UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA
DISSERTAÇÃO DE MESTRADO

Lucas Carrijo de Oliveira

Efeitos da atribuição de pesos a sequências sobre as frequências de aminoácidos em alinhamentos múltiplos de sequências – aplicação em análises de conservação e correlação entre resíduos

BELO HORIZONTE
2016

Lucas Carrijo de Oliveira

Efeitos da atribuição de pesos a sequências sobre as frequências de aminoácidos em alinhamentos múltiplos de sequências – aplicação em análises de conservação e correlação entre resíduos

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do título de Mestre em Bioinformática

Orientador: Prof. Dr. Lucas Bleicher

Belo Horizonte
2016

para meu sobrinho João Gabriel, uma cópia melhorada de mim, certamente a criança mais esperta e carinhosa que eu já conheci.

AGRADECIMENTOS

À minha família, que sempre apoiou e fomentou minhas decisões, e cujo amor e confiança depositados em mim moldaram minha personalidade. Eu não poderia ter recebido educação mais exemplar, baseada no afeto e diálogo, com lições de humanidade e valores independentes de credo ou doutrina.

Ao meu orientador, Lucas Bleicher, exímio professor e pesquisador, cujo estilo de orientação baseado na humildade, abertura ao diálogo e emancipação certamente será perpetuado por seus discípulos.

Aos meus amigos, antigos e novos, que tornam mais alegre e segura essa nova fase da minha vida e sempre estiveram do meu lado nos melhores e piores momentos. Com certeza têm me ensinado mais sobre a vida do que qualquer escola poderia ensinar.

Aos meus colegas de trabalho que, com seu carisma e cooperatividade, tornam bem mais fluidas e agradáveis as árduas tarefas da vida acadêmica.

“Somos uma maneira de o cosmos conhecer a si mesmo.”

– Carl Sagan

RESUMO

Analisando um alinhamento múltiplo de sequências ao nível de resíduos, além das posições conservadas existem outros padrões indicativos de importância funcional que refletem divergência funcional dentro de uma família em decorrência de duplicações gênicas. Em famílias de proteínas homólogas que apresentam subfamílias com especificidades funcionais distintas, algumas posições podem apresentar-se conservadas apenas em uma subfamília particular, ou o aminoácido conservado pode ser diferente para cada subfamília. Isso sugere que seu papel funcional desse resíduo relaciona-se não com com a função global da família, mas sim com especificidades funcionais daquele grupo. Nesses casos, é razoável que tais especificidades não sejam determinadas pela presença de um único resíduo, mas sim por um grupo de resíduos, e esse grupo irá emergir de análises de correlação entre resíduos desde que um número suficiente de proteínas apresentem as mesmas especificidades. Entretanto, algumas famílias de proteínas apresentam subfamílias pouco representadas em número de sequências nos alinhamentos. Ao mesmo tempo, estes costumam vir repletos de sequências redundantes, muitas vezes mutantes ou variantes da mesma sequência, oriundas principalmente de organismos modelo. Essa redundância nos alinhamentos acaba por enviesar análises com caráter estatístico, como são os métodos de correlação. Nesse sentido, o presente trabalho tem por objetivo comparar os efeitos de abordagens distintas que visam a diminuição da redundância em alinhamentos múltiplos de sequências: a atribuição de pesos a sequências e os filtros por identidade máxima. Além disso, o presente trabalho também propõe abordagens para tornar os cálculos de correlação compatíveis com o a atribuição de pesos de sequências, a fim de aperfeiçoar análises de conservação e correlação entre resíduos. A atribuição de pesos a sequências foi capaz de destacar as frequências de aminoácidos específicos de subfamílias pouco amostradas, ao mesmo tempo em que diminuía as frequências de aminoácidos presentes em sequências redundantes. Os cálculos de correlação adaptados ao uso de pesos foram capazes de detectar essas diferenças, oferecendo uma boa alternativa para análises de correlação em alinhamentos pouco representativos da diversidade de proteínas de fato existente na natureza.

ABSTRACT

Analysing a multiple sequece alignment at the residue level, apart from the conserved positions, there are other patterns that are also indicative of functional importance and reflect functional divergence within a homologous protein family due to gene duplication. In families that have subfamilies with distinct functional specificities, some positions can be conserved only in a particular subfamily, or the conserved amino acid can be different for each of the subfamilies. This suggests that the role of this residue relates not to the global function of the family, but to functional specificities of that group. In these cases, it is reasonable that such specificities are not determined by the presence of a single residue, but by a group of residues, and this group will emerge from residue correlation analysis since a sufficient amount of proteins show the same specificities. However, some protein families have subfamilies less represented in terms of amount of sequences in the alignments. Meantime, this alignments use to come full of redundant sequences, many times mutants or variants of the same sequence, originary mainly from model organisms. This redundancy in the alignments tend to introduce bias to analysis with a statistical mean like the correlation methods. In this way, the present work has as objective to compare the effects of distinct approaches aiming the decreasing of redundancy in multiple sequence alignments: sequence weighting and filtering by maximum identity. Besides, this work also proposes approaches to make the correlation calculations compatible with sequence weighting, in order to improve analisys of residue conservation and correlation. Sequence weighting was capable of highlighting frequencies of amino acids specific of less sampled subfamilies, while decreasing the frequencies of amino acids present in redundant sequences. The adapted calculations were capable of detecting such differences, providing a good alternative to conservation and correlation analisys in alignments that are less representative of the actual protein diversity existent in nature.

LISTA DE ABREVIATURAS

A: código de uma letra para o aminoácido alanina (vide contexto)

A: código de uma letra para o nucleotídeo de adenina (vide contexto)

Ala: código de três letras para o aminoácido alanina

AMS: alinhamento múltiplo de sequências

Asn: código de três letras para o aminoácido asparagina

Asp: código de três letras para o aminoácido aspartato

C: código de uma letra para o aminoácido cisteína (vide contexto)

C: código de uma letra para o nucleotídeo de citosina (vide contexto)

C4: denominação para um tipo específico de motivo "dedo de zinco" cujo zinco é coordenado por quatro cisteínas

CAPS: *co-evolution analysis using protein sequences*

CASP: *Critical Assessment of Techniques for Protein Structure Prediction*

Coq7p: um dos componentes do complexo multimérico envolvido na síntese de coenzima-Q

Cys: código de três letras para o aminoácido cisteína

D: código de uma letra para o aminoácido aspartato

DBD: *DNA binding domain*

DCA: *direct coupling analysis*

DNA: *deoxyribonucleic acid*

E: código de uma letra para o aminoácido glutamato

F: código de uma letra para o aminoácido fenilalanina

Fe: sigla para o elemento ferro

G: código de uma letra para o aminoácido glicina (vide contexto)

G: código de uma letra para o nucleotídeo de guanina (vide contexto)

GatCAB: uma amido transferase responsável pela transamidação do glutamil-$tRNA^{Gln}$ em glutaminil-$tRNA^{Gln}$

Glu: código de três letras para o aminoácido glutamato

Gly: código de três letras para o aminoácido glicina

H: código de uma letra para o amino histidina

Her2p: uma das subunidades do complexo GatCAB

His: código de três letras para o aminoácido histidina

HIUase: 5-hidroxi-isourato hidrolase

HMM: *hidden Markov model*

HRE: *hormone response element*

I: código de uma letra para o aminoácido isoleucina

Ile: código de três letras para o aminoácido isoleucina

K: código de uma letra para o aminoácido lisina

L: código de uma letra para o aminoácido leucina

LBD: *ligand binding domain*

Leu: código de três letras para o aminoácido leucina

Lys: código de três letras para o aminoácido lisina

M: código de uma letra para o aminoácido metionina

McBASC: *McLachlan-based substitution correlation*

Met: código de três letras para o aminoácido metionina

Mn: sigla para o elemento manganês

N: código de uma letra para o aminoácido asparagina

NR: *nuclear receptor*

OrthoMCL: método para identificação de genes ortólogos

P: código de uma letra para o aminoácido prolina

P-box: sequência específica de aminoácidos que reconhece HREs no DNA

PDB: *protein data bank*

PFAM: sigla para *protein family database*

Pro: código de três letras para o aminoácido prolina

PSICOV: *protein sparse inverse covariance*

Q: código de uma letra para o aminoácido glutamina

R: código de uma letra para o aminoácido arginina

RMRCM: regularized multinomial regression-based correlated mutations

RNA: *ribonucleic acid*

S: código de uma letra para o aminoácido serina

SCA: *statistical coupling analysis*

SDP: *specificity determining position*

Ser: código de três letras para o aminoácido serina

SOD: superóxido dismutase

T: código de uma letra para o aminoácido treonina (vide contexto)

T: código de uma letra para o nucleotídeo de timina (vide contexto)

T3: triiodotironina

T4: tiroxina

Thr: código de três letras para o aminoácido treonina

tRNA: RNA transportador

Trp: código de três letras para o aminoácido triptofano

TTR: transtirretina

UniProtKB: base de dados sobre proteínas

V: código de uma letra para o aminoácido valina

Val: código de três letras para o aminoácido valina

W: código de uma letra para o aminoácido triptofano

WebLogo: aplicativo online para criação de representações gráficas de perfis de frequências de AMSs

Y: código de uma letra para o aminoácido tirosina

Zn: sigla para o elemento zinco

LISTA DE FIGURAS

SUMÁRIO

1 – INTRODUÇÃO

Os recentes avanços nas tecnologias de sequenciamento *de novo* e ressequenciamento têm gerado um fluxo intenso de dados genômicos (MARX, 2013; CHAGOYEN *et al*., 2015). Consequentemente, temos observado um aumento exponencial no número de sequências de proteínas armazenadas em bases de dados públicas, o qual é ordens de magnitude maior do que o número de proteínas cuja estrutura tridimensional é conhecida ou que dão pistas sobre a sua função (CHAGOYEN *et al*., 2015). Conforme a quantidade de sequências de proteínas conhecidas aumenta exponencialmente, torna-se impossível determinar experimentalmente suas funções biológicas, bem como as regiões particulares dessas proteínas que são responsáveis por tais funções (PIETROSEMOLI *et al*., 2012). Sendo assim, a principal característica da assim chamada "era pós-genômica" é a necessidade de métodos e ferramentas para analisar a imensa quantidade de dados – principalmente na forma de sequências de polímeros biológicos – produzidos na era genômica (PAZOS & BANG, 2006). Isso tem levado a um aumento na dependência de métodos automáticos de anotação, apesar das taxas relativamente altas de erros destes programas, particularmente para famílias multigênicas (FAWAL *et al*., 2014).

Análises de evolução molecular utilizando abordagens computacionais são eficientes para inferir função em proteínas. Isso porque proteínas que compartilham um ancestral comum, denominadas proteínas homólogas, usualmente compartilham uma função geral, enquanto especificidades funcionais podem variar entre subconjuntos de proteínas dentro da família (CHAKRABORTY & CHAKRABARTI, 2014). Fazendo uma abordagem extremamente reducionista da biologia, pode-se dizer que a maioria das informações biológicas, se não todas, podem ser obtidas a partir das sequências de biomoléculas (PAZOS & BANG, 2006; PIETROSEMOLI *et al*., 2012). Uma forma de se obter vantagem sobre a massiva quantidade de dados atualmente disponível é usar métodos de comparação de sequências para coletar e comparar proteínas homólogas. Tal estudo comparativo dos membros de um grupo de proteínas homólogas, também chamado de "superfamília", oferece grande quantidade de informações acerca de características funcionais e estruturais de seus membros (CHAGOYEN *et al*., 2015).

A teoria neutra da evolução (KIMURA, 1979) sugere que várias trocas aleatórias de aminoácidos podem ocorrer em um sítio proteico sem alterar a função geral da proteína,

enquanto apenas alguns sítios estão sob pressão evolutiva, a qual se reflete em uma conservação mais proeminente de determinadas propriedades estruturais ou da sequência (CHAKRABARTI *et al*., 2007; CHAKRABORTY & CHAKRABARTI, 2014). Nesse sentido, alinhamentos múltiplos de sequências (AMS) são extremamente úteis no estudo comparativo de proteínas homólogas, e podem ser pensados como a representação dos resultados de um "experimento de tentativa e erro realizado pela evolução", em que várias mutações são geradas em diferentes sítios, sendo mantidas apenas aquelas que resultarem em proteínas funcionais. São uma importante ferramenta de modelagem cujo desenvolvimento exige uma complexa abordagem que combina problemas computacionais e biológicos (CHAGOYEN *et al*., 2015; CHATZOU, 2015).

Os métodos de AMS se referem a uma série de soluções algorítmicas para o alinhamento de sequências evolutivamente relacionadas, ao passo que leva em conta eventos evolutivos tais como mutações, inserções, deleções e rearranjos sob certas condições. Tais métodos podem ser aplicados a sequências de DNA, RNA ou proteínas, e estão atualmente entre os mais amplamente utilizados em biologia. De fato, um grande número de análises *in silico* depende de métodos de AMS, dentre estes: análises de domínios proteicos, reconstrução filogenética, busca por motivos funcionais e mais uma ampla gama de outras aplicações (CHATZOU, 2015).

Analisando um alinhamento ao nível de resíduos, aqueles que forem evolutivamente correspondentes aparecem na mesma coluna, mais comumente chamada de "posição". Sendo assim, uma posição em um alinhamento pode ser considerada como a representação das mutações "permitidas" pela evolução em resíduos correspondentes de diferentes proteínas homólogas. Nesse sentido, uma posição com um aminoácido invariável, ou seja, uma posição totalmente conservada em um AMS, é interpretada como sendo um resíduo importante para a estrutura e/ou função da proteína, uma vez que ali nenhuma mudança teria sido permitida durante o processo evolutivo (PAZOS & BANG, 2006; PIETROSEMOLI *et al*., 2012; CHAGOYEN *et al*., 2015).

Posições conservadas são o primeiro indicador de funcionalidade e relacionam-se com todos os tipos de sítios funcionais: sítios ativos, de ligação a pequenas moléculas, de interação proteína-proteína, de ligação a ácidos nucleicos, etc. Nem todas as posições conservadas se relacionam com função, muitas delas sendo conservadas devido a requisitos

estruturais, como por exemplo, formação do núcleo estrutural da proteína (PAZOS & BANG, 2006). Posições conservadas podem ser distinguidas, em certa medida, pelo tipo de aminoácido que está sendo mantido. Alguns aminoácidos – como Trp, Leu, Gly e Cys, por exemplo – costumam ter papel estrutural quando conservados, enquanto outros – principalmente apolares, ou de tipos específicos como Asp, Ser, Cys e His – costumam estar presentes em sítios ativos e/ou de ligação (VILLAR *et al*., 1994; PAZOS & BANG, 2006; PIETROSEMOLI *et al*., 2012).

Embora o problema da localização de posições conservadas possa parecer trivial à primeira vista, não existe um único método para tal finalidade (PIETROSEMOLI *et al*., 2012) . Ao contrário, existem várias abordagens distintas que geram diferentes resultados (VALDAR, 2002; CAPRA & SINGH, 2007). Por exemplo, uma posição que contenha tanto Arg quanto Lys e His seria reportada como sendo variável ao se utilizar um método baseado em entropia, enquanto que com outro método que levasse em conta substituições conservativas – aquelas que preservam a característica química do resíduo substituído – tal posição seria reportada como sendo relativamente conservada (PAZOS & BANG, 2006; PIETROSEMOLI *et al*., 2012).

À parte dos casos de total conservação existem outros padrões em posições de um AMS que são igualmente indicativos de importância funcional. Estes padrões apresentam-se na forma de mudanças na conservação ou na taxa evolutiva em determinados sítios, e refletem divergência funcional dentro de uma família em decorrência de duplicações gênicas (GU, 1999; GU, 2001; OHNO, 2013; CHAKRABORTY & CHAKRABARTI, 2014). Se pudermos subdividir uma família de proteínas homólogas em diferentes subgrupos com especificidades funcionais distintas – isto é, diferindo em certos detalhes funcionais ao mesmo tempo em que compartilham a função global da família, comum a todas as proteínas homólogas – , algumas posições podem apresentar-se conservadas apenas em um subgrupo particular, ou o aminoácido conservado pode ser diferente para cada grupo. O fato de que sua conservação é restrita a um subgrupo particular sugere que seu papel funcional relaciona-se com peculiaridades funcionais daquele grupo, não tendo ligação com a função global da família (CHAGOYEN *et al*., 2015).

Em casos em que uma dada família homóloga é populada por proteínas que apresentam características funcionais distintas, é razoável supor que tais características não

sejam determinadas pela presença de um único resíduo, mas sim por um conjunto de resíduos (BLEICHER *et al*., 2011). A existência de mudanças interdependentes em grupos de aminoácidos variáveis inspirou o desenvolvimento de ferramentas para detecção de coevolução molecular em nível de resíduos, as quais tipicamente utilizam um AMS para buscar por mutações correlacionadas (JUAN *et al*., 2013). Tais correlações mutacionais sugerem mudanças compensatórias que ocorrem entre resíduos que sejam de alguma forma associados, estejam eles próximos, em contato direto ou atuando em conjunto em sítios ativos ou de ligação, de maneira que a estabilidade, enovelamento ou função da proteína sejam mantidos (GÖBEL *et al.*, 1994; NEHER, 1994; SHINDYALOV *et al.*, 1994; TAYLOR & HATRICK, 1994). Além disso, é possível estender esses métodos para identificar sítios de interação proteína-proteína ao se buscar mutações correlacionadas entre proteínas distintas que sabidamente interagem (PAZOS *et al*., 1997; TRESS *et al*., 2005; YEANG & HAUSSLER, 2007; BURGER & VAN NIMWEGEN, 2008; WEIGT *et al*., 2009; SCHUG *et al*., 2009; DAGO *et al*., 2012). Paralelamente, desenvolveram-se métodos relacionados para buscar por grandes grupos de resíduos que são co-conservados de maneira específica dentro de subfamílias particulares de proteínas, e assim identificar resíduos que definem propriedades funcionais de uma subfamília, tais como especificidade de ligação ao substrato de uma determinada enzima (CASARI *et al.,* 1995; LICHTARGE *et al.,* 1996).

Uma das primeiras aplicações do conceito de mutações correlacionadas foi a busca por pares de contato, uma vez que observou-se que esses pares de posições fortemente correlacionadas pudessem corresponder a resíduos espacialmente próximos (GÖBEL *et al.*, 1994; SHINDYALOV *et al.*, 1994; NEHER, 1994; TAYLOR & HATRICK, 1994), e informações acerca de tais contatos têm inclusive ajudado na predição da estrutura tridimensional de proteínas (GRAÑA *et al.,* 2005; TRESS & VALENCIA, 2010; MARKS *et al.*, 2011; HOPF *et al.*, 2012; NUGENT & JONES, 2012; SUŁKOWSKA *et al.,* 2012). Desde então, muitos esforços foram feitos para estudar coevolução de resíduos na forma de pares de posições correlacionadas em AMS de famílias de proteínas (BLEICHER *et al.*, 2011; JUAN *et al*., 2013).

A coevolução de resíduos foi originalmente descrita a partir da detecção de pares de posições que apresentavam frequências interdependentes de aminoácidos ou padrões

similares de substituições destes (JUAN *et al*., 2013). Uma abordagem usualmente chamada de McBASC (do termo em inglês “McLachlan-based substitution correlation) constrói tais padrões de substituição de acordo com uma matriz de substituição de aminoácidos pré-calculada, e com isso é capaz de verificar sua similaridade através de uma correlação linear (FODOR & ALDRICH, 2004). Esse método mostra uma pequena porém significativa capacidade de recuperar pares de resíduos em contato físico e tem sido extensivamente testado e comparado com métodos mais recentes, servindo como linha de base para testar a performance de novos métodos (DUNN *et al.*, 2008). O uso de matrizes de substituição para estimar covariação entre posições impõe um cenário um tanto quanto exigente quando se deseja detectar posições muito variáveis como sendo covariantes (JUAN et al., 2013). Consequentemente, o método McBASC tende a detectar como coevoluídas posições bastante conservadas durante a evolução (FODOR & ALDRICH, 2004; JUAN *et al*., 2013). Sem assumir uma natureza bioquímica compensatória específica, essa abordagem oferece apenas uma estimativa grosseira da magnitude da correlação entre posições (JUAN *et al*., 2013). Além de carecer de um limite de confiança definido que permitisse a extrapolação para diferentes casos, a proposta inicial não reportava problemas associados à qualidade do alinhamento, tais como a inclusão de sequências redundantes ou divergentes (OLMEA & VALENCIA, 1997). Foi inspirada nessa abordagem que a análise de coevolução utilizando sequências de proteínas (CAPS na sigla de “co-evolution analysis using protein sequences”) (FARES & TRAVERS, 2006) veio amenizar a influência do *background* de divergência filogenética ao requerer que as correlações ainda sejam detectadas mesmo depois que alguns clados sejam removidos do AMS, além de corrigir a matriz de substituição de modo a considerar a verdadeira divergência entre as sequências. Porém, devido à sua alta demanda computacional, o método CAPS ainda não foi testado em conjuntos de dados em larga escala, tendo sido aplicado somente em casos específicos (JUAN *et al*., 2013).

Uma abordagem chamada de “informação mútua” também foi utilizada para detectar posições covariantes. Enquanto métodos baseados em correlação exploram substituições de aminoácidos entre sequências, informação mútua considera a distribuição de cada aminoácido nas diferentes sequências para uma determinada posição (JUAN *et al*., 2013). Para estimar a informação mútua entre duas posições, Atchley e colaboradores

(1999) usaram uma métrica chamada “associação de posições” (*pa-values* no termo em inglês), que quantifica em que medida a presença de um aminoácido, em uma dada sequência para uma determinada posição, seria uma “boa predição” da presença de algum outro aminoácido na mesma sequência, agora para uma segunda posição (KORBER *et al.*, 1993). Assim, grupos de posições entre as quais encontravam-se os valores de *pa-values* mais altos eram chamados de “cliques” (ATCHLEY *et al.*, 1999). Por observar apenas a significância estatística das covariações observadas, sem considerar quais aminoácidos particulares estão presentes nas mesmas sequências em ambas as posições, a magnitude das mudanças bioquímicas não é levada em conta durante a verificação da similaridade dos padrões mutacionais, uma vez que os diferentes aminoácidos são tratados como diferentes símbolos que não expressam relações de similaridade (JUAN *et al*., 2013). As formulações iniciais dessa abordagem eram vulneráveis a grandes variações nos valores de conservação no AMS (FODOR & ALDRICH, 2004) assim como ao efeito do *background* filogenético, problemas que por sua vez foram tratados em versões subsequentes (DUNN *et al.*, 2008; TILLIER & LUI, 2003; MARTIN *et al.*, 2005) que aumentaram a performance da predição de contatos (DUNN *et al.*, 2008), oferecendo informação importante para a compreensão da estrutura de proteínas (FAIRMAN *et al.*, 2012).

As abordagens McBASC e informação mútua são provavelmente as mais comumente usadas no estudo de coevolução de resíduos, embora outras técnicas tenham sido desenvolvidas ao longo dos anos recentes (JUAN *et al*., 2013). Dentre elas estão os métodos com uma abordagem vetorial, que verificam similaridades entre vetores que representam a presença ou ausência dos 20 diferentes aminoácidos em cada coluna do AMS (ALTSCHUH *et al.*, 1987; OLIVEIRA *et al.*, 2002). Outros métodos usam abordagens filogenéticas que caracterizam a sequência de mudanças evolutivas ocorridas ao longo do tempo partindo da reconstrução de um estado ancestral, a fim de detectar padrões de substituições simultâneas (YEANG & HAUSSLER, 2007; FLEISHMAN *et al.*, 2004; DUTHEIL *et al.*, 2005) ou explicitamente contrastar modelos independentes de evolução e coevolução (POLLOCK *et al.*, 1999; BARKER *et al.*, 2005). Nesse caso, o uso de um modelo aperfeiçoado para coevolução de sequência utilizando um processo de Markov contínuo no tempo (YEANG & HAUSSLER, 2007) representou um importante avanço (JUAN *et al*., 2013). Essas abordagens foram adequadas para estudos de coevolução em

pequenas famílias de proteínas em pequena escala, porém para estudos em larga escala a avaliação de sua performance permanece excessivamente exigente em termos computacionais (JUAN *et al.*, 2013).

Durante as análises de coevolução entre resíduos, um importante obstáculo pode surgir quando mais de duas posições mostram padrões coordenados de substituição. A interdependência evolutiva de cada posição para com uma ou mais posições adicionais leva a uma covariação aparente ou acoplamentos indiretos, cuja agregação pode tornar difícil o reconhecimento das posições diretamente interdependentes (JUAN *et al.*, 2013). Nesse contexto, novas abordagens tornam-se então necessárias para distinguir os acoplamentos diretos dos indiretos, visto que os primeiros são mais confiáveis para predizer resíduos fisicamente próximos na estrutura da proteína (JUAN *et al.*, 2013). Mesmo sendo variados, todos os métodos consideram que o conjunto de acoplamentos diretos entre pares de posições compõem um conjunto maior de posições indiretamente acopladas para formar a rede completa de mutações coordenadas (JUAN *et al.*, 2013). Lapedes e colaboradores (1999) propuseram um primeiro modelo básico no qual consideravam que acoplamentos indiretos não representam interdependência evolutiva e podem ser considerados como sendo covariações par a par não informativas, e então usaram um algoritmo Monte Carlo para inferir o mais simples modelo probabilístico capaz de considerar toda a rede de covariações em um cenário simulado (LAPEDES *et al.*, 1999). Infelizmente, foi necessário que um bom número de publicações recentes (WEIGT *et al.*, 2009; MORCOS *et al.*, 2011) utilizando diferentes estratégias revisitassem o problema para que só então a importância dessa primeira abordagem fosse conhecida (JUAN *et al.*, 2013).

A covariância inversa esparsa de proteína (PSICOV, do termo em inglês “protein sparse inverse covariance”) (JONES *et al.*, 2012) e a análise de acoplamento direto (DCA, sigla em inglês de “direct coupling analysis”) (WEIGT *et al*., 2009; SCHUG *et al*., 2009; DAGO *et al*., 2012) estabeleceram um modelo estatístico global do AMS em termos de acoplamento entre posições e variabilidade posição específica (MORCOS *et al.*, 2011; JONES *et al.*, 2012; JUAN *et al*., 2013). Para resolver o modelo, utilizam-se abordagens heurísticas para obter os valores estimados de interposição direta. No caso do método DCA, esses valores podem finalmente ser transformados em uma formulação baseada em informação mútua (JUAN *et al*., 2013). Uma abordagem relacionada foi utilizada de modo

interessante para realizar buscas por homologia de sequências (BALAKRISHNAN *et al.*, 2011). De maneira alternativa, o método de Burger e van Nimwegen (2008) usa um modelo de rede Bayesiana que inclui dependências condicionais entre os pares, enquanto que a abordagem de mutações correlacionadas baseadas em regressão multinomial regularizada (RMRCM, do inglês "regularized multinomial regression-based correlated mutations") (SREEKUMAR *et al.*, 2011) leva em conta não apenas as dependências entre pares individuais, mas a rede total de dependências (JUAN *et al*., 2013). Embora futuras investigações sejam necessárias, espera-se que abordagens de acoplamento direto sejam capazes de remover influências inespecíficas, tais como o *background* filogenético previamente discutido (JUAN *et al.*, 2008; PAZOS & VALENCIA, 2008; JUAN *et al*., 2013). Além disso, tais contribuições claras são obtidas apenas para famílias de proteínas com milhares de membros (DI LENA *et al.*, 2012). Para AMSs com mais de 1000 sequências, DCA e PSICOV parecem ser superiores ao método de Burger e van Nimwegen (MORCOS *et al.*, 2011; JONES *et al.*, 2012). Apesar disso, ainda é muito cedo para comparar completamente essas metodologias, as quais representam um importante avanço neste campo (JUAN *et al*., 2013). Alguns desses métodos são de fato capazes de predizer contatos entre resíduos afastados na sequência linear de aminoácidos com acurácia suficiente para guiar experimentos *in silico* de enovelamento de proteínas (GRAÑA *et al.*, 2005; TRESS & VALENCIA, 2010; MARKS *et al.*, 2011; HOPF *et al.*, 2012; NUGENT & JONES, 2012; SUŁKOWSKA *et al.*, 2012). Conforme os resultados da recente Avaliação Crítica das Técnicas de Predição de Estrutura de Proteínas (CASP, sigla em inglês para "Critical Assessment of Techniques for Protein Structure Prediction") permanecem altamente inconclusivos, será interessante observar como a aplicação desses novos métodos irá progredir nas mãos da comunidade científica (JUAN *et al*., 2013).

Embora vários métodos que se ocupam em detectar interações par a par deliberadamente excluam grandes grupos de resíduos covariantes por supostamente se tratarem apenas de covariações aparentes, ao contrário, esses grupos de resíduos podem oferecer informação útil acerca do que nem sempre se relaciona com contatos diretos (JUAN *et al.*, 2013). Em muitas famílias de proteínas, como as cinases, suas árvores filogenéticas revelam várias subfamílias que geralmente correspondem a proteínas que apresentam especificidades funcionais distintas – como ligação a diferentes substratos ou

efetores – dentro do contexto da função comum da família inteira (JUAN *et al.,* 2013). Como mencionado anteriormente, enquanto algumas posições são conservadas em todas as sequências – podendo corresponder a resíduos com importância estrutural ou papel catalítico em todas as proteínas da família – outras posições podem manter-se conservadas apenas dentro de subfamílias particulares (JUAN *et al.,* 2013). Uma vez que aparecem agrupados também na estrutura tridimensional da proteína, a esses resíduos específicos de subfamílias geralmente atribui-se a responsabilidade de determinar a funcionalidade específica daquela subfamília (MADABUSHI *et al.,* 2002; DEL SOL MESA *et al.,* 2003) por localizarem-se em sítios de ligação a ligantes e/ou proteínas (CASARI *et al.,* 1995; LICHTARGE *et al.,* 1996; DEL SOL MESA *et al.,* 2003; RAUSELL *et al.,* 2010) ou formarem cadeias alostéricas de resíduos (RODRIGUEZ *et al.,* 2010). Esse padrão de conservação específico de subfamílias costuma refletir em padrões mutacionais correlacionados, o que torna esse fenômeno uma parte essencial do estudo de coevolução molecular (JUAN *et al.,* 2013). Por refletir sua relação com a estrutura das árvores filogenéticas, essas posições eram originalmente chamadas de “posições determinantes de árvore” e mais recentemente foram renomeadas como “posições determinantes de especificidade” (SDPs, do inglês “specificity determining positions”) para destacar seu potencial papel funcional (JUAN *et al.,* 2013).

Enquanto várias ferramentas vêm sendo desenvolvidas especialmente para detectar SDPs (PAZOS & BANG, 2006; DE MELO-MINARDI *et al.,* 2010; BENÍTEZ-PÁEZ *et al.,* 2011; PIETROSEMOLI *et al.,* 2012; TEPPA *et al.,* 2012; CELNIKER *et al.,* 2013; CHAKRABORTY & CHAKRABARTI, 2014; CHAGOYEN *et al.,* 2015), adicionalmente, métodos relacionados baseados em análise de acoplamento estatístico (SCA, do inglês “statistical coupling analysis”) explicitamente buscam por grupos de resíduos coevoluídos que podem contribuir com processos tais como enovelamento de proteínas (SOCOLICH *et al.,* 2005) ou interações alostéricas (LOCKLESS & RANGANATHAN *et al.,* 1999; KASS & HOROVITZ, 2002; SÜEL *et al.,* 2003; REYNOLDS *et al.,* 2011). Esses métodos se diferenciam dos métodos de detecção de SDPs por seu requerimento pouco restritivo para identificar resíduos como sendo específicos de subfamílias particulares de proteínas (JUAN *et al.,* 2013). Métodos de análise de acoplamento estatístico (SCA) são projetados para detectar posições com padrões similares de aminoácidos, embora seu foco usualmente

esteja em padrões funcionalmente associados que realmente definem grupos de resíduos coevoluídos. São desafiadores de se classificar e podem ser considerados como sendo uma combinação de abordagens baseadas em covariação de resíduos e SDPs (JUAN *et al.*, 2013).

Vagamente baseada na mecânica estatística de Boltzman, a primeira implementação do método de SCA (LOCKLESS & RANGANATHAN, 1999) (aqui chamado SCA antigo) propunha que a mudança na frequência de um aminoácido em uma posição produz uma perturbação estatística na frequência de posições evolutivamente acopladas. O método apresentava dois parâmetros do tipo energéticos para quantificar tanto conservação posicional quanto correlação entre posições. O primeiro, denominado ΔG, relacionado com famosa medida conhecida como entropia de sequência (SHENKIN *et al.*, 1991), mede a conservação total em uma posição do alinhamento. O segundo, denominado $\Delta\Delta G$, mede o efeito sobre as frequências dos aminoácidos em uma determinada posição do alinhamento quando se tem um dado aminoácido em outra posição. Por exemplo, $\Delta\Delta G_{i|j = ALA}$ irá medir o quanto a distribuição de aminoácidos na posição i varia quando há uma alanina na posição j (LOCKLESS & RANGANATHAN, 1999). Aqui o conceito de perturbação oferece uma informação muito mais útil ao medir a correlação entre as posições. Isso porque simplesmente afirmar que as posições 25 e 45 são altamente correlacionadas sugere apenas que as duas posições podem estar em contato na estrutura tridimensional. Em vez disso, dizer que a fração de cisteínas na posição 45 aumenta consideravelmente quando há outra cisteína da posição 25 sugere que tal contato possa se tratar de uma ponte dissulfeto (BLEICHER *et al.*, 2011). Entretanto, embora fosse capaz, o método não recuperava tal informação, reportando apenas o $\Delta\Delta G$ total entre duas posições dada a presença de um aminoácido particular em uma delas. Utilizando o mesmo exemplo, o método apenas reportava que a presença de uma cisteína na posição 25 resultava em uma grande variação na distribuição de aminoácidos na posição 45 comparado com a distribuição no alinhamento total (BLEICHER *et al.*, 2011).

Outra propriedade interessante da métrica original de correlação da SCA é o fato de que ela automaticamente leva em conta o tamanho da amostra, uma vez que usa uma distribuição binomial de probabilidades. Ao contrário do que aconteceria para informação mútua, por exemplo, uma dada correlação ocorrendo para 50 sequências em um

alinhamento de 100 sequências não teria uma medida tão alta quanto teria a mesma correlação ocorrendo para 500 sequências em um alinhamento de 1000 sequências (BLEICHER *et al.*, 2011). Entretanto, ela também tem sérias limitações. Supostamente uma alternativa teórica para experimentos como análise de ciclo termodinâmico, a ideia de um parâmetro do tipo energético era usada a fim de "medir o acoplamento energético entre as posições de um AMS" (LOCKLESS & RANGANATHAN *et al.*, 1999). Estudos subsequentes mostraram que algoritmos de mutação correlacionada (incluindo SCA) de fato são capazes de encontrar pares de resíduos espacialmente próximos e que eles tendem a ser termodinamicamente acoplados, o que fazia parecer que havia uma correlação linear entre as duas medidas para o par de resíduos selecionado. Entretanto, há pouca evidência de que o acoplamento termodinâmico seja limitado somente a pares de resíduos obtidos por tais métodos (FODOR & ALDRICH, 2004; CHI *et al.*, 2008). Dekker e colaboradores (2004) propuseram um método baseado em perturbação que oferecia uma medida mais diretamente relacionada com estatísticas de covariação. Esse método baseava-se no cálculo da probabilidade explícita das covariâncias observadas ao invés de usar uma abordagem do tipo energética que nenhuma conexão verdadeira tinha com o conceito real de energia (DEKKER *et al.*, 2004). Esse método, quando comparado com o algoritmo original de SCA, mostrava um aumentado poder preditivo para encontrar contatos nativos ao medir o efeito da perturbação – presença de um certo aminoácido em uma posição – sobre a distribuição de frequências de outro sítio (BLEICHER *et al.*, 2011), utilizando o procedimento empírico do tipo *jackknife* descrito pelos desenvolvedores do SCA para determinar a "menor perturbação significativa" (SÜEL *et al.*, 2003). Pouco depois, visando reduzir resultados espúrios que eventualmente surgissem caso colunas pobremente conservadas fossem incluídas nas análises, Dima e Thirumalai (2006) propuseram um procedimento teoricamente robusto baseado na escolha do menor sub-alinhamento que ainda seria capaz de satisfazer o teorema central do limite (DIMA & THIRUMALAI, 2006).

O método de SCA antigo tem sido amplamente retrabalhado, e suas últimas versões, longe de serem um mero seguimento da abordagem original, representam um método completamente diferente, sendo apenas baseado em uma ideia similar (JUAN *et al.*, 2013). Por exemplo, em artigos mais recentes (HALABI *et al.*, 2009) os autores abandonam o

conceito de perturbação, informando apenas o quão correlacionadas são duas posições, de modo semelhante a outras métricas de correlação já existentes. A versão mais recente do SCA (aqui chamado de SCA novo) (REYNOLDS *et al.*, 2011) é baseada em uma análise multivariada em que são identificados grupos de posições covariantes chamados de "setores de proteínas" (HALABI *et al.*, 2009), formando muitas vezes agrupamentos independentes de aminoácidos com papéis distintos dentro da proteína. Para isso, correlações entre posições são ponderadas de acordo com sua conservação, e uma matriz de correlações ponderadas é reduzida por decomposição espectral. Espera-se com isso que os "componentes principais" obtidos sejam indicativos de conservações específicas de subfamílias, e grupos de posições covariantes são detectados como aqueles que contribuem significantemente com tais componentes (JUAN et al., 2013). Embora essas posições tendam a ser especificamente conservadas dentro de subfamílias, não se trata de SDPs pois elas não necessariamente são diferentemente conservadas nas diferentes subfamílias, ou seja, mais de uma subfamília pode apresentar o mesmo tipo de aminoácido (JUAN et al., 2013). Alguns estudos bastante interessantes em pequena escala têm usado abordagens semelhantes ao método de SCA para identificar redes de resíduos coevoluídos implicados no enovelamento de proteínas (SOCOLICH *et al.*, 2005). Em outros estudos, Süel e colaboradores (2003), submeteram matrizes de correlação (ΔΔG) a métodos de agrupamento a fim de se obter conjuntos de posições auto-correlacionadas, os quais eram postulados como representando "motivos estruturais para comunicação alostérica em proteínas" (SÜEL *et al.*, 2003). Apesar das contribuições significativas do método de SCA, pouca atenção foi prestada para o fato de que conjuntos de posições correlacionadas podem ter significados muito diferentes (BLEICHER *et al.*, 2011). Por exemplo, quando a presença de um determinado aminoácido em uma posição provoca uma perturbação nas frequências de outra posição no sentido de diminuir a frequência de certo aminoácido, diz-se que esses dois aminoácidos são anti-correlacionados. Porém, tanto correlação quanto anti-correlação implicam em valores positivos de ΔΔG. Sendo assim, conjuntos de posições correlacionadas podem ser melhor compreendidos se analisarmos as contribuições individuais de cada tipo de aminoácido, a topologia da rede gerada e a diferenciação entre correlação e anti-correlação (BLEICHER et al., 2011). Além disso, comparações baseadas no SCA antigo e alguns de seus desenvolvimentos subsequentes têm mostrado que essa

abordagem não é particularmente competitiva para predição de contatos em proteínas (FODOR & ALDRICH, 2004; BROWN & BROWN, 2010). Apesar disso o método de SCA novo não tem sido devidamente avaliado em larga escala, sendo a ausência de padrões de avaliação a maior limitação para a aplicação geral de abordagens baseadas em nesse método (JUAN et al., 2013).

A partir de um conjunto de correlações posicionais específicas entre pares de resíduos, é possível construir uma rede cujos nós representam determinados aminoácidos em posições específicas e as arestas denotam o quão correlacionados (ou anti-correlacionados) são esses nós (BLEICHER *et al.*, 2011). No caso de correlação, a relação entre dois nós implica que sequências que contêm o primeiro aminoácido naquela posição também tendem a ter o segundo na outra posição. Ao contrário, no caso de anti-correlação, sequências contendo o primeiro aminoácido naquela posição tendem a não ter o segundo aminoácido na outra posição (BLEICHER *et al.,* 2011). Utilizando a definição de modularidade (NEWMAN & GIRVAN, 2004), que é a medida relacionada a uma rede e a uma divisão dessa rede em grupos, uma comunidade em uma rede seria então um grupo de nós que apresentam fortes conexões entre si, mas pouca ou nenhuma conexão com o restante da rede. Dessa forma, havendo um grande número de arestas (conexões) entre os vértices (nós) do mesmo grupo, mas não muitas entre vértices de diferentes grupos, a modularidade resultante terá um valor alto (BLEICHER *et al.,* 2011). Algoritmos de detecção de estruturas de comunidades em redes têm sido o foco de diferentes grupos desde o início dos anos 2000 (FLAKE *et al.*, 2002; NEWMAN & GIRVAN, 2004; RADICCHI *et al.*, 2004) e podem ser usados para identificar sequências de acordo co a composição de aminoácidos que elas apresentam nas posições correspondentes a cada comunidade (BLEICHER *et al.*, 2011). Algoritmos de detecção de comunidades, portanto, devem ser capazes de maximizar a modularidade, seja por busca exaustiva ou por métodos heurísticos, sendo os últimos necessários para grandes redes. Esses algoritmos podem ser adaptados expandindo-se a definição original de modularidade (GÓMEZ *et al.*, 2009), de modo a especificar a intensidade da conexão entre dois nós, além de incluir arestas direcionais para casos em que se observa relações distintas de A para B e de B para A (BLEICHER *et al.*, 2011). Finalmente, tais algoritmos podem também atribuir pesos negativos para casos em

que os aminoácidos são anti-correlacionados, a fim de destacar o fato de que não deve haver conexão entre dois nós (BLEICHER *et al.*, 2011).

Em 2011, Bleicher e colaboradores propuseram um método simples para explorar e quantificar correlações específicas. O método usa uma distribuição binomial cumulativa para calcular o *p-value* – ou seja, a probabilidade de que seja ao acaso – do aumento (ou diminuição) da frequência de um dado aminoácido em determinada posição, dado que haja um aminoácido específico em outra posição, tendo como probabilidade *a priori* as frequências no alinhamento total. Por exemplo, dado que a frequência de cisteínas na posição 45 para o alinhamento total é de 40%, qual a probabilidade ao acaso de que essa frequência seja de no mínimo 90% no sub-alinhamento em que as sequências contêm outra cisteína na posição 25? Ou no caso de uma anti-correlação, dado que a frequência de aspartatos na posição 45 para o alinhamento total seja de 60%, qual a probabilidade ao acaso de que essa frequência seja de no máximo 10% no mesmo sub-alinhamento contendo uma cisteína na posição 25? Quanto mais improvável for que estes resultados se deem ao acaso, maior será a correlação ou anti-correlação. No exemplo citado, as cisteínas 25 e 45 apareceriam na análise como sendo correlacionadas, enquanto que a cisteína 25 e o aspartato 45 seriam anti-correlacionados. Na rede de aminoácidos correlacionados, as conexões (arestas) recebem o valor do cologaritmo do *p-value*, com sinal positivo em caso de correlação e negativo para anti-correlação, podendo também ser direcionais dependendo de qual aminoácido do par causou a mudança na frequência do outro. Uma vez construída a rede de correlações, o método utiliza algoritmos de detecção de comunidades para encontrar grupos de aminoácidos específicos que tendem a estarem presentes simultaneamente na proteína. Também é medida a "aderência" das sequências às comunidades detectadas, ou seja, quais sequências apresentam um ou outro conjunto de resíduos, oferecendo informação importante acerca de eventuais divergências de função que podem ocorrer dentro de uma família de proteínas (BLEICHER *et al.*, 2011).

Alguns trabalhos têm demonstrado que métodos baseados em correlação que discriminam os aminoácidos envolvidos nos pares correlacionados são bastante eficientes para detectar divergência funcional em famílias de proteínas, especialmente em casos que apresentam aminoácidos mutualmente exclusivos ocupando as mesmas posições. (BLEICHER et al., 2011). Por exemplo, membros da família das Fe/Mn superóxido

dismutases (SODs) podem ser tanto diméricos quanto tetraméricos, e usualmente ligar-se seletivamente a Fe ou Mn no sítio ativo de uma maneira não substituível para apresentar atividade catalítica, sendo que tais propriedades são independentes, ou seja, ambas as SODs (que se ligam a Fe ou Mn) podem ser tanto diméricas quanto tetraméricas (WINTJENS *et al.*, 2008; BACHEGA *et al.*, 2009; BLEICHER et al., 2011). O método de Bleicher e colaboradores foi capaz de detectar grupos correlacionados que aparentemente tinham relação com posições que já haviam sido descritas como determinantes do estado oligomérico e seletividade a metais (WINTJENS *et al.*, 2008), mostrando que posições correlacionadas podem agrupar-se em diferentes conjuntos com propriedades independentes (BACHEGA *et al.*, 2009; BLEICHER et al., 2011). Além do caso das Fe/Mn SODs, consideradas como sendo um "caso ideal" para análises de correlação, no mesmo artigo (BLEICHER et al., 2011) também foram descritos outros dois exemplos com propriedades completamente distintas: a superfamília das peroxidases-catalases e a famíla das lisozimas tipo-C/alfa-lactalbuminas. A primeira pode ser subdividida em três classes com características distintas e seu elevado número de pares correlacionados representava um desafio para os procedimentos de detecção de comunidades. A segunda apresenta tanto membros com atividade de lisozima quanto membros que são unidades regulatórias da enzima lactose sintetase, as quais apresentam independentemente sítios de ligação a cálcio. Seu pequeno número de sequências também desafiava o método no quesito amostragem, uma vez que este usa uma abordagem estatística (BLEICHER et al., 2011). Os resultados para ambos os casos mostraram que a utilização de métricas de correlação resíduo-específicas seguida de análises de comunidades foi capaz de detectar características determinantes de subclasses de proteínas. Isso representava uma vantagem sobre outros métodos previamente descritos que reportavam apenas de maneira geral as interdependências entre posições, sem discriminar os aminoácidos específicos envolvidos (BLEICHER et al., 2011). Vale destacar também, mais especificamente no caso das lisozimas tipo-C/alfa-lactalbuminas, o efeito causado pelo uso de diferentes valores de *cutoff* para os filtros de identidade máxima nessa família, em que novos resíduos emergiam à medida que os valores tornavam-se menos restringentes, ao ponto em que se tornava difícil interpretar prontamente os novos resultados (BLEICHER et al., 2011).

Outro exemplo de aplicação do método de Bleicher e colaboradores (2011) foi sobre a família da proteína Her2p, uma das subunidades do complexo GatCAB, uma amido transferase responsável pela transamidação do glutamil-$tRNA^{Gln}$ em glutaminil-$tRNA^{Gln}$ (FERREIRA-JÚNIOR *et al.*, 2013). O método foi capaz de detectar resíduos funcionalmente importantes, dentre eles, alguns localizados nas interfaces que conectam a Her2p às outras subunidades do complexo, outros que fazem contato direto com o tRNA, além de resíduos mutualmente exclusivos tanto de Her2p quanto de amidases, outro membro homólogo dessa família (FERREIRA-JÚNIOR *et al.*, 2013). Uma aplicação mais recente do método (BUSSO *et al.*, 2015) foi sobre a proteína Coq7p, um dos componentes do complexo multimérico envolvido na síntese de um composto lipofílico essencial para o crescimento aeróbico e fosforilação oxidativa, a coenzima Q. Aqui, foram identificados resíduos que modulam a atividade da proteína, os quais ao serem substituídos inibiam sua atividade devido a alterações estruturais (BUSSO *et al.*, 2015). Estes e outros casos, portanto, demonstram que o método de decomposição de redes de correlação entre aminoácidos proposto por Bleicher e colaboradores se mostra bastante eficaz na detecção de resíduos funcionalmente importantes, inclusive aqueles associados a características específicas de subclasses existentes dentro de famílias de proteínas homólogas.

2 – JUSTIFICATIVA

Um bom exemplo de divergência de função, muito bem descrito em um trabalho recente (AFONSO *et al.*, 2013), são os receptores nucleares (NRs, do inglês “nuclear receptors”), que são fatores de transcrição que regulam uma ampla gama de processos biológicos em metazoários e são ativados por ligantes específicos, geralmente pequenas moléculas hidrofóbicas, e podem estar relacionadas a uma série de doenças (AFONSO *et al.*, 2013). Os NRs apresentam dois domínios, sendo um domínio de ligação ao ligante (LBD, do inglês “ligand binding domain”) e um domínio de ligação a DNA (DBD, do inglês “DNA binding domain”). O domínio LBD é onde se ligam hormônios provocando alterações conformacionais e, consequentemente, o recrutamento de outras proteínas iniciadoras de transcrição. Já o domínio DBD se liga de maneira seletiva a uma região específica do DNA chamada de elemento de resposta a hormônio (HRE, do inglês “hormone response element”), por meio de dois motivos altamente conservados de “dedo de zinco” do tipo C4, um deles contendo uma sequência de aminoácidos conhecida como *P-box* (AFONSO et al., 2013). O motivo *P-box* em humanos usualmente apresenta uma sequência de aminoácidos CEGCKG, exceto pela classe de receptores de 3-cetoesteroides que apresentam uma sequência CGSCKV ou similar, demonstrando a existência de diferentes especificidades em NR-DBDs. Já no clado *nematoda*, por outro lado, esse motivo apresenta a sequência CRACAA, ou a sequência análoga CKACAA, sugerindo que, semelhante ao que acontece nos receptores de 3-cetoesteroides, os receptores nucleares em nematódeos teriam evoluído de modo a desenvolverem diferentes modos de ligação aos HREs no DNA (AFONSO *et al.*, 2013). Aqui é interessante destacar que os receptores de 3-cetoesteroides são exclusivos de vertebrados, sendo suas isoformas fruto da duplicação e diversificação de um receptor de estrógeno ancestral (THORNTON, 2001). Por serem muito recentes na história evolutiva, as sequências correspondentes ao domínio DBD dos receptores de 3-cetoesteroides apresentam alta identidade entre si, sendo portanto cortadas do alinhamento durante o procedimento de filtragem por identidade máxima (AFONSO *et al.*, 2013), semelhante ao que ocorre com a família das lisozimas tipo-C/alfa-lactalbuminas mencionadas anteriormente. Por ser pouco amostrada em número de sequências no alinhamento, características específicas dessa subfamília não podem ser detectadas em análises de correlação como o de Bleicher e colaboradores (2011).

Outro exemplo bastante interessante é o caso da família das Transtirretinas / 5-hidroxi-isourato hidrolases (TTR/HIUase). A transtirretina (TTR) é uma proteína sérica em forma de homotetrâmero que se liga a hormônios tireoidianos e os transportam para todo o organismo, sendo seu principal transportador no cérebro. Também participa indiretamente do carreamento de vitamina A ao acoplar-se a proteínas que se ligam a retinol (POWER *et al.*, 2000), tendo sido verificada também uma atividade críptica de metalopeptidase (LIZ *et al.*, 2004; LIZ *et al.*, 2012). Foi descrita como uma proteína carreadora de tiroxina (T4) em mamíferos eutérios e de triiodotironina (T3) nos demais vertebrados (RICHARDSON, 2014). Mutações nessa proteína podem ocasionar vários tipos de doenças, em especial doenças neurodegenerativas causadas pela formação de fibrilas amiloides, que afetam principalmente nervos periféricos e coração (SARAIVA *et al.*, 2001; PALANINATHAN, 2012). Evidências indicam que o gene da TTR teria se originado durante o surgimento dos vertebrados, a partir de uma duplicação seguida de diversas mutações independentes no gene de uma enzima presente desde bactérias até vertebrados, envolvida no metabolismo de ácido úrico: a 5-hidroxi-isourato hidrolase (HIUase) (RAMAZZINA *et al.*, 2006; ZANOTTI *et al.*, 2006; HENNEBRY, 2009; RICHARDSON, 2014). Justamente por ser tão recente na história evolutiva, assim como nas lisozimas tipo-C/alfa-lactalbuminas e receptores nucleares de 3-cetoesteroides, as transtirretinas apresentam poucas sequências no alinhamento quando comparadas ao grande número de sequências de 5-hidroxi-isourato hidrolases provenientes de organismos de diferentes domínios da vida. Esse problema de amostragem também acaba por impossibilitar análises de correlação que utilizem filtros de identidade máxima durante a preparação dos dados.

Apesar de serem casos de naturezas bastante distintas, tanto a família dos receptores nucleares quanto a das TTR/HIUase têm uma coisa em comum: o fato de que alguma de suas subfamílias apresenta-se pouco representada em número de sequências nos alinhamentos. Ao mesmo tempo, estes costumam vir repletos de sequências redundantes, muitas vezes mutantes ou variantes da mesma sequência, oriundas principalmente de organismos modelo, o que acaba por tornar os alinhamentos pouco representativos da diversidade de proteínas de fato existente na natureza. Essa redundância nos alinhamentos, bem como a baixa amostragem de algumas subfamílias em número de sequências, acaba por enviesar análises com caráter estatístico tais como os métodos de correlação. Isso

porque, de acordo com Bleicher e colaboradores (2011), “se uma dada família de proteínas é populada por proteínas que apresentam propriedades distintas (…), é esperado que tais propriedades possam não ser determinadas pela presença de um único aminoácido, mas sim por um grupo destes – e esse grupo irá emergir de uma análise de correlação *se um número suficiente de proteínas apresenta aquelas propriedades*”.

Uma boa alternativa para solucionar tal problema seria atribuir pesos às sequências utilizando, por exemplo, o método proposto por Henikoff & Henikoff (1994). Esse método tem como característica destacar sequências “raras”, atribuindo pesos maiores a sequências destoantes das demais, ao passo que sequências parecidas entre si recebem pesos menores (VALDAR, 2002). Outra característica importante desse método é que a soma dos pesos individuais de todas as sequências do alinhamento é igual a um. Com isso, para se obter as probabilidades de ocorrência de um determinado aminoácido em uma posição (aqui convenientemente chamadas de “frequências” para facilitar futuras comparações), basta somar os pesos das sequências contendo aquele aminoácido naquela posição. Entretanto, métodos como o de Bleicher e colaboradores (2011), que utilizam uma distribuição binomial cumulativa para medir as correlações, necessitam de valores discretos para a realização dos cálculos (número de ocorrências de um determinado aminoácido numa posição, bem como o tamanho de determinado sub-alinhamento). Ao se atribuir pesos às sequências, não se estará mais trabalhando com frequências (número de ocorrências dividido pelo total), mas sim com probabilidades de ocorrência, ou seja, valores contínuos. Com isso, ao mesmo tempo em que resolve um problema (o da redundância e baixa amostragem de subfamílias em alinhamentos), o método de atribuição de pesos traz consigo outro problema de caráter operacional, fazendo-se necessário, portanto, o uso de uma nova abordagem que seja compatível com o uso de valores contínuos durante os cálculos de correlação.

3 – OBJETIVOS

3.1 – Objetivo Geral

Analisar os efeitos da atribuição de pesos a sequências, com base no seu potencial de aproveitamento de informação e redução do viés filogenético, como estratégia alternativa ao uso de filtros por identidade máxima para diminuição da redundância em alinhamentos múltiplos de sequências, a fim de aperfeiçoar análises de conservação e correlação entre resíduos.

3.2 – Objetivos específicos

✔ Verificar os efeitos da atribuição de pesos a sequências sobre as frequências dos resíduos, comparadas às frequências obtidas a partir de alinhamentos filtrados por identidade máxima.

✔ Desenvolver estratégias compatíveis com a atribuição de pesos a sequências para calcular correlação entre resíduos, especialmente para casos de de famílias homólogas em que haja subfamílias pouco representadas no alinhamento, ou que apresentem sequências redundantes.

✔ Propôr uma estratégia para minimizar a influência do efeito filogenético nas análises de correlação, causado pela alta similaridade entre sequências oriundas de duplicações gênicas recentes na história evolutiva.

4 – METODOLOGIA

4.1 – Conjunto de dados

Foi feito um estudo de caso com duas famílias de proteínas homólogas: a do domínio de ligação a DNA (DBD) dos receptores nucleares (NRs) e a das Transtirretinas/5-hidroxi-isourato hidrolases (TTR/HIUase). Os alinhamentos múltiplos de sequências foram baixados diretamente da base de dados de famílias de proteínas homólogas, o PFAM (FINN *et al.*, 2015).

Para remover fragmentos, utilizou-se um filtro de cobertura mínima em que sequências que apresentassem menos de 80% do tamanho da sequência referência escolhida – geralmente uma sequência bem anotada e, de preferência, com estrutura conhecida – eram cortadas do alinhamento, como descrito por Bleicher e colaboradores (2011). Para evitar o viés relativo a uma sequência de referência específica, pode-se também filtrar fragmentos a partir do percentual (no caso, os mesmos 80%) de cobertura em relação ao tamanho do perfil HMM utilizado como definição do domínio.

Informações acerca das posições funcionalmente importantes para NR-DBDs foram extraídas da literatura (AFONSO *et al.*, 2013). Para as TTR, essas informações foram obtidas a partir de diagramas gerados pelo programa LIGPLOT (WALLACE *et al.*, 1995), os quais podem ser acessados na base de dados PDBsum (LASKOWSKI, 2001) utilizando como entrada de busca a estrutura da transtirretina humana em complexo com T4 (PDBid 2ROX). Para definir quais eram os aminoácidos típicos de cada subfamília nessas posições, as sequências foram submetidas ao programa OrthoMCL (LI *et al.*, 2003) para então serem gerados sub-alinhamentos específicos de cada subfamília. Uma vez de posse desses sub-alinhamentos, era possível verificar quais os resíduos mais frequentes para cada uma das posições funcionalmente importantes.

4.2 – Filtro de identidade máxima e atribuição de pesos

Para resolver o problema da redundância entre sequências com alta similaridade, duas abordagens distintas foram utilizadas. A primeira consiste na aplicação de um filtro de identidade máxima, em que as sequências são comparadas todas com todas e, caso

apresentem identidade acima de determinado valor, aquela de menor tamanho é removida. As sequências foram filtradas com 80% e 70% de identidade máxima.

A segunda abordagem consiste na atribuição de pesos às sequências, evitando assim a eventual perda de informação decorrente do descarte de sequências. Para isso, utilizou-se o algoritmo de Henikoff & Henikoff (1994):

$$w_i = (\Sigma_x^L (K_x n_{xi})^{-1}) / \mathrm{L}$$

em que w_i é o peso da *i*-ésima sequência do alinhamento, L é o número de colunas do alinhamento, x é uma posição do alinhamento, K_x é o número de tipos diferentes de aminoácidos encontrados na posição x, e n_{xi} é o número de vezes em que o aminoácido presente na posição x da sequência i aparece nesta posição x. A soma dos pesos individuais de todas as sequências do alinhamento é igual a um.

4.3 – Perfil de frequências

A seguir, foi gerado um perfil simples contendo as frequências dos aminoácidos presentes em cada posição do alinhamento. Para os alinhamentos filtrados, as frequências eram calculadas somando-se o número de ocorrências de determinado aminoácido em uma posição e dividindo-se pelo total de sequências do alinhamento. Para o alinhamento em que foram atribuídos pesos às sequências, as frequências – ou mais apropriadamente as probabilidades de ocorrência de cada aminoácido – foram obtidas somando-se os pesos das sequências contendo determinado aminoácido naquela posição, conforme a seguir:

$$p_a = \Sigma\, w_i \,,\; i \,\epsilon\, \{ i \mid s_i(x) = a \}$$

em que p_a é a probabilidade de ocorrência do aminoácido a na posição x, $s_i(x)$ é o tipo de aminoácido encontrado na posição x da i-ésima sequência. Em outras palavras, w_i é o peso da sequência i contendo o aminoácido a na posição x (VALDAR, 2002).

Visando uma melhor visualização destes perfis, foram geradas logos das posições funcionalmente importantes através do aplicativo online WebLogo (CROOKS *et al.*, 2004). Para isso, todas as combinações de resíduos presentes nessas posições eram geradas e transformadas em pseudo-sequências, com as quais era gerado um arquivo fasta que seria então submetido ao aplicativo para gerar as logos. Nos alinhamentos sem filtro/pesos ou

filtrados, cada pseudo-sequência era repetida conforme o número de vezes que aquela combinação de aminoácidos naquelas posições aparecia no alinhamento. Nos alinhamentos filtrados, entretanto, o número de repetições se referia à soma dos pesos das sequências contendo tais combinações, multiplicado pelo número total de sequências do alinhamento.

4.4 – Análise de correlação e anti-correlação pelo método atual

Para cada resíduo identificado como sendo funcionalmente importante, foi gerado um sub-alinhamento contendo somente as sequências que apresentam aquele resíduo naquela posição específica. Novos pesos são atribuídos a este sub-alinhamento e um novo perfil de frequências é então gerado. O cálculo do *p-value* continua sendo feito da mesma forma como descrito por Bleicher e colaboradores (2011), utilizando uma distribuição binomial cumulativa, adaptando-se porém os valores utilizados no cálculo. Para calcular a probabilidade (*p-value*) de que a frequência de um aminoácido b na posição y aumente (ou diminua) dada a presença de um aminoácido a na posição x, o número de "tentativas" – que antes era o tamanho do sub-alinhamento contendo o aminoácido a na posição x – agora é o número total de sequências do alinhamento multiplicado pela soma dos pesos originais das sequências contendo este aminoácido nesta posição. O número de "sucessos" – que antes era o número de sequências contendo o aminoácido b na posição y dentro do sub-alinhamento – agora é o número de "tentativas" multiplicado pela soma dos novos pesos das sequências contendo este aminoácido nesta posição dentro do sub-alinhamento. Por fim, a probabilidade *a priori* é a soma dos pesos originais, ou seja, no alinhamento total, das sequências contendo o aminoácido b na posição y. Para que a variação na frequência do aminoácido b na posição y, devido à presença do aminoácido a na posição x, seja considerada significativa, essa frequência deve aumentar para um valor acima de determinado *cutoff* (no caso, 75%) em caso de correlação ou diminuir para um valor abaixo de determinado *cutoff* (25%) em caso de anti-correlação. Além disso, a probabilidade ao acaso (*p-value*) de que essa variação ocorra também deve estar abaixo de um determinado *cutoff* (aqui, 10^{-10}).

4.5 – Análise da distribuição da conservação nos sub-alinhamentos

Como aqui trataríamos de casos de subfamílias pouco amostradas, possivelmente menores que o sub-alinhamento mínimo proposto por Dima e Thirumalai (2006), a distribuição das medidas de conservação para todas as posições de cada sub-alinhamento foi observada a fim de verificar se as correlações nas posições específicas (posições funcionalmente importantes) são de fato significativas quando comparadas ao resto das posições, ou se somente se devem ao efeito filogenético causado pela alta identidade entre as sequências do sub-alinhamento.

A conservação pode ser inferida como sendo o oposto da variabilidade encontrada em uma dada posição. Uma medida de variabilidade bastante difundida é a Entropia de Shannon (SHANNON, 2001), que pode ser calculada conforme a seguir:

$$S_x = - \Sigma_a^K p_a \log_2 p_a$$

sendo S_x a entropia de Shannon observada na posição x, p_a é a probabilidade de ocorrência do aminoácido a na mesma posição x, e K é o conjunto de aminoácidos encontrados nesta posição, como mencionado anteriormente. Outra medida de conservação foi utilizar a frequência do aminoácido mais conservado para cada posição. Assim, para cada sub-alinhamento, a distribuição dos valores de entropia e das frequências dos aminoácidos mais conservados foi calculada e apresentada na forma de histogramas. Somente posições com menos de 20% de *gaps* foram incluídas na análise.

4.6 – Análise de correlação com atenção ao efeito filogenético

Caso fosse verificado, através da distribuição da entropia e das frequências dos aminoácidos mais conservados, que as sequências do sub-alinhamento apresentassem alta identidade entre si, as correlações encontradas não se deveriam a eventos de coevolução entre resíduos, mas sim ao efeito filogenético. Nesse caso, praticamente todas as posições do sub-alinhamento estariam correlacionadas entre si pelo simples fato de que as sequências seriam muito parecidas. Para contornar este problema, foi feita mais uma adaptação ao método original desenvolvido por Bleicher e colaboradores (2011), procedendo-se da mesma maneira como descrito na seção 4.4, porém modificando-se apenas a probabilidade *a priori* da distribuição binomial cumulativa. A nova probabilidade

a priori seria então a média das frequências dos aminoácidos mais conservados no sub-alinhamento para posições que apresentassem menos de 20% de *gaps*. Dessa maneira, caso as sequências do sub-alinhamento apresentassem alta identidade entre si, apenas mudanças significativas de frequência em relação à média do sub-alinhamento seriam detectadas.

5 – RESULTADOS E DISCUSSÃO

As análises dizem respeito às posições do alinhamento. Contudo, como estas podem variar, os resultados foram apresentados utilizando-se sequências referência a partir das quais é possível inferir as posições correspondentes em qualquer alinhamento em que tais sequências estejam presentes. Para os receptores nucleares (NRs) utilizou-se a sequência do domínio de ligação a DNA (DBD) do receptor nuclear humano hRXRα (UniProtKB ID: P19793). Para as transtirretinas (TTR) utilizou-se a sequência da transtirretina humana (UniProtKB ID: P02766), descontando-se os 20 resíduos correspondentes ao peptídeo sinal, conforme consta na estrutura (PDB ID 2rox).

5.1 – Transtirretinas / 5-hidroxi-isourato Hidrolases

5.1.1 – Efeitos da atribuição de pesos sobre as frequências

A figura 1 mostra as representações gráficas (logos) dos perfis de frequências das posições correspondentes a resíduos localizados em sítios de ligação a ligante (T3/T4) ou metal (Zn^{2+}). À primeira vista, é difícil notar diferenças marcantes. Tais diferenças ficam mais evidentes no gráfico da figura 2, que mostra a variação, em relação aos alinhamentos filtrados e sem filtro, nas frequências em cada posição mediante a atribuição de pesos às sequências. Algumas posições são mais afetadas pelos pesos, as quais serão discutidas mais adiante. Porém, de maneira geral, observam-se mais variações positivas do que negativas, principalmente em relação aos alinhamentos filtrados. Isso mostra que o uso de filtros pode levar à perda de informação ao descartar sequências com alta identidade entre si. O uso de pesos, pelo contrário, não só evita a perda dessa informação como também é capaz de destacá-la. A partir daqui, somente os alinhamentos ponderados foram utilizados nas análises.

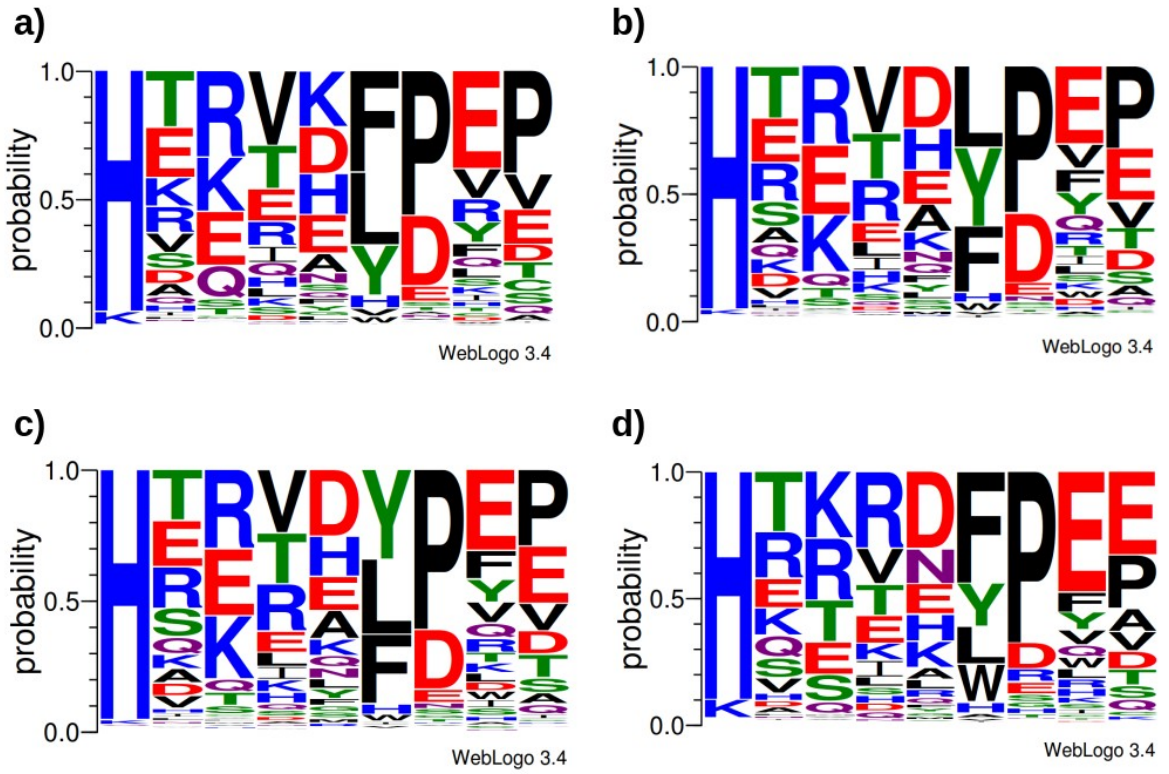

Figura 1: Representações gráficas (logos) dos perfis de frequências do alinhamento total para as posições compreendidas em sítios de ligação. As colunas das logos (da esquerda para a direita) se referem às posições correspondentes aos resíduos 15, 31, 70, 72, 74, 88, 89, 90 e 92 da sequência referência, a Transtirretina Humana (UniProtKB ID: P02766; PDB ID: 2rox). Os perfis se referem, respectivamente, aos alinhamentos: (a) sem pesos e sem filtro; (b) filtrado por identidade máxima de 80%; (c) filtrado por identidade máxima de 70%; (d) ponderado. As logos foram geradas pelo aplicativo online WebLogo (CROOKS *et al.*, 2004).

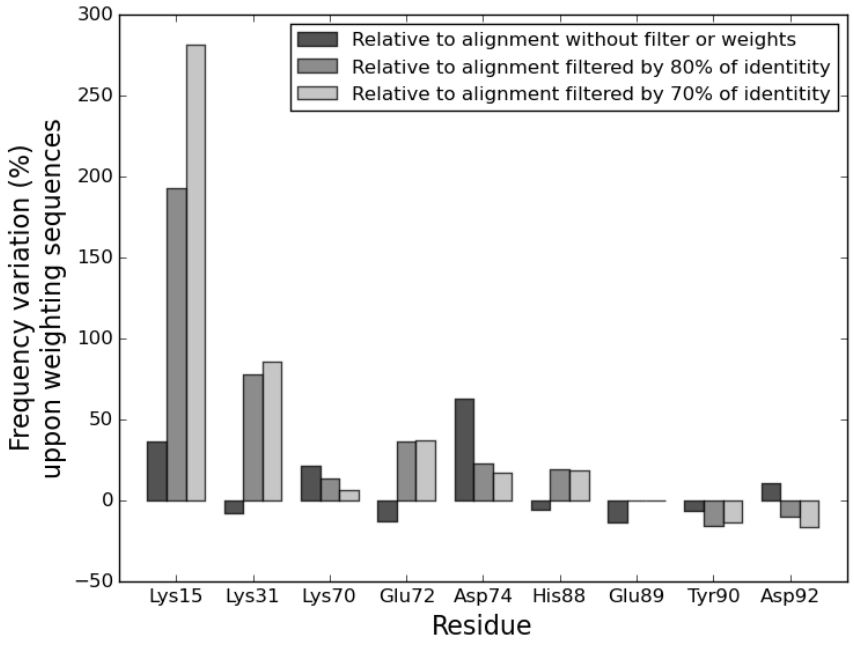

Figura 2: Variação nas frequências mediante o uso de pesos em relação aos alinhamentos filtrados e sem filtro. A variação foi verificada somente para os resíduos mais conservados nas posições que correspondem a sítios de ligação a ligantes ou metais no sub-alinhamento referente ao grupo ortólogo das transtirretinas (Lys15, Lys31, Lys70, Glu72, Asp74, His88, Glu89, Tyr90 e Asp92).

A seguir, dois sub-alinhamentos foram gerados a partir das sequências identificadas como sendo ortólogas pelo método OrthoMCL (LI *et al.*, 2003): um para o grupo ortólogo das 5-hidroxi-isourato hidrolases (HIUase) contendo 3692 das 3893 sequências do alinhamento original, e um para o grupo ortólogo das transtirretinas, contendo apenas 172 sequências. A figura 3 compara os dois novos perfis com o perfil do alinhamento total. A partir desta figura é possível observar o quanto o perfil do grupo das HIUases contribui para o perfil do alinhamento total, enquanto o perfil das transtirretinas se mostra bastante destoante, predominando aminoácidos carregados. Uma posição em especial, a que corresponde ao resíduo 15, se destaca nos perfis. Nela predominam apenas dois aminoácidos: histidina (His, H), nas HIUases, e lisina (Lys, K) nas TTR. Como pode ser visto na figura 2, é justamente na lisina 15 que ocorre o maior aumento de frequência mediante o uso de pesos, o que mostra que de fato essa abordagem pode destacar características específicas de sub-famílias pouco amostradas, como é o caso das TTR.

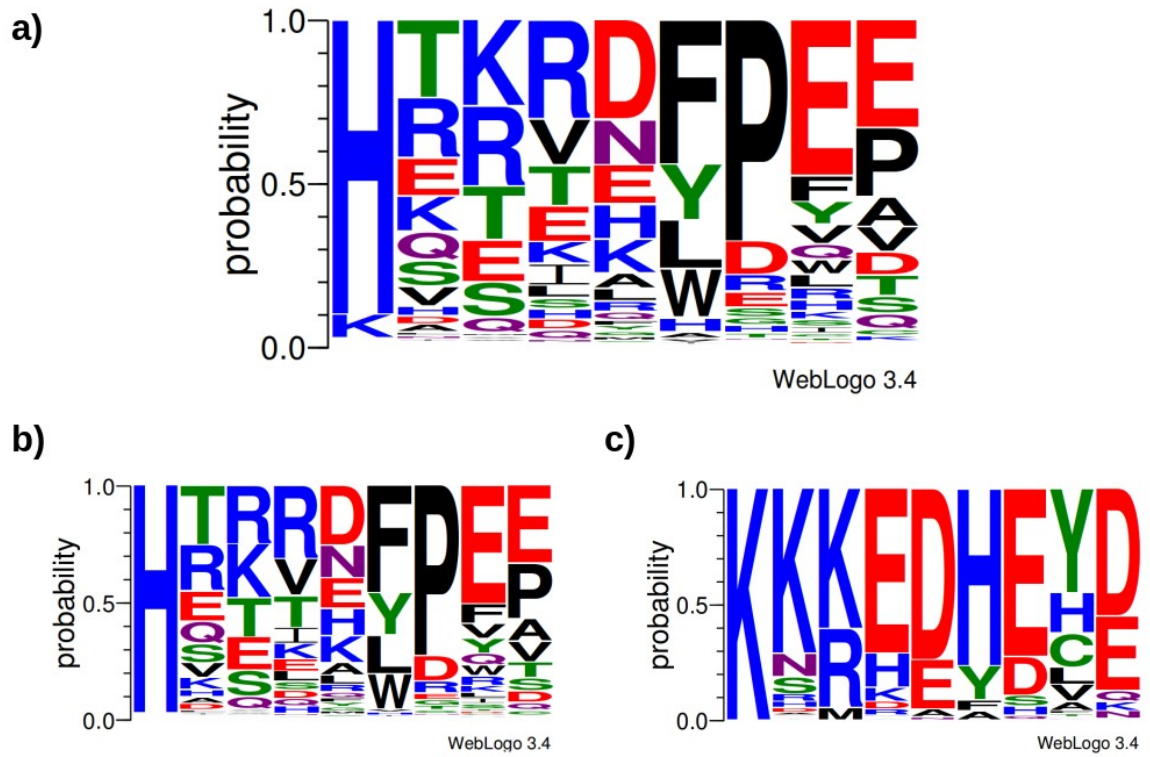

Figura 3: Comparação dos perfis de frequências dos grupos ortólogos em relação ao perfil do alinhamento total. (a) Alinhamento total; (b) 5-hidroxi-isourato hidrolases e (c) transtirretinas. As logos foram geradas pelo aplicativo online WebLogo (CROOKS *et al.*, 2004).

5.1.2 – Correlação entre posições

Uma vez identificados os resíduos típicos de TTR localizados em sítios de ligação, foi analisado o quão correlacionados (ou anti-correlacionados) eles estariam entre si, ou seja, o quanto a presença de um estaria aumentando (ou diminuindo) a frequência do outro. Primeiramente foi feita uma análise visual por meio de logos representando os perfis dos sub-alinhamentos contendo cada um desses resíduos, os quais podem ser vistos na figura 4.

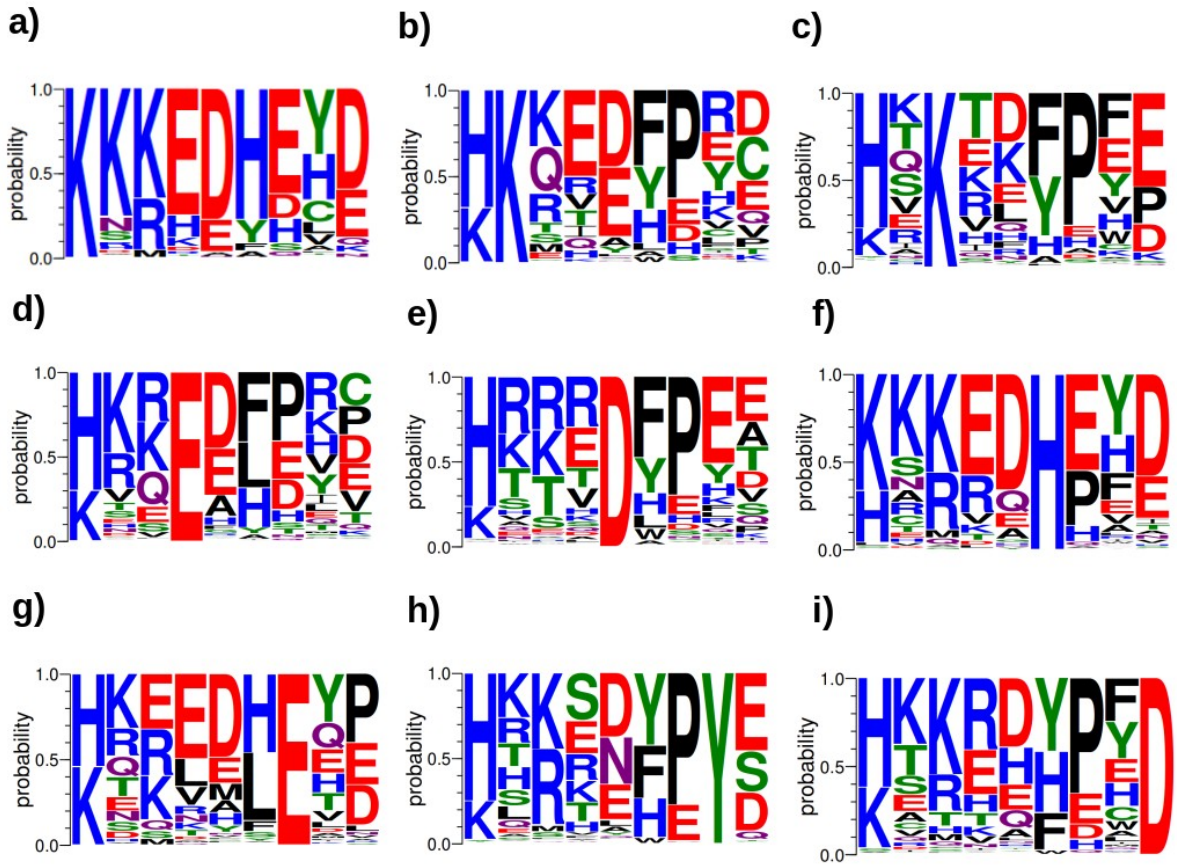

Figura 4: Perfis de frequências dos sub-alinhamentos para os quais todas as sequências contém, respectivamente: (a) lisina na posição 15; (b) lisina na posição 31; (c) lisina na posição 70; (d) glutamato na posição 72; (e) aspartato na posição 74; (f) histidina na posição 88; (g) glutamato na posição 89; (h) tirosina na posição 90 e (i) aspartato na posição 92.

Alguns resíduos se destacam por interferirem de maneira mais significativa nas frequências de outros e por retratarem um perfil semelhante ao do grupo ortólogo das TTR, sendo o principal a própria lisina 15 (fig. 4a) já mencionada anteriormente, seguida pela histidina 88 (fig. 4f). Vale ressaltar aqui que o perfil do sub-alinhamento cujas todas as sequências apresentam lisina na posição 15 (fig. 4a) é praticamente idêntico ao perfil do sub-alinhamento do grupo ortólogo das TTR (fig. 3c), sugerindo que este resíduo seja altamente específico deste grupo, podendo ser inclusive caracterizado como SDP, da mesma forma que a histidina na posição 15 também o seria para o grupo das HIUases.

A seguir, a variação na frequência de cada resíduo foi calculada mediante a presença de cada um dos outros resíduos. A figura 5 mostra dois perfis acompanhados de um gráfico mostrando as frequências no alinhamento total e no sub-alinhamento: um referente ao sub-alinhamento contendo somente lisina na posição 15 e outro contendo histidina na mesma

posição. Estes perfis foram escolhidos devido ao fato de que Lys15 e His15 são resíduos mualmente exclusivos, portanto representariam os casos mais nítidos de correlação (fig. 5a) e anti-correlação (fig. 5b).

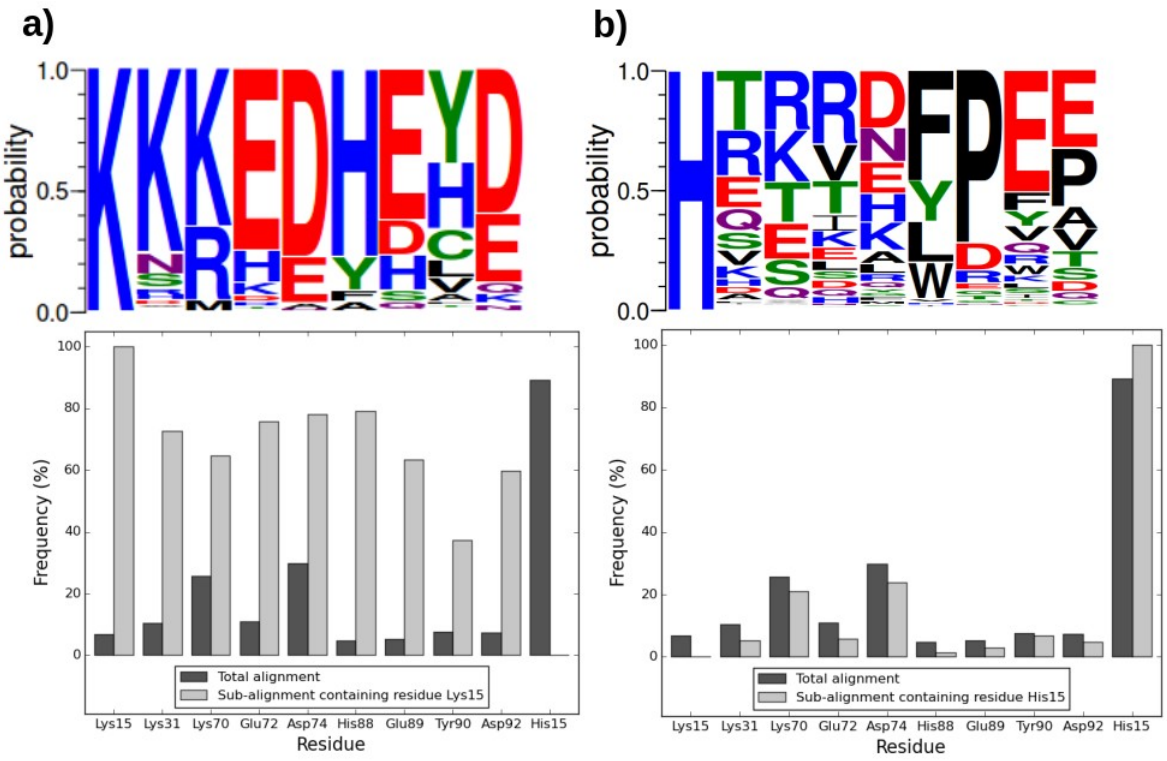

Figura 5: Perfis de frequências dos sub-alinhamentos para os quais todas as sequências contém, respectivamente: (a) lisina na posição 15; (b) histidina ao invés de lisina na mesma posição. Os gráficos abaixo das logos mostram as frequências dos resíduos típicos de TTR para o alinhamento total e para o sub-alinhamento contendo lisina ou histidina na posição 15.

Conforme a variação das frequências mediante a presença de lisina ou histidina na posição 15 satisfazia os critérios para que houvesse correlação ou anti-correlação, o *p-value* era calculado e alguns pares (anti-)correlacionados emergiam das análises. A figura 6 mostra uma pequena rede de correlações formada por pares cuja influência mútua sobre suas frequências era significativa.

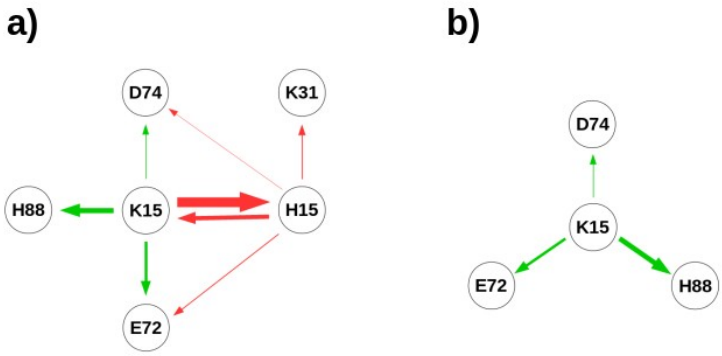

Figura 6: Redes de correlação e anti-correlação entre resíduos compreendidos em posições referentes a sítios de ligação em TTRs. A largura das arestas remete à intensidade da correlação (setas verdes) ou anti-correlação (setas vermelhas). (a) Rede completa. (b) Comunidade específica de TTR vista isoladamente.

Como já era possível observar na figura 4, os resíduos que formam a rede de correlação da figura 6 já esboçavam um maior aumento na frequência devido à presença de lisina na posição 15 ou diminuição na presença de histidina nessa mesma posição. Quando mapeados na estrutura quaternária da TTR (PDB ID 3grb, vide figura 7), para cada uma das quatro cadeias, esses resíduos aparecem em dois sítios distintos: um (Lys15) na entrada do canal onde se liga o hormônio tireoidiano (T3 ou T4), com o qual pode eventualmente formar uma ponte salina (PINTO *et al.*, 2011), e os outros três (Glu72, Asp74 e H88) em um dos sítios de ligação a zinco já descritos (PALMIERI *et al.*, 2010), sendo importante para o reconhecimento de proteínas carreadores de retinol (RBPs, do inglês “retinol binding protein”) por parte das TTRs, bem como na formação de fibrilas amilóides (PALMIERI *et al.*, 2010).

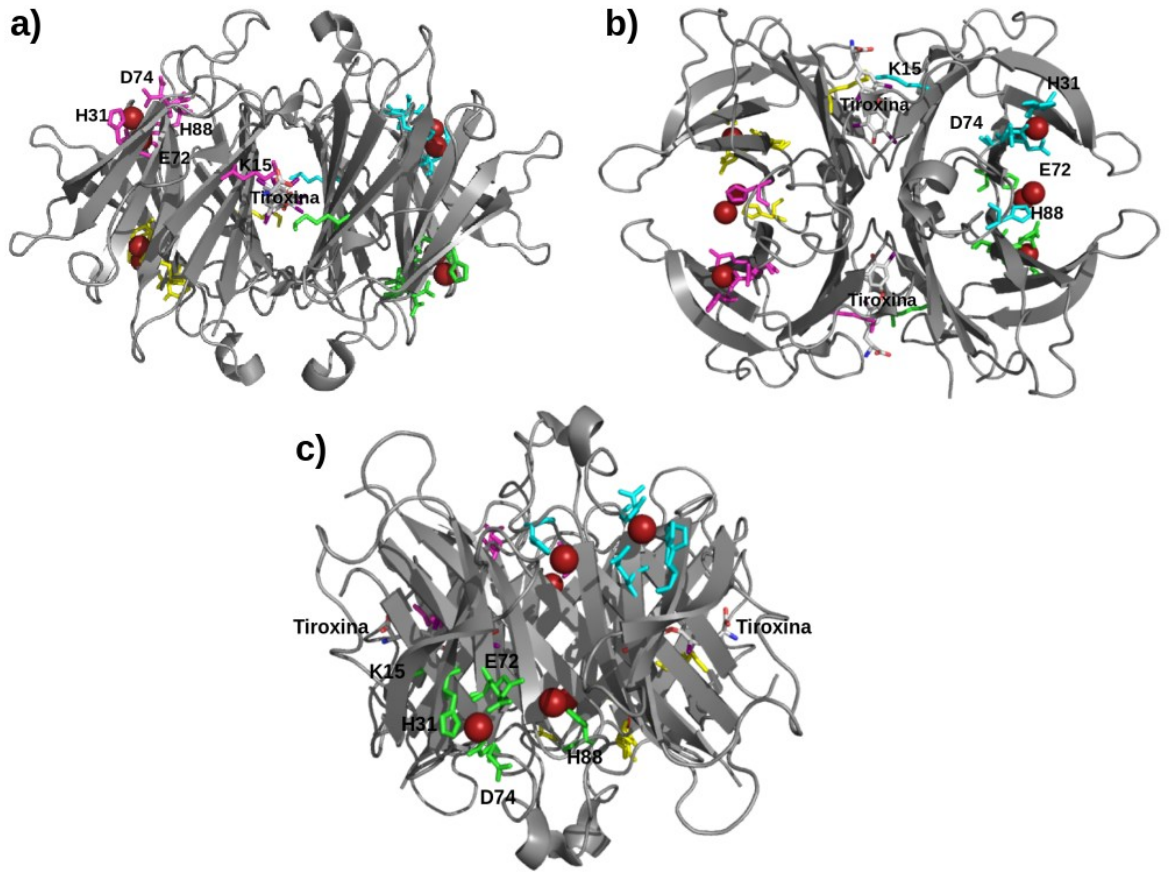

Figura 7: Estrutura quaternária da transtirretina humana (PDB ID 3grb) com destaque para os resíduos encontrados na análise de correlação (exceto pela posição 31, que na transtirretina humana apresenta uma histidina no lugar de lisina). A molécula de tiroxina (T4), entretanto, foi extraída de outra estrutura da mesma proteína (PDB ID 2rox), a qual foi alinhada à primeira para se obter a posição aproximada do ligante. (a) Vista do túnel onde se liga o hormônio tireoidiano (no caso, tiroxina) onde encontram-se quatro resíduos de lisina (Lys15), cada um pertencente a uma cadeia distinta. (b) Vista superior do túnel, de onde vê-se melhor a disposição das duas moléculas de tiroxina. (c) Vista lateral do túnel, de onde se vêem melhor os resíduos (Glu72, Asp74 e His88) que coordenam dois átomos átomo de zinco (esferas marrom-avermelhadas). Os resíduos foram coloridos de acordo com as cadeias às quais pertencem (verde, ciano, amarelo e magenta representando, respectivamente, as cadeias A, B, C e D).

5.1.3 – Conservação nos sub-alinhamentos

A figura 8a mostra a distribuição das maiores frequências, para posições com menos de 20% de *gaps*, no sub-alinhamento referente ao grupo ortólogo das HIUases. A figura 8b mostra a distribuição da entropia de Shannon, para as mesmas posições no mesmo sub-alinhamento. Nota-se que a maioria das posições tem seu aminoácido mais conservado com frequência entre 20 e 30%, ou seja, a maioria das posições apresentam-se pouco conservadas. Isto também pode ser observado pela distribuição da Entropia de Shannon, em

que a maioria das posições apresenta entropia alta, ou seja, são de fato pouco conservadas. Um sub-alinhamento como este dificilmente apresentaria riscos às análises de correlação, pois seria improvável que as correlações detectadas fossem devidas ao efeito filogenético.

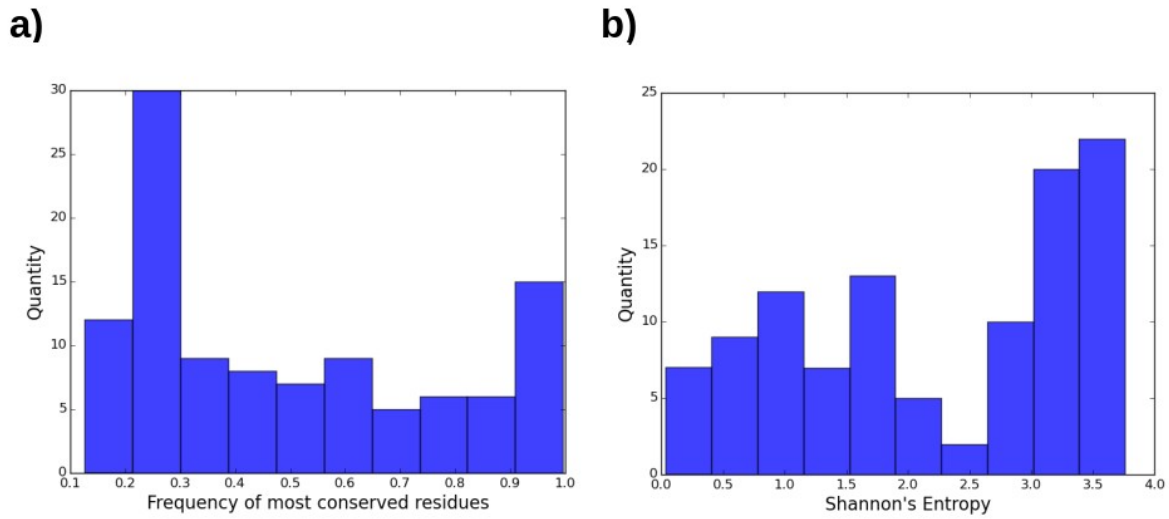

Figura 8: (a) Distribuição das frequências dos aminoácido mais conservados para posições com menos de 20% de *gaps* no sub-alinhamento referente ao grupo ortólogo das HIUases. (b) Distribuição da Entropia de Shannon também para posições com menos de 20% de *gaps* no mesmo sub-alinhamento.

A figura 9, por sua vez, mostra o mesmo que a figura 8, porém para o sub-alinhamento referente ao grupo ortólogo das TTR. Nota-se exatamente o oposto do observado no sub-alinhamento do grupo ortólogo das HIUases, pois aqui a maioria das posições apresentam frequências acima de 90% e entropia baixa, havendo então a predominância de posições conservadas. Devido ao surgimento recente dessa sub-família na história evolutiva, não teria havido tempo suficiente para a diversificação dessas proteínas, o que se reflete na alta identidade entre as sequências desse sub-alinhamento. Sendo assim, as correlações que eventualmente fossem detectadas não seriam devidas a um evento de coevolução entre resíduos, mas sim ao simples fato de que todas as sequências são muito parecidas entre si, ou seja, as correlações seriam devidas ao efeito filogenético.

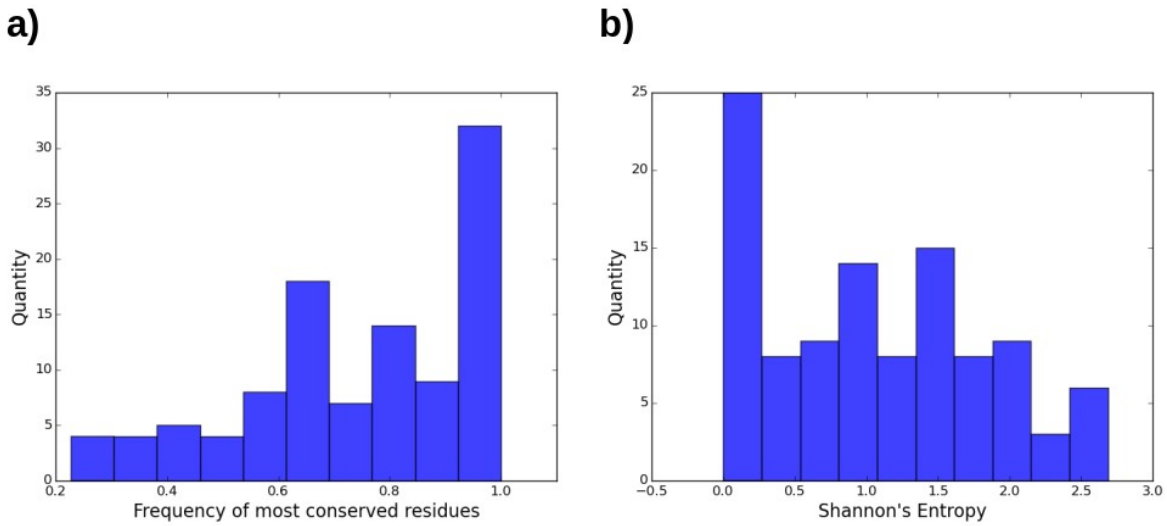

Figura 9: (a) Distribuição das frequências dos aminoácido mais conservados para posições com menos de 20% de *gaps* no sub-alinhamento referente ao grupo ortólogo das TTR. (b) Distribuição da Entropia de Shannon também para posições com menos de 20% de *gaps* no mesmo sub-alinhamento.

5.1.4 – Novas análises de correlação

Para verificar se as correlações encontradas anteriormente são significativas ou se são somente consequência do efeito filogenético, os cálculos de correlação foram refeitos da mesma forma, exceto pela probabilidade *a priori* da distribuição binomial cumulativa, que agora seria substituída pela média das frequências mais altas no sub-alinhamento. As novas correlações, ou melhor, anti-correlações são mostradas na figura 10.

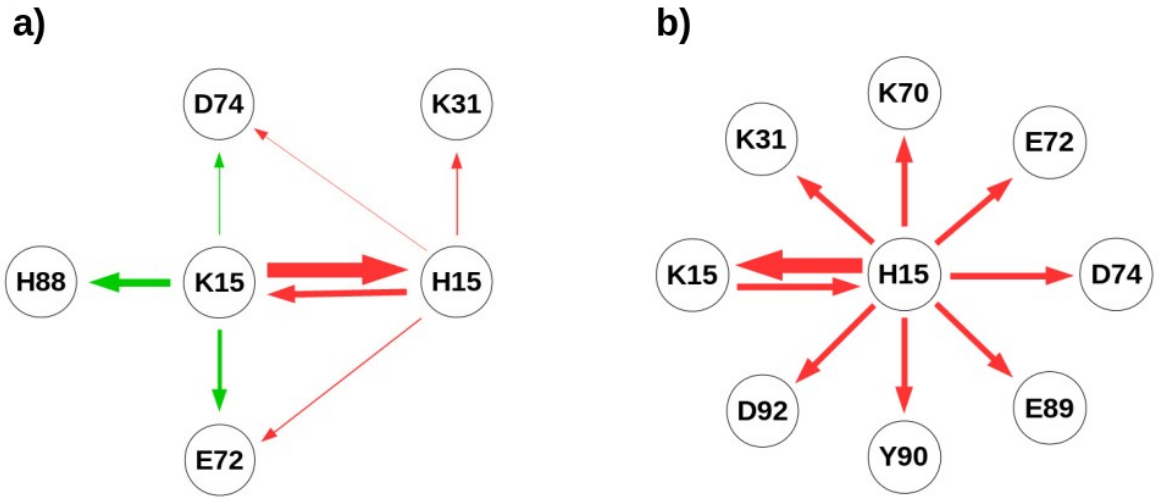

Figura 10: Redes de correlação e anti-correlação entre resíduos compreendidos em sítios de ligação em TTRs. A largura das arestas remete à intensidade da correlação (setas verdes) ou anti-correlação (setas vermelhas). (a) Rede de correlações encontrada anteriormente. (b) Nova rede de (anti-)correlações obtida após levar em consideração o efeito filogenético, conforme discutido no texto.

A partir da figura 10 é possível observar que os resíduos típicos de TTR, embora todos estejam anti-correlacionados com a Hys15, típica de HIUase, eles não mais apresentam-se correlacionados entre si, sugerindo que as correlações positivas encontradas anteriormente não se devessem a uma possível coevolução entre os resíduos de modo a atribuir função às TTR, mas sim porque as TTR apresentam alta similaridade entre si. Não que esses resíduos não tenham importância funcional, afinal eles de fato encontram-se em sítios de ligação e têm lá o seu papel na molécula. O fato é que no presente trabalho as análises foram direcionadas somente para posições específicas selecionadas manualmente, por motivos operacionais. Se ao invés disso as análises tivessem sido direcionadas a todas as posições do alinhamento, provavelmente muito mais resíduos presentes somente nas TTR seriam identificados como correlacionados. Portanto, seria necessário um tempo de divergência muito maior entre as sequências para que fosse possível identificar, dentre todas as correlações observadas, quais seriam realmente devidas a características funcionais e quais seriam devidas somente ao efeito filogenético.

5.2 – Receptores Nucleares – Domínio de ligação a DNA

5.2.1 – Efeitos da atribuição de pesos sobre as frequências

A figura 11 mostra as representações gráficas (logos) dos perfis de frequências das posições correspondentes a sítios específicos (*P-boxes*) de reconhecimento de regiões bem definidas no DNA, os elementos de resposta a hormônio (HRE). Além dos resíduos compreendidos na região *P-box* propriamente dita (posições 152 a 157 da sequência referência), foram selecionados alguns resíduos localizados à esquerda (sentido N-terminal) e à direita (sentido C-terminal) que também interagem com o DNA (vide figura 21) conforme descrito recentemente (AFONSO *et al.*, 2013). É possível observar o quão drástico é o efeito dos filtros sobre o perfil de frequências, comparado com o efeito dos pesos. Isto pode ser melhor visualizado na figura 12, que mostra a variação, em relação aos alinhamentos filtrados e sem filtro, nas frequências em cada posição mediante a atribuição de pesos às sequências. É possível observar diferentes efeitos sobre resíduos mutualmente exclusivos, localizados nas mesmas posições. De maneira geral, observa-se efeitos opostos em relação ao alinhamento original e a aos alinhamentos filtrados. Em outras palavras, quando a variação em relação ao alinhamento original é positiva, esta será negativa em

relação aos alinhamentos filtrados, e vice versa. Isso significa que a atribuição de pesos gera frequências intermediárias entre aquelas observadas no alinhamento original e nos alinhamentos filtrados. Tais resultados demonstram que o uso de pesos é capaz tanto de destacar características pouco representadas, quanto de mascarar características redundantes. A partir daqui, somente os alinhamentos ponderados foram utilizados nas análises.

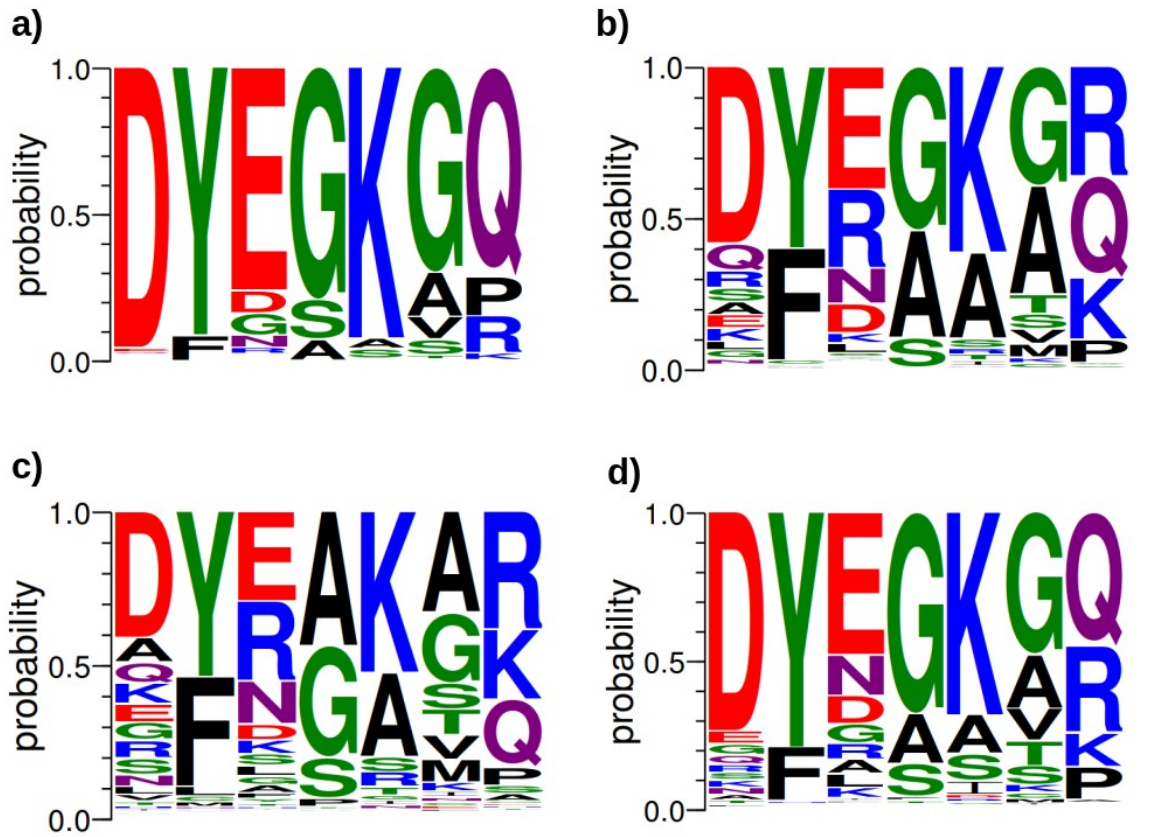

Figura 11: Representações gráficas (logos) dos perfis de frequências (alinhamento total) para as posições compreendidas em sítios de reconhecimento de HRE, os *P-box* (resíduos 152 a 157), e outros resíduos importantes. As colunas das logos (da esquerda para a direita) correspondem às posições 140, 147, 153, 154, 156, 157 e 188 da sequência referência, o receptor nuclear humano hRXRα (UniProtKB ID: P19793, PDB 2NLL). Os perfis se referem, respectivamente, aos alinhamentos: (a) sem pesos e sem filtro; (b) filtrado por identidade máxima de 80%; (c) filtrado por identidade máxima de 70%; (d) ponderado. As logos foram geradas pelo aplicativo online WebLogo (CROOKS *et al.*, 2004).

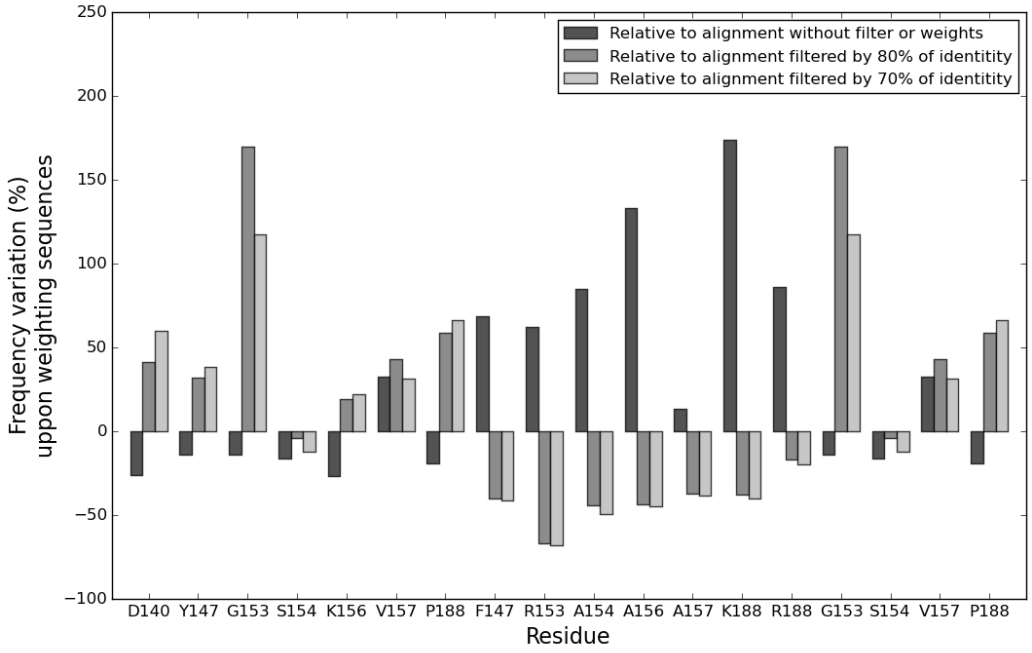

Figura 12: Variação nas frequências mediante o uso de pesos em relação aos alinhamentos filtrados e sem filtro. A variação foi verificada para os resíduos da região *P-box* (e outros resíduos importantes) de humanos e afins (Asp140, Y147, Glu153, Gly154, Lys156, Gly157, Gln188), de nematódeos (Phe147, Arg153, Ala154, Ala156, Ala157 e Lys/Arg188) e receptores esteroidais (Asp140, Y147, Gly153, Ser154, Lys156, Val157 e Pro188).

A seguir, dois sub-alinhamentos foram gerados a partir das sequências identificadas como sendo ortólogas pelo método OrthoMCL (LI *et al.*, 2003): um para o grupo ortólogo dos NRs de humanos (e semelhantes) contendo 698 das 10583 sequências do alinhamento original, e um para o grupo ortólogo dos receptores de 3-cetoesteroides, contendo apenas 131 sequências. As sequências contendo o *P-box* típico de nematódeos foram identificadas como pertencentes a diferentes grupos ortólogos, talvez devido à explosão de duplicações gênicas a partir de um mesmo receptor nuclear ancestral ocorrida neste clado (ROBINSON-RECHAVI *et al*., 2005; AFONSO *et al.,* 2013). Em vez de um único grupo ortólogo para receptores nucleares de nematódeos, um sub-alinhamento contendo o *P-box* típico deste grupo ($CR^{153}A^{154}CA^{156}A^{157}$) foi gerado somente para fins de comparação. A figura 13 compara os três novos perfis com o perfil do alinhamento total. O perfil do grupo dos NRs de humanos (e similares) se destaca no perfil do alinhamento total, com seus aminoácidos sendo os mais conservados. Aqui nota-se que tanto os NRs de humanos (e similares) quanto os receptores de 3-cetoesteroides apresentam Asp140, Tyr147 e Lys156, o que representa

um desafio para a identificação de comunidades específicas de cada grupo. Vale notar também que na posição 147 predominam dois aminoácidos com propriedades físico-químicas semelhantes (Tyr e Phe), podendo então ser considerada como uma posição relativamente conservada para a maioria das sequências. Além disso, o resíduo Pro188, por exemplo, presente em receptores de 3-cetoesteroides, não poderia ter sido identificado em análises de correlação utilizado filtros por identidade máxima, tais como foram identificados os resíduos Asp140, Tyr147 e Gln188 nos NRs de humanos e similares, da mesma forma que o resíduo Phe147 nos NRs de nematódeos (AFONSO *et al.*, 2013). Isso se deve à pouca quantidade de sequências dos receptores de 3-cetoesteroides no alinhamento total.

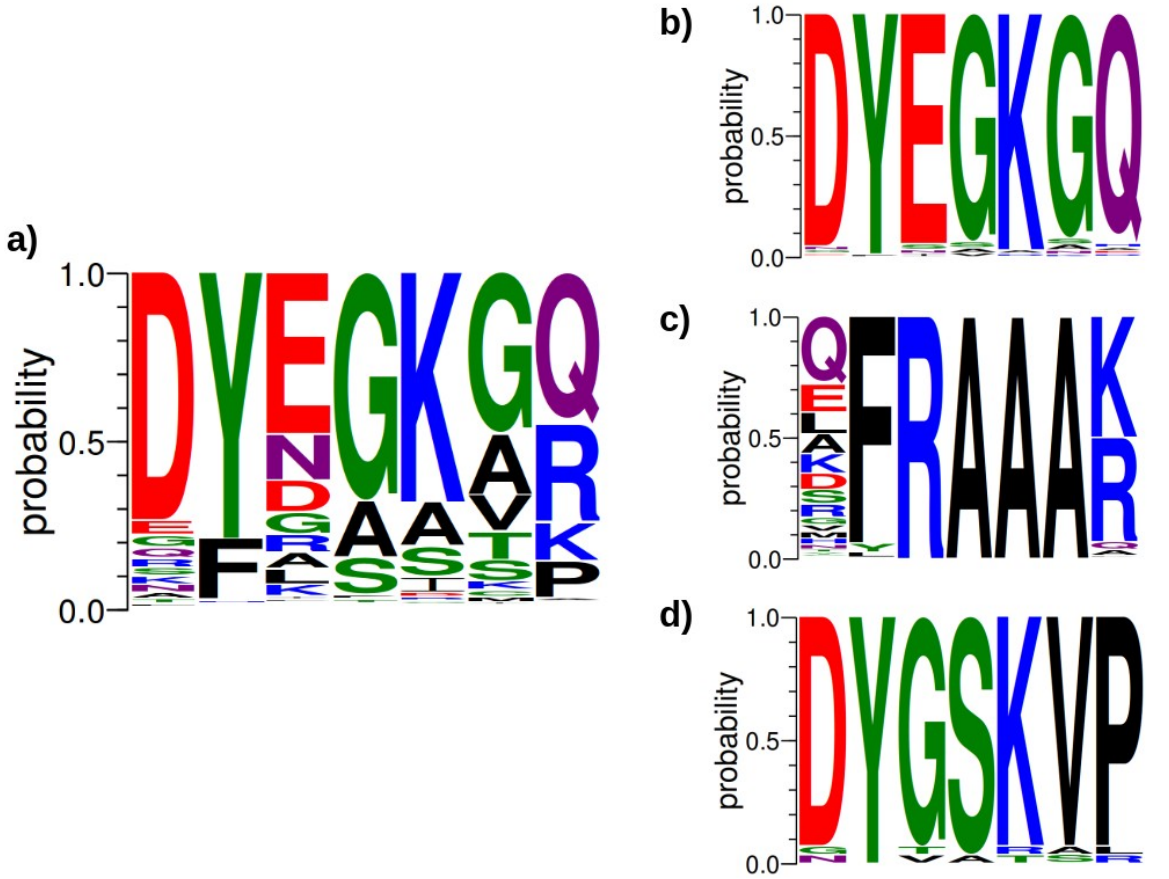

Figura 13: Comparação entre os perfis de frequências dos grupos ortólogos em relação ao perfil do alinhamento total. (a) Alinhamento total; (b) *P-box* (e outros resíduos) de receptores nucleares de humanos (e semelhantes); (c) *P-box* (e outros resíduos) de receptores nucleares de *C. elegans* e outros nematódeos*;* e (d) *P-box* (e outros resíduos) de receptores de 3-cetoesteroides. As logos foram geradas pelo aplicativo online WebLogo (CROOKS *et al.*, 2004). (*) Como as sequências contendo o *P-box* ($CR^{153}A^{154}CA^{156}A^{157}$) de *C. elegans* (b) foram agrupadas em diferentes grupos ortólogos, utilizou-se aqui um sub-alinhamento contendo este perfil, independente do agrupamento obtido por meio do programa OrthoMCL (LI *et al.*, 2003).

5.2.2 – Correlação entre posições

Uma vez de posse dos resíduos localizados nos *P-boxes* (ou próximos a estes) para cada um dos grupos (receptores de 3-cetoesteroides, NRs de nematódeos e NRs de humanos e afins), muitos destes resíduos inclusive sendo mutualmente exclusivos, foi analisado o quão correlacionados (ou anti-correlacionados) eles estariam entre si, ou seja, o quanto a presença de um estaria aumentando (ou diminuindo) a frequência do outro. Primeiramente foi feita uma análise visual por meio de logos representando os perfis dos sub-alinhamentos contendo cada um desses resíduos, os quais podem ser vistos nas figuras 14, 15 e 16.

Na figura 14 temos os perfis para os sub-alinhamentos que possuem algum dos resíduos associados ao motivo *P-box* de humanos e afins. Nota-se que os perfis são muito parecidos entre si, e que todos eles se assemelham ao perfil referente ao grupo ortólogo dos NRs de humanos (figura 13b), o que sugere que os resíduos destacados estejam fortemente correlacionados entre si.

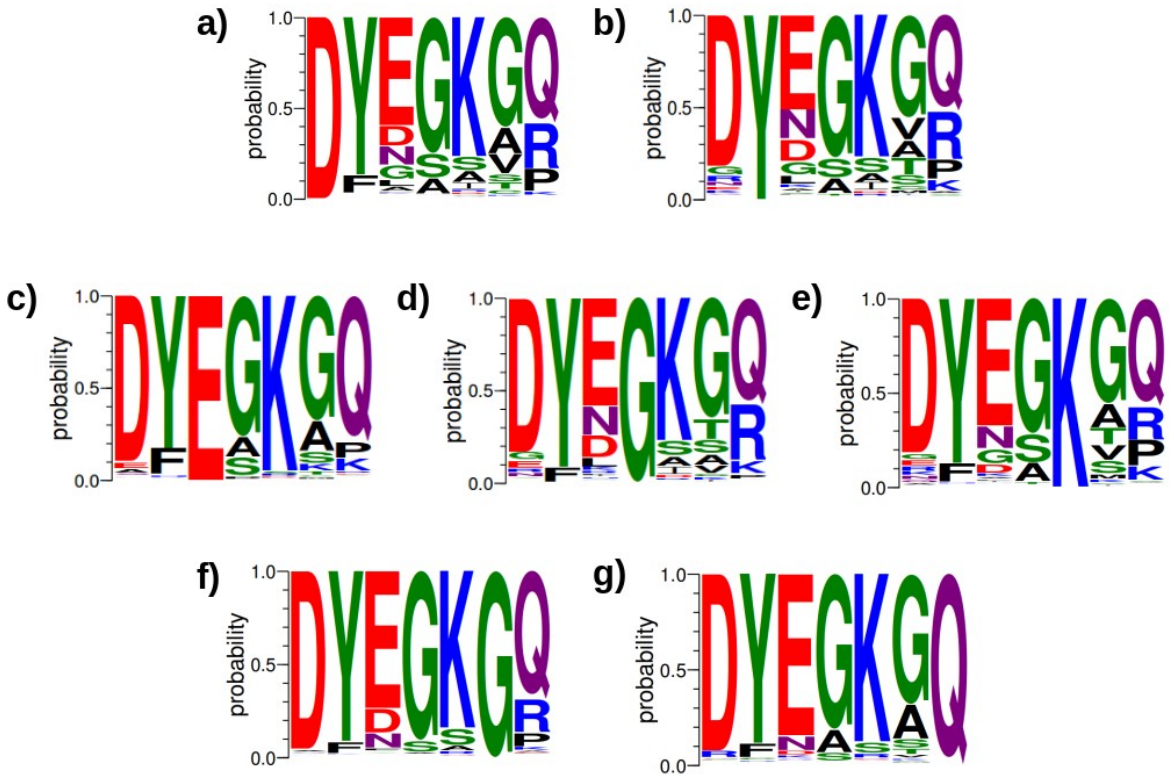

Figura 14: Representações gráficas (logos) dos perfis de frequências da região P-box (e outros resíduos) para sub-alinhamentos cujas todas as sequências possuem algum dos resíduos típicos de NRs de humanos e afins. Os perfis correspondem a sub-alinhamentos contendo, respectivamente: (a) aspartato (Asp, D) na posição

140; (b) tirosina (Tyr, Y) na posição 147; (c) glutamato (Glu, E) na posição 153; (d) glicina (Gly, G) na posição 154; (e) lisina (Lys, K) na posição 156; (f) glicina (Gly, G) na posição 157; e (g) glutamina (Gln, Q) na posição 188.

A figura 15 mostra os perfis para os sub-alinhamentos que contenham algum dos resíduos típicos de NRs de *C. elegans*. Ao contrário do que ocorre nos NRs de humanos, a presença de cada resíduo gera sub-alinhamentos não muito parecidos entre si, tampouco com o sub-alinhamento contendo o *P-box* típico de nematódeos ($R^{153}A^{154}A^{156}V^{157}$), exceto pelos resíduos R153 (figura 15b) e A156 (figura 15d), cujos perfis também apresentam os demais resíduos típicos.

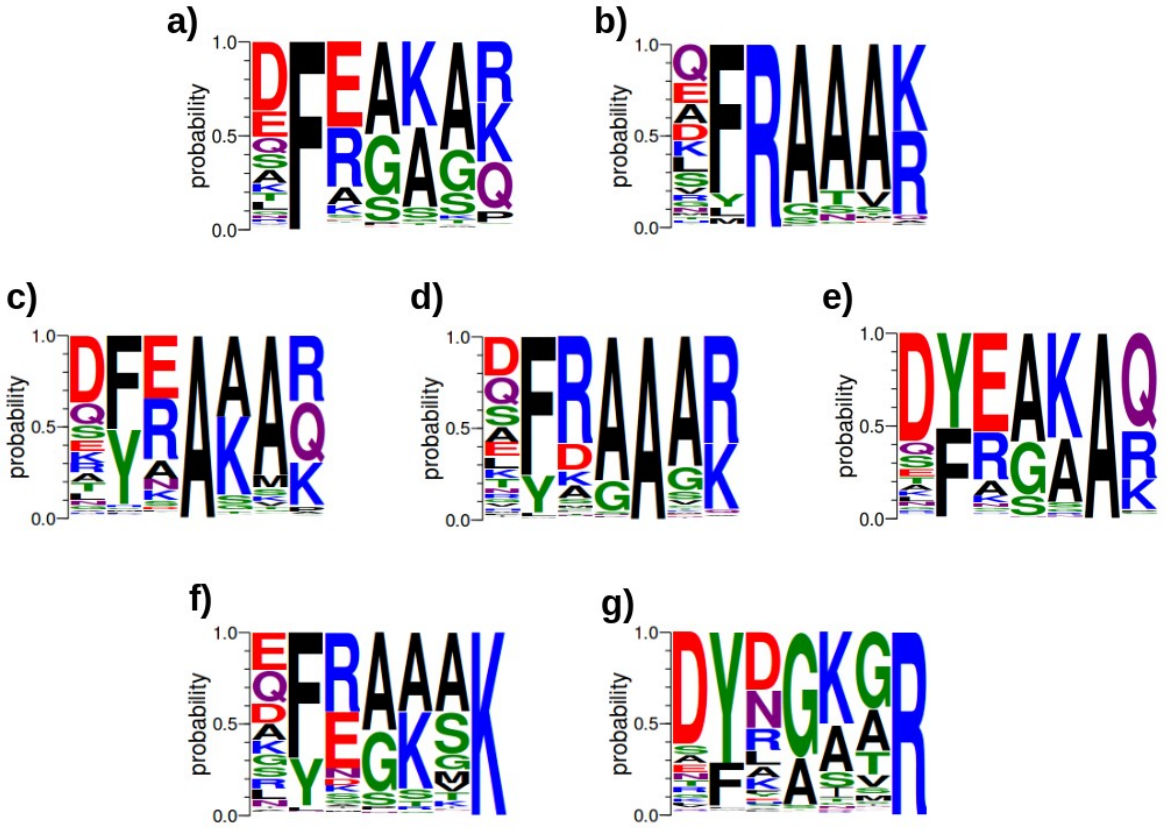

Figura 15: Representações gráficas (logos) dos perfis de frequências da região P-box (e outros resíduos) para sub-alinhamentos cujas todas as sequências possuem algum dos resíduos típicos de NRs de nematódeos. Os perfis correspondem a sub-alinhamentos contendo, respectivamente: (a) fenilalanina (Phe, F) na posição 147; (b) arginina (Arg, R) na posição 153; (c) alanina (Ala, A) na posição 154; (d) alanina (Ala, A) na posição 156; (e) alanina (Ala, A) na posição 157; (f) lisina (Lys, K) na posição 188; e (g) arginina (Arg, R) na posição 188.

Já a figura 16 mostra os perfis para os sub-alinhamentos que contenham algum dos resíduos típicos de receptores de 3-cetoesteroides. Exceto pelos perfis dos sub-

alinhamentos contendo Asp140, Tyr147 e Lys156 (figura 16a, 16b e 16e) os quais também estão presentes em NRs de humanos e com isso acabam tendo mais a ver com estes do que com o grupo ortólogo dos receptores de 3-cetoesteroides, os demais resíduos também parecem estar correlacionados entre si, mesmo que em menor grau.

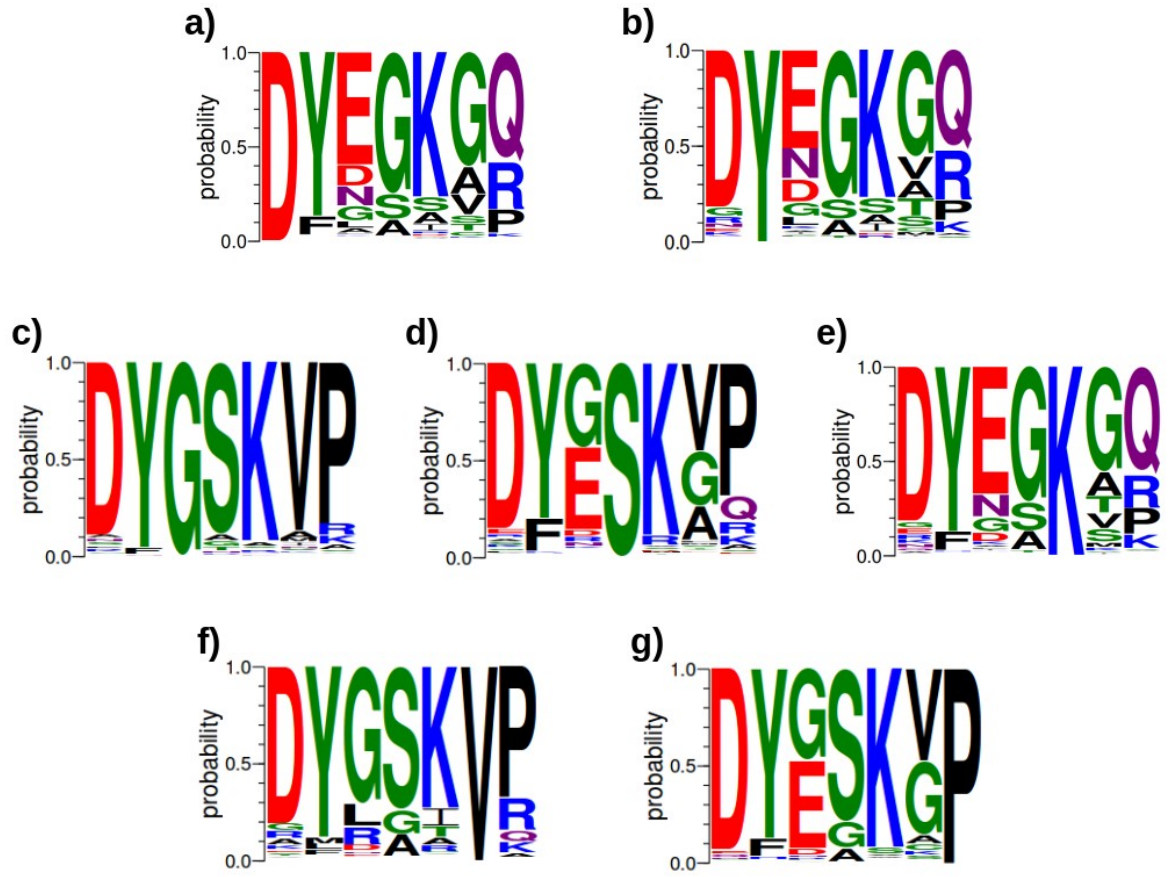

Figura 16: Representações gráficas (logos) dos perfis de frequências da região P-box (e outros resíduos) para sub-alinhamentos cujas todas as sequências possuem algum dos resíduos típicos de receptores de 3-cetoesteroides. Os perfis correspondem a sub-alinhamentos contendo, respectivamente: (a) aspartato (Asp, D) na posição 140; (b) tirosina (Tyr, Y) na posição 147; (c) glicina (Gly, G) na posição 153; (d) serina (Ser, S) na posição 154; (e) lisina (Lys, K) na posição 156; (f) valina (Val, V) na posição 157; e (g) prolina (Pro, P) na posição 188.

A seguir, a variação na frequência de cada resíduo foi calculada mediante a presença de cada um dos outros resíduos. Os gráficos das figuras 17, 18 e 19 mostram as frequências no alinhamento total e nos sub-alinhamentos contendo um determinado aminoácido na posição 153. Esta posição foi escolhida pois os aminoácidos localizados nessa posição parecem gerar sub-alinhamentos mais próximos de cada grupo ortólogo.

Na figura 17 Observa-se um aumento na frequência dos resíduos típicos de NRs de humanos, ao passo que a frequência dos demais aminoácidos tende a diminuir ou permanecer a mesma. O mesmo ocorre na figura 18, porém agora para os resíduos típicos de NRs de nematódeos. A figura 19, por sua vez, mostra que alguns resíduos típicos de NRs de humanos (Asp140, Tyr147 e Lys156), por estarem presentes em ambos os grupos, também têm sua frequência aumentada na presença de resíduos típicos de receptores de 3-cetoesteroides.

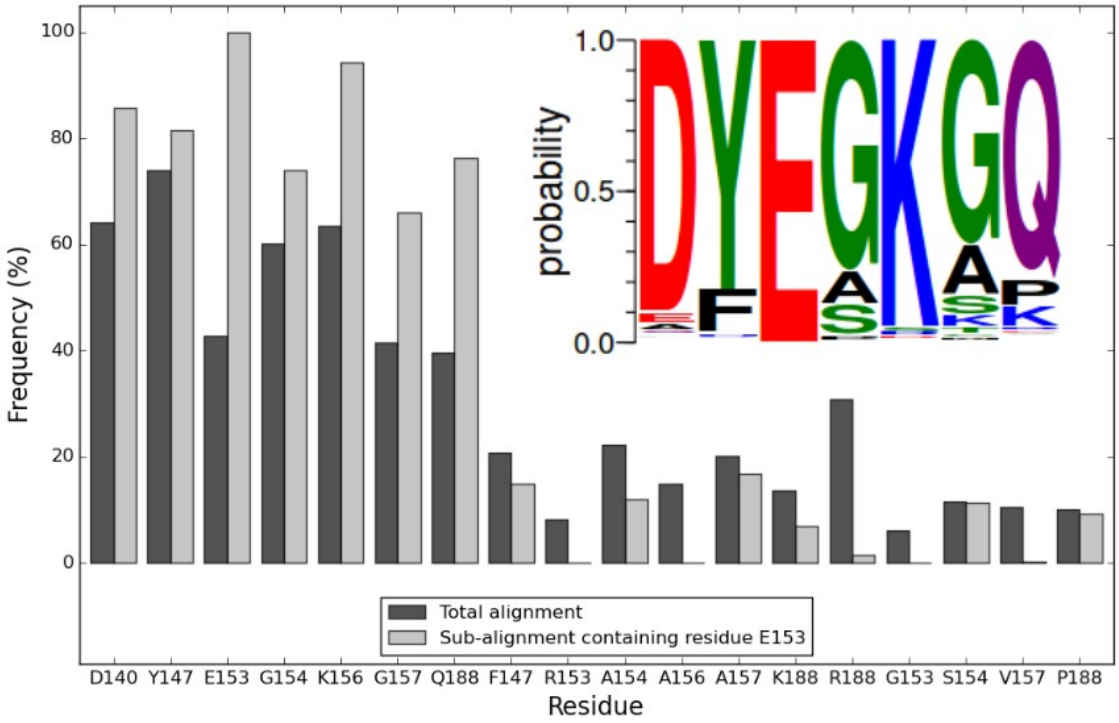

Figura 17: Frequências de resíduos típicos de *P-box* (e outros resíduos) no alinhamento total e no sub-alinhamento contendo apenas glutamato (Glu, E) na posição 153. A logo é a mesma da figura 14c e representa o perfil de frequências para o referido sub-alinhamento.

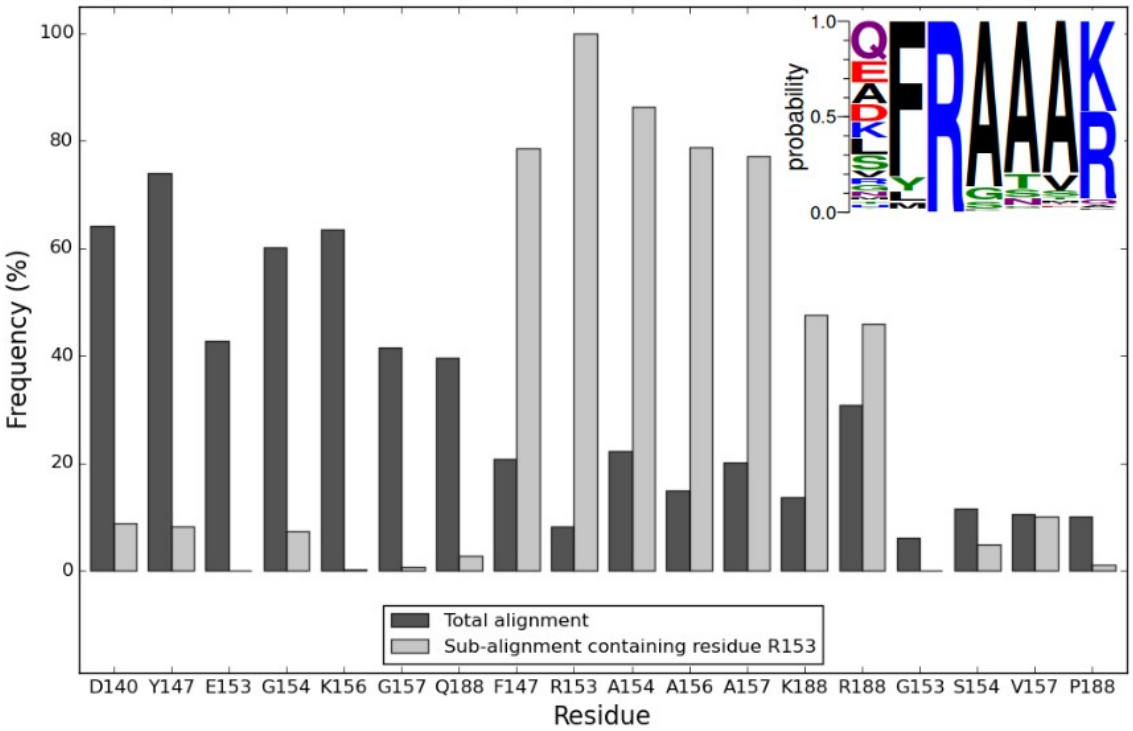

Figura 18: Frequências de resíduos típicos de *P-box* (e outros resíduos) no alinhamento total e no sub-alinhamento contendo apenas arginina (Arg, R) na posição 153. A logo é a mesma da figura 15b e representa o perfil de frequências para o referido sub-alinhamento.

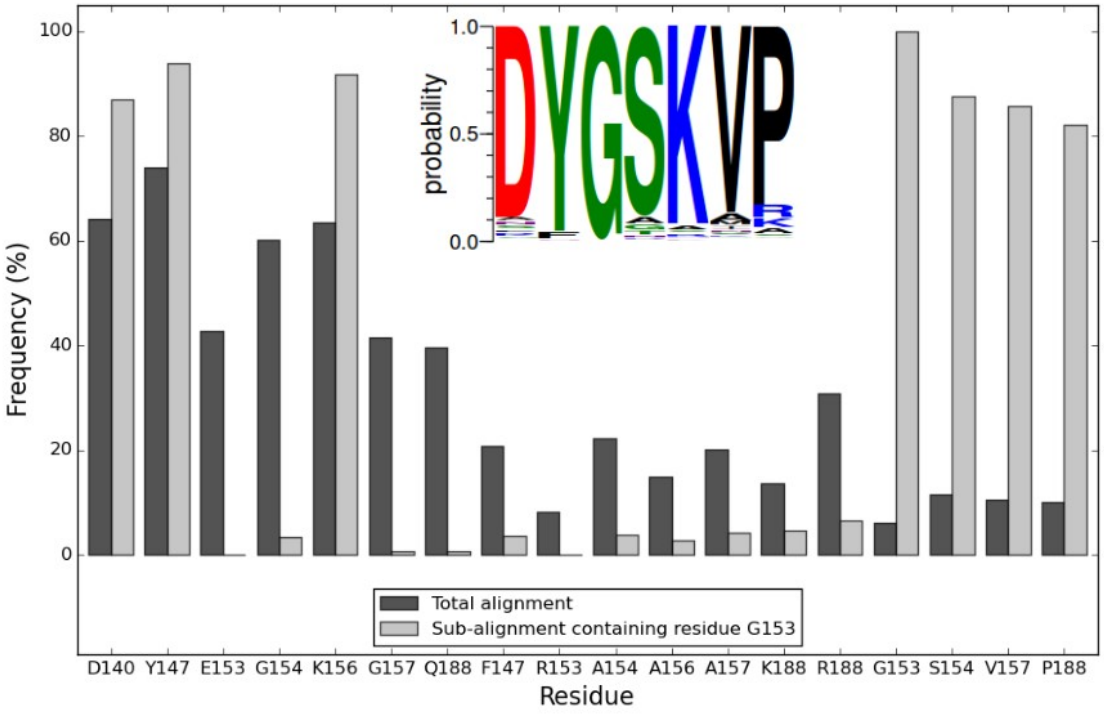

Figura 19: Frequências de resíduos típicos de *P-box* (e outros resíduos) no alinhamento total e no sub-alinhamento contendo apenas glicina (Gly, G) na posição 153. A logo é a mesma da figura 16c e representa o perfil de frequências para o referido sub-alinhamento.

Conforme as variações das frequências mediante a presença de resíduos típicos de um grupo ou outro satisfaziam os critérios para que houvesse correlação ou anti-correlação, o *p-value* era calculado e alguns pares (anti-)correlacionados emergiam das análises. A figura 20 mostra uma pequena rede de correlações formada por pares cuja influência mútua sobre suas frequências era significativa. Entretanto, devido ao fato de que há resíduos diferentes nas mesmas posições, ou seja, mutualmente exclusivos, o alto número de anti-correlações inviabiliza a visualização, tendo sido mostrados na figura 20 somente as correlações positivas.

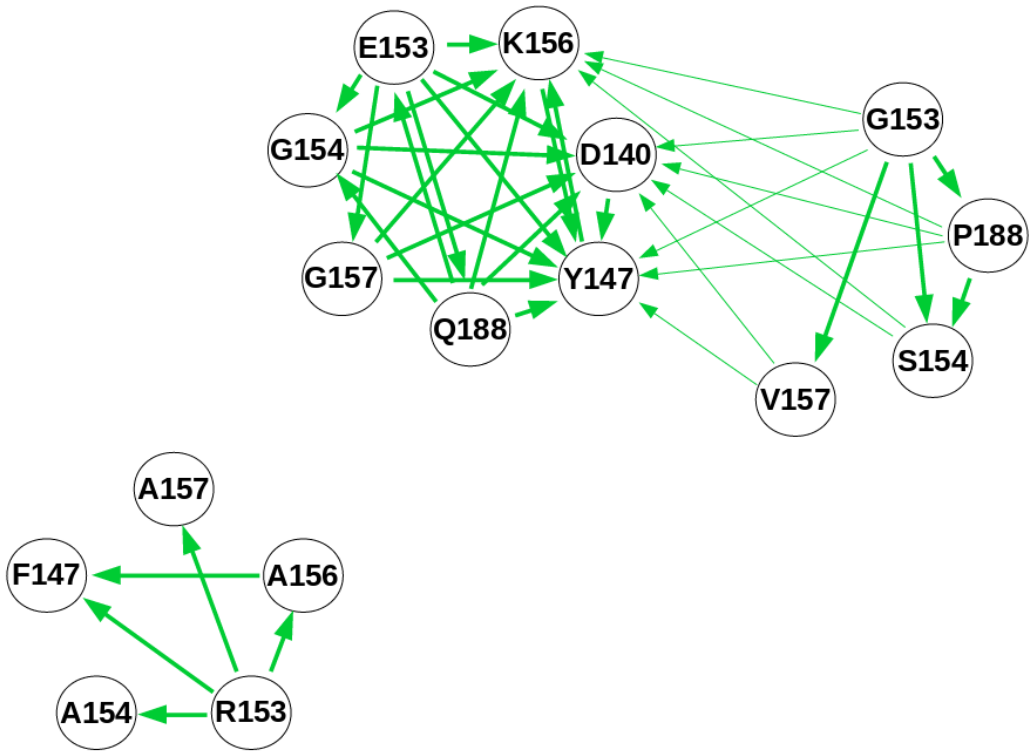

Figura 20: Redes de correlação entre resíduos de *P-box* (e outros resíduos relacionados). Devido ao excessivo número de anti-correlações decorrentes da presença de resíduos mutualmente exclusivos, somente as correlações positivas (setas verdes) foram mostradas. A largura das arestas remete à intensidade da correlação, porém não de maneira proporcional, pois a intensidade das correlações era muito variável.

Aqui vale chamar a atenção para alguns pontos. Primeiramente, a comunidade formada por resíduos típicos de NRs de nemátodos permanece conforme descrito por Afonso e colaboradores (2013). Em segundo lugar, nesse mesmo trabalho, os resíduos típicos de NRs de humanos (figura 20 superior e ao centro) apareciam como duas comunidades separadas. Isso talvez se devesse ao fato de que o uso de filtros reduz significativamente a frequência desses resíduos, como visto na figura 11. Com o uso de pesos, porém, surgem novas correlações entre as duas comunidades outrora detectadas separadamente, formando-se uma nova comunidade com alto número de correlações entre seus membros. Por fim, além de novas conexões, surgem também novos nós nesta rede: os resíduos típicos de receptores esteroidais (figura 20 superior à direita). Entretanto, mesmo com o grande número de anti-correlações entre esta nova comunidade e sua vizinha, ambas permanecem conectadas por três resíduos que apresentam em comum: Asp140, Tyr147 e Lys156. Se considerarmos, porém, que as posições 140 e 147 são relativamente conservadas no alinhamento, e pudéssemos removê-las da rede por não serem exclusivas de

um grupo específico, obteríamos, enfim, três comunidades bem definidas de resíduos específicos para cada um dos três tipos de receptor nuclear aqui investigados. A figura 21 mostra os resíduos típicos de receptores esteroidais mapeados na estrutura, em complexo com o DNA.

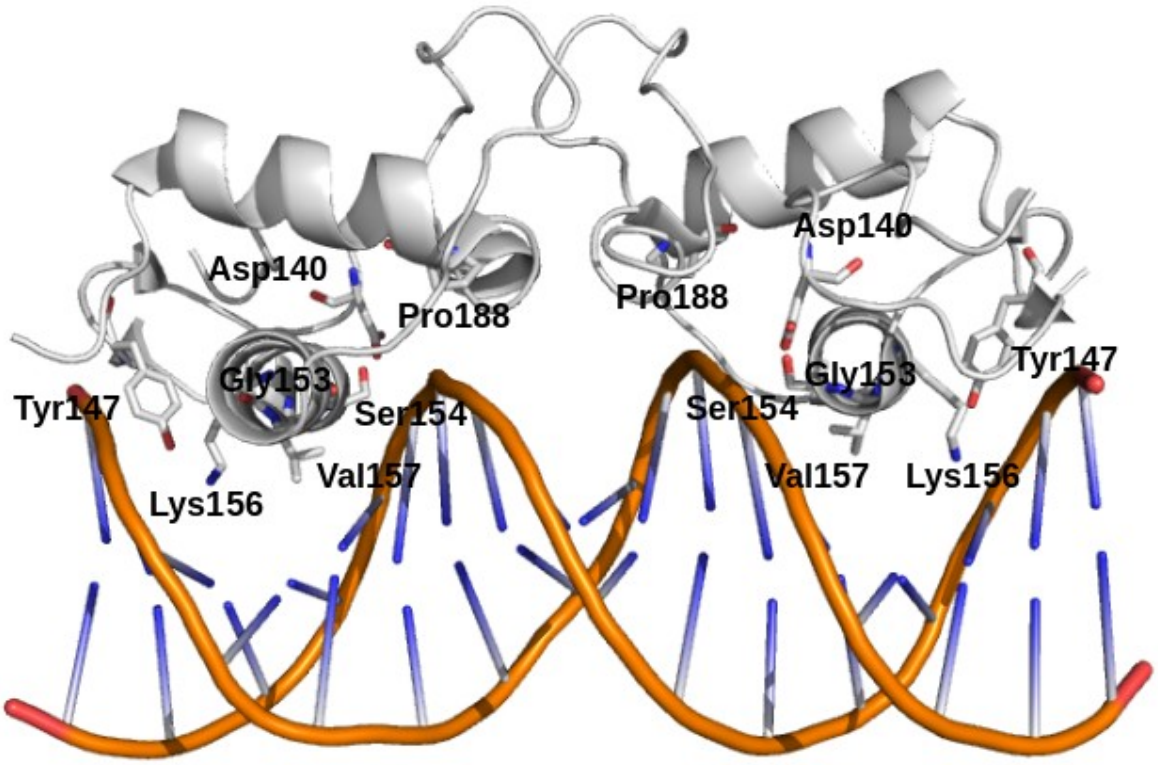

Figura 21: Estrutura cristalográfica do domínio de ligação a DNA (DBD) do receptor de mineralocorticoide (PDB 4tnt) complexado com o DNA. Foi utilizada a numeração da sequência de referência (receptor nuclear humano hRXRα, UniProtKB ID: P19793, PDB 2NLL). Os resíduos $G^{153}S^{154}K^{156}V^{157}$ constituem o motivo *P-box*, embora outros resíduos também participem da interação com o DNA.

5.2.3 – Conservação nos sub-alinhamentos

A figura 22a mostra a distribuição das maiores frequências, para posições com menos de 20% de *gaps*, no sub-alinhamento referente ao grupo ortólogo dos NRs de humanos. A figura 22b mostra a distribuição da entropia de Shannon, para as mesmas posições no mesmo sub-alinhamento. Logo se vê que a maioria das posições é altamente conservada, com seus aminoácidos mais conservados tendo frequências acima de 90% e

entropia baixa na grande maioria. A figura 23 se refere ao sub-alinhamento contendo o *P-box* de nematódeos. Apesar de apresentarem modas opostas, as distribuições de frequências e entropia ainda apontam para um alinhamento redundante. A figura 24 mostra o caso em que este feito encontra-se mais destacado, o grupo ortólogo dos receptores de 3-cetoesteroides.

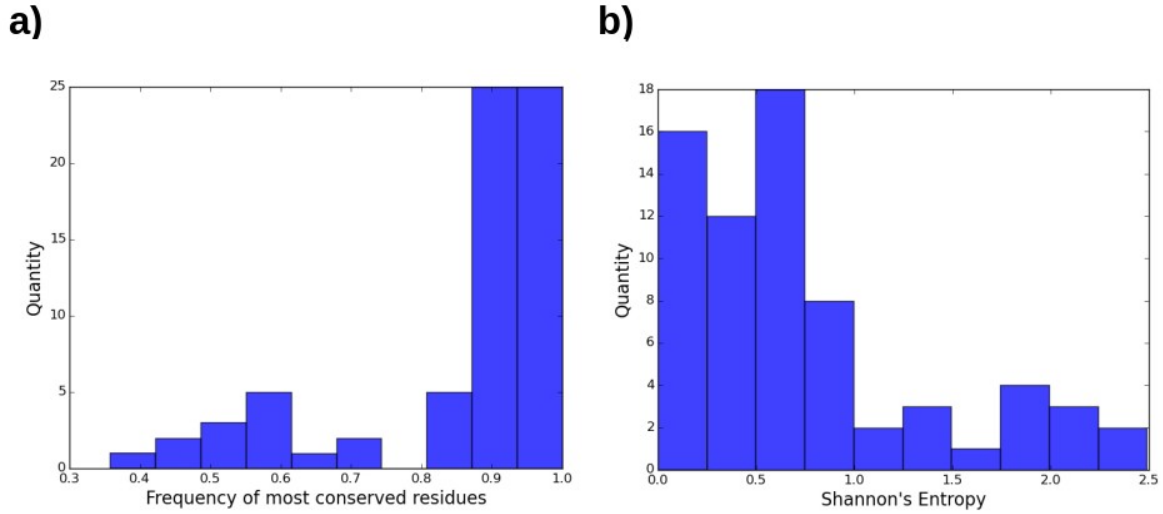

Figura 22: (a) Distribuição das frequências dos aminoácido mais conservados para posições com menos de 20% de *gaps* no sub-alinhamento referente ao grupo ortólogo dos NRs de humanos e afins. (b) Distribuição da Entropia de Shannon também para posições com menos de 20% de *gaps* no mesmo sub-alinhamento.

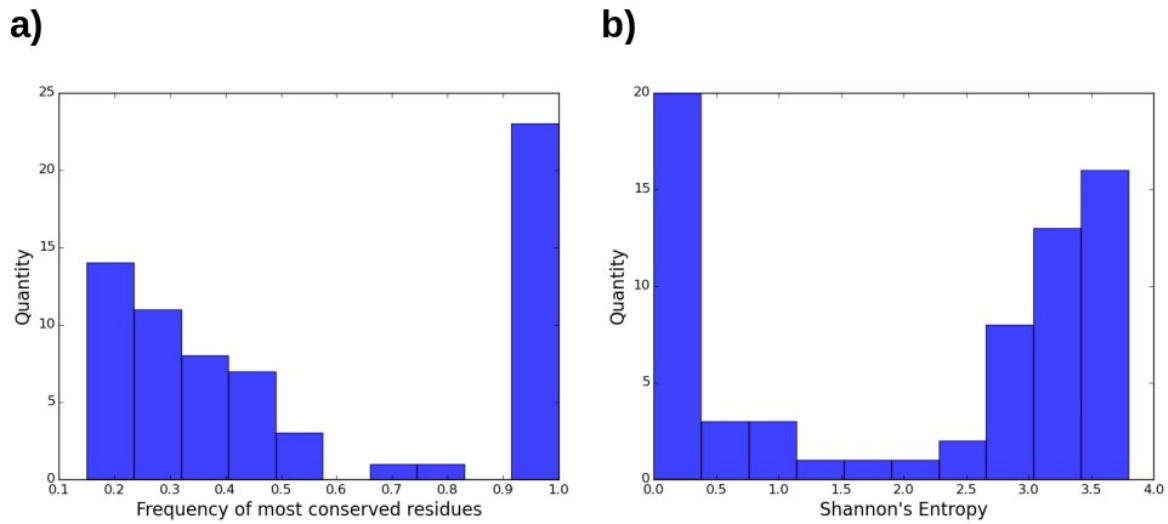

Figura 23: (a) Distribuição das frequências dos aminoácido mais conservados para posições com menos de 20% de *gaps* no sub-alinhamento contendo o *P-box* de *C. elegans* ($CR^{153}A^{154}CA^{156}A^{157}$). (b) Distribuição da Entropia de Shannon também para posições com menos de 20% de *gaps* no mesmo sub-alinhamento.

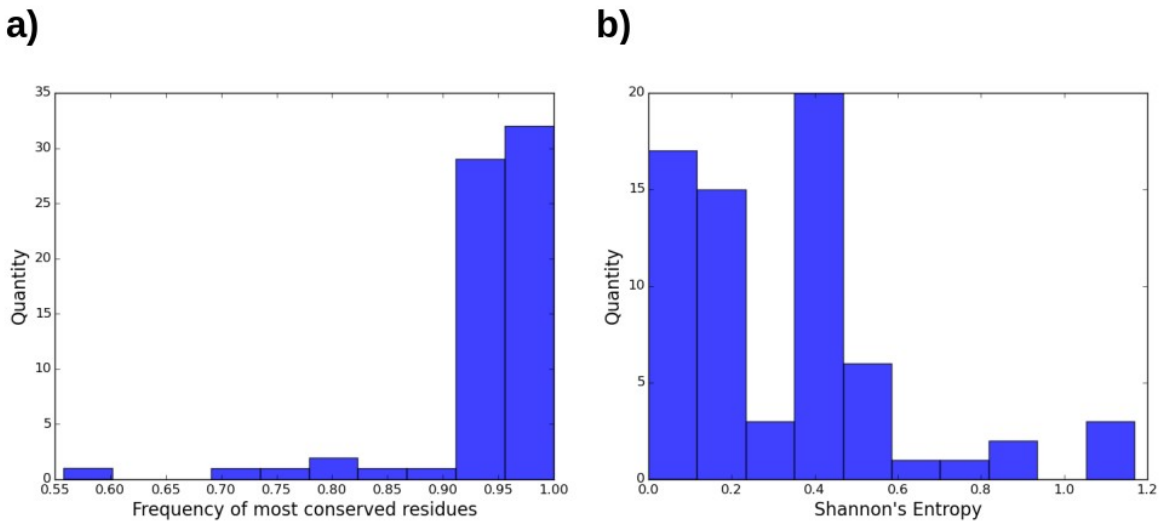

Figura 24: (a) Distribuição das frequências dos aminoácido mais conservados para posições com menos de 20% de *gaps* no sub-alinhamento referente ao grupo ortólogo dos receptores de 3-cetoesteroides. (b) Distribuição da Entropia de Shannon também para posições com menos de 20% de *gaps* no mesmo sub-alinhamento.

Assim como o caso das transtirretinas (TTR), os receptores esteroidais também são muito recentes na história evolutiva, estando presentes somente em vertebrados. Portanto, não houve ainda tempo suficiente para que pudéssemos observar uma diversificação significativa desta subfamília a ponto de sermos capazes de identificar correlações efetivas em meio ao efeito filogenético.

5.2.4 – Novas análises de correlação

Para não restarem dúvidas se as correlações encontradas anteriormente são significativas ou devidas somente ao efeito filogenético, os cálculos de correlação foram refeitos da mesma forma, exceto pela probabilidade *a priori* da distribuição binomial cumulativa, que agora seria substituída pela média das frequências mais altas no sub-alinhamento, como feito no caso das TTR. As novas correlações, não muito destoantes dos primeiros resultados, são mostradas na figura 25, que compara os novos resultados com os anteriores. Logo nota-se uma diminuição na intensidade das correlações entre os resíduos típicos de NRs de nematódeos. Observa-se também uma diminuição no número de conexões entre os resíduos típicos de receptores esteroidais. Pela própria definição de modularidade conclui-se que estes resíduos não formariam uma comunidade bem definida, uma vez que seus membros apresentam poucas conexões entre si e muitas com resíduos da

comunidade vizinha. Ademais, com o aumento da restrição na escolha do *cutoff*, estes resíduos desaparecem quase que completamente da rede.

Mais uma vez, vale ressaltar que no presente trabalho as posições a serem avaliadas foram selecionadas manualmente por questões operacionais, e que, caso a análise fosse extrapolada para as demais posições do alinhamento, certamente apareceriam outras falsas correlações devidas ao efeito filogenético. Portanto, assim como no caso das TTR, o curto período de existência deste grupo ortólogo não oferece uma boa amostragem de sequências que torne possível uma análise de coevolução em nível de resíduos.

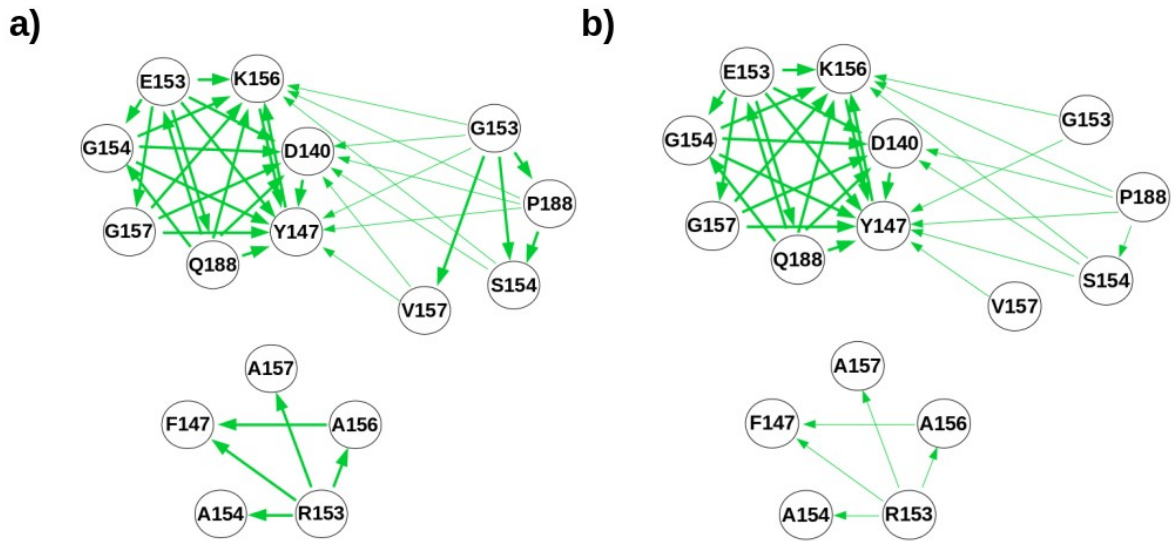

Figura 25: Redes de correlação entre resíduos de *P-box* (e outros resíduos relacionados). Devido ao excessivo número de anti-correlações decorrentes da presença de resíduos mutualmente exclusivos, somente as correlações positivas (setas verdes) foram mostradas. A largura das arestas remete à intensidade da correlação, porém não de maneira proporcional, pois a intensidade das correlações era muito variável. (a) Rede de correlações encontrada anteriormente. (b) Nova rede de correlações obtida após levar em consideração o efeito filogenético, conforme discutido no texto.

CONCLUSÃO

Em suma, o presente trabalho mostrou que, embora subfamílias muito recentes – portanto pouco amostradas em termos de diversidade de sequências – sejam susceptíveis ao efeito filogenético, ainda assim a atribuição de pesos a sequências é capaz de amplificar o sinal oriundo de sequências raras, ao mesmo tempo em que minimiza o efeito de sequências redundantes. Enquanto o uso de filtros geralmente leva a reduções bruscas nas frequências, o uso de pesos geralmente gera frequências intermediárias entre o alinhamento original e os alinhamentos filtrados. Quanto à limitação operacional devida à natureza da distribuição binomial cumulativa, que usa valores inteiros em seus cálculos, esta parece ter sido satisfatoriamente contornada multiplicando-se tais valores discretos pelos valores contínuos obtidos a partir da soma dos pesos de sequências contendo determinado resíduo. Por fim, em relação ao viés introduzido nos alinhamentos pelo efeito filogenético, a abordagem baseada na substituição das probabilidades *a priori* pela média das frequências no sub-alinhamento pareceu ter lidado bem com o problema, reforçando correlações mais fortes e dissolvendo correlações duvidosas. Portanto, a atribuição de pesos às sequências é uma boa prática a ser implementada em análises de conservação, correlação ou qualquer outro método que vise extrair informações a partir de um alinhamento múltiplo de sequências.

REFERÊNCIAS BIBLIOGRÁFICAS

AFONSO, M. Q. L.; DE LIMA, L. HF; BLEICHER, L. Residue correlation networks in nuclear receptors reflect functional specialization and the formation of the nematode-specific P-box. **BMC genomics**, v. 14, n. Suppl 6, p. S1, 2013.

ALTSCHUH, D. *et al.* Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. **Journal of molecular biology**, v. 193, n. 4, p. 693-707, 1987.

ATCHLEY, W. R.; TERHALLE, W.; DRESS, A. Positional dependence, cliques, and predictive motifs in the bHLH protein domain. **Journal of Molecular Evolution**, v. 48, n. 5, p. 501-516, 1999.

BACHEGA, J. F. R. *et al.* Systematic structural studies of iron superoxide dismutases from human parasites and a statistical coupling analysis of metal binding specificity. **Proteins: Structure, Function, and Bioinformatics**, v. 77, n. 1, p. 26-37, 2009.

BALAKRISHNAN, S. *et al*. Learning generative models for protein fold families. **Proteins: Structure, Function, and Bioinformatics**, v. 79, n. 4, p. 1061-1078, 2011.

BARKER, D.; PAGEL, M. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. **PLoS Comput Biol**, v. 1, n. 1, p. e3, 2005.

BENÍTEZ-PÁEZ, A.; CÁRDENAS-BRITO, S.; GUTIÉRREZ, A. J. A practical guide for the computational selection of residues to be experimentally characterized in protein families. **Briefings in bioinformatics**, v. 13, n. 3, p. 329-336, 2011.

BLEICHER, L.; LEMKE, N.; GARRATT, R. C. Using amino acid correlation and community detection algorithms to identify functional determinants in protein families. **PloS one**, v. 6, n. 12, p. e27786, 2011.

BROWN, C. A.; BROWN, K. S. Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my!. **PloS one**, v. 5, n. 6, p. e10779, 2010.

BURGER, L.; VAN NIMWEGEN, E. Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. **Molecular systems biology**, v. 4, n. 1, p. 165, 2008.

BUSSO, C. *et al.* Coq7p relevant residues for protein activity and stability. **Biochimie**, v. 119, p. 92-102, 2015.

CAPRA, J. A.; SINGH, M. Predicting functionally important residues from sequence conservation. **Bioinformatics**, v. 23, n. 15, p. 1875-1882, 2007.

CASARI, G.; SANDER, C.; VALENCIA, A. A method to predict functional residues in proteins. **Nature structural biology**, v. 2, n. 2, p. 171, 1995.

CELNIKER, G. *et al*. ConSurf: using evolutionary data to raise testable hypotheses about protein function. **Israel Journal of Chemistry**, v. 53, n. 3-4, p. 199-206, 2013.

CHAGOYEN, M; GARCÍA-MARTÍN, J. A.; PAZOS, F. Practical analysis of specificity-determining residues in protein families. **Briefings in bioinformatics**, p. bbv045, 2015.

CHAKRABARTI, S.; BRYANT, S. H.; PANCHENKO, Anna R. Functional specificity lies within the properties and evolutionary changes of amino acids. **Journal of molecular biology**, v. 373, n. 3, p. 801-810, 2007.

CHAKRABORTY, A.; CHAKRABARTI, S. A survey on prediction of specificity-determining sites in proteins. **Briefings in bioinformatics**, p. bbt092, 2014.

CHATZOU, M. *et al*. Multiple sequence alignment modeling: methods and applications. **Briefings in bioinformatics**, p. bbv099, 2015.

CHI, C. N. *et al.* Reassessing a sparse energetic network within a single protein domain. **Proceedings of the National Academy of Sciences**, v. 105, n. 12, p. 4679-4684, 2008.

CROOKS, G. E. *et al*. WebLogo: a sequence logo generator. **Genome research**, v. 14, n. 6, p. 1188-1190, 2004.

DAGO, A. E. *et al*. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. **Proceedings of the National Academy of Sciences**, v. 109, n. 26, p. E1733-E1742, 2012.

DEKKER, J. P. *et al.* A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. **Bioinformatics**, v. 20, n. 10, p. 1565-1572, 2004.

DEL SOL MESA, A.; PAZOS, F.; VALENCIA, A. Automatic methods for predicting functionally important residues. **Journal of molecular biology**, v. 326, n. 4, p. 1289-1302, 2003.

DI LENA, P.; NAGATA, K,; BALDI, P. Deep architectures for protein contact map prediction. **Bioinformatics**, v. 28, n. 19, p. 2449-2457, 2012.

DIMA, R. I.; THIRUMALAI, D. Determination of network of residues that regulate allostery in protein families using sequence analysis. **Protein Science**, v. 15, n. 2, p. 258-268, 2006.

DUNN, S. D.; WAHL, L. M.; GLOOR, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. **Bioinformatics**, v. 24, n. 3, p. 333-340, 2008.

DUTHEIL, J. *et al.* A model-based approach for detecting coevolving positions in a molecule. **Molecular biology and evolution**, v. 22, n. 9, p. 1919-1928, 2005.

FAIRMAN, J. W. *et al.* Crystal structures of the outer membrane domain of intimin and invasin from enterohemorrhagic E. coli and enteropathogenic Y. pseudotuberculosis. **Structure**, v. 20, n. 7, p. 1233-1243, 2012.

FARES, M. A.; TRAVERS, S. A. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. **Genetics**, v. 173, n. 1, p. 9-23, 2006.

FAWAL, N. *et al.* Automatic multigenic family annotation: risks and solutions. **Trends in genetics**, v. 30, n. 8, p. 323-325, 2014.

FERREIRA-JÚNIOR, J. R.; BLEICHER, L.; BARROS, M. H. Her2p molecular modeling, mutant analysis and intramitochondrial localization. **Fungal Genetics and Biology**, v. 60, p. 133-139, 2013.

FINN, R. D. *et al*. The Pfam protein families database: towards a more sustainable future. **Nucleic acids research**, p. gkv1344, 2015.

FLAKE, G. W. *et al.* Self-organization and identification of web communities. **Computer**, v. 35, n. 3, p. 66-70, 2002.

FLEISHMAN, S. J.; YIFRACH, O.; BEN-TAL, N. An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. **Journal of molecular biology**, v. 340, n. 2, p. 307-318, 2004.

FODOR, A. A.; ALDRICH, R. W. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. **Proteins: Structure, Function, and Bioinformatics**, v. 56, n. 2, p. 211-221, 2004.

GAUCHER, E. A. *et al.* Predicting functional divergence in protein evolution by site-specific rate shifts. **Trends in biochemical sciences**, v. 27, n. 6, p. 315-321, 2002.

GÖBEL, U. *et al*. Correlated mutations and residue contacts in proteins. **Proteins: Structure, Function, and Bioinformatics**, v. 18, n. 4, p. 309-317, 1994.

GÓMEZ, S.; JENSEN, P.; ARENAS, A. Analysis of community structure in networks of correlated data. **Physical Review E**, v. 80, n. 1, p. 016114, 2009.

GRAÑA, O. *et al*. CASP6 assessment of contact prediction. **Proteins: Structure, Function, and Bioinformatics**, v. 61, n. S7, p. 214-224, 2005.

GU, X. Statistical methods for testing functional divergence after gene duplication. **Molecular biology and evolution**, v. 16, n. 12, p. 1664-1674, 1999.

GU, X. Maximum-likelihood approach for gene family evolution under functional divergence. **Molecular biology and evolution**, v. 18, n. 4, p. 453-464, 2001.

HALABI, N. *et al.* Protein sectors: evolutionary units of three-dimensional structure. **Cell**, v. 138, n. 4, p. 774-786, 2009.

HENNEBRY, S. C. Evolutionary changes to transthyretin: structure and function of a transthyretin- ike ancestral protein. The FEBS journal, v. 276, n. 19, p. 5367, 2009.

HENIKOFF, S.; HENIKOFF, J. G. Position-based sequence weights. **Journal of molecular biology**, v. 243, n. 4, p. 574-578, 1994.

HOPF, T. A. *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. **Cell**, v. 149, n. 7, p. 1607-1621, 2012.

JONES, D. T. *et al*. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. **Bioinformatics**, v. 28, n. 2, p. 184-190, 2012.

JUAN, D.; PAZOS, F.; VALENCIA, A. Co-evolution and co-adaptation in protein networks. **FEBS letters**, v. 582, n. 8, p. 1225-1230, 2008.

JUAN, D.; PAZOS, F.; VALENCIA, A. Emerging methods in protein co-evolution. **Nature Reviews Genetics**, v. 14, n. 4, p. 249-261, 2013.

KASS, I.; HOROVITZ, A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. **Proteins: Structure, Function, and Bioinformatics**, v. 48, n. 4, p. 611-617, 2002.

KIMURA, M. **The neutral theory of molecular evolution**. Cambridge University Press, 1984.

KORBER, B. T. *et al*. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. **Proceedings of the National Academy of Sciences**, v. 90, n. 15, p. 7176-7180, 1993.

LAPEDES, A. S. *et al*. Correlated mutations in models of protein sequences: phylogenetic and structural effects. **Lecture Notes-Monograph Series**, p. 236-256, 1999.

LASKOWSKI, R. A. PDBsum: summaries and analyses of PDB structures. **Nucleic acids research**, v. 29, n. 1, p. 221-222, 2001.

LI, L.; STOECKERT, C. J.; ROOS, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. **Genome research**, v. 13, n. 9, p. 2178-2189, 2003.

LICHTARGE, O.; BOURNE, H. R.; COHEN, F. E. An evolutionary trace method defines binding surfaces common to protein families. **Journal of molecular biology**, v. 257, n. 2, p. 342-358, 1996.

LIZ, M. A. *et al.* Transthyretin, a new cryptic protease. Journal of Biological Chemistry, v. 279, n. 20, p. 21431-21438, 2004.LOCKLESS, S. W.; RANGANATHAN, R. Evolutionarily conserved pathways of energetic connectivity in protein families. **Science**, v. 286, n. 5438, p. 295-299, 1999.

LIZ, M. A. *et al.* Transthyretin is a metallopeptidase with an inducible active site. Biochemical Journal, v. 443, n. 3, p. 769-778, 2012.MADABUSHI, S. *et al.* Structural clusters of evolutionary trace residues are statistically significant and common in proteins. **Journal of molecular biology**, v. 316, n. 1, p. 139-154, 2002.

MARKS, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. **PloS one**, v. 6, n. 12, p. e28766, 2011.

MARTIN, L. C. *et al.* Using information theory to search for co-evolving residues in proteins. **Bioinformatics**, v. 21, n. 22, p. 4116-4124, 2005.

MARX, V. Biology: The big challenges of big data. **Nature**, v. 498, n. 7453, p. 255-260, 2013.

MELO-MINARDI, R. C.; BASTARD, K.; ARTIGUENAVE, F. Identification of subfamily-specific sites based on active sites modeling and clustering. **Bioinformatics**, v. 26, n. 24, p. 3075-3082, 2010.

MORCOS, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. **Proceedings of the National Academy of Sciences**, v. 108, n. 49, p. E1293-E1301, 2011.

NEHER, E. How frequent are correlated changes in families of protein sequences?. **Proceedings of the National Academy of Sciences**, v. 91, n. 1, p. 98-102, 1994.

NEWMAN, M. E. J.; GIRVAN, M. Finding and evaluating community structure in networks. **Physical review E**, v. 69, n. 2, p. 026113, 2004.

NUGENT, T.; JONES, D. T. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. **Proceedings of the National Academy of Sciences**, v. 109, n. 24, p. E1540-E1547, 2012.

OHNO, S. **Evolution by gene duplication**. Springer Science & Business Media, 2013.

OLIVEIRA, L.; PAIVA, A. C.; VRIEND, G. Correlated mutation analyses on very large sequence families. **Chembiochem**, v. 3, n. 10, p. 1010-1017, 2002.

OLMEA, O.; VALENCIA, A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. **Folding and Design**, v. 2, p. S25-S32, 1997.

PALANINATHAN, S. K. Nearly 200 X-ray crystal structures of transthyretin: what do they tell us about this protein and the design of drugs for TTR amyloidoses?. Current medicinal chemistry, v. 19, n. 15, p. 2324, 2012.

PALMIERI, L. C. *et al.* Novel Zn2+-binding Sites in Human Transthyretin IMPLICATIONS FOR AMYLOIDOGENESIS AND RETINOL-BINDING PROTEIN RECOGNITION. **Journal of Biological Chemistry**, v. 285, n. 41, p. 31731-31741, 2010.

PAZOS, F. *et al*. Correlated mutations contain information about protein-protein interaction. **Journal of molecular biology**, v. 271, n. 4, p. 511-523, 1997.

PAZOS, F.; BANG, J. Computational prediction of functionally important regions in proteins. **Current Bioinformatics**, v. 1, n. 1, p. 15-23, 2006.

PAZOS, F.; VALENCIA, A. Protein co-evolution, co-adaptation and interactions. **The EMBO journal**, v. 27, n. 20, p. 2648-2655, 2008.

PIETROSEMOLI, N. *et al.* Computational Prediction of Important Regions in Protein Sequences [Life Sciences]. **IEEE Signal Processing Magazine**, v. 29, p. 143-147, 2012.

PINTO, M. *et al*. Ligand-binding properties of human transthyretin. **Amyloid**, v. 18, n. sup1, p. 51-54, 2011.

POLLOCK, D. D.; TAYLOR, W. R.; GOLDMAN, N. Coevolving protein residues: maximum likelihood identification and relationship to structure. **Journal of molecular biology**, v. 287, n. 1, p. 187-198, 1999.

POWER, D. M. *et al.* Evolution of the thyroid hormone-binding protein, transthyretin. General and comparative endocrinology, v. 119, n. 3, p. 241-255, 2000.RADICCHI, F. *et al*. Defining and identifying communities in networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 101, n. 9, p. 2658-2663, 2004.

RAMAZZINA, I. et al. Completing the uric acid degradation pathway through phylogenetic comparison of whole genomes. Nature chemical biology, v. 2, n. 3, p. 144-148, 2006.RAUSELL, A. *et al.* Protein interactions and ligand binding: from protein subfamilies to functional specificity. **Proceedings of the National Academy of Sciences**, v. 107, n. 5, p. 1995-2000, 2010.

REYNOLDS, K. A.; MCLAUGHLIN, R. N.; RANGANATHAN, R. Hot spots for allosteric regulation on protein surfaces. **Cell**, v. 147, n. 7, p. 1564-1575, 2011.

RICHARDSON, S. J. Tweaking the structure to radically change the function: the evolution of transthyretin from 5-hydroxyisourate hydrolase to triiodothyronine distributor to thyroxine distributor. Frontiers in endocrinology, v. 5, 2014.

ROBINSON-RECHAVI, M. *et al*. Explosive lineage-specific expansion of the orphan nuclear receptor HNF4 in nematodes. **Journal of molecular evolution**, v. 60, n. 5, p. 577-586, 2005.

RODRIGUEZ, G. J. *et al.* Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. **Proceedings of the National Academy of Sciences**, v. 107, n. 17, p. 7787-7792, 2010.

SARAIVA, M. J. M. Transthyretin mutations in hyperthyroxinemia and amyloid diseases. Human mutation, v. 17, n. 6, p. 493-503, 2001.SCHUG, A. *et al.* High-resolution protein complexes from integrating genomic information with molecular simulation.

Proceedings of the National Academy of Sciences, v. 106, n. 52, p. 22124-22129, 2009.

SHANNON, C. E. A mathematical theory of communication. **ACM SIGMOBILE Mobile Computing and Communications Review**, v. 5, n. 1, p. 3-55, 2001.

SHENKIN, P. S.; ERMAN, B.; MASTRANDREA, L. D. Information-theoretical entropy as a measure of sequence variability. **Proteins: Structure, Function, and Bioinformatics**, v. 11, n. 4, p. 297-313, 1991.

SHINDYALOV, I. N.; KOLCHANOV, N. A.; SANDER, C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?. **Protein Engineering**, v. 7, n. 3, p. 349-358, 1994.

SOCOLICH, M. *et al*. Evolutionary information for specifying a protein fold. **Nature**, v. 437, n. 7058, p. 512-518, 2005.

SREEKUMAR, J. *et al*. Correlated mutations via regularized multinomial regression. **BMC bioinformatics**, v. 12, n. 1, p. 1, 2011.

SÜEL, G. M. *et al.* Evolutionarily conserved networks of residues mediate allosteric communication in proteins. **Nature Structural & Molecular Biology**, v. 10, n. 1, p. 59-69, 2003.

SUŁKOWSKA, J. I. *et al.* Genomics-aided structure prediction. **Proceedings of the National Academy of Sciences**, v. 109, n. 26, p. 10340-10345, 2012.

TAYLOR, W. R.; HATRICK, K. Compensating changes in protein multiple sequence alignments. **Protein Engineering**, v. 7, n. 3, p. 341-348, 1994.

TEPPA, E. *et al*. Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. Implication for catalytic residue prediction. **BMC bioinformatics**, v. 13, n. 1, p. 235, 2012.

THOMPSON, J. D. *et al.* The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. **Nucleic acids research**, v. 25, n. 24, p. 4876-4882, 1997.

THORNTON, J. W. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. **Proceedings of the National Academy of Sciences**, v. 98, n. 10, p. 5671-5676, 2001.

TILLIER, E. R.; LUI, T. W. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. **Bioinformatics**, v. 19, n. 6, p. 750-755, 2003.

TRESS, M. *et al*. Scoring docking models with evolutionary information. **Proteins: Structure, Function, and Bioinformatics**, v. 60, n. 2, p. 275-280, 2005.

TRESS, M. L.; VALENCIA, A. Predicted residue–residue contacts can help the scoring of 3D models. **Proteins: Structure, Function, and Bioinformatics**, v. 78, n. 8, p. 1980-1991, 2010.

VALDAR, W. S. J. Scoring residue conservation. **Proteins: Structure, Function, and Bioinformatics**, v. 48, n. 2, p. 227-241, 2002.

VILLAR, H. O.; KAUVAR, L. M. Amino acid preferences at protein binding sites. **FEBS letters**, v. 349, n. 1, p. 125-130, 1994.

WALLACE, A. C.; LASKOWSKI, R. A.; THORNTON, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. **Protein engineering**, v. 8, n. 2, p. 127-134, 1995.

WEIGT, M. *et al.* Identification of direct residue contacts in protein–protein interaction by message passing. **Proceedings of the National Academy of Sciences**, v. 106, n. 1, p. 67-72, 2009.

WINTJENS, R.; GILIS, D.; ROOMAN, M. Mn/Fe superoxide dismutase interaction fingerprints and prediction of oligomerization and metal cofactor from sequence. **Proteins: Structure, Function, and Bioinformatics**, v. 70, n. 4, p. 1564-1577, 2008.

YEANG, C.; HAUSSLER, D. Detecting coevolution in and among protein domains. **PLoS Comput Biol**, v. 3, n. 11, p. e211, 2007.

ZANOTTI, G. et al. Structure of zebra fish HIUase: insights into evolution of an enzyme to a hormone transporter. Journal of molecular biology, v. 363, n. 1, p. 1-9, 2006.

ZVELEBIL, M. J. *et al.* Prediction of protein secondary structure and active sites using the alignment of homologous sequences. **Journal of molecular biology**, v. 195, n. 4, p. 957-961, 1987.