UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

História evolutiva da subfamília FOXP: Análise Evolutiva Molecular e Estrutural em Tetrápodes

LUCAS HENRIQUES VISCARDI

Dissertação submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da UFRGS como requisito parcial para a obtenção do grau de Mestre em Genética e Biologia Molecular.

ORIENTADOR:PROF DR CLAITON BAU

CO-ORIENTADORA:PROFA DRA MARIA CÁTIRA BORTOLINI

PORTO ALEGRE, MARÇO DE 2015

Sumário

Agradecimentos

Às mulheres da minha vida: minha mãe e avó; no qual nunca teria dado um passo em frente para uma conquista sem sua confiança, apoio e afeto. Mas um carinho especial à minha mãe que sempre foi uma guerreira e exemplo a ser seguido. Bem como meu padrasto e amigo Leonardo que é um grande companheiro e junto à ela sempre tem um tempo para me escutar.

Ao meu avô que sempre me apoiou e ajudou de todas as maneiras durante a minha criação e formação profissional. Com muito carinho, o carrego sempre comigo, apesar dele viver sempre no Rio de Janeiro, muito bem acompanhado.

À professora Maria Cátira, por ter acreditado em mim como aluno desde a graduação e ter me incentivado a efetuar o mestrado, no qual pude sempre contar com seu apoio e orientação. Além de ser uma pessoa que serve de referência profissional e como líder de grupo.

Ao professor Claiton Bau, no qual sem ele este mestrado não seria possível. Agradeço de todo coração à oportunidade dada a um aluno até então desconhecido, tendo acreditado em mim e depositado esperança.

Às minhas fiéis amigas e co-orientadoras Luciana Tovo Rodrigues e Vanessa Paixão-Côrtes, que estão sempre comigo nos melhores e piores momentos, não havendo horário ou lugar para resolvermos qualquer problema, dividir qualquer angústia, ou comemorar um conquista.

Ao professor Salzano, pela sua homérica paciência auxiliar não apenas nesta dissertação, mas ao longo de muitos trabalhos já efetuado até então. És com a maior das certezas um grande referencial para mim e para o mundo.

À professora Sidia Callegari-Jacques, que me acolheu de forma excepcional ao longo do programa, sendo uma verdadeira amiga e confidente, além de uma excelente profissional.

Aos meus grandes amigos/colegas do Laboratório de Evolução Humana e Molecular por emprestar muito dos ouvidos ao ter que escutar minha voz ao longo desse tempo todo. Além de estarem ajudando a compor um ambiente de trabalho fenomenal para habitar. Mas um carinho especial para meu grande companheiro de metal e de copo, Dr. Bisso Machado, que com certeza teve e tem um papel muito importante na minha vida dentro e fora da academia.

Ao grupo dos meus irmãos de fé basqueteiros: Adriano, Irving, Flávio, Naej, Jusoan e Renan. Mas dando ênfase ao meu amigo anoréxico, e meu melhor amigo Ândrio da Silva Zafalon, uma pessoa que sempre esteve comigo ao longo de todo esse tempo, dividiu risadas, choros, noites em claras estudando ou fazendo cerveja. E nunca pensou duas vezes em me ajudar, não importa a ocasião que fosse.

Ao meu casal de colegas arqueólogos favoritos que trouxeram a este mundo uma menininha muito linda chamada Alice, que tenho orgulho de ser padrinho. Freddy e Milene, vocês sabem o apreço que tenho por vocês e carinho é gigantesco. Além desse dinamarquês ser um cara que ponho os dois braços mais o tronco no fogo. És um dos maiores exemplos de pessoa que tenho.

Ao meu grande amigo ninja Estevan, que me ajudou muito, seja nas técnicas marciais, comer muito sushi, fazer Le Parkour, ou fazer algo funcionar no terminal do Linux.

E na finaleira, aquela gurizada que faz uma sessão descarrego do trabalho nos finais de semana. Pessoal da BBC Fabrício, Lauro, Lucas e Roberto que é só juntar que dá errado. Porém, uma extensão desse grupo colocara um marco miliário desde 2013 na minha vida, que foi a fundação da Quarta da Gurizada. Meus grandes amigos, André, Giovanni, Lenon, Miguel, Thomás e Renan, que acompanharam integralmente a construção desta dissertação, sempre me dando muito apoio.

Resumo

A família gênica Forkhead P (*FOXP*) tem sido alvo de muitos estudos envolvendo evolução do cérebro e comportamento animal. Destacam-se particularmente as investigações com o gene *FOXP2,* que indicam que mudanças neste gene estariam associadas com a evolução da vocalização em algumas espécies de mamíferos, incluindo o *Homo sapiens*. Recentemente, estudos de desordem intrínseca de proteínas (IDPs) tem ganhado ênfase no contexto evolutivo, visto que uma correlação positiva entre regiões de desordem e altas taxas evolutivas tem sido observada. Através de um conjunto de abordagens que inclui predizer o conteúdo de desordem e os motivos lineares de interação, bem como as taxas evolutivas, buscamos desvendar a historia evolutiva dos genes da subfamília FOXP. Concentramos nossas análises sobre regiões desordenadas das proteínas FOXP1, FOXP2, FOXP3 e FOXP4 encontradas em 77 espécies de tetrápodes. Tais regiões proteicas são normalmente negligenciadas em estudos dessa natureza, pois se localizam fora de seus tradicionais domínios conservados, normalmente associados à função principal da proteína. Sítios apontados estando sob seleção positiva e relaxamento da restrição seletiva mostraram-se *hotspots* importantes para mudanças que podem impactar na capacidade de interação das proteínas. Encontramos que os maiores valores de ω são mais prevalentes em regiões desordenadas que em ordenadas. Ainda, alto e similar valor de desordem (70%) foi encontrado nas 77 proteínas ortólogas de FOXP1, FOXP2, e FOXP4, indicando a manutenção de um "padrão geral" sobre um longo tempo evolutivo. Portanto, a variabilidade tanto de aminoácidos quanto de motivos lineares dentro das regiões de desordem foi marcante. A proteína FOXP3 apresentou menor nível de desordem (30%), mas significante sinal de seleção positiva em alguns sítios. Composição idêntica de resíduo de aminoácido e/ou motivos lineares em espécies filogeneticamente distantes, indica clara convergência molecular, provavelmente associada a pressões seletivas similares. Sucessivamente, nossos achados mostraram uma clara diferença na composição de motivos lineares entre mamíferos e não mamíferos, dando suporte para a importância dos estudos de evolução da interatividade proteica para as compreensões de características taxa-específicas.

Palavras-chaves: Desordem intrínseca de Proteínas, Evolução Molecular, Motivos Lineares, domínio Forkhead, família gênica FOXP.

Abstract

Forkhead Family P (*FOXP*) has been target of many studies about brain and behavior evolution among species. FOXP2 receives special attention in academic society, due associations with vocalization evolution in mammals, including *Homo sapiens*. Recently, intrinsically disorder proteins studies have gained emphasis in the evolutionary context, as positive correlation between disorder regions and higher evolutionary rate has been observed. Through a set of approaches, including disorder and linear motif predictions, as well as estimate evolutionary rates, we aimed to unveil the evolutionary history of FOXP subfamily genes. We focused our analysis over disordered regions of FOXP1, FOXP2, FOXP3 and FOXP4 proteins retrieved in 77 tetrapods. Such protein regions are usually neglected in studies of this nature, for being localized out of the traditional conserved domains, usually associated with the main function of the protein. Sites indicated as under relaxation of selective constrains or positive selection have shown to be important hotspots for changes that can impact in protein interaction capability. Higher ω values are prevalent in disordered regions than in ordered ones. Still, high and similar disorder proportion (~70%) was found among 77 orthologues proteins of FOXP1, FOXP2 and FOXP4, indicating general pattern of disorder maintenance, along tetrapod's evolutionary tree. However, amino acid and linear motifs variability within disordered regions was observed. FOXP3 protein presented lower disorder levels (~30%), when compared with other paralogues, but signal of positive selection was observed in some sites. Identical composition of amino acid residues and/or linear motifs is, probably, associated with similar selective pressure. Successively, our results showed clear differences in linear motif composition between mammals and non-mammals, supporting the importance of evolutionary studies on protein interaction for the understanding of taxa-specifics characteristics.

Key-words: Intrinsically disordered protein, molecular evolution, linear motifs, Forkhead domain, FOXP gene family

Capítulo 1

Introdução

1.1 Considerações Gerais

Desde o início do século XXI, muitos estudos têm focado em abordagens que buscam identificar variações genéticas intraespecíficas, capazes de explicar a variabilidade de fenótipos que caracterizam as mais diversas espécies. Outros são delineados para identificar genótipos taxa-específicos (*i.e,* espécie-específicos, gênero-específicos, *etc.*) que possam explicar a presença de um fenótipo que caracteriza toda uma categoria taxonômica. Neste contexto, destacam-se abordagens que compraram genomas humanos com os de outros primatas, bem como outros animais, o qual permite investigar diferenças e semelhanças moleculares entre estes, na busca de variantes funcionalmente relevantes que possam explicar fenótipos que emergiram a partir de uma história evolutiva parte compartilhada, parte única das espécies investigadas. Dentro deste contexto duas categorias de genes homólogos aparecem em destaque: (a) parálogos, ou seja, grupo de genes que surgem por duplicação a partir de um gene ancestral comum, e (b) ortólogos, grupo de genes que surgem a partir de eventos de especiação (Fitch, 1970). Sendo assim, o estudo das chamadas famílias gênicas revelam tanto a história dos genes quanto das espécies ao longo de uma determinada trajetória evolutiva. Além disso, tais estudos também permitem traçar os mapas da relação entre genótipo e fenótipo, conectados com padrões taxa-específicos. Resgatar esses complexos eventos e interações, no entanto, não é uma tarefa fácil. Por conta disso, mesmo com os avanços na geração de dados genômicos, poucos se sabe sobre a história evolutiva da maioria das famílias gênicas encontradas em espécies animais.

Neste trabalho, através do uso de diversas metodologias de análises de evolução molecular e biologia estrutural funcional, buscamos delinear o cenário evolutivo dos quatro genes *(FOXP1, FOXP2, FOXP3 e FOXP4)* da família *FOXP* ao longo de 77 espécies de tetrápodes. Uma atenção especial foi dada para análises das regiões desordenadas nos fatores de transcrição codificadas por estes genes. Regiões desordenadas caracterizam-se por segmentos protéicos que não apresentarem estrutura secundária e/ou terciária estável, o que lhes permite interações com uma grande gama de proteínas e moléculas devido à sua flexibilidade. Por sua vez, um grande conteúdo de regiões desordenadas caracteriza uma proteína desordenada. Particularmente importante é o conteúdo de desordem em fatores de transcrição de eucariotos visto a enorme complexidade dos sistemas regulatórios nestes organismos. Deste modo, buscamos identificar a proporção de regiões de desordem nas proteínas FOXP1, FOXP2, FOXP3 e FOXP4 e se estas estariam sob algum padrão de pressão seletivo-específico (*e.g.* relaxamento, seleção negativa e ou positiva). Além disso, predizemos também os chamados motivos Lineares de Interação, comuns em proteínas

desordenadas, e o quanto as mudanças de aminoácidos e nestes motivos impactaram a evolução das espécies de tetrápodes investigadas.

1.2 Introdução à Biologia Estrutural

Proteínas são compostas por blocos de 20 aminoácidos caracterizados pela presença de um carbono central ligado a um hidrogênio, um grupo amino primário, um grupo ácido carboxílico e uma cadeia lateral variável, denominada grupo R. Considerando o grupo R que o aminoácido apresenta, é possível identificar suas características particulares e, assim, classifica-lo por seu tamanho, forma ou composição química. No presente trabalhado foi dada atenção principalmente quanto ao último item, composto pelas categorias: neutra-apolar (hidrofóbicas), neutra-polar, ácida e básica (Watson, 2005).

As proteínas são basicamente caracterizadas estruturalmente da seguinte maneira: estrutura primária, secundária e terciária. Enquanto a estrutura primária é a própria sequência de aminoácidos disposta de forma linear, as estruturas secundárias são ordenadas e estabilizadas por ligações de hidrogênio entre grupos da cadeia principal e aqueles não participantes de ligações de hidrogênio. Alfa-hélice e folhas-beta são as formas mais comuns de estruturas secundárias (Watson, 2005). Estruturas terciárias, por sua vez, são fundamentais para a funcionalidade das proteínas, visto que esta determina sua estrutura conformacional tridimensional. Uma das maiores influências sobre o enovelamento proteico, e assim para a formação de estruturas terciárias, pode ser atribuída ao direcionamento interno dos grupos laterais dos aminoácidos hidrofóbicos (apolares). Dessa maneira, quando em solução aquosa, proteínas com um número muito grande de grupos laterais apolares tendem a formar estruturas mais estáveis do que as proteínas formadas majoritariamente por grupos polares (Watson, 2005).

Até recentemente, acreditava-se que apenas com a informação provinda da sequência de aminoácidos primária de uma proteína seria possível prever sua estrutura secundária e terciária (Watson, 2005). Todavia novos estudos vêm demonstrando que muitas proteínas não apresentam estruturas secundárias e terciárias definidas, devido à presença de regiões desordenadas em sua composição. Assim, métodos tradicionais para determinação da estrutura tridimensional das proteínas, tal como a cristalografia acompanhada de raios-X de alta energia para os estudos de difração tornam-se pouco efetivos quando aplicadas às determinadas regiões proteicas.

Sabe-se hoje que muitas proteínas apresentam estruturas secundárias e/ou terciárias instáveis (regiões em desordem), normalmente localizadas fora dos domínios de ligação ao DNA. Pode-se definir, sinteticamente, as chamadas regiões desordenadas (**IDRs**, *intrinsically disordered regions*) como segmentos proteicos enriquecidos por resíduos polares e carregados (Gln, Ser, Por, Glu, Lys), bem como resíduos promotores de desordem (Gly e Ala) (Liu *et al.,* 2006; van der Lee *et al.,*, 2014).

As chamadas proteínas intrinsecamente desordenadas (**IDPs**) apresentam grande conteúdo de regiões em desordem. Desse modo, caracterizam-se por terem suas estruturas tridimensionais instáveis, o que as torna altamente flexíveis. Esse fato permite que as **IDPs** interajam com uma grande variedade de proteínas e moléculas. As **IDPs** apresentam ainda sequências enriquecidas por sítios curtos de ligação e de degradação, conhecidos genericamente como **Motivos Lineares (LMs**, *linear motifs***),** alvos preferenciais de modificações pós-traducionais (Ba, 2014). Estudos mostram que, estando envolvida na resposta de sinalizações rápidas e eficientes, a expressão de **IDPs** dentro das células é altamente controlada para evitar interações indesejáveis (Liu, 2006).

As **IDPs** são enriquecidas por **LMs**, particularmente pelos chamados motivos lineares curtos, constituídos de poucos peptídicos (2 a 15 aminoácidos; *Short Linear Motif,* **SLIM**). Estes motivos são preferencialmente encontrados em regiões desordenadas, estando desse modo, conectados com as características flexíveis das **IDPs** (Gould, 2010; Dinkel, 2013; Disfani, 2012).

Estudos mostram que em eucariotos, 45–50% das proteínas apresentam regiões de desordem composta por 30 ou mais resíduos de aminoácidos. Em bactérias e Arqueas, a proporção cai para 7–30%, demonstrando que proteínas desordenadas são mais frequentes em eucariotos do que procariotos (Dunker, 2000; Oldfield, 2005; Xue, 2010; 2013). Se forem consideradas regiões longas com mais de 50 resíduos, os números seguem igualmente diferenciados, sendo ~20% e ~8% dos proteomas de eucariotos e procariotos desordenados, respectivamente (Dunker, 2000; Marín, Uversky, e Ott, 2013). Essa proporção também é observada considerando-se fatores de transcrição, fato que pode ser devido ao sofisticado e sistema regulatório encontrado em eucariotos, caracterizado por uma grande complexidade quando comparado aos procariotos (Liu *et al*, 2006). Em resumo, a mais alta quantidade de **IDPs** em eucariotos seria o reflexo da complexidade de sinalização e processos regulatórios nesses organismos (Dunker, 2000; Ward, 2004; Tompa *et al.*, 2006). No entanto, a exata dimensão do fenômeno, bem como suas implicações funcionais na maioria das famílias de genes eucarióticos é completamente desconhecida.

1.3 Superfamília *Forkhead*

O nome da família gênica de fatores de transcrição "*forkhead domain*" (FOX) deriva de uma descoberta em 1990 com embriões *knockout* de *Drosophila melanogaster*. O experimento levou ao surgimento de um fenótipo de suas cabeças em formato de forca, dando assim o nome em inglês de *Drosophila melanogaster forkhead* (Coffer e Burgering, 2004). A família *FOX*, já foi identificada nos mais diversos organismos, o que indica sua origem, provavelmente, no ancestral comum entre animais e fungos (Fritzenwanker *et al.,* 2014; Thackray, 2014). A superfamília *FOX* apresenta um total de 24 subfamílias (duas delas específicas de vertebrados), totalizando 43 genes em humanos. Estes genes são caracterizados, principalmente, pela presença do domínio conservado de ligação ao DNA chamado de *Forkhead* ou W*inged-helix,* com aproximadamente 100 aminoácidos (Lam, 2013). Apesar de diversas funções já terem sido associadas aos genes da superfamília *FOX*, alguns papeis principais estão ligados ao desenvolvimento do sistema imune, sistema nervoso central, bem como regulação do eixo hipotalâmico-pituitário-gonadal (Coffer e Burgering, 2004; Lam, 2013; Thackray, 2014).

Devido as importantes funções dos genes *FOX*, tem sido proposto que o aumento da complexidade dos organismos estaria sendo relacionado, pelo menos em parte, pela expansão desta família de fatores de transcrição. Apesar de todas as proteínas FOX apresentarem capacidades semelhantes de ligação ao DNA, a sua funcionalidade pode ser tanto de ativadora como de inibidora da transcrição gênica.

1.4 Subfamília *FOXP*

A subfamília *FOXP* provavelmente é a mais estudada e conhecida dentre todas as que compõem a superfamília *Forkhead*. Por sua característica de formar homo- e heterodímeros, graças ao seu zíper de Leucina, seus quatro genes membros (*FOXP1-FOXP4*) atuam num amplo espectro de funções, incluindo desenvolvimento cerebral, do pulmão, diferenciação de células T, e, como veremos posteriormente, em mecanismos que promovem o desenvolvimento da vocalização, fala e linguagem (Li *et al.,* 2004; Shu, 2007; Brown e Banham, 2009; Enard, 2011; Vernes, 2011).

A expansão da subfamília *FOXP* provavelmente se deu na origem dos vertebrados, visto que seus quatro genes são encontrados em todas as espécies de vertebrados já estudados, enquanto apenas um *FOXP* é encontrado em invertebrados (Santos *et al.,* 2011). Os quatro genes atuam através da formação de dímeros entre eles e outras proteínas, atuando majoritariamente como repressores. Ainda, a ação repressora de *FOXP1, FOXP2*, e *FOXP4* parecem ser significativamente potencializadas quando os mesmos são co-expressos (Mendoza, 2014). Todavia, nenhum gene da subfamília *FOXP* destaca-se tanto na literatura científica (e leiga) quanto o *FOXP2*, visto que este tem sido associado à aprendizagem, fala e linguagem humana.

O gene *FOXP2* ficou mundialmente conhecido com o estudo de caso da família britânica KE, onde os familiares heterozigotos afetados pela mutação Arg553His, no domínio *Forkhead,* apresentavam anormalidades funcionais e estruturais dos circuitos cortico-estriatal e cerebelar, importantes para o aprendizado e execução de movimentos orofaciais articulados. Em outras palavras, os indivíduos afetados da família não conseguiam articular sua fala normalmente (Enard, 2002; Enard, 2011; Teramitsu e White, 2008). Posteriormente, novos casos foram sendo estudados, associados com dificuldades semelhantes (Schön, 2006). A partir das informações dos estudos médicos, Enard *et al.* (2002) publicaram um estudo clássico onde mostram que mutações não-sinônimas de dois aminoácidos (Thr303Asp; Asp325Ser) na proteína FOXP2 estavam fixadas em humanos, mas não apareciam em chimpanzés ou outro grande macaco. A partir destes estudos, emergiu na imprensa leiga, incontáveis matérias sobre o "gene humano da fala", nome pelo qual passou a ser conhecido o gene *FOXP2*.

Outras investigações com *FOXP2* tem providenciado resultados instigantes. Por exemplo, o padrão de expressão do gene é conservado no cérebro durante o desenvolvimento embrionário de mamíferos, pássaros, anfíbios e peixes, sendo associado com funções similares no sistema nervoso central destes grupos de organismos (Li, *et al.* 2007; Teramitsu e White, 2008; Fisher e Scharff, 2009;). Ainda, outros estudos apontaram sinais de seleção positiva no *FOXP2* de humanos, cetáceos, quirópteros e pássaros cantantes (Li *et al.,* 2007; Enard, 2011). Ainda, o *FOXP2* já foi associado com controle da vocalização emocional em filhotes de camundongos, bem como aprendizado através da vocalização em pássaros cantantes e morcegos com ecolocalização (Li *et al.,* 2007; Song *et al.,* 2013).

Atualmente já é conhecido que a proteína FOXP2 encontrada no *Homo sapiens*, e em humanos arcaicos extintos *Homo neanderthalensis* e no espécime de Denisova são idênticas, sendo estimada a origem das mutaçãos (Thr303Asp; Asp325Ser) por volta de 270-400 mil de anos (Krause *et al.,* 2007; Teramitsu e White, 2008; Reich *et al*., 2010; Enard, 2011; Paixão-Cortes *et al*., 2014).

Interessantemente outro trabalho mostrou que, apesar da diferença estar em somente dois aminoácidos, a proteína FOXP2 humana e a do chimpanzé regula de maneira diferenciada 42 genes alvos (Spiteri *et al.,* 2007). A natureza dessas diferenças ainda não é claramente compreendida.

Em humanos, bem como em pássaros cantantes, o FOXP2 é relacionado ao desenvolvimento do circuito cerebral gânglio cortico-basal, ou CBG (Enard, 2011). Supõe-se que CBG esteja envolvido com o aprendizado através de ações de tentativa e erro, estando deste modo relacionado com a via de sinalização da dopamina (Graybiel, 2005). Camundongos transgênicos, homo e heterozigotos para os aminoácidos 303 Asp e 325 Ser encontrados na proteína FOXP2 de espécies extintas e modernas do gênero *Homo*, mostraram alterações nos circuitos CBG, bem como nos níveis de dopamina, plasticidade sináptica, dentre outras alterações (Teramitsu e White, 2008; Enard *et al*., 2009 e 2011; Schreiweis, *et al.* 2014). Tais resultados sustentam a hipótese que o FOXP2-303Asp-325Ser acelere o aprendizado declarativo e processual, relacionando-se, assim, com nossa (e talvez de nossos parentes extintos) capacidade de desenvolver linguagem falada (Schreiweis *et al.,* 2014).

Em resumo, estudos com *FOXP2* têm levado a um maior entendimento sobre o desenvolvimento cerebral, formação de sinapses, adesão celular, aprendizagem, bem como a evolução da fala e linguagem em humanos e processos de vocalização em outras espécies, dentre outros impressionantes achados (Li *et al.,* 2007; Enard, 2011; Song *et al.,* 2013; Mendoza, 2014).

Por outro lado, o conhecimento sobre os demais genes da subfamília não é tão extenso. Sabe-se que, assim como FOXP2, os genes FOXP1 e FOXP4 também apresentam um importante papel no desenvolvimento do cérebro, pulmão e coração dos vertebrados. Por outro lado, FOXP3 está associado com o desenvolvimento e homeostase do sistema imunológico, sendo conhecido como regulador mestre do desenvolvimento e função das células T (TCR) (Benayoun *et al.,* 2011; Andersen *et al.*, 2012).

Outra característica importante da subfamília FOXP é a sua capacidade única entre membros da superfamília FOX, de formar hetero e homodimerização entre os membros FOXPs, que são essenciais para suas capacidades de ligação ao DNA. Essas conexões se dão através do motivo de ligação denominado zíper de Leucina (Li *et al.* 2004; Brown e Banham, 2009). Embora o papel de dimerização não seja bem compreendido (Shu *et al*., 2007), alguns exemplos podem ser citados. Heterodimerização de FOXP1 e FOXP3 é essencial para o desenvolvimento e função das células T, desde que mutações no zíper de Leucina de FOXP3, observadas em pacientes com síndromes autoimunes XAAD / IPEX, impede a interação FOXP1-FOXP3 (Brown e Banham, 2009; Andersen *et al.*,

2012). Além disso, estudos recentes mostraram que, quando FOXP2 é expressa concomitante com FOXP1 ou FOXP4 a repressão dos genes alvos aumenta significativamente (Mendoza *et al.,* 2014). Em resumo, os membros FOXP podem ser definidos como repressor transcricional que necessitam de homo- ou hetero-dimerização para sua funcionalidade. Assim, a compreensão sobre a natureza das estruturas tridimensionais dos FOXPs é extremamente relevante para compreender suas capacidades de interação e amplo espectro de atuações.

Capítulo 2

Objetivos

2.1 Objetivo Geral

Tendo em vista o exposto acima, o presente trabalho buscou delinear o cenário evolutivo dos quatro genes parálogos que constituem a subfamília FOXP, através da investigação de suas respectivas regiões de desordem intrínseca, e da predição de seus Motivos Lineares de ligação a outras proteínas e moléculas. Adicionalmente, buscou-se estimar as taxas evolutivas tanto nas regiões ordenadas quanto de desordem dos 77 genes ortólogos encontrados em espécies de tetrápodes, visando conectar as modificações encontradas com possíveis novidades adaptativas.

2.2 Objetivos específicos

a) Estimar o conteúdo de desordem nas proteínas FOXP1, FOXP2, FOXP3 e FOXP4;

b) Identificar motivos lineares (**LMs**) nessas mesmas proteínas;

c) Predizer as regiões de transição desordem-ordem dentro das proteínas;

d) Identificar as taxas evolutivas dos genes *FOXP1, FOXP2, FOXP3 e FOXP4* e definir o padrão de evolução molecular dos mesmos;

e) Associar as alterações encontradas com características fenotípicas taxa-específicas considerando 77 tetrápodes.

Capítulo 3

FOXP in Tetrapoda: Intrinsically Disordered Regions, Short Linear Motifs and their evolutionary significance

FOXP in Tetrapoda: Intrinsically Disordered Regions, Short Linear Motifs and their evolutionary significance

Accepted Manuscript

Lucas Henriques Viscardi[1], Luciana Tovo-Rodrigues[2], Pamela Paré[1], Nelson Jurandi Rosa Fagundes[1], Francisco Mauro Salzano[1], Vanessa Rodrigues Paixão-Côrtes[3], Claiton Henrique Dotto Bau[1], and Maria Cátira Bortolini[1]

[1]*Programa de Pós-Graduação em Genética e Biologia Molecular, Departamento de Genética, Universidade Federal do Rio Grande do Sul, Caixa Postal 15053, 91501-970 Porto Alegre, RS, Brazil.*

[2]*Programa de Pós-Graduação em Epidemiologia, Universidade Federal de Pelotas, 96020-220 Pelotas, RS, Brazil.*

[3]*Programa de Pós-Graduação em Genética e Biodiversidade, Instituto de Biologia, Universidade Federal da Bahia, Caixa Postal 40170-290, Salvador, BA, Brazil.*

Send correspondence to Prof. Maria Cátira Bortolini.

Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Campus do Vale, Caixa Postal 15053, 91501-970 Porto Alegre, RS, Brazil. E-mail: maria.bortolini@ufrgs.br

Abstract

The *FOXP* subfamily is probably the most extensively characterized subfamily of the forkhead superfamily, playing important roles in development and homeostasis in vertebrates. Intrinsically disorder protein regions (IDRs) are protein segments that exhibit multiple physical interactions and play critical roles in various biological processes, including regulation and signaling. IDRs in proteins may play an important role in the evolvability of genetic systems. In this study, we analyzed 77 orthologous *FOXP* genes/proteins from *Tetrapoda*, regarding protein disorder content and evolutionary rate. We also predicted the number and type of the short linear motifs (SLIMs) in the IDRs. Similar levels of protein disorder (approximately 70%) were found for FOXP1, FOXP2, and FOXP4. However, for FOXP3, which is shorter in length and has a more specific function, the disordered content was lower (30%). Mammals showed higher protein disorders for FOXP1 and FOXP4 than non-mammals. Specific analyses related to linear motifs in the four genes showed also a clear differentiation between FOXPs in mammals and non-mammals. We predicted for the first time the role of IDRs and SLIMs in the FOXP gene family associated with possible adaptive novelties within *Tetrapoda*. For instance, we found gain and loss of important phosphorylation sites in the *Homo* FOXP2 IDR regions, with possible implication for the evolution of human speech.

Keywords: FOXP2, Intrinsically disordered protein regions, forkhead superfamily, short linear motif, molecular evolution.

Abbreviations

FOX, forkhead box; IDR, intrinsically disordered region; IDP, intrinsically disordered protein; SLIM, short linear motif

Introduction

Members of the *Forkhead box* (*FOX*) gene superfamily have been widely associated with organismal development and are identified by their evolutionary conserved forkhead DNA-binding domain (Lam *et al.*, 2013; Morris and Fanucchi, 2016). The *FOXP* subfamily is probably the most extensively characterized subfamily of the forkhead superfamily. The four *FOXP* genes (*FOXP1, FOXP2 FOXP3,* and *FOXP4*) emerged by duplication events during the origin of vertebrates (Santos *et al.*, 2011; Song *et al.*, 2016). Since the duplication events, paralogues *FOXP1*, *FOXP2,* and *FOXP4* have played an important role in brain, lung, heart, and jaw development in vertebrates, while *FOXP3* has been associated with the development and homeostasis of the immune system, since it is described as a master-regulator of $CD4^+$ and $CD25^+$ T-cells (Coffer and Burgering, 2004; Akbar *et al.*, 2007; Takahashi *et al.*, 2008; Benayoun *et al.*, 2011; Andersen *et al.*, 2012; Lam *et al.*, 2013; Cesario *et al.,* 2016).

Undoubtedly, the most widely known member of the *FOXP* subfamily is *FOXP2*, as it has attracted the attention of the scientific community and the general media because of its role in the evolution of speech and vocalization in mammals (Zhang *et al.,* 2002; Li *et al.,* 2007), specially as mutations in this gene promote severe impairment of articulation and grammar in humans (Enard *et al.*, 2002, 2011; Schön *et al.*, 2006; Bowers and Konopka, 2012). *FOXP2* is expressed mostly in the brain, where it plays an important role in synapse formation and cell adhesion, as well as in the specification and differentiation of the lung epithelium and gastrointestinal and cardiovascular tissues (Song *et al.*, 2013).

Evolutionary studies have been successively improved by incorporating new methodological approaches. Analysis of intrinsically disordered regions (IDRs), which is routinely used in medical and structural biology studies, can also be applied in evolutionary studies because of the possible role of IDRs in the evolvability (evolutionary capacity; Pigliucci, 2008) of genetic systems (Neduva and Russell, 2005). IDRs are protein segments rich in hydrophilic, polar, and charged amino acids (glutamine, serine, glutamic acid, arginine, and lysine), as well as glycine, proline, and alanine (Iakoucheva *et al.*, 2004; Liu *et al.*, 2006). IDRs are prevalent in proteins that exhibit multiple physical interactions and play critical roles in various biological processes, including regulation and signaling (Dunker *et al.*, 2000; Nguyen Ba *et al.*, 2012; Forman-Kay and Mittag, 2013). The conformational flexibility of IDRs facilitates exposure of specific residues for modification and binding to other proteins and molecules (Huang and Sarai, 2012; Liu and Huang, 2014). Thus, intrinsically disordered proteins (IDPs) are characterized by high IDR contents and the absence of stable well-folded three-dimensional structures in solutions (Forman-Kay and Mittag, 2013).

Short linear motifs (SLIMs) are short stretches in protein sequences that mediate protein-protein interactions. SLIMs are typically 2–10 amino acids long; however, only 2 or 3 amino acids are essential for interaction with other molecules. SLIMs are common elements in IDRs, and they probably play a significant role in the functioning of these disordered regions (Wagner and Lynch, 2008; Huang and Sarai, 2012; Nguyen Ba *et al.*, 2012; Forman-Kay and Mittag, 2013; Liu and Huang, 2014). The presence of a great number of these motifs in such regions probably confers functional flexibility to this class of proteins (Gould *et al.*, 2010; Disfani *et al.*, 2012; Dinkel *et al.,* 2012, 2014). Furthermore, SLIMs are particularly evolvable because they are poorly conserved between lineages and can appear and disappear through small changes (Wagner and Lynch, 2008). Therefore, changes in SLIMs significantly impact complex regulatory networks (Neduva and Russell, 2005). Thus, analysis of these changes enables the assessment of their importance in the evolutionary trajectory of animals.

In addition to the forkhead, leucine-zipper, and zinc-finger domains, other molecular elements such as IDRs may play crucial roles in the function of FOXP proteins. However, these structures have not been studied extensively. Thus, the present investigation aims to ask how FOXPs structural forms changed through *Tetrapoda* evolution regarding linear motifs composition and disordered content. Furthermore, as *FOXP3* is known to be the only gene among the *FOXP* family playing a role in the immune system, we investigated if a higher evolutionary rate would be observed when compared with other *FOXP*s, and if such rate could be related with higher disordered content.

Material and Methods

Seventy-seven orthologous *FOXP* genes/proteins from tetrapods (Table S1) were considered in the present study. *FOXP* nucleotide sequences were retrieved from the NCBI database using BLASTN with 20,000 Max target sequences. We also used the Ensembl genome database (http://ensembl.org/) for sequence retrieval. The Neanderthal exome (Castellano *et al.*, 2014; http://cdna.eva.mpg.de/) was consulted to verify possible specific changes within the genus *Homo*. However, one protein-coding gene may codify more than one isoform. The presence of many isoforms in the *FOX* genes, caused by alternative splicing, was handled conservatively by choosing only isoforms that clearly resemble the canonical form identified in humans by using UniProt (http://www.uniprot.org/). Incomplete sequences were removed from the analysis. Subsequently, the sequences were aligned using the MAFFT algorithms (standard pattern) implemented in the Guidance web server (http://guidance.tau.ac.il/). The alignments are available in the Supplementary Material. Phylogenetic trees were drawn using FigTree1.4.2. (http://tree.bio.ed.ac.uk/software/figtree/) according to the literature (Meredith *et al.*, 2011; Perelman *et al.*, 2011; Song *et al.*, 2012).

Importantly, while both *FOXP2* and *FOXP4* passed through a standard NsSites test site analysis, for *FOXP3* and *FOXP1* we had to employ distinct data tests. Because of the absence of several base pairs in *Xenopus laevis FOXP1,* we excluded this species. For *FOXP3*, just the mammalian sequences were used because reptilian and amphibian *FOXP3* are shorter and very different, while in birds, *FOXP3* is completely absent (Andersen *et al.*, 2012). In addition, we removed from the analysis a residual N-terminal part of *FOXP3* present only in the mammals *Nomascus leucogenys Papio anubis*, *Chlorocebus sabaeus, Callithrix jacchus*, *Cricetulus griseus*, *Panthera tigris*, *Myotis brandtii*, *Pteropus alecto*, *Chrysochloris asiatica*, and *Dasypus novemcinctus*, as they do not align or resemble other orthologous and known isoforms.

We predicted disordered regions by using the PONDR-FIT metapredictor (Xue *et al.*, 2010). Additionally, the MobiDB server (Potenza *et al.*, 2014) was consulted to check consensus predictions for their disorder content, as provided by a variety of disorder predictors. SLIMs were predicted using the ELM webserver (Dinkel *et al.*, 2012, 2014) considering only the cell nucleus as the cell compartment for biochemical interaction context of FOXP proteins. Given that the linear motifs predicted by ELM can present a high rate of false positives, we considered only ELM in IDR regions and validated such predictions by analyzing the literature on the interactions between linear motifs and their ligands with other transcription factors. Therefore, we considered only linear motifs with confirmed experimental data and/or certainty for ELM reliability annotation. All information regarding the linear motifs was retrieved from the ELM server and from the literature. The ELM server classifies SLIMs into the following four types: protease cleavage sites, protein motif interaction/binding sites, posttranslational modification sites, and subcellular targeting signals (Dinkel *et al.,* 2012). Linear motifs present in the forkhead, leucine-zipper, and zinc-finger domains were not considered because they can represent false positives. Statistical tests comparing sites under purifying selection and/or positive selection within and without disordered regions were performed using WinPepi and SPSS 2.0.

To estimate the molecular evolutionary patterns of *FOXP1*, *FOXP2*, *FOXP3*, and *FOXP4*, we applied phylogeny-based maximum likelihood analysis of ω (non-synonymous/synonymous rate ratio or dN/dS) implemented in the PAML 4.7 package (Yang, 2007). This approach allows the ω ratio to vary among sites while considering several different codon substitution models. A value of $\omega < 1$ indicates potential negative selection, while $\omega = 1$ indicates neutrality, and $\omega > 1$ indicates positive selection. For the NsSites codon substitution model, likelihood ratio tests (LRT) were performed between neutral models (M1a, nearly neutral, M8a, Beta and $\omega = 1$) and models that allow positive selection and/or relaxation of functional constraints (M2a, positive selection and M8, Beta + Selection). Using log values from models M1a, M2a, M8a, and M8, we applied a LRT using HyPhy 2.2.0.

The Branch Site Model was also used to detect if different linear motif composition and disorder scores are reflected in different evolutionary rates among Tetrapoda. The phylogeny was a priori divided into 2 clades, and a LRT was used to evaluate divergences in selective pressures between them, as indicated by different ω ratios. We employed the clade model type D that assumes 2 site classes, which was compared with the neutral model M1a by an LRT with 2 degrees of freedom.

A Bayes empirical Bayes (BEB) approach was considered using CODEML in PAML 4.7 to verify which sites could be under neutral, purifying, or positive selection. The phylogenetic trees used to construct the PAML 4.7 input files were revised as described previously (Meredith et al., 2011; Perelman et al., 2011; Song et al., 2012).

Results and Discussion

FOXP1, FOXP2, FOXP3 and FOXP4 structures and their intrinsic protein disorder content

Our analyses revealed that the three paralogous proteins with similar functions and tissue expression, FOXP1, FOXP2, and FOXP4, had high and similar disorder contents (~70%). In contrast, FOXP3, which plays a role in immune system regulation, presented a lower disorder degree (~30%) relative to its paralogues (Tables 1, S2-S5), according to PONDR-FIT. The patterns of the disordered and ordered regions, as well as the disorder proportion of orthologous proteins, are relatively conserved among taxonomic groups (Tables 1, S2-S5). However, mammals presented a higher degree of protein disorder than all other organisms for FOXP1 and FOXP4 ($P < 0.001$, Table 1). Particularly, amphibians presented a lower degree of disorder for FOXP2 (~64%, Tables 1and S3.1) than the other classes ($P < 0.01$, Table 1). These *FOXP* disorder prediction values are, in general, higher than those obtained by other authors (Andersen *et al.*, 2012), but they used just partial proteins and fewer number of species. Importantly, we must highlight that mammalian larger sample compared to non-mammals may have contributed to these statistical differences in the protein disorder content analysis.

Interestingly, our data reveals that mammals present significantly higher FOXP1 and FOXP4 disorder degrees than the other groups. This finding may be associated with the more complex interaction networks present in mammals, as already proposed for other genetic systems (Disfani *et al.*, 2012), and to a positive correlation between the number of binding partners and disorder scores (Dunker *et al.*, 2000). Thus, it is reasonable to speculate that mammalian FOXP1 and FOXP4 present a larger number of binding partners than the other orthologues investigated here.

FOXP1, FOXP2, FOXP3, and FOXP4 and their interaction sites

Usually, intrinsically disordered proteins are enriched with SLIMs, which play crucial roles in their interaction with other proteins (Tables S6.1-6.4). Here we will briefly describe some selected representative results of the SLIMs compositional analysis. For FOXP1, some of our findings include a Polo-like kinase 1 (PLK) phosphorylation site at position 33 (MOD_PLK), which differentiates Sauropsida (reptiles and birds) from mammals (Tables 2). PLK is involved in events during the cellular cycle (Nakajima *et al.*, 2003; Murakami *et al.,* 2010), suggesting some differences in the FOXP1 phosphorylation pattern during cellular cycle between mammals and Sauropsida.

In the case of FOXP2 (Tables 2), at position 314, mammals have lost one DOC_USP7_1, which interacts with the deubiquitinating enzyme USP7/HAUSP (herpes virus-associated ubiquitin-specific protease) present in all non-mammals, due to a serine to alanine change. Previous studies have demonstrated that the interaction of USP7 with FOX members regulates oxidative stress responses through ubiquitination (van der Horst *et al.*, 2006). Thus, the possible loss of DOC_USP7_1 in mammals could have a functional implication related to response to oxidative stress.

Two known non-synonymous substitutions between humans (*Homo sapiens* and Neanderthals) and chimpanzees (FOXP2 Asn325Ser and Thr303Asn) deserve additional attention, since they were related to human speech (Enard *et al.*, 2002, Krause *et al.*, 2007). One of them (Asn325Ser) promotes the gain of two motifs, MOD_CK1_1 and MOD_GSK3_1, in humans due to the presence of a serine at aligned position 390 (Table 2). Both motifs are promoters of phosphorylation by kinases. Interestingly, carnivores also have a serine at this FOXP2 orthologue position (Zhang *et al.*, 2002), leading to a convergence event of the emergence of both MOD_CK1_1 and MOD_GSK3_1 motifs observed in humans. Cooper (2006) suggested that phosphorylation by kinase C in this FOXP2 region may be related to human behavioral traits such as language. However, the other *Homo* specific substitution at aligned position 368 (Thr303Asn) led to the lost of a phosphorylation site. Changes in phosphorylation patterns can modulate the regulation of transcription factors and their binding affinity to co-activators and DNA. These changes can in turn alter gene expression, cell growth, and differentiation (Iakoucheva *et al.*, 2004). Thus, our results have one very relevant implication: The lost of this phosphorylation site at position 368/303 can have been as important as the gain of the phosphorylation site at position 390/325 for the evolution of human speech. The phenotype implication of the presence of these SLIMS in carnivores is unknown.

For FOXP3, only investigated in mammals (see Material and Methods section), at position 194 (Table 3), a CK1 phosphorylation site (MOD_CK1_1) is predicted for several mammal species, except New World (NW) monkeys (*Saimiri boliviensis* and *Callithrix jacchus*) and *Tarsius syrichta*. Interestingly, these primates present four other linear motifs in this region: MOD_GSK3_1, MOD_ProDkin_1, DEG_SCF_FBW7_1, and DOC_WW_Pin1_4. Therefore, we identified the presence of the same SLIMs in two distinct branches of primates (New World monkeys and Tarsiidae) that live in somewhat similar rainforest environments. As mentioned before, FOXP3 is the only FOXP member playing a role in the immune system, suggesting that at least one of these motifs is associated with the immune response, indicating adaptation through convergence or the maintenance of a primate ancestral state.

Another interesting finding is the sharing of the linear motif LIG_PTAP_UEV_1 between Neanderthals and modern humans due to the Gly175Ser (human position) mutation (Table 4). It has also been suggested that linear motifs mediate interactions between viruses with their hosts (Hagai *et al.,* 2014). In fact, LIG_PTAP_UEV_1 mediates the binding of several cellular and viral proteins to the UEV domain of the class E vacuolar sorting protein Tsg101 (Göttlinger *et al.*,1991) and it is essential for the efficient egress of viral particles from many enveloped RNA viruses (Bieniasz, 2006). Our results indicate that this motif may have played an important role in *Homo* self-immune defense during the Pleistocene.

Regarding *FOXP4*, a striking difference between mammals and Sauropsida was also found (birds and reptiles, Table 2). For instance, the loss of LIG_CtBP_PxDLS_1 in mammals is due to the substitution of a leucine to proline at aligned position 408, probably after the divergence of Synapsids and Sauropsida. Mendoza *et al.* (2015) showed that the presence of the CtBP binding region in the bird *Taeniopygia guttata* has been associated with the potential FOXP4 regulation capacity. This finding for CtBP interaction may be associated with an enhanced potential for transcription repression of FOXP4, known for FOXP1 and FOXP2 (Mendoza *et al.*, 2015). At aligned position 689, almost all non-mammals present a motif that interacts with FHA (LIG_FHA_1 or LIG_FHA_2), while mammals present a DOC_USP7_1 motif.

To better understand the role of SLIMs in evolution, we additionally compared members within the *FOXP* family to verify the number of unique linear motifs in each paralogue (Table 5). The number of predicted types of SLIMs range from 28 (FOXP2) to 39 (FOXP4). Furthermore, FOXP3 presents three unique motifs (DOC_PP2B_1, TRG_NLS_MonoCore_2, and TRG_NLS_MonoExtN_4), FOXP1 presents four (DEG_SCF_FBW7_2, LIG_PCNA_PIPBox_1, LIG_WD40_WDR5_1, and TRG_NES_CRM1_1), while FOXP2 presents no unique SLIM. FOXP4 presents six motifs, among which three (DOC_PP1_RVXF_1, LIG_BRCT_BRAC1_2, and TRG_NLS_MonoExtC_3) are common to almost all species investigated in the current study.

Molecular evolutionary patterns

Evolutionary tests for *FOXP1, FOXP2, FOXP3* and *FOXP4* considering all the tetrapod species investigated in this study indicated that the best log-likelihood model is M1a, which assumes purifying selection and neutral ω values. *FOXP1*, *FOXP2* and *FOXP4* present more than 95% of the sites with ω = 0.03066, 0.01965 and 0.02778, respectively (Table S7), indicating a strong role for purifying selection. *FOXP3* presents 10% of ω values equal to 1, which indicates molecular neutral evolution and/or relaxation of functional constraints.

Additionally, we used the results from the Bayes Empirical Bayes (BEB) test to calculate the posterior probabilities that each codon is from the site class of positive selection (Yang, 2007). The BEB values are only significant for M2 and M8 (that include such selection), therefore this last strategy was only adopted to detect eventual functional sites. Such analysis shows four sites in mammals with ω > 1 with probability > 91%, but the *p* value are not significant (Table S7). Regardless, it is important highlight them: One of the sites inferred with ω = 1.06 (probability = 98.9%) is located at position 194 of *FOXP3* (Table 3), which presents differences in SLIM prediction (MOD_GSK3_1, MOD_ProDkin_1, MOD_CK1_1, DEG_SCF_FBW7_1, and DOC_WW_Pin1_4) in *Saimiri boliviensis, Callithrix jacchus,* and *Tarsius syrichta* when they are compared with the other species. While *Saimiri boliviensis* and *Callithrix jacchus* probably share the same linear motifs because of their clear and relatively recent common origin, in *Tarsius syrichta*, which is phylogenetically more distant may present them because of convergent evolution (Tables S6.3 and 3). MOD_ProdKin_1 is a post-translational modification site phosphorylated by a MAP kinase, while DEG_SCF_FBW7_1 is a degradation site mediated by an important protein complex (Skp, Cullin, F-box containing complex or SCF) that plays a role in checkpoints during the cell cycle (Nguyen Ba *et al.*, 2012). DOC_WW_Pin1_4 interacts with the enzyme Pin1, whose function is also associated with the cell cycle, among others. Additionally, Pin1 regulates the immune response (Gavva *et al.*, 1997; Wulf *et al.*, 2002; Wijchers *et al.*, 2006; Saxena *et al.*, 2010), which is a known function of FOXP3. Again, we identified the presence of the same SLIMs in two distinct primate branches that live in similar environments (rainforest) allow us to infer that a simple neutral model is insufficient to explain this scenario.

In the case of *FOXP4* (Table S8), the Branch Site model indicated that mammals have a ω value 3.7 times higher than non-mammals (0.66102 *versus* 0.18012), a result compatible with relaxation of evolutionary pressures. This striking difference ($p < 0.001$) may be attributed to certain changes such as the absence of the interaction site for CtBP (LIG_CtBP_PxDLS_1) in all mammals (except *Sus scrofa*). Another structural/functional change that can explain the distinct ω values

observed between mammals and non-mammals is the presence of a region rich in glutamine in the mammalian FOXP4, associated to its repression ability.

Conclusion

In summary, our study reveals some important general and more specific findings. For instance, 70% of the disorder content has been retained in FOXP1, FOXP2, and FOXP4 orthologues. Some of the results obtained can be associated with taxa-specific conditions, while others may represent molecular convergence. In fact, we found changes at FOXP3 sites with possible functional implications in the primate branch, including *Homo* genus. Finally, FOXP1 and FOXP4 results show instigating differences between mammals and non-mammals, suggesting its role in the emergence of adaptive novelty inside the *Tetrapoda* taxa. Our results indicate that part of *FOXP* evolutionary "stability" over a long evolutionary period may be attributed to the maintenance of a similar proportion of disordered regions, but not to the amino acid content or linear motifs. Moreover, some of the changes can be suggested as indicating taxa-specific adaptations, since they are probably functional.

Acknowledgements

This research was financially supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Akbar AN, Vukmanovic-Stejic M, Taams LS and Macallan DC (2007) The dynamic co-evolution of memory and regulatory CD4+ T cells in the periphery. Nat Rev Immunol 7: 231–237.

Andersen KG, Nissen JK and Betz AG (2012) Comparative genomics reveals key gain-of-function events in Foxp3 during regulatory T cell evolution. Front Immunol 3: 1–16.

Benayoun BA, Caburet S and Veitia RA (2011) Forkhead transcription factors: Key players in health and disease. Trends Genet 27: 224–232.

Bieniasz PD (2006). Late budding domains and host proteins in enveloped virus release. Virology 344: 55–63.

Bowers JM and Konopka G (2012) The role of the FOXP family of transcription factors in ASD. Dis Markers 33: 251–260.

Castellano S, Parra G, Sánchez-Quinto FA, Racimo F, Kuhlwilm M, Kircher M, Sawyer S, Fu Q, Heinze A, Nickel B, *et al.* (2014) Patterns of coding variation in the complete exomes of three Neandertals. Proc Natl Acad Sci USA 111: 6666–6671.

Cesario JM, Almaidhan AA and Jeong J (2016) Expression of forkhead box transcription factor genes *Foxp1* and *Foxp2* during jaw development. Gene Expr Patterns 20: 111-119.

Coffer PJ and Burgering BMT (2004) Forkhead-box transcription factors and their role in the immune system. Nat Rev Immunol 4: 889–899.

Cooper DL (2006) Broca's arrow: Evolution, prediction, and language in the brain. Anat Rec B New Anat 289: 9–24.

Dinkel H, Michael S, Weatheritt RJ, Davey NE, Roey KV, Altenberg B, *et al.* (2012). ELM - The database of eukaryotic linear motifs. Nucleic Acids Res 40: 1–10.

Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, Toedt G, Uyar B, Seiler M, Budd A, *et al.* (2014) The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res 42: 1–8.

Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN and Kurgan L (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 28: 75–83.

Dunker AK, Obradovic Z, Romero P, Garner EC and Brown CJ (2000) Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform 11: 161–171.

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP and Pääbo S (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418: 869–872.

Enard W (2011) FOXP2 and the role of cortico-basal ganglia circuits in speech and language evolution. Curr Opin Neurobiol 21: 415–424.

Forman-Kay JD and Mittag T (2013) From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. Structure 21: 1492–1499.

Gavva, NR, Gavva R, Ermekova K, Sudol M, and Shen C-K (1997) Interaction of WW Domains with Hematopoietic Transcription Factor p45/NF-E2 and RNA Polymerase II. J Biol Chem 272: 24105-24109.

Göttlinger HG, Dorfman T, Sodroski JG and Haseltine WA (1991) Effect of mutations affecting the p6 gag protein on human immunodeficiency virus particle release. Proc Natl Acad Sci USA 88: 3195–3199.

Gould CM, Diella F, Via A, Puntervoll P, Gemünd C, Chabanis-Davidson S, Michael S, Sayadi A, Bryne JC, Chica C, *et al* (2010) ELM: The status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res 38: 167–180.

Hagai T, Azia A, Babu MM and Andino R (2014) Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. Cell Rep 7: 1729–1739.

Huang H and Sarai A (2012) Analysis of the relationships between evolvability, thermodynamics, and the functions of intrinsically disordered proteins/regions. Comput Biol and Chem 41: 51–57.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z and Dunker AK (2004) The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32: 1037–1049.

Krause J, Lalueza-Fox C, Orlando L, Enard W, Green RE, Burbano HA, Hublin JJ, Hänni C, Fortea J, de la Rasilla M *et al.* (2007) The derived FOXP2 variant of modern humans was shared with Neandertals. Curr Biol 17: 1908–1912.

Lam EW, Brosens JJ, Gomes AR and Koo CY (2013) Forkhead box proteins: Tuning forks for transcriptional harmony. Nat Rev Cancer 13: 482–495.

Li G, Wang J, Rossiter SJ, Jones G and Zhang S (2007) Accelerated FOXP2 evolution in echolocating bats. PLoS One 2: e900.

Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN and Dunker AK (2006) Intrinsic disorder in transcription factors. Biochemistry 45: 6873–6888.

Liu Y, Matthews KS and Bondos SE (2008) Multiple intrinsically disordered sequences alter DNA binding by the homeodomain of the *Drosophila* Hox protein ultrabithorax. J Biol Chem 283: 20874–20887.

Liu Z and Huang Y (2014) Advantages of proteins being disordered. Protein Sci 23: 539–550.

Marín M, Uversky VN and Ott T (2013) Intrinsic disorder in pathogen effectors: Protein flexibility as an evolutionary hallmark in a molecular arms race. Plant Cell 25: 3153–3157.

Mendoza E, Tokarev K, Düring DN, Retamosa EC, Weiss M, Arpenik N and Scharff C (2015) Differential co-expression of FOXP1, FOXP2 and FOXP4 in zebra finch (*Taeniopygia guttata*) song system. J Comp Neurol 523: 1318–1340.

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TL, Stadler T *et al.* (2011) Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334:521–524.

Morris G and Fanucchi S (2016) A key evolutionary mutation enhances DNA binding of the FOXP2 forkhead domain. Biochemistry 55: 1959-1967.

Murakami H, Aiba H, Nakanishi M and Murakami-Tonami Y (2010) Regulation of yeast forkhead transcription factors and FOXM1 by cyclin-dependent and polo-like kinases. Cell Cycle 9: 3233–3242.

Nakajima H, Toyoshima-Morimoto F, Taniguchi E and Nishida E (2003) Identification of a consensus motif for PlK (Polo-like kinase) phosphorylation reveals Myt1 as a Plk1 substrate. J Biol Chem 278: 25277–25280.

Neduva V and Russell RB (2005) Linear motifs: Evolutionary interaction switches. FEBS Lett 579: 3342–3345.

Nguyen Ba, AN, Yeh BJ, van Dyk D, Davidson AR, Andrews BJ, Weiss EL and Moses AM (2012) Proteome-wide discovery of evolutionary conserved sequences in disordered regions. Sci Signal 5: rs1.

Obsil T and Obsilova V (2008) Structure/function relationships underlying regulation of FOXO transcription factors. Oncogene 27: 2263–2275.

Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, *et al.* (2011) A molecular phylogeny of living primates. PLoS Genet 7: 1–17.

Pigliucci M (2008) Is evolvability evolvable? Nat Rev Genet 9: 75–82.

Potenza E, Di Domenico T, Walsh I and Tosatto SC (2014) MobiDB 2.0: An improved database of intrinsically disordered and mobile proteins. Nucleic Acids Res 43: D315–D320.

Santos ME*,* Athanasiadis A, Leitão AB, DuPasquier L and Sucena E (2011) Alternative splicing and gene duplication in the evolution of the FOXP gene subfamily. Mol Bio Evol 28: 237–247.

Saxena U, Owens L, Graham JR, Cooper GM and Hansen U (2010) Prolyl Isomerase Pin1 regulates Transcription Factor LSF (TFCP2) by facilitating dephosphorylation at two-serine-proline motifs. J Biol Chem 285: 31139-31147.

Schön C, Wochnik A, Rössner A, Donow C and Knöchel W (2006) The FOXP subclass in *Xenopus laevis* development. Dev Genes Evol 216: 641–646.

Song S, Liu L, Edwards SV and Wu S (2012) Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci USA 109: 14942–14947.

Song X, Wang Y and Tang Y (2013) Rapid diversification of FOXP2 in teleosts through gene duplication in the teleost-specific whole genome duplication event. PLoS One 8: e83858.

Song X, Tang Y and Wang Y (2016) Genesis of the vertebrate *FoxP* subfamily member genes occurred during two ancestral whole genome duplication events. Gene 588: 156-162

Takahashi K, Liu FC, Oishi T, Mori T, Higo N, Hayashi M, Hirokawa K and Takahashi H (2008) Expression of FOXP2 in the developing monkey forebrain: Comparison with the expression of the genes FOXP1, PBX3, and MEIS2. J Comp Neurol 509: 180–189.

van der Horst A, de Vries-Smits AM, Brenkman AB, van Triest MH, van den Broek N, Colland F, Maurice MM and Burgering BM (2006) FOXP4 transcriptional activity is regulated by monoubiquitination and USP7/HAUSP. Nat Cell Biol 8: 1064–1073.

Wagner GP and Lynch VJ (2008) The gene regulatory logic of transcription factor evolution. Trends Ecol Evol 23: 377–385.

Wjichers PJEC, Burbach JPH and SMIDT MP (2006) In control of biology: of mice, man and Foxes. Biochem J 397: 233-246.

Wulf G, Liou, Y and Ryo A (2002) Role of Pin1 in the regulation of p53 stability and p21 transactivation, and cell cycle chackpoints in response to DNA damage. J Biol Chem 277: 47976-47979.

Xue B, Dunbrack RL, Williams RW, Dunker AK and Uversky VN (2010a) PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. Biochim Biophys Acta 1804: 996–1010.

Xue B, Williams RW, Oldfield CJ, Dunker AK and Uversky VN (2010b) Archaic chaos: Intrinsically disordered proteins in Archaea. BMC Syst Biol 4: S1.

Xue B, Brown CJ, Dunker AK and Uversky VN (2013) Intrinsically disordered regions of p53 family are highly diversified in evolution. Biochim Biophys Acta 1834:725–738.

Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.

Zhang J, Webb DM and Podlaha O (2002) Accelerated protein evolution and origins of human-specific features: FOXP2 as an example. Genetics 162: 1825-1835.

Internet Resources

http://cdna.eva.mpg.de/, last accessed March, 2015.

http://ensembl.org/, last accessed October, 2014.

http://ncbi.nlm.nih.gov, last accessed October, 2014.

Table 1 - Mean disorder proportion for FOXP proteins by class[1]

Class	FOXP2	FOXP4	FOXP1	FOXP3[2]
Mammals	0.7011	0.7321	0.6915	0,3065
Birds	0.7039	0.6858	0.6782	
Reptiles	0.6984	0.6827	0.6713	
Amphibians	0.6305	0.7068	NA	

[1]Mammals showed significant higher proportions than the other groups, as assessed by the Kruskal-Wallis test, for FOXP1and FOXP4 ($P < 0.001$). Additionally, according to the same test, amphibians presented a lower degree of disorder for FOXP2 ($P < 0.01$).

[2]Only mammalian genes were used for the FOXP3 analysis.

NA: Not Available. Since several base pairs in *Xenopus laevis* FOXP1 sequence are missing, we excluded it of the analysis.

Table 2. Linear motifs changes in representative species of *Tetrapoda*, as predicted by ELM.

	Aligned Position	Nucleotide	Amino acid	Grantham Score	*Homo sapiens*	*Pant roglodytes*	*Pan paniscus*	*Mus musculus*	*Taeniopygia guttata*	*Serinus canaria*	*Anolis carolinensis*	*Xenopus laevis*
FOXP1	33	GGT->AGT	Gly->Ser	56	0	0						*
		GGT->GCA	Gly->Ala	60				1				
		GGT->AGC	Gly->Ser	56					$1^{p,r,v}$	$1^{p,r,v}$		
		GGT->GGC	Gly->Gly	Syn							$1^{p,r,v}$	
FOXP2	314	GCA->GCG	Ala->Ala	Syn	0	0		0				
		GCG->CCA	Ala->Pro	27								
		GCG->TCT	Ala->Ser	99					1^{d}	1^{d}	1^{d}	$1^{d,s}$
		GCG->CCC	Ala->Pro	27								
	368	AAC->ACC	Asn->Thr	65	$0^{o4,q3}$	$1^{o4,q4}$		$1^{o4,q4}$	$1^{o4,q4}$	$1^{o4,q4}$	$1^{o4,q4}$	$1^{o4,q4}$
	390	AGT->AAT	Ser->Asn	46	$0^{o,q}$	1		1	1	1	1	0^{o}
FOXP4	408	CCG->CCA	Pro->Pro	Syn	0		0	1				
		CCG->CTG	Pro->Leu	98					1^{i}	1^{i}	1^{i}	
		CCG->TTG	Pro->Leu	98								1^{i}
	689	TCG->TCA	Ser->Ser	Syn	0^{d}		0^{d}					
		TCG->TTG	Ser->Leu	145								
		TCG->GTG	Ser->Val	124				1				
		TCG->ACA	Ser->Thr	58							1^{k}	
		TCG->ACG	Ser->Thr	58					1^{k}	1^{k}		
		TCG->GTC	Ser->Val	124								1^{j}

* indicates gap.

Syn = synonymous changes.

Zero (0) indicates the amino acid present in the Homo sapiens reference sequence, whereas 1 indicates a variant amino acid .Subscribed letters indicate the predicted presence of specific Eukariotic Linear Motifs (see code shown in table S9). Subscribed numbers is the number of times each SLIMs appear.

The nature of modification is not representing an ancestry and descendant relationship. Grantham scores predicted as conservative (0-50), moderately conservative (51-100), moderately radical (101-150) or radical (>151).

Table 3. FOXP3 Linear motifs changes in Mammals, as predicted by ELM.

Aligned Position	**194**						
Nucleotide	GTG->ATG	GTG->ACA	GTG->TTG	GTG->GGG	GTG->GCA	GTG->GCG	GTG->ACG
Amino acid	Val->Met	Val->Thr	Val->Leu	Val->Gly	Val->Ala	Val->Ala	Val->Thr
Grantham Score	21	69	32	109	64	64	69
Homo sapiens	0^{o}						
Pan troglodytes	0^{o}						
Pan paniscus	0^{o}						
Gorilla gorilla	0^{o}						
Pongo abellii	0^{o}						
Pongo pygameus	0^{o}						
Hylobates lar	0^{o}						
Nomascus	0^{o}						
Macaca mulatta	1^{o}						
Papio anubis	1^{o}						
Chlorocebus sabaeus	1^{o}						
Saimiri boliviensis		$1^{b,e,o,q2,w}$					
Callithrix jacchus		$1^{b,e,o,q2,w}$					
Galeopterus variegatus	0^{o}						
Tarsius syrichta		$1^{b,e,o,q2,w}$					
Tupaia chinensis			$1^{m3,j,o}$				
Sorex araneus				1^{o}			
Mus musculus					1^{c}		

Table 4 - FOXP3-specific changes in primates

Organisms	Aligned position	Human position	AA Change	Motifs[1]
Neanderthal and Humans	140	132	Pro->Thr	(+2) DEG_SCF_FBW7_1
	183	175	Gly->Ser	(+) LIG_PTAP_UEV_1
Neanderthal	192	184	Ser->Leu	(-) MOD_CK1_1, (+) DOC_MAPK_1
Catarrhini	278	270	Pro->Ser	(+) MOD_GSK3_1
Haplorhini[2]	82	74	Val->Leu	(-) DOC_WW_Pin1_4, (-) MOD_ProDKin_1
	97	89	Ser->Leu	
	129	121	Arg->His	
	132	124	Asp->Glu	
	181	173	Ser->Asn	(-)DOC_WW_Pin1_4, (-) MOD_ProDKin_1
	246	238	Val->Met	
	262	254	Gly->Ser	
	338	325	Phe->Leu	
	424	411	Phe->Leu	

[1]+: change causes motif gain; -: change causes motif loss.
[2]Excluding *Tarsius syrichta*.

Table 5 - Number of shared and unique short linear motifs (SLIMs) among Tetrapoda FOXPs

Protein	Total type of SLIMs	Number of unique SLIMs	Total SLIMs in *Homo sapiens*	Total SLIMs in *Pan sp.*	Total SLMIs in *Serinus canaria*[1]	Total of species compared
FOXP1	34	4[2]	132	132	135	50
FOXP2	28	0	143	142	140	54
FOXP3	32	3[3]	69	62	-	57
FOXP4	39	6[4]	142	143	160	65

[1]Bird, representing Sauropsida.

[2]DEG_SCF_FBW7_2, LIG_PCNA_PIPBox_1, LIG_WD40_WDR5_1, and TRG_NES_CRM1_1);

[3] DOC_PP2B_1, TRG_NLS_MonoCore_2, and TRG_NLS_MonoExtN_4;

[4] FOXP4 presents six motifs, among which three (DOC_PP1_RVXF_1, LIG_BRCT_BRAC1_2, and TRG_NLS_MonoExtC_3) are common to almost all species investigated in the current study.

Supplementar Material

Table S1. List of Tetrapoda species and the respective sequence codes retrieved to each one of the FOXP subfamily member studied.

Class	Common Name	Species	*FOXP1*	*FOXP2*	*FOXP3*	*FOXP4*
Mammalia	Human	*Homo sapiens*	NM_032682.5[a]	ENST00000350908[b]	ENST00000376207[b]	ENST00000373060[b]
Mammalia	Chimpanzee	*Pan troglodytes*	XM_001140515.3[a]	ENSPTRT00000036314[b]	JX296015.1a	ENSPTRT00000042554.3[b]
Mammalia	Bonobo	*Pan paniscus*	-	-	XM_003807259.2[a]	XM_003833311.2[a]
Mammalia	Gorilla	*Gorilla gorilla gorilla*	XM_004035888.1[a]	AY143180.1[a]	ENSGGOT00000025034[b]	ENSGGOT00000016440[b]
Mammalia	Orangutan	*Pongo abelii*	XM_009238796.1[a]	ENSPPYT00000020887[b]	ENSPPYG00000020341[b]	XM_002816867.3[a]
Mammalia	Pygmy Orangutan	*Pongo pygmaeus*	-	-	JX296014.1[a]	-
Mammalia	Gibbon	*Nomascus leucogenys*	XM_003264887.2[a]	XM_003261215.2[a]	ENSNLET00000010324[b]	XM_003266292.1[a]
Mammalia	White-hand Gibbon	*Hylobates lar*	-	-	JX296016.1[a]	-
Mammalia	Macaque	*Macaca mulatta*	NM_001266321.1[a]	NM_001033021.1[a]	NM_001032918.1[a]	NM_001266091.1[a]
Mammalia	Baboon	*Papio anubis*	XM_009200345.1[a]	XM_009203721.1[a]	ENSPANT00000010875[b]	ENSPANT00000017705[b]
Mammalia	Vervet monkey	*Chlorocebus sabaeus*	XM_007984970.1[a]	XM_007982653.1[a]	XM_007991691.1[a]	XM_007972623.1[a]
Mammalia	Squirrel monkey	*Saimiri boliviensis boliviensis*	XM_003939460.1[a]	XM_003921066.1[a]	XM_010350232.1[a]	XM_003922987.1[a]
Mammalia	Marmoset	*Callithrix jacchus*	XM_002758548.3[a]	XM_002751708.3[a]	GQ284839.1[a]	XM_003732665.2[a]
Mammalia	Sunda flying lemur	*Galeopterus variegatus*	XM_008571774.1[a]	XM_008571504.1[a]	XM_008587131.1[a]	XM_008583550.1[a]

Mammalia	Bushbaby	*Otolemur_garnettii*	-	-	-	ENSOGAT00000010116[b]
Mammalia	Tarsier	*Tarsius syrichta*	-	-	XM_008074655.1[a]	XM_008062590.1[a]
Mammalia	Shrew	*Sorex araneus*	-	-	XM_004606453.1[a]	XM_004605787.1[a]
Mammalia	Chinese tree Shrew	*Tupaia chinensis*	XM_006158050.1[a]	XM_006157413.1[a]	XM_006163498.1[a]	XM_006148586.1[a]
Mammalia	Mouse	*Mus musculus*	ENSMUST00000113326[b]	NM_053242.4[a]	NM_001199348.1[a]	ENSMUST00000097311[b]
Mammalia	Rat	*Rattus norvegicus*	ENSRNOT00000013271[b]	XM_006236115.2[a]	ENSRNOT00000015641[b]	XM_008766868.1[a]
Mammalia	Chinese hamster	*Cricetulus griseus*	XM_007631502.1[a]	XM_007650925.1[a]	XM_007630484.1[a]	XM_003508987.2[a]
Mammalia	Degu	*Octodon degus*	XM_004629502.1[a]	XM_004644864.1[a]	XM_004646275.1[a]	XM_004624157.1[a]
Mammalia	Guinea_pig	*Cavia_porcellus*	-	-	-	XM_003473873.2[a]
Mammalia	Rabbit	*Oryctolagus cuniculus*	XM_002713312.2[a]	NM_001122936.1[a]	ENSOCUT00000010135[b]	-
Mammalia	Pika	*Ochotona princeps*	XM_004581620.1[a]	XM_004592462.1[a]	XM_004587932.1[a]	XM_004590349.1[a]
Mammalia	Sperm whale	*Physeter catodon*	XM_007105389.1[a]	XM_007106509.1[a]	XM_007113265.1[a]	XM_007120287.1[a]
Mammalia	Kiler whale	*Orcinus orca*	-	-	XM_004281924.1[a]	XM_004267622.1[a]
Mammalia	Camel	*Camelus ferus*	-	XM_006184670.1[a]	XM_006195238.1[a]	XM_006202021.1[a]
Mammalia	Alpaca	*Vicugna_pacos*	XM_006196491.1[a]	XM_006206156.1[a]	XM_006213486.1[a]	XM_006202021.1[a]
Mammalia	Cow	*Bos taurus*	-	-	NM_001045933.1[a]	NM_001205524.1[a]
Mammalia	Horse	*Equus caballus*	-	-	NM_001163272.1[a]	ENSECAT00000001439[b]
Mammalia	Pig	*Sus_scrofa*	-	-	-	XM_001926847.2[a]
Mammalia	Weddell seal	*Leptonychotes weddellii*	-	-	XM_006743003.1[a]	XM_006729196.1[a]

Mammalia	Walrus	*Odobenus rosmarus divergens*	XM_004404552.1[a]	XM_004408502.1[a]	XM_004396823.1[a]	XM_004408054.1[a]
Mammalia	Panda	*Ailuropoda melanoleuca*	-	-	XM_002917744.1[a]	XM_002914517.1[a]
Mammalia	Cat	*Felis catus*	-	NM_001113177.1[a]	ENSFCAT00000003824[b]	-
Mammalia	Tiger	*Panthera tigris*	-	XM_007089152.1[a]	XM_007095999.1[a]	-
Mammalia	Ferret	*Mustela_putorius_furo*	-	-	ENSMPUT00000014169[b]	XM_004739938.1[a]
Mammalia	Dog	*Canis lupus familiaris*	-	-	AB501357.1[a]	XM_538914.5[a]
Mammalia	Rhinoceros	*Ceratotherium simum simum*	-	XM_004418795.1[a]	XM_004434916.1[a]	XM_004424166.1[a]
Mammalia	Brandt's myotis	*Myotis brandtii*	XM_005885064.1[a]	XM_005862259.1[a]	XM_005864530.1[a]	XM_005885813.1[a]
Mammalia	Big brown bat	*Eptesicus fuscus*	XM_008154141.1[a]	XM_008152407.1[a]	XM_008158246.1[a]	XM_008158925.1[a]
Mammalia	Black flying fox	*Pteropus alecto*	XM_006917473.1[a]	XM_006910602.1[a]	XM_006922727.1[a]	XM_006926566.1[a]
Mammalia	Star-nosed mole	*Condylura cristata*	XM_004675975.1[a]	XM_004676957.1[a]	XM_004690039.1[a]	XM_004673459.1[a]
Mammalia	Western European hedgehog	*Erinaceus europaeus*	XM_007523308.1[a]	XM_007516144.1[a]	XM_007529426.1[a]	-
Mammalia	Elephant-shrew	*Elephantulus edwardii*	XM_006900488.1[a]	XM_006882655.1[a]	XM_006902719.1[a]	XM_006882009.1[a]
Mammalia	Cape golden mole	*Chrysochloris asiatica*	XM_006874623.1[a]	XM_006859303.1[a]	XM_006876686.1[a]	XM_006860443.1[a]
Mammalia	Lesser hedgehog tenrec	*Echinops telfairi*	XM_004702425.1[a]	XM_004702573.1[a]	XM_004713140.1[a]	-
Mammalia	Aardvark	*Orycteropus afer afer*	XM_007946080.1[a]	XM_007944086.1[a]	XM_007958374.1[a]	XM_007936267.1[a]
Mammalia	Mannatee	*Trichechus manatus latirostris*	XM_004372708.1[a]	XM_004382570.1[a]	XM_004376834.1[a]	XM_004379437.1[a]
Mammalia	Elephant	*Loxodonta africana*	XM_003409786.1[a]	XM_003407220.1[a]	ENSLAFT00000003503[b]	XM_003403923.1[a]
Mammalia	Armadillo	*Dasypus novemcinctus*	-	-	XM_004465020.1[a]	-

Mammalia	Opossum	*Monodelphis domestica*	XM_007500090.1[a]	XM_007504167.1[a]	-	-
Mammalia	Platypus	*Ornithorhynchus_anatinus*	-	-	XM_001507231.3[a]	-
Bird	Duck	*Anas platyrhynchos*	ENSAPLT00000012301[b]	XM_005010980.1[a]	-	XM_005023422.1[a]
Bird	Emperor penguin	*Aptenodytes forsteri*	XM_009284364.1[a]	XM_009286822.1[a]	-	XM_009272821.1[a]
Bird	Anna's hummingbird	*Calypte anna*	XM_008499642.1[a]	XM_008496149.1[a]	-	XM_008499441.1[a]
Bird	Peregrine falcon	*Falco peregriunus*	XM_005229951.1[a]	XM_005242504.1[a]	-	XM_005234510.1[a]
Bird	Flycatcher	*Ficedula albicollis*	XM_005053288.1[a]	XM_005039338.1[a]	-	ENSFALT00000003213[b]
Bird	Chicken	*Gallus gallus*	NM_001024827.1[a]	ENSGALT00000015345[b]	-	XM_003642748.2[a]
Bird	Golden-collared manakin	*Manacus vitellinus*	XM_008925680.1[a]	XM_008932480.1[a]	-	XM_008934298.1[a]
Bird	Budgerigar	*Melopsittacus undulatus*	XM_005149418.1[a]	NM_001281546.1[a]	-	XM_005142863.1[a]
Bird	Ground tit	*Pseudopodoces humilis*	XM_005521985.1[a]	-	XM_005533346.1[a]	XM_005529345.1[a]
Bird	Canary	*Serinus canaria*	XM_009091186.1[a]	XM_009086754.1[a]	-	XM_009098374.1[a]
Bird	Zebra finch	*Taeniopygia_guttata*	ENSTGUT00000010342[b]	NM_001048263.2[a]	-	NM_001279260.1[a]
Bird	White-throated sparrow	*Zonotrichia albicollis*	XM_005491385.1[a]	XM_005481975.1[a]	-	XM_005492836.1[a]
Reptile	Crocodile	*Alligator mississippiensis*	-	-	XM_006261276.1[a]	XM_006275827.1[a]
Reptile	Chinese Crocodile	*Alligator sinensis*	-	-	XM_014523895.1[a]	XM_006025064.1[a]
Reptile	Anole lizard	*Anolis carolinensis*	XM_008105381.1[a]	BK008623.1[a]	XM_008103924.1[a]	XM_008109698.1[a]
Reptile	Western Painted Turtle	*Chrysemys picta bellii*	-	-	-	XM_005308300.2[a]
Reptile	Green Turtle	*Chelonia mydas*	XM_007061459.1[a]	XM_007070534.1[a]	-	XM_007068896.1[a]

Reptile	Chinese softshell turtle	*Pelodiscus_sinensis*	ENSPSIT00000009876[b]	-	-	ENSPSIT00000013410[b]
Reptile	Python	*Python bivittatus*	XM_007421309.1[a]	XM_007428286.1[a]	XM_007420528.1[a]	XM_007426185.1[a]
Amphibious	Xenopus (silurana)	*Xenopus tropicalis*	-	-	XM_002940971.2[a]	NM_001077187.2[a]
Amphibious	Xenopus	*Xenopus laevis*	NM_001095533.1[a]	NM_001095669.1[a]	NM_001127727.1[a]	NM_001095615.1[a]
Amphibious	Emei music-frog	*Babina daunchina**	-	JX543513.1[a]	-	-
Amphibious	Paddle-tail newt	*Pachytriton labiatus**	-	JX543512.1[a]	-	-

Database: a, NCBI; b Ensemble, *individuals used just for disorder analysis.

Table S2.1 Disorder proportion of FOXP1 orthologues.

Species	Disorder Proportion	AA	Order	Class
Homo sapiens	0.692762186	677	Primates	Mammal
Pan troglodytes	0.692762186	677	Primates	Mammal
Gorilla gorilla	0.692762186	677	Primates	Mammal
Pongo abelii	0.692762186	677	Primates	Mammal
Nomascus leucogenys	0.692762186	677	Primates	Mammal
Macaca mulatta	0.692762186	677	Primates	Mammal
Papio anubis	0.692762186	677	Primates	Mammal
Chlorocebus sabaeus	0.692762186	677	Primates	Mammal
Callithrix jacchus	0.692762186	677	Primates	Mammal
Saimiri boliviensis	0.687869822	676	Primates	Mammal
Galeopterus variegatus	0.666174298	677	Dermoptera	Mammal
Tupaia chinensis	0.690828402	676	Scandentia	Mammal
Mus musculus	0.694814815	675	Rodentia	Mammal
Rattus norvegicus	0.703234880	711	Rodentia	Mammal
Cricetulus griseus	0.692532943	683	Rodentia	Mammal
Octodon degus	0.702346041	682	Rodentia	Mammal
Oryctolagus cuniculus	0.697810219	685	Lagomorpha	Mammal
Ochotona princeps	0.696035242	681	Lagomorpha	Mammal
Physeter catodon	0.682020802	673	Cetacea	Mammal
Vicugna pacos	0.692762186	677	Artiodactyla	Mammal
Camelus ferus	0.622974963	679	Artiodactyla	Mammal
Ceratotherium simum simum	0.620029455	679	Perissodactyla	Mammal
Felis catus	0.622974963	679	Carnivora	Mammal
Panthera tigris	0.622974963	679	Carnivora	Mammal
Odobenus rosmarus divergens	0.685376662	677	Carnivora	Mammal

Chrysochloris asiatica	0.690615836	682	Afrosoricida	Mammal
Condylura cristata	0.688693098	681	Soricomorpha	Mammal
Echinops telfairi	0.689807976	677	Afrosoricida	Mammal
Elephantulus edwardii	0.701317716	683	Macroscelidea	Mammal
Eptesicus fuscus	0.693333333	675	Chiroptera	Mammal
Myotis brandtii	0.701624815	677	Chiroptera	Mammal
Pteropus alecto	0.671111111	675	Chiroptera	Mammal
Erinaceus europaeus	0.689910979	674	Erinaceomorpha	Mammal
Monodelphis domestica	0.676514032	677	Didelphimorphia	Mammal
Orycteropus afer afer	0.691285081	677	Tubulidentata	Mammal
Loxodonta africana	0.700292398	684	Proboscidea	Mammal
Trichechus manatus latirostris	0.697810219	685	Sirenia	Mammal
Serinus canaria	0.687134503	684	Passeriformes	Bird
Taeniopygia guttata	0.653521127	710	Passeriformes	Bird
Manacus vitellinus	0.682748538	684	Passeriformes	Bird
Ficedula albicollis	0.678885630	682	Passeriformes	Bird
Melopsittacus undulatus	0.680819912	683	Psittaciformes	Bird
Zonotrichia albicollis	0.681286550	684	Passeriformes	Bird
Falco peregrinus	0.671052632	684	Falconiformes	Bird
Aptenodytes forsteri	0.670571010	683	Sphenisciformes	Bird
Calypte anna	0.689450223	673	Trochiliformes	Bird
Gallus gallus	0.670553936	686	Galliformes	Bird
Anas platyrhynchos	0.693997072	683	Anseriformes	Bird
Python bivittatus	0.678311499	687	Squamata	Reptilia
Anolis carolinensis	0.701601164	687	Squamata	Reptilia
Pelodiscus sinensis	0.649851632	674	Testudines	Reptilia
Chelonia mydas	0.655325444	676	Testudines	Reptilia
Xenopus laevis	0.600346021	578	Anura	Amphibia

Table S2.2. Ordered regions for FOXP1 orthologues.

Species	Structured Region			
Homo sapiens		220-234	302-364	440-574
Pan troglodytes		220-234	302-364	440-574
Gorilla gorilla		220-234	302-364	440-574
Pongo abelii		220-234	302-364	440-574
Nomascus leucogenys		220-234	302-364	440-574
Macaca mulatta		220-234	302-364	440-574
Papio anubis		220-234	302-364	440-574
Chloroscebus sabaeus		220-234	302-364	440-574
Saimiri boliviensis boliviensis	52-55	219-233	302-363	439-573
Callithrix jacchus		220-234	302-359	440-574
Galeopterus variegatus	52-57	220-234	302-364	440-587
Tupaia chinensis		220-235	301-364	440-573
Mus musculus		218-230	300-362	438-572
Cricetulus griseus		226-240	308-370	446-580
Rattus novergicus		254-268	336-398	474-610
Oryctolagus cuniculus		228-241	310-372	448-582
Ochotona princeps		225-239	307-369	445-578
Octodon degus		227-238	309-369	445-579
Eptesicus fuscus		220-228	299-362	438-578
Myotis brandtii		222-230	301-364	440-574
Pteropus alecto	49-62	218-232	300-362	438-572
Ceratotherium simum simum	16-28 37-73	222-236	305-366	442-576
Felis catus	16-28 39-72	222-236	304-366	442-576
Panthera tigris	16-28 39-72	222-236	304-366	442-576
Odobemus rosmarus divergens	53-57	220-233	301-364	440-574

Erinaceus europaeus			217-231	298-361	437-571
Physeter catodon		52-56	216-229	298-354	436-572
Camelus ferus	13-29	36-69	222-360		442-575
Vicugna pacos		52-56	220-233	302-358	440-573
Echinops telfairi			220-236	301-364	441-574
Trichechus manatus			229-244	310-367	449-581
Orycteropus afe afer			220-235	301-364	441-574
Elephantulus edwardii			226-239	307-370	447-577
Chrysochloris asiatica			224-240	306-369	446-579
Condylura cristata			224-238	306-368	444-584
Loxodonta africana			227-241	308-365	447-581
Monodelphis domestica		52-63	218-228	302-364	441-576
Ficedula albicollis		47-57	225-237	308-364	445-581
Taeniopygia guttata		42-79	253-265	336-392	473-609
Falco peregrinus		44-59	227-239	310-366	447-583
Calypte anna			216-228	299-355	436-572
Aptenodytes forsteri		47-57	227-239	310-366	447-588
Zonotrichia albicollis		47-57	227-238	309-365	447-583
Manacus vitellinus		51-60	227-239	310-366	447-583
Serinus canaria		52-58	227-239	310-366	447-583
Melopsittacus undulatus		47-56	226-238	308-364	446-582
Gallus gallus		47-56	229-241	312-368	449-592
Anas platyrhynchos			226-238	309-365	446-582
Python bivittatus		47-56	230-244	311-378	450-582
Anolis carolinensis			230-246	312-364	450-582
Chelonia mydas		47-58	217-232	302-364	437-580
Pelodiscus sinensis		44-62	217-230	300-356	435-580
Xenopus laevis			121-161	202-258	341-473

Table S3.1 Disorder proportion for FOXP2 orthologues.

Species	Disorder Proportion	AA	Order	Class
Homo sapiens	0.702097902	715	Primates	Mammal
Pan troglodytes	0.702513966	716	Primates	Mammal
Gorilla gorilla	0.701262272	713	Primates	Mammal
Pongo abelii	0.684431978	713	Primates	Mammal
Nomascus leucogenys	0.701262272	713	Primates	Mammal
Macaca mulatta	0.701680672	714	Primates	Mammal
Papio anubis	0.701680672	714	Primates	Mammal
Chlorocebus sabaeus	0.701680672	714	Primates	Mammal
Saimiri boliviensis boliviensis	0.701680672	714	Primates	Mammal
Callithrix jacchus	0.705469845	713	Primates	Mammal
Galeopterus variegatus	0.690834473	731	Dermoptera	Mammal
Tupaia chinensis	0.701680672	714	Scandentia	Mammal
Mus musculus	0.701680672	714	Rodentia	Mammal
Rattus norvegicus	0.701408450	710	Rodentia	Mammal
Cricetulus griseus	0.704323570	717	Rodentia	Mammal
Octodon degus	0.708913649	718	Rodentia	Mammal
Oryctolagus cuniculus	0.702513966	716	Lagomorpha	Mammal
Ochotona princeps	0.705382436	706	Lagomorpha	Mammal
Physeter catodon	0.689265537	708	Cetacea	Mammal
Vicugna pacos	0.704225352	710	Artiodactyla	Mammal
Camelus ferus	0.704225352	710	Artiodactyla	Mammal
Ceratotherium simum simum	0.701262272	713	Perissodactyla	Mammal
Felis catus	0.698727015	707	Carnivora	Mammal
Panthera tigris	0.689404934	689	Carnivora	Mammal
Odobenus rosmarus divergens	0.699576869	709	Carnivora	Mammal

Chrysochloris asiatica	0.701262272	713	Afrosoricida	Mammal
Echinops telfairi	0.711888111	715	Afrosoricida	Mammal
Elephantulus edwardii	0.700842697	712	Macroscelidea	Mammal
Eptesicus fuscus	0.704481793	714	Chiroptera	Mammal
Myotis brandtii	0.704067321	713	Chiroptera	Mammal
Pteropus alecto	0.704323570	717	Chiroptera	Mammal
Erinaceus europaeus	0.703389830	708	Erinaceomorpha	Mammal
Condylura cristata	0.701680672	714	Soricomorpha	Mammal
Orycteropus afer afer	0.698033707	712	Tubulidentata	Mammal
Loxodonta africana	0.697609001	711	Proboscidea	Mammal
Trichechus manatus latirostris	0.705801105	724	Sirenia	Mammal
Monodelphis domestica	0.700987306	709	Didelphimorphia	Mammal
Taeniopygia guttata	0.701828410	711	Passeriformes	Bird
Serinus canaria	0.702247191	712	Passeriformes	Bird
Manacus vitellinus	0.698727016	707	Passeriformes	Bird
Ficedula albicollis	0.712328767	730	Passeriformes	Bird
Melopsittacus undulatus	0.700987306	709	Psittaciformes	Bird
Zonotrichia albicollis	0.700987306	709	Passeriformes	Bird
Falco peregrinus	0.715862069	725	Falconiformes	Bird
Aptenodytes forsteri	0.700564972	708	Sphenisciformes	Bird
Calypte anna	0.707015131	727	Trochiliformes	Bird
Gallus gallus	0.700564971	708	Galliformes	Bird
Anas platyrhynchos	0.702247191	712	Anseriformes	Bird
Python bivittatus	0.681818182	704	Squamata	Reptilia
Anolis carolinensis	0.710124827	721	Squamata	Reptilia
Pelodiscus sinensis	0.699290780	705	Testudines	Reptilia
Chelonia mydas	0.702549575	706	Testudines	Reptilia
Xenopus laevis	0.644475920	706	Anura	Amphibia

Xenopus tropicalis	0.616531165	738	Anura	Amphibia
Babina daunchina	0.656862745	714	Anura	Amphibia
Pachytriton labiatus	0.607142857	728	Caudata	Amphibia

Table S3.2. Ordered Regions For FOXP2 orthologues.

Species	Structured Region		
Homo sapiens	259-280	341-390	484-628
Pan troglodytes	259-280	342-391	485-629
Gorilla gorilla	257-278	339-388	482-626
Pongo abelii	257-288	339-388	468-626
Nomascus leucogenys	258-278	339-388	482-626
Macaca mulatta	258-289	340-389	483-627
Papio anubis	258-279	340-389	483-627
Chloroscebus sabaeus	258-279	340-389	483-618
Saimiri boliviensis boliviensis	258-279	340-389	483-627
Callithrix jacchus	257-278	339-388	482-617
Galeopterus variegatus 87-99	275-296	357-406	500-644
Tupaia chinensis	258-279	340-389	483-627
Mus musculus	258-289	340-389	483-627
Cricetulus griseus	262-282	343-392	486-630
Rattus novergicus	253-273	336-385	479-627
Oryctolagus cuniculus	260-281	342-391	485-629
Ochotona princeps	251-271	332-381	475-610
Octodon degus	263-280	344-393	487-631
Eptesicus fuscus	258-281	341-389	483-618
Myotis brandtii	257-280	340-389	483-618
Pteropus alecto	261-282	344-392	486-630
Ceratotherium simum simum	257-278	339-388	482-627
Felis catus	251-272	333-382	476-620
Panthera tigris	231-252	315-364	458-602
Odobemus rosmarus divergens	253-274	335-384	478-622

Erinaceus europaeus	252-271	335-383			477-621
Physeter catodon	252-274	334-383			477-621
Camelus ferus	254-275	336-385			479-614
Vicugna pacos	254-275	336-385			479-614
Echinops telfairi	260-277	341-390			484-615
Trichechus manatus	268-289	350-399			493-637
Orycteropus afe afer	256-277	339-387			481-637
Elephantulus edwardii	256-277	338-387			480-625
Chrysochloris asiatica	258-278	339-388			482-626
Condylura cristata	258-279	340-389			483-627
Loxodonta africana	255-276	335-386			480-624
Monodelphis domestica	253-272	335-384			478-622
Ficedula albicolis	274-293	358-405			499-643
Taeniopygia guttata	255-274	337-386			480-624
Falco peregrinus	270-288	351-395			494-638
Calypte anna	272-291	354-404			496-640
Aptenodytes forsteri	252-271	334-383			477-621
Zonotrichia albicollis	253-272	335-384			478-622
Manacus vitellinus	250-273	334-381			476-620
Serinus canaria	256-275	338-387			481-625
Melopsittacus undulatus	253-272	335-384			478-622
Gallus gallus	252-271	334-383			477-621
Anas platyrhynchos	256-275	338-387			481-625
Python bivittatus	248-267	330-379			473-624
Anolis carolinensis	265-284	347-391			491-634
Chelonia mydas	250-269	332-380			475-619
Pelodiscus sinensis	249-268	331-380			474-619
Xenopus laevis	250-273	333-381	461-466	475-631	655-668

Table S4.1. Disorder Proportion for FOXP3 orthologues.

Species	Disorder Proportion	AA	Order	Class
Homo sapiens	0.280742459	431	Primates	Mammal
Pan troglodytes	0.276102088	431	Primates	Mammal
Pan paniscus	0.276102088	431	Primates	Mammal
Gorilla gorilla	0.273781903	431	Primates	Mammal
Pongo abellii	0.299303944	431	Primates	Mammal
Pongo pygameus	0.299303944	431	Primates	Mammal
Hylobates lar	0.278422274	431	Primates	Mammal
Nomascus leucogenys	0.276744186	430	Primates	Mammal
Macaca mulatta	0.294663573	431	Primates	Mammal
Papio anubis	0.287703016	431	Primates	Mammal
Chlorocebus sabaeus	0.315545244	431	Primates	Mammal
Saimiri boliviensis boliviensis	0.341067285	431	Primates	Mammal
Callithrix jacchus	0.338747100	431	Primates	Mammal
Galeopterus variegatus	0.274418605	430	Dermoptera	Mammal
Tarsius syrichta	0.274418605	430	Primates	Mammal
Tupaia chinensis	0.362186788	439	Scandentia	Mammal
Mus musculus	0.258741259	429	Rodentia	Mammal
Cricetulus griseus	0.338785047	428	Rodentia	Mammal
Rattus norvegicus	0.254079254	429	Rodentia	Mammal
Octodon degus	0.266203704	432	Rodentia	Mammal
Oryctolagus cuniculus	0.335648148	432	Lagomorpha	Mammal
Ochotona princeps	0.315668203	434	Lagomorpha	Mammal
Physeter catodon	0.313225058	431	Cetacea	Mammal
Orcinus orca	0.313225058	431	Cetacea	Mammal
Camelus ferus	0.269141531	431	Artiodactyla	Mammal

Bos taurus	0.310904872	431	Artiodactyla	Mammal
Equus caballus	0.395348837	430	Perissodactyla	Mammal
Ailuropoda melanoleuca	0.394859813	428	Carnivora	Mammal
Felis catus	0.365740741	432	Carnivora	Mammal
Canis lupus familiaris	0.374418605	430	Carnivora	Mammal
Vicugna pacos	0.271461717	431	Artiodactyla	Mammal
Panthera tigris	0.390697674	430	Carnivora	Mammal
Mustela putorius furo	0.360277136	433	Carnivora	Mammal
Odobenus rosmarus divergens	0.321759259	432	Carnivora	Mammal
Leptonychotes weddellii	0.374133949	433	Carnivora	Mammal
Ceratotherium simum simum	0.329466357	431	Perissodactyla	Mammal
Eptesicus fuscus	0.264501160	431	Chiroptera	Mammal
Myotis brandtii	0.308584687	431	Chiroptera	Mammal
Pteropus alecto	0.271461717	431	Chiroptera	Mammal
Condylura cristata	0.287383178	428	Soricomorpha	Mammal
Chrysochloris asiatica	0.320185615	431	Afrosoricida	Mammal
Erinaceus europaeus	0.294252874	435	Erinaceomorpha	Mammal
Elephantulus edwardii	0.299303944	431	Macroscelidea	Mammal
Echinops telfairi	0.314814815	432	Afrosoricida	Mammal
Sorex araneus	0.331018519	432	Soricomorpha	Mammal
Orycteropus afer afer	0.354988399	431	Tubulidentata	Mammal
Loxodonta africana	0.309133489	427	Proboscidea	Mammal
Trichechus manatus latirostris	0.365967366	429	Sirenia	Mammal
Dasypus novemcinctus	0.338747100	431	Cingulata	Mammal
Ornithorhynchus anatinus	0.198598131	428	Monotremata	Mammal
Monodelphis domestica	0.166666667	462	Didelphimorphia	Mammal

Table S4.2. Ordered regions for FOXP3 orthologues.

Species	Structured Region			
Homo sapiens	89-182	189-270		278-413
Pan troglodytes	89-184	189-270		278-413
Pan paniscus	89-184	189-270		278-413
Gorilla gorilla	89-184	189-270		278-413
Pongo abellii	89-259			281-413
Pongo pygameus	89-259			281-413
Hylobates lar	89-182	189-270		277-413
Nomascus leucogenys	89-184	189-269		276-412
Macaca mulatta	89-182	190-263		278-413
Papio anubis	89-182	190-263		278-413
Chlorocebus sabaeus	89-182	190-259	278-289	300-413
Saimiri boliviensis boliviensis	90-175	192-260	279-290	301-413
Callithrix jacchus	89-174	191-259	278-289	300-413
Galeopterus variegatus	88-108	123-183		188-413
Tarsius syrichta	89-173			190-412
Tupaia chinensis	98-115	138-266		287-421
Sorex araneus	94-108	136-183		190-415
Mus musculus	89-107	111-269		276-415
Rattus norvegicus	92-297			302-415
Cricetulus griseus	91-106	123-257	279-295	301-414
Octodon degus	91-289			300-417
Oryctolagus cuniculus	90-252			302-415
Ochotona princeps	93-122	127-175	193-297	305-417
Physeter catodon	89-182		190-258	278-413
Orcinus orca	89-182		190-258	278-413

Camelus ferus	90-264			277-414
Bos taurus	90-182	189-258		277-413
Ailuropoda melanoleuca	94-109	134-174	193-266	284-410
Felis catus	97-103	128-171	190-270	276-415
Canis lupus familiaris	91-104	131-171	190-270	277-411
Vicugna pacos	90-271			277-414
Panthera tigris	96-104	123-171	190-270	293-413
Mustela putorius furo	100-106	135-174		191-415
Odobenus rosmarus divergens	93-117	131-183		191-414
Leptonychotes weddellii	91-118	130-183	190-260	300-415
Equus caballus	89-106	125-171	191-263	292-413
Ceratotherium simum simum	92-171		192-265	277-414
Eptesicus fuscus	90-183			190-415
Myotis brandtii	93-107	124-184		189-415
Pteropus alecto	89-182			190-415
Condylura cristata	90-171			189-411
Chrysochloris asiatica	97-107	124-270		279-414
Erinaceus europaeus	96-111	116-188	193-274	281-418
Elephantulus edwardii	20-53	127-265	277-290	300-415
Echinops telfairi	92-119	127-172	190-265	278-414
Orycteropus afer afer	97-104	126-174	189-271	277-414
Loxodonta africana	94-170		190-268	274-412
Trichechus manatus latirostris	123-175		190-257	276-412
Dasypus novemcinctus	92-104	133-182		189-413
Monodelphis domestica	24-44			85-450

Table S5.1. Disorder proportion for FOXP4.

Species	Disorder Proportion	AA	Order	Class
Homo sapiens	0.726470588	680	Primates	Mammal
Pan paniscus	0.726470588	680	Primates	Mammal
Gorilla gorilla	0.726470588	680	Primates	Mammal
Pongo abelli	0.726470588	680	Primates	Mammal
Nomascus leucogenys	0.730882353	680	Primates	Mammal
Macaca mulatta	0.735294118	680	Primates	Mammal
Papio anubis	0.735294118	680	Primates	Mammal
Chlorocebus sabaeus	0.735294118	680	Primates	Mammal
Saimiri boliviensis boliviensis	0.733823529	680	Primates	Mammal
Callithrix jacchus	0.713864307	678	Primates	Mammal
Otolemur gernettii	0.728340675	681	Primates	Mammal
Galeopterus variegatus	0.722873900	682	Dermoptera	Mammal
Tupaia chinensis	0.730421687	664	Scandentia	Mammal
Mus musculus	0.745985401	685	Rodentia	Mammal
Rattus norvegicus	0.748538012	684	Rodentia	Mammal
Cricetulus griseus	0.740145985	685	Rodentia	Mammal
Cavia porcellus	0.740203193	689	Rodentia	Mammal
Octodon degus	0.733137830	682	Rodentia	Mammal
Chrysochloris asiatica	0.752186589	686	Afrosoricida	Mammal
Bos taurus	0.745985401	685	Artiodactyla	Mammal
Camelus ferus	0.742358079	687	Artiodactyla	Mammal
Sus scrofa	0.714701601	687	Artiodactyla	Mammal
Vicugna pacos	0.742358079	687	Artiodactyla	Mammal
Ailuropoda melanoleuca	0.698142415	646	Carnivora	Mammal

Canis lupus	0.740088106	681	Carnivora	Mammal
Leptonychotes weddelli	0.719530103	681	Carnivora	Mammal
Mustela putorius	0.741176471	680	Carnivora	Mammal
Odobenus rosmarus	0.725000000	680	Carnivora	Mammal
Orcinus orca	0.733624454	687	Cetacea	Mammal
Physeter catodon	0.736842105	684	Cetacea	Mammal
Erinaceus europaeus	0.750000000	704	Erinaceomorpha	Mammal
Ochotona princeps	0.734306569	685	Lagomorpha	Mammal
Elephantulus edwardii	0.746334311	682	Macroscelidea	Mammal
Pseudopodoces humilis	0.679104478	670	Passeriformes	Mammal
Ceratotherium simum	0.729809104	681	Perissodactyla	Mammal
Equus caballus	0.748538012	684	Perissodactyla	Mammal
Loxodonta africana	0.740524781	686	Proboscidea	Mammal
Trichechus manatus latirostris	0.740524781	686	Sirenia	Mammal
Eptesicus fuscus	0.743024963	681	Chiroptera	Mammal
Myotis brandtii	0.744493392	681	Chiroptera	Mammal
Pteropus alecto	0.746705710	683	Chiroptera	Mammal
Condylura cristata	0.696925329	683	Soricomorpha	Mammal
Sorex araneus	0.751803752	693	Soricomorpha	Mammal
Orycteropus afer afer	0.740145985	685	Tubulidentata	Mammal
Serinus canaria	0.685543964	671	Passeriformes	Bird
Taeniopygia guttata	0.697604790	668	Passeriformes	Bird
Falco peregriunus	0.697014925	670	Falconiformes	Bird
Calypte anna	0.681470138	653	Trochiliformes	Bird
Aptenodytes forsteri	0.674772036	658	Sphenisciformes	Bird
Gallus gallus	0.689759036	664	Galliformes	Bird
Anas platyrhynchos	0.674174174	666	Anseriformes	Bird
Alligator mississipiensis	0.695522388	670	Crocodilia	Reptilia

Alligator sinensis	0.693693694	666	Crocodilia	Reptilia
Python bivittatus	0.678624813	669	Squamata	Reptilia
Anolis carolinensis	0.639344262	671	Squamata	Reptilia
Chelonia mydas	0.693693694	666	Testudines	Reptilia
Chrysemys picta bellii	0.695067265	669	Testudines	Reptilia
Pelodiscus sinensis	0.680303030	660	Testudines	Reptilia
Xenopus tropicalis	0.708527132	645	Anura	Amphibia
Xenopus laevis	0.705148206	641	Anura	Amphibia

Table S5.2. Ordered regions for FOXP4.

Species	Structured Region					
Homo sapiens	66-71	219-228	308-355	391-396	452-542	558-581
Pan paniscus	66-71	219-228	308-355	391-396	452-542	558-581
Gorilla gorilla	66-71	219-228	308-355	391-396	452-542	558-581
Pongo abelli	66-71	219-228	308-355	391-396	452-542	558-581
Nomascus leucogenys	68-71	219-228	308-355	391-396	453-542	558-581
Macaca mulatta		219-228	308-355	391-396	452-542	558-581
Papio anubis		219-228	308-355	391-396	452-542	558-581
Chlorocebus sabaeus		219-228	308-355	391-396	452-542	558-581
Callithrix jacchus		219-228	308-355	388-396	451-541	550-582
Saimiri boliviensis		219-228	308-351	391-396	452-542	558-581
Otolemur garnettii		220-229	309-356	392-397	454-543	559-589
Galeopterus variegatus		220-229	309-354	387-398	454-543	559-588
Tupaia chinensis	67-74	223-230	310-357		438-527	543-566
Mus musculus		224-230	314-361	396-402	458-547	563-585
Rattus norvegicus		223-225	313-360	395-401	456-546	562-582
Cricetulus griseus		224-234	314-361	396-402	458-547	563-583
Octodon degus			312-358	384-400	454-544	560-583
Cavia porcellus			319-363	391-407	461-551	567-589
Ochotona princeps		225-232	309-359	396-400	458-547	563-589
Orcinus orca		224-236	315-362	398-404	459-549	656-586
Physeter catodon		223-232	310-359	395-400	456-546	562-584
Bos taurus		224-233	313-358		458-547	563-586
Vicugna pacos		226-235	315-362	398-403	460-549	565-587
Camelus ferus		226-235	315-362	398-403	460-549	565-587
Sus scrofa		226-235	315-363	388-401	460-549	565-595

Canis lupus			220-229	309-356	392-397	454-543	559-581
Mustela putorius furo			219-228	308-355	391-395	453-542	558-580
Ailuropoda melanoleuca	21-37	85-94		273-321	357-362	422-508	524-547
Odobenus rosmarus			219-228	308-355	391-397	453-542	558-588
Leptonychotes weddelli			221-230	310-357	387-398	454-543	559-589
Eptesicus fuscus			220-229	309-356	392-395	454-543	559-581
Myotis brandtii			220-229	309-356	392-395	454-543	559-580
Pteropus alecto			222-227	311-358	394-399	456-545	561-583
Equus caballus			224-231	311-358	393-397	459-545	561-583
Ceratotherium simum			220-229	303-352	391-398	454-543	559-582
Chrysochloris asiatica			225-233	314-356	397-400	463-548	564-590
Condylura cristata			222-231	308-360	382-410	455-545	561-583
Elephantulus edwardii			222-229	309-352	395-398	457-544	560-586
Erinaceus europaeus			234-243	317-368		477-566	582-605
Sorex araneus			228-237	314-360		466-555	571-594
Orycteropus afer afer			225-232	312-359	395-400	462-547	563-589
Loxodonta africana		71-78		314-361	397-401	459-548	564-586
Trichechus manatus latirostris			226-234	313-360	396-400	462-548	564-587
Taeniopygia guttata		63-66	212-233	302-347		437-532	548-579
Serinus canaria		64-71	214-236	305-350		437-535	551-583
Pseudopodoces humilis		64-71	213-235	303-349		435-534	550-582
Falco peregriunus			213-235	304-349		435-534	550-582
Calypte anna			196-221	287-332		418-517	533-565
Aptenodytes forsteri		51-58	201-223	292-337		423-522	538-570
Gallus gallus			207-229	298-343		430-528	544-576
Anas platyrhynchos		56-59	208-230	299-344			430-569
Alligator mississipiensis			213-235	299-349		439-534	550-582
Alligator sinensis			209-231	295-345		435-530	546-578

Python bivittatus	62-69	212-234	298-348		434-533	549-580
Anolis carolinensis	64-70	214-249	297-350	376-385	436-535	551-582
Chrysemys picta bellii		211-233	301-348		438-535	548-580
Chelonia mydas		210-232	300-347		437-534	547-580
Pelodiscus sinensis		207-229	297-344		434-531	544-586
Xenopus tropicalis			280-331		411-517	530-558
Xenopus laevis			276-327		406-513	526-544

Table S6.1. Whole protein comparison for FOXP1 linear motifs content.

	Linear Motifs	CLV_C14_Caspase3-7	DEG_APCC_DBOX_1	DEG_SCF_FBW7_2	DOC_CKS1_1	DOC_CYCLIN_1	DOC_MAPK_1	DOC_PP2B_2	DOC_USP7_1	DOC_USP7_2	DOC_WW_Pin1_4	LIG_14-3-3_2	LIG_14-3-3_3	LIG_BRCT_BRCA1_1	LIG_CtBP_PxDLS_1	LIG_FHA_1	LIG_FHA_2	LIG_NRBOX	LIG_PCNA_PIPBox_1	LIG_PTAP_UEV_1	LIG_SUMO_SBM_1	LIG_SUMO_SBM_2	LIG_WD40_WDR5_1	MOD_CDK_1	MOD_CK1_1	MOD_CK2_1	MOD_GSK3_1	MOD_NEK2_1	MOD_NEK2_2	MOD_PIKK_1	MOD_PKA_1	MOD_PKA_2	MOD_PLK	MOD_ProDKin_1	MOD_SUMO	TRG_NES_CRM1_1	Total
Mammals	*Homo sapiens*	1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	8	22	10	1	2	2	3	1	15	2	1	140
	Pan troglodytes	1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	8	22	10	1	2	2	3	1	15	2	1	140
	Gorilla gorilla	1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	8	22	10	1	2	2	3	1	15	2	1	140
	Pongo abelii	1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	8	22	10	1	2	2	3	1	15	2	1	140
	Nomascus leucogenys	1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	8	22	10	1	2	2	3	1	15	2	1	140
	Macaca mulatta	1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	8	22	10	1	2	2	3	1	15	2	1	140
	Papio anubis	1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	8	22	10	1	2	2	3	1	15	2	1	140
	Chlorocebus sabaeus	1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	8	22	10	1	2	2	3	1	15	2	1	140
	Saimiri boliviensis	1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	8	22	10	1	2	2	3	1	15	2	1	140
	Callithrix jacchus	1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	8	22	10	1	2	2	3	1	15	2	1	140
	Galeopterus variegatus	1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	9	23	10	1	2	2	3	1	15	2	1	142
	Tupaia chinensis	1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	7	24	10	1	2	2	3	1	15	2	1	141

Mammals

Mus musculus		1	1	7	1	3	1	4		14	1	4	1	2	3	2	1	1	1	2			4	15	8	23	11	1	2	2	4	1	14	2	1	138
Rattus norvegicus		1	1	7	1	3	1	3		13	1	4	1	2	4	3	1	1	1	2			3	15	8	22	10	1	2	2	4	1	13	2	1	134
Cricetulus griseus		1	1	7	1	3	1	3		14	1	4	1	2	4	2	1	1	1	2			4	14	8	21	11	1	2	2	4	1	14	2	1	135
Octodon degus		1	1	7	2	3	1	3		14	1	4	1	2	5	2	1	1	1	2			4	12	8	21	10	1	2	2	3	1	14	2	1	133
Oryctolagus cuniculus		1	1	7	2	3	1	3		15	1	4	1	2	8	2	1	1	1	2		1	4	14	9	23	11	1	2	2	3	1	15	2	1	145
Ochotona princeps		1	1	7	2	3	1	4		16	1	4	1	2	8	2	1	1	1	2		1	4	14	9	24	11	1	2	2	3	1	16	2	1	149
Physeter catodon		1	1	7	2	3	1	3		15	1	4	1	2	5	2	1	1	1	2			4	14	7	25	10	1	2	2	3	1	15	2	1	140
Vicugna pacos		1	1	7	2	3	1	3		16	1	4	1	2	6	2	1	1	1	2		1	4	14	8	22	10	1	3	2	3	1	16	2	1	143
Odobenus rosmarus divergens		1		6	2	3	1	3		14	1	4	1	2	5	1	1	1	1	2	1		3	14	9	23	10	1	2	2	3	1	14	2	1	135
Myotis brandtii	1	1	1	7	2	3	2	4		14	1	4	1	2	3	2	1	1	1	2		1	4	11	7	19	7	1	2	2	3	1	14	2	1	128
Eptesicus fuscus	1	1	1	7	3	3	2	4		14	1	3	1	2	2	2	1	1	1	2		1	4	11	6	19	6	1	2	2	3	1	14	2	1	125
Pteropus alecto		1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	13	9	21	10	1	2	2	3	1	15	2	1	139
Erinaceus europaeus		1	1	7	1	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	14	9	23	10	1	2	2	3	1	15	2	1	141
Condylura cristata		1	1	7	2	3	1	3		15	1	4	1	2	6	2	1	1	1	2		1	4	12	9	23	11	1	2	2	3	1	15	2	1	141
Echinops telfairi		1	1	8	1	3	1	3		14	1	4	1	2	6	2	1	1		2			3	12	7	19	10	1	2	2	3	2	14	2	1	130
Chrysochloris asiatica		2	2	8	2	3	1	4		16	1	4	1	2	6	2	1	1		2		1	4	12	9	23	11		2	2	3	1	16	2	1	145
Elephantulus		1	1	8	2	3	1	3		15	1	4	1	2	4	1	1	1		2		1	4	13	7	20	9	1	5	2	3	1	15	2	1	135

	edwardii																																					
	Orycteropus afer afer	1	1	8	1	3	1	3		15	1	4	1	2	6	2	1	1		2		1	3	12	9	21	11	1	2	2	3	1	15	2	1	137
	Trichechus manatus latirostris	1	1	8	1	3	1	3		16	1	4	1	2	7	1	1	1		2	2	1	4	15	8	23	10	1	2	2	3	1	16	2	1	145
	Loxodonta africana	1	1	7	1	3	1	3		14	1	4	1	2	5	3	1	1		2	2	1	4	14	9	23	10	1	3	2	3	1	14	2	1	141
	Monodelphis domestica	1	1	9	2	3	1	3		18	1	4	1	2	4	2	1	1		2			4	12	6	20	9		2	2	3	1	18	2	1	136
	Serinus canaria	1	1	8	1	3	1	3		16		4	1	2	7	3	1	1	1	2	2		4	12	8	20	10	1	2	2	3	2	16	2	1	141
	Zonotrichia albicollis	1	1	8	1	3	1	3	1	16		4	1	2	7	3	1	1	1	2	2		4	12	7	20	10	1	2	2	3	2	16	2	1	141
	Ficedula albicollis	1	1	8	1	3	1	3		16		4	1	2	7	3	1	1	1	2	2		4	12	8	20	10	1	2	2	3	2	16	2	1	141
	Taeniopygia guttata	1	1	8	2	3	1	3		17		4	1	2	7	3	1	1	1	2	2		4	13	8	21	10	1	2	2	3	2	17	2	1	146
	Manacus vitellinus	1	1	8	2	3	1	3		16		4	1	2	7	3	1	1	1	2	2		4	12	8	21	10	1	2	2	3	2	16	2	1	143
Birds	*Melopsittacus undulatus*	1	1	8	1	3	1	3		16		3	1	2	7	3	1	1	1	2	2		4	12	8	20	10	1	2	2	3	2	16	2	1	140
	Falco peregrinus	1	1	8	1	3	1	3		16		4	1	2	7	3	1	1	1	2	2		4	12	8	20	10	1	2	2	3	2	16	2	1	141
	Aptenodytes forsteri	1	1	8	1	3	1	3		16		4	1	2	8	3	1	1	1	2	2		4	12	8	20	10	1	2	2	3	2	16	2	1	142
	Calypte anna	1	1	8	1	3	1	3		15		4	1	2	7	3	1	1	1	2	2		3	12	7	19	10	1	2	2	3	2	15	2	1	136
	Anas platyrhynchos	1	1	8	2	3	1	3		16		4	1	2	7	3	1	1	1	2	2		4	12	8	21	10	1	3	2	3	2	16	2	1	144

	Gallus gallus		1	1	8	1	3	1	3		16		4	1	2	7	3	1	1	1	2	2		4	12	8	20	10	1	2	2	3	2	16	2	1	141
	Python bivittatus		1	2	7	1	3	1	4		14		3	1	2	4	4	1	1	1	2	1		3	12	7	21	9		2	2	3	2	14	2	1	131
	Anolis carolinensis		1	1	7	1	3	1	3		15	1	4	1	2	6	3	1	1	1	2	2		3	14	9	23	11	1	2	2	3	2	15	2	1	144
	Chelonia mydas	1	1	1	9	1	3	1	3		17		5	1	2	7	3	1	1		2	2		4	12	7	21	10	1	1	1	2	3	17	2	1	143
	Pelodiscus sinensis	1	1	1	9	1	3	1	3		17		4	1	2	7	3	1	1		2	2		4	12	7	21	10	1	1	1	2	3	17	2	1	142
Amphibian	*Xenopus laevis*	1	2		6	1	1		6		13		3	2	2	8	3	1	1	1	2	3		2	10	5	17	9		3	2	3	1	13	2	1	124

Table S6.2. Whole protein comparison for FOXP2 linear motifs content.

	Linear Motifs	CLV_C14_Caspase3-7	DEG_APCC_DBOX_1	DEG_SIAH_1	DOC_CKS1_1	DOC_CYCLIN_1	DOC_MAPK_1	DOC_PP2B_2	DOC_USP7_1	DOC_WW_Pin1_4	LIG_14-3-3_3	LIG_CtBP_PxDLS_1	LIG_FHA_1	LIG_FHA_2	LIG_NRBOX	LIG_PCNA_PIPBox_1	LIG_PTAP_UEV_1	LIG_SUMO_SBM_1	LIG_SUMO_SBM_2	MOD_CK1_1	MOD_CK2_1	MOD_GSK3_1	MOD_NEK2_1	MOD_NEK2_2	MOD_PIKK_1	MOD_PKA_1	MOD_PKA_2	MOD_PLK	MOD_ProDKin_1	MOD_SUMO	TRG_NES_CRM1_1	Total
Mammals	*Homo sapiens*	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	16	7	33	7	1	3	3	6	1	16	1		150
	Pan troglodytes	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Gorilla gorilla	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Pongo abelii	2	1	1	6	4	1	1	3	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		148
	Nomascus leucogenys	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Macaca mulatta	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Papio anubis	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Chlorocebus sabaeus	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Saimiri boliviensis	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Callithrix jacchus	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Galeopterus variegatus	3	1	1	6	3			4	16	1	2	6	2	1		1	2	4	15	7	33	7	1	3	3	6	1	16	1		146
	Tupaia chinensis	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Mus musculus	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Rattus norvegicus	2		1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	4	3	6	1	16	1		149

	Cricetulus griseus	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	32	7	1	3	3	6	1	16	1	148
	Octodon degus	2	1	1	6	4	1	1	4	16		2	5	2	1		1	2	6	16	7	32	7	1	3	3	6	1	16	1	148
	Oryctolagus cuniculus	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1	149
	Ochotona princeps	2	1		6	4	2	1	5	16	1	2	5	2	1	1	1	2	6	14	7	32	7	1	3	3	6	2	16	1	150
	Physeter catodon	2	1	1	6	4	1	1	5	17	1	2	5	2	1		1	2	6	14	7	30	7	1	3	3	6	1	17	1	148
	Camelus ferus	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	4	3	6	1	16	1	150
	Vicugna pacos	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	4	3	6	1	16	1	150
	Ceratotherium simum simum	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	14	7	31	7	1	3	3	6	1	16	1	146
	Odobenus rosmarus divergens	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	16	7	34	7	1	3	3	6	1	16	1	151
	Panthera tigris	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	16	7	34	7	1	3	3	6	1	16	1	151
	Felis catus	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	16	7	34	7	1	3	3	6	1	16	1	151
	Myotis brandtii	3	1	1	6	4	1	1	5	17	1	2	5	3	1		1	3	6	13	7	34	7	1	3	3	6	1	17	1	154
	Eptesicus fuscus	3	1	1	6	4	1	1	5	17	1	2	5	3	1		1	3	6	13	7	34	7	1	3	3	6	1	17	1	154
	Pteropus alecto	2	1	1	6	4	1	1	4	16	1	2	6	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1	150
	Erinaceus europaeus	3	1	1	6	4	1	1	4	16	1	2	6	2	1		1	2	6	14	7	35	7	1	3	3	6	1	16	1	152
	Condylura cristata	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1	149
	Echinops telfairi	1	1	1	6	4	1	1	5	16	1	2	5	2	1		1	2	6	15	7	31	7	1	3	3	6	1	16	1	147
	Chrysochloris asiatica	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	16	7	33	7	1	3	3	6	1	16	1	150

Mammals

	Elephantulus edwardii	2	1	1	6	4	1	1	3	16	1	2	5	2	1		1	2	6	15	7	32	7	1	3	3	6	1	16	1		147
	Orycteropus afer afer	3	1	1	6	4	1	1	4	16	1	2	5	1	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Trichechus manatus latirostris	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Loxodonta africana	2	1	1	6	4	1	1	4	16	1	2	5	2	1		1	2	6	15	7	33	7	1	3	3	6	1	16	1		149
	Monodelphis domestica	2	1	1	6	4	1	1	5	17		2	4	2	1		1	2	4	13	7	32	7	1	3	3	6	1	17	1		145
Birds	*Serinus canaria*	2	1	1	6	4	1	1	5	16		2	5	2	1		1	2	6	13	7	33	7	1	3	3	6	1	16	1		147
	Zonotrichia albicollis	2	1	1	6	4	1	1	5	16		2	5	2	1		1	2	6	13	7	33	7	1	3	3	6	1	16	1		147
	Ficedula albicollis	2	1	1	6	4	1	1	5	16		2	5	2	1		1	2	6	13	6	33	7	1	3	3	6	1	16	1		146
	Taeniopygia guttata	2	1	1	6	4	1	1	5	16		2	5	2	1		1	2	6	14	7	33	7	1	3	3	6	1	16	1		148
	Manacus vitellinus	2	1	1	6	4	1	1	5	16		2	5	2	1		1	3	6	13	7	32	6	1	3	3	6	1	16	1	2	148
	Melopsittacus undulatus	2	1	1	6	4	1	1	5	16		2	5	2	1		1	2	6	13	7	33	7	1	3	3	6	1	16	1		147
	Falco peregrinus	2	1	1	6	4	1	1	5	16		2	5	2	1		1	2	6	13	7	33	7	1	3	3	6	1	16	1	1	148
	Aptenodytes forsteri	2	1	1	6	4	1	1	5	16		2	5	2	1		1	2	6	13	7	33	7	1	3	3	6	1	16	1		147
	Calypte anna	2	2	1	6	4	1	1	5	16		2	5	2	1		1	3	6	13	7	33	7	1	3	3	6	1	16	1	1	150
	Anas platyrhynchos	2	1	1	6	4	1	1	4	16		2	5	2	1		1	2	6	13	7	33	7	1	3	3	6	1	16	1		146
	Gallus gallus	2	1	1	6	4	1	1	4	16		2	5	2	1		1	2	6	13	7	33	7	1	3	3	6	1	16	1		146

Reptilia	*Python bivittatus*	2	1	1	6	4	1	1	4	16		2	5	3	1		1	1	6	13	7	30	6	1	3	3	6	1	16	1		142
	Anolis carolinensis	3	1	1	6	5	1	1	5	16		2	5	2	1		1	2	6	15	7	31	8	1	3	3	6	1	16	1		150
	Chelonia mydas	2	2	1	6	4	1	1	6	16		2	5	2	1		1	2	6	13	7	33	7	1	3	3	6	1	16	1	1	150
	Pelodiscus sinensis	2	1	1	6	4	1	1	6	16		2	5	2	1		1	2	6	13	7	32	6	1	2	3	6	1	16	1		145
Amphibian	*Xenopus laevis*	2	1	1	6	4	1	1	4	15	1	3	8	2	1		1	2	6	15	7	32	8	1	3	3	6	1	15	1		151

Table S6.3. Whole protein comparison for FOXP3 linear motifs content.

Linear Motfis	CLV_C14_Caspase3-7	DEG_SCF_FBW7_1	DOC_CKS1_1	DOC_CYCLIN_1	DOC_MAPK_1	DOC_PP2B_1	DOC_PP2B_2	DOC_USP7_1	DOC_WW_Pin1_4	LIG_14-3-3_2	LIG_14-3-3_3	LIG_BRCT_BRCA1_1	LIG_FHA_1	LIG_FHA_2	LIG_MYND_1	LIG_NRBOX	LIG_PTAP_UEV_1	LIG_RRM_PRI_1	LIG_SUMO_SBM_1	LIG_SUMO_SBM_2	MOD_CDK_1	MOD_CK1_1	MOD_CK2_1	MOD_GSK3_1	MOD_NEK2_1	MOD_NEK2_2	MOD_PIKK_1	MOD_PKA_1	MOD_PKA_2	MOD_PLK	MOD_ProDKin_1	TRG_NES_CRM1_1	TRG_NLS_MonoCore_2	TRG_NLS_MonoExtN_4	Total
Homo sapiens	2	2				1	3	7	6	1	1		2	2	1	1	2			1		6	2	9	3	1	3	2	2	1	6		1	1	69
Homo neanderthalensis	2	2			1		2	7	5	1	1		1	2	1		2					4	2	9	3	1	3	2	2	1	5	1	1	1	62
Pan troglodytes	2					1	3	6	5	1	1		2	2	1	1	1			1		6	2	8	3	1	3	2	2	1	5		1	1	62
Pan paniscus	2					1	3	6	5	1	1		2	2	1	1	1			1		6	2	8	3	1	3	2	2	1	5		1	1	62
Gorilla gorilla	2				1	1	3	6	5	1	1		2	2	1	1				1		6	2	9	3	1	3	2	2	1	5		1	1	63
Pongo abelii	2					1	3	6	5	1	1		3	2	1	1	1			1		7	2	8	3	1	3	2	2	1	5		1	1	64
Pongo pygmaeus	2					1	3	6	5	1	1		3	2	1	1	1			1		7	2	8	3	1	3	2	2	1	5		1	1	64
Nomascus leucogenys	2					1	3	5	5	1	1		3	2	1	1	1			1		6	2	9	3	1	3	2	2	1	5		1	1	63
Hylobates lar	2					1	3	6	5	1	1		3	2	1	1	2			1		6	2	9	3	1	3	2	2	1	5		1	1	65
Macaca mulatta	2			1		1	3	6	6	1	1		2	2	1	1	1		1	2		7	2	8	4	1	3	2	2	1	6	1	1	1	70
Papio anubis	2					1	3	6	5	1	1		2	2	1	1	1			1		7	2	8	3	1	3	2	2	1	5		1	1	63
Chlorocebus sabaeus	2					1	3	6	5	1	1		2	2	1	1	1		1	1		7	2	8	3	1	3	2	2	1	5	1	1	1	65

Saimiri boliviensis	2	1				1	3	6	6	1	1		3	2	1	1	1			1		8	2	11	1	1	3	2	2	2	6		1	1	70
Callithrix jacchus	2	1				1	2	6	6	1	1		3	2	1	1	1			1		7	2	10	2	1	3	2	2	2	6		1	1	68
Tarsius syrichta	2	1				1	2	4	7	1			3	1	1	1	1			3		5	1	10	3	1	3	2	1	1	7		1	1	64
Galeopterus variegatus	2			1		1	3	5	7	1			3	2	1	1			1	1		5	2	7	3	1	3	2	2	1	7		1	1	64
Tupaia chinensis	2		1		1	1	3	5	7	1	1		4	2		1	1			4		14	2	16	4	1	3	2	3	1	7		1	1	89
Mus musculus	2					1	4	3	4		1		2	1	1	1				1	1	3	2	6	4	1	3	2	1		4		1	1	50
Rattus norvegicus	2					1	4	5	5		2		2	2	2	1				1	1	3	3	8	5	2	3	2	1		5		1	1	62
Cricetulus griseus	2					1	4	5	5		1		2	2	1	1				1	1	2	3	5	5	2	3	1	1		5		1	1	55
Octodon degus	2		2	1		1	4	8	9	2	2	1	3	2	1	1	1		2	5		5	3	9	4	1		1	2	1	9		1	1	84
Oryctolagus cuniculus	2					1	3	5	7	1	1		4	2	1	1	1			3		6	3	9	3	2	3	2	4	2	7		1	1	75
Ochotona princeps	2	1				1	3	9	8	1	1		4	2	1	1	1			3	1	10	3	15	2	2	3	2	4	2	8		1	1	92
Physeter catodon	1			1		1	3	6	9	1	2		2	2	1	1			2	2		5	2	10	2	1	2	2	2	1	9		1	1	72
Orcinus orca	1			1		1	3	6	8	1	2		2	2	1	1			2	2		5	2	9	2	1	2	2	2	1	8		1	1	69
Bos taurus	2			1		1	3	5	10	1	1		2	2	1	1			2	2		6	2	13	3	1	2	2	2	1	10		1	1	78
Camelus ferus	2			1		1	3	5	7	1	1		2	3	1	1		1	1	2		5	3	8	2	1	3	2	2	1	7		1	1	68
Vicugna pacos	2			1		1	3	5	7	1	1		2	3	1	1		1	1	2		5	3	8	2	1	3	2	2	1	7		1	1	68
Ceratotherium simum simum	3	3					3	8	9	1	1		2	2	1	1	1		1	1		6	2	15	3	1	2	2	2	2	9		1	1	83

Equus caballus	1				1	3	7	9	1			2	2	1	1	2		1	1		6	3	10	4	1	2	2	2	1	9		1	1	74
Leptonychotes weddellii	1		1		1	3	7	10		1		2	3	1	1	1		1	1	1	4	4	9	2	1	2	2	2		10		1	1	73
Odobenus rosmarus divergens	1		2		1	3	7	11	1	1		2	2	1	1	1	1	1	2	1	5	3	12	3	1	2	2	1	1	11		1	1	82
Mustela putorius furo	1		1		1	3	8	11		1		3	2	1	1	1	1	1	2	1	8	3	16	3	1	2	2	1		11		1	1	88
Ailuropoda melanoleuca	1		1		1	3	9	10	1	1		2	2	1	1			1	1	1	9	3	15	4	1	2	2	1	1	10		1	1	86
Canis lupus familiaris	1		1		1	3	8	9	1			2	2	2	1	1		1	1	1	6	3	13	3	1	2	2	2	1	9		1	1	79
Panthera tigris	1		2		1	3	7	10	1			2	2	1	1	1		1	1	1	6	3	10	3	1	2	2	2	1	10		1	1	77
Felis catus	1		1		1	3	6	10	1			2	1	1	1	1		1	1	1	6	2	12	3	1	2	2	2	1	10		1	1	75
Myotis brandtii	1			1		4	5	6		1		2	2	1	1			2	2		4	2	10	3	1	2	2	2	1	6		1	1	63
Eptesicus fuscus	1			1		3	5	6		2	2	4	3	1	1			2	2		2	3	12	3	1	2	2	1	2	6		1	1	69
Pteropus alecto	2				1	3	5	7	1	1		2	2		1	1		2	3		5	2	10	4	1	3	2	2	1	7		1	1	70
Erinaceus europaeus	1		1	1	1	2	8	7				4	2	1	1				1	1	7	2	12	2	1	3	1	1	1	7		1	1	70
Sorex araneus					1	2	5	9	1	3	2	2	2	1	1			1	1	1	8	3	8	3		3	2	3	1	9		1	1	74
Condylura cristata	1				1	2	4	8	1	1		2	2	1	1	1		1	1	1	5	2	8	3	1	3	2	2	1	8		1	1	65
Dasypus novemcinctus	2				1	3	6	7	1			2	2	1	1			1	1		7	2	6	3	1	3	2	3	1	7		1	1	65
Echinops telfairi		1	1		1	4	6	8		2	1	3	1	2	1			1	1		10	2	13	4	1	2	2	2		8		1	1	79
Chrysochloris	1		1		1	4	4	8	1	1		2	2	2	1			1	1	1	6	3	11	3	1	2	2	2	1	8		1	1	72

asiatica																																			
Elephantulus edwardii	1		1	1			2	7	8	1	1		3	2	1	1			2		1	4	2	10	3	1	2	2	2	1	8	1	1	1	70
Orycteropus afer afer	1	1				1	3	7	8	1	1		2	2	2	1			1	1		6	2	14	3	1	2	2	1	1	8	1	1	1	75
Trichechus manatus latirostris	1			1		1	3	6	8		1		2	2	2				1	1		6	2	9	3	1	3	1	2		8	1	1	1	67
Loxodonta africana	1			1			3	3	5	1	1	1	2	2	1					1		7	2	9	4	1	3	2	2	1	5	1	1	1	61
Monodelphis domestica	1	2	1	4			3	8	6		1	1	8	2		1			1			3	2	14	3		1	1	3	1	6		2	2	77
Ornitorhynchus anatinus	2	2	1	2	1		4	6	12		4	3	4	2		1			1	1		12	5	20	5	1	5	2	3		12			1	112

Table S6.4. Whole protein comparison for FOXP4 linear motifs content.

	Linear Motifs	CLV_C14_Caspase3-7	DEG_APCC_DBOX_1	DEG_ODPH_VHL_1	DEG_SCF_FBW7_1	DEG_SIAH_1	DOC_ANK_TNKS_1	DOC_CKS1_1	DOC_CYCLIN_1	DOC_MAPK_1	DOC_PIKK_1	DOC_PP1_RVXF_1	DOC_PP2B_2	DOC_USP7_1	DOC_USP7_2	DOC_WW_Pin1_4	LIG_14-3-3_2	LIG_14-3-3_3	LIG_BRCT_BRCA1_1	LIG_BRCT_BRCA1_2	LIG_CtBP_PxDLS_1	LIG_FHA_1	LIG_FHA_2	LIG_MAD2	LIG_MYND_1	LIG_NRBOX	LIG_SUMO_SBM_1	LIG_SUMO_SBM_2	MOD_CDK_1	MOD_CK1_1	MOD_CK2_1	MOD_GSK3_1	MOD_NEK2_1	MOD_NEK2_2	MOD_PIKK_1	MOD_PKA_1	MOD_PKA_2	MOD_PKB_1	MOD_PLK	MOD_ProDKin_1	MOD_SUMO	TRG_NES_CRM1_1	TRG_NLS_MonoExtC_3	Total
Mammals	*Homo sapiens*	4	1			1		3	2	1		1	4	9		14	1		2	1			3		1	1	2	2	3	10	8	14	5	3	2	4	6		2	14	2		1	127
	Pan paniscus	4	1			1		3	2	1		1	4	8		14	1		2	1			3		1	1	2	2	3	10	8	16	5	2	2	4	6		1	14	2		1	126
	Gorilla gorilla	4	1			1		3	2	1		1	4	10		14	1		2	1			3		1	1	2	2	3	11	8	14	5	3	2	4	6		2	14	2		1	129
	Pongo abelii	4	1		2	1		4	2	1		1	3	9		15	1		2	1			3		1	1	2	2	3	10	8	15	5	3	2	4	6		2	15	2		1	132
	Nomascus leucogenys	4	1			1		3	2	1		1	4	9		14	1		2	1			3		1	1	2	2	3	10	8	15	5	3	2	4	6		2	14	2		1	128
	Macaca mulatta	4	1		2	1		4	2	1		1	3	8		15	1		2	1			3		1	1	2	2	3	11	7	16	5	3	2	4	6		2	15	2		1	132
	Papio anubis	4	1		2	1		4	2	1		1	3	8		15	1		2	1			3		1	1	2	2	3	11	7	16	5	3	2	4	6		2	15	2		1	132
	Chlorocebus sabaeus	4	1		2	1		4	2	1		1	3	7		15	1		2	1			3		1	1	2	2	3	11	7	16	5	3	2	4	6		2	15	2		1	131
	Saimiri boliviensis	4	1			1		3	2	1		1	4	10		14	1		2	1			3		1	1	2	2	3	12	7	15	5	3	3	4	5		2	14	2		1	130
	Callithrix jacchus	4	1			1		3	2	1		1	3	9		1	1		1				3		1	1	2	2	3	1	7	1	6	3	3	4	6		2	1	2		1	127

Mammals

											3														2		6							3				
Otolemur garnettii	4	1	1		3	2	1	1	5	8	14	1		2	1		1	3	2	1	2	2	3	12	7	12	5	2	2	4	6	2	14	2		1	127
Galeopterus variegatus	4	1	1		3	2	1	1	4	9	14	1		2	1			3	2	1	2	2	3	16	7	16	5	2	3	4	6	2	14	2	1	1	136
Tupaia chinensis	2	1	1		3	2	1	1	5	5	15	1		1				2	1	1	2	2	3	9	7	11	5	3	2	4	6	2	15	2		1	116
Mus musculus	1	1	1		3	2	1	1	5	7	13	1		1	1			1	1	1	2	2	4	12	8	13	4	2	5	3	5	1	13	1		1	117
Rattus norvegicus	2	1	1		3	2	1	1	4	6	15	1		1	1			1	1	1	2	2	4	11	8	14	3	2	5	3	5	1	15	1		1	119
Cricetulus griseus	2	1	1		3	3	2	1	4	6	13	1		2	1		1	1	1	1	2	2	4	12	7	15	4	3	4	4	7	2	13	1		1	125
Octodon degus	1	1	1		4	2	2	1	3	6	15	1	1	1			1	1	1	1	2	5	4	9	7	15	6	3	4	4	6	2	15	2		1	128
Cavia porcellus	1	1			3	2	1	1	3	8	14	1	1	1				1	1	1	2	2	3	12	7	13	5	3	4	4	6	2	14	2		1	120
Ochotona princeps	3	1	1	1	3	2	1	1	4	7	15	1	1	2	1			2	2	1	2	2	3	13	7	14	5	3	3	4	6	2	15	2		1	131
Physeter catodon	2	1			4	2	1	1	4	9	15	1		2	1			1	2	1	2	2	4	12	7	12	5	3	3	3	5	2	15	2		1	125
Orcinus orca	2	1	1		4	2	1	1	5	9	16	1		2	1			1	2	1	2	2	4	12	7	11	5	3	2	4	6	2	16	2		1	129
Bos taurus	2	1	1		4	2	1	1	3	7	15	1		2	1			1	1	1	1	2	4	13	7	13	5	3	3	4	7	2	15	2		1	126
Sus scrofa	2	1	1	1	4	2	1	1	4	7	15	1	1	2	1	1		1	2	1	2	2	4	12	7	11	5	3	2	4	7	2	15	2		1	128
Camelus ferus	2	1	1		4	2	1	1	4	8	1	1		2	1		1	1	2	1	2	5	4	1	7	1	4	3	2	4	6	2	1	2		1	132

											5												4		3							5			
Mammals	*Vicugna pacos*	2	1	1		4	2	1	1	4	8	15	1		2	1	1	1	2	1	2	5	4	14	7	13	4	3	2	4	6	2	15	2	1	132
	Ceratotherium simum simum	2	1	1	1	4	2	1	1	4	8	15	1		2	1		1	1	1	2	2	4	13	7	13	5	2	3	4	7	2	15	2	1	129
	Equus caballus	3	1	1		4	2	1	1	4	9	15	2		2	1		2	2	1	2	2	4	12	7	14	5	2	2	4	6	2	15	2	1	131
	Leptonychotes weddellii	3	1	1		4	2	1	1	3	7	15	1		1			2	2	1	2	1	4	14	8	12	5	2	3	4	7	2	15	2	1	127
	Odobenus rosmarus divergens	3	1	1		4	2	1	1	4	8	15	1		2	1		2	2	1	2	1	4	15	8	13	5	2	3	4	7	2	15	2	1	133
	Mustela putorius furo	2	1	1		4	2	1	1	4	8	15	1		2	1		1	2	1	2	2	4	15	7	11	5	2	1	4	7	2	15	2	1	127
	Ailuropoda melanoleuca	2	1	1		3	2	1	1	3	8	14	1		2	1		1	2	1	2	2	4	12	7	13	5	2	4	4	6	2	14	2	1	124
	Canis lupus familiaris	2	1	1		4	3	1	1	3	7	15	1		2	1		1	2	1	2	2	4	14	8	12	5	2	4	4	6	2	15	2	1	129
	Myotis brandtii	2	1	1		4	2	1	1	4	7	15	1		2	1		1	2	1	2	2	4	13	6	13	5	3	3	4	6	2	15	2	1	127
	Eptesicus fuscus	2	1	1		4	2	1	1	4	6	15	1		2	1		1	2	1	2	2	4	13	6	13	5	3	3	4	6	2	15	2	1	126
	Pteropus alecto	2	1	1		4	2	1	1	4	7	15	1		2	1		1	2	1	2	2	4	11	6	11	5	3	2	4	7	2	15	2	1	123
	Erinaceus europaeus	2	1	1		4	2	2	1	3	10	14	1	1			2	2	1	1	3	6	3	17	7	21	5	2	4	3	7	2	14	2	1	145
	Sorex araneus	2	1	1		4	2	1	1	4	11	15	1					1	1	1	1	1	2	12	7	12	5	3	2	3	5	2	15	2	1	119

	Condylura cristata	2	1		1	4	2	1		1	4	7		14	1		1				1	2	1	2	2	4	11	7	10	6	2	3	4	7		1	14	2		1	119
	Chrysochloris asiatica	2	1		1	4	2	1		1	4	5	1	16	1		2	1		2	1	3	1	3	4	4	11	8	13	6	3	1	4	6	1	2	16	2		1	134
	Elephantulus edwardii	2	1		1	4	2	1		1	4	6		16	1		1	1		1	2	2	1	2	4	4	8	7	10	5	3	2	4	6		2	16	2		1	123
	Orycteropus afer afer	2	1		1	4	2	1		1	5	7		15	1		2	1		2	1	2	1	2	4	4	11	7	13	5	3	3	4	6	1	2	15	2		1	132
	Trichechus manatus latirostris	2	1		1	4	2	1		1	4	8		16	1		2	1		2	1	2	1	2	4	4	12	7	12	5	3	3	4	7	1	2	16	2		1	135
	Loxodonta africana	3	1		1	4	2	1		1	4	9		16	1		2	1		2	2	2	1	2	4	4	12	7	20	5	3	4	4	5		2	16	2		1	144
Birds	*Serinus canaria*	1	1			5	2	1	2	1	1	2		16	1	1	3		1	3	3	1	1	2		3	11	8	26	9	2	3	3	5		3	16	2		1	140
	Pseudopodoces humilis	1	1			5	2	1	2	1	1	2		16	1	1	3		1	2	3	1	1	2		3	11	8	25	10	2	2	3	5		3	16	2		1	138
	Taeniopygia guttata	1	1			5	2	1	2	1	1	3		15	1	1	3		1	1	3	1	1	2	1	3	11	8	22	8	2	2	4	6		3	15	2	1	1	135
	Falco peregrinus	1	1			5	2	1	2	1	1	2		16	1	1	3		1	3	4	1	1	2		3	12	9	25	9	2	4	3	5		3	16	2		1	143
	Aptenodytes forsteri	1	1			5	2	1	2	1	1	3		16	1	1	3		1	2	4	1	1	2		3	12	9	25	9	2	3	3	5		3	16	2		1	142
	Calypte anna	1	1		1	5	2	1	2	1	1	3		16	1	2	3		1	2	4	1	1	2		3	12	9	23	8	2	3	3	6		3	16	2		1	142
	Anas platyrhynchos	1	1	1		5	4	1	2	1	2	3		15	1	1	4		1	2	3	2	1	2		3	12	7	24	8	2	2	3	5		2	15	2	1	1	140
	Gallus gallus	1	1			5	2	1	2	1	1	2		15	1	1	3		1	3	3	1	1	2		3	12	9	26	8	2	4	3	5		3	15	2		1	140

Reptilia	*Alligator mississippiensis*	1	1	5	2	1	2	1	1	1	15	1	1	3	1	3	4		1	1	2	3	11	9	24	9	2	4	3	5	2	15	2	1	137
	Alligator sinensis	1	1	5	2	1	2	1	1	1	15	1	1	3	1	3	4		1	1	2	3	11	9	24	9	2	4	3	5	2	15	2	1	137
	Python bivittatus	1	1	4	2	1	2	1	1	1	15	1	1	3	1	3	5		1	1	2	2	12	9	25	9	1	4	3	5	2	15	2	1	137
	Anolis carolinensis	1	1	3	2	1	2	1	1	1	15	1		2	1	3	4		1	1	2	2	12	8	23	9	1	3	3	5	1	15	2	1	128
	Chrysemys picta bellii	1	1	4	2	1	2	1	1	1	14		1	2	1	3	4		1	1	2	3	13	9	25	9	1	4	2	4	2	14	2	1	132
	Chelonia mydas	1	1	4	2	1	2	1	1	1	14		1	2	1	3	4		1	1	2	3	12	9	25	10	1	4	2	4	2	14	2	1	132
	Pelodiscus sinensis	1	1	4	2	1		1	1	1	14		1	2	1	4	3		1	1	2	3	12	8	23	10	1	4	2	4	2	14	2	1	127
Amphibian	*Xenopus laevis*	1	1	4	4	1		1		5	15	2			1	5	4	1	1	1	2	2	15	9	23	6	1	5	2	4	1	15	1	1	134
	Xenopus tropicalis	3	1	5	3	1		1	1	7	16	1			1	5	5	1	1	1	2	2	16	8	26	7	2	5	2	4	2	16	1	1	147

Table S7. Estimated parameters under different codon substitution models for forkhead P sub-family genes[1].

	Model	*dN/dS*	Estimated parameters	Sites with ω>1[2]	ℓ	*P value*
FOXP1	M1a: neutral	0.0731	p_0 = 0.95618, (p_1 = 0.04382) (ω_0 = 0.03066), (ω_1 = 1.00000)		−13857.331728	1
	M2a: selection	0.0731	p_0 = 0.95618, p_1 = 0.04382, (p_2 = 0.00000) (ω_0 = 0.03066), (ω_1 = 1.00000), ω_2 = 11.32611		−13857.331728	
	M8a: β & ω=1	0.0553	p_0 = 0.99268, (p1 = 0.00732), p = 0.15813 q = 2.81242, ω = 1.00000		−13799.608758	1
	M8: β & ω	0.0553	p_0 = 0.99268, (p1 = 0.00732), p = 0.15813, q = 2.81243, ω = 1.00000		−13799.608758	
FOXP2	M1a: neutral	0.0615	p_0 = 0.95735, (p_1 = 0.04265) (ω_0 = 0.01965), (ω_1 = 1.00000)		−12558.167089	0.9295
	M2a: selection	0.0628	p_0 = 0.95601, p_1 = 0.01529, (p_2 = 0.02870) (ω_0 = 0.01965), (ω_1 = 1.00000), ω_2 = 1.00000		−12558.093967	
	M8a: β & ω=1	0.0458	p_0 = 0.99999, (p_1 = 0.00001), p = 0.08897 q = 1.61951, ω = 1.00000		−12524.222980	1
	M8: β & ω	0.0455	p_0 = 0.99999, (p1 = 0.00001), p = 0.09061, q = 1.66060, ω = 1.00001		−12524.233000	

[1]p_0 = proportion of sites where ω < 1, p_1 = proportion of sites where ω = 1, and p_2 = proportion of sites where ω > 1 (selection models only), ω_0 < 1 (negative selection), ω_1 = 1 (neutral selection), and ω_2 > 1 (positive selection), likelihood ratio tests were performed between neutral models (M1a - Nearly Neutral, and M7 - Beta) and models that identify positive selection (M2a - Selection, and M8 – Beta + Selection), the comparisons between M1 vs. M2 and M7 vs. M8 have 2 degrees of freedom. Parentheses indicate fixed parameters.

[2]Position and protein IUPAC code.

Table S7. Estimated parameters under different codon substitution models for forkhead P sub-family genes[1] (continued).

	Model	*dN/dS*	Estimated parameters	Sites with ω>1[2]	*ℓ*	*P value*
FOXP3	M1a: neutral	0.1834	p_0 = 0.89272, (p_1 = 0.10728) (ω_0 = 0.08525), (ω_1 = 1.00000)		−12510.124897	1
	M2a: selection	0.1834	p_0 = 0.89272, p_1 = 0.04720, (p_2 = 0.06008) (ω_0 = 0.08525), (ω_1 = 1.00000), ω_2 = 1.00000		−12510.124897	
	M8a: β & ω=1	0.1518	p_0 = 0.95683, (p_1 = 0.04317), p = 0.48347 q = 3.64447, ω = 1.00000		−12424.107542	0.9137
	M8: β & ω	0.1532	p_0 = 0.96152, (p_1 = 0.03848), p = 0.47274, q = 3.45516, ω = 1.06357	183S, 184A, 194V, and 301V	−12424.017242	
FOXP4	M1a: neutral	0.0487	p_0 = 0.97844, (p_1 = 0.02156) (ω_0 = 0.02778), (ω_1 = 1.00000)		−18552.897747	1
	M2a: selection	0.0487	p_0 = 0.97843, p_1 = 0.02157, (p_2 = 0.00000) (ω_0 = 0.02778), (ω_1 = 1.00000), ω_2 = 6.24385		−18552.897760	
	M8a: β & ω=1	0.0384	P_0 = 0.99789, (p_1 = 0.00211), p = 0.23183 q = 5.59727, ω = 1.00000		-18335.456738	1
	M8: β & ω	0.0392	p_0 = 0.99999, (p_1 = 0.00001), p = 0.22328 q = 4.99479, ω = 2.82598		−18340.025059	

[1] p_0 = proportion of sites where ω < 1, p_1 = proportion of sites where ω = 1, and p_2 = proportion of sites where ω > 1 (selection models only), ω_0 < 1 (negative selection), ω_1 = 1 (neutral selection), and ω_2 > 1 (positive selection), likelihood ratio tests were performed between neutral models (M1a - Nearly Neutral, and M7 - Beta) and models that identify positive selection (M2a - Selection, and M8 – Beta + Selection), the comparisons between M1 vs. M2 and M7 vs. M8 have 2 degrees of freedom. Parentheses indicate fixed parameters.

[2] Position and protein IUPAC code.

Table S8. Branch site model for *FOXP* genes.

		Model Clade D (free-ratio model, admit selection)				**M1a (neutral)**		**Probability**
		Proportion	Clade 1[b]	Clade 2[f]	LogL	Proportion	LogL	M1a vs. Clade
FOXP1[2]	Mammals[f] vs. non-mammals[b]	P_0 = 0.70045 P_1 = 0.24561 P_2 = 0.05394	ω_0 = 0.00176 ω_1 = 0.13116 ω_2 = 0.68705	ω_0 = 0.00176 ω_1 = 0.13116 ω_2 = 0.60693	−15531.277393	P_0 = 0.93230 P_1 = 0.06770 ω_0 = 0.03010	−15594.616874	1
						ω_1 = 1.00000		
FOXP4	Mammals[f] vs. non-mammals[b]	P_0 = 0.70046 P_1 = 0.25887 P_2 = 0.04067	ω_0 = 0.00627 ω_1 = 0.11160 ω_2 = 0.18012	ω_0 = 0.00627 ω_1 = 0.11160 ω_2 = 0.66102	−22189.482510	P_0 = 0.95450 P_1 = 0.04550 ω_0 = 0.03293	−22463.434676	<0.001
						ω_1 = 1.00000		

Table S9. Eukaryotic linear motifs and their functions.

ELM	Evidence Class	Logical	Reability	Code	Probability	ELM Description	Functional Description
CLV_C14_Caspase3-7	Experimental	TP	Likely	a	0.0030937	The amino acids around the caspase-3 and -7 cleavage site are named N- to C-terminal: P4, P3, P3, P2, P1, P-1. The scissile bond between P1 and P-1 is cleaved by caspase-3 and -7, whereas positions P4 to P-1 are important for substrate specificity and recognition. P1 is always an aspartate (D), while P-1 is usually a small amino acid. Proline (P) as secondary alpha-amino acid is not accepted at P-1. An in vitro kinetic study argues for small amino acids, phenylalanine (F) or tyrosine and no ionic amino acids at P-1 (Stennicke,2000). The regular expression allows small amino acids at P-1. Other residues are still described but data was not valid enough to create an additional regular expression. The backbone of amino acids at P2 and P3 is stabilised by hydrogen (H) bonds allowing caspase-3 and -7 a broad spectrum of amino acids at these positions. At P2 non-polar amino acids (valine (V), leucine (L), P) are preferred because of possible interactions with a hydrophobic pocket. Threonine (T) is also very common. At P3 glutamate (E) is preferred because of an additional H-bond. However other amino acids like serine (S) or L are still common. In the regular expression P3 is not specified, except for the prohibition of P, because caspase-3 and -7 accept a variety of amino acids at P3. D is strongly preferred at P4 due to strong H-bond interactions, followed by S, T, and E. Crystal structures with pentapeptides argue for a preference for hydrophobic residues at P5 because of hydrophobic interactions with two F residues in case of caspase-3. This site is missing in caspase-7 (Fu,2008). The regular expression does not include P5 because caspase-3 cleaves also substrates with non-hydrophobic residues at P5. Nevertheless a hydrophobic residue at P5 is a hint that the protein is rather a caspase-3 substrate then a caspase-7 one. Based on the observed variations at P4-P2, the regular expression will on the one hand produce false positives and on the other hand not match all described cleavage sites.	The proteases caspases-3 and -7 play an important role in programmed cell death (apoptosis). Cleavage of the caspase substrates results in characteristic morphological features of apoptotic cell death, including membrane blebbing, pyknotic nuclei, cell rounding, and formation of apoptotic vesicles. Caspases recognise their substrates by a cleavage motif. The amino acids of the substrate around the caspase cleavage site are named N- to C-terminal: P4, P3, P2, P1, P-1. The scissile bond between the essential aspartate at P1 and P-1, usually a small amino acid, is cleaved by caspase-3 and -7, whereas positions P4 to P-1 are important for substrate specificity and recognition.
DEG_SCF_FBW7_1	Experimental	TP	Certain	b	0.0007138	FBW7 (also called FBXW7, hCdc4 or hSel10) is a member of a family of F-box proteins that binds via WD40 beta propeller to its substrates after their phospho-degron motifs (also named CPDs, i.e. Cdc4 phospho-degrons) have been doubly phosphorylated (Hao,2007, Welcker,2008). The core of the motif is TPxxS, preceded by a variable number of hydrophobic residues. The motif is used in cell cycle regulation: the widely conserved G2 phase-specific cyclin E destruction by FBW7 was first described in yeast. The Thr is often phosphorylated by GSK3, after priming at the other P-site, linking the FBW7 activity with the mitogenic signalling pathway. In some instances the Thr may alternatively be targeted by other kinases such as CDKs. Interestingly, many of the known FBW7 substrates are proto-oncogenes with key roles in the regulation of cell division, differentiation and growth. However, a proposed FBW7 phosphodegron in the key cell state monitor mTor (Mao,2008) does not match the diphosphorylated motif. The phosphodegron in v-Jun is mutated and inactivated, enhancing oncogenicity by preventing its destruction. Some variant motifs substitute a Glu residue for the second phosphosite, e.g. in SV40 large T, and this variant is represented by the alternative pattern in ELM.	Several phosphodegrons are required for cell state-dependent recognition of regulatory proteins by SCF complexes via repeat domains of associated F box proteins (FBPs) and their subsequent ubiquitin-mediated degradation. The SCF-FBW7 and the SCF-betaTrCP1 motifs, contain two phosphorylated residues, which are recognised via a WD40 domain. For example, the SCF-FBW7 degron TPxxS is found in cyclin E, which is required for the G1/S transition. The SCF-betaTrCP1 degron DSGxxS operates in a broader range of cell regulation. For example, NF-kappa-B inhibitors are phosphorylated and destroyed under immune stimulation while beta-catenin is degraded in the absence of Wnt signalling. Skp2, another FBP, recognises cell cycle regulators via its leucine-rich repeat. In case of the single-phosphorylated DEG_SCF_SKP2-CKS1_1 motif, Skp2 requires additional binding of Cks1 for recognition. So far, only a few cell cycle inhibitors, including p27Kip1 that is mainly involved in G1 arrest, have been found to carry this degron.
DOC_PP2B_2	Experimental	TP	Certain	c	0.0022964	This PP2B-docking motif is defined by four amino acids. There is an absolute requirement for leucine in the first position of the motif and for proline in the last position. However, the viral protein A238L (O36972), which has been shown to prevent recognition of LxvP-containing substrates by calcineurin, contains a lysine residue instead of a proline in the last position (4F0Z) (Grigoriu,2013). The second position can accommodate a variety of amino acids. The third position generally contains a valine, but other hydrophobic residues have also been observed. Some motif instances are immediately preceded by an aromatic residue, which can further strengthen the interaction. Upon binding of Ca2+, the CNA subunit undergoes a conformational change, exposing the hydrophobic motif-binding pocket that is located at the interface of the CNA and CNB subunits. Therefore, LxvP sites can only interact with activated calcineurin. The hydrophobic pocket includes two CNA residues (W352, F356) and three CNB residues (L115, M118, V119), which mediate binding to substrates containing an LxvP motif. Immunosuppressants have been shown to bind to the hydrophobic pocket in a similar way. Upon binding to the hydrophobic pocket, the motif adopts a conformation in which it is almost parallel to the α-helix of CNA binding to CNB. The proline in the motif is predicted to interact with the aromatic residues in CNA.	Calcineurin (PP2B) is a Ca2+- and calmodulin-regulated serine/threonine protein phosphatase known to affect cell biological function mainly in yeast and mammalian systems but is found in all Eukaryotes. It regulates a number of different pathways, including activation of the NFAT family of transcription factors, regulation of axonal guidance, the Ca2+-dependent migration of neutrophils, synaptic plasticity, and apoptosis. The effectiveness of dephosphorylation depends not only on the proximal position of calcineurin-binding site to the dephosphorylation site but also on the flexibility of the calcineurin active site. Calcineurin has been shown to interact with two distinct linear motifs: one docking motif (DOC_PP2B_1) that directly interacts with the catalytic subunit A (CNA) of calcineurin, and a second docking motif (DOC_PP2B_2) that binds to both CNA and calcineurin B (CNB), the regulatory Ca2+-binding subunit of calcineurin.
DOC_USP7_1	Experimental	TP	Certain	d	0.0123885	Targeting motif found in USP7 substrates, docking to the NTD domain. The USP7 N-terminal domain is a TRAF-like domain but with different sequence specificity to the classical TRAF domain. The motif identified in p53 and MDM2 recognizes the same surface groove in USP7. But MDM2s make more extensive contacts than p53, leading to stronger affinity. The motif identified in these proteins can therefore be categorized as either a lower affinity motif or high affinity motif, depending on the extent of contact. The general pattern of the motif is a simple P..S and will have frequent matches in cellular proteins but it is not clear how numerous are the USP7 substrates. CAUTION: This motif has poor discrimination and is likely to be poorly predictive.	USP7, also known as HAUSP, is a deubiquitinating enzyme which cleaves ubiquitin moieties from its substrates. The USP7-mediated deubiquitination of P53, Mdm2 and USP7 inhibition by the herpesviral proteins EBNA1 and ICP0 shows its importance in the regulation of cell survival pathways and controlling key cellular processes important for viral infection. The N-terminal TRAF-like domain of USP7 is responsible for substrate recognition and nuclear localization while the catalytic core domain is required for the deubiquitinating activity.

DOC_WW_Pin1_4	Experimental	TP	Certain	e	0.0154332	Proline-directed phosphorylation is a major cellular signalling mechanism. Proline directed kinases include Map kinases and cyclin-dependent kinases. Many p(S/T)P sites act as binding motifs for recruiting proteins or placing enzymes close to their substrates. Proline isomerisation may confound the function of these sites but also provides a mechanism for their regulation. This requires a peptidyl-prolyl cis/trans isomerase (PPIase). The phospho-specific PPIase is Pin1 (for Protein interacting with NIMA 1). It belongs to the parvulin subfamily of PPIases. Pin1 homologs are highly conserved in eukaryotes. Pin1 regulates diverse cellular process, including growth-signal response, cell-cycle progression, cellular stress response, neuronal function and immune response. Pin1 is an 18 kDa protein with two domains: the N-terminal WW domain and the C-terminal PPIase domain connected by a flexible linker. The WW domain acts as the binding module to bind its substrate p(S/T)P site and bring the catalytic PPIase domain close to its substrate where it catalyses cis-trans conversions. The Pin1 WW domain interaction with substrates occurs through a conserved phosphate-binding pocket, present in Pin1 and it homologues. The pSer/Thr-Pro peptides all bind to the WW domain in the trans conformation. Only WW domains which have an arginine in loop1 will be able to bind the sequence with core motif p(S/T)P. The phosphorylated residue is recognized by the Arginine as well as by backbone interactions. The Pro residue enters a pocket bounded by the second Tryptophan.	WW Domains are small but abundant domains found in diverse regulatory situations. The binding peptide motifs appear always to involve proline residues. Specific motifs vary for different WW domains and in some cases must be phosphorylated on a serine or threonine.
Lig_14-3-3_2	Experimental	TP	Certain	f	0.0015569	The longer mode 2 interacting phospho-motif for 14-3-3 proteins with key conservation RxxxS#p where # is a conserved hydrophobic position and p is a semiconserved Pro. Pro is excluded -1 and +1 of the pSer due to tight backbone interactions. Pro is usually depicted as required at +2 but (unlike mode 1) it is neither strongly conserved nor making tight contact to 14-3-3. Instead the +1 residue is always a conserved hydrophobic with an extensive packing face. Other residue preferences in the x positions are likely to affect binding affinity and some combinations may be disfavoured. The typical kinases phosphorylating mode 2 peptides are likely to have a positive charge preference preceding the phosphorylated residue.	14-3-3 proteins interact with specific phosphoserine or phosphothreonine containing motifs.
Lig_14-3-3_3	Experimental	TP	Certain	g	0.004948	Consensus derived from reported natural interactors which do not match the Mode 1 and Mode 2 ligands. Key conserved residues are missing. While the sequence range of 14-3-3 binders is certainly not fully defined, a pattern derived from outliers as here may be poorly predictive and matches should be treated with CAUTION. Validation is paramount.	14-3-3 proteins interact with specific phosphoserine or phosphothreonine containing motifs.
LIG_BRCT_BRCA1_1	Experimental	TP	Certain	h	0.0019115	The LIG_BRCT_BRCA1_1 motif binds with low affinity to the BRCT domain of BRCA1. The motif has the consensus sequence S..F and these residues are specially recognized by the binding pocket in the BRCT domains. The high affinity motif has an additional bound lysine residue (S..F.K).	BRCT domains are protein modules mainly found in Eukaryota. BRCT domains are present in proteins that are associated with DNA damage response. They recognize and bind specific phosphorylated serine (pS) sequences. This phospho-protein mediated interaction of the BRCT domain has a central role in cell-cycle check point and DNA repair functions.
LIG_CtBP_PxDLS_1	Experimental	TP	Likely	i	0.0001173	The PxDLS motif pattern is based on the conservation of reported sequence instances together with the structure of the CtBP domain in complex with a PxDLS peptide (1HL3). Beta-augmentation at the sheet edge places the peptide sidechains in specific places on the CtBP surface. Pro at position 1 makes an H-bond to strand edge backbone and fits in a hydrophobic pocket that will not accommodate other residues (Gly with no side chain may be the least disruptive). Position 2 contributes to beta augmentation so that the semi-conserved sidechain is placed in a shallow hydrophobic pocket, which also allows Glu due to proximity of surface positive charge. Position 3 is most often Asp, probably due to favourable charged residue proximity but is surface accessible and accepts some changes. Position 4 contributes to beta augmentation so that the sidechain enters a deep hydrophobic groove that fits to Leu and would probably allow Met but reject most other residues. Position 5 has a Ser-Thr preference but appears to accept R (as in HDACs) and some other mostly small residues. Following the core peptide there are clear preferences for Lys or Arg but these are not a strict requirement. However the conserved GLDLSKK motif in Hic1 is reported to bind CtBP but lacks Pro: Therefore for Gly at position 1 which must weaken the interaction, the motif in ELM requires C-terminal positive charge compensation.	The PxDLS motif is present in a number of nuclear proteins, including certain transcription factors and HDACs, that recruit CtBP (C-terminal binding protein) into nuclear complexes. CtBP has a generally repressive effect on transcription and must be removed to active CtBP regulated genes.
LIG_FHA_1	Experimental	TP	Certain	j	0.0086622	LIG_FHA_1 motifs are short phosphothreonine modules binding FHA domains with large aliphatic amino acids at the pT+3 position. The motif has the consensus sequence of T..[IVL]. Proteins with FHA domains having this preference include the checkpoint kinase chk2 (Li,2002) and DNA repair protein rad9 (Byeon,2001).	The FHA domain is a signal transduction module which recognizes phosphothreonine containing peptides on the ligand proteins. FHA domains partake in many signalling processes but are especially prevalent in nuclear proteins that are involved in cell cycle checkpoint, DNA repair and transcriptional regulation.
LIG_FHA_2	Experimental	TP	Certain	k	0.0082864	LIG_FHA_2 motifs are short phosphothreonine peptide modules contains acidic amino acids at the pT+3 position. The motif has the consensus sequence of T..[ED]. FHA domains with this preference are found in checkpoint/repair proteins MRC1 and Rad9 of fungi and metazoan Xrcc1 (Luo,2004) and Xrcc4 (Koch,2004).	The FHA domain is a signal transduction module which recognizes phosphothreonine containing peptides on the ligand proteins. FHA domains partake in many signalling processes but are especially prevalent in nuclear proteins that are involved in cell cycle checkpoint, DNA repair and transcriptional regulation.

LIG_MYND_1	Experimental	TP	Certain	l	0.0006499	The PxLxP motif is recognized by a subset of MYND domain containing proteins. The target specificity is mainly due to the highly charged surface of the MYND domain with a positive face consisting of C-terminal residues on one side, and a negatively charged region on the other side. The positive charge of these side chains could mediate long-range electrostatic interactions with the negative charges that may flank the PxLxP motif. So the interaction might require a longer region, while the PxLxP motif constitutes the core motif.	The MYND domain is a zinc binding domain that is involved in protein-protein interactions mainly in the context of transcriptional regulation. It is named after Myeloid, Nervy, and DEAF-1, which are the three most characterized proteins that contain the MYND domain. Only a small number of MYND domain containing proteins have been identified and they are involved in various biological processes such as cell proliferation, apoptosis, adhesion, migration, and tumorigenesis and oxygen homeostasis. MYND domain typically binds a proline-rich motif in their interacting partners, however they have different binding specificities.
LIG_SUMO_SBM_2	Experimental	TP	Likely	m	0.018383	Minimal region common to all known SBM instances and required for SBM function is the hydrophobic patch consisting of 3 hydrophobic residues (I, V, or L) or 4 residues consisting of 3 hydrophobic residues (I,V, or L) and a variable residue at 2nd or 3rd positions. Contrary to LIG_SUMO_SBM_1, in the inverted SBM, the hydrophobic core is preceded by a variable length stretch of phosphorylatable residues (mainly Serine) and acidic residues. Both the phoshorylatable residues (when phosphorylated) and the acidic residues increase the affinity of binding to SUMO due to interactions with basic residues of SUMO on SBM interaction interface. Inverted SBM forms a beta sheet pairing with the second beta sheet of SUMO in anti-parallel orientation (eg RanBP2 (P49792)) (Reverter,2005).	Non-covalent binding to SUMO proteins is mediated via SUMO-binding motif (SBM). Sumo-interacting proteins predominantly function in the nucleus. SBM is essential for a variety of cellular processes including transcriptional regulation, sub-nuclear localization, nuclear body assembly, and anti-viral response. Viral proteins are also known to utilize such processes via their SBMs upon host cell invasion.
LIG_WD40_WDR5_1	Experimental	TP	Certain	n	0.0001462	The conserved WDR5-binding motif in the unstructured acidic C-terminal tail region of RbBP5 interacts on the edge of the beta-propeller between blades 5-6 on the opposite face of the WDR5 protein where MLL1 binds with the Win motif (LIG_WD40_WDR5_WIN_1, LIG_WD40_WDR5_WIN_2, LIG_WD40_WDR5_WIN_3). The motif is negatively charged due to multiple acidic residues, while the RbBP5-binding site is positively charged, allowing accommodation of the peptide. The acidic residues include one aspartate between the two conserved hydrophobic positions, and up to four consecutive aspartate and/or glutamate residues (generally a minimum of two acidic residues randomly distributed at these four positions is observed) N-terminal to the first conserved hydrophobic position. After binding there is a negatively charged complex. The RbBP5-binding site has hydrophobic features allowing two conserved valine residues to bind into two pockets. The first pocket binds Val-375 in human RbBP5, and is made up of Tyr-228 and Leu-240 and the aliphatic part of the Lys-250 side chain. The second pocket is formed by Phe-266, Val-268 and Leu-288 and interacts with Val-377. Polar interactions between the WDR5 residues Asn-225 and Gln-289 and the main chain help stabilize the binding of the valine residues into the hydrophobic pockets. Not all acidic residues contributing to the negativity of the peptide interact with the WDR5 surface as they are orientated away from the protein, however they may be of importance due to their negativity (Odho,2010) (2XL2). The motif is highly conserved among eukaryotes, especially amongst vertebrates and mammals, however among fungi some variances can be observed, which are captured in a separate motif specific for fungal species (LIG_WD40_WDR5_2).	In the nuclei of eukaryotic cells, DNA is complexed with histones into nucleosomes. Post-translational modification of histones regulates their interactions with DNA and other nuclear proteins, and is important for the control of cellular processes such as gene transcription, cell cycle progression and DNA repair. One important modification is the methylation of H3 histones at lysine 4, which is located in the intrinsically disordered N-terminal H3 region. Methylation of H3K4 is catalyzed by members of the Set1/MLL protein family, whose activity depends on their assembly in a multi-protein histone methylation complex. The WD40 repeat domain protein WDR5 plays a key role in H3K4 methylation by acting as a scaffold protein for the assembly of the core histone methylation complex, which is conserved through evolution. The recruitment of different complex subunits by WDR5 depends on distinct motifs in WDR5-binding partners, including the catalytic Set1/MLL subunits and the accessory protein RbBP5.
MOD_CK1_1	Experimental	TP	Certain	o	0.0170407	CK1 Phosphorylation site	Motif recognised by CK1 for Ser/Thr phosphorylation
MOD_CK2_1	Experimental	TP	Certain	p	0.0145681	The main determinant of CK2 phosphorylation specificity is a negative charge 3 positions after the modification residue.	Motif recognised by CK2 for Ser/Thr phosphorylation
MOD_GSK3_1	Experimental	TP	Certain	q	0.0267866	The GSK3 phosphorylation recognition site is based primarily on vertebrate data. GSK3-family kinases are widespread in Eukaryota but in most cases the specificity has not been confirmed.	Site recognised by GSK3 for Ser/Thr phosphorylation.
MOD_NEK2_1	Experimental	TP	Certain	r	0.0097983	The optimal substrate motif of NEK2 was determined by positional scanning oriented peptide library screening (PS-OPLS) (Alexander,2011). The substrate motif targeted by NEK2 for phosphorylation shows the strongest amino acid selectivity in the -3 and +2 positions (relative to the Ser/Thr residue that is phosphorylated by NEK2). Having a preferred residue in the -3 position might compensate for the occurrence of less favorable residues in the +1 and +2 positions and vice versa. To encode this information, two variants of the motif have been defined, based on the selectivity for particular residues observed in the PS-OPLS experiment (Alexander,2011). The first variant of the NEK2 substrate phosphorylation motif contains Phe, Leu or Met in the -3 position. The occurrence of one of these residues, which are highly preferred to other hydrophobic residues, might compensate for less favored amino acids in the +1 and +2 positions. Hydrophilic amino acids as well as Ile and Val are strongly disfavored in the -3 position. In the -2 position, all amino acids except Pro are tolerated, although there is a slight preference for basic and hydrophobic residues. Similarly, in the -1 position there is no strong selectivity except for a strong discrimination against Pro. Also in the +1 position Pro is disfavored, together with the acidic Glu and Asp residues. The strong selection against Pro in this position allows discrimination against Cdk1 mitotic kinase phosphorylation sites. In the +2 position, the acidic residues Glu and Asp are not allowed, while a Pro residue is again accepted in this position. Although there seems to be a preference for particular hydrophobic residues in the +1 position and for specific amino acids in the +2 position, the occurrence of less favored residues in these two positions is allowed in this variant of the motif, as this might be compensated by the presence of the strongly selected Phe, Leu or Met in the -3 position.	The NEK protein kinases are Never in mitosis A (NimA)-related kinases that belong to the NEK Serine/Threonine protein kinase family. NEKs have been identified in many Eukaryotes, where they play a critical role in cell cycle control. The closest mammalian NimA homologue NEK2 is a core component of the human centrosome and its activity and expression peak in S and G2 phase, during which it interacts with and phosphorylates several centrosomal proteins. NEK2 has many cell cycle-related functions, including cell cycle progression, spindle pole formation, microtubule anchoring, centriolar cohesion, cilia formation and chromatin condensation.
MOD_PIKK_1	Experimental	TP	Certain	s	0.0092301	The PIKK family member proteins specifically phosphorylate the (ST)Q motif in their substrates. The glutamine adjacent to the target serine-threonine is critical for the substrate recognition.	The phosphoinositide-3-OH-kinase related kinases (PIKKs) are atypical protein kinases exclusive to eukaryotes. The PIKK members are large proteins with Ser/Thr kinase activity serving important roles in DNA repair and DNA damage checkpoints. The three PIKK proteins with repair and checkpoint functions in mammalian cells are: DNA-PK (DNA-dependent protein kinase), ATM (ataxia telangiectasia mutated), and ATR (ATM and Rad3 related).

MOD_PKA_1	Experimental	TP	Certain	t	0.0023151	The optimal phosphorylation site preference for PKA is RR.(ST) - Lysine in the basic sites may weaken the prefence. This site is also targeted by other basophilic kinases of the AGC group, including PKG and PKC isoforms. AGC group kinases do not tolerate Pro at position +1. Lys at the Arg positions may be allowed in some weaker sites. Some specificity determinants may be present at the less conserved non-basic sites.	Motifs phosphorylated by a subset of AGC group kinases including PKA that all have similar sequence specificity.
MOD_PKA_2	Experimental	TP	Certain	u	0.0094575	Secondary preference for PKA-type AGC kinase phosphorylation with a single Arg at p-2. This motif is probably more often targeted by other basophilic kinases of the AGC group, including PAK1 and PKC isoforms: These kinases actually show a stronger preference at p-2 than PKA, which has the strongest basophilic preference at p-3. AGC group kinases do not tolerate Pro at position +1. It is likely that some specificity determinants distinguishing among these kinases may be present at the less conserved non-basic sites.	Motifs phosphorylated by a subset of AGC group kinases including PKA that all have similar sequence specificity.
MOD_PLK	Experimental	TP	Certain	v	0.0060146	Site phosphorylated by the Polo-like kinase.	Site recognised and phosphorylated by the Polo-like kinase.
MOD_ProDKin_1	Experimental	TP	Certain	w	0.0154332	(S/T)P site at which Pro-directed kinases such as P38 MAP Kinase phosphorylate a Ser/Thr residue in various signal transduction pathways. These kinases require a docking motif to be activated, so the site is not meaningful by itself. MAP Kinase homologues are found widely in eukaryotes, including protozoa, but experimental data is from higher eukaryotes, mainly metazoa.	Site at which MAP Kinase phosphorylates substrates.
DOC_ANK_TNKS_1	Experimental	TP	Certain	x	0.0003538	Tankyrase1 and Tankyrase 2 are closely related poly (ADP-ribose) polymerases. They have both enzymatic and scaffolding activities. Tankyrases are multidomain proteins. The C-terminal PARP domain catalyzes the poly-(ADP-ribosyl)ation of substrate proteins, whereas the ankyrin repeat domain interacts with the substrates and brings them to PARP domain. There are five ankyrin triple repeat clusters (ARCs) present in Tankyrase and, except ARC3, all are predicted to act as independent binding unit. The ARCs interact with substrates through their tankyrase binding motifs. All ARCs having similar substrate recognition abilities and the overall binding mode of substrates to an ARC of Tankyrase is highly conserved, with the binding peptide lying in a surface groove. Each ARC is able to bind the same set of proteins with different binding affinity. So they are redundant in function. The Tankyrase binding motif is usually depicted as RxxPDG. The arginine and glycine residues at positions 1 and 6 are highly conserved acting as critical anchor residues of the motif. The R enters the deepest part of the groove making charged interactions. The G is sandwiched in the narrowest part of the groove. The P and D positions are more variable. The motif is experimentally verified in nearly 17 proteins including 3BP2, AXIN1, TRF1, IRAP and NUMA1.	Tankyrases belong to the poly (ADP-ribose) polymerase (PARP) family of proteins, which function by catalyzing the covalent linkage of ADP-ribose polymers onto target proteins and thereby regulating their ubiquitylation, stability and function. The human genome encodes two similar tankyrases. TNKS and TNKS2 both recruit a variety of substrates with a broad range of functions. Each shows some variation in their conservation and tissue expression. Moreover both Tankyrases contain N-terminal ankyrin repeats. These ankyrin repeats form five domains known as ankyrin repeat clusters (ARCs) and each can serve as a discrete binding site for its binding partners. Tankyrase plays important roles in many biological functions including telomere length regulation, insulin signalling and centrosome function. Dysregulation of 3BP2 recognition by tankyrase causes a human disease knows as cherubism. Tankyrase is also considered to be a potential candidate for a telomere-directed anticancer target.
MOD_PKB_1	Experimental	TP	Certain	y	0.0006034	Phosphorylation site preference R.R.(ST) targeted by some basophilic kinases of the AGC group, including PKB and p70S6K. AGC group kinases do not tolerate Pro at position +1. Lys at the Arg positions may be allowed in some weaker sites. Some specificity determinants may occur at the less conserved non-basic sites: e.g Phe at position +2 is reported to be an optimal PKB site but a poor one for p60S6K.	Motif phosphorylated by a subset of AGC group kinases including PKB that all have similar sequence specificity.
LIG_MAD2	Experimental	TP	Certain	z	0.0001011	Mitotic spindle checkpoint protein MAD2 binding motif	Mitotic spindle checkpoint protein MAD2 binding motif

CLV: Cleavage Sites; DEG: Degradation sites; DOCK: Docking sites; Ligand binding sites; MOD: post-translational modification sites;TP: True Positive.

Capítulo 4

Conclusão

A discussão específica e completa referente aos resultados obtidos neste trabalho encontram-se no manuscrito a ser submetido em revista científica de circulação internacional (ver capítulo 3). Sendo assim, neste capítulo somente uma conclusão geral referente ao tema será apresentada. Mesmo assim, alguma sobreposição é inevitável com o texto apresentado no manuscrito.

Apesar de uma porção significativa da sociedade acadêmica voltar-se para estudos dos homeodomínios proteicos, bem como outras regiões estruturadas, as regiões desordenadas das proteínas (**IDRs**) estão revelando-se importantes para a regulação e funcionalidade das mesmas (Liu *et al.,* 2008). Além disso, as **IDRs** são mais prevalentes em fatores de transcrição, do que em outros tipos de proteínas, apresentando também uma correlação positiva com a complexidade dos mesmos (Dunker, *et al.*, 2000; Oldfield *et al.*, 2005; Liu *et al*. 2006; Liu, Matthews e Bondos, 2008; Disfani, *et al*. 2012; Huang, e Sarai 2012; Marín *et al.,* 2013; Xue *et al.*, 2010; 2013), fatos que nos instigaram a investigar seu papel ao longo da evolução da importante subfamília de fatores de transcrição FOXP.

Dentre os parálogos aqui estudados, interessantemente, as proteínas FOXP1, FOXP2 e FOXP4, que apresentam função e padrão de expressão relativamente semelhante, tem similar nível de desordem (~70%). Interessantemente, tal nível elevado de desordem está sendo mantido, provavelmente há 800 milhões de anos, considerado o ortólogo FOXP de *Apis mellifera*, usado aqui como representante da proteína ancestral. Assim, é possível especular que desde a separação dos invertebrados dos vertebrados, passando pelos vários eventos de duplicação e especiação, ao menos três genes FOXP seguiram um caminho evolutivo, conservando através da seleção natural, seu nível plasticidade. Entretanto, o conteúdo de aminoácidos e motivos lineares destas regiões desordenadas é distinto, particularmente entre mamíferos e não mamíferos. Por outro lado, em FOXP3 o nível de desordem é mais baixo (~30%), possivelmente associado com suas funcionalidades mais específicas.

Como já comentado, há uma correlação positiva entre valores altos de desordem e o número de motivos lineares e a complexidade dos organismos (Dunker, *et al*. 2005; Disfani, *et al*.2012). Curiosamente, mamíferos apresentaram maiores valores de desordem para FOXP1 e FOXP4 quando comparando com não mamíferos ($p<0.001$), o que nos permite sugerir que mamíferos apresentariam interações mais complexas para a rede que envolve FOXP1 e FOXP4.

Diferenças no tempo e local de expressão dos genes FOXP1, FOXP2 e FOXP4 entre pássaros e mamíferos já foram apontadas em estudos anteriores (Mendoza, *et al*. 2014).

Como esperado, a grande parte dos genes da família FOXP estão evoluindo sob um modelo majoritário de seleção purificadora. Considerando altos valores de ω, a maioria dos sítios se sobrepõe com regiões altamente desordenadas. O quadro, no entanto é distinto para a proteína FOXP3, no qual apresentou seleção positiva, provavelmente no ramo dos primatas. Curiosamente, o único, dentre os cinco sítios, em que há sobreposição de seleção positiva e região desordenada, a probabilidade de seleção é na ordem de 99%. Estes achados, de uma forma geral, sugerem que a habilidade de evoluir (evolvabilidade) está ocorrendo, ao menos em parte, dentro das regiões desordenadas, similar a outros sistemas genéticos, como já descrito por Xue *et al,* (2013) e Montanari *et al*, (2011).

Nossos achados mostram, também, que diversas mudanças nos ortólogos aqui estudados têm implicações em adaptações taxa-específicas, visto que são provavelmente funcionais.

Por exemplo, macacos do Novo Mundo e Tarsiidae vivem em florestas tropicais, e apresentam quatro motivos lineares idênticos em FOXP3. Esses organismos teriam ganhado estes motivos devido à presença de uma Treonina na posição 194 alinhada do FOXP3. Este sítio está sobre seleção positiva, com probabilidade 99%. Ao menos um destes motivos está associado com resposta imune, indicando uma provável convergência adaptativa, que pode estar relacionada a similares pressões seletivas em ambientes de floresta tropical.

Redes regulatórias complexas podem surgir e desaparecer através de pequenas mudanças nos aminoácidos. Entretanto, alguma forma de seleção estabilizadora é existente, visto que sítios importantes para interação podem ser perdidos em uma posição, porém são ganhos em outras (Ba e Moses, 2010; Ba, 2014). Nossos resultados, como um todo, estão indicando que a estabilidade da família FOXP ao longo de um período evolutivo, está claramente conduzida pela proporção de desordem, mas não quanto ao seu conteúdo de aminoácidos e motivos lineares, e que muitas mudanças descritas podem ser associadas com adaptações taxa-específicas.

Em resumo, a sequência e a arquitetura do domínio Forkhead de ligação ao DNA, bem como outros domínios das proteínas FOXPs, estão sendo relativamente conservados ao longo da evolução, o que provavelmente garante a estabilidade do sistema e sua funcionalidade básica. No entanto, além de funções similares e básicas, FOXP1, FOXP2, FOXP3 e FOXP4 também atuam sobre funções específicas. Aqui demostramos que pelo menos parte do

caminho molecular envolvido com novas funções pode ter se dado através da presença de motivos lineares específicos.

Referências Bibliográficas

Andersen, K.G. *et al.*, Comparative Genomics Reveals Key Gain-of-Function Events in *FOXP3* During Regulatory T Cell Evolution (2012). Frontiers in Immunology, 3:e113.

Ba, A.N.N. (2014) Molecular Evolution of Posttranslational Regulation in Intrinsically Disordered Regions. Graduate Department of Cell and Systems Biology. Tese de PhD. University of Toronto.

Benayoun, B. *et al.,* (2011) Forkhead Transcription factors: Key Players in health and disease. Trends in genetics, 27: 224-232.

Brown, P. J. e Banham, A.H. (2009) *Mus musculus* forkhead box P1. Transcription Factor Encyclopedia. http://www.cisreg.ca/cgi-bin/tfe/articles.pl?tfid=480.

Coffer, P.J. e Burgering, B.M.T. (2004) Forkhead-Box Transcriptionfactors and Their Role In The Immune System. Nature reviews Immunology.4:889-899.

Dinkel, H. *et al.* (2013) The Eukaryotic Linear Motif Resource ELM: 10 Years and Counting. Nucleic Acids Research. 42:D259-D266.

Disfani, F.M. *et al.* (2012) MoRFpred, a Computational Tool for Sequence-Based and Characterization of Short Disorder-to-Order Transitioning Binding Regions in Proteins. Bioinformatics. 28: i75-i83.

Dunker, A.K. *et al.* (2000) Intrinsic Protein Disorder in Complete Genomes. Genome Inform Ser Workshop Genome Inform, 11:161-171.

Santos, E. M. *et al*. (2011) Alternative Splicing and Gene Duplication in the Evolution of the FoxP Gene Subfamily. Molecular Biology and Evolution, 28: 237-247.

Enard, W. *et al*. (2002) Molecular Evolution of FOXP2, a Gene Involved in Speech and Language. Nature, 418:869-872.

Enard, W (2011) *FOXP2* and the Role of Cortico-basal Ganglia Circuits in Speech and Language Evolution.Current Opinion in Neurobiology, 21:415-424.

Fisher, S.E. e Scharff, C. (2009) *FOXP2* as a Molecular Window into Speech and Language. Trends in Genetics, 25: 166-177.

Fitch, W. M. (1970) Distinguishing Homologous from Analogous Proteins. Systematic Zoology, 19:99-113.

Fritzenwanker, J.H. *et al* (2014) The Fox/Forkhead Transcription Factor Family of the Hemichordate *Saccoglossus kowalevskii.* EvoDevo. 5:17.

Gould, C. M, Diella F., Via, A. *et al.* (2010). ELM: The Status of the 2010 Eukaryotic Linear Motif resource. Nucleic Acids Research, 38:D167-180.

Graybiel, A. M. (2005) The Basal Ganglia: Learning New Tricks and Loving it. Current Opinion in Neurobiology, 15: 638-644.

Huang, H. e Sarai, A. (2012) Analysis of the Relationships Between Evolvability, Thermodynamics, and the Functions of Intrinsically Disordered Proteins/Regions. Computational Biology and Chemistry, 41: 51-57.

Krause, J. *et al* (2007) The derived *FOXP2* variant of modern humans was shared with Neandertals. Current Biology, 17: 1908-1912.

Lam, E.W.F *et al.,* (2013) Forkhead box Proteins: Tuning Forks for Transcriptional Harmny. Nature, 13:482-495.

Li, G. *et al.* (2007) Accelerated FoxP2 Evolution in Echolocating Bats. Plos One. 2(9): e900.

Li, S. *et al.,* (2004) Transcriptional and DNA Binding Activity of The *FOXP1*/2/4 Family is Modulated Heterotypic and Homotypic Interactions. Molecular and Cellular Biology, 24:809-822.

Liu, J *et al*. (2006) Intrinsic Disorder in Transcription Factors. Biochemistry, 45: 6873-6888.

Liu, Y. *et al.,* (2008). Multiple Intrinsically Disordered Sequences Alter DNA Binding by the Homeodomain of the *Drosophila* Hox Protein Ultrabithorax. Journal of Biological Chemistry, 283: 20874-20887.

Marín, M., *et al.,* (2013) Intrinsic Disorder in Pathogen Effectors: Protein Flexibility as an Evolutionary Hallmark in a Molecular Arms Race. The Plant Cell, 25: 3153–3157.

Mendoza, E. *et al.,* (2014) Differential Co-expression of FoxP1, FoxP2 and FoxP4 in the Zebra Finch (*Taeniopygia guttata*) Song System. Journal of Comparative Neurology. *In press.*doi: 10.1002/cne.23731.

Oldfield, C.J. *et al.,* (2005) Coupled Folding and Binding with Alpha-Helix-Forming Molecular Recognition Elements. Biochemistry, 44:12454-12470.

Paixão-Côrtes, V.R. *et al.,* (2012) The Cognitive Ability of Extinct Hominins: Bringing Down the Hierarchy Using Genomic Evidences. American Journal of Human Biology, 25: 702-705

Reich, D. *et al*., (2010) Genetic History of an Archaic Hominin Group From Denisova Cave in Siberia. Nature, 468:1053-1060.

Schön, C. *et al.,* (2006) The FoxP Subclass in *Xenopus laevis* development. Development genes and Evolution, 216:641-646

Schreiweis, C. *et al*. (2014) Humanized Foxp2 Accelerates Learning by Enhancing Transitions from Declarative to Procedural Performance. PNAS, 111: 14253-14258.

Shu, W. *et al*. (2007) Foxp2 and Foxp1 Cooperatively Regulate Lung and Esophagus Development, 134:1991-2000.

Song, X. *et al*., (2013) Rapid Diversification of *FOXP2*in Teleosts Through Gene Duplication in the Teleost-Specific Whole Genome Duplication Event. Plos one, 8: e83858.

Spiteri, E. *et al.,* (2007) Identification of the Transcriptional Targets of *FOXP2,* a Gene Linked to Speech and Language, in Developing Human Brain. American Journal of Human Genetics, 81: 1144–1157.

Teramitsu, I. e White, S.A. (2008) Motor Learning: The *FOXP2* Puzzle Piece. Current Biology, 18:335-337.

Thackray, V.G. (2014) Fox Tales: Regulation of Gonadotropin Gene Expression by Forkhead Transcription Factors. Molecular and Cellular Endocrinology, 385:62-70.

Tompa, P. *et al.,* (2006) Prevalent Structural Disorder in E.coli and S.cerevisiae Proteomes. J Proteome Research, 5: 1996-2000.

Van der Lee, R. *et al*., (2014). Classification of Intrinsically Disordered Regions and Proteins. Chemical Reviews, 114: 6589-6631.

Vernes, S.C. *et al*., (2011). *FOXP2* Regulates Gene Networks Implicated in Neurite Outgrowth in the Developing Brain. Plos genetics, 7: e1992145.

Ward, J.J. *et al.* (2004) Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. Journal of Molecular Biology, 337:635-645.

Watson, J. D. *et al.* (2006) Biologia Molecular do Gene, Porto Alegre, 5ª ed. Artmed.

Xue, B *et al.* (2010) Archaic chaos: intrinsically disordered proteins in Archaea. BMC Systems Biology, 4:1-21.

Xue, B *et al.* (2013) Intrinsically Disordered Regions of p53 Family are Highly Diversified in Evolution. Biochimica et Biophysica Acta, 1834:725-738.