a síntese protéica por uma partícula que auxilia na passagem para fora das membranas e, após a passagem, o peptídeo sinal é clivado pela enzima *sinal peptidase* tanto em bactérias quanto em eucariotos (Voet & Voet 1995a).

A estrutura comum de peptídeos sinais de bactérias e de eucariotos presentes em várias proteínas é normalmente descrita como uma região N-terminal carregada positivamente, seguida de uma região hidrofóbica e uma região C-terminal neutra, mas polar (von Heijne 1983; von Heijne 1985). A regra (-3,-1) especifica que os resíduos nas posições -3 e -1 (relativas ao sítio de clivagem) devem ser pequenos e neutros para que a clivagem ocorra corretamente (von Heijne 1983; von Heijne 1985).

Tordai e Patthy (2004) analisaram a distribuição de íntrons nos primeiros 100 aminoácidos das proteínas e observaram excesso de íntrons de fase um na vizinhança dos sítios de clivagem dos peptídeos sinais (Figura 1.15). Proteínas

que não contêm peptídeo sinal apresentam distribuição homogênea de íntrons de fase um (Figura 1.15) (Tordai & Patthy 2004).

Figura 1.15: (A) Distribuição de tamanho de peptídeo sinal. Distribuição de íntrons de fase zero, um e dois nos primeiros 100 resíduos de aminoácido de proteínas sem (B) e com (C) peptídeo sinal. Figura e legenda reproduzidas da Figura 1 de Tordai e Patthy (2004). Note que o pico de íntrons de fase um das proteínas contendo peptídeo sinal (C) coincidem com a borda C-terminal dos peptídeos(A).

Os autores sugeriram que alguma característica dos peptídeos sinal deve favorecer a inserção de íntrons de fase um na vizinhança dos sítios de clivagem

(Tordai & Patthy 2004). Sugeriram ainda que as seqüências de aminoácido que flanqueiam os sítios de clivagem são enriquecidas com sítios de *proto-splice*, o que levaria a uma freqüência elevada de eventos de inserção de íntrons na vizinhança do peptídeo sinal.

O argumento deles é baseado na existência de sítios de *proto-splice* restritos aos nucleotídeos AG|G e no fato de que tais seqüências são traduzidas em aminoácidos mais freqüentemente observados nas regiões de junção éxon-íntron (Tomita et al. 1996). Por exemplo, íntrons de fase dois normalmente estão localizados dentro de um códon de arginina (AGG), enquanto que íntrons de fase um estão localizados dentro de um dos quatro códons (GGN) que codificam glicina (Tomita et al. 1996).

Os autores se referiram à análise de matriz de freqüência de aminoácidos (Tabela 1.4) observada em sítios de clivagem de peptídeos sinais de eucariotos (Nielsen et al. 1997). Esta matriz indica que as posições -1, -3, -4 e -5 são significativamente enriquecidas com resíduos de glicina (Tabela 1.4). Dessa maneira, o pico de íntrons de fase um observados na vizinhança dos peptídeos sinais seria uma consequência da grande quantidade de sítios de *proto-splice* decorrentes da grande quantidade de resíduos de glicinas existentes nestas regiões (Tordai & Patthy 2004).

Tabela 1.4: Matriz de frequência de aminoácidos nas posições -13 a +2 relativa aos sítios de clivagem de peptídeos sinal encontrados em eucariotos. Frequência de cada aminoácido encontrado em cada uma das posições e frequência esperada (esp) para todas as posições calculada a partir da frequência média observada de cada aminoácido (Nielsen et al. 1997). Em vermelho, estão apresentados os valores observados de glicina nas posições -1, -3, -4 e -5 que são muito maiores que o valor esperado.

	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	esp.
A	101	112	106	100	158	128	107	149	146	107	258	80	458	141	55	73.86
C	36	32	34	31	42	66	57	39	30	34	68	21	50	23	27	27.14
D	2	0	3	0	3	2	4	5	25	14	4	44	6	68	66	28.42
E	1	2	4	5	4	6	13	9	28	27	7	66	6	92	88	36.57
F	75	67	80	81	65	58	92	67	25	34	7	46	6	34	28	33.80
G	39	24	26	34	36	50	40	29	108	125	74	38	184	52	57	53.42
H	5	3	4	2	4	10	14	7	23	12	5	53	2	23	22	14.76
I	73	67	82	59	52	52	47	69	22	43	41	18	4	37	34	33.45
K	4	1	1	2	0	3	3	3	20	19	7	24	4	52	40	31.95
L	376	432	397	449	386	329	333	280	98	147	65	165	21	60	51	116.43
M	26	15	21	19	11	30	20	11	5	12	9	14	3	14	12	27.84
N	2	4	4	6	7	7	8	7	25	14	9	37	9	26	53	26.56
P	23	13	17	9	12	13	24	60	98	74	7	17	27	11	166	44.53
Q	3	6	4	9	11	19	18	13	55	40	5	72	15	135	42	30.72
R	7	4	4	2	3	7	9	6	34	38	5	56	20	33	40	34.01
S	57	47	40	44	65	76	67	66	117	97	142	112	141	91	88	65.26
T	32	39	30	38	38	49	42	35	70	70	103	44	43	39	57	45.02
V	112	124	127	100	99	86	79	129	60	72	191	45	8	50	56	56.25
W	16	13	20	11	9	13	24	21	13	13	2	33	2	9	6	12.12
Y	17	3	6	9	6	7	10	6	8	19	2	26	2	21	23	18.49

No entanto, analisando cuidadosamente as proteínas humanas que contêm peptídeo sinal notamos que muitas delas são proteínas extracelulares ou transmembranares. Sabe-se que a formação de novos genes através de *shuffling* de éxons simétricos pode gerar aumento da frequência de íntrons de uma determinada fase (ver explicação na seção 1.5) (Patthy 1987). Já foi descrito também que muitas das proteínas criadas durante a evolução de metazoários são extracelulares e transmembranares (Patthy 1996; Patthy 1999). Sabemos

também que estas proteínas foram formadas através de *shuffling* de domínios correspondentes a éxons 1-1 (Patthy 1999; Patthy 2003; Tordai et al. 2005). Os autores do estudo mencionado acima sugeriram que o excesso de íntrons de fase um encontrado na vizinhança da região C-terminal do peptídeo sinal pode ser produto de inserção preferencial de íntrons, mas não mencionariam que eventos de *exon-shuffling* dos peptídeos sinal propriamente ditos possam gerar tal excesso de íntrons de fase um nestas proteínas (Tordai & Patthy 2004).

Considerando todos os dados sobre *exon-shuffling* de proteínas criadas durante a evolução de eucariotos somado ao fato de *exon-shuffling* de seqüências alvo já terem sido observados anteriormente não podemos eliminar a hipótese de que *exon-shuffling* possa ter sido importante na geração deste excesso de íntrons de fase um na vizinhança de sítios de clivagem de peptídeo sinal (Long et al. 1996).

Patthy (1999), em um trabalho anterior, mostra que a evolução de proteases plasmáticas fornecem um dos exemplos de evidência inequívoca de *exon-shuffling*. Estas proteases de coagulação sanguínea contêm domínios homólogos a algumas fibronectinas, domínios que, por sua vez, são codificados por éxons 1-1. Nestes mesmos trabalhos, Patthy (1996, 1999) sugere uma explicação plausível para tal padrão: “o íntron de fase um presente entre o peptídeo sinal e o domínio seguinte da protease ancestral deveria ser um bom receptor para éxons 1-1”. Ele ainda sugere que tal evento tenha causado a proliferação de íntrons de fase um nestas proteínas. Como no caso destas proteases, muitas outras proteínas que contêm íntrons de fase um na borda do peptídeo sinal podem ter sofrido eventos de *exon-shuffling* envolvendo este mesmo íntron. Dessa maneira, a abundância dos íntrons de fase um na borda

dos sítios de clivagem podem também ser resultado de *shuffling* de éxons 1-1 e não apenas produto da inserção de íntrons em sítios de *proto-splice*.

1.12. Objetivo geral

O trabalho desenvolvido nesta tese envolve análises da distribuição de éxons, íntrons e domínios com o objetivo de compreender melhor como as proteínas atuais foram formadas e discutir como tais características estão relacionadas com a origem e a evolução dos íntrons e do fenômeno de *exon-shuffling*.

1.12.1 Objetivos específicos

Nossas análises focalizaram duas questões não abordadas em estudos anteriores:

- Avaliação da distribuição de *exon-shuffling* "antigo" e "moderno" ao longo das proteínas através da análise da distribuição de excesso de éxons simétricos 0-0 e 1-1 e da análise da distribuição de domínios antigos e modernos.
- O papel do fenômeno de *exon-shuffling* em proteínas contendo peptídeos sinal, relacionando o fenômeno à alta frequência de íntrons de fase um na borda do sítio de clivagem destes peptídeos.

1.13. Apresentação da tese

As análises realizadas produziram dois trabalhos completos e portanto estão apresentados em capítulos separados em forma de artigo e manuscrito (Capítulos 3 e 4, respectivamente).

Dessa forma, nos abstivemos de detalhar os Materiais e Métodos e a Discussão de cada trabalho em seus respectivos capítulos da tese. O primeiro capítulo de Resultados da tese (Capítulo 3) se encontra reduzido a análises complementares da distribuição de excesso de exons simétricos e domínios modernos e antigos que não foram publicadas. O Capítulo 5 apresenta uma discussão complementar às discussões já apresentadas no artigo e no manuscrito apresentados nos Capítulos 3 e 4.

Capítulo 2: Materiais e Métodos

Normalmente, os trabalhos desenvolvidos através da biologia computacional e da bioinformática utilizam como materiais, seqüências de DNA e de aminoácidos depositados em bancos de dados públicos; e como métodos, programas computacionais disponíveis ou desenvolvidos durante o projeto para manipulação destes dados. Por isso, dividimos este capítulo em duas seções: materiais, isto é, seqüências e bancos de dados utilizados; e métodos, que descreve os programas e algoritmos utilizados e desenvolvidos durante a execução do trabalho.

2.1. Materiais

2.1.1. ExInt: banco de dados de éxons e íntrons

Usamos o banco de dados ExInt versão 133 (Sakharkar et al. 2002) para analisar as posições e fases dos íntrons. Este banco de dados contém dados sobre a estrutura gênica de 134.422 seqüências que contém íntrons provenientes da versão 133 do Genbank. As seqüências presentes no banco são seqüências de aminoácidos correspondentes às proteínas produzida por cada gene contido no Genbank. Produtos gênicos diferentes provenientes do mesmo gene são considerados seqüências diferentes dentro do banco. Cada seqüência contida no banco está apresentada em formato Fasta que apresenta em seu cabeçalho as informações de identidade do gene, número, posição e fase dos seus íntrons em relação a sua seqüência de aminoácidos.

2.1.1.1. ExInt não redundante

O ExInt é construído a partir do Genbank que é um repositório de seqüências redundantes, contendo múltiplas cópias, completas e parciais, de várias seqüências (Sakharkar et al. 2002). Eliminamos a redundância do banco através da retirada das seqüências que apresentavam identidade relativa maior ou igual a 99% em alinhamentos par a par com outras seqüências do banco (ver descrição dos métodos de agrupamento e definição de identidade relativa na seção 2.2.2.). Basicamente, a partir dos alinhamentos par a par, construímos agrupamentos de seqüências e selecionamos apenas uma seqüência de cada agrupamento (aquela com maior número de exons) para construir o conjunto não redundante do ExInt que denominamos de ExInt 99%.

Além de seqüências parciais e duplicadas, o ExInt contém muitos genes ortólogos e parálogos, isto é, múltiplas cópias de certos genes com uma origem em comum. Em algumas de nossas análises se tornou necessário o uso de subconjuntos do banco contendo apenas uma seqüência representante de cada conjunto de genes homólogos. Para isto, agrupamos os genes em famílias e selecionamos uma seqüência (aquela com maior número de éxons) para construir o banco de dados não redundante. Realizamos diferentes agrupamentos de famílias gênicas baseadas no grau de identidade apresentada entre as seqüências. Por exemplo, construímos um banco de dados denominado ExInt 60%, constituído de uma seqüência de cada família. Os genes agrupados em cada família apresentavam identidade relativa mínima de 60% em alinhamentos par a par com outro gene da mesma família. Construímos, então, diferentes subconjuntos não redundantes denominados ExInt 20, 40, 60, 80 e

90% (ver descrição dos métodos de agrupamento e definição de identidade relativa na seção 2.2.2.).

2.1.1.2. ExInt experimental

O Genbank é um repositório de seqüências de DNA e de aminoácidos que contém muitas seqüências provenientes de projetos genomas e outros projetos de sequenciamento em larga escala. Por isso, muitos dos genes descritos no Genbank não foram caracterizados por meio de experimentos clássicos de genética molecular, e sim, preditos através de programas de computação. Desta maneira, muitos destes genes podem conter erros de anotação ou ser simples artefatos. No intuito de construir um subconjunto do ExInt que apresentasse o menor número possível de genes com erros de anotação, eliminamos do banco aqueles que apresentassem, em sua linha de definição, palavras características de genes preditos ou provenientes de projetos de sequenciamento de genomas como: *BAC*, *PAC*, *Cosmid* e *Chromosome*. Selecionamos um conjunto de 48.863 genes que não apresentam tais palavras, o qual denominamos de ExInt experimental.

Sabemos que este procedimento de seleção pode não eliminar todos os genes preditos porque é baseado apenas em uma busca de palavras. No entanto, notamos que reduzimos drasticamente o nosso número de seqüências no subconjunto experimental do ExInt em relação ao ExInt total. Apesar de não podermos saber se eliminamos todos os genes preditos na construção do ExInt experimental, podemos avaliar se o número de genes deste subconjunto é comparável ao número de seqüências existentes em bancos que utilizam

procedimentos de alta qualidade de anotação como o SWISS-PROT (Boeckmann et al. 2003). Comparamos o número de seqüências humanas do ExInt experimental com o número de seqüências humanas existentes no SWISS-PROT. No ExInt experimental existem 7.824 seqüências humanas enquanto que na versão 42 do SWISS-PROT continham aproximadamente 10.000 seqüências humanas (ver estatísticas em <http://www.expasy.org/sprot/relnotes/relstat.html>). Esta versão do SWISS-PROT foi produzida em 2003 e corresponde a data mais próxima da produção da versão 133 do Genbank. Apesar de não termos comparado o número de seqüências de todos os organismos contidos no SWISS-PROT e no ExInt experimental, acreditamos que este procedimento pode ser utilizado para se ter uma estimativa da eliminação de seqüências mal anotadas através de nossa busca de palavras. Isto porque muitos genes humanos foram preditos por programas computacionais a partir de seqüências genômicas. Esta comparação mostra que o número de seqüências presentes no nosso subconjunto experimental humano é menor do que o número de seqüências existentes em banco de dados de proteínas humanas com alta qualidade de anotação e que nosso procedimento de construção do ExInt experimental provavelmente deve ter eliminado a maioria de seqüências anotadas com baixa qualidade.

2.1.2. SWISS-PROT

Utilizamos em nosso trabalho o conjunto de proteínas de bactérias e humanas. As proteínas de bactérias descritas no SWISS-PROT foram utilizadas como banco de dados de seqüências de procariotos, para buscas de similaridade

das proteínas do ExInt e das proteínas humanas do próprio SWISS-PROT, nos dois trabalhos apresentados nesta tese. Como a anotação do SWISS-PROT fornece a posição exata do peptídeo sinal em cada proteína humana, proteínas humanas descritas no SWISS-PROT foram utilizadas no trabalho sobre o papel do *exon-shuffling* em seqüências com peptídeo sinal (Capítulo 4). Utilizamos as versões 41 e 47 nos trabalhos apresentados nos Capítulos 3 e 4, respectivamente (Boeckmann et al. 2003).

2.1.3. *PFam*

O PFam é um banco de dados de domínios protéicos, cada um dos quais é caracterizado por um alinhamento semente, que contém membros representativos de cada domínio (Bateman et al. 2004). Para permitir o alinhamento e a busca por seqüências homólogas em outras proteínas, um HMM (“Hidden Markov Model”) é construído a partir de cada alinhamento semente, através de um processo onde cada coluna no alinhamento múltiplo é transformado em um sistema de pontuação específico para cada posição (Eddy 1998). Os domínios presentes no PFam são divididos em duas seções, PFam-A e PFam-B (Bateman et al. 1999). O PFam-A contém famílias caracterizadas por alinhamentos sementes derivados de várias fontes, todos curados e anotados manualmente (Bateman et al. 1999). Além do alinhamento semente, o PFam-A inclui alinhamentos múltiplos de todas as regiões homólogas detectadas no banco de dados UniProt usando os HMMs (Bateman et al. 1999; Bateman et al. 2000). O Pfam-B, por sua vez, é construído através do agrupamento de todas as regiões de proteínas do UniProt que não estão presentes no Pfam-A (Bateman et

al. 1999). Em nossas análises utilizamos apenas o Pfam-A (versão 12) para busca de domínios nas proteínas estudadas.

Os domínios são classificados segundo a sua distribuição taxonômica no Pfam (Bateman et al. 2004). Em nosso trabalho, dividimos os domínios em antigos e modernos. Antigos são aqueles presentes em proteínas de procariotos e eucariotos. Enquanto que os modernos são aqueles presentes exclusivamente em proteínas de eucariotos.

Em seu trabalho sobre domínios correspondentes a éxons, Kaessman e colaboradores (2002) utilizaram classificação similar para subdividir domínios em modernos e antigos em seu trabalho sobre domínios correspondentes a éxons. Porém, diferentemente de nossa classificação, eles determinaram que domínios modernos são exclusivos de metazoários e domínios antigos são aqueles compartilhados por metazoários e procariotos.

2.2. Métodos

2.2.1. Programas de alinhamentos de seqüências

Ao longo de todo o trabalho, usamos dois programas de alinhamento local: BLAST e FASTA (Altschul et al. 1997; Pearson & Lipman 1988). Estes programas foram usados na construção dos agrupamentos de seqüências similares, na busca de seqüências similares a seqüências procarióticas e para outros fins. As duas principais diferenças entre estes dois programas são a eficiência e o tipo de alinhamento final produzido. O BLAST é um programa mais rápido e fornece todos os melhores alinhamentos locais que identifica, entre duas seqüências. O

FASTA não é tão rápido quanto o BLAST, mas é mais sensível na detecção de homólogos remotos e fornece somente um único alinhamento para cada par de seqüências comparadas. Apesar de ser um programa de alinhamento local, o programa FASTA inclui um passo adicional no cálculo de similaridade par a par inicial que permite que várias regiões separadas de alinhamento sejam unidas para a construção de um alinhamento único (Pearson & Lipman 1988).

O tipo de resultado fornecido pelo FASTA é vantajoso para o cálculo de identidade total entre duas seqüências, isto é, ao longo de toda a extensão das seqüências. O cálculo de identidade relativa (que utiliza a identidade total como base - veja seção 2.2.2.), usado para agrupar seqüências homólogas seria mais complicado se calculado através do somatório das identidades de cada alinhamento produzido pelo BLAST, pois este pode gerar vários alinhamentos diferentes com sobreposições.

2.2.2. Método de construção de agrupamentos de seqüências do ExInt por níveis de identidade em alinhamento produzidos pelo programa FASTA

No trabalho apresentado no Capítulo 3, utilizamos um método de agrupamento de seqüências similar aos usados em trabalhos anteriores baseados em critérios de identidade (Long et al. 1995). Primeiro, ordenamos por comprimento decrescente todas as seqüências de aminoácidos do ExInt. Alinhamos par a par todas as seqüências ordenadas, usando o programa FASTA (Pearson & Lipman 1988). Calculamos a identidade relativa entre cada par de seqüências multiplicando a identidade do alinhamento propriamente dito (percentual de aminoácidos iguais) pelo tamanho do alinhamento e dividindo este

produto pelo tamanho da menor seqüência do par alinhado (Long et al. 1995). O agrupamento propriamente dito é feito seqüencialmente, avaliando primeiro as seqüências mais longas. Se o alinhamento de uma seqüência com outra de comprimento maior apresentar identidade relativa maior ou igual a identidade mínima estipulada, a seqüência passa a pertencer ao mesmo agrupamento que a seqüência maior. Caso contrário, esta seqüência será a primeira de um novo agrupamento. Nenhuma seqüência pode pertencer a dois agrupamentos diferentes.

Este método não permite que sejam reunidas no mesmo conjunto seqüências que não tenham identidade nenhuma com a maior proteína do grupo (Figura 2.1A). Métodos em que as seqüências precisam apresentar identidade mínima com apenas uma seqüência do grupo podem gerar conjuntos contendo muitas proteínas sem similaridade alguma (Figura 2.1B). Principalmente quando valores de identidades mínima muito baixos são utilizados, como 20% ou 40%.

Figura 2.1: Métodos de agrupamento de seqüências protéicas. (A) Agrupamento por seleção de seqüências que apresentam uma identidade mínima em alinhamentos par a par com a maior proteína do grupo (em vermelho). (B) Agrupamento por seleção de seqüências que apresentam uma identidade mínima com qualquer outra seqüência do grupo.

2.2.3. Método de construção de agrupamentos de seqüências do SWISS-PROT por identidade e cobertura em alinhamentos produzidos pelo programa BLAST

Para seleção de famílias gênicas que codificam proteínas humanas do SWISS-PROT contida no trabalho apresentado no Capítulo 4, agrupamos as seqüências que apresentavam uma cobertura maior que 90% nos alinhamentos par a par entre elas (programa BLAST (Altschul et al. 1997)). A cobertura foi calculada através da divisão do tamanho da região alinhada pelo tamanho da maior seqüência de cada par. Considerando cada par de proteínas, uma família foi definida como um conjunto de proteínas onde cada uma delas está relacionada com pelo menos uma outra do mesmo grupo (Figura 2.1B). Nenhuma proteína pode ser membro de mais do que uma família. Diferentemente do método de agrupamento realizado com as proteínas do ExInt, este método pode agrupar muitas seqüências sem nenhuma identidade na mesma família gênica. No entanto, como o critério de cobertura é alto (90% da maior seqüência do par alinhado), os agrupamentos, na maioria dos casos, correspondem a seqüências parálogas de uma mesma espécie (Kopelman et al. 2005).

2.2.4. Simulações

As análises descritas no artigo contido no Capítulo 3 incluíram duas simulações diferentes. Nas seções 2.4.1 e 2.4.2, explicamos mais detalhadamente os métodos aplicados nessas simulações.

2.2.4.1. Simulação da distribuição de éxons simétricos 0-0 e 1-1 ao longo das proteínas contendo domínios antigos e modernos

Para verificar se a diferença na distribuição do excesso de éxons simétricos 0-0 e 1-1 ao longo das três regiões das proteínas (0-33; 34-66; 67-100%) é significativa, realizamos uma simulação. Cinco mil iterações foram realizadas, onde somente a distribuição dos éxons ao longo das proteínas podia variar aleatoriamente. Todos os outros parâmetros como número e posição dos íntrons ao longo das proteínas, a região de localização dos domínios antigos e modernos, a frequência total da fase dos íntrons (Tabela 2.1) e a frequência total dos diferentes tipos de éxons (simétricos ou não simétricos) (Tabela 2.2) foram as mesmas que as observadas no dado real.

Tabela 2.1: Frequência total de íntrons observada nas regiões contendo domínios nas proteínas do ExInt que apresentam domínios antigos e modernos.

Fase do íntron	0	1	2
Frequência	48 %	30%	22%

Tabela 2.2: Frequência total de éxons observada nas regiões contendo domínios nas proteínas do ExInt que apresentam domínios antigos e modernos.

		Fase do íntron 3'		
		0	1	2
Fase do íntron 5'	0	28%	10%	10%
	1	11%	13%	6%
	2	11%	6%	5%

Contabilizamos o número de iterações das simulações que apresentavam distribuição de excesso de éxons simétricos similares ao observado no dado real. As distribuições foram consideradas similares quando apresentavam excesso de simétricos 1-1 maior nas extremidades do que no centro e excesso de éxons 0-0 maior no centro do que nas extremidades. A diferença de porcentagem de excesso entre as extremidades e o centro deveria ser 3,5% ou maior para os éxons 1-1 e 6,0% ou maior para os éxons 0-0. Esses valores foram determinados porque correspondem à metade da menor diferença observada entre as extremidades e o centro observadas no dado real. Tal procedimento possibilita que casos de simulação que apresentem diferenças pequenas entre as regiões, como 0,5%, não sejam contados como casos similares ao dado real.

2.2.4.2. Simulação da frequência de domínios antigos e modernos correspondentes a éxons 0-0 e 1-1.

No artigo apresentado no Capítulo 3, nós realizamos uma análise para verificar se a frequência de unidades de *shuffling* 0-0 e 1-1 (éxons ou conjunto de éxons correspondentes a domínios) em domínios antigos e modernos correspondia à frequência esperada pela associação aleatória da idade dos domínios e da simetria das unidades de *shuffling*. Os números observados e esperados de éxons simétricos (0-0 e 1-1) correspondentes a domínios antigos e modernos foram comparados entre si. Os valores esperados foram calculados de acordo com as frequências de domínios 0-0 e 1-1 observadas considerando o conjunto de domínios total, antigos e modernos somados. Simulações de 10.000 iterações foram realizadas para verificar se os números observados eram

significativamente diferentes dos esperados. Nestas simulações, os números totais de domínios antigos e modernos eram mantidos, assim como o número dos diferentes tipos de éxons (simétricos e não simétricos) também não variava. O único parâmetro que variava aleatoriamente na simulação era a associação de um domínio, moderno ou antigo, a um éxon flanqueado por determinados íntrons. Dessa forma, em cada iteração os domínios 0-0 e 1-1 tinham suas frequências calculadas independentes da idade dos domínios. Casos em que os domínios antigos 0-0 e os domínios modernos 1-1 fossem maiores ou iguais aos observados eram contabilizados para medir a significância dos resultados obtidos. Assim como os casos em que o número de domínios modernos 0-0 e o número de domínios antigos 1-1 fossem menor ou igual aos observados.

2.2.5. *Determinação da representação de termos de Ontologia Gênica (Gene Ontology) em genes humanos que codificam proteínas com e sem peptídeo sinal*

Proteínas receptoras de matriz extracelular, de adesão celular, imunoglobulinas e receptores com atividade de tirosina quinase são tidas como formadas por eventos de *exon-shuffling* (Patthy 1999; Patthy 2003). No manuscrito apresentado no Capítulo 4 (Material suplementar - Seção 2) analisamos os termos de ontologia gênica de seqüências humanas para determinar se os termos relacionados a eventos de *shuffling* estão super-representados nos genes que codificam proteínas com peptídeo sinal em relação àquelas que não apresentam tais seqüências. O programa GStat (Beissbarth & Speed 2004) é uma ferramenta de busca de termos de ontologia gênica estatisticamente significativos em uma lista de genes. Comparamos os dois grupos de genes (que codificam proteínas com e sem

peptídeo sinal) no programa GOstat para procurar os termos estatisticamente (limite máximo de $P = 0,1$) super-representados nos genes que codificam proteínas com peptídeo sinal.

**CAPÍTULO 3: SINAIS DE *EXON-SHUFFLING* ANTIGO E MODERNO ESTÃO
CORRELACIONADOS COM A
DISTRIBUIÇÃO DE DOMÍNIOS ANTIGOS E
MODERNOS AO LONGO DAS PROTEÍNAS**

Estudos anteriores mostraram que dois tipos de *shuffling* de éxons correspondentes a domínios teriam ocorrido em períodos diferentes da evolução (de Souza et al. 1996; de Souza et al. 1998; Kaessmann et al. 2002; Long et al. 1995; Patthy 1999; Patthy 2003). O enriquecimento de domínios antigos e modernos com éxons 0-0 e 1-1, respectivamente, sugerem que o tipo de *shuffling* predominante em períodos mais antigos teria sido o de éxons 0-0, enquanto que o tipo mais freqüente em períodos mais recentes teria envolvido éxons 1-1 (Kaessmann et al. 2002).

No artigo apresentado neste capítulo estudamos a distribuição dos eventos dos diferentes tipos de *exon-shuffling* (moderno e antigo, 0-0 e 1-1) ao longo das proteínas. Para isto, analisamos a distribuição do excesso de éxons simétricos 0-0 e 1-1 e a distribuição dos domínios ao longo das mesmas.

A principal conclusão do nosso trabalho (apresentado na seção 3.1) é a correlação entre a distribuição de domínios antigos e modernos e os sinais de *exon-shuffling* ao longo das proteínas. Mostramos que, em proteínas contendo domínios antigos e modernos concomitantemente, os domínios antigos são encontrados em maior abundância nas regiões centrais das seqüências, enquanto que domínios modernos são encontrados mais freqüentemente nas extremidades das proteínas. Observamos também que os excessos de éxons simétricos 0-0 e 1-1 que são os indícios de ocorrência de *exon-shuffling* antigo e moderno, respectivamente, apresentam distribuições similares. Observamos que o excesso de éxons 1-1 é significativamente maior nas extremidades das proteínas, enquanto que o excesso de éxons 0-0 é maior na região central.

Este capítulo está subdividido em duas partes: o artigo que já publicado no *Journal of Molecular Evolution* (seção 3.1) e algumas análises complementares

sobre o estudo da distribuição de domínios antigos e modernos e da distribuição de excesso de éxons simétricos ao longo das proteínas (seção 3.2).

3.1. Artigo intitulado: *Signs of ancient and modern exon-shuffling are correlated to the distribution of ancient and modern domains along protein*