UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA
Programa de Pós-Graduação em Ciências Biológicas (Bioquímica)

GUSTAVO STARVAGGI FRANÇA

História evolutiva de *exon shuffling* em eucariotos

São Paulo

26/11/2009

GUSTAVO STARVAGGI FRANÇA

História evolutiva de *exon shuffling* em eucariotos

Dissertação apresentada ao Instituto de Química da Universidade de São Paulo para obtenção do Título de Mestre em Ciências (Bioquímica)

Orientador: Prof. Dr. Sandro José de Souza

São Paulo

2009

Gustavo Starvaggi França

História evolutiva de *exon shuffling* em eucariotos

Dissertação apresentada ao Instituto de Química da Universidade de São Paulo para obtenção do Título de Mestre em Ciências (Bioquímica)

Aprovado em: _____

Banca Examinadora

Prof. Dr. _____

Instituição: _____

Assinatura: _____

Prof. Dr. _____

Instituição: _____

Assinatura: _____

Prof. Dr. _____

Instituição: _____

Assinatura: _____

AGRADECIMENTOS

Gostaria de agradecer especialmente aos meus pais Luiz e Anna Gina e ao meu irmão Guilherme, por todo o suporte e amor dedicados a mim. Agradeço ao meu Tio Roberto, que me fez perceber aspectos sutis que me nortearam para um propósito de vida maior. Pelos mesmos motivos agradeço à Márcia Regina, por sua especial sensibilidade e demonstração de carinho. Também sou grato à minha avó Maria Alice, à minha Tia Suzana, Irene, aos meus Tios Dorival e Loredana e aos meus primos Fábio, Fabiana e Felipe. Muito obrigado pelos momentos alegres que temos passado juntos!

Meus sinceros agradecimentos ao meu orientador, Sandro, por ter me aceitado como aluno e ter sugerido um tema de estudo instigante e de particular interesse. Agradeço pela confiança depositada em mim, pela liberdade que tive para desenvolver o trabalho, por suas idéias e sugestões sempre criativas, e pelas oportunidades oferecidas. Obrigado pelo incentivo e palavras de apoio. Foi um período de grande crescimento pessoal e profissional.

Sou imensamente grato ao amigo e colaborador Douglas Cancherini, pela constante ajuda técnica, pelas discussões proveitosas e comentários inteligentes. Teria sido muito mais difícil sem a sua ajuda. Obrigado!

Devo agradecer pela ajuda e agradável convivência a todas as pessoas que fizeram ou fazem parte do Laboratório de Biologia Computacional. Ao Pedro, Júlia, Robson, Suzana, Rodrigo, Patrícia, Jorge, Daniel, Ana Cláudia, Renata, André, Eduardo, Elza, Micha e Carmen. Todos contribuíram de alguma forma para a realização desse trabalho.

Agradeço ao Instituto Ludwig de Pesquisa sobre o Câncer e a todos os seus funcionários, por ceder seu espaço e proporcionar ótimas condições de trabalho.

Ao Hospital Alemão Oswaldo Cruz.

Ao Instituto de Química da USP, professores e funcionários.

À FAPESP pelo suporte financeiro.

RESUMO

França, G.S. **História evolutiva de *exon shuffling* em eucariotos.** 2009. 82p. Dissertação de Mestrado. Programa de Pós-Graduação em Bioquímica. Instituto de Química, Universidade de São Paulo, São Paulo.

Exon shuffling foi primeiramente proposto por Walter Gilbert em 1978 como um mecanismo em que exons de diferentes genes podem ser combinados, levando à formação de novos genes. O mecanismo de *exon shuffling* é favorecido por recombinações intrônicas e está correlacionado com a simetria de exons. Evidências deste mecanismo provém de análises de combinações de fases de introns, correlações entre bordas de exons e de domínios protéicos e da recorrência de domínios em diversas proteínas. Dessa forma, a evolução de proteínas formadas por *exon shuffling* pode ser inferida considerando a organização exon-intron dos genes, o padrão de combinações de fases de introns e a organização de domínios nas proteínas. Neste sentido, regiões protéicas que possivelmente foram originadas por eventos de *exon shuffling* foram identificadas através de análises em larga escala em diferentes espécies eucarióticas. A estratégia foi baseada no alinhamento entre todas as proteínas anotadas de uma determinada espécie e a verificação da presença de introns e suas respectivas fases em torno das regiões alinhadas. Nós verificamos que eventos de *exon shuffling* em eucariotos antigos, de origem anterior aos Metazoa, são predominantemente simétricos 0-0, enquanto nos metazoários a predominância é de unidades simétricas 1-1. Esses dados confirmam idéias anteriores de que a transição para a multicelularidade animal foi marcada pelo embaralhamento extensivo de exons e domínios 1-1. O metazoário basal *Trichoplax adhaerens* pode ser considerado um representante desta transição, evidenciada

pelas freqüências balanceadas de regiões simétricas 0-0 e 1-1. O sinal de flanqueamento por introns em torno das bordas de domínios protéicos confirmou os resultados obtidos através dos alinhamentos, com a prevalência de domínios 0-0 em não metazoários e 1-1 em metazaoários. Um agrupamento hierárquico de domínios flanqueados por introns foi construído, permitindo identificar domínios ou grupos de domínios com evidência de expansões em períodos específicos, como nos vertebrados. Por fim, os genes envolvidos em eventos de *exon shuffling* foram analisados quanto ao enriquecimento em termos do *Gene Ontology*. Os resultados indicaram que este mecanismo contribuiu significativamente para a formação de genes relacionados com uma grande diversidade de termos, alguns dos quais envolvidos diretamente com características de metazoários e vertebrados, tais como matriz extracelular, adesão, coagulação sangüínea, processos do sistema imune e sistema nervoso.

Palavras-chave: *exon shuffling*, domínios, introns, metazoários, evolução.

ABSTRACT

França, G.S. **Evolutionary history of exon shuffling in eukaryotes.** 2009. 82p. Master Thesis. Graduate Program in Biochemistry. Instituto de Química, Universidade de São Paulo, São Paulo.

Exon shuffling was first proposed by Walter Gilbert in 1979 as a mechanism in which exons from different genes could be combined to lead the creation of new genes. The mechanism of exon shuffling is favored by intronic recombinations and it is correlated with symmetry of exons. Evidence of this mechanism come from analyses of intron phase combinations, correlations between the borders of exons and domains and domain recurrence in several proteins. Taking this into account, the evolution of proteins formed by exon shuffling can be inferred regarding the exon-intron organization of the genes, the pattern of intron phase combinations and the protein domain organization. In this sense, protein regions that were probably arose by exon shuffling events were identified through a large scale analysis in several eukaryotic species. The strategy was based on alignments between all annotated proteins from a given species. Then, the aligned regions were verified in respect with intron phase combinations surrounding them. We have found that exon shuffling events in early eukaryotes are preferentially symmetric of phase 0, while in metazoans, the preference is for 1-1 symmetric units. These data confirms previous ideas that the transition to animal multicellularity was marked by extensive 1-1 exon shuffling. The basal metazoan *Trichoplax adhaerens* is a representative of this transition, evidenced by the balanced frequencies of 0-0 and 1-1 symmetric regions. The signal of intron flanking around the borders of protein domains corroborated previous analyses, showing that non metazoans have higher frequencies of 0-0

domains and metazoans have higher frequencies of 1-1 domains. A hierarchical clustering of domains flanked by introns was built, allowing us to identify domains or groups of domains with evidence of expansions during specific periods, such as in vertebrates. Finally, genes involved in exon shuffling events were analyzed regarding the Gene Ontology enriched terms. The results indicated that this mechanism significantly contributed to the creation of genes related with a large diversity of terms, some of them are directly involved with features of metazoans and vertebrates, such as extracellular matrix, cell adhesion, blood coagulation and immune and nervous system processes.

Keywords: exon shuffling, domains, introns, metazoans, evolution.

SUMÁRIO

1. Introdução

1.1. Organização exon-intron dos genes eucarióticos

O termo “gene” foi proposto por Wilhem Johannsen em 1909 baseado no conceito desenvolvido por Gregor Mendel em 1866. Ao estudar o cruzamento entre plantas, Mendel observou que certas características eram transmitidas aos seus descendentes através de unidades ou fatores discretos hereditários. Atualmente um gene é definido pelo *Human Genome Nomenclature Organization* como “um segmento de DNA que contribui para o fenótipo/função. Na ausência de demonstração de função, um gene pode ser caracterizado pela sua seqüência, transcrição ou homologia” (Wain *et al*. 2002). O *Sequence Ontology Consortium* define gene como “região localizável de uma seqüência genômica, correspondendo a uma unidade hereditária, que é associada à regiões regulatórias, regiões transcritas e/ou outras regiões funcionais” (Pearson, 2006). Dicussões mais aprofundadas e as implicações dos dados recentes sobre o conceito de gene podem ser vistas em Gerstein *et al*. 2007.

Os genes bacterianos geralmente são colineares ao seu produto protéico e normalmente organizados em unidades denominadas operons. Os genes eucarióticos por sua vez, possuem regiões intervenientes às seqüências codificantes. As regiões intervenientes são removidas do transcrito primário durante o processamento do RNA através do mecanismo de *splicing* e são chamadas de introns. As seqüências codificantes, presentes no RNA maduro, são chamadas de exons (figura 1). Embora em alguns eucariotos unicelulares uma pequena fração dos genes contém introns e alguns genes de eucariotos multicelulares não serem

interrompidos, a organização exon-intron pode ser considerada uma característica universal dos genes eucarióticos (Lewin, 2007).

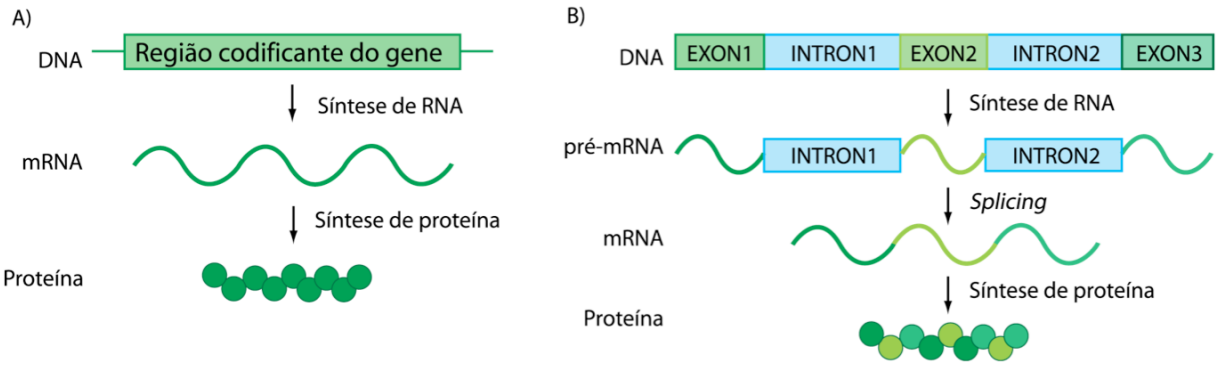


Figura 1. Genes procarióticos e eucarióticos. A) Estrutura básica dos genes procarióticos. A região codificante do gene é colinear com a molécula de mRNA. B) Estrutura dos genes eucarióticos. As seqüências codificantes (exons) são interrompidas por seqüências não codificantes (introns) que são removidas por *splicing* durante o processamento do pré-mRNA. Modificado de Lewin (2007).

O tamanho e o número de introns varia enormemente nas espécies. Cerca de 96% dos genes de *Saccharomyces cervisae* por exemplo, não são interrompidos por introns. Já nos insetos e mamíferos, apenas 6% dos genes não contém introns (Lewin, 2007). Em alguns eucariotos unicelulares (e.g *Giardia lamblia*) existem um ou dois introns por genoma, enquanto nos vertebrados a média é de cinco a oito introns por gene (Koonin, 2009). Além disso, é comum encontrarmos introns com tamanhos variando de 200 pares de base até dezenas de quilobases. Por outro lado, os exons são muito mais conservados que os introns, não apresentam significativas variações de tamanho e em média codificam para 50 aminoácidos (Lewin, 2007).

A origem e evolução dos introns tem sido objeto de estudo desde a sua descoberta. Ao contrário do que se acreditava na década de 80 e início dos anos 90, os introns não são mais considerados “DNA lixo”, termo utilizado para se referir à

sua aparente ausência de função. Os introns podem atuar em diversos processos, como na regulação da expressão gênica e origem de microRNAs, participação no mecanismo de *splicing*, recombinação, sinais de exportação do mRNA do núcleo para o citoplasma, regulação da estabilidade do mRNA e na origem de novos genes. (Fedorova e Fedorov, 2003).

Mudanças na organização exon-intron estão sujeitas a grandes variações linhagem-específicas e estão relacionadas com a complexidade estrutural e funcional dos genomas (Yandell *et al.* 2006). O estudo destas variações é fundamental para a compreensão da evolução dos genomas e como novos genes podem ser originados.

1.2. Origem de novos genes

No início da era genômica os pesquisadores perceberam que o tamanho dos genomas, assim como o número de genes, não estão diretamente relacionados com a complexidade fenotípica das espécies. Este suposto contra-senso entre o número de genes e a complexidade é conhecido como *G-value* ou *N-value paradox* (Betrán e Long, 2002; Claverie, 2001). A chave para a interpretação da complexidade biológica não está ligada somente ao número de genes que uma espécie possui, mas na intrincada rede de interações entre os produtos gênicos, mecanismos regulatórios, diversidade de transcritos, etc. (Szathmáry *et al.* 2001).

A evolução geralmente se utiliza de formas preexistentes para a criação de novidades, e a origem de novos genes está fundamentada neste princípio. Baseado na revisão de Long *et al*. (2003), a seguir são descritos sucintamente os principais mecanismos pelos quais novos genes podem ser formados. Embora

descritos separadamente, em muitos casos os genes podem ser formados pela combinação dos diferentes mecanismos apresentados (ver resumo na figura 2).

1.2.1. Duplicação gênica

Uma região cromossômica contendo um gene pode sofrer rearranjos através de *crossing over* desigual em regiões homólogas, reproduzindo parte de um gene, uma cópia inteira ou vários genes em seqüência. Duplicações também podem ocorrer por eventos de retrotransposição. A duplicação de cromossomos ou genomas inteiros também é possível principalmente devido à erros de segregação durante a meiose, fenômeno muito observado em plantas. Duplicações resultam na formação de famílias gênicas e genes pertencentes à mesma família são referidos como parálogos. (Zhang, 2003). As conseqüências evolutivas destes processos são essenciais para a geração de diversidade. A nova cópia de um gene, sujeita à ação da seleção natural, pode sofrer três destinos principais: i) acumular mutações que levam à inativação, gerando um pseudogene ou ii) aquisição de uma nova função (neofuncionalização) e iii) especialização das duas cópias, desempenhando funções complementares (subfuncionalização) (Hurles, 2004).

1.2.2. Elementos móveis

A integração de elementos móveis em genes nucleares, como elementos Alu e outros, pode alterar a composição do gene hospedeiro e formar um novo transcrito pela aquisição de mutações ou alteração do sítio de *splicing* (Long *et al.,* 2003). Um novo gene também pode ser formado através da transcrição reversa de seu mRNA e

reintegração em uma nova localização genômica. Este mecanismo é mediado por elementos chamados de retrotransposons. Como a retrotransposição geralmente não reintegra o promotor original, a funcionalidade da nova cópia depende da inserção junto à uma região regulatória, e por isso o gene pode adquirir uma função diferente daquela apresentada por seu gene parental (Long *et al.*, 2003).

1.2.3. Transferência lateral de genes

Transferência lateral é o termo que se aplica quando um gene não é herdado de forma vertical, ou seja, da geração parental para seus descendentes. Este fenômeno é muito comum em procariotos, quando uma região genômica é transferida de um organismo para outro através de conjugação, transdução ou transformação (Ochman *et al.,* 2000). Recentemente foram reportados casos de transferência lateral em eucariotos e um dos exemplos é a integração de grande parte do genoma do endossimbionte *Wolbachia pipentis* em espécies de *Drosophila*. Porém, os mecanismos envolvidos nestes casos são ainda desconhecidos (Hotopp *et al.,* 2007).

1.2.4. Fusão ou fissão gênica

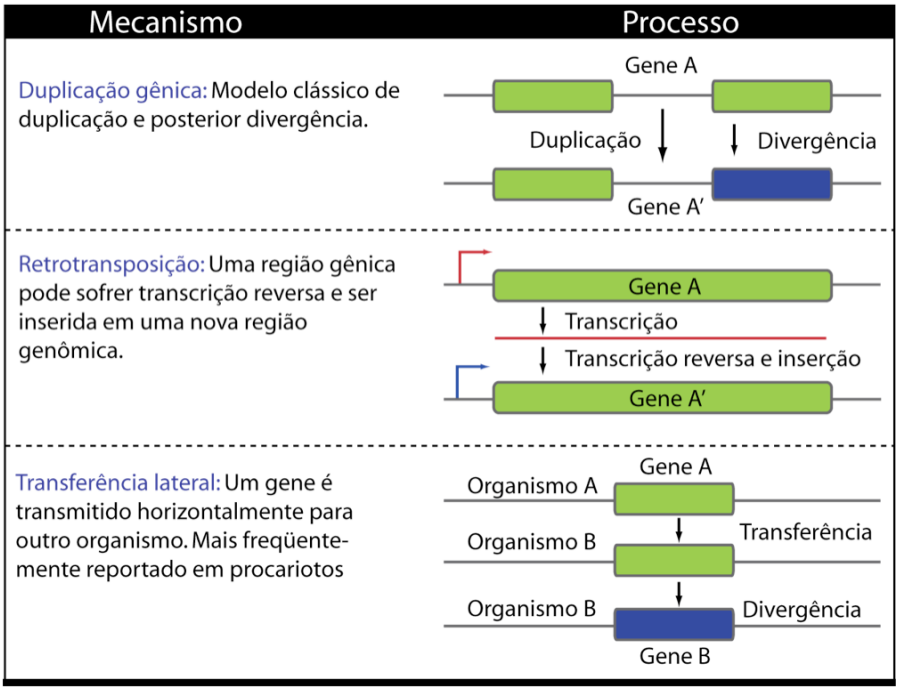
Um ou mais genes podem se fusionar e tornarem-se parte do mesmo transcrito através de deleções ou mutações na região do códon de terminação. Os genes também podem sofrer quebras (fissão) em duas ou mais partes. Sabe-se muito pouco sobre o mecanismo pelo qual a fissão ocorre. (Chandrasekaran e Betrán, 2008).

1.2.5. Origem "de novo"

Novos genes podem ser formados a partir do recrutamento de regiões não codificantes. Embora os mecanismos envolvidos sejam desconhecidos, uma das possibilidades é o acúmulo de mutações pontuais em regiões não codificantes que passam então a ser codificantes (Zhou *et al.*, 2008).

1.2.6. Exon shuffling

Através da recombinação espúria entre introns ou por retrotransposição, exons e grupos de exons podem ser embaralhados ou inseridos em genes distintos. *Exon shuffling* é descrito como um dos principais mecanismos geradores de diversidade protéica (Patthy, 1996). Mais detalhes sobre este mecanismo serão discutidos ao longo deste trabalho.



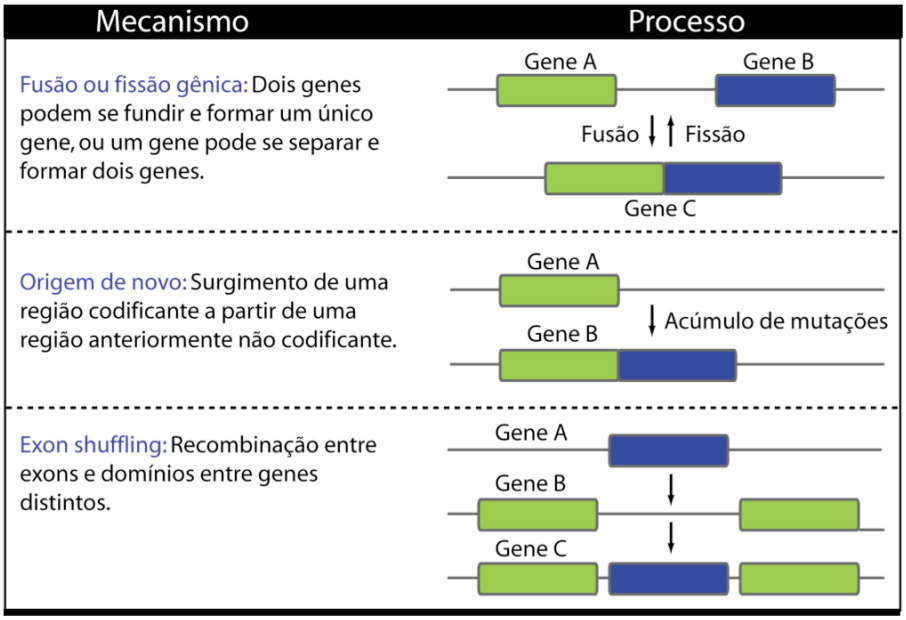


Figura 2. Ilustracão dos principais mecanismos de origem de novos genes. Modificado de Long *et al.* (2003).

1.3. Origem e evolução dos introns

Os introns são as regiões interpostas aos exons, normalmente removidas do transcrito de RNA e não contribuem para a formação das proteínas. Existem quatro principais grupos de introns. Os introns do Grupo I são encontrados em genomas de organelas e bactérias e em RNAs ribossômicos no núcleo de protistas e fungos. Os introns do Grupo II estão presentes em genomas de bactérias, em organelas de plantas, fungos e protistas e no genoma de *Methanosarcina*, gênero representante de archaea. Estes introns estão ausentes em genomas nucleares e nas mitocôndrias de animais. Tanto os introns do Grupo I como os do Grupo II (*in vitro*) são capazes de catalisar sua auto-excisão (Rodriguez-Trelles *et al.,* 2006).

Os introns de *splicing* nuclear estão presentes nos genomas nucleares de eucariotos e sua remoção depende de um complexo celular formado por RNAs e centenas de proteínas chamado spliceossomo (Lewin, 2007). Os mecanismos envolvidos na excisão dos introns do Grupo II guardam similaridades com os introns de *splicing* nuclear, sugerindo que estes possam ter se originado a partir da transferência e proliferação de introns do Grupo II das mitocôndrias (Koonin, 2006). Ainda existem os introns presentes nos genes nucleares de tRNAs que são removidos por um mecanismo muito diferente dos introns de *splicing* nuclear (Rodríguez-Trelles *et al.*, 2006).

Logo após a descoberta dos introns em 1977 por Phillip Sharp e Richard Roberts, muitos autores questionaram como e quando os introns surgiram e que funções desempenham nas células, dando origem a intensos debates na literatura sobre a evolução dos introns de *splicing* nuclear.

1.3.1. Teoria "introns-early"

A teoria *introns-early* original propõe que a maioria dos introns atuais dos eucariotos estava presente no ancestral comum dos eucariotos e procariotos (progenoto). A presença dos introns no progenoto teria papel fundamental na construção das primeiras proteínas por facilitar a recombinação de exons ou módulos protéicos (*exon shuffling*), acelerando o processo evolutivo (Doolittle, 1978; Gilbert, 1978). Assim, os primeiros genes seriam formados pela combinação de mini-genes ou exons. Esta idéia é conhecida como *The Exon Theory of Genes*. (Gilbert, 1987). A ausência dos introns nos procariotos é explicada pela pressão seletiva em maximizar a taxa de replicação do DNA, economizando energia metabólica em

organismos com altas taxas de reprodução. A teoria *introns-early* é compatível com um cenário em que a presença dos introns foi crucial para o surgimento dos primeiros genes no progenoto. Parte dos introns dos eucariotos seriam remanescentes dos primeiros genes enquanto os procariotos sofreram perda massiva de introns (Roy e Gilbert, 2006).

1.3.2. Evidências que suportam a teoria "introns-early"

As primeiras evidências a favor da teoria *introns-early* vieram de análises da estrutura do gene da triose fosfato isomerase (TPI). Este gene apresenta alto grau de conservação entre espécies de eucariotos e as posições de introns também são conservadas entre vertebrados e plantas (Marchionni e Gilbert, 1986). Utilizando a definição de módulos protéicos de Go (1981), foi possível correlacionar as bordas dos módulos com as posições dos introns. Previu-se a existência de um intron ao redor do aminoácido 64, sugerindo a correlação entre introns e módulos protéicos. A presença deste intron foi posteriormente confirmada em *Culex tarsalis* (Tittiger *et al.*, 1993). Estes resultados iniciaram um período de fortes evidências a favor da teoria *introns-early*, com o principal argumento de que genes antigos, cuja origem antecede a divergência dos eucariotos, teriam sido formados pela combinação de módulos protéicos através de recombinações intrônicas (discutido por De Souza, 2003).

Uma das predições da teoria é que a distribuição de fases de introns não é aleatória. Fase de intron é a posição em que o intron reside dentro dos códons. Introns de fase 0 estão inseridos entre dois códons, introns de fase 1 estão entre o primeiro e o segundo nucleotídeo do códon e introns de fase 2 residem entre o segundo e o terceiro nucleotídeo do códon (ver figura 1.4A).

Long e colaboradores (1995) testaram as predições da teoria *introns-early* considerando a distribuição de fases de introns ao redor de regiões ancestrais conservadas (RACs), que correspondem à regiões gênicas conservadas entre eucariotos e procariotos. Os autores observaram que a distribuição de fases de introns não é aleatória, sendo 48% de introns de fase 0, 30% de fase 1 e 22% de fase 2. Além disso, observou-se que existe excesso de exons simétricos (exons flanqueados por introns de mesma fase) nas RACs, dando suporte à idéia de que *exon shuffling* foi primordial para a construção de genes antigos.

Outra importante contribuição para o debate surgiu com o trabalho de De Souza *et al*. (1998). Neste estudo os autores realizaram testes com uma amostra de 44 proteínas eucarióticas com estruturas tridimensionais determinadas que compartilham regiões conservadas com procariotos. Os autores observaram que somente as posições de introns de fase 0 estão correlacionadas com as bordas de módulos de proteínas antigas. Os módulos foram baseados na definição de Go (1981) que correspondem à subregiões compactas de uma cadeia polipeptídica cujas distâncias entre os carbonos alfa apresentam determinado diâmetro. Go (1981) observou correspondência entre exons do gene da hemoglobina e módulos de 28Å (Angstroms) de diâmetro. Estes módulos representam unidades de *folding* independentes (Panchenko *et al.*, 1996).

De acordo com a previsão da teoria *introns early*, espera-se encontrar introns entre os módulos como resultado de eventos de *exon shuffling*. O trabalho de De Souza *et al.* (1998) apresenta o que os autores denominam de Teoria Sintética da Evolução dos Introns, assumindo que uma porção dos introns (cerca de 30%) teria origem no progenoto, sendo a maioria de fase 0, com posições correlacionadas com módulos de 21, 27 e 33Å, enquanto os demais introns foram inseridos tardiamente

durante a evolução dos eucariotos. Fedorov *et al*. (2001) observaram os mesmos padrões de correlação entre bordas de módulos protéicos e a presença de introns de fase 0, desta vez com uma amostra de 276 proteínas antigas em módulos de 28Å. Este resultado foi explorado novamente por Fedorov *et al.* (2003) utilizando posições de introns com forte sinal de conservação entre vertebrados, invertebrados, fungos, plantas e protistas, o que caracteriza um conjunto definitivamente antigo de introns.

1.3.3. Teoria "introns-late"

A teoria *introns-late* assume que os introns foram adquiridos tardiamente, com ganhos contínuos de introns durante a evolução dos eucariotos (Cavalier-Smith, 1991; Palmer e Logsdon, 1991; Logsdon, 1998). O aparecimento dos introns nos eucariotos representa um panorama mais parcimonioso do que a proposta *introns-early*, que assume a perda dos introns nos procariotos, entretanto, a parcimônia é um critério estatístico e não deve ser necessariamente aplicado em eventos de singular natureza como a origem dos introns (Koonin, 2006).

1.3.4. Evidências que suportam a teoria "introns-late"

Os argumentos utilizados pela teoria *introns-late* baseiam-se principalmente na distribuição filogenética e no processo de ganho de introns. Foi observado que os introns do grupo II, presentes em eubactérias e em genomas de mitocôndrias e cloroplastos, apresentam mecanismos de *splicing* semelhantes aos introns de *splicing* nuclear (Rogers, 1990). O fato de alguns destes introns serem considerados elementos móveis por possuírem capacidade de retrotransposição (Robart e

Zimmerly, 2005) e a idéia de que a célula eucariótica descende de uma célula bacteriana, serviram para postular que os introns de *splicing* nuclear e a maquinaria de *splicing* não poderiam ser funcionais nos primeiros genes. A razão para isto é que os introns teriam emergido nas células eucarióticas através da co-evolução de proteínas eucarióticas e a invasão de introns provenientes de um endossimbionte mitocondrial nos genes nucleares da célula hospedeira (Rogozin *et al.*, 2005; Koonin, 2006; Rodríguez-Trelles *et al.*, 2006).

A descoberta inicial da presença do intron no gene da TPI em *Culex* foi um forte indício de correlação entre introns e módulos protéicos, favorencendo a teoria *introns-early*. A questão a saber era se este intron tinha origem antiga ou se representava uma inserção recente. A presença deste intron em outros táxons poderia revelar sua origem antiga, contudo, verificou-se que este intron estava ausente no gene da TPI de outras espécies de insetos, além disso, foram detectados outros sete introns adicionais neste gene, sugerindo inserções recentes (Logsdon *et al.*, 1995).

Um dos pontos apresentados pela teoria *introns-early* é o excesso de introns de fase 0 e que estes introns teriam sua origem no progenoto (De Souza *et al.*, 1998). A explicação da teoria *introns-late* provém da hipótese de que o excesso de introns de fase 0 é decorrente da inserção preferencial de introns em sítios específicos denominados “proto-splice sites” (Qiu *et al.*, 2004). A inserção após o terceiro nucleotídeo do códon, que geralmente é uma posição sinônima, exerceria pouca influência negativa na eficiência do *splicing*, permitindo a fixação preferencial destes introns (Sverdolov *et al.*, 2003).

1.3.5. Considerações sobre o debate "introns-early/late"

Nos últimos anos, evidências à favor das teorias *introns-early* e *introns-late* foram sendo acumuladas, contudo, a questão sobre a presença ou ausência dos introns no progenoto e o seu papel na montagem dos primeiros genes ainda é um mistério a ser desvendado pela biologia. A densidade de introns nos eucariotos é extremamente variável e não apresenta um padrão filogenético claro. Nem sempre eucariotos primitivos são depletados e eucariotos recentes são ricos em introns, isso implica em episódios recorrentes de perda e ganho de introns (Roy e Gilbert, 2005; Jeffares, 2006; Roy e Gilbert, 2006).

Análises sobre a conservação das posições de introns mostram alguns padrões importantes. Rogozin *et al*. (2003) verificaram a presença ou ausência de introns ortólogos em espécies de animais, plantas, fungos e protistas. De acordo com os autores, cerca de 24% dos introns humanos são compartilhados com a planta *Arabidopsis thaliana*. O número de introns humanos compartilhados com plantas é de duas a três vezes maior do que em insetos e nematódeos, indicando perda de introns nestas duas linhagens. Aproximadamente um terço dos introns do protista *Plasmodium falciparum* têm correspondência com algum outro grupo dos eucariotos analisados. Isto sugere que muitos introns têm sido conservados há pelo menos 1.5 bilhões de anos e que possivelmente o ancestral comum dos eucariotos possuia um genoma rico em introns. Raible e colaboradores (2005) chegaram a conclusões parecidas no que se refere ao ancestral comum dos animais. Comparando as posições de introns do anelídeo *Platynereis dumerilii*, os autores observaram que cerca de 60% dos introns encontravam-se em posições idênticas aos ortólogos humanos.

O trabalho de Carmel *et al.* (2007) apresenta um modelo probabilístico que determina taxas de ganhos e perdas de introns em 19 espécies eucarióticas. O estudo revelou de três modos distintos na dinâmica de evolução dos introns. O modo balanceado aparentemente atua em todas as linhagens, em que as taxas de ganho e perda de introns são correlacionadas positivamente. O segundo modo envolve elevadas taxas de perda de introns, experimentada por exemplo por fungos, insetos e nematódeos. O terceiro modo é representado por altas taxas de ganho de introns, associado à eventos de especial impacto na evolução dos eucariotos, como a origem dos animais. Os resultados do trabalho sugerem que uma densidade relativamente alta de introns foi adquirida muito cedo durante a evolução dos eucariotos (> 2.15 introns por kb) e no ancestral dos organismos multicelulares (~3.39 introns por kb).

Os estudos de genômica comparativa têm demonstrado que o ancestral comum dos eucariotos provavelmente era relativamente rico em introns. Esta hipótese é reforçada pela descoberta de um ou poucos introns em eucariotos extremamente primitivos, como *Giardia* (Nixon *et al.*, 2002) e *Trichomonas* (Vanacová *et al.*, 2005). Não há evidências empíricas que validem a hipótese original *introns-early*, que presume a existência de numerosos introns nos primeiros estágios da vida, porém, a visão radical da teoria *introns-late*, em que maioria dos introns presentes nos eucariotos são relativamente recentes e que foram adquiridos por um processo contínuo de inserção não é mais aceita. De acordo com Koonin (2006), os introns de *splicing* nuclear surgiram de um *pool* inicial (provavelmente introns do Grupo II e retroelementos) que invadiram uma célula hospedeira, sendo que a principal contribuição dos introns para a complexidade biológica se deu após o surgimento dos eucariotos. Segundo o autor, não há “perdedores” neste debate e

ambas as hipóteses trouxeram contribuições fundamentais para o entendimento atual da evolução dos introns.

1.4. *Exon shuffling*

A descoberta da organização exon-intron dos genes eucarióticos representou um marco para a biologia molecular e abriu portas para novas interpretações em relação às funções e evolução dos genes e genomas. Pouco tempo após a descoberta dos introns, Walter Gilbert (1978) em seu artigo entitulado "*Why genes in pieces?"* sugeriu a denominação de introns para as regiões "intragênicas" e exons para as regiões "expressas" dos genes. Gilbert vislumbrou o uso alternativo de exons, hoje conhecido como *splicing* alternativo, e o mais importante para este trabalho, a possibilidade de recombinações intrônicas entre genes diferentes, o que poderia eventualmente causar a inserção ou o embaralhamento de exons. De acordo com Gilbert, os introns podem ser caracterizados como *hot spots* de recombinações. As razões para este argumento é que os introns são comparativamente muito maiores que os exons, estão sujeitos à menor pressão seletiva por não serem codificantes, e apresentam longas seqüências repetitivas. A idéia de Gilbert é que as recombinações entre introns permite o embaralhamento de unidades codificantes entre os genes, acelerando o processo evolutivo e a geração de novidades funcionais. Este mecanismo é conhecido como *exon shuffling.*

1.4.1. Exon shuffling: Mecanismos

São conhecidos basicamente dois mecanismos através dos quais exons podem ser inseridos em regiões distintas e podendo originar um novo gene. Os mecanismos principais são a retrotransposição e a recombinação ilegítima entre introns.

1.4.1.1. Retrotransposição

Retrotransposons são seqüências de DNA que podem produzir cópias de si mesmas através da transcrição reversa da sua molécula de RNA e eventualmente serem reintroduzidas no genoma (Eickbush, 1999). Acredita-se que cerca de 40% do genoma humano seja resultado de eventos de retrotransposições (Lander *et al.,* 2001). Embora a introdução destes elementos em certas regiões possa causar efeitos deletérios, em alguns casos as conseqüências podem ser benéficas. Moran e colaboradores (1999) realizaram ensaios envolvendo a retrotransposição do elemento LINE-1 (L1). O elemento L1 contém uma região 5' não traduzida (5'UTR), duas fases abertas de leitura (ORF1 e ORF2) e uma região 3'UTR que termina em uma cauda poli(A). Os autores observaram que após a transcrição, o elemento L1 inserido no intron de um gene continha uma cauda poli(A) correspondente à cauda do gene hospedeiro, e não a própria cauda do L1. Isto sugeriu que a cauda poli(A) do elemento L1 apresenta fraco sinal, e portanto a transcrição pode avançar até a cauda poli(A) do gene hospedeiro. Durante este processo, um exon ou grupos de exons podem ser transcritos junto com o L1. Ao sofrer a transcrição reversa para uma molécula de cDNA, o L1 pode se integrar em outro gene carregando consigo

exons de um gene “doador”. A figura 3 ilustra como um novo gene pode surgir através deste mecanismo. A origem do gene *jingwei* de *Drosophila* é um exemplo de formação de um gene quimérico a partir do recrutamento de seqüências retrotranspostas, neste caso, o *jingwei* contém a retroseqüência do gene da álcool desidrogenase (*Adh*). (Long e Langley, 1993).

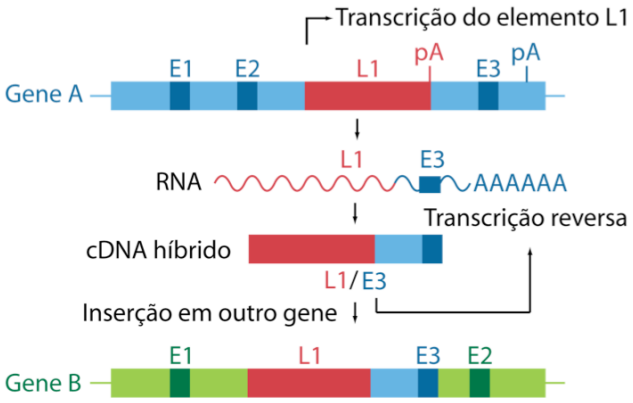


Figura 3. *Exon shuffling* através de retrotransposição. O elemento L1 possui fraco sinal de poli(A) e a transcrição pode continuar até que um novo sinal de poli(A) seja encontrado. O transcrito resultante possui seqüências não só do elemento L1 como por exemplo o exon 3 do gene A. A transcrição reversa deste segmento dá origem a uma molécula de cDNA híbrida que pode ser inserida em outro gene, dando origem ao gene B. Modificado de Strachan e Read (2003).

1.4.1.2. Recombinação ilegítima

Recombinação ilegítima pode ser definida como a recombinação entre seqüências não homólogas ou que apresentam pequenos trechos de similaridade. Estas recombinações ocorrem com freqüência em procariotos e eucariotos produzindo rearranjos no genoma (van Rijk e Bloemendal, 2003). Através deste mecanismo, exons ou grupos de exons podem ser inseridos em outros genes através da recombinação entre os introns que flanqueiam os exons e os introns do

gene aceptor, criando eventualmente um novo gene. Recombinações intrônicas também podem provocar duplicações e/ou deleções dentro de um gene, comum por exemplo em genes de receptores de LDL (Patthy, 1999a). Um aspecto importante dos introns de *splicing* nuclear é que somente um pequeno segmento é essencial para o *splicing*, que são as seqüências conservadas nas porções 5' e 3' das bordas exon/intron. As demais porções dos introns são bastante tolerantes à deleções e inserções. Isto faz com que os sítios de *splicing* de introns diferentes sejam equivalentes e assim, mesmo após as recombinações, o *splicing* não é afetado. Outra característica que facilita a recombinação é a presença de seqüências repetitivas, como repetições Alu, podendo aumentar as chances de pareamento entre introns não homólogos (Patthy, 1999a).

O sucesso de um evento de *exon shuffling* através de recombinações intrônicas depende da compatibilidade de fases de intron. Conforme foi definido na seção 1.3.2, fase de intron é a posição dos introns dentro dos códons (figura 4A). Considerando as fases dos introns que flanqueiam os exons, nove combinações de fases são possíveis. Exons flanqueados por introns de mesma fase são chamados de simétricos (figura 4B), quando flanqueados por introns de fases diferentes são chamados de exons assimétricos (figura 4C).

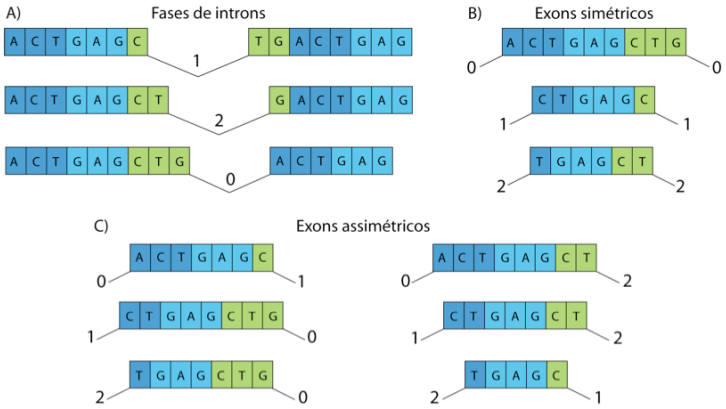


Figura 4. Ilustração representando as fases de introns e simetria de exons. A) Fases de introns. Trincas de cores diferentes correspondem aos códons dos exons. Introns de fase 0 situam-se entre dois códons. Introns de fase 1 estão entre o primeiro e o segundo nucleotídeo e introns de fase 2 residem entre o segundo e o terceiro nucleotídeo deo códon. B) Exons simétricos possuem introns flanqueadores de mesma fase. C) Exons assimétricos possuem introns flanqueadores de fases distintas. Retirado de Kolkman e Stemmer (2001).

O quadro de leitura do gene somente é preservado integralmente nos casos onde ocorre a inserção de exons simétricos em introns de fase idêntica, ou seja, os introns recombinantes devem possuir as mesmas fases (Patthy, 1987). Quando ocorre a inserção de exons assimétricos, o quadro de leitura do gene aceptor é alterado durante a tradução, tornando menos provável a geração de um produto funcional. Por isso, espera-se que eventos de *exon shuffling* envolvendo exons simétricos tenham maiores chances de serem mantidos pela seleção natural (figura 5). Long *et al.* (1995) mostraram que existe excesso estatisticamente significativo de simetria para exons e grupos de até seis exons, este resultado é interpretado como forte evidência de sucessivos eventos de *exon shuffling* durante a evolução. Também é considerado *exon shuffling* em alguns casos quando ocorrem duplicações internas dentro dos genes, gerando repetições em tandem de exons ou conjuntos de exons. Este fenômeno é relativamente comum e existem diversos casos reportados na literatura (Björklund *et al.*, 2006).

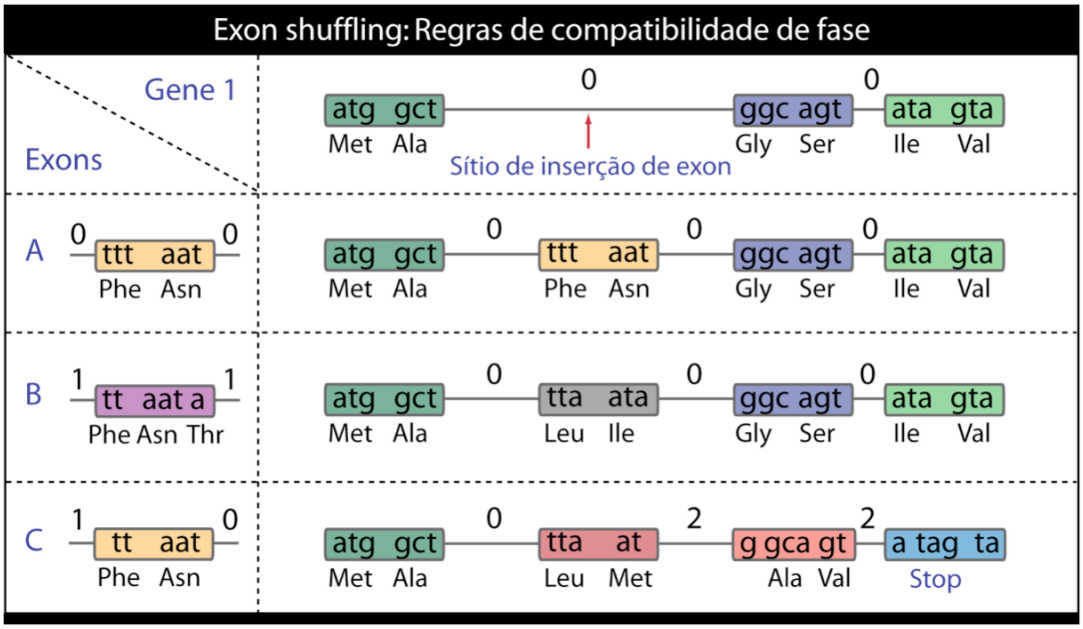


Figura 5. Representação do mecanismo de *exon shuffling* envolvendo exons simétricos e assimétricos. A) A inserção de um exon simétrico em um intron de mesma fase não altera o quadro de leitura do Gene 1. B) A inserção de um exon simétrico de fase diferente do intron aceptor altera a leitura do exon inserido. C) A inserção de um exon assimétrico altera todo o restante do quadro de leitura do Gene 1. Modificado de Vibranovski (2005a).

1.4.2. Exon shuffling e domínios protéicos

De acordo com a hipótese proposta por Gilbert, os exons primordiais corresponderiam à módulos protéicos. Eventos de duplicações e embaralhamento destas unidades resultaria na expressão de novas proteínas que combinam diversos módulos protéicos (Gilbert, 1987). Evidências para esta hipótese vieram principalmente de estudos baseados nas posições e fases de introns ao redor de módulos protéicos definidos estruturalmente, mostrando que existe correlação entre exons e módulos de proteínas antigas (De Souza *et al.*, 1996; De Souza *et al.*, 1998). Alguns outros casos foram descritos, como a correspondência entre exons e o domínio N-terminal do gene *jingwei* de *Drosophila* (Long *et al.*, 1999). Contudo, estes estudos foram realizados em períodos em que a disponibilidade dos dados

genômicos era infinitamente menor do que a atual, e por isso concentram-se em casos individuais ou em poucos genes.

Domínios protéicos podem ser definidos como unidades evolutivas que apresentam configuração estrutural independente e assumem funções específicas (Söding e Lupas, 2003). Os domínios geralmente apresentam padrões de seqüências conservadas de aminoácidos que podem ser utilizados na definição ou identificação de domínios, como por exemplo os domínios Pfam (Finn *et al.*, 2008). Liu e Grigoriev (2004) investigaram correlações exon-domínio em escala genômica. Os autores demonstraram que existe correlação fortemente significativa entre as bordas de domínios e as bordas dos exons em várias espécies de eucariotos e na maioria dos casos os domínios são simétricos, preferencialmente circundados por introns de fase um. Resultados similares foram obtidos por Kaessmann *et al.* (2002) considerando domínios flanqueados por introns, mostrando que existe enriquecimento de domínios simétricos codificados por grupos de até quatro exons (ver esquema na figura 6).

Estas observações indicam claramente que *shuffling* de exons ou grupos de exons, pode promover novas combinações de domínios e contribuir para aumentar o repertório protéico, tornando os genomas mais complexos e funcionalmente diversos (Chotia *et al.*, 2003). O potencial do fenômeno de *exon shuffling* para a aquisição de novas formas é vasto e inclusive alguns estudos utilizam abordagens experimentais *in vitro* para a construção de proteínas visando produtos com potencial terapêutico (Kolkman e Stemmer, 2001).

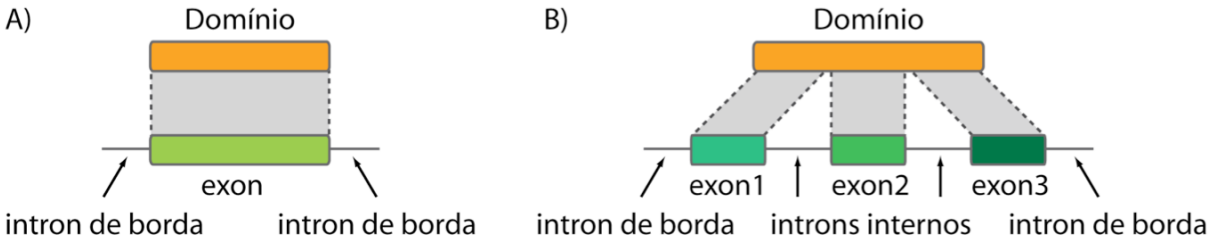


Figura 6. Esquema representativo das correlações entre exons e domínios protéicos. A) Correlação entre domínio e seu exon codificante. B) Correlação entre domínios e grupos de exons. Kaessmann *et al.* (2002) observaram excesso de domínios simétricos rem relação aos introns de borda tanto no caso "A" como no caso "B", com grupos de até quatro exons. Modificado de Kaessmann *et al.* (2002).

1.4.3. Exon shuffling e a evolução de proteínas multidomínios

A maioria das proteínas é formada por um ou vários domínios protéicos. Cerca de 70% das proteínas de archaea, bacteria e eucariota possuem pelo menos um domínio Pfam (domínios identificados por padrões de seqüência) (Ekman *et al.,* 2005). Os domínios representam unidades básicas diretamente relacionadas com a estrutura e função geral da proteína. O rearranjo de domínios obtido principalmente por fusão ou fissão gênica, duplicação e recombinação é um processo extremamente importante em termos evolutivos porque permite a criação de proteínas multidomínios e o surgimento de novas funções. A incorporação de um domínio numa proteína não apenas pode agregar a função do domínio inserido, como alterar substancialmente a conformação tridimensional da proteína, modificando as interfaces entre os domínios, criando sítios ativos, etc. (Bashton e Chotia, 2007). Além disso, sabe-se que os domínios medeiam interações entre proteínas. A aquisição de novas combinações pode modificar redes de interações e alterar vias de sinalizações (Bornberg-Bauer *et al.*, 2005). Desta forma é fácil

perceber como o rearranjo de domínios é um eficiente mecanismo de inovação e seu enorme potencial para gerar sistemas e formas biologicamente complexas.

A distribuição de proteínas multidomínios nos três reinos (archaea, bacteria e eukarya) varia com a metodologia e as bases de dados utilizadas, mas o ponto comum é que os eucariotos possuem mais proteínas multidomínios do que os procariotos. (Koonin *et al.,* 2002). Dentro dos eucariotos, a freqüência de proteínas com mais de um domínio é maior nos metazoários[1], com cerca 39%, nos eucariotos unicelulares a freqüência é de 32%. A proporção de proteínas multidomínios decresce no seguinte sentido: metazoa > plantas > fungi ~ protozoa > bacteria > archaea. A correlação entre o número de proteínas e o número de domínios constituintes também é maior nos metazoários, proteínas com mais de 10 domínios é aproximadamente nove vezes mais freqüente que em archaea (Tordai *et al.,* 2005).

Exon shuffling é um dos principais mecanismos de rearranjo de domínios e criação de proteínas multidomínios. Em trabalhos realizados principalmente na década de 90, Patthy identificou diversas proteínas cuja estrutura de domínios, posições e fases de introns indicam claramente origem por *exon shuffling*. São proteínas comumente presentes na matriz extracelular e proteínas transmembrana com domínios voltados para a matriz, exemplos incluem lamininas, fibronectinas, colágeno, fatores de coagulação sangüínea, proteínas do complemento, receptores de LDL, fatores de crescimento, etc. A grande maioria das proteínas identificadas apresentam estrutura modular, formadas por domínios e repetições de domínios

---

[1] Metazoa é um grupo monofilético representado por animais multicelulares, inclui grupos como poríferos e vertebrados (Lake, 1990).

flanqueados simétricamente por introns de fase 1, um exemplo típico pode ser observado na figura 7 (Patthy, 1987; Patthy, 1996; Patthy, 1999b; Patthy, 2003).

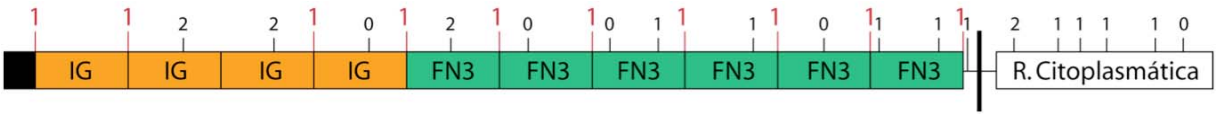


Figura 7. Estrutura do gene receptor de netrina humano. Os retângulos laranja e verde indicam os domínios imunoglobulina e fibronectina do tipo III respectivamente. As barras vermelhas indicam as posições e fases dos introns ao redor dos domínios e as barras pretas as posições e fases dos demais introns. O retângulo preto indica a região do peptídeo sinal e a barra vertical preta a região transmembrana. Nota-se que todas as repetições dos domínios IG e FNIII são flanqueados por introns de fase 1. Reproduzido de Patthy (1999b).

Estudos em larga escala utilizando dados de vários genomas têm contribuído de forma decisiva para o entendimento do papel do fenômeno de *exon shuffling* na evolução das proteínas. Kaessmann e colaboradores (2002) observaram que domínios flanqueados por introns são fortemente enriquecidos pela classe 1-1. Esta classe de domínios é super-representada em proteínas exclusivas de metazoários, enquanto a classe 0-0 mostrou-se enriquecida em proteínas antigas, compartilhadas entre procariotos e eucariotos, sugerindo possíveis eventos de *exon shuffling* no progenoto.

Através de análises comparativas entre nove genomas, Liu *et al.* (2005) mostraram que domínios cujas bordas estão correlacionadas com bordas de exons são preferencialmente simétricos 1-1 e tendem a ser mais abundantes e vastamente distribuídos em relação aos demais domínios, o que sugere um processo de seleção positiva durante a evolução possivelmente devido às funções que desempenham e a capacidade de estarem distribuídos em diversas arquiteturas de domínios. Esta capacidade pode ser caracterizada como a

mobilidade dos domínios. Tordai e co-autores (2005) mediram a mobilidade através do número de arquiteturas locais de domínios em uma proteína, ou seja, quantos domínios vizinhos diferentes situados à 5’ e à 3’ um dado domínio possui. Estas arquiteturas locais podem ser chamadas de tripletes, pois englobam o domínio de interesse e seus domínios vizinhos. Os domínios simétricos 1-1 estão entre os primeiros no “ranking” de mobilidade. O trabalho de Ekman *et al.* (2007) revelou que as principais mudanças e a criação de novas arquiteturas de domínios ocorreram na linhagem dos metazoários, com contribuição fundamental do grupo de domínios correlacionados com bordas de exons.

2. Objetivos

2.1. Objetivos gerais

Com a atual disponibilidade dos genomas de espécies que ocupam posições filogenéticas de extremo interesse e o crescente avanço no desenvolvimento de ferramentas de bioinformática, o objetivo deste trabalho foi o de explorar e determinar, através de análises em larga escala, quais os padrões evolutivos do fenômeno de *exon shuffling* em diferentes espécies eucarióticas.

2.2. Objetivos específicos

- Identificar regiões protéicas que possivelmente foram originadas por *exon shuffling* e determinar os padrões de simetria de fases de introns em torno destas regiões.
- Identificar domínios protéicos flanqueados por introns e inferir em que períodos durante a evolução de eucariotos há evidências de expansão por *exon shuffling*.
- Investigar aspectos funcionais e a relação entre *exon shuffling* e o surgimento de características biológicas.

3. Materiais e métodos

3.1. Espécies utilizadas no estudo e obtenção dos arquivos

As espécies que compõem o estudo foram selecionadas considerando três fatores principais: i) espécies com genomas completamente seqüenciados e a disponibilização das seqüências de proteínas preditas, ii) espécies relativamente ricas em introns, com média maior que quatro introns por gene e iii) espécies representativas dos principais grupos de eucariotos. No anexo B estão detalhadas as informações sobre os grupos taxonômicos aos quais as espécies pertencem e as bases de dados públicas das quais foram obtidas as seqüências de proteínas e arquivos de anotação genômica.

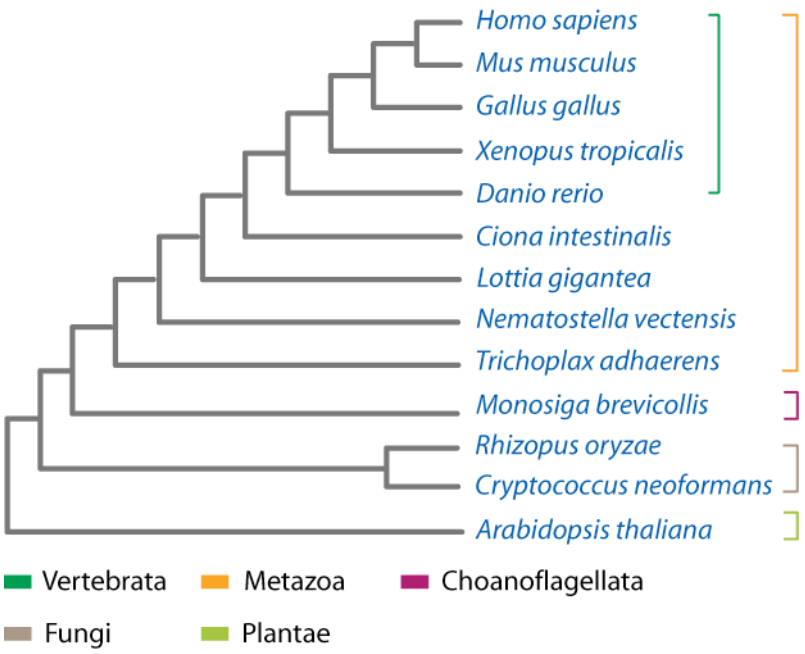


Figura 8. Espécies utilizadas e seus respectivos grupos taxonômicos. Mais detalhes sobre as espécies e a obtenção dos arquivos de anotação genômica e seqüências protéicas podem ser vistos no no anexo B.

3.2. Determinação das posições e fases de introns

As posições dos introns foram derivadas dos arquivos de anotação genômica através de *scripts* de Perl. As posições foram mapeadas nas proteínas a partir das coordenadas das CDSs (*Coding Sequences*) especificadas para cada transcrito. As fases de introns foram determinadas de acordo com as suas posições nos códons. Na prática, se um intron estiver localizado na posição 120 em relação à seqüência codificante, significa que entre os códons que correspondem aos aminoácidos 39 e 40 (120/3) existe um intron de fase 0. Se estiver na posição 121, o intron está mapeado no aminoácido 40 em fase 1 (entre o primeiro e o segundo nucleotídeo do códon). Por fim, se o intron estiver na posição 122 em relação às coordenadas da CDS, significa que existe um intron que interrompe o códon do aminoácido 40 entre o segundo e o terceiro nucleotídeo (fase 2). Para cada espécie, foram determinadas as freqüências de fases de introns e o excesso de simetria de exons em relação ao esperado considerando somente a maior proteína de cada gene.

O excesso de simetria de exons para cada classe foi calculado como sendo: $|N.E - On|/N.E$. Onde N é o total de exons, E é a freqüência esperada para a classe, On é o número de exons observados para a classe. A freqüência esperada (E) é $Pi.Pj$, onde Pi é a freqüência de introns de fase i e Pj é a freqüência de introns de fase j.

3.3. Alinhamento das proteínas

Dado o conjunto de seqüências protéicas para cada espécie, foi realizado o alinhamento local entre todas as proteínas de um determinado proteoma utilizando o

programa Blastp (versão 2.2.17) (Altschul, *et al.*, 1997) admitindo-se valor de *e-value* menor que 10 e filtro de baixa complexidade ativado. Apenas as proteínas de maior tamanho para cada gene foram alinhadas. Com o objetivo de identificar regiões similares compartilhadas por proteínas sem relação de homologia, foram considerados apenas alinhamentos que possuem identidade igual ou superior à 30% e tamanho da região alinhada igual ou inferior à 40% do tamanho da menor proteína do par alinhado. O critério de identidade visa a obtenção de regiões conservadas (HSPs), conforme sugerido por outros autores (Brenner, *et al.*, 1998; Mewes, *et al.*, 1997). A restrição em relação ao tamanho da região alinhada é um procedimento que tem por objetivo evitar que alinhamentos entre proteínas parálogas sejam incluídos nas análises. Após a obtenção dos alinhamentos filtrados com os critérios acima, para cada espécie foram criados arquivos com os identificadores das proteínas alinhadas e coordenadas de início e fim dos HSPs.

3.4. Identificação de regiões possivelmente originadas por *exon shuffling*

Uma região protéica cuja origem possivelmente se deu através de um evento de *exon shuffling* deve apresentar similaridade com outras proteínas não homólogas e possuir sinal evidente de flanquemento por introns em torno de suas bordas. As regiões conservadas, identificadas através dos alinhamentos, foram verificadas quanto a presença de introns em torno de ambas as extremidades dos HSPs nas duas proteínas do par alinhado. Considerou-se o intervalo de 1 aminoácido em direção ao interior do HSP e 10 aminoácidos em direção à região externa ao HSP (ver figura 9). Este intervalo é semelhante ao utilizado por outros trabalhos (Liu e

Grigoriev, 2004; Liu *et al.*, 2005). Nos casos em que existe mais de um intron em torno da borda verificada, considerou-se aquele mais próximo em relação às extremidades de início ou fim do HSP.

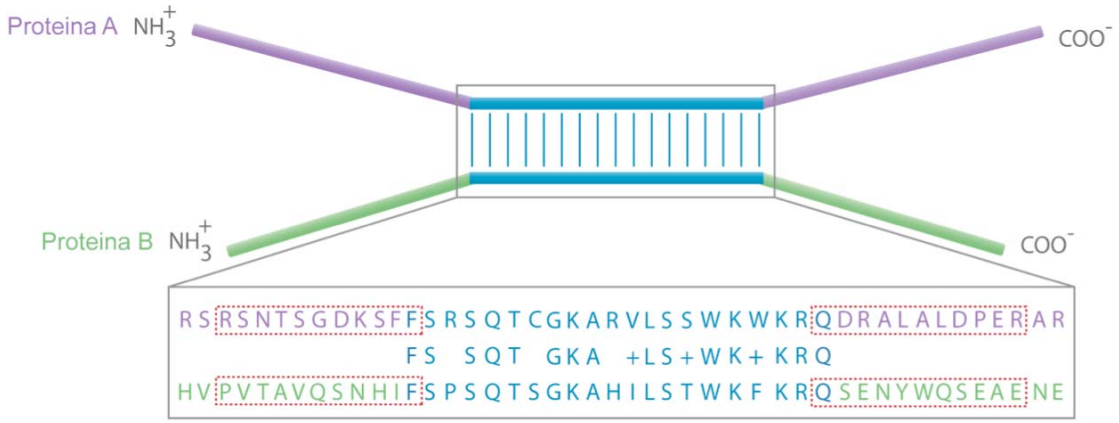


Figura 9. Esquema representativo da estratégia utilizada para a identificação de regiões possivelmente originadas por *exon shuffling*. A região azul corresponde à porção alinhada (HSP) com identidade igual ou maior que 30% entre duas proteínas não homólogas A e B. Foi verificada a presença de introns em torno das quatro bordas do HSP em um intervalo indicado pelas linhas vermelhas, sendo 10 aminoácidos em direção à região externa ao alinhamento e 1 aminoácido em direção à região interna ao alinhamento.

3.5. Identificação de domínios sobrepostos às regiões conservadas.

Para cada proteína de todas as espécies estudadas, foram mapeadas as posições dos domínios protéicos da base Pfam versão 22 (Finn *et al.*, 2008), identificados através do programa HMMER (Eddy, 1998). Foram considerados somente domínios que pertencem à divisão Pfam-A Exigiu-se um valor de *e-value* < 0,01 para a inclusão de determinada ocorrência de domínio.

Após a determinação das posições dos domínios, foi verificada a sobreposição dos domínios em relação às regiões conservadas e flanqueadas por introns. Foram considerados somente casos em que o domínio estivesse totalmente

sobreposto, ou seja, domínios inteiramente contidos na região compreendida pelos HSPs. A distribuição de simetria de fases de introns ao redor das regiões conservadas que contém e não contém domínios sobrepostos foi então calculada.

3.6. Identificação de domínios flanqueados por introns

As ocorrências de domínios Pfam-A mapeados nas proteínas foram verificadas quanto à presença de introns ao redor de suas bordas de início e fim. Os critérios para esta verificação foram os mesmos utilizados para os HSPs (ver seção 3.4), exigindo-se a presença de introns em ambas as extremidades do domínio.

3.7. Correção do sinal de flanqueamento por introns em torno de domínios devido ao efeito da paralogia

As ocorrências de domínios flanqueados por introns sofreram um controle devido ao fenômeno de duplicação gênica. Para tanto, foi necessário definir a relação de paralogia entre as proteínas. Para as espécies *H. sapiens*, *M. musculus*, *G. gallus*, *X. tropicallis* e *D. rerio*, foram utilizadas listas de proteínas parálogas obtidas através da ferramenta BioMart (Smedley *et al.*, 2009). Para as demais espécies, a relação de paralogia foi definida localmente. Duas proteínas foram consideradas parálogas quando o alinhamento através do Blastp apresentou *e-value* $< 10^{-6}$, identidade igual ou superior à 30% e comprimento da região alinhada igual ou superior à 70% do tamanho da maior proteína do par alinhado. Critérios semelhantes foram implementados em outros trabalhos (Makova e Li, 2003; Suyama *et al.*, 2006).

Após a determinação dos pares parálogos, as proteínas foram agrupadas em famílias através do método chamado *single linkage clustering*, que consiste em agrupar numa mesma família, proteínas que satisfaçam os critérios com algum dos membros já existentes de determinada família. Por exemplo, se as proteínas A e B são consideradas parálogas e a proteína C é paráloga de A ou B, as proteínas A, B e C são agrupadas na mesma família.

Para cada tipo de domínio, foi determinada uma freqüência de flanqueamento por introns para as nove combinações de fase possíveis. Para cada família de proteínas foi contado o número de ocorrências de um dado domínio e dividido pelo número total de membros da família que possui tal domínio. A soma das razões para todas as famílias representa o total ponderado de ocorrências do domínio. As ocorrências de domínios flanqueados foram determinadas da mesma forma, verificando o total de ocorrências de domínios flanqueados para cada família em relação ao número de membros da família que possui tal domínio. A soma das razões para todas as famílias representa o total ponderado de ocorrências flanqueadas. A freqüência de flanqueamento para um certo tipo de domínio foi definida como a razão entre o total ponderado de ocorrências flanqueadas e o total ponderado de ocorrências.

3.8. Construção do agrupamento hierárquico de domínios flanqueados por introns

As freqüências de domínios flanqueados por introns em suas diferentes combinações de fase, determinadas pelo método descrito acima, foram utilizadas para a construção de um agrupamento hierárquico. O agrupamento foi construído

com o pacote estatístico R (versão 2.7.1) (http://www.r-project.com) através do programa *hclust,* que utilizou distâncias euclidianas. Foram considerados apenas domínios com mais de duas ocorrências flanqueadas por introns e além disso, apenas domínios que apresentaram mais de 10% de suas ocorrências sendo flanqueadas por introns foram considerados. O sinal de flanqueamento para cada domínio foi considerado como sendo a porcentagem de ocorrências de determinado tipo de simetria em relação ao total de ocorrências flanqueadas.

3.9. Análise de enriquecimento de termos do *Gene Ontology*

O conjunto de genes que possui regiões conservadas e flanqueadas por introns foi utilizado para a análise de enriquecimento de termos do *Gene Ontology* (G.O) (The Gene Ontology Consortium, 2000).

A anotação dos termos foi obtida para *Homo sapiens* através da ferramenta BioMart (Smedley *et al.*, 2009), para as espécies *N. vectensis*, *T. adhaerens* (http://www.jgi.doe.gov), e para *A. thaliana* (ftp://ftp.arabidopsis.org/home/tair/Ontologies/).

A análise de enriquecimento foi feita utilizando o programa Cytoscape (Shannon *et al.*, 2003) e a ferramenta BINGO (Maere *et al.,* 2005). O conjunto de genes envolvidos em *exon shuffling* foi comparado com o conjunto total de genes que possuem termos anotados através do teste exato de Fischer com nível de significância $p < 0,05$.

4. Resultados

4.1. Distribuição de fases de introns

As posições e fases dos introns foram mapeadas nos genes codificantes de proteínas nas diferentes espécies estudadas. Conforme demonstrado por Fedorov *et al.* (1992) e Long *et al.* (1995), as freqüências para as três fases diferem entre si. Introns de fase 0 são mais abundantes que introns de fase 1 e estes mais abundantes que introns de fase 2 em todas as espécies analisadas (figura 10).

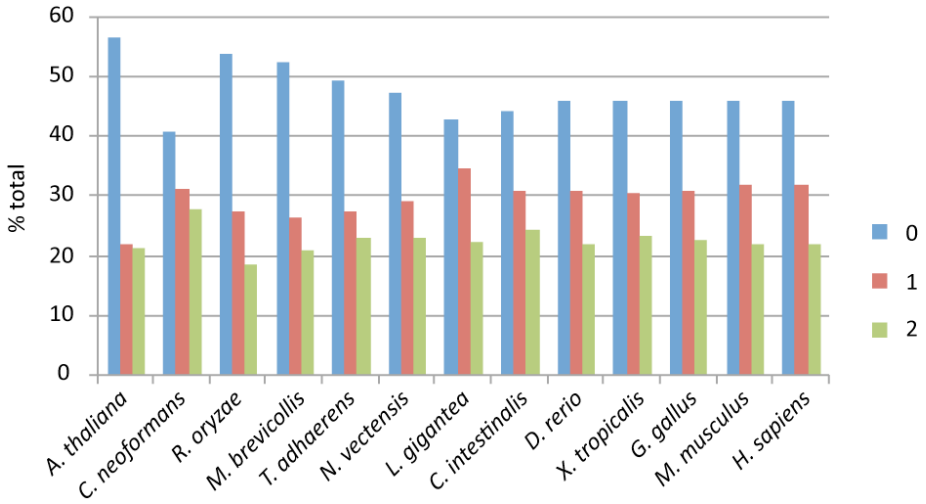


Figura 10. Distribuição de fases de introns em espécies eucarióticas.

Todas as freqüências obtidas foram similares àquelas encontradas por Nguyen *et al.* (2006). As espécies com as maiores freqüencias de introns de fase 0 são *A. thaliana* (56,5%), *R. oryzae* e *Monosiga brevicollis* (52,6%). A tendência que se observa é metazoários apresentarem menor freqüência de introns de fase 0 ($\overline{X}$ = 46%) em relação aos não metazoárias (*A. thaliana*, *R. oryzae* e *M. brevicollis*). Por outro lado, a freqüência de introns de fase 1 é maior nos metazoários, com média de

31%, enquanto em *A. thaliana, R. oryzae* e *M. brevicollis* as freqüências são de 21,9% e 27,6% e 26,4% respectivamente. As freqüências de introns de fase 2 são semelhantes em todas as espécies analisadas ($\overline{X}$ = 23%), com valores muito próximos em plantas (21,5%) e humanos (21,9%). Em fungos, a densidade de introns é bastante variável. *S. cerevisiae* por exemplo, possui média de 1 a 2 introns por gene, enquanto *C. neorformans* apresenta em média 7 introns por gene (Nielsen *et al*., 2004). É possível que as diferenças nas freqüências de fases entre *C. neoformans* e *R. oryzae* sejam conseqüência da dinâmica de ganho e perda de introns linhagem-específicas.

4.2. Simetria de exons

O excesso de exons simétricos é interpretado como evidência de *exon shuffling* (Long *et al*., 1995). Dessa forma, as freqüências de exons simétricos e assimétricos foram determinadas e seu excesso foi calculado em relação ao esperado (figura 11). Nota-se que os excessos são observados predominantemente para as classes simétricas (0-0, 1-1 e 2-2), sendo que a classe 1-1 apresenta as maiores porcentagens.

Os excessos para as três classes de exons simétricos são evidentes em todas as espécies de metazoários, com médias de 10%, 16,8% e 7,7% para as classes 0-0, 1-1 e 2-2 respectivamente. Esses valores são consistentes com o trabalho de Long *et al*. (1995). *M. brevicollis* e *C. neoformans* apresentam um excesso marginal, de 2,8% e 2,7% para classe 0-0, 1,2% e 2,2% para classe 1-1 e -0,4% e 4,6% para a classe 2-2 respectivamente. *R. oryzae* apresenta um padrão anômalo em relação as demais espécies, com excessos de classes assimétricas.

Novamente, a dinâmica de introns linhagem-específica pode ser o fator responsável por esta diferença. Ao contrário de fungos e *M. brevicollis*, a planta *A. thaliana* apresentou excesso comparável aos metazoários para as três classes simétricas: 9,2% para exons 0-0, 9,3% para 1-1 e 13% para classe 2-2. Contudo, estes excessos são evidências indiretas e não representam um estimativa real do fenômeno de *exon shuffling*.

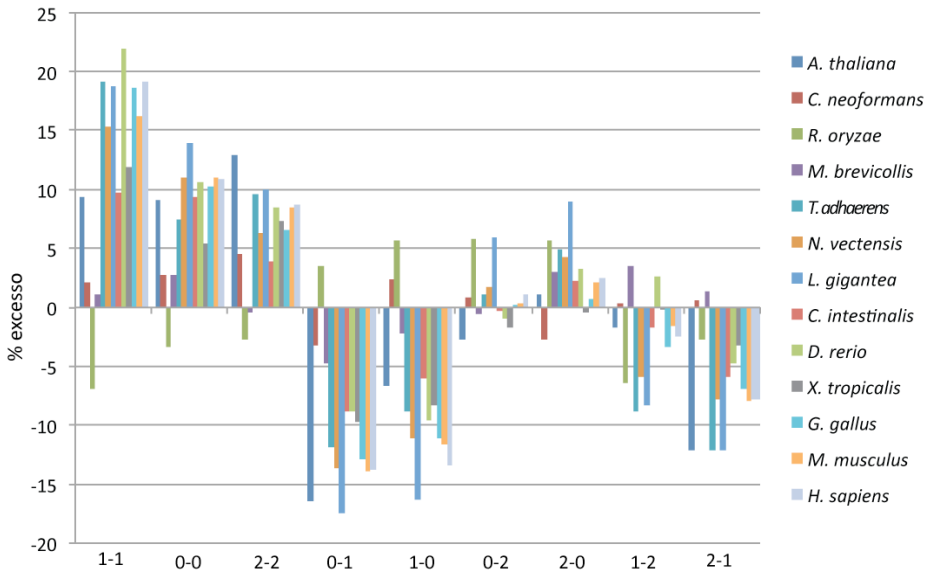


Figura 11. Simetria de exons. Excessos de exons simétricos e assimétricos em relação à freqüência esperada.

4.3. *Exon shuffling* em eucariotos

4.3.1. Geração de um catálogo de regiões protéicas originadas por exon shuffling

Para se obter evidências diretas de *exon shuffling*, é necessário mostrar que: I) duas regiões conservadas estão presentes em proteínas não homólogas e II) a aquisição das regiões conservadas foram mediadas por recombinações intrônicas,

portanto, a presença de introns em torno das regiões conservadas são indicativas deste processo (Patthy, 1999b). Com base nessas duas premissas, foi gerado um catálogo de regiões protéicas possivelmente originadas por *exon shuffling* através do alinhamento de proteínas e a verificação do flanqueamento por introns em torno das bordas das regiões conservadas (ver critérios na seção 3.4 e representação esquemática da estratégia na figura 9). A tabela 1 mostra os números obtidos em cada espécie analisada.

Tabela 1. Número total de genes e regiões conservadas. Estão sumarizados o número total de genes preditos para cada espécie, o número de genes que possuem regiões com evidência de *exon shuffling* (ES), a porcentagem de genes com evidência de ES e o número total de regiões conservadas e flanqueadas por introns (HSPs).

Espécies	Genes	Genes em ES	Genes em ES (%)	HSPs
H. sapiens	23.943	1.546	6,4	2.884
M. musculus	24.496	1.037	4,2	1.610
G. gallus	16.736	899	5,4	1.144
X. tropicalis	18.025	790	4,4	976
D. rerio	21.322	911	4,3	1.139
C. intestinalis	14.180	442	3,2	386
L. gigantea	23.851	764	3,2	956
N. vectensis	27.273	818	3,0	984
T. adhaerens	11.520	483	4,2	349
M. brevicollis	9.196	233	2,5	192
R. oryzae	17.467	148	0,8	110
C. neoformans	6.475	23	0,3	19
A. thaliana	26.814	272	1,0	185

Enquanto 6,4% dos genes humanos possui evidência de ao menos um evento de *exon shuffling*, este valor é extremamente inferior em plantas (1%), fungos (0,3% e 0,8%) e coanoflagelados (2,5%). Os números mostram que a freqüência de eventos de *exon shuffling* é maior em metazoários. O número de HSPs em *T. adhaerens*, o metazoário mais primitivo analisado, é cerca de duas vezes maior que em *A. thaliana* (349/185). Em *H. sapiens* este valor é aproximadamente 15 vezes maior (2.884/185).

É possível que o número de genes e regiões com evidência de *exon shuffling* estejam subestimados em razão da restringência dos critérios utilizados. Isso desfavorece a identificação de eventos principalmente nas espécies mais primitivas, em virtude da menor conservação do sinal da presença de introns em torno das das regiões conservadas.

4.3.2. Padrões de simetria de "exon shuffling" em eucariotos.

O sucesso dos eventos de *exon shuffling* depende das compatibilidades de fases de introns. Espera-se que exons ou grupos de exons simétricos sejam inseridos com maior probabilidade de êxito por não alterarem o quadro de leitura do gene. Por isso, as regiões conservadas foram analisadas quanto à distribuição de simetria de fases dos introns flanqueadores (figura 12). Cerca de 60% dos casos correspondem a unidades simétricas, indicando que a estratégia utilizada foi adequada para a identificação das regiões. O padrão obtido mostra que a classe 0-0 é predominante em não metazoários (*M. brevicollis* e *A. thaliana*), enquanto nos metazoários a classe predominante é 1-1 ($\overline{X}$ = 41%). *C. neoformans* não apresentou número suficiente de casos para a análise. A maioria dos casos em *R. oryzae* são assimétricos, representando casos menos confiáveis.

Nota-se que as freqüências de simetria no cnidário *N. vectensis* são equivalentes às observadas em *H. sapiens,* enquanto *T. adhaerens* apresenta um padrão intermediário, com 23% dos casos simétricos de fase 0 e 25% simétricos de fase 1.

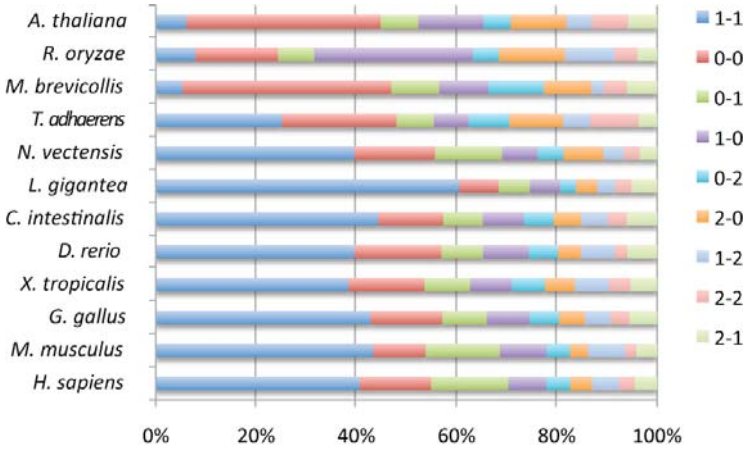


Figura 12. Freqüências de simetria de fases de introns em regiões conservadas. Metazoários apresentam maior freqüência de eventos que envolvem unidades 1-1. Regiões flanqueadas por introns em espécies não metazoárias (*M. brevicollis* e *A. thaliana*) são predominantemente simétricas de fase 0.

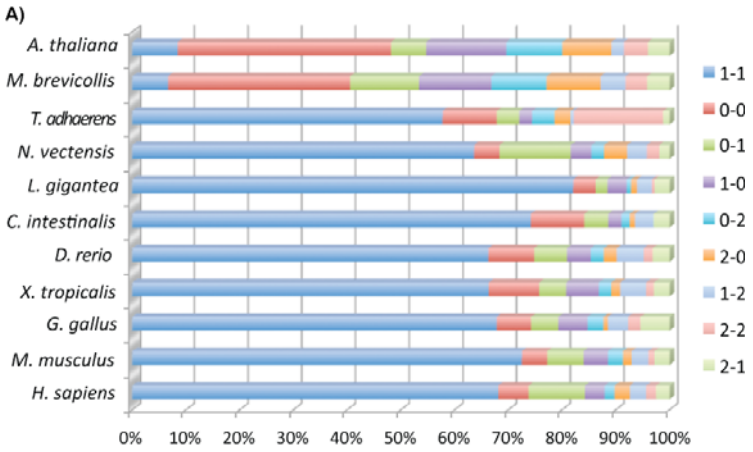
4.3.3. Padrões de simetria de "exon shuffling" que envolvem domínios

As regiões envolvidas em *exon shuffling* foram analisadas quanto à sobreposição com domínios protéicos. As ocorrências de domínios Pfam-A foram mapeadas nas proteínas e posteriormente identificados os casos em que houve total sobreposição com as regiões flanqueadas por introns. Em metazoários, aproximadamente 43% das regiões identificadas apresentaram domínios sobrepostos. Em *M. brevicollis*, 38% (N = 107) das regiões conservadas contiveram domínios, em *A. thaliana* este valor foi de apenas 11% (N = 32) .Em fungos não foram encontrados casos de regiões sobrepostas com os critérios utilizados.

As regiões que apresentaram sobreposição de domínios foram verificadas quanto ao seu padrão de simetria de fases de introns. A freqüência de unidades simétricas 1-1 em metazoários foi ainda maior nos casos em que houve sobreposição, perfazendo em média 68% das ocorrências em Metazoa (figura 13A).

Em *A. thaliana*, *M. brevicollis* e fungos, poucos casos de regiões sobrepostas foram identificados, impossibilitando a determinação dos padrões de simetria. Para contornar esse problema, o procedimento de identificação de eventos de *exon shuffling* foi refeito utilizando critérios mais relaxados, exigindo-se a presença de introns em três das quatro bordas dos HSPs. Confirmando os resultados obtidos anteriormente, a classe 0-0 foi predominante em plantas (39%) e coanoflagelados (34%) (figura 13A). Em fungos, nenhum caso de sobreposição foi encontrado.

As regiões não totalmente sobrepostas a domínios apresentaram maior freqüência de classes assimétricas, sugerindo que este conjunto de dados pode estar enriquecido com eventos falso-positivos (figura 13B). Ainda assim, as classes simétricas 1-1 e 0-0 constituem cerca de 40% dos casos. É possivel que essas regiões correspondam a domínios identificados por outros métodos ou domínios ainda não caracterizados.



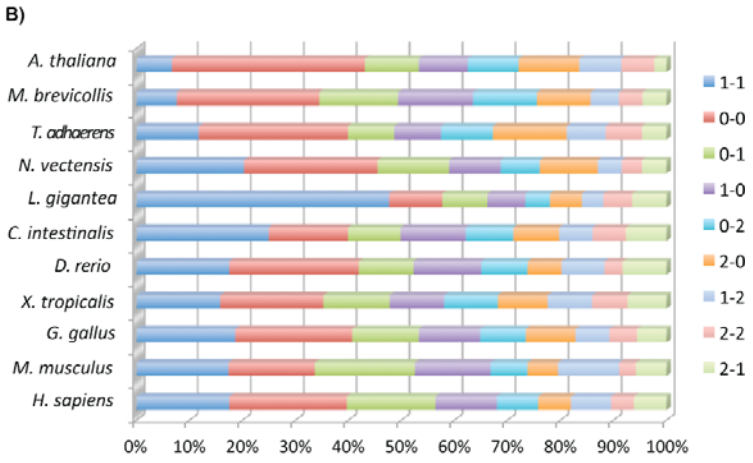


Figura 13. Padrão de simetria de fases de introns em regiões conservadas com e sem sobreposição de domínios. A) Freqüências de simetria em regiões que contém domínios. A presença de domínios está relacionada com a maior freqüência de unidades simétricas 1-1 em metazoários. Nas espécies não metazoárias, as regiões são preferencialmente simétricas de fase 0. **B)** Freqüências de simetria em regiões que não contém domínios. Nestas regiões há predomínio de classes assimétricas. *Os dados de *A. thaliana* e *M. brevicollis* foram gerados exigindo-se a presença de introns em três das quatro bordas dos HSPs.

4.4. Domínios flanqueados por introns

4.4.1. Freqüências de domínios simétricos em eucariotos

Além das evidências de *exon shuffling* obtidas a partir do alinhamento entre proteínas, uma outra abordagem utilizada para evidenciar este fenômeno foi a verificação da presença de introns em torno das bordas de domínios protéicos. Um passo importante nesta análise foi a tentativa de minimizar o efeito da paralogia na contagem das ocorrências de domínios. Os critérios estão detalhados nas seções 3.6 e 3.7.

As freqüências de domínios simétricos confirmam os dados anteriores. A classe 0-0 é predominante em não metazoários (*A. thaliana* e *M. brevicollis*), enquanto a transição para os metazoários está associada à expansão de domínios

1-1. As baixas freqüências de 2-2 indicam que *exon shuffling* envolvendo esta classe de simetria é extremamente raro.

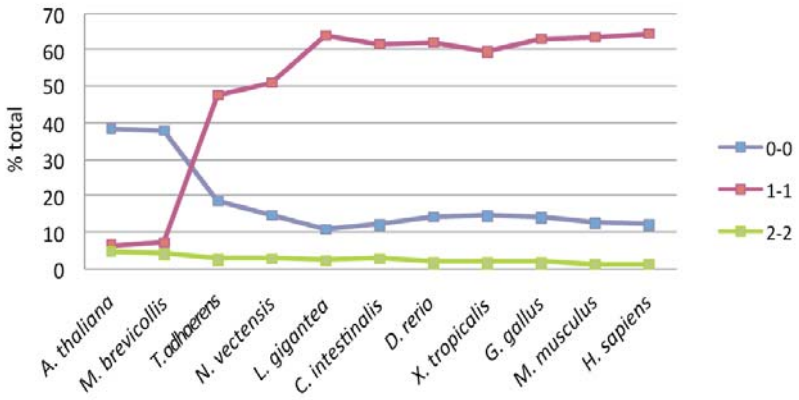


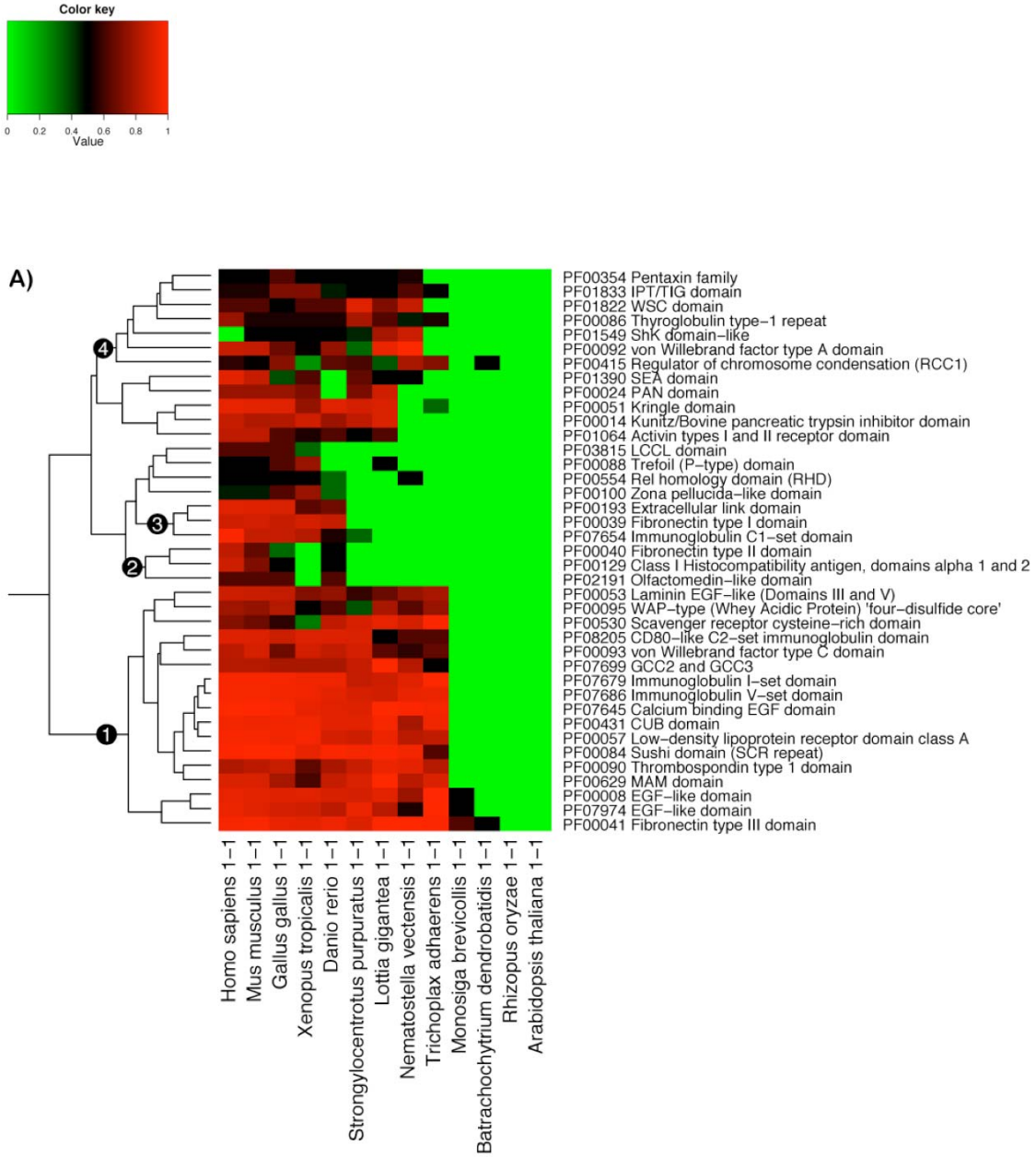
Figura 14. Freqüências de ocorrências de domínios simétricos. Domínios 0-0 são predominantes em não metazoários. O aumento na freqüência de 1-1 está associado à emergência de Metazoa. As freqüências de ocorrências simétricas foram calculadas em relação ao total de ocorrências de domínios flanqueados por introns.

4.4.2. Agrupamento hierárquico dos domínios flanqueados por introns

Os domínios protéicos e suas respectivas classes de simetria foram caracterizadas através da construção de um agrupamento hierárquico, usando as freqüências de flanqueamento por introns para cada domínio nas diferentes espécies. Para simplificar a apresentação dos resultados, apenas os agrupamentos de domínios 0-0 e 1-1 serão mostrados. O agrupamento completo e o mapa de calor contendo todos os domínios identificados podem ser vistos no anexo C.

O agrupamento hierárquico permitiu analisar a evolução dos domínios em termos de sua expansão por *exon shuffling*. Alguns resultados importantes foram obtidos em relação aos domínios 1-1: I) a maioria dos domínios (ramos 1 e 4 da figura 15A) compartilha o sinal de flanqueamento por introns em todos os metazoários. II) Os domínios “EGF-like”, “Fibronectin type III” e “RCC1”, apesar do sinal reduzido, possuem ocorrências flanqueadas em coanoflagelados e fungos. III)

Os domínios dos ramos 2 e 3 (figura 15A) representam os casos mais confiáveis de expansão específica em vertebrados. IV) Os domínios da figura 15B apresentam sinal de flanqueamento em deuterostômios. V) Os domínios “Class II Histocompatibility antigen, alpha/beta” e “Repeat of unknown function (DUF1220)” (figura 15B) apresentam forte sinal de flanqueamento em *H. sapiens*, sugerindo expansão extremamente recente, possivelmente após a divergência dos primatas.



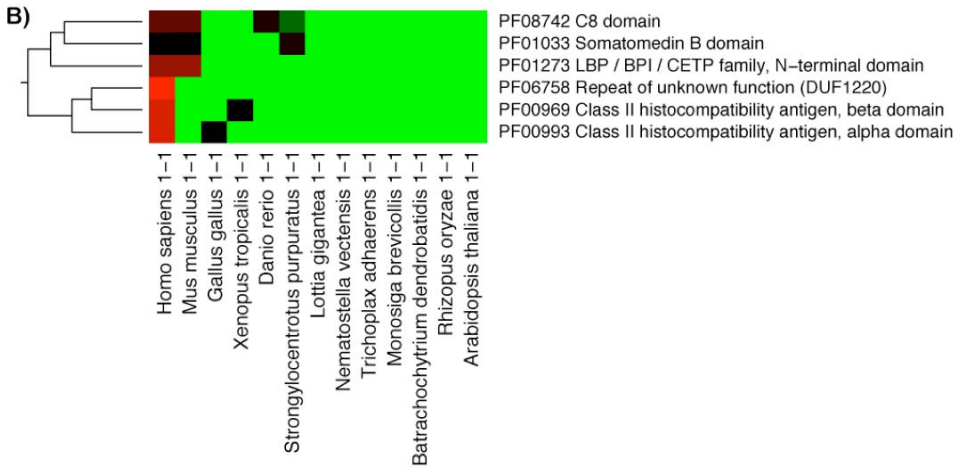


Figura 15. Mapa de calor e agrupamento hierárquico de domínios simétricos 1-1. A) Grupo de domínios flanqueados em metazoários. **B)** Grupo de domínios flanqueados principalmente em deuterostômios. Como indicado pela legenda do mapa, as regiões em vermelho e verde correspondem respectivamente, a frações elevadas e reduzidas de flanqueamento de domínios pro introns. O equinodermo *Strongylocentrotus purpuratus* foi utilizado em substituição a *Ciona intestinalis*, devido ao fenômeno de perda de introns sofrido pelo urocordado (Putnam *et al.*, 2007). O fungo *Batrachochytrium dendrobatis* também foi adicionado às análises.

Para os domínios tipicamente 0-0 (figura 16), algumas considerações podem ser feitas: I) Coanoflagelados e fungos tendem a apresentar maior diversidade de domínios 0-0 quando comparados com a classe 1-1, reforçando a idéia de que o embaralhamento dessa classe de domínios é mais antigo, anterior à divergência dos animais. II) Os domínios “Collagen triple helix”, “7 transmembrane receptor” e “Latrophilin/C1 like GPS” são flanqueados por introns apenas em metazoários. III) O domínio “Calpain inhibitor” é o caso mais convincente de expansão em vertebrados. IV) O domínio “Beta/Gamma crystallin”, apesar do sinal reduzido de flanqueamento, também foi identificado por Kaessmann *et al.* (2002) em análises de *domain shuffling*.

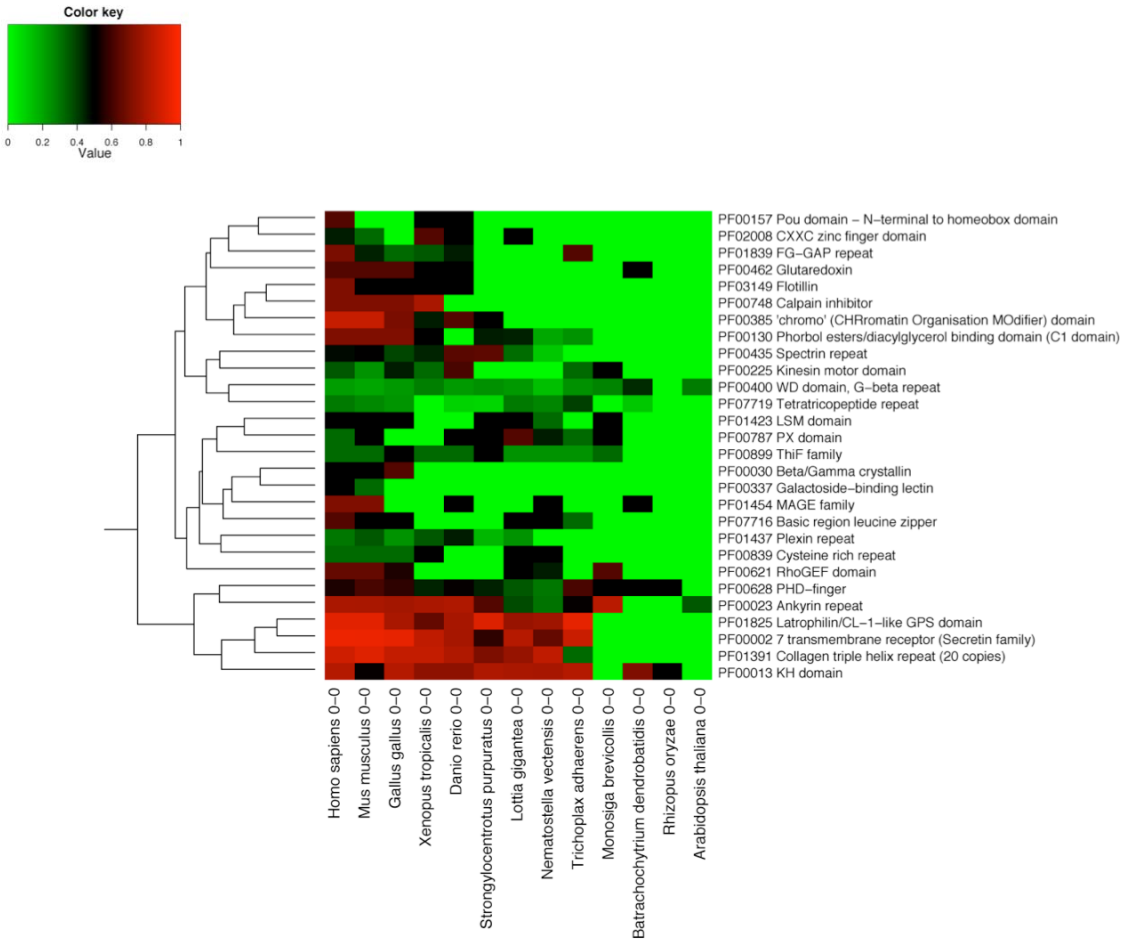


Figura 16. Mapa de calor e agrupamento hierárquico de domínios simétricos 0-0.

Um grupo de três domínios simétricos 2-2 foi identificado, dos quais o “Peptide hormone” é o caso mais convincente, encontrado apenas em vertebrados (anexo C).

Domínios assimétricos também foram identificados (ver anexo C), incluindo as classes 0-1 (*e.g* “KRAB box”), 1-0 (*e.g* “Calponin homology”) e 2-1 (*e.g* “SH2”). O mecanismo de expansão desses domínios não está claro e deve ser investigado com mais profundidade.

4.5. Categorias do *Gene Ontology* enriquecidas em genes envolvidos em *exon shuffling*

Com o intuito de investigar as implicações funcionais relacionadas ao fenômeno de *exon shuffling*, o conjunto de genes que possuem regiões conservadas e flanqueadas por introns foi utilizado para a determinação de quais termos do *Gene Ontology* (The Gene Ontology Consortium, 2000) estão enriquecidos. Como parâmetro de comparação, o conjunto total de genes com termos anotados foi utilizado. A análise foi feita para as espécies *H. sapiens*, *N. vectensis*, *T.adhaerens* e *A. thaliana*. Duas abordagens foram conduzidas nesta etapa. Na primeira análise foram considerados os termos do “GO Slim generic”, que compreende um subconjunto de termos gerais do *Gene Ontology*, fornecendo uma visão ampla e geral da ontologia dos termos enriquecidos. A figura 17 mostra os resultados obtidos para *A. thaliana*. Pode-se concluir que os genes que apresentam evidência de *exon shuffling* em plantas exercem sua atividade predominantemente no núcleo. As funções moleculares mais evidentes são a ativação de fatores de transcrição e atividade de proteína quinase, um dos componentes chave na transdução de sinal e comunicação celular.

Assim como em *A. thaliana*, termos referentes à ligação ao DNA e proteínas, transcrição e atividade de proteína quinase estão sobre-representados em *H. sapiens*. Entretanto, a grande inovação tem a ver com o papel de *exon shuffling* na construção de proteínas da matriz extracelular, proteínas localizadas na membrana plasmática, morfogênese e processos relacionados aos animais multicelulares (figura 18).

Figura 17. Termos enriquecidos do *Gene Ontology* em *A. thaliana*. O tamanho dos círculos é proporcional ao número de genes anotados com determinado termo. De acordo com a legenda, a coloração indica valores de *p*. Os dados foram gerados exigindo-se a presença de introns em três das quatro bordas dos HSPs. A análise foi feita usando os termos do "G.O slim generic".

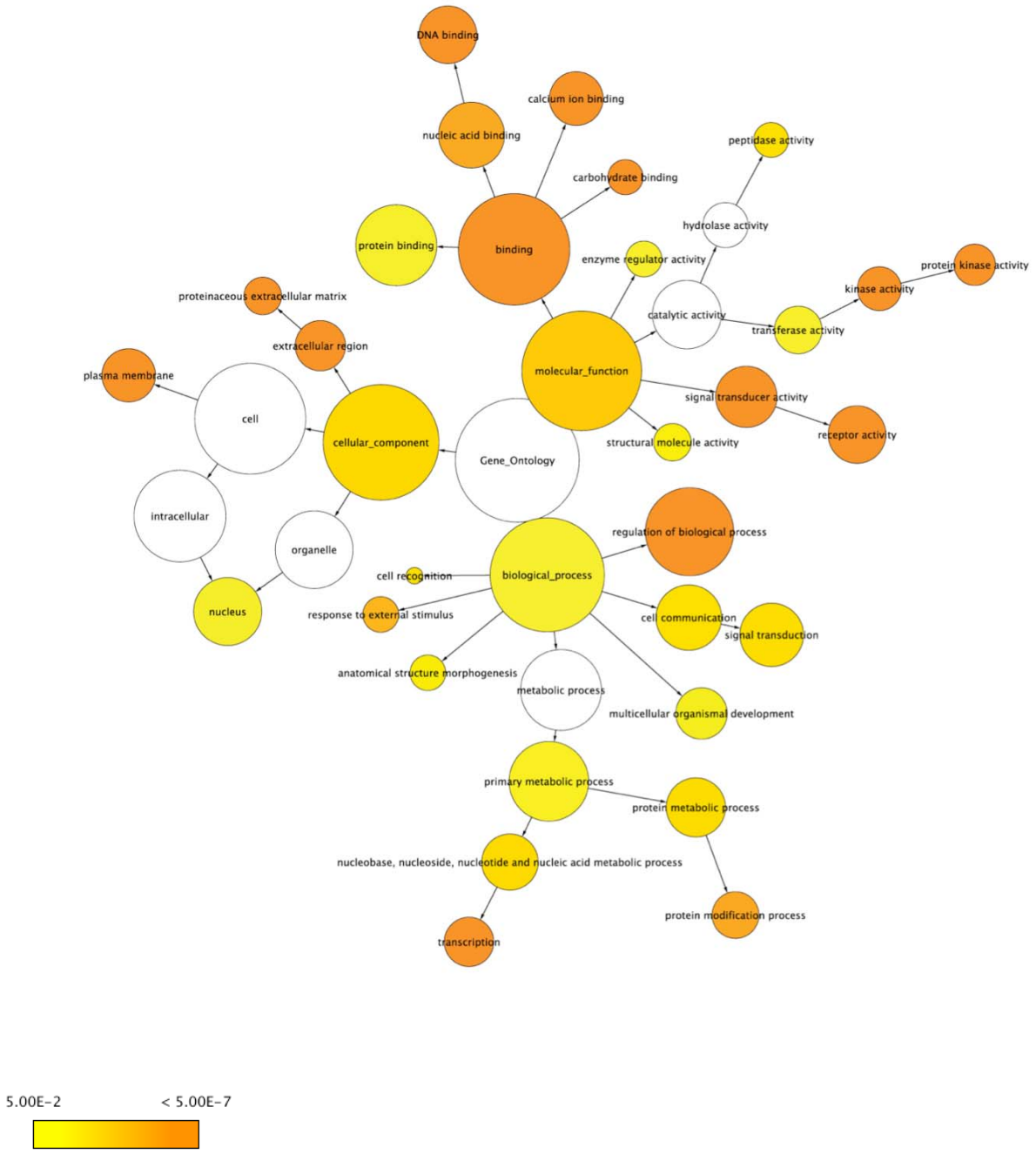


Figura 18. Termos enriquecidos do *Gene Ontology* em *H. sapiens.* O tamanho dos círculos é proporcional ao número de genes anotados com determinado termo. De acordo com a legenda, as cores indicam valores de *p*. A análise foi feita usando os termos do “G.O slim generic”.

A segunda abordagem procurou investigar o enriquecimento de termos específicos em metazoários. Para isso, todo o conjunto de termos do *Gene Ontology* foi considerado. Devido a complexa relação hierárquica entre os termos,

principalmente em humanos, no quadro 1 estão listados apenas alguns dos termos significativamente enriquecidos. As relações entre todos os termos sobre-representados podem ser vistas nos anexos D, E e F, correspondendo às três categorias principais do G.O: processo biológico, função molecular e componente celular, respectivamente.

	Categoria G.O	Termo	*p-value*
H. sapiens	Função molecular	calcium ion binding	6,00E-45
		protein tyrosine kinase activity	2,00E-21
		extracellular matrix structural constituent	2,00E-20
		zinc ion binding	2,00E-19
		vascular endothelial growth factor receptor activity	1,00E-13
		MHC class II receptor activity	2,50E-10
		scavenger receptor activity	1,60E-06
		brain specific angiogenesis inhibitor activity	3,00E-06
		low density lipoprotein binding	8,00E-05
	Processo biológico	cell adhesion	1,00E-40
		phosphate transport	5,60E-23
		antigen processing and presentation	6,00E-11
		blood coagulation	3,00E-10
		membrane invagination	7,00E-09
		immune response	1,00E-06
		complement activation	1,00E-06
		axon guidance	3,00E-03
		neuropeptide signaling pathway	8,00E-03
	Componente celular	extracellular matrix	4,00E-31
		cellular component collagen	4,00E-18
		basement membrane	5,50E-06
N. vectensis	Função molecular	calcium ion binding	2,00E-51
		kinase activity	2,70E-02
	Processo biológico	phosphate transport	1,00E-07
		cell adhesion	8,00E-04
	Componente celular	extracellular matrix	3,00E-02
		cellular component collagen	3,00E-02
T. adhaerens	Função molecular	calcium ion binding	2,00E-09
		transmembrane receptor activity	6,00E-07
		scavenger receptor activity	9,00E-05
	Processo biológico	neuropeptide signaling pathway	2,00E-06
		cell adhesion	1,00E-05
	Componente celular	membrane	7,00E-03

Quadro 1. Termos enriquecidos do *Gene Ontology* em metazoários.

Como já foi observado por Patthy (1999b), *exon shuffling* está intrinsecamente associado ao surgimento de proteínas de matriz extracelular e adesão celular. Aqui fica evidente que esse processo teve início num período muito primitivo da história evolutiva dos metazoários, como pode ser visto pelo enriquecimento em “collagen”, “cell adhesion” e “extracellular matrix” em *T. adhaerens* e *N. vectensis*. Em humanos, a análise revela que a expansão de eventos de *exon shuffling*, principalmente de unidades simétricas 1-1, está associada com o aumento de complexidade em metazoários derivados. Isso se reflete na aquisição de genes que participam da coagulação sangüínea, ativação do complemento e sistema imune, sistema nervoso assim como diversos outros processos e funções (ver anexos D, E e F).

5. Discussão

5.1. Considerações sobre fases de introns e excesso de exons simétricos

5.1.1. Fases de introns

As freqüências de fases de introns foram obtidas com a intenção de verificar a correspondência com os dados da literatura. Os dados foram muito próximos aos observados por outros autores (Long *et al.*, 1995; Nguyen *et al.*, 2006). A não uniformidade das freqüências para as três fases é conhecida desde o trabalho de Fedorov e colaboradores (1992). Este fato têm sido interpretado de duas maneiras, de acordo com as teorias de evolução dos introns.

Uma das interpretações para o excesso de introns de fase 0 tem relação com a contribuição de eventos de *exon shuffling* para o desvio da proporção esperada de 0,33 para cada uma das fases (Fedorov *et al.*, 1992). Esta hipótese fundamenta-se na origem antiga dos introns proposta pela teoria *introns early*. Os genes primitivos codificariam para pequenos peptídeos independentes e novos genes poderiam ser formados através do embaralhamento destas unidades, facilitado pela recombinação de introns primordiais (Gilbert, 1987). A principal evidência para este argumento é a existência da correlação entre posições de introns e módulos protéicos em proteínas antigas (compartilhadas com procariotos) (De Souza *et al.*, 1996). Esta correlação mostrou-se válida somente para introns de fase 0, sugerindo que o excesso observado é decorrente de uma população antiga de introns, predominantemente de fase 0, que intermediaram eventos de *exon shuffling* antes da divergência dos

eucariotos. A outra parcela dos introns (~60%) teria sido adquirida após a divergência dos eucariotos (De Souza *et al.*, 1998).

A explicação alternativa para a predominância de introns de fase 0 trabalha com a hipótese de que os introns são preferencialmente inseridos em sítios específicos. Dibb e Newman (1989) observaram que os introns ganhos em genes das famílias das *actinas* e *tubulinas* foram inseridos preferencialmente em seqüências consenso (A/C)AG|GT ("I" representa o sítio de inserção), denominada de "proto-splice sites". Na verdade esta seqüência é praticamente idêntica às seqüências exônicas adjacentes aos sítios de *splicing*.

O trabalho de Sverdlov *et al.* (2004) testou a hipótese desta seqüência consenso ter evoluído por convergência. Utilizando um conjunto de códons conservados em proteínas ortólogas de eucariotos, assumiu-se que estas regiões possivelmente estão sob forte seleção purificadora em nível protéico, e portanto são refratárias à mutações que selecionam para uma maior eficiência do sinal de *splicing*. A análise dos nucleotídeos em torno dos introns que interrompem os códons conservados revelou a existência do "proto-splice site", reforçando a idéia de que os introns foram inicialmente inseridos em sítios específicos. Através de um modelo estatístico que prevê eventos de ganhos e perdas de introns, Qiu e colaboradores (2004) concluíram que a distribuição de fases de introns recentemente ganhos são predominantemente de fase 0, e que a distribuição de fases obtida pelo modelo não difere significativamente da distribuição observada na natureza. O trabalho de Qiu *et al.* também mostrou que as inserções de introns ocorrem preferencialmente nos "proto-splice sites". Resultados semelhantes foram obtidos por Nguyen *et al.* (2006).

As diferenças nas freqüências de fases de introns estão sujeitas a diferentes fenômenos. Duplicações gênicas e de exons, *exon shuffling*, a inserção de introns em sítios preferenciais, a dinâmica de perda e ganho de introns e a possível presença de introns no progenoto parecem todos ter influência na distribuição de fases de introns.

5.1.2. Excesso de exons simétricos

Considerando os mecanismos propostos, eventos de *exon shuffling* tendem a ocorrem com maior sucesso quando há o envolvimento de unidades simétricas inseridas em introns de mesma fase, de tal modo que o quadro de leitura do gene não seja alterado. Long *et al*. (1995) utilizaram um conjunto de 13.042 exons de diversas espécies e obtiveram as freqüências para as diferentes combinações de fases. O estudo mostrou que exons ou grupos de até seis exons simétricos ocorrem com excesso estatisticamente significativo em relação às freqüências esperadas para as três classes de simetria (0-0, 1-1 e 2-2), sendo que a classe 1-1 apresenta a maior quantidade de excesso. Os autores atribuem este excesso de exons simétricos como uma manifestação do fenômeno de *exon shuffling*.

A figura 11 confirma os resultados obtidos por Long *et al.* (1995). As três classes de simetria foram encontradas em excesso em todas as espécies de metazoários estudadas. Curiosamente, a planta *A. thaliana* também apresentou excesso de exons simétricos, inclusive para a classe 1-1. Entretanto, o excesso de exons simétricos é uma evidência indireta *exon shuffling*. Todos os dados provenientes de evidências diretas (regiões conservadas e domínios flanqueados por introns), não apontam para a ocorrência de *shuffling* extensivo 1-1 em plantas e

2-2 em todas as espécies. Por isso, o excesso de unidades simétricas não deve ser interpretado como conseqüência exclusiva de *exon shuffling*. É possível que uma parte deste excesso seja explicado por duplicações de genes e exons. No genoma humano por exemplo, estima-se que pelo menos 6% dos exons são duplicados e que este fenômeno interfere nas distribuições de fases de introns e simetria de exons (Fedorov *et al.*, 1998).

5.2. *Exon shuffling* em eucariotos

Uma das questões centrais do processo evolutivo é entender como ocorre o surgimento de novas estruturas e funções. A evolução opera essencialmente através da utilização ou combinação de formas pré-existentes, resultando eventualmente na aquisição de características e sistemas mais elaborados (Jacob, 1977). O reuso de exons e domínios é de extrema relevância para o surgimento de novidades evolutivas. Neste sentido, foi gerarado um catálogo de regiões protéicas provavelmente envolvidas em *exon shuffling* em 13 espécies eucarióticas. Como resultado de uma busca extensiva, as freqüências de simetria de domínios e regiões conservadas mostraram que eventos de *exon shuffling* mais antigos são predominantemente do tipo 0-0. Esta observação foi particularmente mais evidente para *A. thaliana* e *M. brevicollis*. Por outro lado, o número de genes e regiões que apresentaram evidências diretas de *exon shuffling* indica que a construção de proteínas através deste mecanismo se deu com maior intensidade a partir da emergência dos metazoários, havendo uma clara expansão de regiões simétricas 1-1. Esta idéia teve início com os trabalhos de Patthy, que identificou inúmeras proteínas animais originadas por este mecanismo (Patthy 1985, Patthy, 1996, Patthy, 1999b).

Os metazoários formam um grupo monofilético que compartilha um ancestral comum com Choanoflagellata. Análises filogenéticas de genes nucleares, genomas mitocondriais e genoma nuclear apontam os coanoflagelados (e.g *M. brevicollis*) como o grupo mais próximo dos metazoários (Burger *et al.*, 2003; Lang *et al.*, 2002; King *et al.*, 2008). *Trichoplax adhaerens* é a única espécie reconhecida do filo Placozoa. Estudos recentes têm demonstrado que Placozoa é um verdadeiro metazoário ('eumetazoa') que divergiu anteriormente aos cnidários (e.g *N. vectensis*), enquanto as esponjas representam o grupo mais primitivo dos animais (Ender e Schierwater, 2003; Dellaporta *et al.*, 2006; Srivastava *et al.*, 2008).

Em que momento da história evolutiva dos animais se deu o início desta expansão? Essa pergunta pôde ser respondida de forma mais precisa com este trabalho. O padrão de simetria de regiões conservadas e de domínios protéicos revelou que o extensivo embaralhamento de exons e domínios simétricos de fase 1 teve início no período entre a divergência de coanoflagelados e placozoários. As freqüências balanceadas de 0-0 e 1-1 mostraram que *T. adhaerens* é um representante dessa transição e reforça de forma contundente a idéia proposta por Patthy (1999b), de que a passagem para a multicelularidade animal foi acompanhada da expansão de unidades 1-1.

Kaessmann e colaboradores (2002) verificaram que domínios simétricos de fase 0 estão sobre-representados entre os domínios antigos (compartilhados entre procariotos e eucariotos). Ao estudar as preferências de localização de domínios antigos 0-0 e modernos 1-1 (presentes somente em eucariotos), Vibranovski *et al.* (2005b) chegaram a um modelo de que *exon shuffling* antigo teria contribuído para a formação das porções centrais das proteínas enquanto os domínios modernos 1-1 teriam sido adicionados posteriormente nas extremidades.

Em metazoários, aproximadamente 40% dos casos putativos de *exon shuffling* apresentaram sobreposição com domínios protéicos, sendo que a grande maioria (~70%) são simétricos de fase 1, corroborando os resultados obtidos por Liu *et al.* (2005). Os trabalho concentrou-se em estabelecer correlações entre bordas de exons e domínios, determinando um “score” estatístico para cada domínio de acordo com o grau de correlação com as bordas de seus exons codificantes. Os domínios freqüentemente sobrepostos às regiões envolvidas em *exon shuffling* e aqueles identificados pelo agrupamento hierárquico correspondem, em sua maior parte, aos domínios com “scores” significativamente mais altos reportados pelo trabalho de Liu *et al*. (2005). Os autores sugerem que estes domínios foram positivamente selecionados durante a evolução, expandindo-se em abundância e distribuição numa tendência crescente de invertebrados para vertebrados.

5.3. Expansões de domínios em metazoários

Um dos aspectos investigados no trabalho foi o sinal de flanqueamento por introns ao redor de domínios protéicos. Esta análise forneceu indícios sobre o período de expansão de domínios durante a evolução animal. Uma das principais conclusões é que a maior parte dos domínios 1-1 têm sido embaralhados desde o surgimento dos primeiros metazoários. Isso sugere que o período de transição para a multicelularidade foi marcado pelo embaralhamento acelerado deste grupo de domínios, possivelmente facilitado pelas elevadas taxas de ganho de introns na linhagem que deu origem aos animais (Carmel *et al.*, 2007).

Um período crucial na evolução dos animais foi a origem dos vertebrados. Este período também foi marcado pelo embaralhamento de certos domínios. O domínio

“Immunoglobulin C-1 set” por exemplo, é praticamente exclusivo de proteínas do sistema imune, nas cadeias leve e pesada de imunoglobulinas e em vários receptores de células T (Pammer e Cresswell, 1998; Rodosevich e Ono, 2003). O domínio “Xlink” está presente na proteína da cartilagem chamada “Link”, e em proteínas de adesão e migração celular (Barta *et al.*, 1993). Os domínios “Fibronectin type I e II”, em combinação com “Fibronectin type III”, compõem a fibronectina, uma proteína extremamente importante no desenvolvimento dos vertebrados, participando de inúmeros processos como adesão, diferenciação, coagulação, etc (Pankov e Yamada, 2002). O aumento na complexidade das estruturas gênicas não está relacionada somente à expansão de domínios, mas também o modo em que se combinam para formar um novo produto. Kawashima e colaboradores (2009) identificaram várias combinações novas que deram origem a em genes envolvidos com estruturas específicas de cordados, como endóstilo, membrana de Reissner, tubo neural e notocorda.

O domínio DUF1220 (“domain of unknown function”) chama a atenção por apresentar forte sinal de flanqueamento simétrico 1-1 apenas em *H. sapiens* (figura 15B). Este domínio foi recentemente descoberto em uma nova família de genes específica de primatas, a NBPF (“neuroblastoma breakpoint family”) (Vandepoele *et al.*, 2005). DUF1220 ocorre em repetições em tandem, com seu número de cópias aumentado nos primatas mais próximos à *H. sapiens*. O maior número de cópias deste domínio é encontrado na espécie humana. Ele é altamente expresso em regiões cerebrais associadas à áreas de funções cognitivas superiores, preferencialmente no corpo celular e dendritos dos neurônios (Popesco *et al.*, 2006). O papel de *exon shuffling* na evolução deste domínio e as funções desempenhadas por ele merecem ser investigadas com mais profundidade em estudos futuros.

5.4. Fatores que podem influenciar a expansão de domínios

Certos domínios protéicos são extremamente móveis e podem ser encontrados em uma grande diversidade de combinações ou arquiteturas[2]. A capacidade de um certo domínio se combinar com vários outros, contribuindo para a formação de proteínas multidomínios, pode ser chamada de "promiscuidade" (Basu *et al.*, 2008). De acordo com a classificação de Basu e colaboradores (2008), a maioria dos domínios identificados neste trabalho são considerados promíscuos. Alguns autores acreditam que os rearranjos de domínios são fruto de eventos aleatórios, independentes da função ou estrutura do domínio (Vogel *et al.*, 2005). A promiscuidade estaria relacionada com a antiguidade. Domínios mais antigos, presentes nos três reinos (Archaea, Bacteria e Eukarya) teriam tido maior tempo para se expandir (Apic *et al.*, 2001). Uma visão alternativa propõe que as combinações de domínios não é aleatória e estão sujeitas à pressões seletivas. Essa visão é suportada pelas seguintes observações: i) apenas uma pequena fração de todas as possíveis combinações são encontradas (Doolittle *et al.*, 1995) ii) a ordem dos domínios da extremidade N- para C- terminal tende a ser conservada (Vogel *et al.*, 2004) iii) certas combinações de domínos sofrem mais duplicações do que o esperado por chance (Apic *et al.*, 2003).

As razões pelas quais alguns domínios são mais promíscuos que outros ainda são desconhecidas. Tordai *et al.* (2005) verificaram que domínios de tamanho menor ocorrem em uma maior diversidade de arquiteturas. Os domínios que apresentam interfaces menores tendem a apresentar *folding* independente, uma condição que

---

[2] Arquitetura de domínios refere-se à organização seqüencial ou linear de domínios numa dada proteína.

interfere no *folding* dos domínios vizinhos e passa a ser especialmente importante em proteínas multidomínios (Han *et al.*, 2007). A promiscuidade está correlacionada positivamente com o número de interações físicas entre domínios, demonstrando uma associação de dependência funcional, principalmente relacionadas com transdução de sinal, estruturas extracelulares e sinalização célula-célula (Basu *et al.*, 2008).

Dessa forma, forças seletivas parecem governar a composição e organização dos domínios nas proteínas. Mecanismos de seleção devem ter influenciado na expansão preferencial de domínios 1-1 em metazoários, os quais estão fortemente enriquecidos em proteínas de matriz extracelular, por exemplo. A seleção pode ter atuado no tamanho, função ou estrutura dos domínios. Introns de fase 1 tendem a interromper códons de glicina (Fedorov *et al.*, 2001). É possível que esta preferência confira alguma vantagem seletiva em aspectos estruturais, promovendo uma adequação mais rápida do domínio quando inserido num novo contexto protéico.

5.5. *Exon shuffling* e características biológicas de metazoários

Uma das principais conseqüências do mecanismo de *exon shuffling* é que a aquisição de domínios ou o rearranjo de unidades protéicas promove a criação de proteínas que desempenham funções e interações moleculares diferentes. A relação entre *exon shuffling* e a presença de certas características biológicas foi investigada através da análise de termos enriquecidos do *Gene Ontology* para os genes que apresentaram evidência de *exon shuffling*.

A comparação dos termos enriquecidos em plantas e animais sugere que eventos de *exon shuffling* antigos podem estar mais associados a atividades de

proteina quinase e transcrição, por exemplo. Por outro lado, *exon shuffling* em metazoários, principalmente de regiões 1-1, está intimamente relacionado com proteínas da matriz extracelular, adesão celular, coagulação sangüínea, ativação do complemento, geração de neurônios etc.

A matriz extracelular (MEC) é composta de colágenos, proteoglicanas e glicoproteínas que formam uma complexa rede de suporte estrutural, organizacional e de orientação das células (Bosman e Stamenkovic, 2003). Além disso, a organização dos animais multicelulares depende de moléculas que promovam coesão intercelular e entre as células e o meio externo. Embora *T. adhaerens* não tenha membrana basal e MEC descritas (Schierwater, 2005; Srivastava, *et al.*, 2008) e *N. vectensis* não possua uma MEC tal como os animais triblásticos[3] (Huxley-Jones *et al.*, 2007), os termos “extracellular matrix”, “collagen” e “cell adhesion” apresentaram-se enriquecidos nessas duas espécies. Diversos genes humanos envolvidos na adesão célula-célula, célula-matriz e na composição da membrana basal apresentaram homólogos em *T. adhaerens* ou *N. vectensis*. Dentre os genes identificados, podem ser citados “collagen IV”, “laminin-α, -β e -γ”, “nidogen”, “heparan sulfate proteoglycan”, “contactin associated protein-like”, “thrombospondin“, “integrin-α e -β” “tenascin”, “fibrillin”, “agrin” entre outros.

Ao mesmo tempo que parte dos genes de MEC e adesão são bastante conservados em animais primitivos, *exon shuffling* também contribuiu para a origem de genes recentes. As matrilinas formam uma família de quatro membros. As proteínas possuem cópias dos domínios “Von Willebrand type A” e “EGF-like”, todas simétricas de fase 1. São moléculas que surgiram na linhagem dos deuterostômios

---

[3] Animais triblásticos são aqueles que possuem três camadas germinativas: ectoderme, mesoderme e endoderme.

(Huxley-Jones *et al.*, 2007) e participam na formação de estruturas fibrilares e filamentosas, expressas principalmente na cartilagem (Wagener *et al.*, 2005). Outro exemplo é o gene “aggrecan”, cuja função é promover resistência à compressão na cartilagem, uma estrutura única dos vertebrados (Watanabe e Yamada, 1999). A proteína possui os domínios “Immunoglobulin-V-set”, “Xlink”, “C-type lectin”, “Sushi”, EGF-like e EGF-extracellular. O par IG-V-set/Xlink é uma combinação nova, adquirida por *domain shuffling* no ancestral dos vertebrados (Kawashima *et al.*, 2009).

Assim como Patthy (1985), genes que participam da coagulação e fibrinólise, tais como os fatores VII/IX/X/II, V/VIII, XIII, XII, proteínas C e Z, proteína S, plasminogênio e urokinase, também foram identificados. A maioria dessas proteínas são serina proteases que contém os domínios “GLA” e “Peptidase S1” (domínio com atividade proteolítica) associados aos domínios tipicamente envolvidos em *shuffling* (“Fibronectin type III, II e I”, “EGF”, “PAN”, “Sushi”, “Kringle”) (Jiang e Doolittle, 2003). Jiang e Doolittle (2003) buscaram por ortólogos de 26 genes de mamíferos envolvidos na coagulação e fibrinólise nos genomas do peixe *Fugu rubripes* e do urocordado *Ciona intestinalis*. Foram encontrados homólogos para 21 genes em *Fugu rubripes* e nenhum em *Ciona intestinalis*, mostrando que o repertório de genes da coagulação sangüínea foi originado após a divergência dos urocordados, havendo intensa participação de eventos de duplicação gênica e *domain shuffling*.

Termos como “generation of neurons“, “axon guidance” e “neuropeptide signaling pathway” estão enriquecidos no conjunto de genes com evidência de *exon shuffling*, dentre os quais alguns são conservados entre *H. sapiens*, *N. vectensis* e *T. adhaerens*. O gene CELSR3 por exemplo, codifica para uma caderina essencial na migração de neurônios no cérebro anterior (Ying *et al.*, 2009). O gene SEMA5A, possivelmente envolvido na orientação dos axônios (Hilario *et al.*, 2009), e NCAM1,

uma molécula envolvida em diversos processos do sistema nervoso, apresentaram homólogos em *T. adhaerens* e *N. vectensis*. Netrinas (NTN1) e receptores de netrinas (UNC5), proteínas que participam no direcionamento dos axônios (Dickson, 2002), também foram identificados. É possível que essas moléculas exerçam um papel sensorial primitivo, possibilitando a exploração e percepção do ambiente externo. Genes de origem mais recente, relacionados com o sistema nervoso, também foram identificados. Com base no alinhamento entre proteínas, o gene BAI1 aparentemente é encontrado apenas em vertebrados. Este gene pode ter papel na adesão celular e transdução de sinal, funcionando como um potente inibidor da angiogênese no cérebro (Duda *et al.*, 2002). O MDGA1 possui homólogo em *Ciona intestinalis* e parece ter surgido nos cordados. Sua função pode estar envolvida com a migração neuronal na camada superficial do neocórtex (Takeuchi e O'Leary, 2006).

Os mecanismos genômicos e seletivos que dirigem os rearranjos de exons e domínios, como a inserção de um domínio afeta o caráter estrutural e funcional da proteína aceptora, e como os rearranjos podem influenciar na evolução de caracteres fenotípicos ainda são questões mal compreendidas. A abordagem desses assuntos poderá ser de grande impacto para a compreensão não apenas de mecanismos evolutivos fundamentais, como questões biológicas práticas e aplicadas.

6. Conclusões

Dada a importância do fenômeno de *exon shuffling* como mecanismo gerador de diversidade funcional, a proposta deste trabalho foi gerar um catálogo de regiões protéicas cuja origem se deu por *exon shuffling* e estudar aspectos desse mecanismo durante a evolução dos eucariotos. A determinação dos padrões de simetria de regiões conservadas e de domínios protéicos mostrou que em espécies não metazoárias, eventos de *exon shuffling* são menos freqüentes e predominantemente do tipo 0-0. Em metazoários, *exon shuffling* tornou-se mais abundante, especialmente devido a expansão de unidades simétricas 1-1. *Trichoplax adhaerens* apresentou um padrão intermediário de freqüências de exons 0-0 e 1-1, indicando que a expansão de eventos 1-1 está claramente associada ao aparecimento dos primeiros animais multicelulares. Casos de *exon shuffling* de classe 2-2 foram pouco evidentes. Grupos de domínios protéicos com sinal de flanqueamento por introns em todos os metazoários e especificamente em vertebrados foram identificados. Estes domínios estão presentes principalmente em proteínas relacionadas com características específicas dos animais como a matriz extracelular e adesão. Os rearranjos destes domínios também contribuíram para o aumento da complexidade, principalmente em vertebrados, através da montagem de proteínas de coagulação sanguínea, sistema imune, geração de neurônios, etc.

REFERÊNCIAS

Altschul, S.F, Madden, T.L., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, v. 25, n. 17, p. 3389-3402.

Apic, G., Gough, J., Teichmann, S.A. (2001). Domain combinations in Archaeal, Eubacterial and Eukaryotic proteomes. **Journal of Molecular Biology**, v. 310. 311-325.

Apic, G., Huber, W., Teichmann, S.A. (2003). Multi-domain protein families and domain pairs: comparison with known structures and random model of domain recombination. **Journal of Structural and Functional Genomics**, v. 4, p. 67-78.

Barta, E., Deák, F., Kiss, I. (1993). Evolution of the hyaluronan-binding module of link protein. **Biochemical Journal**, v. 292, p. 947-949.

Bashton, M., Chotia, C. (2007). The generation of new protein functions by the combination of domains. **Structure**, v. 15, p. 85-99.

Basu, M.K., Carmel, L., Rogozin, I.B., Koonin, E.V. (2008a). Evolution of domain promiscuity in eukaryotes. **Genome Research**, v. 18, p. 449-461.

Betrán, E. Long, M. (2002). Expansion of genome coding regions by acquisition of new genes. **Genetica**, v. 115, p. 65-80.

Björklund, A.K., Ekman, D., Elofsson, A. (2006). Expansion of protein domain repeats. **PLOS Computational Biology**, v. 2, p. 959-970.

Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S.K., Teichmann, S.A., Weiner III, J. (2005). The evolution of domain arrangements in protein and interaction networks. **CMLS Cellular and Molecular Life Sciences**, v. 62, p. 435-445.

Bosman, F.T., Stamenkovic, I. (2003). Functional structure and composition of the extracellular matrix. **Journal of Pathology**, v. 200, p. 423-428.

Burger, G., Forget, L., Zhu, Y., Gray, M.W., Lang, B.F. (2003). Unique mithocondrial genome architecture in unicellular relatives of animals. **Proc. Natl. Acad. Sci. USA**, v. 100, p. 892-897.

Carmel, L., Wolf, Y.I., Rogozin, I.B., Koonin, E.V. (2007). Three distinct modes of intron dynamics in the evolution of eukaryotes. **Genome Research**, v. 17, p. 1034-1044.

Cavalier-Smith, T. (1991). Intron phylogeny: a new hypothesis. **Trends in Genetics**, v. 7, p. 145-148.

Chandrasekaran, C., Betrán, E. (2008). Origins of new genes and pseudogenes. **Nature Education**, 1(1).

Chotia, C., Gough, J., Vogel, C., Teichmann, S.A. (2003). Evolution of the protein repertoire. **Science**, v. 300, p. 1701-1703.

Claverie, J.M. (2001). Gene number. What if there are only 30,000 human genes? **Science**, v.291, 1255-1257.

De Souza, S.J., Long, M., Schoenbach, L., Roy, S.W., Gilbert, W. (1996). Intron positions correlate with module boundaries in ancient proteins. **Proc. Natl. Acad. Sci. USA**, v. 93, n. 25, p. 14632-14636.

De Souza, S.J., Long, M., Klein, R.J., Roy, S.W., Lin, S., Gilbert, W. (1998). Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. **Proc. Natl. Acad. Sci. USA**, v. 95, p. 5094-5099.

De Souza, S.J. (2003). The emergence of a synthetic theory of intron evolution. **Genetica**, v. 118, p. 117-121.

Dibb, N.J., Newman, A.J. (1989). Evidence that introns arose at proto-splice sites. **EMBO Journal**, v. 8, n. 7, p. 2015-2021.

Doolittle, R.F. (2005). The multiplicity of domains in proteins. **Annual Review of Biochemistry**, v. 64, p. 287-314.

Doolittle, W.F. (1978). Genes in pieces: were they ever together? **Nature**, v. 272, p. 581-582.

Duda, D.G., Sunamura, M., Lozonschi, L., Yokoyama, T., Yatsuoka, T., Motoi, F., Horii, A., Tani, K., Asano, S., Nakamura, Y., Matsuno, S. (2002). Overexpression of the p53-inducible brain-specific angiogenesis inhibitor 1 suppresses efficiently tumor angiogenesis. **British Journal of Cancer**, v. 86, p. 490-496.

Eddy, S.R. (1998). HMMER User's guide: Biological sequence analysis using profile Hidden Markov Models, version 2.1.1: Washington University School of Medicine.

Eickbush, T.H. (1999). Exon shuffling in retrospect. **Science**, v. 283, p. 1465-1467.

Ekman, D., Bjorklund, A.K., Frey-Scott, J., Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. **Journal of Molecular Biology**, v. 341, n. 1, p. 231-243.

Ekman, D., Björklund, A.K., Elofsson, A. (2007). Quantification of the elevated rate of domain rearrangements in Metazoa. **Journal of Molecular Biology**, v. 372, p. 1337-1348.

Ender, A., Schierwater, B. (2003). Placozoa are not derived cnidarians: evidence from molecular morphology. **Molecular Biology and Evolution**, v. 20, p. 130-134.

Fedorov, A., Suboch, G., Bujakov, M., Fedorova, L. (1992). Analysis of nonuniformity in intron phase distribution. **Nucleic Acids Research**, v. 20, n. 10, p. 2553-2557.

Fedorov, A., Fedorova, L., Starshenko, V., Filatov, V., Grigor'ev, E. (1998). Influence of exon duplication on intron and exon phase distribution. **Journal of Molecular Evolution**, v. 46, p. 263-271.

Fedorov, A., Cao, X., Saxonov, S. De Souza, S.J., Roy, S.W., Gilbert, W. (2001). Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. **Proc. Natl. Acad. Sci. USA**, v. 98, 13177-13182.

Fedorov, A., Roy, S.W., Cao, X., Gilbert, W. (2003). Phylogenetically older introns strongly correlate with module boundaries in ancient proteins. **Genome Research**, v. 13, p. 1155-1157.

Fedorova, L., Fedorov, A. (2003). Introns in gene evolution. **Genetica**, v. 118, p. 123-131.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, J.S., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., Bateman, A. (2008). The Pfam protein families database. **Nucleic Acids Research**, v. 36, p. 281-288.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissmann, S., Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. **Genome Research**, v. 17, p. 669-681.

Gilbert, W. (1978). Why genes in pieces? **Nature**, v. 271, p. 501.

Gilbert, W. (1987). The exon theory of genes. **Cold Spring Harbor Symposia on Quantitative Biology**, v. 52, p. 901-905.

Go, M. (1981). Correlation of DNA exonic regions with protein structural units in haemoglobin. **Nature**, v. 291, p. 677-679.

Han, J.H., Batey, S., Nickson, A.A., Teichmann, S.A., Clarke, J. (2007). The folding and evolution of multidomain proteins. **Nature Reviews Molecular Cell Biology**, v. 8, p. 319-330.

Hotopp, J. *et al*. (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. **Science**, v. 317, p. 1753-1756.

Hurles, M. (2004). Gene duplication: the genomic trade in spare parts. **Plos Biology**, v. 2, p. 900-904.

Huxley-Jones, J., Robertson, D.L., Boot-Handford, R.P. (2007). On the origins of the extracellular matrix in vertebrates. **Matrix Biology**, v. 26, p. 2-11.

Jacob, F. (1977). Evolution and tinkering. **Science**, v. 196, n. 4295, p. 1161-1166.

Jeffares, D.C., Mourier, T., Penny, D. (2006). The biology of intron gain and loss. **Trends in Genetics**, v. 22, p. 16-22.

Jiang, Y., Doolittle, R.F. (2003). The evolution of vertebrate blood coagulation viewed from a comparison of pupper fish and sea squirt genomes. **Proc. Natl. Acad. Sci. USA**, v. 100, n. 13, p. 7527-7532.

Kaessmann, H., Zöllner, S., Nekrutenko, A., Li, W.H. (2002). Signatures of domain shuffling in the human genome. **Genome Research**, v. 12, p. 1642-1650.

Kawashima, T., Kawashima, S., Tanaka, C., Murai, M., Yoneda, M., Putnam., N.H., Rokhsar, D.S., Kanehisa, M., Satoh, N., Wada, H. (2009). Domain shuffling and the evolution of vertebrates. **Genome Research**, v. 19, n. 8, p. 1393-1403.

King, N., Westbrook, M.J., Young, S.L. *et al*. (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. **Nature**, v. 451, p. 783-788.

Kolkman, J.A., Stemmer, W. P.C. (2001). Directed evolution of proteins by exon shuffling. **Nature Biotechnology**, v. 19, p. 423-428.

Koonin, E.V., Wolf, Y.I, Karev, G.P. (2002). The structure of the protein universe and genome evolution. **Nature**, v. 420, p. 218-223.

Koonin, E. V. (2006). The origins of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? **Biology Direct**, 1:22.

Koonin, E.V. (2009). Evolution of genome architecture. **The International Journal of Biochemistry & Cell Biology**, v. 41, p. 298-306.

Lang, B.F., O'Kelly, C., Nerad, T., Gray, M.W., Burger, G. (2002). The closest unicellular relatives of animals. **Current Biology**, v. 12, p. 1773-1778.

Lake, J.A. (1990). Origin of Metazoa. **Proc. Natl. Acad. Sci. USA**, v. 87, p. 763-766.

Lander, E.S. Linton, L.M., Nusbaum, C. *et al*. (2001). Initial sequencing and analysis of the human genome. **Nature**, v. 409, p. 860-921.

Lewin, B. (2007). Genes IX. Jones and Bartlett Publishers, Sudbury, Massachusetts, 892 p.

Liu, M., Grigoriev, A. (2004). Protein domains correlate strongly with exons in multiple eukaryote genomes: evidence of exon shuffling? **Trends in Genetics**, v. 20, n.9, p. 399-403.

Liu, M., Walch, H, Wu, S., Grigoriev, A. (2005). Significant expansion of exon-bordering domains during animal proteome evolution. **Nucleic Acids Research**, v. 33, n. 1, p. 95-105.

Logsdon Jr, J.M., Tyshenko, M.G., Dixon, C., D-Jafari, J., Walker, V.K., Palmer, J.D. (1995). Seven newly discovered intron positions i the triose-phosphate isomerase gene: evidence for the introns-late theory. **Proc. Natl. Acad. Sci. USA**, v. 92, p. 8507-8511.

Logsdon Jr, J.M., (1998). The recent origins of spliceosomal introns revisited. **Current Opinion in Genetics & Development**, v. 8, p. 637-648.

Long, M., Langley, C.H. (1993). Natural selection and the origino f jingway, a chimeric processed functional gene in *Drosophila*. **Science**, v. 293, p. 91-95.

Long, M., Rosenberg, C., Gilbert, W. (1995). Intron phase correlations and the evolution of the intron/exon structure of genes. **Proc. Natl. Acad. Sci. USA**, v. 95, p. 219-223.

Long, M., Wang, W., Zhang, J. (1999). Origin of new genes and source of N-terminal domain of the chimerical gene, *jingwei*, in *Drosophila*. **Gene**, v. 238, p. 135-141.

Long, M., Betrán, E., Thornton, K., Wang, W. (2003). The origin of new genes: glimpses from the Young and old. **Nature Reviews Genetics**, v. 4, 865-875.

Maere, S., Heymans, K., Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. **Bioinformatics**, v. 21, p. 3448-3449.

Marchionni, M., Gilbert, W. (1986). The triose phosphate isomerase gene from maize: introns antedate the plant-animal divergence. **Cell**, v. 46, p. 133-141.

Dellaporta, S.L., Xu, A., Sagasser, S., Jakob., W., Moreno, M.A., Buss, L.W., Schierwater, B. (2006). Mitochondrial genome of *Trichoplax adhaerens* supports Placozoa as the basal lower metazoan phylum. **Proc. Natl. Acad. Sci. USA**, v. 103, n. 23, p. 8751-8756.

Moran, J.V., DeBerardinis, R.J., Kazazian Jr, H.H. (1999). Exon shuffling by L1 retrotransposition. **Science**, v. 283, p. 1530-1534.

Nguyen, H.D., Yoshihama, M., Kenmochi, N. (2006). Phase distribution of spliceosomal introns: implications for intron origin. **BMC Evolutionary Biology**, 6:69.

Nielsen, C.B., Friedman, B., Birren, B., Burge, C.B., Galagan, J.E. (2004). Patterns of intron gain and loss in Fungi. **PLOS Biology**, 2:12.

Nixon, J.E.J., Wang, A., Morrison, H.G. *et al.* A spliceosomal intron in *Giardia lamblia*. **Proc. Natl. Acad. Sci. USA**, v. 99, p. 3701-3705.

Ochman, H., Lawrence, J.G., Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. **Nature**, v. 405, p. 299-304.

Palmer, J.D., Logddon, J.M.Jr. (1991). The recent origin of introns. **Current Opinion in Genetics & Development**, v.1, 470-477.

Pammer, E., Cresswell, P. (1998). Mechanisms of MHC class I – restricted antigen processing. **Annual Review of Immunology**, v. 16, p. 323-358.

Panchenko, A. R., Luthey-Schulte, Z., Wolynes, P.G. (1996). Foldons, protein structural modules, and exons. **Proc. Natl. Acad. Sci. USA**, v. 93, p. 2008-2013.

Pankov, R., Yamada, K.M. (2002). Fibronectin at a glance. **Journal of Cell Science**, v. 115, p. 3861-3863.

Patthy, L. (1985). Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. **Cell**, v. 41, p. 657-663.

Patthy, L. (1987). Intron-dependent evolution: preferred types of exons and introns. **FEBS Letters**, v. 214, p. 1-7.

Patthy, L. (1996). Exon shuffling and other ways of module Exchange. **Matrix Biology**, v. 15, p. 301-310.

Patthy, L. (1999a). Protein Evolution. Blackwell Science, Oxford, 228 p.

Patthy, L. (1999b). Genome evolution and the evolution of exon suffling- a review. **Gene**, v. 238, 103-114.

Patthy, L. (2003). Modular assembly of genes and the evolution of new functions. **Genetica**, v. 118, p. 217-231.

Pearson, H. (2006). Genetics: What is a gene? **Nature**, v. 238, p. 398-401.

Popesco, M.C., Maclaren, E.J., Hopkins, J., Dumas, L., Cox, M., Meltesen, L., McGavran, L., Wyckoff, G.J., Sikela, J.M. (2006). Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. **Science**, v. 313, p. 1304-1307.

Putnam, N.H., Srivastava, M., Hellsten, U., *et al.* (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. **Science**, v. 317, p. 86-94.

Qiu, W., Schisler, N., Stoltzfus, A. (2004). The evolutionary gain of spliceosomal introns: sequence and phase preferences. **Molecular Biology and Evolution**, v. 21, p. 1252-1263.

Radosevich, M., Ono, S.J. (2003). Novel mechanisms of class II major histocompatibility complex gene regulation. **Immunology Research**, v. 27, n. 1, p. 85-106.

Raible, F., Tessmar-Raible, K., Osoegawa, K. *et al.* (2005). Vertebrate-type intron rich genes in marine annelid Platynereis dumerilii. **Science**, v. 310, p. 1325-1326.

Robart, A.R., Zimmerly, S. (2005). Group II intron retroelements: funtion and diversity. **Cytogenetic and Genome Research**, v. 110, p. 589-597.

Rodriguez-Trelles, F., Tarrío, R., Ayala, F.J. (2006). Origins and evolution of spliceosomal introns. **Annual Review of Genetics**, v. 40, p47-76.

Rogers, J.H. (1990). The role of introns in evolution. **FEBS Letters**, v. 268, p. 339-343.

Rogozin, I.B., Wolf., Y.I., Sorokin, A.V., Mirkin, B.G., Koonin, E.V. (2003). Remarkable interkingdon conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. **Current Biology**, v. 13, p. 1512-1517.

Rogozin, I.B., Sverdlov, A.V., Babenko, V.N., Koonin, E.V. (2005). Analysis of evolution of exon-intron structure of eukaryotic genes. **Briefings in Bioinformatics**, v. 6, p. 118-134.

Roy, S.W., Gilbert, W. (2005). Rates of intron loss and gain: implications for early eukaryotic evolution. **Proc. Natl. Acad. Sci. USA**, v. 19, p. 5773-5778.

Roy, S.W., Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. **Nature Reviews Genetics**, v. 7, p. 211-221.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schiwikowski, B., Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. **Genome Research**, v. 13, n. 11, p. 2498-2504.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., Kasprzyk, A. (2009). BioMart – biological queries made easy. BMC Genomics, v. 10:22.

Söding, J., Lupas, A.N. (2003). More than sum of their parts: on the evolution of proteins from peptides. **Bioessays**, v. 25, 837-846.

Srivastava, M., Begovic, E., Chapman, J. *et al.* (2008). The *Trichoplax adhaerens* genome and the nature of placozoans. **Nature**, v. 454, p. 955-960.

Sverdlov, A.V., Rogozin, I.B., Babenko, V.N, Koonin, E.V. (2003). Evidence of splice signal migration from exon to intron during intron evolution. **Current Biology**, v. 13, p. 2170-2174.

Strachan, T., Read, A.P. (2003). Human molecular genetics. John Wiley & Sons, New York, 2. ed.

Sverdlov, A.V., Rogozin, I.B., Babenko, V.N., Koonin, E.V. (2004). Reconstruction of ancestral protosplice sites. **Current Biology**, v. 14, p. 1505-1508.

Szathmáry, E., Jordan, F., Pál, C. (2001). Can genes explain biological complexity? **Science**, v. 292, p. 1315-1316.

The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. **Nature Genetics**, v. 25, p. 25-29.

Tittiger, C.S., Whyard, S., Walker, V.K. (1993). A novel intron site in the triose phosphate isomerase gene from the mosquito *Culex tarsalis*. **Nature**, v. 361, p. 470-472.

Tordai, H., Nagy, A., Farkas, K., Banyai, L., Patthy, L. (2005). Modules, multidomain proteins and organismic complexity. **The FEBS Journal**, v. 272, p. 5064-5078.

Vandepoele, K., Van Roy, N., Staes, K., Speleman, F., Van Roy, F. (2005). A novel gene family NBPF: Intricate structure generated by gene duplications during primate evolution. **Molecular Biology and Evolution**, v. 22, n. 11, p. 2265-2274.

Van Rijk, A., Bloemendal, H. (2003). Molecular mechanisms of exon shuffling: illegitimate recombination. **Genetica**, v. 118, p. 245-249.

Vanacová, S., Yan, W., Carlton, J.M., Jonhnson, P.J. (2005). Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis.* **Proc. Natl. Acad. Sci. USA**, v. 102, p. 4430-4435.

Vibranovski, M.D. (2005a). O Papel do fenômeno de “exon shuffling” antigo e moderno na evolução de proteínas. São Paulo, Universidade de São Paulo, Programa de Pós Graduação em Bioquímica, Tese de Doutorado, 145 p.

Vibranovski, M.D., Sakabe, N.J., De Oliveira, R.S., De Souza, S.J. (2005b). Signs of ancient and modern exon shuffling are correlated to the distribution of ancient and modern domains along proteins. **Journal of Molecular Evolution**, v. 61, p. 341-350.

Vogel, C., Berzuini, C., Bashton, M., Gough, J., Teichmann, S.A. (2004). Supra-domains - evolutionary units larger than single protein domains. **Journal of Molecular Biology**, v. 336, p. 809-823.

Vogel, C., Teichmann, S.A., Pereira-Leal, J. (2005). The relationship between domain duplication and recombination. **Journal of Molecular Biology**, v. 346, p. 355-365.

Wagener, R., Ehlen, H.W.A., Ko, Y., Kobbe, B., Mann, H.H., Sengle., G., Paulsson, M. (2005). The matrilins – adaptor proteins in the extracellular matrix. **FEBS Letters**, v. 579, p. 3323-3329.

Wain, H.M., Bruford, E. A., Lovering, R. C., Lush, M.J., Wright, M. W., Povey, S. (2002). Guidelines for human gene nomenclature. **Genomics**, v. 79, p. 464-470.

Watanabe, H., Yamada, Y. (1999). Mice lacking link protein develop dwarfism and craniofacial abnormalities. **Nature Genetics**, v. 21, p. 225-229.

Yandell, M., Mungall, C.J., Smith, C., Prochnik, S., Kaminker, J., Hartzell, G., Lewis, S., Rubin, G.M. (2006). Large-scale trends in the evolution of gene structures within 11 animal genomes. **Plos Computational Biology**, v. 2, p. 113-125.

Zhang, J. (2003). Evolution by gene duplication. **Trends in Ecology and Evolution**, v. 18, p. 292-298.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., Wang, W. (2008). On the origino f new genes in *Drosophila*. **Genome Research**, v. 18, p. 1446-1455.

LISTA DE ANEXOS

ANEXO A – Súmula curricular

DADOS PESSOAIS

Nome: Gustavo Starvaggi França

Local e data de nascimento: São Paulo – SP, 17 de Abril de 1984.

EDUCAÇÃO

2007 – 2009

Mestrado em Bioquímica e Biologia Molecular - Departamento de Bioquímica do Instituto de Química da Universidade de São Paulo / Laboratório de Biologia Computacional do Instituto Ludwig de Pesquisa sobre o Câncer.

2002 – 2005

Bacharelado em Ciências Biológicas – Universidade Federal do Paraná – Curitiba-PR.

OCUPAÇÃO

Bolsista de Mestrado – FAPESP, março de 2007 – fevereiro de 2008.

PUBLICAÇÕES

Manuscritos em preparação

França, G.S., Cancherini, D.V., De Souza, S.J. Evolutionary history of exon shuffling in eukaryotes. *Em preparação*.

Cancherini, D.V., França, G.S., De Souza, S.J. The role of exon shuffling in shaping protein-protein interaction networks. *Em preparação*.

Trabalhos em congressos

Cancherini, D.V., França, G.S., De Souza, S.J. Exon shuffling and the evolution of protein-protein interaction networks. Gordon Research Conference in Quantitative Genetics and Genomics. Galveston – Texas, 2009.

França, G.S., Cancherini, D.V., De Souza, S.J. Evolutionary history of exon shuffling. Fourth International Conference of the Brazilian Association for Bioinformatics and Computational Biology. Salvador – Bahia, 2008.

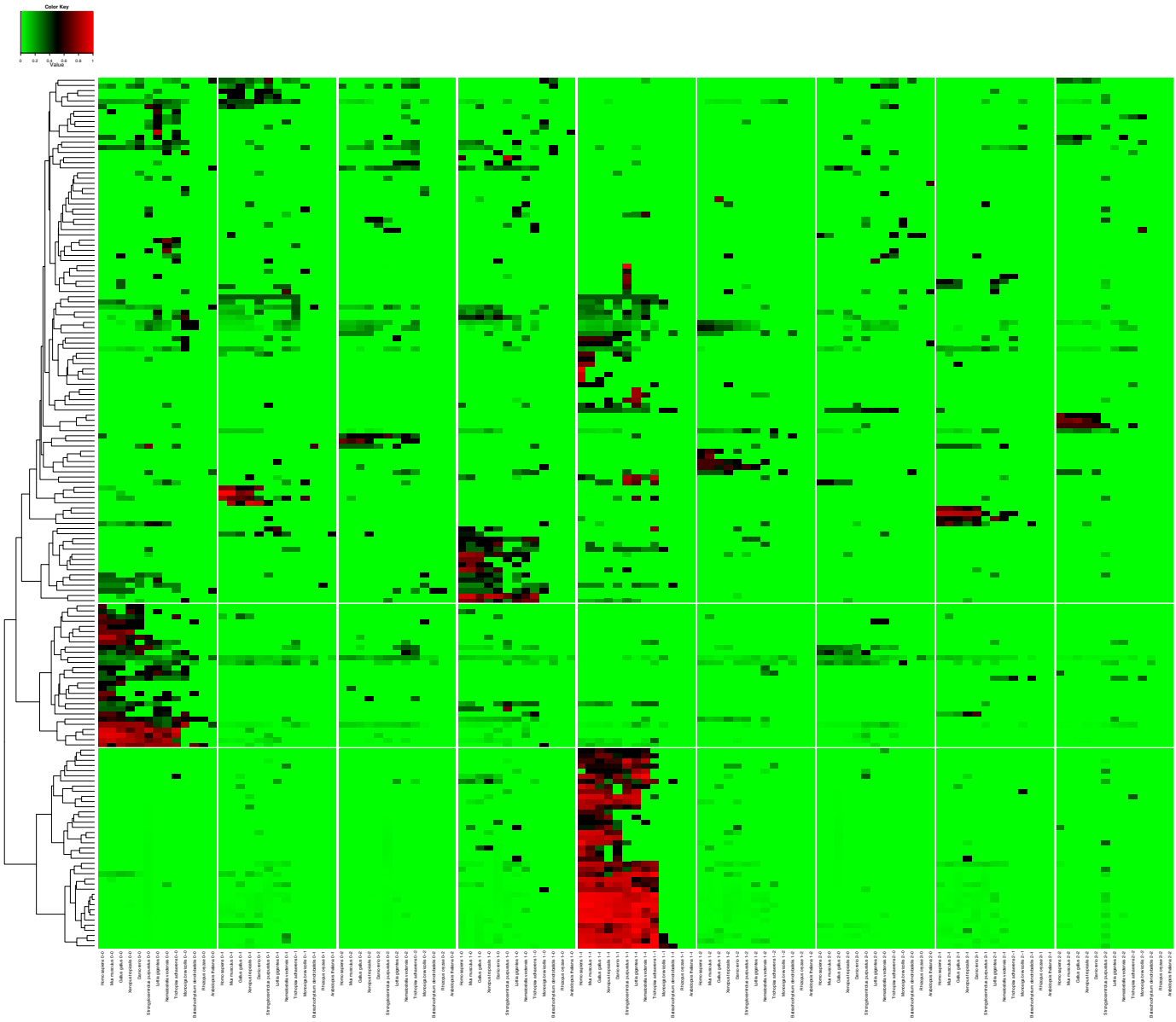
ANEXO B – Tabela das espécies utilizadas, informações taxonômicas e fonte dos arquivos de anotação genômica e seqüências protéicas .

Anex

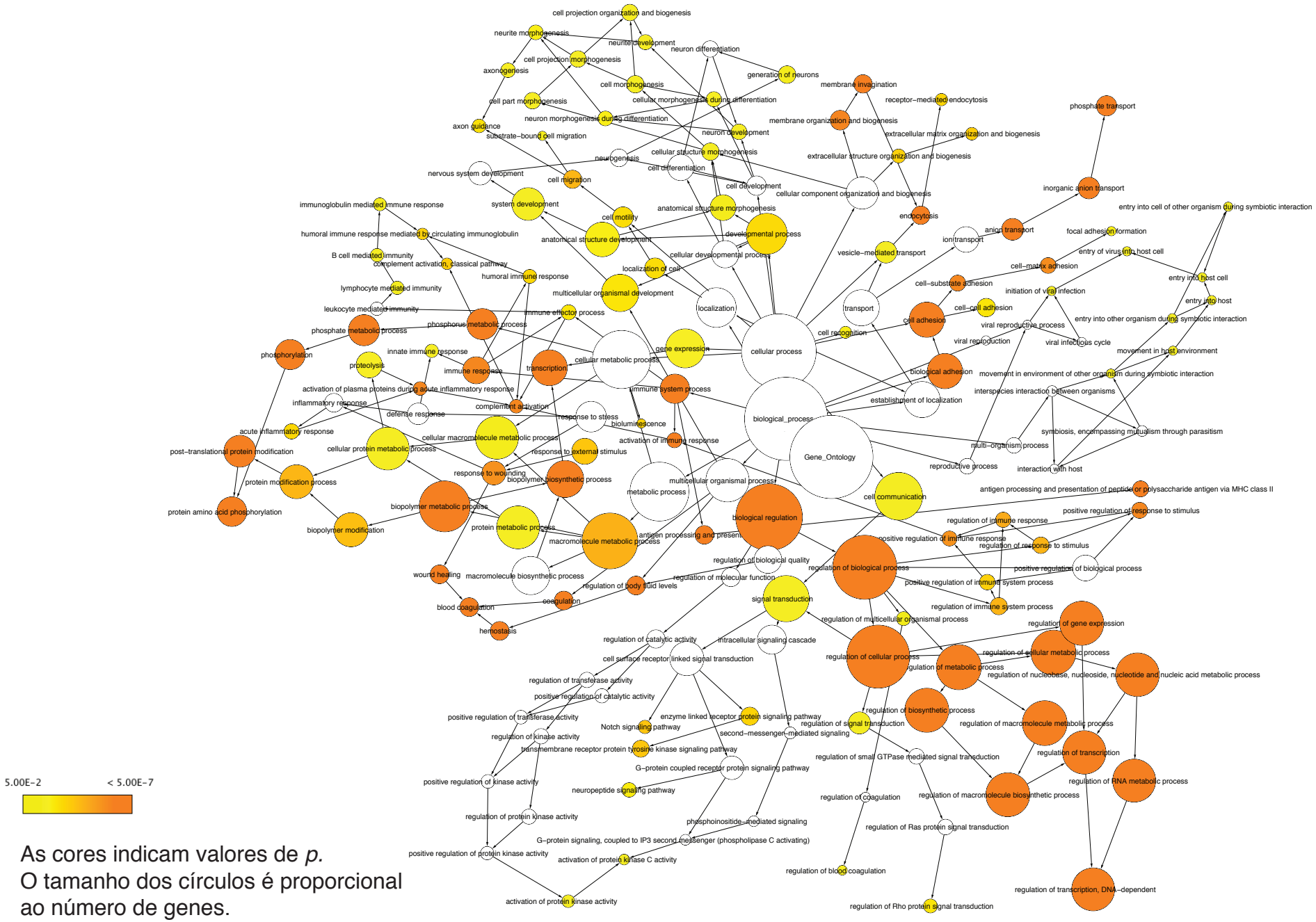
Espécie	Reino/Filo/Classe	Fonte
Homo sapiens	Metazoa/Chordata/Mammalia	http://www.ensembl.org/info/data/ftp/index.html
Mus musculus	Metazoa/Chordata/Mammalia	http://www.ensembl.org/info/data/ftp/index.html
Gallus gallus	Metazoa/Chordata/Aves	http://www.ensembl.org/info/data/ftp/index.html
Xenopus tropicalis	Metazoa/Chordata/Amphibia	http://www.ensembl.org/info/data/ftp/index.html
Danio rerio	Metazoa/Chordata/Actinopterygii	http://www.ensembl.org/info/data/ftp/index.html
Ciona intestinalis	Metazoa/Chordata/Ascidiacea	http://www.ensembl.org/info/data/ftp/index.html
Strongylocentrotus purpuratus	Metazoa/Echinodermata/Echinoidea	ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Spurpuratus/fasta/
Lottia gigantea	Metazoa/Mollusca/Gastropoda	http://genome.jgi-psf.org/Lotgi1/Lotgi1.download.ftp.html
Nematostella vectensis	Metazoa/Cnidaria/Anthozoa	http://genome.jgi-psf.org/Nemve1/Nemve1.download.ftp.html
Trichoplax adhaerens	Metazoa/Placozoa/	http://genome.jgi-psf.org/Triad1/Triad1.download.ftp.html
Monosiga brevicollis	Protista/Choanozoa/Choanoflagelatta	http://genome.jgi-psf.org/Monbr1/Monbr1.download.ftp.html
Rhizopus oryzae	Fungi/Zygomycota/Zygomycetes	http://www.broadinstitute.org/annotation/genome/rhizopus_oryzae
Cryptococcus neoformans	Fungi/Basidiomycota	ftp://ftp.ncbi.nih.gov/genomes/Fungi/Cryptococcus_neoformans_var_JEC21/
Batrachochytrium dendrobatis	Fungi/Chytridiomycota/Chytridiomycetes	http://genome.jgi-psf.org/Batde5/Batde5.download.ftp.html
Arabidopsis thaliana	Plantae/Tracheophyta/Magnoliopsida	ftp://ftp.ncbi.nih.gov/genomes/Arabidopsis_thaliana/GNOMON/

Anexo C – Agrupamento hierárquico e mapa de calor de domínios flanqueados por introns.

Legenda: Como indicado pela legenda do mapa, as regiões em vermelho e verde correspondem respectivamente, a frações elevadas e reduzidas de flanqueamento de domínios pro introns. O equinodermo *Strongylocentrotus purpuratus* foi utilizado em substituição a *Ciona intestinalis*, devido ao fenômeno de perda de introns sofrido pelo urocordado (Putnam *et al.*, 2007). O fungo *Batrachochytrium dendrobatis* também foi adicionado às análises.

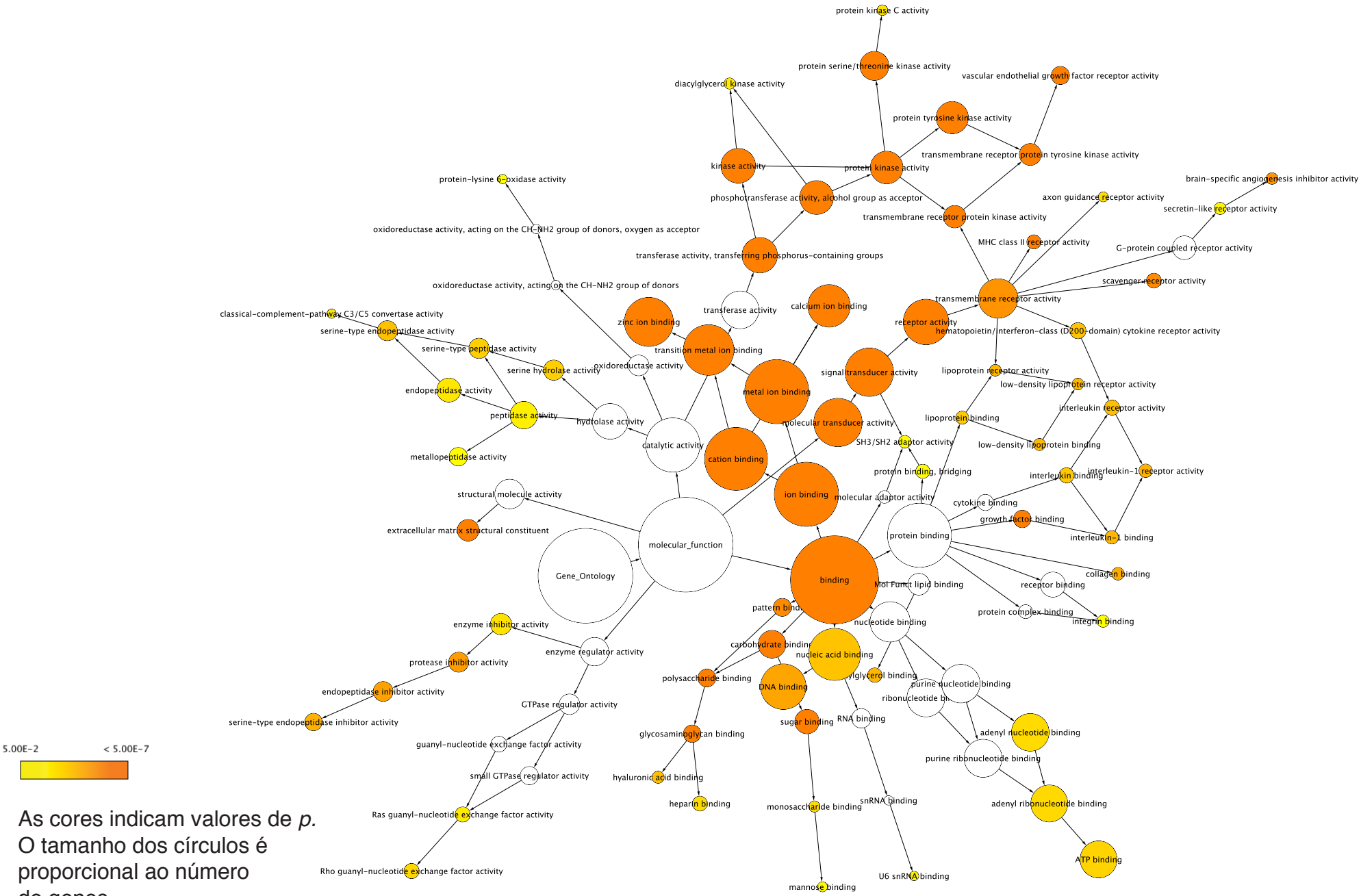

Color Key
0 0.2 0.4 0.6 0.8 1
Value
PF00046 Myb–like DNA–binding domain
PF00270 DEAD/DEAH box helicase
PF00567 Tudor domain
PF06119 Nitrogen–like
PF00169 PH domain
PF00788 Ras association (RalGDS/AF–6) domain
PF00227 Proteasome subunit, peptide complex small chain
PF00270 DNA2/NAM7
PF07728 ATPase family associated with various cellular activities (AAA)
PF00400 WD40/G10 family
PF06184 Sapozin–like type B, region 1
PF01424 R3H domain
PF00153 Mitochondrial carrier protein
PF00270 DEAD box family
PF04851 Type III restriction enzyme, res subunit
PF00084 von Willebrand factor type D domain
PF00646 Gelsolin repeat
PF01851 Proteasome/cyclosome repeat
PF04548 AIG1 family
PF01535 PPR repeat
PF03596 Eukaryotic protein of unknown function, DUF292
PF00350 Dynamin family
PF00270 DEAD box helicase
PF00171 Beta defensin
PF00561 alpha/beta hydrolase fold
PF00107 Zinc–binding dehydrogenase
PF00074 FG/B type C domain
PF00106 short chain dehydrogenase
PF00653 kR domain
PF00270 DEAD box
PF00147 CHCH domain
PF02142 MGS–like domain
PF04482 Slime mold cyclic AMP receptor
PF02176 TRAF–type zinc finger
PF03372 PDRM central domain
PF00149 Calcineurin–like phosphoesterase
PF00270 DEAD box
PF07960 Male sterility protein
PF00791 ZUS domain
PF00050 Kazal–type serine protease inhibitor domain
PF07648 Kazal–type serine protease inhibitor domain
PF07690 Major Facilitator SuperFamily
PF00270 SNARE domain
PF00270 DEAD box family
PF00397 WW domain
PF00536 SAM domain (Sterile alpha motif)
PF07647 SAM domain (Sterile alpha motif)
PF07653 Variant SH3 domain
PF00018 SH3 domain
PF00498 FHA domain
PF00270 DEAD box
PF07717 Domain of unknown function (DUF1605)
PF00595 PDZ domain (Also known as DHR or GLGF)
PF08740 C8 domain
PF01033 Somatomedin B domain
PF07172 LBP / BPI / CETP family, N–terminal domain
PF06758 Repeat of unknown function (DUF1205)
PF00966 Class II histocompatibility antigen, beta domain
PF00969 Class II histocompatibility antigen, alpha domain
PF00270 DEK motif
PF01660 C8 module
PF01607 Chitin binding Peritrophin–A domain
PF02822 Aristaless family
PF01553 Acyltransferase
PF00270 DEAD box
PF00535 Pre–SET motif
PF00105 Peptide hormone
PF04482 Prenyltransferase and squalene oxidase repeat
PF00058 Low–density lipoprotein receptor repeat class B
PF01363 FYVE zinc finger
PF00916 Sulfate transporter family
PF00532 BRCA C Terminus (BRCT) domain
PF00048 Small cytokines (intercrine/chemokine), interleukin–8 like
PF07690 Major Facilitator SuperFamily, DUF820
PF00616 Regulator of G protein signaling domain
PF00433 Protein kinase C terminal domain
PF00004 ATPase family associated with various cellular activities (AAA)
PF01477 PLAT/LM3 domain
PF00096 Zinc kinuclea
PF00061 Lipocalin / cytosolic fatty–acid binding protein family
PF00270 DEAD box
PF00008 Laminin G–type domain
PF07562 Nine Cysteines Domain of family 3 GPCR
PF00556 RpS5 domain
PF00017 Sh2 domain
PF07546 BM domain
PF04410 LIM domain
PF00270 DEAD box
PF00076 Double–stranded RNA binding motif
PF00096 Zinc finger, ZZ type
PF00625 Guanylate kinase
PF01602 Adaptin N terminal region
PF01826 Trypsin inhibitor like cysteine rich domain
PF06027 GDSL–like Lipase/Acylhydrolase
PF00270 DEAD box conserved region
PF00888 Cullin
PF00566 TBC domain
PF00418 Tau and MAP protein, tubulin–binding repeat
PF06627 UBA/TS–N domain
PF00018 Inositol–trisk–beta N–terminal domain
PF07783 Hormone receptor domain
PF00096 Zinc finger, C2H2 type domain
PF00172 Fhu domain – N–terminal to homeobox domain
PF00096 CXXC zinc finger domain
PF01839 FG–GAP repeat
PF00462 Glutaredoxin
PF00146 Fibilin
PF00748 Caspain inhibitor
PF00168 Protein kinase C conservation Organisation MOdifier) domain
PF00130 Phorbol esters/diacylglycerol binding domain (C1 domain)
PF00435 Spectrin repeat
PF00225 Kinesin motor domain
PF00400 WD domain, G–beta repeat
PF07714 Tetratricopeptide repeat
PF01405 LBM domain
PF00270 DEAD box
PF00867 TUF family
PF00023 Beta/Gamma crystallin
PF00337 Galactoside–binding lectin
PF01454 MAGE family
PF07716 Basic region leucine zipper
PF01437 Plexin repeat
PF08820 Cysteine rich repeat
PF00621 RhoGEF domain
PF00270 DEAD box
PF00023 Ankyrin repeat
PF01605 Lanosphin/CL–1–like GPS domain
PF00002 7 transmembrane receptor (Secretin family)
PF01391 Collagen triple helix repeat (20 copies)
PF00210 NH domain
PF00018 Pertinum family
PF00270 DEAD box family
PF00092 VWC domain
PF00086 Thyroglobulin type–1 repeat
PF01846 9N4 domain–like
PF00084 von Willebrand factor type A domain
PF00415 Regulator of chromosome condensation (RCC1)
PF01390 SEA domain
PF00270 DEAD box
PF00051 Kringle domain
PF00014 Kunitz/Bovine pancreatic trypsin inhibitor domain
PF01846 Actvin type I and II receptor domain
PF08815 LCCL domain
PF00096 Trefoil (P–type) domain
PF00533 Ref homology domain (RHD)
PF00018 Pertinum family domain
PF00102 Extracellular link domain
PF00059 Fibronectin type I domain
PF07654 Immunoglobulin C1–set domain
PF00040 Fibronectin type II domain
PF00030 Class I Histocompatibility antigen, domains alpha 1 and 2
PF02191 Olfactomedin–like domain
PF00008 EGF–like domain (Pfam Accession III and V)
PF00094 WAP–type (Whey Acidic Protein) 'four–disrfide core'
PF00530 Scavenger receptor cysteine–rich domain
PF08205 C2H2–like C2–set immunoglobulin domain
PF00086 von Willebrand factor type C domain
PF07686 GCC2 and GCC3
PF07679 Immunoglobulin I–set domain
PF00018 Pertinum family domain
PF07645 Calcium binding EGF domain
PF00431 CUB domain
PF00057 Low–density lipoprotein receptor domain class A
PF00084 Sushi domain (SCR repeat)
PF00090 Thrombospondin type 1 domain
PF00635 MAM domain
PF00008 EGF–like domain
PF07974 EGF–like domain
PF00041 Fibronectin type III domain
Homo sapiens 0-0
Mus musculus 0-0
Gallus gallus 0-0
Xenopus tropicalis 0-0
Danio rerio 0-0
Strongylocentrotus purpuratus 0-0
Lottia gigantea 0-0
Nematostella vectensis 0-0
Trichoplax adhaerens 0-0
Taiwanofungus camphoratus 0-0
Blastocladiella emersonii 0-0
Fitzroyia cirrosa 0-0
Amphimedon queenslandica 0-0
Homo sapiens 1-0
Mus musculus 1-0
Gallus gallus 1-0
Xenopus tropicalis 1-0
Danio rerio 1-0
Strongylocentrotus purpuratus 1-0
Lottia gigantea 1-0
Nematostella vectensis 1-0
Trichoplax adhaerens 1-0
Taiwanofungus camphoratus 1-0
Blastocladiella emersonii 1-0
Fitzroyia cirrosa 1-0
Amphimedon queenslandica 1-0
Homo sapiens 1-1
Mus musculus 1-1
Gallus gallus 1-1
Xenopus tropicalis 1-1
Danio rerio 1-1
Strongylocentrotus purpuratus 1-1
Lottia gigantea 1-1
Nematostella vectensis 1-1
Trichoplax adhaerens 1-1
Taiwanofungus camphoratus 1-1
Blastocladiella emersonii 1-1
Fitzroyia cirrosa 1-1
Amphimedon queenslandica 1-1
Homo sapiens 1-2
Mus musculus 1-2
Gallus gallus 1-2
Xenopus tropicalis 1-2
Danio rerio 1-2
Strongylocentrotus purpuratus 1-2
Lottia gigantea 1-2
Nematostella vectensis 1-2
Trichoplax adhaerens 1-2
Taiwanofungus camphoratus 1-2
Blastocladiella emersonii 1-2
Fitzroyia cirrosa 1-2
Amphimedon queenslandica 1-2
Homo sapiens 2-0
Mus musculus 2-0
Gallus gallus 2-0
Xenopus tropicalis 2-0
Danio rerio 2-0
Strongylocentrotus purpuratus 2-0
Lottia gigantea 2-0
Nematostella vectensis 2-0
Trichoplax adhaerens 2-0
Taiwanofungus camphoratus 2-0
Blastocladiella emersonii 2-0
Fitzroyia cirrosa 2-0
Amphimedon queenslandica 2-0
Homo sapiens 2-1
Mus musculus 2-1
Gallus gallus 2-1
Xenopus tropicalis 2-1
Danio rerio 2-1
Strongylocentrotus purpuratus 2-1
Lottia gigantea 2-1
Nematostella vectensis 2-1
Trichoplax adhaerens 2-1
Taiwanofungus camphoratus 2-1
Blastocladiella emersonii 2-1
Fitzroyia cirrosa 2-1
Amphimedon queenslandica 2-1
Homo sapiens 2-2
Mus musculus 2-2
Gallus gallus 2-2
Xenopus tropicalis 2-2
Danio rerio 2-2
Strongylocentrotus purpuratus 2-2
Lottia gigantea 2-2
Nematostella vectensis 2-2
Trichoplax adhaerens 2-2
Taiwanofungus camphoratus 2-2
Blastocladiella emersonii 2-2
Fitzroyia cirrosa 2-2
Amphimedon queenslandica 2-2

Anexo D - Termos enriquecidos do Gene Ontology em *H. sapiens*: Processo biológico



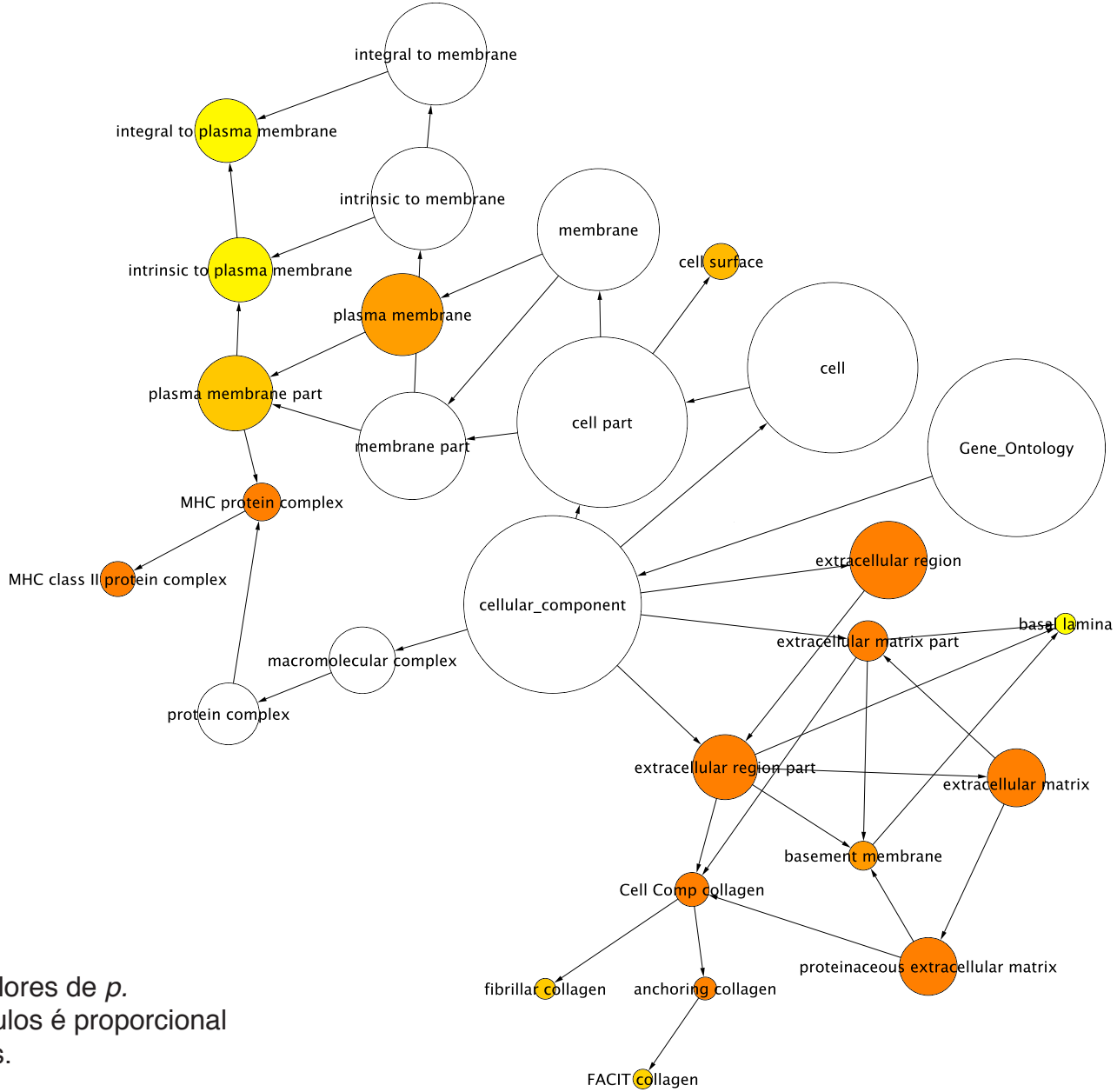
As cores indicam valores de *p*.
O tamanho dos círculos é proporcional ao número de genes.

Anexo E - Termos enriquecidos do Gene Ontology em *H. sapiens*: Função molecular



As cores indicam valores de *p*.
O tamanho dos círculos é proporcional ao número de genes.

Anexo F - Termos enriquecidos do Gene Ontology em *H. sapiens*: Componente celular



As cores indicam valores de *p*.
O tamanho dos círculos é proporcional ao número de genes.