

Universidade de São Paulo  
Programa Interunidades em Bioinformática

Diego Trindade de Souza

Origem de genes recentes, uma abordagem  
com PSSMs deterioradas e arquiteturas de  
domínio proteico

São Paulo,  
2016

Diego Trindade de Souza

Origem de genes recentes, uma abordagem  
com PSSMs deterioradas e arquiteturas de  
domínio proteico

Tese de doutorado apresentada ao Programa de  
Pós-Graduação em Bioinformática da Universidade  
de São Paulo como parte dos requisitos para a  
obtenção do título de Doutor em Bioinformática.

Área de Concentração: Bioinformática.

Orientador: Sergio Russo Matioli.

Processo FAPESP: 2013/06592-7

Universidade de São Paulo  
Programa Interunidades em Bioinformática  
São Paulo, 2016

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada à fonte.

**Dados Internacionais de Catalogação na Publicação (CIP)**

Souza, Diego Trindade De

Origem de genes recentes, uma abordagem com PSSMs deterioradas e arquiteturas de domínio proteico / Diego Trindade de Souza; orientador Sergio Russo Matioli. - São Paulo, 2002.

74 páginas.

Tese (Doutorado) – Programa Interunidades em Bioinformática da Universidade de São Paulo, 2016.

1. Filoestratigrafia. 2. Detecção de novos genes. 3. Bioinformática.

**SOUZA, DT. Origem de genes recentes, uma abordagem com PSSMs deterioradas e arquiteturas de domínio proteico.** Tese de doutorado apresentada ao Programa de Pós-Graduação em Bioinformática da Universidade de São Paulo como parte dos requisitos para a obtenção do título de Doutor em Bioinformática.

Aprovado em: 6 de outubro de 2016.

Banca Examinadora

Nome dos Participantes da Banca	Função	Sigla da CPG	Resultado
Sergio Russo Matioli	Presidente	IB - USP	<u>APROVADO</u>
João Marcelo Pereira Alves	Titular	ICB - USP	<u>APROVADO</u>
João Carlos Setubal	Titular	IQ - USP	<u>APROVADO</u>
Marcelo Ribeiro da Silva Briones	Titular	UNIFESP - Externo	<u>APROVADO</u>
Francisco Prosdocimi	Suplente	UFRJ - Externo	<u>APROVADO</u>

João Marcelo Pereira Alves

João Carlos Setubal

Marcelo Ribeiro da Silva Briones

Francisco Prosdocimi

Sergio Russo Matioli  
Presidente da Comissão Julgadora

Aos meus pais, Francisco e Terezinha, aos meus irmãos, Fabiane e Bruno, e à Juliana com muito carinho.

## Agradecimentos

Ao Prof. Dr. Sergio Russo Matioli, por ter sido um excelente orientador e amigo.

Ao Prof. Dr. José Miguel Ortega, pelo apoio, ensinamentos e amizade.

Aos professores e funcionários do Programa de Pós-Graduação em Bioinformática pela atenção e ensinamentos.

Aos antigos e atuais colegas do Laboratório de Biodados da UFMG e do Laboratório de Bioinformática do Instituto de Biociências da USP pelas discussões científicas e amizade.

Aos colegas e amigos do Programa de Pós-Graduação em Bioinformática, em especial ao Lucas, companheiro de longa data, pelas discussões científicas e, principalmente, amizade.

Às agências de fomento: FINEP, CNPq, CAPES e FAPESP pelo apoio financeiro e bolsa de estudo que viabilizou o desenvolvimento deste trabalho.

A todos aqueles, que participaram direta ou indiretamente da confecção deste trabalho e que, por uma falha, não se encontram citados aqui.

Muito Obrigado!

## Resumo

SOUZA, DT. Origem de genes recentes, uma abordagem com PSSMs deterioradas e arquiteturas de domínio proteico. 2016. Tese (Doutorado) – Programa Interunidades em Bioinformática da Universidade de São Paulo, 2016.

A origem dos novos genes é um processo importante para a evolução dos organismos, pois ela fornece fontes singulares para a inovação evolutiva. As abordagens que mostram como esses novos genes surgem e adquirem novas funções no curso da evolução são de fundamental importância, por exemplo, elas podem ajudar a correlacionar mutações com alterações metabólicas, fisiológicas e/ou morfológicas, indicando quais mutações podem ser importantes funcionalmente. Recentemente, uma nova abordagem, nomeada de filoestratigrafia, foi desenvolvida para estabelecer origem evolutiva dos genes. Neste método a emergência de novas sequências de um nó filogenético particular em uma linhagem ancestral-descente é inferida geralmente utilizando algoritmos de similaridade. No presente trabalho, nós fizemos uma pesquisa filoestratigráfica de dois bancos de dados de domínios proteicos, CATH e Pfam, para todas as entradas humanas descrevemos a origem dos domínios e arquiteturas humanas. Também realizamos uma nova abordagem para refinar os resultados por Male-PSI-BLAST, em um estudo de caso dos domínios príons e ADHs. A análise das duas bases de dados mostrou que existiram três períodos importantes de aparecimento de domínios proteicos humanos: a origem do organismo celular, Eucarioto e Euteleostomi, nos quais há um elevado número de surgimento de novos genes na linhagem ancestral-descente de humanos. Quando analisamos o aparecimento de arquiteturas, elas são evidentemente mais recentes que o aparecimento de domínios, embora, em seu conteúdo, possa haver domínios muito antigos misturados com domínios novos. Não notamos nenhuma tendência de acréscimo de novos domínios para arquiteturas que compreendem domínios antigos ou recentes. Para medir o grau de versatilidade de domínio, nós utilizamos a frequência ponderada de bigrama, uma combinação específica de dois domínios adjacentes. O teste de correlação de *Spearman* mostrou que existe uma baixa correlação negativa entre a idade de domínios e índices de versatilidade. Em um estudo de caso, demonstramos que é possível caracterizar descontinuidades evolutivas nos resultados de Male-PSI-BLAST entre domínios que surgiram a partir de outros. Pela primeira vez, a origem de todos os domínios e arquiteturas proteicas presentes nas bases de dados estudadas foi descrita, fornecendo um cenário evolutivo que pode ser mais explorado a partir das abordagens aqui desenvolvidas.

## Abstract

SOUZA, DT. Origin of recent genes, an approach with deteriorated PSSMs and protein domain architectures. 2016. Tese (Doutorado) – Programa Interunidades em Bioinformática da Universidade de São Paulo, 2016.

The origin of new genes is an important process for the evolution of organisms because they provide singular sources for evolutionary innovation. The approaches that show how these new genes arise and acquire new functions in the course of evolution are of fundamental importance: they can help to correlate mutations with metabolic, physiological, and morphological changes, indicating which mutations are likely to be functionally important. Recently, a new approach, named phylostratigraphy, was developed to establish the evolutionary origin of the genes. In this method the emergence of novel sequences at a particular phylogenetic node in a descendent-ancestor lineage is infer usually by using the similarity search algorithm. In the present work, we did a phylostratigraphical search of two protein domain databases, CATH and Pfam, for all human entries and depicted the origin of human domains and architectures. We also conducted a new approach to refine results by Male-PSI-BLAST in a case study of prions and ADH's domains. The analysis of two databases showed that there are three important periods of appearance of human gene domains: the origin of cellular organism, Eukaryote, and Euteleostomi appear to be important for production of new genes at the ancestor-descendent lineages that lead to the human species. However, when we analyze the appearance of architectures, they are by far more recent than the appearance of domains, although they might contain very ancient domains mixed with recent ones. We did not notice a bias of addition of new domains to architectures comprising either ancient or recent domains. To measure the degree of domain versatility, we used the weighted bigram frequency, where bigram is defined as a specific combination of two adjacent domains. The Spearman correlation test showed that there is a low negative correlation between the age of domains and versatility indexes. In the study of case, we have demonstrated that it is possible to characterize evolutionary ruptures in results of Male-PSI-BLAST between domains that emerged from others. For the first time the origin of all protein domains and architectures was depicted, providing an evolutionary scenario that can be further explored.

## Sumário

1.- Introdução.....	10
1.1.- Origem dos genes .....	10
1.2.- Homologia .....	15
1.3.- Domínios e arquiteturas.....	20
1.4.- Filoestratigrafia.....	21
2.- Objetivos .....	25
3.- Material e métodos .....	26
4.- Resultados .....	32
4.1.- Análise filoestratigráfica das entradas CATH.....	32
4.2.- Análise filoestratigráfica das entradas Pfam.....	44
4.3.- Estudo de caso: príons e ADH's .....	49
5.- Discussão.....	52
5.1.- Análise filoestratigráfica das entradas CATH.....	52
5.2.- Análise filoestratigráfica das entradas Pfam.....	56
5.3.- Estudo de caso: príons e ADHs .....	59
6.- Conclusões .....	61
7.- Referências bibliográficas .....	62

## 1.- Introdução

### 1.1.- Origem dos genes

Um dos maiores desafios na era pós-genômica é a compreensão da origem evolutiva dos genomas. As diferenças na quantidade de genes entre os organismos existentes revelam que a origem de novos genes é um fenômeno biológico geral (ZHOU & WANG, 2008). O nascimento de novos genes com novas funções, desde a década de 1930, tem sido considerado como um importante fator que contribui para a inovação na evolução (KAESSMANN, 2010). Novos genes são as fontes de inovações genéticas nos genomas e a origem destes genes tem sido atribuída a uma variedade de processos no nível molecular (CARDOSO-MOREIRA & LONG, 2012; KAESSMANN 2010; ZHOU & WANG, 2008; LONG *et al.*, 2003). A seguir, será feita uma breve revisão das principais características dos processos envolvidos na origem de novos genes.

**a.- Duplicação gênica.** Os primeiros trabalhos que propuseram a existência do mecanismo da duplicação gênica na origem de novos genes foram desenvolvidos na década de 1930 por Haldane (1933) e Muller (1935). Entretanto, foi a partir de uma revisão produzida por Ohno (1970) que o papel da duplicação gênica na evolução foi mais amplamente discutido e aceito. A partir daí, esse mecanismo começou a ser considerado como o principal processo gerador de novidades gênicas.

A presença de uma segunda cópia de um gene é um evento singular para a evolução do genoma, já que permite que um dos genes duplicados mantenha a função original, enquanto o outro possa acumular mutações que ocasionalmente podem gerar uma nova função (LONG, 2001). O modelo proposto por Ohno (1973) considera que a maior parte dos genes originados pelo mecanismo de duplicação poderia rapidamente acumular mutações, o que acarretaria em pseudogenização, ou seja, na formação de cópias não funcionais de genes ativos. Entretanto, atualmente é conhecido que, mesmo que uma parte dos genes originados por duplicação tenham sofrido pseudogenização, vários deles podem ter se fixado e mantido sua funcionalidade através de processos, tais como: neofuncionalização, subfuncionalização ou evolução de redundância (ZHOU & WANG, 2008).

A neofuncionalização acontece quando um gene ganha uma nova função, enquanto a outra cópia mantém a função ancestral. Nesta situação, a função inicial de um gene é assegurada por um dos exemplares e outra cópia pode ser submetida a uma seleção purificadora mais relaxada e, portanto, acumular mutações que em outras condições seriam

eliminadas na totalidade (ZHOU & WANG, 2008). A subfuncionalização se dá quando a função de um gene ancestral é dividida entre as cópias duplicadas, principalmente pela degeneração complementar dos elementos regulatórios (LYNCH & FORCE, 2000). Neste caso, a união das atividades e padrões de expressão dos dois genes duplicados pode resultar na função original do gene ancestral. Por fim, a evolução por redundância acontece quando o aumento na quantidade do gene causada por uma duplicação é favorecido por seleção natural e, portanto, as duas cópias do gene, que executam funções iguais, coexistem inalteradas no genoma.

A duplicação pode ser parcial ou completa de um gene, ou mesmo de vários genes em um segmento do genoma, mas a duplicação também pode ser gerada por um evento maior em que há polissomia ou poliploidização (DING *et al.*, 2012). Um ponto a ser destacado é que a natureza da duplicação, seja ela parcial ou completa, pode influenciar o futuro do gene originado. A duplicação parcial de um gene cujo produto apresenta muitos parceiros de interação, por exemplo, genes pertencentes a uma via de sinalização ou reguladores transpcionais, poderia perturbar a estequiometria do processo e provavelmente a duplicação seria selecionada contra. Ao contrário, na duplicação de todo o genoma, o equilíbrio na quantidade dos genes seria mantido, assim como suas relações dentro da rede gênica, por conseguinte, a duplicação poderia ser retida (HUFTON *et al.*, 2009).

**b.- Embaralhamento de éxons.** Outro mecanismo gerador de novos genes é o embaralhamento de éxons (*exon shuffling*, em inglês), processo que origina novas combinações de domínios a partir de estruturas mais simples. Neste mecanismo, dois ou mais éxons de diferentes genes podem ser reunidos ectopicamente, recombinação entre genes não homólogos, ou o mesmo o éxon pode ser duplicado, para criar uma estrutura éxon-ítron diferente (ITOH *et al.*, 2007).

Este mecanismo foi proposto inicialmente por Walter Gilbert (1978) e atualmente é considerado como uma das principais forças evolutivas inovadoras para os genomas e proteomas dos eucariotos. Estima-se que cerca de 19% dos éxons de genes eucarióticos podem ter sido formados por embaralhamento de éxons (LONG *et al.*, 2003).

Os mesmos mecanismos que produzem a duplicação parcial do gene também podem gerar recombinação ectópica dos éxons. Consequentemente, o embaralhamento de éxons parece ter uma grande importância na evolução de genomas complexos, uma vez que estes contêm muitos ítrons e elementos repetitivos que facilitam os eventos de recombinação intrônicas (KEREN *et al.*, 2010).

**c.- Elementos transponíveis.** Este mecanismo ocorre quando uma sequência de DNA altera a posição dentro do genoma, em alguns casos criando ou invertendo mutações ou alterando o tamanho do genoma na célula (WESSLER, 2006). Desde a descoberta, em 1950, destes "genes saltadores", pela ganhadora do Prêmio Nobel Barbara McClintock, o conhecimento e compreensão dos elementos móveis aumentaram drasticamente na comunidade científica. A atual visão considera os elementos transponíveis como um dos mais importantes mecanismos com papel relevante na evolução por promover e controlar processos tais como, rearranjos cromossômicos, expressão de genes, adaptação populacional e especiação (BIÉMONT, 2010).

A classificação, atual e mais aceita, dos elementos transponíveis os divide em duas classes, dependendo da natureza da transposição intermediária: Transposons de Classe-I ou *retrotransposons* que usam o RNA como intermediário e a transcrição reversa em uma maneira de "copiar e colar". E Transposons de Classe-II ou *DNA Transposons* que usam o DNA como intermediário, em um mecanismo do tipo "cortar e colar". Estas classes ainda podem ser divididas com base no mecanismo de transposição, semelhança de sequência ou estrutura para incluir novas subclasse que possam surgir em uma nomenclatura sugerida por Wicker e seu grupo (2007) e Kapitonov & Jurka (2008).

Os elementos transponíveis podem ainda ser classificados como autônomos, se eles codificam as proteínas necessárias para o processo de transcrição, e não autônomos quando não possuem as ORFs necessárias para este mecanismo, sendo necessária a presença de outros elementos transponíveis para sua mobilização (BIÉMONT, 2010). Independentemente de qual seja o mecanismo, o resultado é uma estrutura químérica, seja por ser uma região de codificação transposta com uma nova sequência reguladora ou por ser uma região de codificação transposta com um novo fragmento de proteína que é recrutada do sítio alvo, fazendo que esta estrutura tenha uma função diferente daquela do gene parental (LONG *et al.*, 2003).

**d.- Transferência lateral de genes.** A transferência lateral de genes é descrita pela transferência não sexual da informação genética entre organismos diferentes ou entre organelas e núcleo (DING *et al.*, 2012). Este mecanismo é comum em bactérias e tido como um dos principais mecanismos para evolução dos procariotos (FROST *et al.*, 2005). Apesar da transferência lateral de gene resultar na troca de genes homólogos, há evidências de que ela também pode atuar no recrutamento de novos genes e consequentemente fornecer novos fenótipos (LONG *et al.*, 2003).

A transferência lateral de gene foi descrita pela primeira vez por Freeman (1951) em uma publicação que demonstrou a transferência de um gene viral em *Corynebacterium diphtheriae* resultou em virulência a partir de uma cepa não virulenta (FREEMAN, 1951). Oito anos depois no Japão, Ochiai e seu grupo (1959) demonstraram a transferência da resistência aos antibióticos entre espécies diferentes de bactérias. E em meados da década de 1980, foi sugerido que este mecanismo poderia estar envolvido na formação da história evolutiva desde o início da vida na Terra (SYVANEN, 1985).

A importância da transferência lateral de genes na evolução de bactérias foi estabelecida há muito tempo (BOUCHER et al., 2003). Este mecanismo também tem sido documentado frequentemente em organismos fagocíticos e eucariotos parasitários unicelulares (KEELING & PALMER, 2008). Em animais e plantas, a transferência lateral de genes parece estar limitada a eventos associados com endossimbiose ou parasitismo (HOTOPP et al., 2007).

**e.- Origem *de novo*.** A origem *de novo* é o mecanismo pelo qual novos genes surgem a partir de sequências não codificantes do DNA. A origem *de novo* de genes codificadores de proteínas inteiras foi considerada por muito tempo improvável. Em uma influente revisão, François Jacob (1977) afirmou que a probabilidade de uma proteína funcional originar pelo mecanismo *de novo* por associação aleatória de aminoácidos é quase nula e, por conseguinte, a origem de novas sequências nucleotídicas não poderia ter qualquer importância na produção de novas informações.

Muito embora a verdadeira origem *de novo* de novas sequências de genes a partir de regiões supostamente não codificadoras seja rara, pesquisas conseguiram identificar alguns genes candidatos a terem sido originados dessa forma em *Drosophila* (LEVINE et al., 2006; ZHOU et al., 2008) e leveduras (CAI et al., 2008). O fato é que a identificação dos genes originados pelo mecanismo *de novo*, é em geral difícil. A principal razão é que os novos genes e/ou o seu DNA não codificante ancestral correspondente são susceptíveis de evoluir rapidamente, devido à seleção direcional e à falta de constrangimento funcional, respectivamente. Consequentemente isso dificulta a detecção de similaridade entre as sequências. Além disso, sempre há a possibilidade de que o gene candidato ter sido originado pelo mecanismo *de novo*, tenha sido transferido a partir de um organismo cuja linhagem foi extinta e que, portanto, não há a possibilidade desse fato ter sido documentado. A tabela 1 resume os principais processos pelos quais se originam novos genes.

Tabela 1 | Mecanismo molecular para a surgimento de novos genes\*

Mecanismo	Processo	Exemplos	Comentários
Embaralhamento de éxon <sup>1</sup> : recombinação ectópica dos éxons e domínios a partir de genes distintos		AFGPs, BC1RNA, BC200RNA	~19% dos exons nos genes de eucariotos têm sido formados por embaralhamento de exons
Duplicação de genes <sup>2</sup> : modelo clássico de duplicação com divergência		CGβ, Cid, RNASE1B	Muitos genes duplicados provavelmente desenvolveram novas funções
Retroposição <sup>3</sup> : novos genes duplicados são criados em novas posições genômicas por transcrição reversa ou por outros processos		PGAM3, Pgk2, PMCHL1, PMCHL2, Sphinx	1% de DNA humano é retroposto para novas localizações genômicas
Elemento móvel <sup>4</sup> : um elemento móvel a sequência é diretamente recrutados por genes hospedeiros		HLA-DR-1, human DAF, lungkerine mRNA, mNSC1 mRNA	Gera 4% de novos exons nos genes que codificam proteínas humanas
Transferência lateral de genes <sup>5</sup> : um gene é horizontalmente transmitidos entre os organismos		acylneuraminate lysase, Escherichia coli mutU and mutS	Na maioria das vezes relatado em procariotas e recentemente relatados em plantas
Gene de fusão/ cisão <sup>6</sup> : dois genes adjacentes se fundem em um único gene ou um gene se divide em dois genes		Fatty-acid synthesis enzymes, Kua-UEV, Sdic	Envolvido na formação de ~ 0,5% de genes procarióticos
Origem <i>De novo</i> <sup>7</sup> : uma região codificadora origina a partir de uma região genômica que anteriormente não era codificante		AFGPs, BC1RNA, BC200RNA	Raro para originação de um gene; pode não ser raro para origem parcial dos genes

AFGP, antifreeze glycoprotein; CGβ, chorionic gonadotropin β polypeptide; Cid, centromere identifier; DAF, decay-accelerating factor; HLA-DR-1, major histocompatibility complex DR1; PGAM3, phosphoglycerate mutase 3; Pgk2, phosphoglycerate kinase 2; PMCHL, pro-melanin-concentrating hormone-like; RNASE, ribonuclease; Sdic, sperm-specific dynein intermediate chain; UEV, tumour susceptibility gene. \*Tabela adaptada de Long et al (2003). <sup>1</sup>Patthy, 1996; Long e Langley, 1993; Paulding et al, 2003; Patthy, 1995; Long et al, 1995; Gilbert et al, 1997; Long, 2001; Long et al, 1998; Javaud et al, 2003. <sup>2</sup>Ohno, 1970; Kimura, 1983; Prince e Pickett, 2002; Zhang et al, 2002; Maston e Ruvolo, 2002; Henikoff et al, 2001; Malik e Henikoff, 2001; Hughes, 2000. <sup>3</sup>Wang et al, 2002; Betrán e Long, 2002a; Betrán e Long, 2002b; Pickeral et al, 2000; McCarrey, 1987; McCarrey, 1990; McCarrey, 1994; Courseaux e Nahon, 2001; Goodier et al, 2000; Brosius, 1999; Brosius, 2003. <sup>4</sup>Nekrutenko e Li, 2001; Sorek et al, 2002; Makalowski, 2000; Lorenc e Makalowski, 2003. <sup>5</sup>Ochman, 2001; de Koning et al, 2000; Bergthorsson et al, 2003; Ragan, 2001. <sup>6</sup>Thomson et al, 2000; Nurminsky et al, 1998; Ranz et al, 2003; McCarthy e Hardie, 1984; Snel et al, 2000. <sup>7</sup>Chen et al, 1997a; Chen et al, 1997b; Martignetti e Brosius, 1993a; Martignetti e Brosius, 1993b.

## 1.2.- Homologia

A homologia é um conceito fundamental da Biologia. Historicamente, o conceito de homologia modificou-se profundamente com a aceitação universal da evolução biológica dentro do meio científico. O termo homologia foi inicialmente cunhado por Richard Owen (1848), para se referir a estruturas similares de organismos diferentes baseado no conceito de "arquétipo", ou plano estrutural do corpo desses organismos. Posteriormente, o termo "homologia" foi redefinido com base na evolução biológica para designar a mesma estrutura que estava presente no ancestral comum a esses organismos. Assim, atualmente nos referimos à homologia de monômeros em uma macromolécula como o estabelecimento da hipótese de que dois monômeros em duas sequências macromoleculares distintas foram o mesmo monômero de uma macromolécula de um indivíduo ancestral comum aos indivíduos que possuem as sequências que estão sendo analisadas.

Entretanto esta definição não especifica o cenário da relação de homologia, por isso Walter Fitch (1970) introduziu dois novos termos: ortologia e paralogia. As sequências homólogas são ortólogas se são inferidas por descendência de um mesmo ancestral comum separado por um evento de especiação. Quando uma espécie diverge para duas espécies separadas, as cópias de um único gene nas duas espécies resultantes são referidas como sendo ortólogas. Contrariamente, sequências homólogas são parálogas se elas originaram por um evento de duplicação dentro do genoma. Para eventos de duplicação de genes, se um gene num organismo é duplicado para ocupar duas posições diferentes no mesmo genoma, então as duas cópias são parálogas.

A distinção entre ortólogos e parálogos é importante, uma vez que os dois tipos de homologia têm diferentes implicações funcionais. Genes ortólogos, geralmente apresentam funções semelhantes entre os organismos, ainda que existam exceções a esta regra. Por outro lado, parálogos tendem a executar diferentes funções biológicas, já que o evento de duplicação normalmente permite que uma das novas cópias evoluir sem restrição enquanto a outra cópia mantém as funções originais. A equivalência funcional de ortólogos é muito útil, uma vez que pode ser utilizada para anotação da função do gene, extrapolando a função dos genes pertencentes a diferentes organismos. Parálogos, ao contrário, podem ser utilizados para estudar inovação. Além disso, esta distinção entre os homólogos é essencial para a inferência filogenética, que requer ortólogos, mas não parálogos (ALTENHOFF & DESSIMOZ, 2012).

Estes termos simples foram estendidos para explicar cenários evolutivos mais complexos. Coortólogos são genes duplicados seguidos de especiação que, além disso,

originam dois ou mais genes que são coletivamente ortólogos de um ou mais genes em outra linhagem (SYVANEN, 1985). Parálogos também podem ser divididos em *outparalogs* e *inparalogs*, que poderiam ser traduzidos como exoparálogos e endoparálogos, e são definidos como genes que sofreram duplicação antes e depois do processo de especiação, respectivamente (REMM et al, 2001; ALEXEYENKO et al, 2006). Existem ainda cenários evolutivos mais complicados, tais como: pseudo-ortólogos, que são referidos como exoparálogos que foram incorretamente considerados como ortólogos devido a uma perda diferencial na linhagem dos organismos analisados (ALTENHOFF & DESSIMOZ, 2012); xenólogos e pseudoparálogos que ocorrem por um evento de transferência lateral de gene com e sem deslocamento do gene inicial, respectivamente (KOONIN, 2005); "epactólogos" ("*epaktologs*") são proteínas que estão relacionadas apenas através de aquisição do mesmo tipo de domínios móveis (NAGY et al., 2011); Há ainda os "ohnólogos" ("*ohnologs*", em homenagem a Ohno) que são definidos como genes parálogos originados por eventos de duplicação de todo o genoma (WOLFE, 2000).

Os parágrafos acima mostram o quanto importante é a homologia para os estudos em Biologia e suas diferentes subdivisões para o entendimento do cenário evolutivo ao qual uma determinada sequência possa ter passado. Para que possam ser atribuídos os processos evolutivos que contribuíram na organização atual de um gene, é necessário que as comparações entre as sequências macromoleculares possam indicar possíveis indícios que sustentem a hipótese da homologia.

No campo da bioinformática, vários programas e algoritmos foram criados para hipotetizar homologias entre sequências de aminoácidos ou nucleotídeos baseados em similaridades entre as sequências. Embora a homologia e a similaridade sejam conceitos distintos, somente se pode supor a homologia de estruturas baseando-se na sua similaridade. Para que se possa estimar a similaridade entre sequências macromoleculares, normalmente é necessário que se realize o alinhamento entre as sequências, que é o estabelecimento da hipótese de homologia para cada sítio das sequências consideradas. Existem dois tipos de alinhamento: global e local. Alinhadores globais tentam alinhar cada resíduo presente nas sequências e são mais úteis quando as sequências presentes no conjunto analisado tem tamanho aproximado e são semelhantes. Os alinhadores locais, por sua vez, são adequados para sequências diferentes que contenham regiões de semelhança ou motivos de sequências semelhantes dentro do contexto da sequência maior.

Os alinhadores também podem ser classificados como par-a-par ou múltiplo. Em alinhadores par-a-par a busca só pode ser realizada entre duas sequências de cada vez, mas eles são eficientes para calcular e são muitas vezes utilizados para métodos que não exigem extrema precisão. Alinhador de múltiplas sequências é uma extensão dos alinhadores par-a-par, já que incorporam mais de duas sequências por vez. Vários métodos de alinhamento tentam alinhar todas as sequências de um determinado conjunto. Alinhamentos múltiplos são muitas vezes utilizados na identificação de regiões de sequência conservada entre um grupo de sequências em que o pesquisador quer testar a hipótese de ser evolutivamente relacionada. Abaixo será feita uma breve revisão de alguns dos mais importantes alinhadores comumente utilizados.

**a.- BLAST.** É, de longe, a técnica mais amplamente utilizada para se detectar similaridade entre sequências primárias de proteínas ou de DNA. BLAST é o acrônimo de *Basic Local Alignment Search Tool* (ALTSCHUL et al, 1990). A busca com BLAST permite ao pesquisador comparar uma sequência de consulta ("query", em inglês, que passaremos a traduzir como "quesito" ou "sequência quesito") com um banco de dados de sequências e identificar as sequências presentes neste banco de dados que se assemelham com a sequência quesito acima de um limiar de similaridade.

A busca se inicia com um pequeno subconjunto de letras obtidas a partir da sequência quesito, conhecidas como palavra consulta. Em seguida, o BLAST procura não somente as palavras consultas obtidas a partir da sequência quesito, mas também as palavras consultas relacionadas em que substituições conservativas possam ter sido introduzidas, pois estas correspondências podem ser biologicamente informativas. As matrizes de pontuação são utilizadas para determinar quais palavras consultas são relacionadas, estas palavras relacionadas são chamadas de vizinhança ou palavras vizinhas. Neste caso, a matriz de pontuação mais comumente utilizada é BLOSUM62. As palavras consultas relacionadas devem satisfazer o requisito de pontuação acima do limiar de pontuação da vizinhança, chamado de T.

O alinhamento é iniciado com as palavras que têm pontuação acima do limiar T, em ambas as direções, registrando a pontuação acumulativa resultante de correspondência, não correspondência e lacunas, até o alinhamento local de tamanho máximo. Em seguida, o BLAST lista todos os pares de segmentos em que os pontos obtidos sejam maiores que o ponto de corte S. O alinhamento resultante é chamado de pares de segmentos de alta pontuação ou HSP (sigla em inglês de *high-scoring segment pairs*). Uma vez que o algoritmo

do BLAST percorre sistematicamente toda a sequência, utilizando todas as possíveis palavras consultas, é possível que mais de uma HSP seja encontrada para quaisquer pares de sequências.

De posse das HSPs identificadas, o próximo passo é determinar se o alinhamento resultante tem valor relevante estatisticamente. Para isso, o BLAST calcula o valor de expectativa (*E-value*) para avaliar valor de confiança estatística, este cálculo representa o número de vezes que uma HSP poderia ser encontrada puramente ao acaso. Colocando de outra forma, o *E-value* fornece uma medida que indica se a HSP obtida é um falso positivo. Baixos valores de *E-value* implicam maiores probabilidades de significado biológico.

Em alguns casos, o BLAST pode encontrar duas ou mais regiões HSPs que podem ser combinadas para formar um alinhamento maior, neste caso lacunas podem ser introduzidas para formar um alinhamento maior. A avaliação da significância das HSPs combinadas pode ser feita pelo método de Poisson ou da soma dos pontos. Por fim, quando o *E-value* para uma sequência similar encontrada no banco de dados satisfaz o limiar selecionado pelo usuário, o resultado é então reportado.

**b.- PSI-BLAST.** A variação do algoritmo BLAST conhecida como PSI-BLAST (sigla em inglês de *Position-Specific Iterative BLAST*) é adequada para identificar proteínas distorcamente relacionadas, alinhando sequências mesmo que haja pouca similaridade entre as sequências comparadas (ALTSCHUL et al., 1997). O PSI-BLAST utiliza matrizes de pontuação de posição específica ou PSSM (sigla em inglês, *Position Specific Scoring Matrix*) que são representações numéricas derivadas de um alinhamento múltiplo e são utilizadas para encontrar sequências pouco similares. Nela a pontuação de substituição de aminoácidos é indicada separadamente para cada posição em um alinhamento múltiplo de sequências de proteína. Isso quer dizer, que uma substituição de Tyr-Trp na posição A de um alinhamento pode receber uma pontuação muito diferente do que a mesma substituição na posição B. Isto contrasta com as matrizes de substituição independentes de posição, tais como PAM e BLOSUM em que a mesma substituição de Tyr-Trp recebe a mesma pontuação não importando a posição que ela ocorra.

Em PSSMs, geralmente são mostrados valores inteiros positivos ou negativos. Pontuações positivas indicam que uma dada substituição de um aminoácido é mais frequente no alinhamento esperado ao acaso. Valores negativos indicam que a substituição ocorre com menos frequência do que o esperado. Valores positivos elevados indicam resíduos funcionais

críticos, que podem ser resíduos de sítio ativos ou resíduos necessários para outras interações intermoleculares.

Então, a PSSM obtida a partir de um alinhamento múltiplo de uma família proteica conterá a informação que representa as características comuns a este determinado conjunto de sequência. Os números na PSSM representam o alinhamento múltiplo que, por sua vez, refletem as probabilidades de um dado aminoácido ocorrer em cada posição na sequência. Além disso, a PSSM reflete também o efeito conservativo ou não conservativo das substituições em cada posição no alinhamento. Deste modo, ao utilizar a PSSM, o PSI-BLAST faz uso das características comuns entre as sequências que estão embutidas na matriz para procurar por sequências com pouca similaridade, permitindo a identificação de sequências distantes, mas que possivelmente podem estar relacionadas.

O funcionamento do BLAST e do PSI-BLAST tem várias características em comum. O PSI-BLAST inicia-se realizando BLASTp com a sequência quesito, como descrito no tópico anterior. Essa busca resulta em um conjunto de sequências que tem o *E-value* abaixo do limiar determinado pelo usuário. Este conjunto de sequências juntamente com a sequência quesito é utilizado para construir a PSSM de forma automática. A PSSM passa então a ser utilizada como consulta para busca no banco de dados de sequências. O PSI-BLAST neste momento utiliza as características coletivas das sequências para realizar novas buscas. E a cada iteração do PSI-BLAST novas sequências podem encontradas e, assim, incorporadas ao resultado fazendo com que a PSSM modifique as probabilidades contidas a cada iteração. O processo do PSI-BLAST continua pelo número de iteração limite determinado pelo usuário ou até que a procura convirja, o que significa que nenhuma sequência nova foi encontrada.

**b.- HMMER.** HMMER é um pacote de softwares livre comumente utilizado para a análise de sequências (EDDY, 1998). Este programa pode ser usado para substituir o BLAST e PSI-BLAST para pesquisar em banco de dados de proteínas com uma única sequência quesito. Devido à natureza dos modelos probabilísticos subjacentes, o HMMER tem a habilidade de detectar homólogos mais distanamente relacionados com a sequência quesito. HMMER compara a sequência quesito com um perfil em que é atribuído um sistema de pontuação específico para todas as combinações possíveis de lacunas, correspondências e não correspondências, para determinar o alinhamento múltiplo de sequência mais provável. Este perfil é um modelo probabilístico chamado de perfil de modelos ocultos de Markov ou perfil-HMM (sigla em inglês de *hidden Markov models*).

A vantagem de usar HMMs é que eles têm uma base probabilística formal e esta base probabilística permite fazer coisas que outros métodos heurísticos não fazem facilmente. Uma das características mais importantes é que os HMMs têm uma teoria consistente para definir pontuações de inserções, deleções e substituições. Este é um dos pontos que diferencia os HMMs das PSSMs que utilizam penalidades arbitrárias para abertura e extensão de uma lacuna.

Nos termos de um HMM típico, os estados observados são as colunas de alinhamento individuais e os estados ocultos representam a sequência ancestral presumida a partir da qual as demais sequências teriam se originado. O algoritmo do HMM não apenas calcula um alinhamento melhor pontuado, mas sim uma soma de probabilidades obtidas ao longo da montagem do alinhamento. Para o HMMER, as sequências consultas são consideradas similares às sequências que produziram o perfil-HMM, desde que obtenham pontuação significativamente melhor ao perfil-HMM em comparação com o modelo nulo.

### **1.3.- Domínios e arquiteturas**

Domínios de proteínas são unidades estruturais, funcionais e evolutivas das proteínas e as arquiteturas de domínios proteicos (ADPs) são os arranjos lineares dos domínios em proteínas individuais (ZHANG *et al*, 2012). As ADPs, por sua vez, podem ser simples ou compostas, dependendo da quantidade de domínios presentes (FONG *et al*, 2007; TEICHMANN *et al*, 1998). Existem métodos computacionais que permitem estabelecer a homologia através do reconhecimento das ADPs. Alguns agrupamentos de genes homólogos reúnem proteínas com ADPs idênticas ou semelhantes, enquanto que proteínas que possuem ADPs distintas não são normalmente agrupadas (ZHANG *et al*, 2012).

Estudos têm conseguido identificar, por alinhamento de sequências, a presença de eventos de embaralhamento de exons, mecanismo que pode gerar os embaralhamentos das ADPs, em vários organismos (FRANÇA *et al.*, 2012; AL-BALOOL *et al.*, 2011). Entretanto, até o que sabemos no momento, ainda não se investigou sistematicamente se genes recentes repetem arquiteturas ou topologias de domínios já existentes em genes ancestrais, ou quais delas são originadas pelo reagrupamento *de novo* dos domínios pré-existentes. A presença ou ausência de ADPs nos proteomas de 174 organismos cujas sequências genômicas são conhecidas foi utilizada com sucesso para a reconstrução filogenética dos organismos analisados (YANG *et al*, 2005). Nesse estudo, no entanto, a presença ou ausência dos domínios foi verificada independentemente das ADPs existentes.

Existem atualmente vários bancos de dados (BD) em que são mostradas as relações de domínios e arquiteturas. Entretanto, nenhuma delas, pelo que sabemos, traz as possíveis relações de ancestralidade e descendência entre essas estruturas. Dentre essas bases de dados destacaremos abaixo algumas principais.

O Pfam é um banco de dados de famílias de proteínas conservadas e domínios no qual cada um deles é representado por um conjunto de alinhamentos múltiplos de sequência por HMM (FINN et al, 2010).

O BD CATH utiliza uma classificação hierárquica de domínios de proteínas baseada no conteúdo das estruturas das proteínas e tem sido produzido e mantido por uma curadoria realizada caso a caso por especialistas, auxiliada por algoritmos de classificação e previsão, tais como a comparação estrutural e métodos baseados em HMM (SILLITOE et al, 2013).

O BD de domínios conservados (sigle em inglês, CDD) consiste em uma coleção bem anotada de alinhamentos múltiplos de sequências que estão disponíveis como PSSMs para rápida identificação de domínios conservados em sequências de proteínas através de RPS-BLAST (“*Reversed Position Specific BLAST*”) (MARCHLER-BAUER et al, 2013).

O BD InterPro integra modelos preditivos, ou assinaturas, a partir de vários repositórios, no qual o objetivo da plataforma é “combinar dados obtidos de vários BD para fornecer um único recurso através do qual os cientistas podem acessar informações abrangentes sobre famílias de proteínas, domínios e locais funcionais” (HUNTER et al, 2012).

A ferramenta de busca CDART realiza buscas baseadas nas arquiteturas de domínios. Nesta ferramenta, a arquitetura é definida como a ordem sequencial de domínios conservados nas proteínas, sendo que seu algoritmo encontra semelhanças significativas de proteínas através de distâncias evolutivas utilizando perfis de domínios proteicos, a qual é mais sensível que a simples similaridade entre sequências (GEER et al, 2002).

#### **1.4.- Filoestratigrafia**

Em 2007, foi introduzido o termo filoestratigrafia para se referir ao método para datar a emergência de genes ou famílias gênicas (DOMAZET-LOSO *et al.* 2007), apesar de o método preceder o termo e também ter sido usado para abordar uma variedade de questões. A abordagem filoestratigráfica tipicamente utiliza métodos que tentam inferir homologia, sendo o BLAST o método mais comumente utilizado, para buscar por espécies que contenham determinada sequência (CAPRA *et al.*, 2013). Nesta abordagem são selecionados grupos

taxonômicos ascendentes a partir de uma espécie focal procurando para cada gene o táxon mais antigo o qual existem homólogos à sequência quesito (DOMAZET-LOSO *et al.* 2007). Em princípio, a sensibilidade desta abordagem na detecção de outros eventos para origem de gene (por exemplo, transferência horizontal de genes) depende da sua aplicação (ou seja, que genomas estão sendo comparados) e parametrização (por exemplo, os critérios de pesquisa BLAST).

As análises filoestratigráficas têm mostrado que, em comparação com genes mais antigos, os genes recentes: evoluem mais rápido (ALBÀ & CASTRESANA, 2005), têm menor expressão (WOLF *et al.* 2009; CAI & PETROV, 2010), codificam proteínas mais curtas (WOLF *et al.* 2009), estão sujeitos a fraca seleção purificadora e têm seleção positiva mais intensa (CAI & PETROV, 2010), têm menor probabilidade de estar envolvida com doenças humanas (DOMAZET-LOSO & TAUTZ, 2008), são menos frequentemente expressas durante o estágio filotípico, etapa no desenvolvimento embriônico de um animal quando mais se assemelha a outras espécies (DOMAZET-LOSO & TAUTZ 2010), têm diferentes sinônimos de viés de utilização de códon (PRAT *et al.* 2009) e são menos metilados (KELLER & YI, 2014). O método também tem sido utilizado para investigar a origem de genes pelo mecanismo *de novo* (CARVUNIS *et al.* 2012) e o ciclo de vida dos genes (ABRUSÁN 2013).

Entretanto, outros estudos criticaram o fato desses estudos anteriores usarem a ferramenta de alinhamento BLAST para datar o surgimento evolutivo de genes ou famílias gênicas, pois esta ferramenta tem uma conhecida limitação em inferir homologia entre homólogos distantes e que consequentemente pode subestimar a idade dos genes (MOYERS & ZHANG, 2014; ALBÀ & CASTRESANA, 2007). Moyers e Zhang (2014) também sugerem que métodos tais como HMMER (FINN *et al.*, 2011) e PSI-BLAST (ALTSCHUL *et al.*, 1997) poderiam obter melhores resultados para estimar a idades dos genes em comparação ao BLAST.

O fato é que, após uma grande quantidade de tempo, a similaridade entre os genes homólogos torna-se diminuta ao ponto da hipótese de homologia tender a ser descartada, causando o agrupamento de cada um dos produtos da duplicação gênica em grupos ortólogos diferentes. É plausível que um gene presente exclusivamente em mamíferos, por exemplo, possa ter originado de duplicação, mutações e recombinações a partir de um ou mais genes encontrados nos ancestrais de mamíferos e em outros grupos. E se a similaridade entre os

genes de mamíferos e de não mamíferos for muito baixa, as bases de dados de genes ortólogos falharão em sugerir a homologia do gene de mamíferos com genes de outros grupos.

O objetivo do presente trabalho é o desenvolvimento de uma metodologia alternativa àquelas que estão disponíveis atualmente e que permita aprofundar temporalmente a origem de grupos de genes órfãos, ou seja, de genes cuja homologia com genes que os originaram seja ainda desconhecida. Uma dessas metodologias é o alinhamento com PSSMs, usada para obter qual a localização mais provável do motivo que elas representam.

Em agrupamentos de ortólogos *stricto sensu*, monitoramos a pontuação que as PSSMs dão à sequência quesito, pois sabemos que as matrizes podem, em casos extremos após múltiplas iterações, passar a pontuar tão pouco a ponto de a própria sequência quesito desaparecer nos resultados do PSI-BLAST. Quando isso acontece, dizemos que a PSSM está completamente deteriorada (RIBEIRO, 2013).

Entretanto, já foi mostrado que proteínas que não eram consideradas como relacionadas, passaram a sê-lo com o emprego de PSSMs em processo inicial de deterioração. Por exemplo, as buscas que utilizaram uma enzima como quesito passaram a conter enzimas com níveis superiores de número E.C. (sigla em inglês para *Enzyme Commission number*, sistema de classificação hierárquico de enzimas baseado nas reações que são catalisadas), o que caracteriza semelhança funcional (COELHO JR et al, 2012). Outros resultados neste mesmo estudo conseguidos com a sobreposição das estruturas secundárias deduzidas reforçaram esses achados, pois grupos formados por agrupamento hierárquico aglomerativo com a métrica de sobreposição de estrutura secundária, resultam em associações idênticas àquelas obtidas com as PSSMs deterioradas.

E somente a caracterização das filogenias que envolvem os trechos correspondentes a cada um dos domínios da proteína em questão poderia resultar na solução da origem do gene correspondente. Considerando que o embaralhamento de exons deve ter gerado arquiteturas que são combinações de domínios, as abordagens com PSSMs ou com a sobreposição de estrutura secundária precisam ser adaptadas, pois o embaralhamento causa descontinuidade nos alinhamentos de sequência ou da estrutura secundária.

Esta proposta visa estudar genes que estão anotados exclusivamente em clados recentes e investigar sua origem, utilizando métricas adaptadas para buscas com baixa similaridade como PSI-BLAST com deterioração de PSSM em combinação com o rearranjo

de domínios. Para esta abordagem será necessário implementar bases locais integradas de domínios, sejam elas baseadas somente em sequências ou ainda integradas com topologias.

## **2.- Objetivos**

### **2.1.- Objetivo Geral**

Criar um banco de dados local com mapeamento de domínios, arquiteturas de domínios, alinhamentos com baixa similaridade e estudar a origem de genes exclusivos de clados recentes.

### **2.2.- Objetivos específicos:**

2.2.1.- Carregar em banco de dados local sequências obtidas de organismos com proteomas completos;

2.2.2.- Mapear nesses proteomas: domínios Pfam, arquiteturas e topologias CATH;

2.2.3.- Mapear arquiteturas utilizando Pfam e CATH, nos proteomas; por meio de agrupamentos de ortólogos.

2.2.4.- Traçar origem de arquiteturas complexas recentes a partir de arquiteturas simples e ancestrais, determinando a ancestralidade dos domínios e arquiteturas e propondo prováveis origens de genes presentes em clados mais recentes.

2.2.5.- Aplicar métricas de associação de homólogos com baixa similaridade com o software PSI-BLAST, acompanhando a deterioração da PSSM; propor a origem de genes recentes a partir de “hits” de PSI-BLAST mais ancestrais.

### **3.- Material e métodos**

#### **3.1.- Material**

##### **3.1.1. Dados biológicos**

As sequências polipeptídicas utilizadas nestas análises foram recuperadas de diferentes BDs. Utilizamos o BD CATH versão 3.5, disponível no endereço “[ftp://ftp.biochem.ucl.ac.uk/pub/cath/v3\\_5\\_0/](ftp://ftp.biochem.ucl.ac.uk/pub/cath/v3_5_0/)”, para obter as sequências de domínios proteicos. Sequência de famílias proteicas foram obtidas do BD Pfam versão 27, disponível no site “<ftp://ftp.ebi.ac.uk/pub/databases/Pfam>”. As sequências de PDB, outubro de 2013, foram obtidas no endereço “[ftp://ftp.wwpdb.org/pub/pdb/derived\\_data/pdb\\_seqres.txt](ftp://ftp.wwpdb.org/pub/pdb/derived_data/pdb_seqres.txt)”. As sequências das proteínas humanas foram recuperadas do BD UniProt, versão 24 (agosto de 2013), disponível em “<http://www.uniprot.org/downloads>”.

##### **3.1.1.2.- Grupos de proteínas ortólogas**

Utilizamos o BD de grupos de ortólogos UEKO, versão maio de 2013, obtida no endereço “<http://maxixe.icb.ufmg.br/ueko/>”. UEKO DB é um banco de dados desenvolvido e mantido pelo laboratório de biodados da UFMG e corresponde à base KEGG Orthology (KO) enriquecida com membros do uniref50.

##### **3.1.2.- Hardware**

Este trabalho foi executado em duas plataformas: uma com arquitetura Intel(R) Core (TM) 2 Quad CPU Q6600 2.40 Ghz de frequência de clock e 3 Gb de memória RAM no sistema operacional Linux; A outra, com arquitetura Intel(R) Xeon(R) CPU E5645 2.40 GHz e 64 Gb de memória no sistema operacional Linux.

##### **3.1.3.- Software**

###### **3.1.3.1.- BLAST+ versão 2.2.28+ desenvolvido pelo NCBI para Linux**

BLAST+ é um conjunto de ferramentas BLAST que utiliza o conjunto de ferramentas NCBI C#. As aplicações BLAST+ têm uma série de melhorias de desempenho e recursos sobre as aplicações BLAST e são distribuídas para os usuários que desejam executar o BLAST localmente. Podem ser baixadas a partir do endereço “<http://blast.ncbi.nlm.nih.gov/>”.

###### **3.1.3.2.- Máquina Virtual Java (JVM)**

O ambiente de execução de um dos *softwares* aqui produzido requer a instalação da

máquina virtual Java versão 6 ou posterior disponível no endereço “<http://java.com>”. Foi utilizada a máquina virtual OpenJDK 6 pré-instalada no sistema operacional Linux.

O Eclipse (disponível em “<http://www.eclipse.org/>”) é uma plataforma universal de código aberto para integração de ferramentas, construída para criar ambientes integrados de desenvolvimento (IDE). É uma ferramenta muito popular, escrita na linguagem Java, na qual o conjunto de pontos de extensões de funcionalidades permite a manipulação de recursos, entre muitas outras funcionalidades.

#### 3.1.3.4.- R versão 3.0.2

R é um ambiente de software livre, de código aberto, para análises estatísticas e produção de gráficos, disponível em “<http://www.r-project.org/>”. Algumas análises desta pesquisa envolvem agrupamentos e cálculos estatísticos, sendo este software utilizado para executar estas tarefas.

#### 3.1.3.5. Perl versão 5.14.2

Perl é uma linguagem de programação multiplataforma que incorpora recursos de outras linguagens tais como: C, Shell *scripting* (sh), AWK e SED (YATES, 2009). Ela fornece poderosos recursos para processamento de texto, sem os limites de comprimento de dados arbitrários de muitas ferramentas contemporâneas de linha de comando Unix (CHRISTIANSEN et al, 2012) que facilitam manipulação de arquivos de texto. Por esse motivo as manipulações de texto foram executadas utilizando essa ferramenta.

#### 3.1.3.6.- Banco de Dados MySQL

O software para gerenciamento de banco de dados utilizado foi o MySQL versão 5.6, obtido no endereço “<http://mysql.com>” junto a um componente para comunicação com aplicativos Java chamado Connector/J, disponível no mesmo endereço.

#### 3.1.3.7.- LCA *web service*

O último ancestral comum (LCA, sigla em inglês, “*Last Common Ancestor*”) é um serviço que se destina a ser uma forma rápida para encontrar o último ancestral comum entre dois ou mais táxons. Este serviço foi construído a fim de ajudar o estudo sobre as origens genéticas e está disponível em “<http://biodados.icb.ufmg.br/services/#lcaws>”. Este *web service* permite requisição no formato de protocolo simples de acesso aos objetos (SOAP, sigla em inglês de *Simple Object Access Protocol*) contendo apenas um conjunto de identificadores (ID) taxonômicos.

### 3.1.3.8.- SeedServer

Este aplicativo *Web* (disponível em “<http://maxixe.icb.ufmg.br:8080/BOWSWeb/>”) foi projetado para fornecer ortólogos presumidos a uma sequência semente (*seed*) qualquer, utilizando-se de uma estratégia que cria grupos de proteínas coortólogas e parálogas internas para cada organismo. Fundamentalmente, este *software* processa a busca em três passos: inicialmente baseia-se no método de agrupamento de sequências *Seed Linkage* descrito por Barbosa-Silva *et al* (2008) que busca pela melhor sequência alvo bidirecional (*bidirectional best hit*, BBH) entre os pares de ortólogos mais prováveis entre pelo menos dois organismos diferentes; Em seguida é feita uma consulta no banco de dados UEKO para recrutamento de ortólogos veiculados. E por fim ocorre uma validação por PSSM supervisionada.

A principal vantagem do SeedServer é contornar a necessidade de genomas completos para grupo ortólogos presumido. Para melhorar o desempenho, o SeedServer utiliza apenas entradas completas UniProt e não fragmentos.

### 3.1.3.9.- MaLe-PSI-BLAST web application

MaLe-PSI-BLAST é um aplicativo web que utiliza buscas com o PSI-BLAST supervisionadas por técnicas de aprendizado de máquinas e mineração de texto. Este aplicativo está disponível no endereço “<http://maxixe.icb.ufmg.br:8080/BOWSWeb/>” (RIBEIRO, 2013).

### 3.1.3.10.- Phyl-SPST

Phyl-SPST é um programa que analisa as semelhanças entre sequências e infere a evolução de famílias de proteínas (LEONARDI *et al.*, 2008). Seu algoritmo faz parte de uma classe de programas que fazem análises das sequências sem alinhamento. Nele é feita uma estimativa de um conjunto de contextos esparsos e a probabilidade de transição entre as sequências. Um contexto esparsoso é uma sequência curta de um subconjunto de símbolos, em um dado alfabeto, que são relevantes para predizer qualquer símbolo na sequência, dado que os símbolos anteriores pertencem aos subconjuntos de contexto. Pode ser obtido em <http://www.ime.usp.br/numec/softwares/phyl-spst/>.

### 3.1.3.11.- TimeTree

TimeTree é uma plataforma web na qual é possível obter o tempo de divergência entre os organismos permitindo comparar a filogenia com a história de outros organismos e com a história planetária, tais como a geologia, o clima, os impactos relacionados (HEDGES *et al.*,

2006). Ela está disponível no endereço web “<http://www.timetree.org>”.

### **3.2.- Processamentos e análises**

#### **3.2.1.- Anotação de domínios topológicos**

Utilizamos as sequências ATOM do BD CATH, sequências que têm registros no arquivo do PDB, e as alinhamos contra o BD PDB com a ferramenta *blast2sequence* do pacote de ferramentas NCBI+. Recuperamos então as sequências da região do PDB que alinhavam com as sequências ATOM. Esse procedimento foi importante, pois as sequências do BD CATH não são contínuas. Em seguida as 2.626 sequências de domínios superfamília do CATH distintas, que agora estavam contínuas, foram mapeadas em todas as 73.972 entradas de proteínas humanas que continham o “*status complete*” presente no BD UniProt. Esse mapeamento foi produzido com a ferramenta BLASTp com parâmetros padrão. Foram mantidas para processamento posterior todas as sequências que continham cobertura e identidade maiores ou iguais a 80%. O LCA foi estimado utilizando o LCA *web service* verificando-se o grupo de ortólogos do UEKO.

#### **3.2.2.- Determinando o LCA por SeedServer**

Houve proteínas cujos grupos de ortólogos foram estabelecidos como mais recentes que a origem dos Tetrapodas. Estas foram analisadas mais profundamente com o SeedServer com parâmetros *default*. E, com os grupos de ortólogos produzidos por essa ferramenta, foi feita uma nova verificação de LCA.

#### **3.2.3.- Determinando o LCA por MaLe-PSI-BLAST**

Investigamos com o Male-PSI-BLAST a origem do domínio CATH. O resultado dessa abordagem foi verificado com o LCA *web service* e relacionados com grupos de ortólogos do UEKO e com os termos do Gene Ontology.

#### **3.2.4.- Avaliando a continuidade de homologia entre as sequências**

O grupo de sequências obtidas pelo Male-PSI-BLAST foi submetida a uma análise de todos contra todos pelo programa PHYL-SPST. A matriz de distância gerada pelo PHYL-SPST foi submetida a uma análise de seleção de número ótimo de grupos pelo pacote pamk (HENNIG, 2014). Em seguida foi plotado um histograma que evidenciou descontinuidades entre as sequências.

### 3.2.5.- Calculando os índices de promiscuidade

Foi adotada, para medida de índice de promiscuidade de domínio ( $\pi_i$ ) a frequência ponderada de bigrama – um termo padrão para pares de palavras adjacentes em linguística computacional (MANNING & SCHÜTZE, 1999) – proposta por Basu e seus colaboradores (2008). Esta foi originalmente derivada da fórmula de ganho de informação Kullback-Leibler:

$$\pi_i = \beta_i \times \log \left( \frac{\beta_i}{f_i} \right) \quad (1)$$

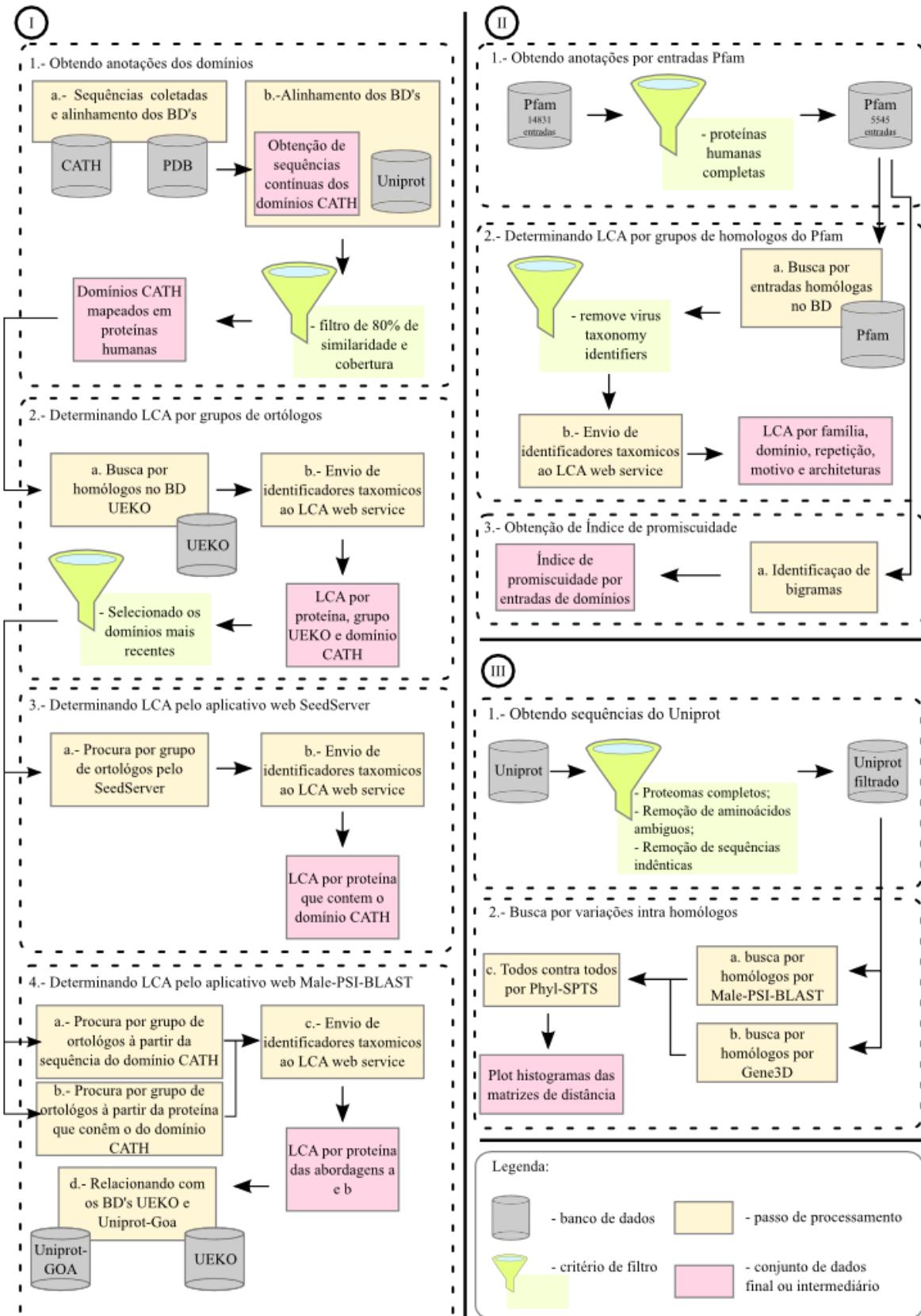
A frequência de bigrama ( $\beta_i$ ) é:

$$\beta_i = \frac{T_i}{\frac{1}{2} \sum_{j=1}^t T_j} \quad (2)$$

onde  $t$  é o número de domínios diferentes.  $T_i$  é o número de vizinhos exclusivos do domínio  $i$  e  $f_i$  é a frequência do domínio  $i$  no genoma, calculado como  $n_i / N$ , em que  $n_i$  é a contagem total do domínio  $i$  e  $N$  é o número total de domínios detectados no dado genoma:

$$N = \sum_{i=1}^t n_i \quad (3)$$

O fluxograma a seguir (Figura 1) sumariza as ações que foram tomadas para processamento e análise dos dados de origem gênica gerados com a metodologia proposta neste trabalho. Os códigos aqui produzidos estão disponíveis em <http://bioinfo.ib.usp.br/diego/source>.



**Figura 1.**- Fluxograma descrevendo os métodos de análises das entradas (I) CATH, (II) Pfam e (III) estudo de caso com os príons e ADHs.

## 4.- Resultados

Os resultados obtidos estão apresentados em três partes. A primeira parte contém as análises derivadas da BD CATH, cujos resultados foram apresentados no *X-Meeting* em 2013 e no *II International Symposium On Evolutionary Biology* em 2014. A segunda parte descreve as análises da BD Pfam, cujos resultados foram apresentados no *X-Meeting* 2014 e futuramente apresentado no *62º Brazilian-International Congress of Genetics* em setembro de 2016. E na terceira parte estão mostrados os resultados obtidos das análises de um estudo de casos com as proteínas de príons e ADHs. Em princípio, cada uma dessas partes corresponderá a um artigo que será submetido em breve para publicação.

### 4.1.- Análise filoestratigráfica das entradas CATH

A busca por alinhamento local entre os domínios CATH e as proteínas humanas presentes no BD UniProt resultou em um conjunto de 11.547 proteínas humanas mapeadas com os domínios CATH. É importante ressaltar que, nesta abordagem, os domínios CATH de todos os organismos presentes no BD foram alinhados contra o proteoma humano e esta abordagem por si só já expandiu o banco de dados, por mapear os domínios em proteínas que não têm estrutura tridimensional definida em humanos.

Mapeadas as proteínas humanas com os domínios de superfamília do CATH foi possível então relacioná-las com o banco de dados de grupos de ortólogos UEKO e assim determinar o último ancestral comum dentre os grupos de ortólogos. Das 11.547 proteínas humanas que contêm os domínios de superfamília CATH, 7.421 proteínas estão inseridas em pelo menos um grupo de ortólogos do UEKO; e o restante das proteínas mapeadas, 4.126, não estão contidas em qualquer grupo do BD UEKO.

Das proteínas presentes no BD UEKO, foram encontrados, entre os clados mais recentes, Euteleostomi, Tetrapoda, Eutheria, Catarrini e Hominoidea, com respectivamente, 1.801, 232, 488, 2 e 2 proteínas que se originaram ao longo da linhagem humana e que contêm o domínio CATH. Estas proteínas estão contidas em, respectivamente, 373, 74, 55, 2 e 2 grupos de ortólogos. Esta contagem é interessante, já que alguns grupos de ortólogos são super-representados, causado por um viés relativo ao número de estudos de alguns grupos de ortólogos.

Além disso, foi realizada uma contagem pela ocorrência mais filogeneticamente basal dos domínios CATH, já que existem domínios que são mapeados em diferentes grupos de

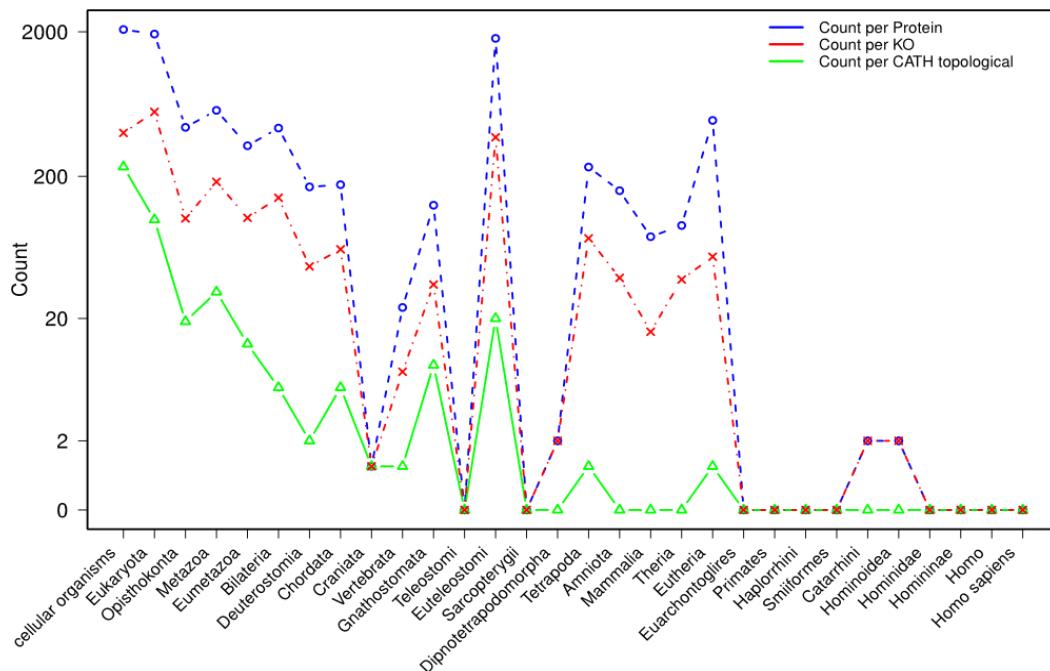
ortólogos do UEKO que, por sua vez, podem ter diferentes ancestralidades. A Tabela 2 exemplifica a situação descrita anteriormente para o domínio 3.10.450.10 que está contido em diversas proteínas e, por conseguinte, podem ou não estar contido em diferentes grupos de ortólogos do UEKO e suas respectivas origens.

**Tabela 2.-** Exemplo de mapeamento de proteínas e domínios CATH nos grupos de ortólogos do UEKO com suas respectivas ancestralidades.

Uniprot accession	Superfamília CATH	KO	LCA taxid	LCA nome	LCA level	LCA rank
P28325	3.10.450.10	K13897	314295	Hominoidea	26	superfamily
P28325	3.10.450.10	K13898	9347	Eutheria	20	no rank
P01034	3.10.450.10	K13898	9347	Eutheria	20	no rank
P01034	3.10.450.10	K13899	32523	Tetrapoda	16	no rank
P28325	3.10.450.10	K13900	314295	Hominoidea	26	superfamily
P28325	3.10.450.10	K13901	9526	Catarrini	25	parvorder
P01040	3.10.450.10	K13907	33154	Opisthokonta	3	no rank
P04080	3.10.450.10	K13907	33154	Opisthokonta	3	no rank
Q6IB90	3.10.450.10	K13907	33154	Opisthokonta	3	no rank
Q76LA1	3.10.450.10	K13907	33154	Opisthokonta	3	no rank

A contagem pela ocorrência mais filogeneticamente basal dos domínios resultou que as proteínas originadas em Hominoidea e Catarrini, que contém os domínios 3.10.450.10, que ainda não tem nome atribuído, e 2.60.40.10, Imunoglobulinas, respectivamente, também contêm domínios que estão presentes tanto em proteínas humanas quanto em organismos no qual o ancestral comum entre eles e os humanos mais recentes derivou na separação do clado Opisthokonta, que é um ancestral mais distante. A Figura 2 mostra todas as contagens dessa abordagem.

Entretanto, foram encontrados 20, 1 e 1 domínios que tiveram seu primeiro aparecimento em Euteleostomi, Tetrapoda e Eutheria, respectivamente, os quais são clados bem mais recentes na evolução da linhagem humana. Os domínios mais recentes que estão contidos no UEKO tiveram as origens detectadas, sendo um em Tetrapoda, o domínio 1.10.790.10, proteína prón principal, e um em Eutheria, 2.40.15.10, produto oncogene P14tcl1.



**Figura 2.-** Ganho de novos domínios na linhagem ancestral-descente de humanos baseados nos grupos de ortólogos do UEKO. Em azul contagem por proteínas, em vermelho contagem por grupos de ortólogos do Kegg e em verde contagem entradas do BD CATH.

Utilizamos as proteínas cujos domínios CATH tiveram origem recente, desde a divisão que originou aos Euteleostomi, para proceder as análises com o *SeedServer*. A abordagem que utilizou *SeedServer* mostrou uma proximidade com os resultados obtidos com os grupos de ortólogos, entretanto alguns domínios tiveram os níveis de ancestralidades baixadas, como foi o caso do domínio 2.40.15.10, produto oncogene P14tcl1, detectada a origem em Eutheria pelos grupos de ortólogos e no método que utiliza o *SeedServer* teve sua ancestralidade determinada em Mammalia. Outros domínios tiveram a origem detectada em Euteleostomi e passaram a ser Gnathostomata. A Tabela 3 mostra os domínios e suas respectivas ancestralidades, encontradas com a ferramenta *SeedServer*.

**Tabela 3.-** Ganho de novos domínios na linhagem ancestral de humanos baseados nos grupos de ortólogos produzido pelo *SeedServer*.

LCA	Nível LCA	Entrada CATH	Nome
Mammalia	18	2.40.15.10	Proto-oncogene - Oncogene Product P14tcl1
Tetrapoda	16	1.10.790.10	Major Prion Protein
		1.10.168.10	Phosducin, domain 2
		1.10.532.10	Transcription Factor, Stat-4;
		1.10.870.10	HLA-dr Antigens Associated Invariant Chain; Chain A
		2.20.60.10	Heparin-binding Growth Factor, Midkine; Chain A
		2.30.90.10	Heparin-binding Growth Factor, Midkine; Chain A-C-terminal Domain
		3.50.30	Glucose Oxidase; domain 1
		3.90.1290.10	Plakin repeat
Euteleostomi	13	4.10.51.10	Cytochrome C Oxidase, chain K
		4.10.630.10	Nuclear receptor coactivator src-1
		4.10.760.10	Agouti Related Protein; Chain A
		4.10.800.10	Invariant Chain; Chain I
		4.10.81	Cytochrome C Oxidase; Chain M
		3.30.500	Murine Class I Major Histocompatibility Complex, H2-DB; Chain A, domain 1
		4.10.10.10	Metallothionein Isoform II
		2.170.300.10	Tie2 ligand-binding domain superfamily
		4.10.260	G Protein Gi Gamma 2
		1.10.100.10	Insulin-like, subunit E
Gnathostomata	11	3.10.320.10	Class II Histocompatibility Antigen, M Beta Chain; Chain B, domain 1
		1.20.1250	Growth Hormone; Chain: A;
		4.10.1220.10	EGF-type module

Entre as proteínas que contêm os domínios mais recentes, escolhemos para estudo de caso as proteínas príons para verificar a fundo como a abordagem se comporta para determinar a origem dos domínios.

O resultado do *SeedServer* mostrou que as proteínas que contêm os domínios príon produzem três grupos de ortólogos em duas ancestralidades distintas, um grupo que tem origem em Tetrapoda e dois com origem em Theria. Nos agrupamentos de ortólogos

produzidos pelo *SeedServer* dentre os que tiveram sua origem em Theria, um grupo (grupo 1) foi formado pela proteína B3KQX7 que contêm proteínas as quais possuem principalmente como descrição “proteína prón principal” e o outro grupo (grupo 2) foi formado pelas proteínas A7U7M4, A7U7M2, Q9UKY0 e Q27H88 as quais estavam em um grupo que continha principalmente a descrição “proteína duplicata do tipo prón”. O terceiro agrupamento de proteínas (grupo 3) resultante é composto pelas proteínas Q53YK7, P04156, Q5U0K3, B2R5Q9, Q6FGR8, A1YVW6, B4DJ65, B2NI04, B4DDS1, B2NI05, B4DI53, O75942, e Q6SES1 com origem em Tetrapoda e tem em geral a descrição de “proteína prón principal”.

O próximo passo foi investigar com Male-PSI-BLAST, a origem do próprio domínio prón de modo permissivo, no qual se permitiu que a PSSM pudesse iniciar o processo de deterioração. Para acompanhar possíveis detecções de proteínas pertencentes a grupos de ortólogos ou funções moleculares distintas, foi realizado o LCA par a par entre a sequência quesito e as proteínas presentes no resultado do alinhamento, e também as relacionando com as funções moleculares do *Gene Ontology* (GO) e os grupos do UEKO. Neste caso, funções moleculares e grupos de ortólogos distintos aos da sequência quesito poderiam indicar uma possível relação com o domínio que deu origem à sequência quesito.

E notavelmente os dois domínios próns distintos presentes no BD CATH foram mapeados com LCA em Tetrapoda e Theria, em acordo com os resultados obtidos pelo *SeedServer* para busca de ortólogos. É importante enfatizar que nesta análise o Male-PSI-BLAST convergiu e consequentemente não houve deterioração da PSSM.

Os domínios da base CATH dos tipos “proteína prón principal” mapearam em um único grupo de KO, K05634 o qual é dedicado ao prón. Enquanto os domínios “proteína duplicata do tipo prón” não foram mapeados em nenhum grupo de ortólogo. Entretanto os dois compartilham as anotações do termo do GO, GO:0005507, ligador de íon cobre. Além disso, as proteínas encontradas que contêm o domínio “proteína prón principal” estão também associados a outros termos GO que podem ser resumidos como “proteínas ligadoras” e estão descritos na tabela 4. As tabelas 4 e 5 mostram amplitude de ancestralidade comum dos domínios “proteína prón principal” e “proteína duplicata do tipo prón”, relacionando com os grupos de ortólogos e os termos de ontologia gênica.

**Tabela 4.-** Amplitude dos domínios “proteína prón principal” na ancestralidade da linhagem humana relacionada com grupos de ortólogos e ontologia gênica.

nível LCA	nome LCA	Presença	Termos GO	UEKO
1	cellular organisms	não		
2	Eukaryota	não		
3	Opisthokonta	não		
4	Metazoa	não		
5	Eumetazoa	não		
6	Bilateria	não		
7	Deuterostomia	não		
8	Chordata	não		
9	Craniata	não		
10	Vertebrata	não		
11	Gnathostomata	não		
12	Teleostomi	não		
13	Euteleostomi	não		
14	Sarcopterygii	não		
15	Dipnotetrapodomorpha	não		
16	Tetrapoda	sim		K05634
17	Amniota	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802	
18	Mammalia	não		
19	Theria	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802	K05634
20	Eutheria	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802	K05634
21	Boreoeutheria	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
22	Euarchontoglires	sim	GO:0005507; GO:0005515; GO:0008017; GO:0015631; GO:0042802; GO:0043008; GO:0046872; GO:0051087	K05634
23	Primates	sim		K05634
24	Haplorrhini	não		
25	Simiiformes	sim	GO:0005507; GO:0008017 ; GO:0015631; GO:0042802 ; GO:0046872	K05634
26	Catarrhini	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634

27	Hominoidea	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
28	Hominidae	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
29	Homininae	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
30	Homo	não		
31	Homo sapiens	sim	GO:0005507; GO:0005515; GO:0008017; GO:0015631; GO:0042802; GO:0043008; GO:0046872; GO:0051087	K05634

**Tabela 5.-** Amplitude dos domínios “proteína duplicata do tipo prón” na ancestralidade da linhagem humana relacionada com grupos de ortólogos e ontologia gênica

nível LCA	nome LCA	Presença	Termos GO	UEKO
1	cellular organisms	não		
2	Eukaryota	não		
3	Opisthokonta	não		
4	Metazoa	não		
5	Eumetazoa	não		
6	Bilateria	não		
7	Deuterostomia	não		
8	Chordata	não		
9	Craniata	não		
10	Vertebrata	não		
11	Gnathostomata	não		
12	Teleostomi	não		
13	Euteleostomi	não		
14	Sarcopterygii	não		
15	Dipnotetrapodomorpha	não		
16	Tetrapoda	não		
17	Amniota	não		
18	Mammalia	não		
19	Theria	sim	GO:0005507	
20	Eutheria	sim		
21	Boreoeutheria	sim	GO:0005507	
22	Euarchontoglires	sim	GO:0005507	

23	Primates	não		
24	Haplorrhini	não		
25	Simiiformes	sim	GO:0005507	
26	Catarrhini	sim	GO:0005507	
27	Hominoidea	sim		
28	Hominidae	não		
29	Homininae	sim	GO:0005507	
30	Homo	não		
31	Homo sapiens	sim	GO:0005507	

Foi usada também a sequência total de aminoácidos das proteínas que contêm os domínios príon para procurar com Male-PSI-BLAST por proteínas que possivelmente serviram como arcabouço que manteve o domínio príon quando ele apareceu pela primeira vez. Para sumarizar os resultados foram utilizados os agrupamentos produzidos pela ferramenta SeedServer. O Male-PSI-BLAST detectou a origem das proteínas pertencentes ao grupo 1 em Tetrapoda, ao grupo 2 em Theria e ao grupo 3 em Eukaryota. As tabelas 6, 7 e 8 descrevem os resultados obtidos pelo Male-PSI-BLAST nos grupos 1, 2 e 3, respectivamente.

**Tabela 6.-** Amplitude do grupo 1 “proteína príon principal” na ancestralidade da linhagem humana relacionada com grupos de ortólogos e ontologia gênica.

nível LCA	nome LCA	Presença	Termos GO	UEKO
1	cellular organisms	não		
2	Eukaryota	não		
3	Opisthokonta	não		
4	Metazoa	não		
5	Eumetazoa	não		
6	Bilateria	não		
7	Deuterostomia	não		
8	Chordata	não		
9	Craniata	não		
10	Vertebrata	não		
11	Gnathostomata	não		
12	Teleostomi	não		
13	Euteleostomi	não		

14	Sarcopterygii	não		
15	Dipnotetrapodomorpha	não		
16	Tetrapoda	sim		K05634
17	Amniota	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802	
18	Mammalia	não		
19	Theria	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802	K05634
20	Eutheria	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802	K05634
21	Boreoeutheria	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
22	Euarchontoglires	sim	GO:0005507; GO:0005515; GO:0008017; GO:0015631; GO:0042802; GO:0043008; GO:0046872; GO:0051087	K05634
23	Primates	sim		K05634
24	Haplorrhini	não		
25	Simiiformes	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
26	Catarrhini	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
27	Hominoidea	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
28	Hominidae	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
29	Homininae	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
30	Homo	não		
31	Homo sapiens	sim	GO:0005507; GO:0005515; GO:0008017; GO:0015631; GO:0042802; GO:0043008; GO:0046872; GO:0051087	K05634

**Tabela 7.-** Amplitude do grupo 2 “proteína duplicata do tipo príon” na ancestralidade da linhagem humana relacionada com grupos de ortólogos e ontologia gênica.

nível LCA	nome LCA	Presença	Termos GO	UEKO
1	cellular organisms	não		
2	Eukaryota	não		
3	Opisthokonta	não		
4	Metazoa	não		

5	Eumetazoa	não		
6	Bilateria	não		
7	Deuterostomia	não		
8	Chordata	não		
9	Craniata	não		
10	Vertebrata	não		
11	Gnathostomata	não		
12	Teleostomi	não		
13	Euteleostomi	não		
14	Sarcopterygii	não		
15	Dipnotetrapodomorpha	não		
16	Tetrapoda	não		
17	Amniota	não		
18	Mammalia	não		
19	Theria	sim	GO:0005507	
20	Eutheria	sim		
21	Boreoeutheria	sim	GO:0005507	
22	Euarchontoglires	sim	GO:0005507	
23	Primates	não		
24	Haplorrhini	não		
25	Simiiformes	sim	GO:0005507	
26	Catarrhini	sim	GO:0005507	
27	Hominoidea	sim		
28	Hominidae	não		
29	Homininae	sim	GO:0005507	
30	Homo	não		
31	Homo sapiens	sim	GO:0005507	

**Tabela 8.-** Amplitude do grupo 3 “proteína príon principal” na ancestralidade da linhagem humana relacionada com grupos de ortólogos e ontologia gênica.

nível LCA	nome LCA	Presença	Termos GO	UEKO
1	cellular organisms	sim	GO:0000166; GO:0003676; GO:0003677; GO:0003723; GO:0003824; GO:0004252; GO:0009982; GO:0016853;	K06178; K08372

			GO:0016866	
2	Eukaryota	sim	GO:0000166; GO:0003674; GO:0003676; GO:0005199	K12898
3	Opisthokonta	sim		
4	Metazoa	não		
5	Eumetazoa	não		
6	Bilateria	sim	GO:0000166; GO:0003676; GO:0003723; GO:0008061	K12741
7	Deuterostomia	não		
8	Chordata	não		
9	Craniata	não		
10	Vertebrata	não		
11	Gnathostomata	não		
12	Teleostomi	não		
13	Euteleostomi	sim	GO:0000166; GO:0003676; GO:0005509; GO:0005544	K12741; K13158
14	Sarcopterygii	não		
15	Dipnotetrapodomorpha	não		
16	Tetrapoda	sim		K05634
17	Amniota	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802	
18	Mammalia	não		
19	Theria	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802	K05634
20	Eutheria	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802	K05634
21	Boreoeutheria	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
22	Euarchontoglires	sim	GO:0005507; GO:0005515; GO:0008017; GO:0015631; GO:0042802; GO:0043008; GO:0046872; GO:0051087	K05634
23	Primates	sim		K05634
24	Haplorrhini	não		
25	Simiiformes	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
26	Catarrhini	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634

27	Hominoidea	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
28	Hominidae	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
29	Homininae	sim	GO:0005507; GO:0008017; GO:0015631; GO:0042802; GO:0046872	K05634
30	Homo	não		
31	Homo sapiens	sim	GO:0005507; GO:0005515; GO:0008017; GO:0015631; GO:0042802; GO:0043008; GO:0046872; GO:0051087	K05634

Além disso, para os grupos 1 e 2 as proteínas mapeadas permaneceram com o mesmo grupo de ortólogos e termos de ontologia gênica. Entretanto, para o grupo 3 as proteínas mais ancestrais encontradas com o Male-PSI-BLAST pertenciam a quatro distintos grupos de KO, K12898, K12741, K13158 e K05634, mas todos iniciados com “ribonucleoproteínas nucleares heterogêneas”. E os arcabouços das proteínas mais ancestrais compartilham termos GO de função moleculares diferentes (GO:0000166 e GO:0003676, nucleotídeo e ligante de ácido nucleico), enquanto os mais recentes, que contêm o domínio prón, compartilham os termos GO:0005507, GO:0008017, GO:0015631 e GO:0042802 (Cobre, microtúbulos e proteínas idênticas ligadoras).

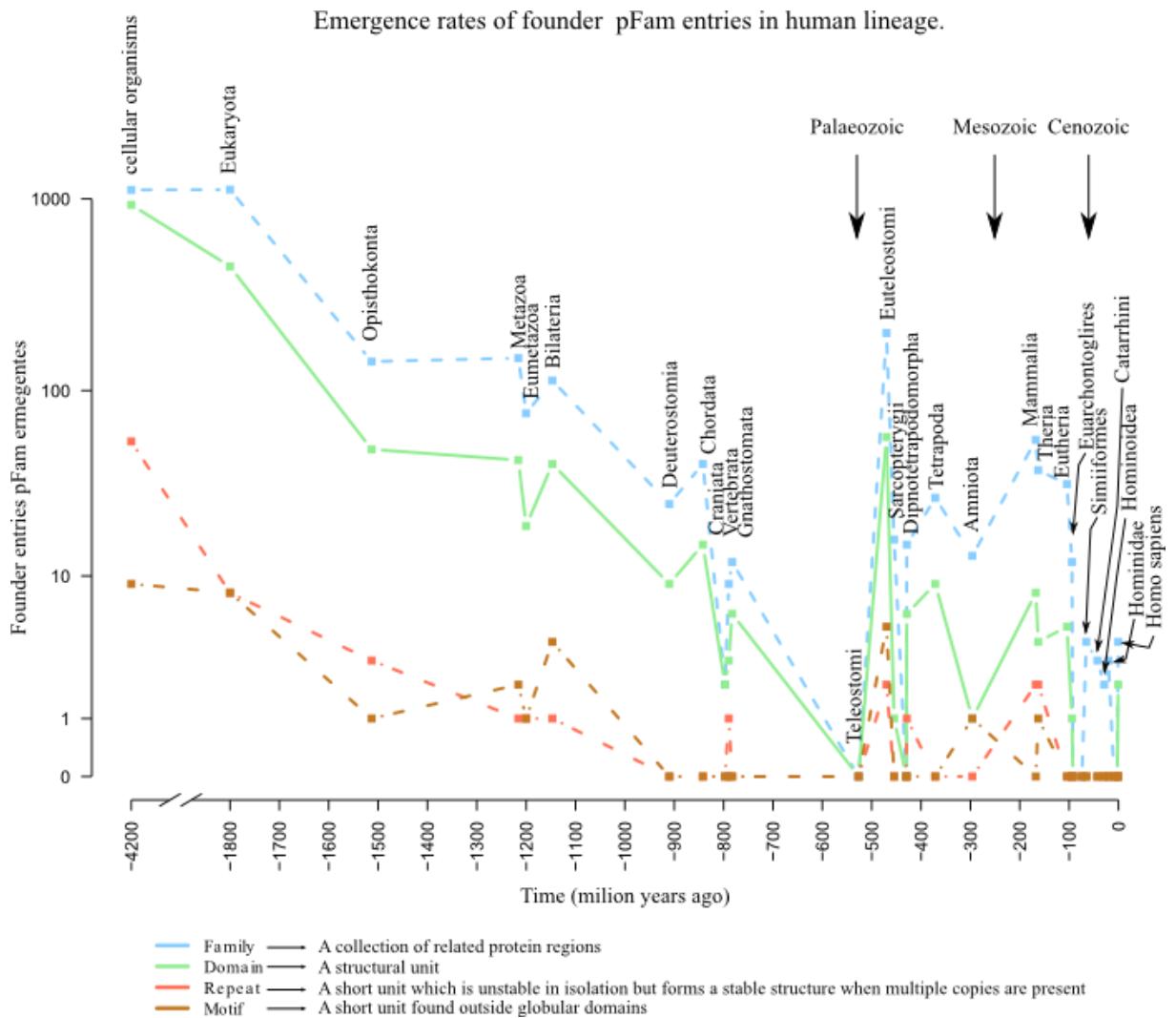
#### 4.2.- Análise filoestratigráfica das entradas Pfam

Nesta etapa realizamos uma pesquisa filoestratigráfica das entradas humanas presentes no Pfam rastreando a origem dos domínios e arquiteturas. O BD Pfam contém 14.831 famílias do tipo Pfam-A, entradas curadas contendo um pequeno conjunto de membros representativos das famílias. Em seguida, foi aplicado um filtro das entradas que estejam presentes no proteoma humano e que tenham status “complete” no BD UniProt. Esse procedimento reduziu o número de entradas para 5.845.

Os identificadores taxonômicos das entradas homólogas do Pfam foram enviados ao LCA *web service*. Neste ponto, os identificadores de vírus foram removidos, e com as contagens das origens das entradas foi produzido o gráfico correspondente à figura 3.

A determinação da origem das proteínas baseadas no grupo de entradas homólogas do Pfam mostrou ter perfil similar com as análises dos domínios CATH, dados apresentados na seção anterior, evidenciados principalmente por três picos proeminentes os quais somados representam 74% das entradas presentes no Pfam. São eles, um pico na origem da vida, um na origem do super-reino eucarionte e o outro na origem do clado Euteleostomi. A Figura 3 apresenta distribuição quanto à origem das entradas presentes no Pfam. Entretanto, no que se refere às entradas recentes, quando comparadas aos domínios CATH, o Pfam apresenta um número maior de entradas surgindo desde o organismo que deu origem aos Euteleostomi, 532 (11%) entradas, inclusive com entradas restritas ao ser humano.

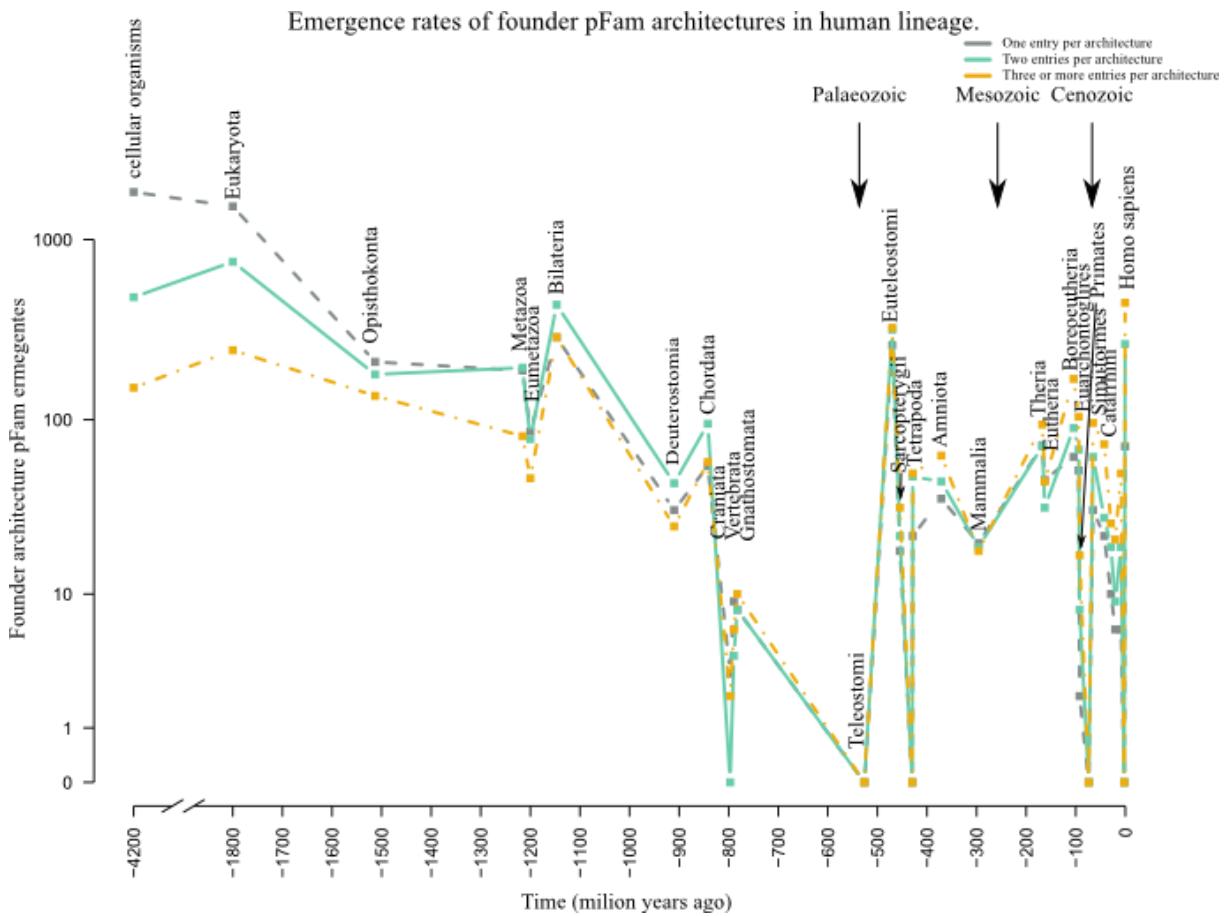
Uma abordagem similar foi aplicada às entradas de arquiteturas do Pfam, conjuntos de domínios que estão presentes na proteína. Para facilitar as análises, categorizamos as arquiteturas em três classes: aquelas que contêm um domínio, as que contêm dois domínios e as que contêm três ou mais domínios. A distribuição destes domínios é mostrada na figura 4.



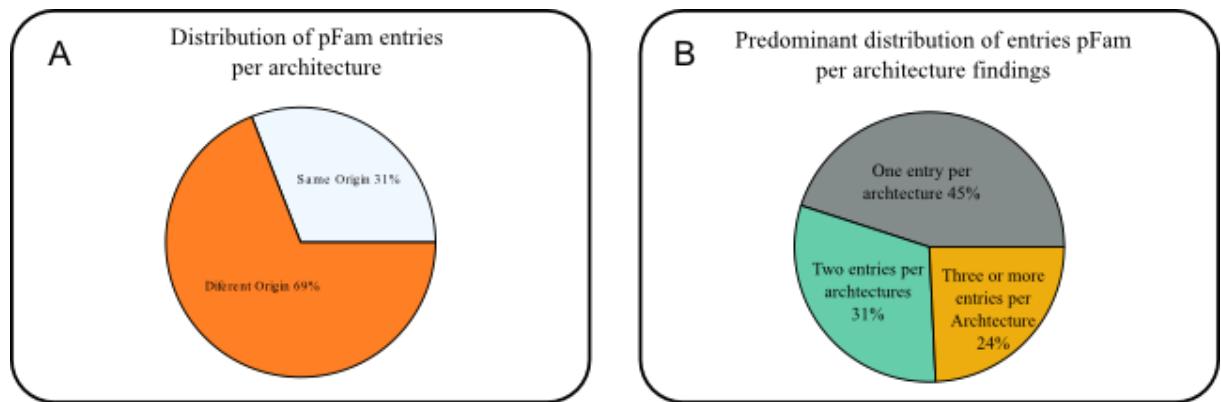
**Figura 3** - Ganho de novas entradas Pfam, na linhagem ancestral de humanos baseados nos grupos de entradas homólogas presente neste banco de dados. Em azul contagem das emergências de novas famílias, em verde contagem por domínios, em vermelho contagem por repetições e em marrom contagem das emergências de motivos.

As análises filoestratigráficas das arquiteturas Pfam mostram existir um número ainda maior de arquiteturas recentes, sendo 31% das arquiteturas compartilhadas entre os Euteleostomi. A figura 5 mostra a taxa de arquiteturas emergentes do Pfam.

Foi realizado um experimento que relaciona o surgimento da arquitetura e a origem do domínio mais recente presente nesta arquitetura. O resultado, para essa análise, mostrou, em geral, não existir uma “preferência” de incorporação de domínios nas arquiteturas em relação à origem dos domínios (Figura 6). Outro ponto evidenciado por essa análise foi que 31% das arquiteturas surgem na mesma época que seu domínio mais recente.



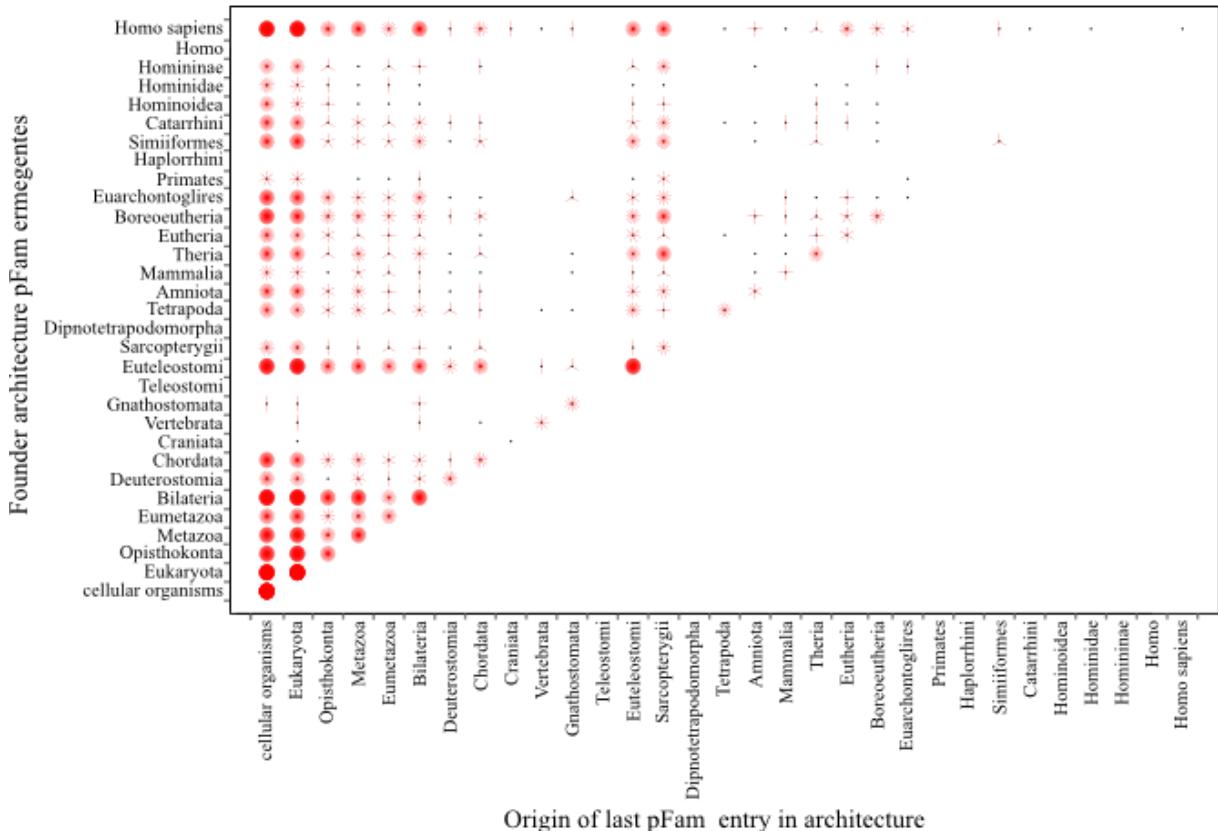
**Figura 4** - Ganhos de novas arquiteturas Pfam, na linhagem ancestral de humanos baseados nos grupos de arquiteturas homólogas presente neste banco de dados. Em cinza contagem de arquiteturas que contêm somente um domínio por arquitetura, em verde contagem das arquiteturas que contêm dois domínios por arquitetura e em amarelo contagem das arquiteturas que contêm três ou mais domínios por arquitetura.



**Figura 5 – (A)** Distribuição das entradas Pfam por arquitetura quanto à origem. Em laranja grupo de arquiteturas em que a origem dela é diferente da origem do domínio mais recente

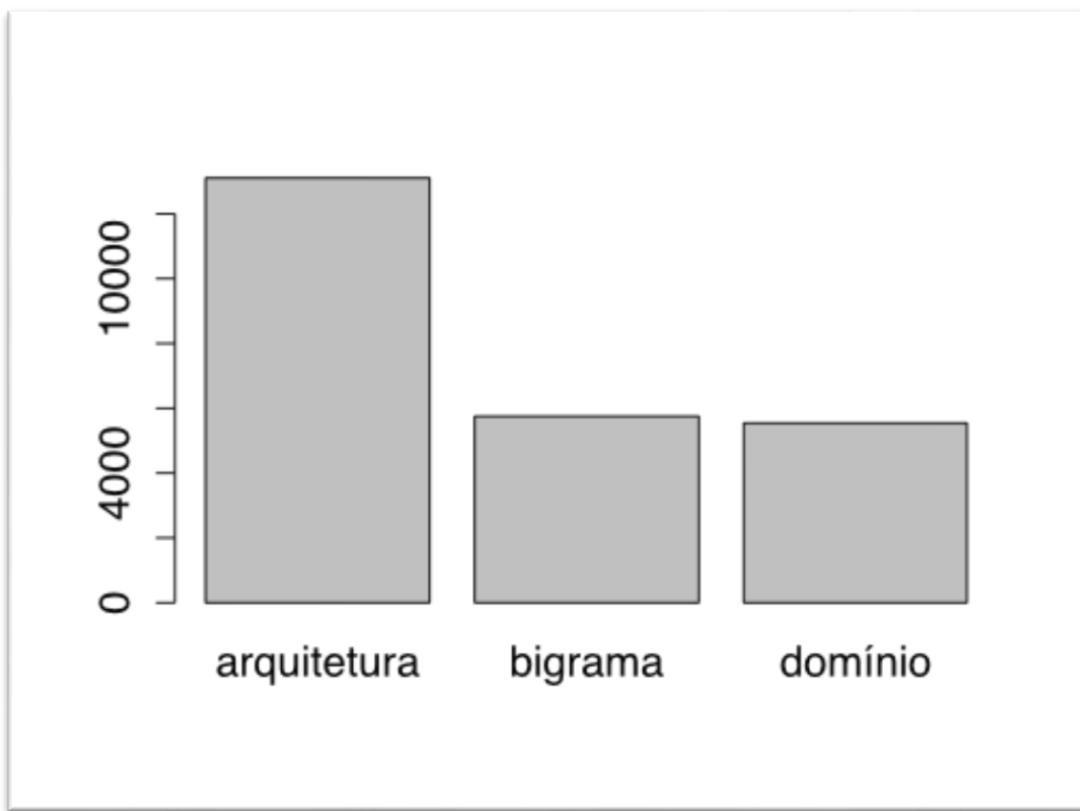
(legenda Fig. 5 continuação) presente na mesma arquitetura. Em azul, grupo que a origem da arquitetura é a mesma do domínio mais recente; (B) Distribuição das entradas Pfam por arquitetura. Em cinza, arquiteturas com somente um domínio, em verde, arquiteturas com dois domínios e em amarelo, arquiteturas com três ou mais domínios.

### Density distribution of Architecture emergence versus last domain origin in the architecture



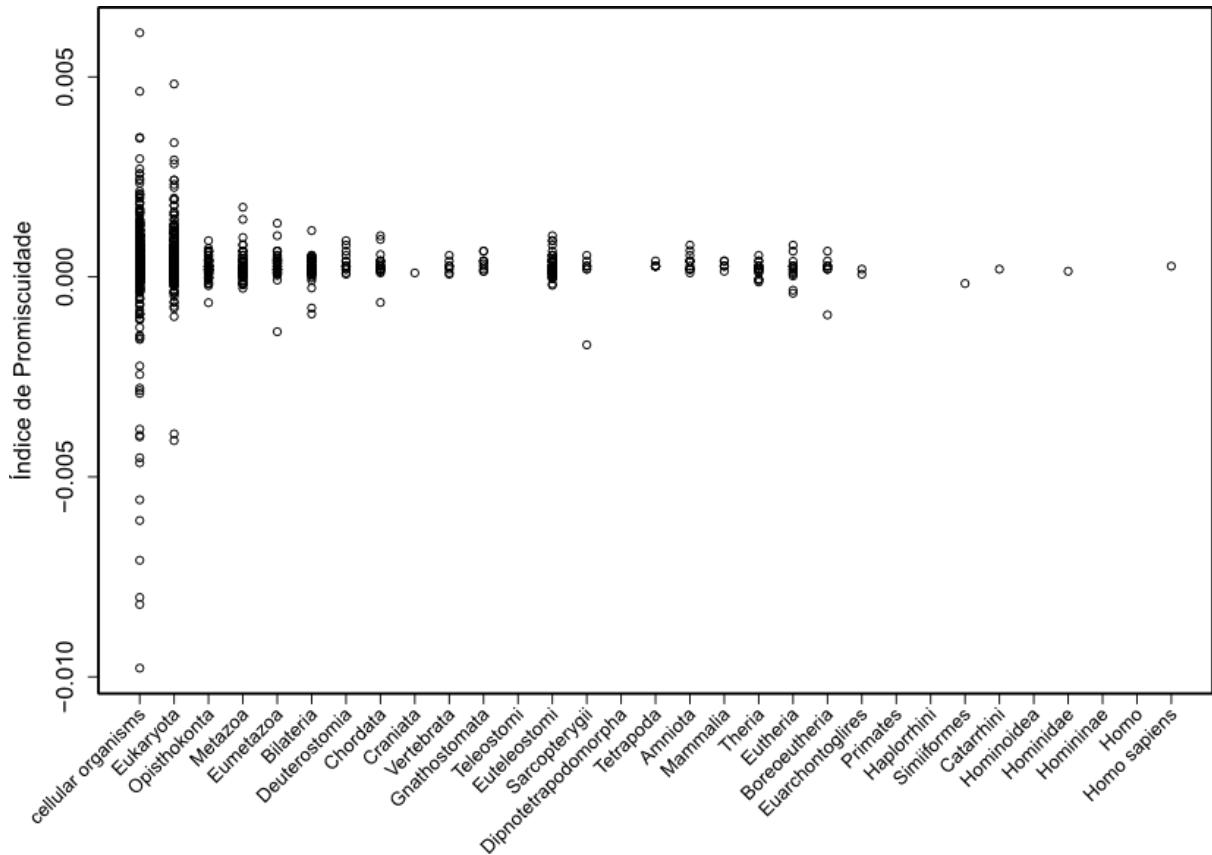
**Figura 6 –** Densidade de distribuição das arquiteturas emergentes versus a origem do último domínio presente nas arquiteturas. Traços em vermelho indicam a origem da estrutura.

Utilizando as sequências de domínios que compõem as arquiteturas de domínios foi possível identificar todos os pares de domínios que são vizinhos, também chamados de bigramas presente no proteoma humano. O número de bigramas de domínios encontrado no proteoma humano foi de 5751. O número de bigramas encontrado apresenta uma tênue diferença entre os 5544 domínios identificados nas proteínas humanas (Figura 7).



**Figura 7** – Contagem de arquiteturas, bigramas e domínios presentes no BD Pfam.

Em seguida, foi examinado se existe alguma tendência para domínios formarem combinações estáveis com outros domínios, uma característica referida como "promiscuidade", em relação a origem dos domínios. Para isso, foi utilizada a métrica ponderada da frequência de bigrama, proposta por Basu e seus colaboradores (2008), a qual é normalizada pela frequência de domínio, para definir a promiscuidade relativa dos domínios que aparecem em proteínas multidomínios. Os domínios foram agrupados em 31 categorias com base na ancestralidade, obtidas por análises filoestratigráficas das entradas de domínios presentes no banco de dados Pfam. Das 31 categorias de idade é possível evidenciar por teste de correlação de Spearman que existe uma baixa correlação negativa de aproximadamente -0,043, entre a idade dos domínios e os índices de promiscuidade deles. Isso indica que, quanto mais antigo, mais promíscuo o domínio é. A figura 8 mostra a variação dos índices de promiscuidade em relação à origem dos domínios.



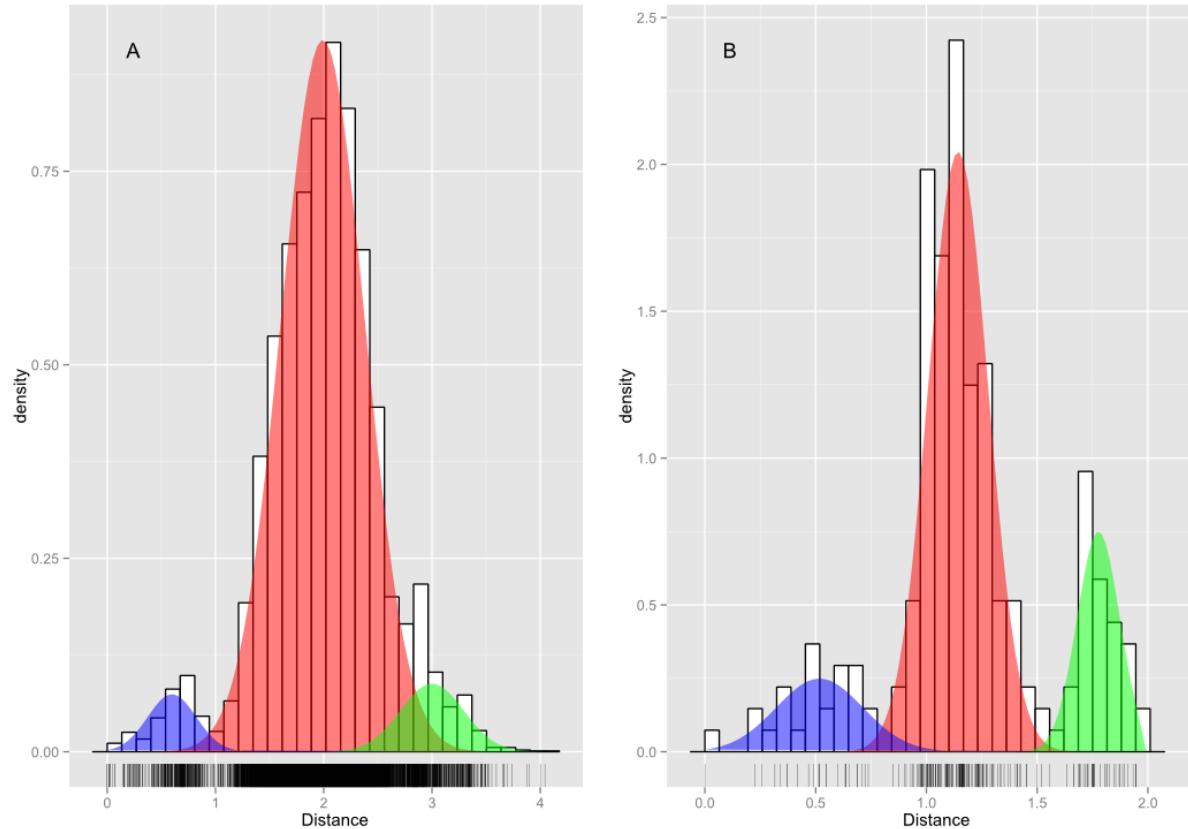
**Figura 8** – Relação entre os índices de promiscuidade dos domínios Pfam e as suas respectivas origens na linhagem ancestral-descendente de *H. sapiens*.

#### 4.3.- Estudo de caso: príons e ADH's

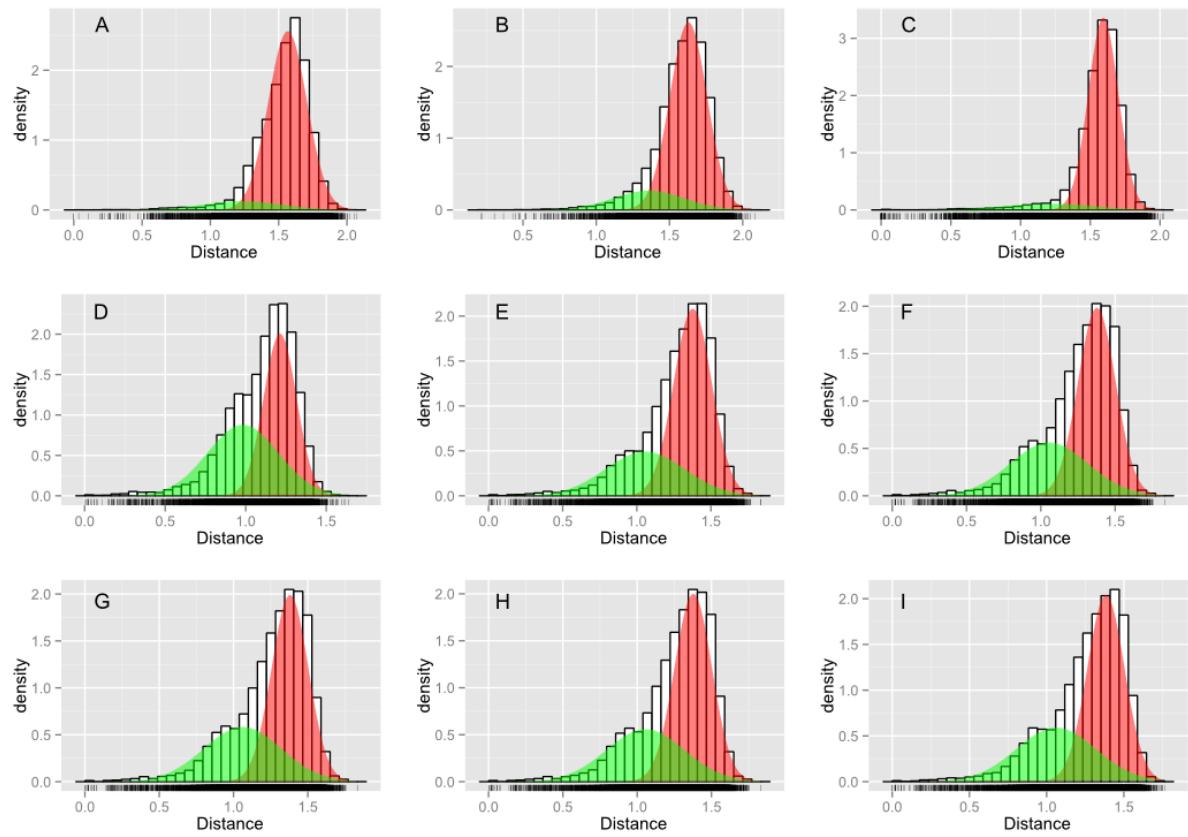
Em uma tentativa de aprimorar a busca pela origem do domínio CATH, realizamos um estudo de caso com os domínios príons (o qual é conhecidamente recente) e os domínios ADHs (que também tem sua origem antiga bem conhecida). Esses domínios foram escolhidos, porque eles foram profundamente estudados o que resultou em uma vasta literatura sobre eles.

Nessa abordagem, utilizamos os resultados obtidos pelo Male-PSI-BLAST com algumas modificações no banco de dados utilizado para o alinhamento. A primeira delas foi usada somente proteínas de organismos com genomas completos. A segunda modificação só foi utilizada sequências de um mesmo sequenciamento. Na terceira modificação as sequências idênticas foram removidas do banco de dados, neste procedimento se utilizou do BD Uniparc do UniProt para realizar este procedimento. E por fim, foram removidas as sequências que continham aminoácidos ambíguos B (asparagina ou ácido aspártico), Z (glutamina ou ácido glutâmico), J (leucina ou isoleucina) e X (aminoácidos inespecífico ou desconhecido).

Com o banco de dados refeito, os príons P04156 e Q9UKY0, e as ADH's, P00325, P00326, P07327, P08319, P11766, P40394, Q00796, Q08257 e Q14914, foram submetidos à busca de ortólogos com o Male-PSI-BLAST. Em seguida, foi feita uma abordagem de todos contra todos utilizando o aplicativo PHYL-SPST. E assim, foi produzido o histograma evidenciando os grupos mais correlacionados. Os príons apresentaram um padrão descontínuo com módulos distinguíveis quando comparados às ADH's (mostrados nas Figuras 9 e 10).



**Figura 9** - Histograma da matriz de distância todos contra todos dos príons (A) P04156 e (B) Q9UKY0 realizado pelo PHYL-SPST



**Figura 10** - Histograma da matriz de distância todos contra todos das ADHs (A) Q08257, (B) Q14914, (C) Q00796, (D) P11766, (E) P40394, (F) P00325, (G) P00326, (H) P07327 e (I) P08319 realizado pelo PHYL-SPST.

## 5.- Discussão

### 5.1.- Análise filoestratigráfica das entradas CATH

A árvore da vida exibe uma vasta diversidade biológica e podemos considerar que a variação genômica segue essa diversidade fenotípica podendo até mesmo ser considerada subjacente a ela. A evolução dos genomas é impulsionada por várias forças, tais como modificações em sequências codificantes ou não codificantes, além do surgimento de novos genes (CHEN et al, 2013). Programas utilizados para hipotetizar a homologia fazem uma grande suposição comum. Se duas sequências macromoleculares são semelhantes o suficiente em alguma medida, elas são homólogas, ou seja, elas hipoteticamente compartilham uma sequência ancestral. Os pesquisadores têm liberdade para decidir em qual limiar a semelhança deve ser utilizada como critério para o estabelecimento da homologia.

Em um organismo cada gene originou em certo momento da evolução, alguns são antigos, outros são mais recentes e outros estão em meio a processos de geração ou perda. Desde que foi reportado o primeiro gene recente, o *jingwei*, por Long e Langley (1993), os estudos de novos genes têm proporcionado uma nova perspectiva sobre os processos moleculares que originaram a estrutura inicial dos genes (KAESSMANN, 2010; KAESSMANN et al, 2009), a evolução das funções moleculares (TAUTZ & DOMAZET-LOŠO, 2010; DING et al, 2010), além das taxas globais dos padrões e mecanismos de origem de um novo gene (LONG et al, 2003; ZHANG et al 2011).

Abordagens filoestratigráficas têm sido utilizadas de forma produtiva em estudos que variam desde análise de famílias específicas de genes à análise estatística em escala genômica. Os estudos de famílias específicas de genes focam a origem de genes em relação, por exemplo, a uma doença específica, a uma via metabólica ou ao estágio embrionário (DOMAZET-LOSO & TAUTZ, 2008; DOMAZET-LOSO & TAUTZ 2010). As análises estatísticas em escala genômica procuram, por sua vez, por parametrizações gerais, relacionando a origem do gene com, por exemplo, taxas de substituições, expressão gênica, tamanho ou variabilidade (ALBÀ & CASTRESANA, 2005; WOLF et al. 2009; CAI & PETROV, 2010; PRAT et al. 2009).

Aqui reportamos como se deu a origem dos domínios do BD CATH com base nas proteínas que continham os domínios catalogados nos grupos do BD UEKO. E essa é a primeira iniciativa, em nosso conhecimento, para mapear domínios conhecidos de proteínas humanas e determinar a primeira ocorrência deles ao longo da linhagem ancestral dos

domínios CATH. Embora a maior parte dos domínios de superfamília do CATH, 52,6%, apareça em proteínas partilhadas com as bactérias e/ou archaea, alguns deles são recentes. A origem de alguns clados mostra uma evidente produção de novos domínios em grupos, tais como, os eucariotos, os eumetazoários e os Euteleostomi (Figura 2).

Cabe enfatizar alguns pontos que diferenciam o método aqui aplicado com os demais trabalhos que utilizaram a abordagem filoestratigráfica. Estudos filoestratigráficos, em sua ampla maioria, utilizam o BLAST para hipotetizar a homologia entre as sequências. Nesta abordagem o parâmetro *E-value* foi ajustado de forma mais permissiva, geralmente com limiar de 1E-3 (DOMAZET-LOSO & TAUTZ, 2003). Estes estudos geralmente não focam em domínios.

Estudos filoestratigráficos com domínios geralmente são utilizados para indicar a presença e a ausência dos dobramentos, estruturas secundárias, em cada genoma (LIN & GERSTEIN, 2000) ou o número de cópias de um determinado dobramento em cada genoma (CAETANO-ANOLLÉS & CAETANO-ANOLLÉS, 2003). O padrão de ocorrência é então utilizado para gerar árvores filogenéticas com topologias possíveis sugerindo que o modelo subjacente para construir tais árvores das novas dobras ou superfamílias que originaram nos genomas se mantém.

Poucos estudos fazem busca pela origem das proteínas utilizando os bancos de dados de ortólogos existentes, tais como o KEGG. Donnard e seus colaboradores (2011), por abordagem de mineração de texto, construíram a via regulatória do desenvolvimento embrionário pré-implantação e para a análise de origem dos genes pertencente a esta via foi utilizado o BD KEGG. Recentemente, Kirsch e Chechik (2016), em um estudo que visava identificar a relação entre o padrão de expressão dos genes de diferentes regiões do cérebro humano com a origem deles, também utilizaram da mesma estratégia que utilizamos. Em seu método para busca de origem das proteínas os autores também utilizaram o BD KEGG. Foi somente na fase final da redação dessa tese que tomamos conhecimento dessa publicação, feita há alguns dias. Esses autores alegam ter verificado correlações positivas para os parâmetros estudados.

Aqui utilizamos os domínios CATH para fazer busca por origem destes domínios em uma abordagem pouco diferenciada da proposta por Domazet-Loso e Tautz (2003). O uso do BD UEKO trouxe ganhos em eficiência, pois não foi necessário fazer busca por ortólogos com o BLAST. Além disso, agora nosso banco de dados local possui a origem para todos os domínios CATH, o que, por sua vez, possibilitará estudos futuros com diferentes abordagens

relacionando a origem de domínios com diferentes aspectos. Em outro momento, também realizamos a busca por sequências similares com a ferramenta *SeedServer* para comparar os resultados obtidos, que por sua vez acrescenta robustez para nossas análises. Em seguida, realizamos a busca com Male-PSI-BLAST para relacionar proteínas mais recentes com as mais antigas e destacar sua possível relação. Entre os domínios recentes, aqui encontrados, está o domínio Prón, que por suas características únicas como proteína patogênica, tais quais, as várias formas de encefalopatias espongiformes transmissíveis em mamíferos, é provavelmente uma das proteínas mais estudadas (VAN RHEEDE et al, 2003). De fato, devido ao amplo estudo com essa família de proteínas desde os anos iniciais da década 2000, já se conhecia bem a amplitude de organismos que continham os príons, tais como, aves, répteis, anfíbios e possivelmente peixes, em adição a todos os mamíferos (AGUZZI et al, 2008). Colocado nos termos filoestratigráficos, isso quer dizer, que a origem do Prón se deu no organismo que deu origem aos Tetrápodas, quando não consideramos a existência de prón em peixe, ou no organismo que deu origem aos Euteleostomi, quando consideramos que existe prón em peixes.

Príons parecem proteger contra a morte celular programada e a apoptose mediado pela proteína BAX (VAN RHEEDE et al, 2003; BOUNHAR et al, 2001). Os príons são proteínas de ligação ao cobre que podem ter atividade de superóxido dismutase. Portanto, a função do prón seria proteger contra danos oxidativos e contribuir para a homeostase sináptica (BROWN et al, 1999; BROWN, 2001), visto que a superóxido dismutase participa desses papéis. A exposição aos íons Cu<sup>+2</sup> promove a endocitose dos príons (PERERA et al, 2000). Além disso, existe evidência que o prón é um receptor de superfície celular para transdução de sinal acoplado à tirosina quinase Fyn (MOUILLET-RICHARD et al, 2000). O prón se alterna em estar presente entre a superfície celular e as vesículas de endossomos revestidas por clatrinas, como fazem a maioria dos receptores de superfície celular. É uma mudança conformacional no prón que dá origem à sua forma patogênica.

Com a abordagem aplicada nesse trabalho, conseguimos verificar que houve duas épocas importantes relacionadas ao aparecimento e à evolução do domínio prón. A primeira época importante foi o surgimento do próprio domínio prón quando da separação do clado Tetrapoda dos demais cordados, sendo que o domínio característico do prón formou-se a partir do arcabouço proteico de uma “ribonucleoproteína nuclear heterogênea”. É importante enfatizar que, somente com as sequências dos domínios do BD CATH, não conseguimos

detectar o príon de peixes, o que fez com que a origem do príon ficasse em Tetrapoda e não em Euteleostomi, como teria ocorrido caso tivéssemos encontrado os ditos príons de peixe.

Outros pesquisadores também já questionaram a existência de príons em peixes, dada a sua tênue similaridade de sequência com os príon dos outros organismos aliada a falta de testes bioquímicos que comprovem de fato a sua existência (AGUZZI et al, 2008). Além disso, a abordagem que utilizou o Male-PSI-BLAST com a sequência total das proteínas que contêm os domínios príon conseguiu recuperar os príons presumidos de peixes, o que mostra que os príons de peixes estão mais relacionados com a sua região extradomínio das proteínas.

Outro ponto que é preciso que se ressalte, é que não há estudos que relacionam a origem dos príons com as ribonucleoproteínas nucleares heterogêneas. Ehsani e seu grupo (2011) mostraram que os príons podem ter surgido de elementos retrotransponíveis da família ZIP, mas nossa abordagem com Male-PSI-BLAST detectou similaridade com proteínas pertencentes a diferentes grupos de ortólogos de ribonucleoproteínas nucleares heterogêneas do KEGG em grupos de organismos cuja ancestralidade com os seres humanos se dá em Euteleostomi e outros clados mais basais. Outro estudo também sugere esta homologia, King e seu grupo (2012), utilizando algoritmos de detecção de príons, mostraram que um grupo de proteínas humanas ligadoras ao RNA contêm domínios presumidos de príons.

Há de se destacar também o ponto sobre a preocupação da utilização do BLASTp, discutidas principalmente por Moyer e Zhang (2014), para capturar por sequências de homólogos mais remotos e, consequentemente, realizar a datação errada para origem de um gene. Aqui em nossas análises as três abordagens utilizadas mostraram resultados similares no estudo com os domínios príons que é corroborada com a diversidade de organismos já conhecida (AGUZZI et al, 2008). A diferença se deu com o uso do Male-PSI-BLAST utilizando a sequência completa das proteínas (Q53YK7, P04156, Q5U0K3, B2R5Q9, Q6FGR8, A1YVW6, B4DJ65, B2NI04, B4DDS1, B2NI05, B4DI53, O75942, e Q6SES1) que contêm o domínio príon atribuindo uma tênue relação entre o príon e ribonucleoproteínas nucleares heterogêneas.

## 5.2.- Análise filoestratigráfica das entradas Pfam

O Pfam é um banco de dados amplamente usado de famílias proteicas e, apesar de conter dados advindos de métodos probabilísticos que utilizam cadeias ocultas de Markov, os quais são conhecidos por serem mais sensíveis para busca de homólogos mais remotos, os padrões de surgimento das entradas do Pfam foram similares àqueles obtidos com as análises dos domínios CATH mostrados na seção 5.1 deste trabalho em que utilizamos a ferramenta BLAST e os grupos de ortólogos do UEKO. Com picos proeminentes de surgimentos de domínios no surgimento dos organismos celulares, dos Eucariotos e dos Euteleostomi.

Algumas teorias sugerem eventos moleculares que poderiam explicar estes picos proeminentes propiciando o surgimento dos domínios. RIVERA & LAKE (2004) defendem a teoria quimérica de fusão genômica que deu a origem aos Eucariotos e DEHAL & BOORE (2005) sugerem que pode ter havido duas duplicações genômicas consecutivas que deram origem aos Euteleostomi, o que já tinha sido hipotetizado na década de 1960 e 1970 pelo grupo de Susumu Ohno (OHNO et al, 1967, OHNO, 1970) através de estudos citogenéticos.

Algumas entradas do Pfam, PF15399, PF00523, PF0061, PF15143, PF15263, PF00500, e PF14313, apresentaram LCA bastante recente, existindo somente entre os humanos. Essas entradas, em sua maioria, estão associadas a partículas virais e domínios com função desconhecida. O mecanismo de transferência lateral é comum entre os procariotos e os eucariontes simples. Este processo permite aos organismos compartilhar rapidamente um gene de resistência a antibióticos, por exemplo. Entretanto, ainda é contestada a existência de transferência lateral entre organismos complexos, tais como os primatas.

O estudo realizado por CRISP et al. (2015) mostrou que pode haver 145 genes em humanos provenientes de transferência lateral. Embora este não seja o foco deste trabalho, principalmente pelo fato que excluímos vírus de nossas análises, conseguimos observar alguns desses domínios muito recentes, que podem ter vindo de transferência lateral, já que existem em humanos e são compartilhados entre as demais espécies de eucariotos cujos genomas foram completamente sequenciados. Alternativamente, estes domínios são frutos de uma anotação errada obtida a partir de contaminação das amostras de DNA empregadas no sequenciamento dos genomas humanos. Com relação à hipótese de contaminação, os estudos de metagenômica em amostras de tecidos humanos mostram cada vez mais a presença de extensas floras microbianas (CHO & BLASER, 2012, por exemplo), algo que não era esperado em passado recente quando houve a grande disseminação de sequenciamento de genomas humanos.

Outro ponto evidenciado por nossas análises foi que as arquiteturas tendem a ser mais recentes que os domínios que as compõem, embora haja arquiteturas que surgiram na mesma época que seu último domínio. Além disso, parece não haver um viés de utilização de domínios relacionado ao surgimento deles, o que é esperado pela teoria que prevê a acomodação, por seleção natural, do viés de utilização de códons da região transferida ao viés presente no organismo hospedeiro, especialmente depois de decorrido um tempo suficientemente longo.

Pesquisas têm tentado agregar mais características dos genomas que vão além da informação contida na sequência em si, atributos como composição dos genes e ordem dos genes também são utilizadas para construir árvores filogenéticas que, geralmente, combinam o sinal filogenético com os sinais que refletem os estilos de vida dos organismos, em comparação, por exemplo, a perda de genes em paralelo de parasitas (WOLF et al, 2004; SNEL et al, 2005; WANG & CAETANO-ANOLLÉS, 2006). Particularmente, as combinações de domínios têm sido usadas como uma característica filogenética para resolver problemas filogenéticos difíceis, por exemplo, o dilema Coelomata-Ecdysozoa na evolução de animais (WOLF et al, 2004; WANG & CAETANO-ANOLLÉS, 2006).

Aqui, utilizamos os índices de promiscuidade dos domínios Pfam e aqueles relacionados com os dados de origem dos mesmos, no proteoma humano. De fato, foi mostrado que existe uma baixa correlação entre os índices de promiscuidade com o tempo de origem dos domínios. Entretanto, é razoável supor que domínios mais antigos são susceptíveis a desempenhar um papel dominante na organização das arquiteturas humanas atuais, devido ao fato que os domínios mais antigos apresentaram os índices de promiscuidade mais altos.

Outros autores, utilizando uma classe específica de proteínas da matriz extracelular, mostraram que domínios antigos pertencentes a esta classe também possuem índices de promiscuidades mais elevados do que domínios originados em clados mais recentes (CROMAR et al, 2014).

Basu e seus colaboradores (2008) mostraram que é possível haver uma íntima relação entre o aumento do número de domínios promíscuos e o aumento da complexidade do organismo, além de uma forte dependência linear entre o número de domínios promíscuos e o número de tipos de domínio. Ainda neste estudo, os autores mostraram que a função biológica pode também atuar na seleção de arquiteturas com domínios promíscuos, sendo as transduções

de sinais, estruturas extracelulares e sinalização célula a célula, as funções moleculares que mais contêm domínios promíscuos (BASU et al, 2008).

Existe grande interesse científico em estudos que visam definir a contribuição dos diferentes mecanismos genéticos pelos quais as arquiteturas de domínios foram formadas no nível genômico. Estudos com proteínas multidomínios sugerem que as arquiteturas de domínios se formaram em grande parte por mecanismos de fusão gênica, expansão de repetições e perdas de domínios em regiões terminais (BORNBERG-BAUER e ALBÀ, 2013).

Ademais, outros resultados têm sugerido que mudanças de posições de domínios em arquiteturas existentes juntamente com uma taxa relativamente baixa, mas existente, de surgimento de novos domínios é um modo favorável de adaptação na qual a evolução molecular pode rapidamente reutilizar módulos estabelecidos sem destruir completamente uma rede celular intrincada e util.

Exemplos disso são propostos nos estudos subsequentes em que são mostradas as relações entre os mecanismos moleculares e a formação da arquitetura. Fusões de genes podem trazer domínios que catalisam fases sucessivas de uma via metabólica em estreita proximidade espacial e sob o controle de transcrição conjunta (FALB et al, 2008). Fusão no meio de domínio pode implicar pressões seletivas e possivelmente pode estar relacionada à susceptibilidade a doença (ROGER & HART, 2012). Em várias famílias de fatores de transcrição, o recrutamento de um novo domínio efetor iniciou o processo para uma mudança funcional que, posteriormente, levou a uma expansão de uma nova subfamília (AMOUTZIAS et al, 2004; GRAMZOW et al, 2010). Domínios podem adquirir novas funções se recrutados por famílias existentes e estes rearranjos podem acelerar subfuncionalização da sinalização (JIN et al, 2009; GRASSI et al, 2010).

Associado a isto tudo, ainda há o fato que semelhança entre duas proteínas no nível de arquitetura de domínios pode agregar valor a métodos de anotação de sequências baseadas em similaridade e de classificação de famílias ou em subfamílias, além dos ganhos evidentes em velocidade de processamento, dadas as características das sequências que neste nível são compostas de domínios ao invés de nucleotídeos ou aminoácidos.

Entretanto, ainda é escasso o número de ferramentas as quais analisam a semelhança entre o conteúdo de domínios proteicos em arquiteturas de domínios, entre os notórios algoritmos se destacam: o RADS/RAMPAGE, em que eles realizam um alinhamento global

entre as sequências de domínios em cada proteína (TERRAPON et al, 2014); o MDAT utiliza uma matriz de similaridade de domínio para pontuar pares de domínio e alinha as arquiteturas de domínio através de um método de alinhamento progressivo (KEMENA et al, 2015); e o ADASS que compara e classifica as arquiteturas de domínio por reconhecer semelhança entre as arquiteturas de domínio mesmo em proteínas cuja similaridade entre as sequências seja tênue, por métodos de análises de sequência livre de alinhamento (SYAMALADEVI et al, 2013). Logicamente as vantagens e desvantagens inerentes a cada método, também são as mesmas verificadas em métodos semelhantes de análises de sequências de nucleotídeos ou aminoácidos.

Outro método pelos quais os pesquisadores utilizam para conduzir a uma melhor compreensão dos processos que criam novas arquiteturas de proteínas são as redes de coocorrências, nas quais são estudados, por exemplo, domínios que ocorrem em diferentes arquiteturas. Wuchty (2001) estudando rede de coocorrência a partir dos dados de diferentes bases de dados de domínios proteicos mostrou que a rede resultante não possui as características de gráficos aleatórios.

### **5.3.- Estudo de caso: príons e ADHs**

Este estudo foi uma tentativa de remontar o caminho percorrido pelo Male-PSI-BLAST ao inferir homologia entre as sequências. Nesta abordagem se esperava que se o Male-PSI-BLAST conseguisse inferir homologia de um domínio ancestral, mas relacionado, a distribuição de estimativas de distâncias entre pares de sequências de resíduos de aminoácidos apresentaria com múltiplas modas. E caso o Male-PSI-BLAST não conseguisse inferir homologia de um grupo ancestral relacionado à distribuição se apresentaria com uma única moda.

Em seção anterior (5.1), mostramos que o Male-PSI-BLAST conseguiu inferir homologia entre os príons e as RBP's, as quais possivelmente podem estar relacionadas aos príons quanto à origem deles. Em contrapartida, as ADH's pertencem a uma família de proteínas ubíqua entre os organismos mais diversos.

O aspecto multimodal apresentado no histograma em que foram analisadas as sequências dos príons pode ser possivelmente atribuído a uma descontinuidade devido ao surgimento independente do domínio em, pelo menos, duas ocasiões. Príons são conhecidos por ter, em sua composição, regiões com poliglutaminas que são regiões de baixa

complexidade que, consequentemente, geram falsos positivos para a hipótese de homologia pelo simples alinhamento entre sequências idênticas de poliglutaminas. De fato, repetições nucleotídicas CpG (CAG ou CTG), conhecidas como ilhas CpG, são regiões reconhecidas como sendo instáveis mutacionalmente. A molécula da Huntingtina, quando possui acima de um determinado número de repetições CAG pode causar a doença de Huntington (BUTLAND et al, 2007). Há outras proteínas com trechos poliglutamínicos que são mutacionalmente instáveis, como a FMRP, envolvida na síndrome do X-frágil (KHANDJIAN, 1999).

O aumento dos sequenciamentos em larga escala de todo o genoma tem permitido aumentar e refinar as análises das tendências gerais da evolução das proteínas. A idade dos genes tem grande poder de evidenciar correlações entre a história dos genes e suas funções no contexto da evolução dos organismos, da dinâmica ecológica e dos processos bioquímicos.

Entretanto, um problema fundamental nestas análises é devido ao fato que os genes não têm uma idade única e bem definida. E o problema se torna ainda pior para genes que possuem estruturas complexas, nos quais existe um arranjo de domínios em que pode haver vários eventos que poderiam ser considerados como origem do gene. O gene *Jingwei* se originou pela combinação dos mecanismos: embaralhamento de exons, retrotransposição e duplicação (LONG & LANGLEY, 1993). Para o *Jingwei*, todos estes mecanismos moleculares foram identificados por comparação de sequência e de funcionalidade, dada a semelhança entre *Jingwei* e seus genes parentais.

Aliado a isso, ainda existe a necessidade de métodos estatísticos para identificar as tendências dos dados evolutivos. Aqui, nos propusemos uma abordagem na qual é possível distinguir eventos evolutivos recentes com base na continuidade entre os membros da família gênica encontrada pelo Male-PSI-BLAST.

De modo geral, nossos resultados relembram a teoria de bricolagem molecular proposta por Jacob (1977), em que ele afirma que, por toda a vida são compartilhadas as mesmas moléculas orgânicas e vias metabólicas semelhantes, é mais provável que as novas proteínas funcionais surjam a partir de um rearranjo de elementos genéticos em oposição ao surgimento de proteínas pelo mecanismo *de novo*.

## 6.- Conclusões

Utilizando a abordagem de detecção da origem de domínios de superfamília do CATH com base na diversidade de organismos presentes em grupos de ortólogos, verificou-se que existem estágios para origem de novos genes ao longo da linhagem descente-ancestral dos *H. sapiens*, embora a maior parte dos domínios apareça em proteínas partilhadas com as bactérias e/ou archaea.

Curiosamente, tanto para as análises com domínios de superfamília do CATH, baseadas em grupos de ortólogos do UEKO, quanto para as análises com os domínios Pfam, que utiliza métodos baseados em Cadeia Oculta de Markov, mostraram que as origens dos clados Eucarioto e Euteleostomi, além da origem da vida, são proeminentes para origem de novos domínios.

O domínio Príon está entre os domínios recentes encontrados em humanos e possivelmente formou-se a partir do arcabouço proteico de uma “ribonucleoproteína nuclear heterogênea”. Este fato é corroborado pelo estudo realizado por King e seu grupo (2012), que utilizando algoritmos de detecção de príons, mostrou que um grupo de proteínas humanas ligantes a RNA contêm domínios presumidos de príons, mas que deixaram de relatar que este conjunto de proteínas poderia estar relacionado quanto à origem da proteína príon.

Nossas análises também evidenciaram que as arquiteturas tendem a ser mais recentes que os domínios que as compõem, embora haja arquiteturas que surgiram na mesma época que seu domínio mais recente constituinte.

Além disso, não há algum viés de utilização de domínios relacionado ao surgimento deles. E, por conseguinte, os domínios mais antigos tendem a ser mais promíscuos, embora haja domínios antigos que não sejam tão promíscuos.

Por fim, a descontinuidade mostrada na distribuição de estimativas de distâncias entre pares de sequências de resíduos de aminoácidos obtidas a partir dos resultados dos experimentos com Male-PSI-BLAST tem potencial de mostrar grupos antigos, mas relacionados, com relação a sua origem.

## 7.- Referências bibliográficas

- AGUZZI, Adriano; BAUMANN, Frank; BREMER, Juliane. The prion's elusive reason for being. **Annu. Rev. Neurosci.**, v. 31, p. 439-477, 2008.
- ALBÀ SOLER, Mar; CASTRESANA, José. On homology searches by protein Blast and the characterization of the age of genes. **BMC Evol. Biol.** 2007; 7: 53, 2007.
- ALEXEYENKO, Andrey et al. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. **Bioinformatics**, v. 22, n. 14, p. e9-e15, 2006.
- ALTENHOFF, Adrian M.; DESSIMOZ, Christophe. Inferring orthology and paralogy. **Evolutionary Genomics: Statistical and Computational Methods, Volume 1**, p. 259-279, 2012.
- ALTSCHUL, Stephen F. et al. Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403-410, 1990.
- ALTSCHUL, Stephen F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic acids research**, v. 25, n. 17, p. 3389-3402, 1997.
- AMOUTZIAS, Gregory D. et al. Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. **EMBO reports**, v. 5, n. 3, p. 274-279, 2004.
- ASHTON, Elaine. The Timeline of Perl and its Culture. 2001.
- BARBOSA-SILVA, Adriano et al. Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence. **BMC bioinformatics**, v. 9, n. 1, p. 1, 2008.
- BERGTHORSSON, Ulfar et al. Widespread horizontal transfer of mitochondrial genes in flowering plants. **Nature**, v. 424, n. 6945, p. 197-201, 2003.
- BETRÁN, Esther et al. Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene. **Molecular biology and evolution**, v. 19, n. 5, p. 654-663, 2002.
- BETRÁN, Esther; LONG, Manyuan. Expansion of genome coding regions by acquisition of new genes. **Genetica**, v. 115, n. 1, p. 65-80, 2002.
- BIÉMONT, Christian. A brief history of the status of transposable elements: from junk DNA to major players in evolution. **Genetics**, v. 186, n. 4, p. 1085-1093, 2010.

- BORNBERG-BAUER, Erich; ALBÀ, M. Mar. Dynamics and adaptive benefits of modular protein evolution. **Current opinion in structural biology**, v. 23, n. 3, p. 459-466, 2013.
- BOUCHER, Yan et al. Lateral gene transfer and the origins of prokaryotic groups. **Annual review of genetics**, v. 37, n. 1, p. 283-328, 2003.
- BOUNHAR, Younes et al. Prion protein protects human neurons against Bax-mediated apoptosis. **Journal of Biological Chemistry**, v. 276, n. 42, p. 39145-39149, 2001.
- BROSius, J. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. **Gene**, v. 238, n. 1, p. 115-134, 1999.
- BROSius, Jürgen. The contribution of RNAs and retroposition to evolutionary novelties. In: **Origin and Evolution of New Gene Functions**. Springer Netherlands, 2003. p. 99-116.
- BROWN, David R. et al. Normal prion protein has an activity like that of superoxide dismutase. **Biochemical Journal**, v. 344, n. 1, p. 1-5, 1999.
- BROWN, David R. Prion and prejudice: normal protein and the synapse. **Trends in neurosciences**, v. 24, n. 2, p. 85-90, 2001.
- BUTLAND, Stefanie L. et al. CAG-encoded polyglutamine length polymorphism in the human genome. **BMC genomics**, v. 8, n. 1, p. 1, 2007.
- CAETANO-ANOLLÉS, Gustavo; CAETANO-ANOLLÉS, Derek. An evolutionarily structured universe of protein architecture. **Genome research**, v. 13, n. 7, p. 1563-1571, 2003.
- CAI, Jing et al. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. **Genetics**, v. 179, n. 1, p. 487-496, 2008.
- CARDOSO-MOREIRA, Margarida; LONG, Manyuan. The origin and evolution of new genes. In: ANISIMOVA, M. **Evolutionary Genomics: Statistical and Computational Methods**. Suiça: Humana Press; 2012. v. 2. Cap 7, p. 161-86
- CHEN, Liangbiao; DEVRIES, Arthur L.; CHENG, Chi-Hing C. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. **Proceedings of the National Academy of Sciences**, v. 94, n. 8, p. 3817-3822, 1997.

- CHEN, Sidi; KRINSKY, Benjamin H.; LONG, Manyuan. New genes as drivers of phenotypic evolution. **Nature Reviews Genetics**, v. 14, n. 9, p. 645-660, 2013.
- CHENG, Jianlin et al. SCRATCH: a protein structure and structural feature prediction server. **Nucleic acids research**, v. 33, n. suppl 2, p. W72-W76, 2005
- CHO, Ilseung; BLASER, Martin J. The human microbiome: at the interface of health and disease. **Nature Reviews Genetics**, v. 13, n. 4, p. 260-270, 2012.
- CHRISTIANSEN, Tom et al. An Overview of Perl. In: \_\_\_\_\_. Programming Perl: Unmatched power for text processing and scripting. EUA: O'Reilly Media, 2012. Cap 1.
- COELHO Jr; SOUZA, DT; RIBEIRO, DT & ORTEGA, JM The human proteome resembles a matrioshka but some protein architectures are more popular in each member of the set. Res. 58º. Cong. Brasil. Genet., 2012.
- COURSEAUX, Anouk; NAHON, Jean-Louis. Birth of two chimeric genes in the Hominidae lineage. **Science**, v. 291, n. 5507, p. 1293-1297, 2001.
- CRISP, Alastair et al. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. **Genome Biol**, v. 16, n. 50, p. 10.1186, 2015.
- CROMAR, Graham et al. New Tricks for "Old" Domains: How Novel Architectures and Promiscuous Hubs Contributed to the Organization and Evolution of the ECM. **Genome biology and evolution**, v. 6, n. 10, p. 2897-2917, 2014.
- CUFF, Alison L. et al. Extending CATH: increasing coverage of the protein structure universe and linking structure with function. **Nucleic acids research**, v. 39, n. suppl 1, p. D420-D426, 2011.
- DE KONING, Audrey P. et al. Lateral gene transfer and metabolic adaptation in the human parasite *Trichomonas vaginalis*. **Molecular biology and evolution**, v. 17, n. 11, p. 1769-1773, 2000.
- DEHAL, Paramvir; BOORE, Jeffrey L. Two rounds of whole genome duplication in the ancestral vertebrate. **PLoS Biol**, v. 3, n. 10, p. e314, 2005.
- DING, Yun et al. A young Drosophila duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. **PLoS Genet**, v. 6, n. 12, p. e1001255, 2010.

- DING, Yun; ZHOU, Qi; WANG, Wen. Origins of new genes and evolution of their novel functions. **Annual Review of Ecology, Evolution, and Systematics**, v. 43, p. 345-363, 2012.
- DOMAZET-LOŠO, Tomislav; TAUTZ, Diethard. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. **Nature**, v. 468, n. 7325, p. 815-818, 2010.
- DOMAZET-LOŠO, Tomislav; TAUTZ, Diethard. An ancient evolutionary origin of genes associated with human genetic diseases. **Molecular biology and evolution**, v. 25, n. 12, p. 2699-2707, 2008.
- DOMAZET-LOSO, Tomislav; TAUTZ, Diethard. An evolutionary analysis of orphan genes in *Drosophila*. **Genome Research**, v. 13, n. 10, p. 2213-2219, 2003.
- DONNARD, Elisa et al. Preimplantation development regulatory pathway construction through a text-mining approach. **BMC genomics**, v. 12, n. Suppl 4, p. S3, 2011.
- EDDY, Sean R... Profile hidden Markov models. **Bioinformatics**, v. 14, n. 9, p. 755-763, 1998.
- EHSANI, Sepehr et al. Evidence for retrogene origins of the prion gene family. **PloS one**, v. 6, n. 10, p. e26800, 2011.
- ELHAIK, Eran; SABATH, Niv; GRAUR, Dan. The “inverse relationship between evolutionary rate and age of mammalian genes” is an artifact of increased genetic distance with rate of evolution and time of divergence. **Molecular biology and evolution**, v. 23, n. 1, p. 1-3, 2006.
- EVINE, Mia T. et al. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. **Proceedings of the National Academy of Sciences**, v. 103, n. 26, p. 9935-9939, 2006.
- FALB, Michaela et al. Metabolism of halophilic archaea. **Extremophiles**, v. 12, n. 2, p. 177-196, 2008.
- FINN, Robert D. et al. The Pfam protein families database. **Nucleic acids research**, v. 38, n. suppl 1, p. D211-D222, 2012.
- FINN, Robert D.; CLEMENTS, Jody; EDDY, Sean R. HMMER web server: interactive sequence similarity searching. **Nucleic acids research**, p. gkr367, 2011.

- FITCH, Walter M. Distinguishing homologous from analogous proteins. **Systematic Biology**, v. 19, n. 2, p. 99-113, 1970.
- FONG, Jessica H. et al. Modeling the evolution of protein domain architectures using maximum parsimony. **Journal of molecular biology**, v. 366, n. 1, p. 307-315, 2007.
- FREEMAN, Victor J. Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. **Journal of bacteriology**, v. 61, n. 6, p. 675, 1951.
- FROST, Laura S. et al. Mobile genetic elements: the agents of open source evolution. **Nature Reviews Microbiology**, v. 3, n. 9, p. 722-732, 2005.
- GEER, Lewis Y. et al. CDART: protein homology by domain architecture. **Genome research**, v. 12, n. 10, p. 1619-1623, 2002.
- GILBERT, Walter. Why genes in pieces?. **Nature**, v. 271, n. 5645, p. 501, 1978.
- GILBERT, Walter; DE SOUZA, Sandro J.; LONG, Manyuan. Origin of genes. **Proceedings of the National Academy of Sciences**, v. 94, n. 15, p. 7698-7703, 1997.
- GOODIER, John L.; OSTERTAG, Eric M.; KAZAZIAN JR, Haig H. Transduction of 3'-flanking sequences is common in L1 retrotransposition. **Human molecular genetics**, v. 9, n. 4, p. 653-657, 2000.
- GRAMZOW, Lydia; RITZ, Markus S.; THEIßEN, Günter. On the origin of MADS-domain transcription factors. **Trends in genetics**, v. 26, n. 4, p. 149-153, 2010.
- GRASSI, Luigi et al. Identity and divergence of protein domain architectures after the yeast whole-genome duplication event. **Molecular BioSystems**, v. 6, n. 11, p. 2305-2315, 2010.
- HALDANE, J. B. S. The part played by recurrent mutation in evolution. **American Naturalist**, p. 5-19, 1933.
- HEDGES, S. Blair; DUDLEY, Joel; KUMAR, Sudhir. TimeTree: a public knowledge-base of divergence times among organisms. **Bioinformatics**, v. 22, n. 23, p. 2971-2972, 2006.
- HENIKOFF, Steven; AHMAD, Kami; MALIK, Harmit S. The centromere paradox: stable inheritance with rapidly evolving DNA. **Science**, v. 293, n. 5532, p. 1098-1102, 2001.
- HENNIG, Christian. Package ‘fpc’. 2014.
- HOTOPP, Julie C. Dunning et al. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. **Science**, v. 317, n. 5845, p. 1753-1756, 2007.

- HUFTON, Andrew L. et al. Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention. **Genome research**, v. 19, n. 11, p. 2036-2051, 2009.
- HUGHES, Austin L. **Adaptive evolution of genes and genomes**. Oxford University Press, USA, 2000
- ITOH, Masumi et al. Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. **Genome biology**, v. 8, n. 6, p. R121, 2007.
- JACOB, François. Evolution and tinkering. 1977.
- JAVAUD, Christophe et al. The fucosyltransferase gene family: an amazing summary of the underlying mechanisms of gene evolution. In: **Origin and Evolution of New Gene Functions**. Springer Netherlands, 2003. p. 157-170.
- JIN, Jing et al. Eukaryotic protein domains as functional units of cellular evolution. **Sci Signal**, v. 2, n. 98, p. 1-18, 2009.
- KAESSMANN, Henrik. Origins, evolution, and phenotypic impact of new genes. **Genome research**, v. 20, n. 10, p. 1313-1326, 2010.
- KAESSMANN, Henrik; VINCKENBOSCH, Nicolas; LONG, Manyuan. RNA-based gene duplication: mechanistic and evolutionary insights. **Nature Reviews Genetics**, v. 10, n. 1, p. 19-31, 2009.
- KAPITONOV, Vladimir V.; JURKA, Jerzy. A universal classification of eukaryotic transposable elements implemented in Repbase. **Nature Reviews Genetics**, v. 9, n. 5, p. 411-412, 2008.
- KAWASHIMA, Shuichi; KANEHISA, Minoru. AAindex: amino acid index database. **Nucleic acids research**, v. 28, n. 1, p. 374-374, 2000.
- KEELING, Patrick J.; PALMER, Jeffrey D. Horizontal gene transfer in eukaryotic evolution. **Nature Reviews Genetics**, v. 9, n. 8, p. 605-618, 2008.
- KEMENA, Carsten; BITARD-FEILDEL, Tristan; BORNBERG-BAUER, Erich. MDAT- Aligning multiple domain arrangements. **BMC bioinformatics**, v. 16, n. 1, p. 1, 2015.
- KEREN, Hadas; LEV-MAOR, Galit; AST, Gil. Alternative splicing and evolution: diversification, exon definition and function. **Nature Reviews Genetics**, v. 11, n. 5, p. 345-355, 2010.

- KHANDJIAN, Edouard W. Biology of the fragile X mental retardation protein, an RNA-binding protein. **Biochemistry and Cell Biology**, v. 77, n. 4, p. 331-342, 1999.
- KIMURA, Motoo. **The neutral theory of molecular evolution**. Cambridge University Press, 1984.
- KING, Oliver D.; GITLER, Aaron D.; SHORTER, James. The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease. **Brain research**, v. 1462, p. 61-80, 2012.
- KIRSCH, Lior; CHECHIK, Gal. On Expression Patterns and Developmental Origin of Human Brain Regions. **PLOS Comput Biol**, v. 12, n. 8, p. e1005064, 2016.
- KOONIN, Eugene V. Orthologs, paralogs, and evolutionary genomics 1. **Annu. Rev. Genet.**, v. 39, p. 309-338, 2005.
- LEONARDI, Florencia et al. Detecting phylogenetic relations out from sparse context trees. **arXiv preprint arXiv:0804.4279**, 2008.
- LI, Hongwei et al. Insight into role of selection in the evolution of polyglutamine tracts in humans. **PloS one**, v. 7, n. 7, p. e41167, 2012.
- LIN, Jimmy; GERSTEIN, Mark. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. **Genome Research**, v. 10, n. 6, p. 808-818, 2000.
- LONG, Manyuan et al. Relationship between “proto-splice sites” and intron phases: evidence from dicodon analysis. **Proceedings of the National Academy of Sciences**, v. 95, n. 1, p. 219-223, 1998.
- LONG, Manyuan et al. The origin of new genes: glimpses from the young and old. **Nature Reviews Genetics**, v. 4, n. 11, p. 865-875, 2003.
- LONG, Manyuan. Evolution of novel genes. **Current opinion in genetics & development**, v. 11, n. 6, p. 673-680, 2001.
- LONG, Manyuan; LANGLEY, Charles H. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. **Science**, v. 260, n. 5104, p. 91, 1993

- LONG, Manyuan; LANGLEY, Charles H. Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. **SCIENCE-NEW YORK THEN WASHINGTON-**, v. 260, p. 91-91, 1993.
- LONG, Manyuan; ROSENBERG, Carl; GILBERT, Walter. Intron phase correlations and the evolution of the intron/exon structure of genes. **Proceedings of the National Academy of Sciences**, v. 92, n. 26, p. 12495-12499, 1995.
- LORENC, Anna; MAKALOWSKI, Wojciech. Transposable elements and vertebrate protein diversity. **Genetica**, v. 118, n. 2-3, p. 183-191, 2003.
- LYNCH, Michael; FORCE, Allan. The probability of duplicate gene preservation by subfunctionalization. **Genetics**, v. 154, n. 1, p. 459-473, 2000.
- MAKALOWSKI, Wojciech. Genomic scrap yard: how genomes utilize all that junk. **Gene**, v. 259, n. 1, p. 61-67, 2000.
- MALIK, Harmit S.; HENIKOFF, Steven. Adaptive evolution of Cid, a centromere-specific histone in *Drosophila*. **Genetics**, v. 157, n. 3, p. 1293-1298, 2001.
- MANNING, Christopher D.; SCHÜTZE, Hinrich. **Foundations of statistical natural language processing**. Cambridge: MIT press, 1999.
- MARCHLER-BAUER, Aron et al. CDD: a Conserved Domain Database for the functional annotation of proteins. **Nucleic acids research**, v. 39, n. suppl 1, p. D225-D229, 2011.
- MARTIGNETTI, John A.; BROSIUS, JURGEN. BC200 RNA: a neural RNA polymerase III product encoded by a monomeric Alu element. **Proceedings of the National Academy of Sciences**, v. 90, n. 24, p. 11563-11567, 1993.
- MARTIGNETTI, John A.; BROSIUS, JURGEN. Neural BC1 RNA as an evolutionary marker: guinea pig remains a rodent. **Proceedings of the National Academy of Sciences**, v. 90, n. 20, p. 9698-9702, 1993.
- MASTON, Glenn A.; RUVOLO, Maryellen. Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. **Molecular Biology and Evolution**, v. 19, n. 3, p. 320-335, 2002.
- MCCARREY, John R. Evolution of tissue-specific gene expression in mammals. **BioScience**, v. 44, n. 1, p. 20-27, 1994.

- MCCARREY, John R. Molecular evolution of the human Pgk-2 retroposon. **Nucleic acids research**, v. 18, n. 4, p. 949-955, 1990.
- MCCARREY, John R. Nucleotide sequence of the promoter region of a tissue-specific human retroposon: comparison with its housekeeping progenitor. **Gene**, v. 61, n. 3, p. 291-298, 1987.
- MCCARTHY, Alun D.; HARDIE, D. Grahame. Fatty acid synthase—an example of protein evolution by gene fusion. **Trends in Biochemical Sciences**, v. 9, n. 2, p. 60-63, 1984.
- MOUILLET-RICHARD, S. et al. Signal transduction through prion protein. **Science**, v. 289, n. 5486, p. 1925-1928, 2000.
- MOYERS, Bryan A.; ZHANG, Jianzhi. Phylostratigraphic bias creates spurious patterns of genome evolution. **Molecular biology and evolution**, p. msu286, 2014.
- MULLER, H. J. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. **Genetica**, v. 17, n. 3, p. 237-252, 1935.
- NAGY, Alinda; BÁNYAI, László; PATTHY, László. Reassessing domain architecture evolution of metazoan proteins: major impact of errors caused by confusing paralogs and epiktologs. **Genes**, v. 2, n. 3, p. 516-561, 2011.
- NEKRUTENKO, Anton; LI, Wen-Hsiung. Transposable elements are found in a large number of human protein-coding genes. **TRENDS in Genetics**, v. 17, n. 11, p. 619-621, 2001.
- NURMINSKY, Dmitry I. et al. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. **Nature**, v. 396, n. 6711, p. 572-575, 1998.
- OCHIAI, K. et al. Inheritance of drug resistance (and its transfer) between *Shigella* strains and Between *Shigella* and *E. coli* strains. **Hihon Iji Shimpur,(in Japanese)**, v. 34, p. 1861, 1959.
- OCHMAN, Howard. Lateral and oblique gene transfer. **Current opinion in genetics & development**, v. 11, n. 6, p. 616-619, 2001.
- OHNO, Susumu et al. Diploid-tetraploid relationship among old-world members of the fish family Cyprinidae. **Chromosoma**, v. 23, n. 1, p. 1-9, 1967.
- OHNO, Susumu. Ancient linkage groups and frozen accidents. **Nature**, v. 244, p. 259-262, 1973.
- OHNO, Susumu. **Evolution by gene duplication**. Springer Science & Business Media, 1970.

- OWEN, Richard. **On the archetype and homologies of the vertebrate skeleton.** van Voorst, 1848.
- PATTHY, László. Exon shuffling and other ways of module exchange. **Matrix biology**, v. 15, n. 5, p. 301-310, 1996.
- PATTHY, László. **Protein evolution by exon-shuffling.** Springer, 1995.
- PAULDING, Charles A.; RUVOLO, Maryellen; HABER, Daniel A. The Tre2 (USP6) oncogene is a hominoid-specific gene. **Proceedings of the National Academy of Sciences**, v. 100, n. 5, p. 2507-2511, 2003.
- PERERA, W.; Sumudhu S.; HOOPER, Nigel M. Ablation of the metal ion-induced endocytosis of the prion protein by disease-associated mutation of the octarepeat region. **Current Biology**, v. 11, n. 7, p. 519-523, 2001.
- PICKERAL, Oxana K. et al. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. **Genome research**, v. 10, n. 4, p. 411-415, 2000.
- PRINCE, Victoria E.; PICKETT, F. Bryan. Splitting pairs: the diverging fates of duplicated genes. **Nature Reviews Genetics**, v. 3, n. 11, p. 827-837, 2002.
- RAGAN, Mark A. On surrogate methods for detecting lateral gene transfer. **FEMS Microbiology letters**, v. 201, n. 2, p. 187-191, 2001.
- RANZ, José María et al. Origin and evolution of a new gene expressed in the Drosophila sperm axoneme. In: **Origin and Evolution of New Gene Functions.** Springer Netherlands, 2003. p. 233-244.
- REMM, Maito; STORM, Christian EV; SONNHAMMER, Erik LL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. **Journal of molecular biology**, v. 314, n. 5, p. 1041-1052, 2001.
- RIBEIRO, HAL. Desenvolvimento de um serviço de análise de sequências utilizando um modelo baseado em atributos de resultados de PSI-BLAST. 2013. Tese (Doutorado em Bioinformática) – Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte. 2013
- RIVERA, Maria C.; LAKE, James A. The ring of life provides evidence for a genome fusion origin of eukaryotes. **Nature**, v. 431, n. 7005, p. 152-155, 2004.

- ROGERS, Rebekah L.; HARTL, Daniel L. Chimeric genes as a source of rapid evolution in *Drosophila melanogaster*. **Molecular biology and evolution**, v. 29, n. 2, p. 517-529, 2012.
- ROST, Burkhard; SANDER, Chris; SCHNEIDER, Reinhard. Redefining the goals of protein secondary structure prediction. **Journal of molecular biology**, v. 235, n. 1, p. 13-26, 1994.
- SCHULT, Daniel A.; SWART, P. Exploring network structure, dynamics, and function using NetworkX. In: **Proceedings of the 7th Python in Science Conferences (SciPy 2008)**. 2008. p. 11-16.
- SNEL, Berend; BORK, Peer; HUYNEN, Martijn. Genome evolution: gene fusion versus gene fission. **Trends in genetics**, v. 16, n. 1, p. 9-10, 2000.
- SNEL, Berend; HUYNEN, Martijn A.; DUTILH, Bas E. Genome trees and the nature of genome evolution. **Annu. Rev. Microbiol.**, v. 59, p. 191-209, 2005.
- SOREK, Rotem; AST, Gil; GRAUR, Dan. Alu-containing exons are alternatively spliced. **Genome research**, v. 12, n. 7, p. 1060-1067, 2002.
- SYAMALADEVI, Divya P.; JOSHI, Adwait; SOWDHAMINI, Ramanathan. An alignment-free domain architecture similarity search (ADASS) algorithm for inferring homology between multi-domain proteins. **Bioinformation**, v. 9, n. 10, p. 491, 2013.
- SYVANEN, Michael. Cross-species gene transfer; implications for a new theory of evolution. **Journal of theoretical Biology**, v. 112, n. 2, p. 333-343, 1985.
- TAUTZ, Diethard; DOMAZET-LOŠO, Tomislav. The evolutionary origin of orphan genes. **Nature Reviews Genetics**, v. 12, n. 10, p. 692-702, 2011.
- TEICHMANN, Sarah A.; PARK, Jong; CHOTHIA, Cyrus. Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. **Proceedings of the National Academy of Sciences**, v. 95, n. 25, p. 14658-14663, 1998.
- TERRAPON, Nicolas et al. Rapid similarity search of proteins using alignments of domain arrangements. **Bioinformatics**, v. 30, n. 2, p. 274-281, 2014.

- THOMSON, Timothy M. et al. Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. **Genome research**, v. 10, n. 11, p. 1743-1756, 2000.
- UNIPROT CONSORTIUM et al. UniProt: a hub for protein information. **Nucleic acids research**, p. gku989, 2014.
- VAN RHEEDE, Teun et al. Molecular evolution of the mammalian prion protein. **Molecular biology and evolution**, v. 20, n. 1, p. 111-121, 2003.
- VAN RHEEDE, Teun et al. Molecular evolution of the mammalian prion protein. **Molecular biology and evolution**, v. 20, n. 1, p. 111-121, 2003.
- WALSH, J. Bruce. How often do duplicated genes evolve new functions?. **Genetics**, v. 139, n. 1, p. 421-428, 1995.
- WANG, Minglei; CAETANO-ANOLLÉS, Gustavo. Global phylogeny determined by the combination of protein domains in proteomes. **Molecular biology and evolution**, v. 23, n. 12, p. 2444-2454, 2006.
- WANG, Wen et al. Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*. **Proceedings of the National Academy of Sciences**, v. 99, n. 7, p. 4448-4453, 2002.
- WESSLER, Susan R. Eukaryotic transposable elements: teaching old genomes new tricks. **The implicit genome**, p. 138-165, 2006.
- WICKER, Thomas et al. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, n. 12, p. 973-982, 2007.
- WOLF, Yuri I.; ROGOZIN, Igor B.; KOONIN, Eugene V. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. **Genome Research**, v. 14, n. 1, p. 29-36, 2004.
- WOLFE, Ken. Robustness? it's not where you think it is. **Nature genetics**, v. 25, n. 1, 2000
- WU, Dong-Dong; IRWIN, David M.; ZHANG, Ya-Ping. De novo origin of human protein-coding genes. **PLoS Genet**, v. 7, n. 11, p. e1002379, 2011.
- WUCHTY, Stefan. Scale-free behavior in protein domain networks. **Molecular biology and evolution**, v. 18, n. 9, p. 1694-1702, 2001.

- YANG, Song; DOOLITTLE, Russell F.; BOURNE, Philip E. Phylogeny determined by protein domain content. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 2, p. 373-378, 2005.
- ZEMLA, Adam et al. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. **Proteins: Structure, Function, and Bioinformatics**, v. 34, n. 2, p. 220-223, 1999.
- ZHANG, Jianzhi; ZHANG, Ya-ping; ROSENBERG, Helene F. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. **Nature genetics**, v. 30, n. 4, p. 411-415, 2002.
- ZHANG, Xue-Cheng et al. Evolutionary dynamics of protein domain architecture in plants. **BMC evolutionary biology**, v. 12, n. 1, p. 1, 2012.
- ZHANG, Yong E. et al. Accelerated recruitment of new brain development genes into the human genome. **PLoS Biol**, v. 9, n. 10, p. e1001179, 2011.
- ZHOU, Qi et al. On the origin of new genes in Drosophila. **Genome research**, v. 18, n. 9, p. 1446-1455, 2008.
- ZHOU, Qi; WANG, Wen. On the origin and evolution of new genes—a genomic and experimental perspective. **Journal of Genetics and Genomics**, v. 35, n. 11, p. 639-648, 2008.