

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS

RICARDO ASSUNÇÃO VIALLE

**Evidências de mudanças estruturais proteicas em transições
macroevolutivas**

BELO HORIZONTE – MG

2017

RICARDO ASSUNÇÃO VIALLE

**Evidências de mudanças estruturais proteicas em transições
macroevolutivas**

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito para a obtenção do título de Doutor em Bioinformática

Orientador: Dr. José Miguel Ortega

Coorientador: Dr. Roberto Tadeu Raittz

BELO HORIZONTE – MG

2017

043 Vialle, Ricardo Assunção.
Evidências de mudanças estruturais proteicas em transições
macroevolutivas [manuscrito] / Ricardo Assunção Vialle. – 2017.

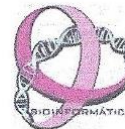
94 f. : il. ; 29,5 cm.

Orientador: Dr. José Miguel Ortega. Coorientador: Dr. Roberto Tadeu Raittz.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Macroevolução. 2. Proteínas. 3. Filogenia - Teses. 4. Genômica comparativa. 5. Bioinformática - Teses. I. Ortega, José Miguel. II. Raittz, Roberto Tadeu. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 575.112



"Evidências de mudanças estruturais proteicas em transições macroevolutivas"

Ricardo Assunção Vialle

Tese aprovada pela banca examinadora constituída pelos Professores:



Prof. José Miguel Ortega - Orientador
UFMG




Prof. Roberto Tadeu Raittz - Co-Orientador
UFPR



Prof. Aristoteles Góes Neto
UFMG



Prof. Elio Anthony Cino
UFMG



Prof. Carlos Henrique da Silveira
Universidade Federal de Itajubá



Profa Cristiane Neri Nobre
ICEI/PUC Minas



Prof. Douglas Eduardo Valente Pires
Fundação Oswaldo Cruz/Centro de Pesquisa

Belo Horizonte, 17 de fevereiro de 2017.

À minha amada esposa, Katia.

Agradecimentos

Agradeço primeiramente a toda minha família, em especial aos meus pais Solange e Ubiratan e ao meu irmão Rodrigo, por todo apoio e incentivo ao longo de toda a minha vida.

À minha querida Katia, amor da minha vida, pela qual tenho profunda admiração, por estar sempre presente.

Aos meus professores orientadores Miguel e Roberto pela oportunidade e pelo apoio em todos os momentos.

Ao programa de pós-graduação em Bioinformática da UFMG, a todos os professores que tive o prazer de conhecer e à Sheila por toda atenção e disponibilidade em ajudar sempre que foi preciso.

A todos meus amigos que conheci por meio do Laboratório de Biodados. Ao Super KRAV, presente em todos os momentos. À Verônica pelas montagens magníficas. Ao Assis pelas discussões “não convencionais”. Ao Lucas e Marcele pela parceria e boas partidas de WAR. Ao Diego e Rayson pela ótima companhia. Ao Carlos e Tetsu, sempre dispostos a ajudar. Ao Velloso pela parceria no desenvolvimento do BOWS. E a Beatriz, Daniel, Edgar, Elisson, Fenícia, Fernandinha, Gabriel, Lissur, Luís, Rafael, Raquel e Thaís pela amizade e boa convivência.

Aos novos amigos que tive o prazer de conhecer ao longo desses anos. Em especial ao Fernando, Matheus, Vinícios, Camila, Rose, Marcio, Tutu e Junia. E também ao “não tão novo” amigo Vitor, que mesmo de longe esteve presente.

Ao Nick Goldman pela oportunidade de trabalhar junto ao seu grupo de pesquisa e a todos os membros grupo, especialmente ao Asif por acompanhar de perto o desenvolvimento do trabalho.

Agradeço aos órgãos financiadores da bolsa de estudo e fomento CNPq, CAPES, Fapemig e ao Programa Ciências sem Fronteiras.

E a todos que, direta ou indiretamente, contribuíram para eu concretizar este trabalho, sou, sinceramente,

Muito grato!

“Ao infinito... e além!”
Buzz Lightyear

Sumário

Lista de figuras.....	v
Lista de tabelas.....	vii
Lista de abreviaturas e siglas.....	viii
Resumo	ix
Abstract.....	x
1. Introdução.....	1
1.1. Origem de genes e macroevolução	1
1.2. Origem “de novo” de genes	6
1.3. Aminoácidos e proteínas.....	7
1.4. Métodos de predição de estrutura secundária	11
2. Justificativa e objetivo	13
3. Materiais e métodos.....	14
3.1. Dados de estruturas de proteínas	14
3.2. Predições de estruturas secundárias	15
3.3. Determinação do consenso das predições	15
3.4. Dados de proteomas completos e de genes com origem de novo	15
3.5. Dados de domínios estruturais.....	17
3.6. Dados de origem dos genes	18
3.7. Integração e manipulação dos dados	19
3.8. Análise de enriquecimento funcional.....	19
4. Resultados.....	20
4.1. Análise dos métodos de predição de estrutura secundária	20
4.2. Análise evolutiva com dados do PDB.....	29
4.3. Análise evolutiva com dados de predição	38
4.3.1. Análise de Homo sapiens	41
4.3.2. Proteínas com origem de novo e ORFs de regiões não codificadoras	54
5. Discussão	58
5.1. Limitações do PDB e predições de estrutura secundária	58
5.2. Mudanças estruturais relacionadas à idade evolutiva dos genes	59
5.3. Origem de novo tem suporte com dados de estrutura secundária	63
6. Conclusão.....	65
7. Produção científica	67
Referências.....	68

Lista de figuras

Figura 1 - Abordagens para identificação de novos genes.	2
Figura 2 - Linha do tempo com origem dos clados da linhagem humana.	4
Figura 3 - Terminologia para paralogia e ortologia.	5
Figura 4 - Tipos de conformações de estrutura secundária.	9
Figura 5 - Composição da estrutura secundária em estruturas do PDB e métodos de predição.	22
Figura 6 - Comprimentos de segmentos de estrutura secundária em estruturas do PDB e métodos de predição.	23
Figura 7 - Correlação entre distribuições de estrutura secundária de estruturas do PDB e métodos de predição.	24
Figura 8 - Distribuição estrutural de sequências de Homo sapiens por método de predição.	25
Figura 9 - Correlação entre distribuições de estrutura secundária de predições do proteoma de Homo sapiens.	26
Figura 10 - Distribuição estrutural de sequências permutadas de Homo sapiens por método de predição.	27
Figura 11 - Correlação entre distribuições de estrutura secundária de predições de sequências permutadas do proteoma de Homo sapiens.	28
Figura 12 - Correlação entre distribuições de estrutura secundária entre predições de sequências permutadas e não permutadas de Homo sapiens.	29
Figura 13 - Composição de estrutura secundária em estruturas do PDB por domínio taxonômico.	30
Figura 14 - Distribuição alfa-beta em estruturas do PDB por domínio taxonômico.	31
Figura 15 - Comprimento de estruturas do PDB por domínio taxonômico.	31
Figura 16 - Cobertura de estruturas do PDB em relação a sequências do UniProt por domínio taxonômico.	32
Figura 17 - Comprimentos de segmentos de estrutura secundária em estruturas do PDB por domínio taxonômico.	32
Figura 18 - Classificação de domínios CATH para sequências do PDB por domínio taxonômico.	33
Figura 19 - Composição de estrutura secundária em estruturas do PDB de Homo sapiens por idade evolutiva.	34
Figura 20 - Distribuição de alfa hélices e fitas beta em estruturas de Homo sapiens presentes no PDB por idade evolutiva.	35
Figura 21 - Comprimento de estruturas PDB de Homo sapiens por idade evolutiva.	35
Figura 22 - Cobertura de estruturas do PDB de Homo sapiens em relação às sequências do UniProt por idade evolutiva.	36
Figura 23 - Classificação de domínios CATH para sequências de Homo sapiens presentes no PDB por idade evolutiva.	36
Figura 24 - Uso de aminoácidos em sequências de Homo sapiens presentes no PDB com idade em Euk_Bac_Arch versus Verteb_Mammalia.	37
Figura 25 - Composição estrutural dos proteomas de referência.	39
Figura 26 - Análise de componentes principais (PCA) da estrutura secundária dos proteomas de referência.	40
Figura 27 - Classificação por idade evolutiva do proteoma humano.	42

Figura 28 - Comprimento de sequências do proteoma humano por idade evolutiva.	43
Figura 29 - Uso de aminoácidos em sequências do proteoma de Homo sapiens com idade em Cellular_organisms e Mammalia.	44
Figura 30 - Enriquecimento de termos de processos do Gene Ontology.....	46
Figura 31 - Enriquecimento funcional de termos do Gene Ontology.....	47
Figura 32 - Enriquecimento de termos de componentes do Gene Ontology.....	48
Figura 33 - Enriquecimento de termos do Gene Ontology para proteínas pouco estruturadas.	49
Figura 34 - Classificação de domínios CATH para sequências do proteoma de Homo sapiens preditas no Gene3D por idade evolutiva.	50
Figura 35 - Cobertura de domínios por idade evolutiva.	51
Figura 36 - Composição estrutural de domínios e regiões extradomínios.	52
Figura 37 - Uso de aminoácidos em regiões de domínios e extradomínios.....	54
Figura 38 - Uso de aminoácidos em regiões de domínios e extradomínios por idade evolutiva.....	54
Figura 39 - Composição alfa-beta de sequências de primatas com origem de novo.	55
Figura 40 - Estrutura secundária em regiões não codificadoras.....	56
Figura 41 - Uso de aminoácidos em regiões não codificadoras.	57
Figura 42 - Determinantes de aminoácidos que definem diferenças estruturais e funcionais entre as proteínas ordenadas e intrinsecamente desordenadas.	62

Lista de tabelas

Tabela 1 - Preferências conformacionais de aminoácidos.....	10
Tabela 2 - Estados de estrutura secundária definidas no DSSP.	14
Tabela 3 - Lista das 66 espécies de referência utilizadas em Quest for Orthologs (QfO). ...	16
Tabela 4 - Genes com origem de novo descobertos recentemente.....	17
Tabela 5 - Frequência do total de resíduos nos três tipos de situações para formação de consenso.	21
Tabela 6 - Acurácia dos métodos de predição de estrutura secundária.	21
Tabela 7 - Composição de estrutura secundária do PDB e métodos de predição.	23
Tabela 8 - Sequências de estruturas do PDB por domínio taxonômico.....	31
Tabela 9 - Estruturas do PDB de Homo sapiens por idade evolutiva.	35
Tabela 10 - Uso de aminoácidos em sequências de Homo sapiens presentes no PDB com idade em Euk_Bac_Arch e Verteb_Mammalia.	37
Tabela 11 - Sequências de Homo sapiens por idade evolutiva	42
Tabela 12 - Uso de aminoácidos em sequências do proteoma de Homo sapiens com idade em Cellular_organisms e Mammalia.	45
Tabela 13 - Composição estrutural de regiões de domínios e extra-domínios por idade evolutiva.....	53

Lista de abreviaturas e siglas

BLAST	Basic Local Alignment Search Tool
CATH	Class, Architecture, Topology, Homology
DSSP	Dictionary of Protein Secondary Structure
FT	Fatores de transcrição
HMM	Hidden Markov Model
IDP	Intrinsically Disordered Protein
LCA	Lowest Common Ancestor
MRCA	Most Recent Common Ancestor
ORF	Open Reading Frame
PDB	Protein Data Bank
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
PSSM	Position-Specific Scoring Matrix
PTM	Post-Translational Modification
QfO	Quest for Orthologs
SIFTS	Structure Integration with Function, Taxonomy and Sequence
SOV	Segment Overlap
SVM	Support Vector Machine
UTR	Untranslated Region
VAST	Vector Alignment Search Tool

Resumo

Proteínas são polímeros lineares de aminoácidos que apresentam uma enorme variedade de estruturas e funções. Sua versatilidade está relacionada à diversidade e ordenação dos resíduos encontrados em suas cadeias. A cadeia de polímeros (estrutura primária) enovela-se em estruturas secundárias e terciárias podendo formar multímeros, conhecidas como estruturas quaternárias. Dados genômicos e estruturais permitem novas descobertas sobre como proteínas provavelmente evoluíram a partir de um pequeno conjunto inicial de domínios e arranjos de domínios. Alguns genes codificadores de proteínas parecem ter surgido *de novo* a partir de partes randômicas do DNA, outros parecem existir por bilhões de anos, virtualmente inalterados. Neste trabalho investigamos se existe um viés estrutural relacionado ao período de origem das proteínas. Conhecendo a época de origem das proteínas, determinada através da técnica de filoestratigrafia, analisamos dados de estrutura secundária, obtida tanto a partir de dados experimentais, como a partir de predições. Notamos que proteínas mais recentes apresentam menor conteúdo de estrutura secundária e maior diversidade no teor de alfa-hélice e fitas beta. Diferenças no uso de aminoácidos em regiões de domínios e extradomínios podem explicar as mudanças observadas. Notamos também, que as mudanças mais acentuadas ocorrem em proteínas com origem a partir de *Opisthokonta*. Funções relacionadas à ligação ao DNA e a fatores de transcrição são enriquecidas tanto em proteínas originadas neste clado, como em proteínas com baixa composição de estrutura secundária. Além disso, investigamos como o viés observado pode ser estendido para auxiliar a compreensão de processos de origem de novas proteínas. Realizando predições de ORFs deduzidas de regiões provavelmente não codificadoras, observamos que possíveis cadeias podem ser obtidas com conteúdo de estrutura encontrado em proteínas naturais, dando suporte à hipótese de origem *de novo* de genes, no que se refere a esse requerimento estrutural.

Palavras-chave: Macroevolução, Filoestratigrafia, Estrutura secundária de proteínas, Origem *de novo* de genes.

Abstract

Proteins are linear polymers of amino acids involved in a huge range of different structures and functions within the cell. Its versatility is related to the diversity and ordering of residues found in its side chains. The polymer chain (primary structure) folds into secondary and tertiary structures, and may form multimers, known as quaternary structures. Genomic and structural data allow new discoveries about how proteins probably evolved from a small initial set of domains and domain arrangements. Some protein-coding genes appear to have arisen from random parts of DNA, others seem to exist for billions of years, virtually unchanged. In this work, we investigate if exists a structural bias related to the period of origin of the proteins. Knowing the time of origin of proteins, determined by phylostratigraphic methods, we analyzed data from secondary structure, obtained both from experimental data and prediction data. We found that more recent proteins have generally fewer structured content and, at same time, shows more diverse structural content of alpha helices and beta strands. Differences in the use of amino acids in domains and inter-domains regions may explain these changes. We noticed that the most significant changes occur in proteins originating from *Opisthokonta*. Functions of the DNA binding and transcription factors are enriched in proteins originated in this clade and in proteins with lower secondary structure composition. Furthermore, we investigated if the observed bias can be extended to understanding processes of novel proteins origins. Predictions of ORFs derived from noncoding regions showed that the secondary structure composition of these sequences are compatible with those found in natural proteins. Thus, the source of new genes is compatible with the readily achieved by translation of these sequences, given support for *de novo* hypothesis.

Keywords: Macroevolution, Phylostratigraphy, Protein secondary structure, *de novo* gene origin.

1. Introdução

1.1. Origem de genes e macroevolução

A origem de novos genes é um importante fator para inovação em todos os organismos. A duplicação gênica foi o primeiro mecanismo proposto para explicar a origem gênica. Susumu Ohno, em 1970, compilou em seu livro “*Evolution by gene duplication*” diversos casos onde mostrava de maneira convincente a relevância evolutiva de genes originados por duplicação. Neste trabalho ele afirma que sem a duplicação gênica a criação dos metazoários, vertebrados e mamíferos seria impossível a partir de organismos unicelulares (OHNO, 1970). Duplicações podem ser descritas envolvendo um genoma inteiro, grandes segmentos de um genoma, genes individuais, éxons, ou até mesmo partes específicas de éxons (BETRÁN; LONG, 2002). Por meio desse mecanismo, um gene duplicado pode desenvolver novas funções enquanto a cópia ancestral mantém as funções originais, sendo muitas dessas novas funções extremamente importantes para o desenvolvimento em vários organismos (PRINCE; PICKETT, 2002). São diversos os mecanismos que podem gerar a duplicação gênica: cruzamento desigual, transposições envolvendo elementos transponíveis, retrotransposição, duplicações segmentadas e duplicações de genoma inteiro (HUANG; BURNS; BOEKE, 2012; JIANG et al., 2004; KAESSMANN; VINCKENBOSCH; LONG, 2009; MORGANTE et al., 2005; RANZ et al., 2007; ZHANG, 2003). Por muito tempo pensou-se que todos os genes modernos fossem formas derivadas de outros genes através desse processo. Hoje, vários outros mecanismos são conhecidos por estar envolvidos na origem gênica, tais como: embaralhamento de éxons, retroposons, elementos móveis, transferência lateral, fusão e fissão gênica e origem *de novo* (LONG et al., 2003).

Todas essas novas sequências aumentam a complexidade e a diversidade de genomas. Com o passar do tempo, as variações nas populações, governadas por interações ambientais, resultam em mudanças observadas em longo prazo em genótipos e fenótipos (SIMAKOV; LARSSON; ARENDT, 2013). No estudo de evolução biológica, tais mudanças são geralmente classificadas em termos de microevolução e macroevolução. Microevolução refere-se ao processo de mudanças ao longo do tempo em espécies ou populações específicas. Como exemplo, podemos citar mudanças genéticas ou fenotípicas, mudanças na distribuição geográfica ou entre interações com outras espécies. Já o termo macroevolução refere-se aos processos relacionados à longos períodos de tempo, geralmente envolvendo origem e extinção de espécies bem como mudanças genéticas e

fenotípicas de longo prazo (BOKMA, 2015). Neste sentido, o processo de origem de novos genes é um processo de microevolução (LONG et al., 2013).

Abordagens de análise comparativa entre genes e genomas podem ser empregadas na identificação de novos genes. Novos genes podem ser definidos a partir da observação da sua presença em um grupo monofilético (grupo de táxons descendentes de um ancestral comum) e sua ausência em grupos externos (LONG et al., 2013). Genes vizinhos podem ser utilizados para auxiliar nessa descoberta. Com informações de múltiplos genomas, é possível realizar análises de sintenia (conservação na ordem dos genes) para identificar novos genes em espécies relacionadas (Figura 1). Buscas por similaridade também podem auxiliar na identificação de novos genes. A ausência de genes semelhantes encontrados em outros organismos (gene órfãos) pode ser um sinal de novidade (MCLYSAGHT; HURST, 2016). Duplicações também podem ser identificadas por meio de comparações com todos os genes do genoma (VINCKENBOSCH; DUPANLOUP; KAESSMANN, 2006). Já para saber se um novo gene é funcional ou não, informações sobre taxas de substituições sinônimas (dS) e não sinônimas (dN) podem ser utilizadas. Sinais de seleção purificadora ($dN/dS < 1$) podem indicar que o novo gene está sob pressão seletiva (LONG et al., 2013). Outra opção é realizar análises de polimorfismos procurando por desvios em espectros de frequência alélica. Somado a isso, informações de expressão gênica e nocautes podem ser definitivos na determinação da funcionalidade do novo gene (MCLYSAGHT; HURST, 2016).

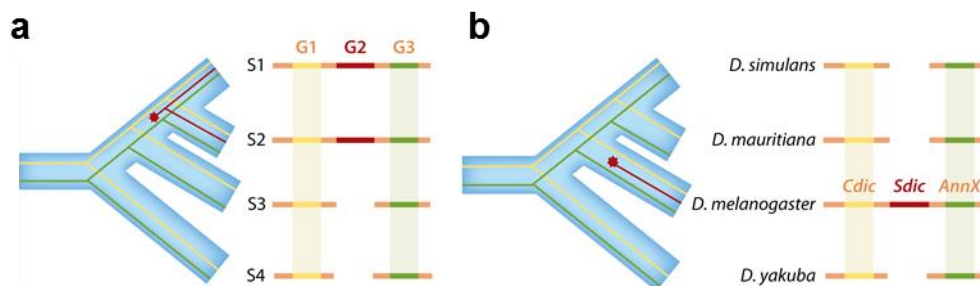


Figura 1 - Abordagens para identificação de novos genes. (a), refere-se ao procedimento baseado apenas na ocorrência ou ausência de genes em táxons relacionados. O relacionamento filogenético das espécies S1 à S4 é indicado pela árvore azul. Os relacionamentos entre os genes G1 (em amarelo), G2 (em vermelho) e G3 (em verde) é indicado pelas linhas internas. O alinhamento dos genomas S1 à S4, mostra que o novo gene (G2) está presente em S1 e S2 e ausente em S3 e S4. Isso indica que G2 teve origem no ancestral comum de S1 e S2, entre os genes antigos G1 e G3 (estrela vermelha). **(b)**, mostra um exemplo utilizando sintenia de alinhamentos. *Sdic* existe somente em *Drosophila melanogaster*. Neste caso, *Sdic* teve origem através de uma recombinação de duplicações nos genes vizinhos, *Cdic* e *AnnX*. Adaptado de (LONG et al., 2013).

À medida com que os genes são originados (independente da forma como foram gerados) e fixados nas espécies, uma assinatura com respeito a sua idade evolutiva é

deixada no genoma (YIN et al., 2016). A abordagem estatística para a reconstrução das tendências macroevolutivas baseadas no princípio do gene fundador e no surgimento de famílias proteicas é conhecido como filoestratigrafia (KAHL, 2015). A filoestratigrafia utiliza a ideia de equilíbrio pontuado de genes, no qual novos genes sofrem uma rápida divergência seguida de uma subsequente redução na evolução das sequências (DOMAZET-LOŠO et al., 2017). Esse método permite estudar sistematicamente as características dos genes no decorrer do tempo, sendo possível observar padrões de surgimento de novos genes e estimar suas idades evolutivas (DOMAZET-LOSO; BRAJKOVIĆ; TAUTZ, 2007; NEME; TAUTZ, 2013). Além disso, estudos como o *TimeTree* (BLAIR HEDGES; KUMAR, 2009), permitem relacionar classificações taxonômicas com estimativas dos períodos de origem de cada clado, permitindo assim, datar a origem dos genes desde a origem da vida até as espécies atuais, como exemplificado na Figura 2. Dessa forma, a filoestratigrafia vem sendo aplicada para investigar as mais variadas questões, tais como: modelos para origem gênica (CARVUNIS et al., 2012), desenvolvimento de tipos celulares a partir de células tronco (HEMMRICH et al., 2012), ciclo de vida celular (ABRUSÁN, 2013), desenvolvimento do cérebro e sistemas sensoriais em vertebrados (SESTAK et al., 2013; ŠESTAK; DOMAZET-LOŠO, 2015), velocidade evolutiva de genes em mamíferos (ALBÀ; CASTRESANA, 2005), relações entre expressão e idade gênica (WOLF et al., 2009), seleção adaptativa (CAI; PETROV, 2010), doenças em humanos (DOMAZET-LOSO; TAUTZ, 2008), desenvolvimento embrionário (DOMAZET-LOŠO; TAUTZ, 2010) e uso de códons (PRAT et al., 2009). Geralmente, para efetuar tal tarefa, realizam-se buscas por similaridade com o BLAST (ALTSCHUL et al., 1990) utilizando parâmetros de corte bastante relaxados. Um gene tem sua origem determinada no ancestral comum das linhagens em que é detectado. Dessa forma, a confiabilidade da datação depende da correta identificação dos homólogos (SCHLÖTTERER, 2015).

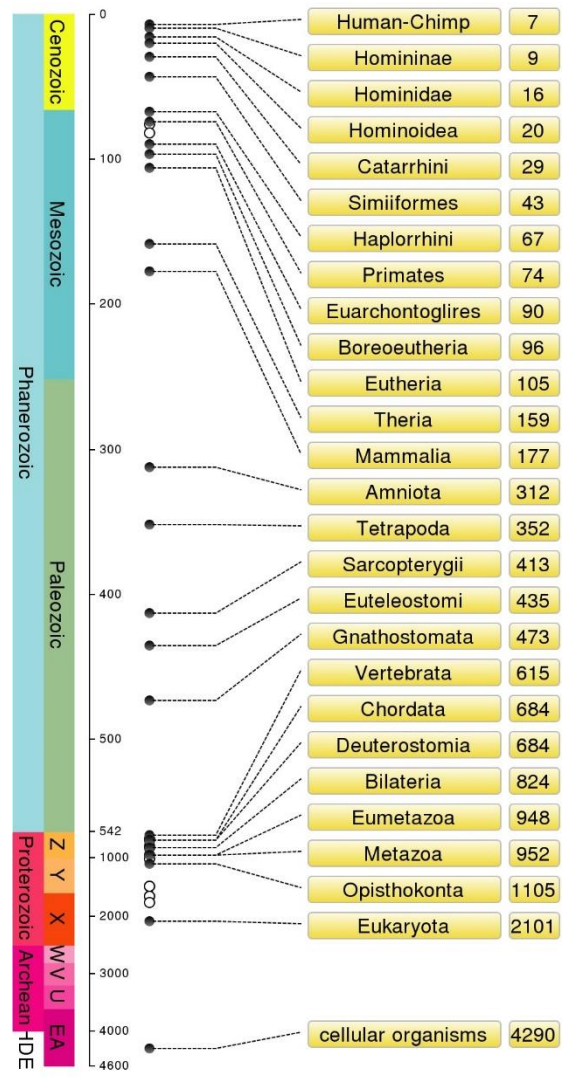


Figura 2 - Linha do tempo com origem dos clados da linhagem humana. As estimativas de origem de cada clado foram compiladas de diversos estudos por (BLAIR HEDGES; KUMAR, 2009). Os números representam milhões de anos. As barras laterais representam éons e eras geológicas. Figura obtida em timetree.org (acessado em 23/02/2017).

Homologia é um termo que se refere à existência de similaridade entre duas sequências de nucleotídeos ou aminoácidos, decorrente de uma origem evolutiva comum (KAHL, 2015). Logo, genes são ditos homólogos quando possuem um ancestral comum (ALTENHOFF; DESSIMOZ, 2012). Incluso nesse contexto, existem os conceitos de ortólogos e parálogos. Genes ortólogos são genes originados devido a um evento de especiação fruto de um único gene em um ancestral comum. Já genes parálogos são genes originados devido a um evento de duplicação dentro da mesma linhagem (ALTENHOFF; DESSIMOZ, 2012). Esses termos podem ser estendidos para cenários mais complexos como exemplificado na Figura 3. Nota-se, portanto, que devido à possibilidade de um evento de duplicação ter ocorrido posteriormente a um evento de especiação (acarretando em múltiplas cópias da mesma sequência em organismos diferentes), relações de ortologia

podem ser classificadas em um para um, um para muitos, muitos para um e muitos para muitos (KOONIN, 2005). Além disso, genes ortólogos tendem a ser mais conservados do que parálogos (CHEN; ZHANG, 2012; GABALDÓN; KOONIN, 2013) e a possuir funções semelhantes entre todos os organismos, enquanto parálogos tendem a ter funções distintas, a depender da divergência ou de quão remota é a duplicação (JENSEN, 2001). Essa equivalência funcional dos ortólogos pode ser útil quando se busca inferir funções de genes presentes em outros organismos (ALTENHOFF; DESSIMOZ, 2012).

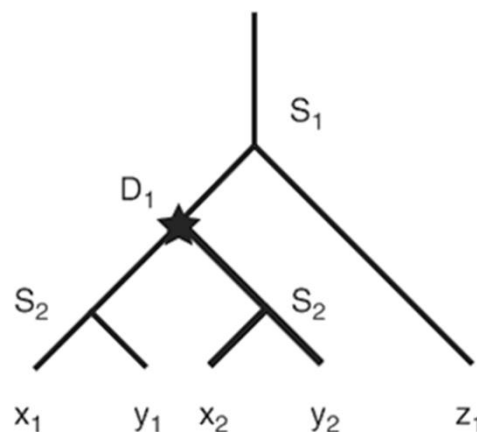


Figura 3 - Terminologia para paralogia e ortologia. A figura mostra um cenário envolvendo eventos de duplicação (D) e especiação (S). Ortólogos são genes que emergiram de um evento de especiação, como por exemplo os pares x₁,y₁ ou x₂,z₁. Parálogos são genes que emergiram de um evento de duplicação, como por exemplo os pares x₁,x₂ ou x₁,y₂. Dentro deste contexto, genes parálogos podem ser classificados como in-parálogos ou out-parálogos. In-parálogos, refere-se à genes parálogos em que houve uma duplicação após um evento de especiação de referência. Logo, os genes x₁ e x₂ são in-parálogos com relação ao evento de especiação S₁. Out-parálogos são parálogos onde ocorre um evento de duplicação antes de um evento de especiação de referência. Assim, x₁ e y₂ são out-parálogos em relação a S₂. Há também o conceito de co-ortólogos, que define a relação entre três genes onde dois deles são in-parálogos com relação a um evento de especiação para com o terceiro. No caso, x₁ e x₂ são co-ortólogos em relação a z. Adaptado de (ALTENHOFF; DESSIMOZ, 2012).

Métodos baseados em buscas por similaridade utilizando o BLAST oferecem a vantagem de poderem ser aplicadas na identificação de possíveis novos genes em larga escala. Entretanto são suscetíveis a falsos negativos (MCLYSAGHT; HURST, 2016). Como as buscas com o BLAST são baseadas em heurísticas, mesmo homólogos verdadeiros podem deixar de ser detectados quando a similaridade fica abaixo de certo valor de corte. Isto ocorre especialmente em genes pequenos e de evolução rápida (MOYERS; ZHANG, 2015, 2016). Por outro lado, o BLAST também infere muitos erros quando agrupa famílias gênicas, tornando a época de origem virtualmente mais remota. Assim, para identificar eventos de cladogênese, opções de métodos para inferência de ortólogos e técnicas de agrupamento mais sofisticadas podem melhorar a datação da origem gênica. Ortólogos, por usarem como base a comparação entre espécies, tendem a ser mais conservados do que

parálogos (CHEN; ZHANG, 2012; GABALDÓN; KOONIN, 2013). O uso de grupos de ortólogos para determinação de idade gênica tipicamente considera um grupo de ortólogos como todas as linhagens descendentes do nó de especiação mais distante, ou a divergência entre os dois homólogos mais distantes. Dessa forma, a idade do grupo de genes é definida pelo ancestral comum mais recente (*Most Recent Common Ancestor*, MRCA; ou LCA, *Lowest Common Ancestor*, termo derivado da teoria de grafos) das espécies encontradas nesse grupo. Um estudo recente comparou diversos métodos de inferência de ortólogos com foco na filoestratigrafia e notou diferenças significativas entre os algoritmos (LIEBESKIND; MCWHITE; MARCOTTE, 2016). Portanto, ao utilizar essas medições temos que ter em mente que são aproximações, e que seus resultados devem ser utilizados com cautela (DOMAZET-LOŠO et al., 2017).

1.2. Origem “*de novo*” de genes

A origem de genes *de novo* refere-se a quando uma região antes não codificadora passa a codificar um novo gene. Estudos recentes têm descoberto genes derivados a partir desse mecanismo em organismos desde plantas, leveduras, insetos, mamíferos, primatas e até humanos (MCLYSAGHT; GUERZONI, 2015). Genes originados *de novo* tendem a ser pequenos, simples e geralmente são expressos em tecidos associados à reprodução masculina (LEVINE et al., 2006; MCLYSAGHT; GUERZONI, 2015; REINHARDT et al., 2013; ZHAO et al., 2014). Além disso, estudos de genética de populações mostraram que genes originados *de novo* têm origem contínua e muitos ainda são polimórficos (ZHAO et al., 2014). Um dos modelos que busca explicar como os genes *de novo* se originam, utiliza a hipótese da utilização de próto-genes transitórios que são gerados pela tradução de regiões não gênicas. Essas atividades apresentam um potencial adaptativo e podem até mesmo ser mais prevalentes que esporádicas duplicações gênicas (CARVUNIS et al., 2012).

O nascimento de genes *de novo* envolve dois fatores: a aquisição de uma ORF (*Open Reading Frame*) e a adição de sinais regulatórios necessários para a transcrição (SCHLÖTTERER, 2015). No entanto, não há necessariamente uma ordem particular para que esses eventos ocorram (REINHARDT et al., 2013). Certos modelos descrevem regiões transcritas do genoma que adquirem ORFs através de mutações subsequentes (MCLYSAGHT; GUERZONI, 2015). Assim, sequências de RNAs, como lncRNAs, podem dar origem a novos peptídeos (REINHARDT et al., 2013; RUIZ-ORERA et al., 2014; XIE et al., 2012). Por outro lado, outros modelos consideram que ORFs pré-existentes podem eventualmente ser expressas e originarem novas proteínas (MCLYSAGHT; GUERZONI,

2015). Alterações no DNA em regiões próximas às ORFs favorecendo sítios de ligação para fatores de transcrição (elementos cis) podem induzir sua expressão (KAESSMANN, 2010).

Sequências de proteínas derivadas de novas ORFs podem ser consideradas arbitrárias (MCLYSAGHT; GUERZONI, 2015). Em casos em que a proteína passa a ser expressa abruptamente em altos níveis, é mais provável que ela tenha efeitos negativos e seja removida com a seleção (KAESSMANN, 2010; LEVINE et al., 2006). No entanto, cenários em que há uma grande quantidade de próto-genes expressos em níveis baixos podem apresentar maiores chances destes se tornarem genes fixos (WILSON; MASEL, 2011). Novas funções apresentadas por genes *de novo* podem ser raras, no entanto foi mostrado que sequências aleatórias ligadas a sequências funcionais podem causar efeitos favoráveis (HAYASHI et al., 2003). O mais esperado é que genes originados *de novo* se tornem funcionais, porém não necessariamente essenciais (MCLYSAGHT; HURST, 2016).

1.3. Aminoácidos e proteínas

O Dogma Central da Biologia Molecular descreve a maneira com que a informação é transferida através de um sistema biológico: o DNA pode ser copiado para DNA (Replicação), a informação do DNA pode ser copiada para um mRNA (Transcrição) e as proteínas podem ser sintetizadas a partir da informação no mRNA (Tradução) (CRICK, 1970). No processo de tradução, regiões no início (extremidade 5') e fim (extremidade 3') do mRNA não são traduzidas. Por esse motivo, essas regiões são conhecidas como UTRs (*Untranslated Regions*). UTRs possuem sítios de ligação e estruturas que influenciam na eficiência da tradução e na vida útil do mRNA (WILKIE; DICKSON; GRAY, 2003). O código genético é determinado por tríplexes de nucleotídeos (conhecidos como códons), cada tríplex codifica um aminoácido (com exceção dos códons de parada, que indicam o fim da tradução). Dessa forma, a combinação dos nucleotídeos gera 64 códons possíveis, destes, três são códons de parada no código universal, e os demais codificam 20 aminoácidos (LOH; SONG, 2010). Anticódon presentes em RNAs de transferência (tRNAs) reconhecem e interagem diretamente com os códons no mRNA (PHIZICKY; HOPPER, 2010). Os tRNAs carregados são ligados covalentemente a aminoácidos específicos, provendo uma ligação física entre o códon e seu respectivo aminoácido. Quando diferentes códons codificam o mesmo aminoácido, são ditos sinônimos (BUDD, 2012). Além disso, a escolha da base inicial utilizada na tradução determina a fase de leitura do código genético. Dada a composição em tripletos do códon, a tradução em fase pode ser entendida como aquela que tem o início da tradução na primeira posição do códon. Assim, mudanças na fase (ou *frameshifts*) virtualmente sempre resultam em proteínas totalmente diferentes (BUDD, 2012).

Os aminoácidos são compostos de um grupo amino ($-NH_2$), um grupo carboxila ($-COOH$) e uma cadeia lateral específica (BERG; TYMOCZKO; STRYER, 2002). Diferentes aminoácidos podem ser produzidos pela célula, no entanto, somente 22 podem ser incorporados “naturalmente” em proteínas (chamados proteogênicos). Destes, 20 são derivados a partir do código genético e dois podem ser incorporados por mecanismos especiais de tradução (AMBROGELLY; PALIOURA; SÖLL, 2007). Alterações químicas em aminoácidos após a tradução da proteína são conhecidas como “modificações pós-traducionais” ou PTMs (*Post-Translational Modifications*). Essas alterações podem ser introduzidas via modificações enzimáticas (FARLEY; LINK, 2009) ou não enzimáticas (CLOOS; CHRISTGAU, 2002), tendo papel crucial na mediação da função da proteína.

As proteínas são polímeros lineares de aminoácidos unidos por ligações peptídicas entre o nitrogênio do grupo amino e o carbono do grupo carboxila do aminoácido subsequente (BUDD, 2012). As proteínas apresentam uma enorme variedade de estruturas e funções (PETSKO; RINGE, 2004). Sua versatilidade funcional está relacionada à diversidade e a ordenação dos resíduos encontrados nas cadeias laterais (GUTTERIDGE; THORNTON, 2005). Em termos estruturais, proteínas podem ser classificadas em quatro níveis: estrutura primária, secundária, terciária e quaternária. A estrutura primária refere-se portanto à sequência de aminoácidos propriamente dita; a estrutura secundária corresponde ao arranjo espacial de aminoácidos próximos entre si, por exemplo, as alfa-hélices e as folhas beta; a estrutura terciária descreve a conformação da proteína inteira, onde as proteínas hidrossolúveis se enovelam em estruturas compactas com o interior (via de regra) apolar; e a estrutura quaternária são as conformações de duas ou mais cadeias polipeptídicas (BERG; TYMOCZKO; STRYER, 2002).

As conformações das proteínas são estabilizadas principalmente por forças não covalentes fracas como interações de van der Waals e ligações de hidrogênio. Em alguns casos, em que a proteína não está exposta a um ambiente redutor (que favorecem grupos $-SH$ livres ao invés de pontes S-S), pontes dissulfeto entre resíduos de cisteína podem ser encontradas. Padrões repetitivos originados pelas interações entre aminoácidos vizinhos e moléculas do solvente constituem a chamada estrutura secundária da proteína. São definidos três tipos principais de estrutura secundária (Figura 4): (i) as hélices, das quais a mais comum é a alfa-hélice; (ii) as folhas beta, que assumem duas formas quando se agrupam, paralela e antiparalela; e as alças (ou voltas), nas quais a cadeia é forçada a reverter de direção possibilitando conformações compactas (PETSKO; RINGE, 2004). A estrutura secundária de alça é mais simples, geralmente envolvendo quatro resíduos. Ela consiste em uma ligação de hidrogênio entre um oxigênio da carbonila de um resíduo (n) com o próton da amida de outro resíduo, três posições adiante ($n + 3$). Prolina e glicina são prevalentes neste tipo de conformação (CAMPBELL-PLATT, 2011).

As alfa-hélices são a estrutura secundária mais comum encontrada em proteínas enoveladas, possuindo uma estrutura espiralada envolvendo uma única cadeia polipeptídica (PETSKO; RINGE, 2004). A sua estabilidade é determinada por ligações de hidrogênio paralelas ao eixo da hélice, onde o oxigênio do grupo carbonila de cada resíduo (n) liga-se ao nitrogênio do grupo amida dos resíduos quatro posições à frente ($n + 4$), com exceção dos resíduos N e C da última volta, que fazem ligação apenas para o interior da estrutura. Para cada volta da hélice existem 3,6 resíduos sob os ângulos de $\varphi = -57^\circ$ e $\psi = -48^\circ$ (CAMPBELL-PLATT, 2011). A espiral pode adotar tanto sentido horário como anti-horário, no entanto, devido a todos os aminoácidos (com exceção da glicina) adotarem a configuração estereoquímica L (onde o grupo $\alpha\text{-NH}_3^+$ está projetado para a esquerda), acaba-se favorecendo com que as espirais adotem o sentido horário (PETSKO; RINGE, 2004). Alfa-hélices podem conter uma variedade de número de resíduos. Com mais ou menos propensão para alguns resíduos específicos (Tabela 1). A prolina, por não conter um grupo N-H, geralmente não é encontrada em alfa-hélices. No entanto, outras conformações de hélice podem apresentar conformações ricas em prolina, como a tripla hélice do colágeno e sequências de poliprolinas (PETSKO; RINGE, 2004).

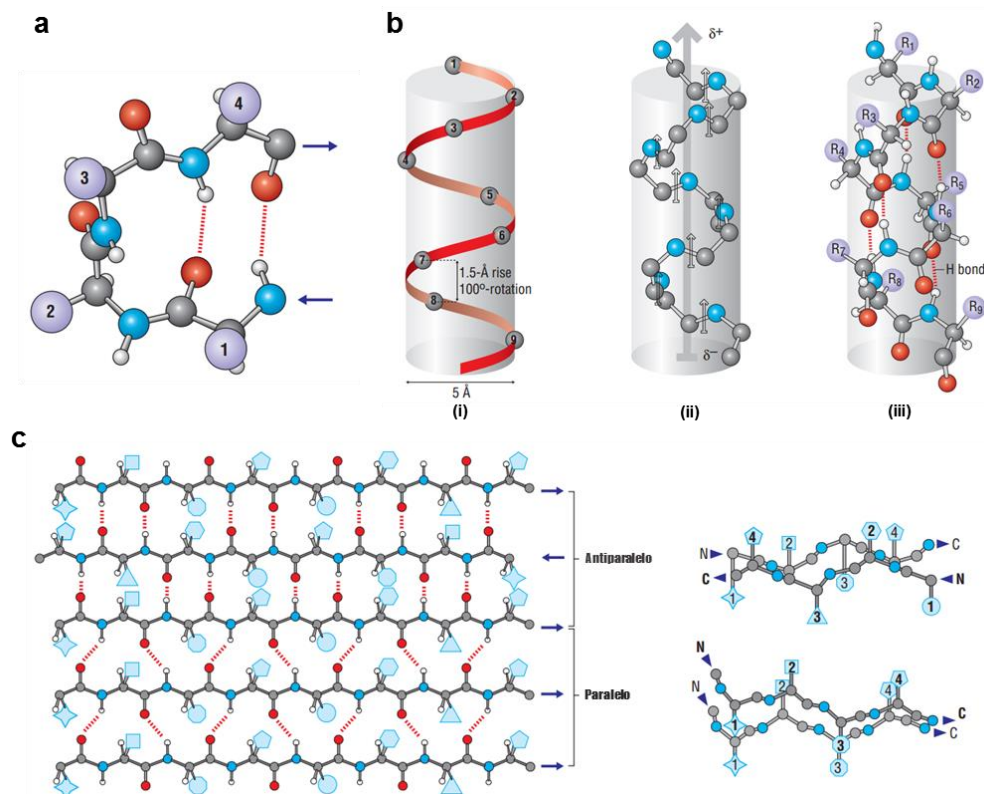


Figura 4 - Tipos de conformações de estrutura secundária. (a), mostra uma estrutura típica de alça. (b), mostra a estrutura de alfa-hélice: (i) exibe somente os carbonos alfa; (ii) mostra os carbonos do esqueleto com o momento dipolar; (iii) mostra a estrutura completa com esqueleto e ligações de hidrogênio. (c), mostra a estrutura de folha beta, a esquerda são mostradas folhas beta contendo segmentos paralelos e antiparalelos. É possível notar ligações de hidrogênio mais lineares em folhas antiparalelas. A direita é mostrada a visão das extremidades. Adaptado de (PETSKO; RINGE, 2004).

Ao contrário das hélices, as folhas beta apresentam uma conformação quase que totalmente estendida. Ligações de hidrogênio perpendiculares à cadeia se formam entre partes distintas de uma cadeia simples dobrada (ligação intracadeia) ou entre diferentes cadeias (ligação intercadeia) (CAMPBELL-PLATT, 2011). As folhas beta são formadas por duas ou mais fitas beta que podem inclusive estar localizadas a longas distâncias na sequência da proteína. As fitas podem correr na mesma direção (paralelas) ou em direções opostas (antiparalelas). Folhas formadas por fitas paralelas são raramente encontradas em tamanhos pequenos e frequentemente estão enterradas na proteína. Enquanto fitas antiparalelas geralmente estão expostas ao ambiente aquoso. Além disso, fitas paralelas são descontínuas, com conexões entre elas (comumente alfa-hélices). Já fitas antiparalelas geralmente apresentam alças nas conexões apresentando conformações mais estáveis. Aminoácidos como a valina e a isoleucina podem ser facilmente acomodados em estruturas de folha beta (PETSKO; RINGE, 2004).

Tabela 1 - Preferências conformacionais de aminoácidos.

Aminoácidos	Preferência*		
	Alfa-Hélice	Fita Beta	Alça
Glu / E	1.59	0.52	1.01
Ala / A	1.41	0.72	0.82
Leu / L	1.34	1.22	0.57
Met / M	1.30	1.14	0.52
Gln / Q	1.27	0.98	0.84
Lys / K	1.23	0.69	1.07
Arg / R	1.21	0.84	0.90
His / H	1.05	0.80	0.81
Val / V	0.90	1.87	0.41
Ile / I	1.09	1.67	0.47
Tyr / Y	0.74	1.45	0.76
Cys / C	0.66	1.40	0.54
Trp / W	1.02	1.35	0.65
Phe / F	1.16	1.33	0.59
Thr / T	0.76	1.17	0.90
Gly / G	0.43	0.58	1.77
Asn / N	0.76	0.48	1.34
Pro / P	0.34	0.31	1.32
Ser / S	0.57	0.96	1.22
Asp / D	0.99	0.39	1.24

*Frequências normalizadas para cada conformação calculadas a partir da fração de ocorrência de cada resíduo de aminoácido. A ocorrência aleatória de um aminoácido em particular em uma conformação tem valor igual à unidade. Valores maiores indicam a preferência para um tipo de estrutura secundária específica.

Adaptado de (WILLIAMS et al., 1987).

A organização destes elementos de estrutura secundária influi na estrutura terciária da proteína. Geralmente estas estruturas apresentam conformações específicas e estáveis.

Específicas no sentido de que uma outra molécula de proteína com a mesma sequência irá formar a mesma estrutura (ou pelo menos muito semelhante). E estável, pois a estrutura tende a permanecer em uma conformação semelhante durante o tempo. Domínios globulares são um exemplo desse tipo de estrutura (BUDD, 2012). Por outro lado, há casos em que proteínas não apresentam estruturas específicas e estáveis. Estas proteínas são chamadas intrinsecamente desordenadas (IDPs), em uma definição mais formal, são proteínas que não possuem um único ponto de equilíbrio estrutural bem definido (UVERSKY, 2014). Diversas funções são associadas às IDPs, principalmente envolvendo funções de sinalização (DYSON; JANE DYSON; WRIGHT, 2005). Além disso, são observados diversos vieses com relação a fatores evolutivos (WARD et al., 2004). Com a complexidade da evolução, torna-se possível que regiões desestruturadas, no entanto, se estruturam ao se ligarem a outras proteínas.

1.4. Métodos de predição de estrutura secundária

A predição de estrutura secundária tem suas origens em 1951 quando Linus Pauling, Robert Corey e Herman Branson realizaram a predição de conformações de hélices e folhas em proteínas antes mesmo da primeira estrutura ter sido determinada (PAULING; COREY; BRANSON, 1951). Desde então diversas novas metodologias e aprimoramentos foram desenvolvidos para a predição de estruturas secundárias. Três gerações de técnicas de predição podem ser classificadas (ROST, 2001). Na primeira geração, estruturas secundárias eram preditas de acordo com as frequências observadas de cada aminoácido para os respectivos estados estruturais (como exemplificado na Tabela 1). O método mais representativo desta geração foi desenvolvido por Chou e Fasman, e combinava propensões com heurísticas (CHOU; FASMAN, 1974). A segunda geração utilizava a ideia de janelas deslizantes para capturar informações das vizinhanças dos resíduos. Estes métodos aplicavam diferentes técnicas para inferir as conformações dos resíduos, tais como: informações estatísticas (ARNOLD et al., 1992; GARNIER; OSGUTHORPE; ROBSON, 1978; KABAT; WU, 1973), teoria de grafos (MITCHELL et al., 1990), redes neurais (BOHR et al., 1988; HOLLEY; KARPLUS, 1989), regressão logística (MUGGLETON; KING; STERNBERG, 1992) e métodos de vizinhança (YI; LANDER, 1993). A terceira geração agregou informações evolutivas derivadas de alinhamentos múltiplos de sequências (ROST; SANDER, 1993b; ZVELEBIL et al., 1987). Técnicas empregadas neste contexto envolvem máquinas de vetores de suporte (*Support Vector Machines*, SVMs) (HUA; SUN, 2001; WARD et al., 2003), redes Bayesianas (YAO; ZHU; SHE, 2008), modelos de Markov (AYDIN; ALTUNBASAK; BORODOVSKY, 2006), campos aleatórios condicionais

(LIU et al., 2004) e, apresentando melhores resultados, redes neurais (DOR; ZHOU, 2007; HEFFERNAN et al., 2015; JONES, 1999; ROST; SANDER, 1993b; WANG et al., 2016a).

Hoje, a acurácia dos métodos de predição de estrutura secundária se aproxima do seu limite teórico (entre 88 e 90%) (YANG et al., 2016). Métodos que representam o estado da arte em predição de estrutura secundária utilizam perfis de sequências, tais como o das matrizes de pontuação posição-específica (*Position-Specific Scoring Matrix*, PSSM) do PSI-BLAST (ALTSCHUL et al., 1997), para obter informações de estruturas conservadas entre sequências homólogas. Métodos como o SSpro (MAGNAN; BALDI, 2014) alcançam acurácias ainda mais altas ao simplesmente utilizar as próprias estruturas secundárias presentes em sequências homólogas. Entretanto, infelizmente, a maior parte das sequências de proteínas ainda não possuem estruturas de homólogos definidas, limitando tal abordagem (YANG et al., 2016). Dessa forma, métodos de predição sofisticados são necessários para obter resultados satisfatórios na maioria dos casos.

2. Justificativa e objetivo

São conhecidos diversos fatores relacionados ao período de origem dos genes, desde alterações no uso de códons (PRAT et al., 2009) até o surgimento de genes cancerígenos (DOMAZET-LOSO; TAUTZ, 2010). Questões relacionadas à estrutura de proteínas também têm sido investigadas, geralmente com foco em proteínas intrinsecamente desordenadas (WARD et al., 2004). Entretanto, pouco tem sido estudado com respeito às relações entre a composição de elementos estruturais de proteínas e sua ancestralidade.

Com o aumento de informações de genomas completos disponíveis e o desenvolvimento de novas tecnologias e metodologias de bioinformática, torna-se possível realizar estudos em larga escala para investigar mudanças macroevolutivas ao nível de estruturas de proteínas. Utilizando informações de estrutura secundária, podemos investigar mudanças na composição de elementos de estrutura secundária em proteínas com origem antiga e recente.

Podemos também, examinar se fatores determinantes de estrutura, como o uso de aminoácidos e regiões de domínios, estão envolvidos nas diferenças observadas e qual é o seu papel na evolução das proteínas. Além disso, podemos investigar os mecanismos que regem a origem de novos genes. Predições de estrutura secundária podem ser utilizadas para investigar a viabilidade de proteínas ter origem a partir de regiões não codificadoras. Levantando questionamentos com respeito aos modelos que explicam a origem *de novo* de genes.

Dessa forma este trabalho tem como objetivo principal, investigar questões macroevolutivas relacionadas a estrutura secundária de proteínas. Para tal, utilizaremos informações de estruturas definidas experimentalmente, bem como estruturas provenientes de métodos de predição, que serão relacionadas à informações funcionais e estruturais. Além disso, utilizaremos técnicas de filoestratigrafia para classificar proteomas de organismos de acordo com a sua idade evolutiva. Com isso será possível analisar se existem diferenças ao longo da evolução, quais são os fatores que influenciam estas diferenças, quais funções e processos biológicos estão envolvidos e como isso se estende para a compreensão de processos de origem de novas proteínas.

3. Materiais e métodos

3.1. Dados de estruturas de proteínas

Dados de sequências do PDB (*Protein Data Bank*) e suas sequências de estrutura secundária (também conhecidas como DSSP) foram obtidas através da base de dados PDBFINDER2 (TOUW et al., 2015) (acessível em <http://swift.cmbi.ru.nl/gv/pdbfinder/>). Esta base corrige alguns problemas de inconsistência encontrados em arquivos do PDB além de ser de fácil utilização. A base de dados utilizada foi obtida em julho de 2016 e possui 119.590 entradas únicas do PDB e um total de 323.830 sequências de cadeias. Destas sequências, foram removidas entradas sem informações de estrutura secundária (indicada no campo DSSP, do arquivo "PDBFIND2.TXT"), sequências com menos de 50 resíduos e entradas de estruturas redundantes. Para seleção das estruturas não redundantes utilizamos o serviço VAST (*Vector Alignment Search Tool*) (MADEJ et al., 2014), que disponibiliza essas informações pré-computadas para a base de dados do PDB (nr-PDB, acessível em <https://structure.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html>, dados de abril de 2016). Utilizamos o parâmetro de menor valor de similaridade disponível (valor $p = 10^{-7}$) para obter a base menos redundante possível. Após os filtros restaram 15.362 sequências, contendo 14.485 PDBs únicos.

Os dados de estrutura secundária do DSSP possuem oito estados de conformações (Tabela 2) que foram convertidas para três estados de acordo com estudos anteriores (JONES, 1999; ROST; SANDER, 1993a). Dessa forma, H e G são consideradas hélices (H); E e B são consideradas fitas (E); e todos os demais são consideradas irregulares (C).

Tabela 2 - Estados de estrutura secundária definidas no DSSP.

Estrutura secundária	Descrição
G	Hélice de 3 voltas (hélice 3_{10}). Comprimento mínimo de 3 resíduos.
H	Hélice de 4 voltas (alfa-hélice). Comprimento mínimo de 4 resíduos.
I	Hélice de 5 voltas (hélice π). Comprimento mínimo de 5 resíduos.
E	Fita beta paralela ou antiparalela. Comprimento mínimo de 2 resíduos.
B	Resíduo isolado de fita-beta.
T	Volta com pontes de hidrogênio. (3, 4 ou 5 voltas).
S	Dobramento ou curva. Não realiza pontes de hidrogênio.
C	Desordenada (<i>coil</i>). Resíduos que não são classificados em nenhuma das anteriores.

3.2. Predições de estruturas secundárias

Utilizamos três ferramentas para predição de estrutura secundária: PSIPRED v4.0 (JONES, 1999); DeepCNF v1.02 (WANG et al., 2016b); e SSpro 5.2 SCRATCH release 1.1 (MAGNAN; BALDI, 2014). O funcionamento das três ferramentas é semelhante. Buscas utilizando o PSI-BLAST (ALTSCHUL et al., 1997) são realizadas contra um banco de dados não redundante para confecção de matrizes de pontuação por posição específica. Estas matrizes são então utilizadas como características de entrada em uma rede neural treinada para classificação dos estados de estrutura secundária. Utilizamos somente as predições para três estados (entretanto o SSpro e o DeepCNF têm a opção para classificação em oito estados). Tanto o SSpro como o DeepCNF disponibilizam as bases de dados não redundantes em seus pacotes. Para o SSpro utilizamos uma versão do UniRef50 (SUZEK et al., 2015) fornecida no pacote. Já o DeepCNF fornece versões do nr (COORDINATORS, 2016), e utilizamos a versão com 90% de identidade (nr90). Para o PSIPRED utilizamos também o nr90. As predições foram realizadas tanto para estruturas não redundantes do PDB (descrito no item anterior) como para dados de proteomas completos.

3.3. Determinação do consenso das predições

Para casos que temos a predição das proteínas determinada pelos três métodos, definimos uma predição consenso de acordo com (ALBRECHT et al., 2003). Considerando que cada resíduo pode apresentar um dos três estados estruturais (H, E, C), o resíduo consenso é definido como aquele que ocorre na maioria dos preditores. Três possibilidades podem ocorrer: (i) casos de maioria absoluta, onde os três preditores concordam na estrutura do resíduo; (ii) casos de maioria simples, onde dois preditores apresentam a mesma estrutura e um preditor apresenta uma estrutura diferente; e (iii) casos de empate, quando cada preditor apresenta uma estrutura diferente. Nestes casos o consenso é definido como estrutura irregular (C).

3.4. Dados de proteomas completos e de genes com origem *de novo*

Sequências de proteomas completos foram obtidas através da base do consórcio *Quest for Orthologs* (QfO) que disponibiliza 66 proteomas de referência (ALTENHOFF et al., 2016; SONNHAMMER et al., 2014) (Tabela 3). Utilizamos a versão 2016_04, acessível em http://www.ebi.ac.uk/reference_proteomes. A base de dados fornece tanto sequências de aminoácidos como de nucleotídeos, além de arquivos com identificadores mapeados referentes a outras bases de dados. Sequências de aminoácidos foram utilizadas

diretamente nas predições de estrutura secundária. Já as sequências de nucleotídeos, foram utilizadas para análises de ORFs em regiões de *frameshifts* e antisenso. Selecionamos ORFs com pelo menos 150 pares de bases que foram posteriormente traduzidas em sequências de aminoácidos para realizarmos a predição.

Tabela 3 - Lista das 66 espécies de referência utilizadas em *Quest for Orthologs* (QfO).

Espécie	NCBI TaxId	Domínio	Táxon	Genes
<i>Halobacterium salinarum</i>	64091	Archaea	Archaea	2415
<i>Korarchaeum cryptofilum</i>	374847	Archaea	Archaea	1599
<i>Methanocaldococcus jannaschii</i>	243232	Archaea	Archaea	1787
<i>Methanosarcina acetivorans</i>	188937	Archaea	Archaea	4296
<i>Sulfolobus solfataricus</i>	273057	Archaea	Archaea	2924
<i>Thermococcus kodakaraensis</i>	69014	Archaea	Archaea	2290
<i>Aquifex aeolicus</i>	224324	Bacteria	Bacteria	1552
<i>Bacillus subtilis</i>	224308	Bacteria	Bacteria	4197
<i>Bacteroides thetaiotaomicron</i>	226186	Bacteria	Bacteria	4775
<i>Bradyrhizobium japonicum</i>	224911	Bacteria	Bacteria	8024
<i>Chlamydia trachomatis</i>	272561	Bacteria	Bacteria	895
<i>Chloroflexus aurantiacus</i>	324602	Bacteria	Bacteria	3819
<i>Deinococcus radiodurans</i>	243230	Bacteria	Bacteria	3079
<i>Dictyoglomus turgidum</i>	515635	Bacteria	Bacteria	1731
<i>Escherichia coli</i>	83333	Bacteria	Bacteria	4306
<i>Fusobacterium nucleatum</i>	190304	Bacteria	Bacteria	2043
<i>Geobacter sulfurreducens</i>	243231	Bacteria	Bacteria	3395
<i>Gloeobacter violaceus</i>	251221	Bacteria	Bacteria	4318
<i>Leptospira interrogans</i>	189518	Bacteria	Bacteria	3418
<i>Mycobacterium tuberculosis</i>	83332	Bacteria	Bacteria	3987
<i>Pseudomonas aeruginosa</i>	208964	Bacteria	Bacteria	5550
<i>Rhodopirellula baltica</i>	243090	Bacteria	Bacteria	6999
<i>Streptomyces coelicolor</i>	100226	Bacteria	Bacteria	8005
<i>Synechocystis</i>	1111708	Bacteria	Bacteria	3424
<i>Thermodesulfobrio yellowstonii</i>	289376	Bacteria	Bacteria	1970
<i>Thermotoga maritima</i>	243274	Bacteria	Bacteria	1851
<i>Dictyostellium discoideum</i>	44689	Eukaryota	Amoebozoa	12731
<i>Monosiga brevicollis</i>	81824	Eukaryota	Choanozoa	9188
<i>Candida albicans</i>	237561	Eukaryota	Dikarya	8264
<i>Cryptococcus neoformans</i>	214684	Eukaryota	Dikarya	6602
<i>Neosartorya fumigata</i>	330879	Eukaryota	Dikarya	9649
<i>Neurospora crassa</i>	367110	Eukaryota	Dikarya	9756
<i>Phaeosphaeria nodorum</i>	321614	Eukaryota	Dikarya	15993
<i>Saccharomyces cerevisiae</i>	559292	Eukaryota	Dikarya	6721
<i>Schizosaccharomyces pombe</i>	284812	Eukaryota	Dikarya	5121
<i>Sclerotinia sclerotiorum</i>	665079	Eukaryota	Dikarya	14400
<i>Ustilago maydis</i>	237631	Eukaryota	Dikarya	6788
<i>Yarrowia lipolytica</i>	284591	Eukaryota	Dikarya	6448
<i>Arabidopsis thaliana</i>	3702	Eukaryota	Embryophyta	27064
<i>Physcomitrella patens</i>	3218	Eukaryota	Embryophyta	34793
<i>Anopheles gambiae</i>	7165	Eukaryota	Eumetazoa	11988
<i>Bos taurus</i>	9913	Eukaryota	Eumetazoa	20055
<i>Branchiostoma floridae</i>	7739	Eukaryota	Eumetazoa	28538
<i>Caenorhabditis elegans</i>	6239	Eukaryota	Eumetazoa	20137
<i>Canis familiaris</i>	9615	Eukaryota	Eumetazoa	19644
<i>Ciona intestinalis</i>	7719	Eukaryota	Eumetazoa	16641
<i>Danio rerio</i>	7955	Eukaryota	Eumetazoa	24821
<i>Drosophila melanogaster</i>	7227	Eukaryota	Eumetazoa	13707
<i>Gallus gallus</i>	9031	Eukaryota	Eumetazoa	15775
<i>Homo sapiens</i>	9606	Eukaryota	Eumetazoa	21006
<i>Ixodes scapularis</i>	6945	Eukaryota	Eumetazoa	20463
<i>Macaca mulatta</i>	9544	Eukaryota	Eumetazoa	21726
<i>Monodelphis domestica</i>	13616	Eukaryota	Eumetazoa	21181
<i>Mus musculus</i>	10090	Eukaryota	Eumetazoa	22136
<i>Nematostella vectensis</i>	45351	Eukaryota	Eumetazoa	24428
<i>Ornithorhynchus anatinus</i>	9258	Eukaryota	Eumetazoa	21122
<i>Pan troglodytes</i>	9598	Eukaryota	Eumetazoa	18656
<i>Rattus norvegicus</i>	10116	Eukaryota	Eumetazoa	21330
<i>Schistosoma mansoni</i>	6183	Eukaryota	Eumetazoa	10716
<i>Takifugu rubripes</i>	31033	Eukaryota	Eumetazoa	18492
<i>Xenopus tropicalis</i>	8364	Eukaryota	Eumetazoa	18252
<i>Giardia intestinalis</i>	184922	Eukaryota	Excavata	7154
<i>Leishmania major</i>	5664	Eukaryota	Excavata	8031
<i>Trichomonas vaginalis</i>	5722	Eukaryota	Excavata	50188
<i>Plasmodium falciparum</i>	36329	Eukaryota	Halvaria	5159
<i>Thalassiosira pseudonana</i>	35128	Eukaryota	Halvaria	11706

Sequências de genes com origem *de novo* foram obtidas através do Ensembl Browser (acessível em <http://www.ensembl.org/index.html>) a partir dos identificadores obtidos nos artigos referenciados na lista compilada por (MCLYSAGHT; GUERZONI, 2015) (Tabela 4).

Tabela 4 - Genes com origem *de novo* descobertos recentemente.

Organismos	Número de genes	Exemplos e comentários	Referências
Primatas	15	PART1; carcinogênese da próstata	TOLL-RIERA et al., 2009
Hominóides	24	Transcrição no cerebelo	XIE et al., 2012
Hominídeos	1	NCYM; Patogênese do neuroblastoma	SUENAGA et al., 2014
<i>H. sapiens</i>	3	CLLU1; suprarregulado em leucemia linfocítica crônica	KNOWLES; MCLYSAGHT, 2009
<i>H. sapiens</i>	1	FLJ33706(C20orf203); expresso no cérebro; proteína encontrada em neurônios	LI et al., 2010
<i>H. sapiens</i>	60	----	WU; IRWIN; ZHANG, 2011
<i>H. sapiens</i>	1	PBOV1; associado a câncer de mama e glioma	SAMUSIK et al., 2013
<i>H. sapiens</i>	1	ESRG; essencial para a manutenção da pluripotência	WANG et al., 2014

Adaptado de (MCLYSAGHT; GUERZONI, 2015).

3.5. Dados de domínios estruturais

Informações sobre domínios estruturais foram obtidas através da base de dados CATH (SILLITOE et al., 2015). Utilizamos a versão 4.1, obtida em julho de 2016 (acessível em <http://www.cathdb.info/download>). O CATH classifica domínios estruturais de forma hierárquica a partir de dados obtidos no PDB. Domínios individuais que apresentam evidências de possuírem o mesmo ancestral comum são agrupados em uma superfamília de homólogos (nível H). Quando membros de uma superfamília apresentam a mesma conformação, estas são classificadas com a mesma topologia (nível T). Topologias que apresentam arranjos de estruturas secundárias semelhantes são agrupadas na mesma arquitetura (nível A). Note-se que o termo arquitetura não se refere a conjuntos de domínios, como em outras bases. E, por fim, 40 arquiteturas são classificadas em quatro classes principais (nível C) baseadas de acordo com seu conteúdo de estrutura secundária (principalmente alfa, principalmente beta, misto de alfa e beta e pouca estrutura secundária).

A identificação dos domínios CATH utiliza um código individual baseado no código do PDB. Assim, por exemplo, a entrada “1gk8E02” é um domínio encontrado na cadeia “E” da estrutura “1gk8” do PDB. O número no final é sequencial e indica a ocorrência de domínios na mesma cadeia. Já a classificação deste domínio na hierarquia do CATH segue um código com quatro números que representam cada um dos níveis. Dessa forma, o domínio “1gk8E02”, por exemplo, é classificado na classe 3 (misto de alfa e beta), arquitetura 3.20 (barril alfa beta), topologia 3.20.20 (barril TIM) e superfamília de homólogos

3.20.20.110 (rubisco). Utilizamos neste trabalho somente informações sobre classes e arquiteturas CATH. A descrição contendo os identificadores do PDB e seus respectivos identificadores foi obtida no arquivo “cath-domain-list-v4_1_0.txt” e a descrição dos códigos foi obtida através do arquivo “cath-names-v4_1_0.txt”.

Para as sequências de proteomas completos, que não possuem informações sobre domínios prontamente acessíveis, utilizamos a base de dados Gene3D (LAM et al., 2016) (acessível em <http://download.cathdb.info/gene3d/>). O Gene3D fornece predições de domínios para sequências de proteínas do Ensembl e UniProtKB. As predições são determinadas através de perfis de modelos ocultos de Markov (HMMs). Utilizamos a versão 14.0.0 que segue como referência a versão 4.1 do CATH. O arquivo “arch_schema_cath.tsv” possui informações dos números de acesso do UniProt relacionados com os domínios preditos. Neste arquivo são fornecidas tanto as coordenadas do domínio na sequência como os identificadores das superfamílias do CATH. Em nossas análises definimos regiões intradomínios e extradomínios, baseado nestas coordenadas. Consideramos como regiões intradomínios somente segmentos de domínios individuais, logo as medidas utilizadas em nossas análises (porcentagem de estrutura secundária e uso de aminoácidos) são referentes a cada domínio específico. Já para regiões extradomínio, consideramos todas os segmentos da proteína que não foram mapeados como domínios. Assim, para cada proteína temos somente uma sequência de extradomínio (com exceção de proteínas em que o domínio corresponde à totalidade da sequência). Sequências com menos de 50 resíduos de aminoácidos, tanto de intradomínios como de extradomínios, foram descartadas das análises.

3.6. Dados de origem dos genes

As informações sobre origem de genes foram obtidas a partir do estudo de (LIEBESKIND; MCWHITE; MARCOTTE, 2016) (acessível em <http://geneages.org/>). Neste estudo, a idade dos genes é estimada pelo consenso de 13 métodos para inferência de ortologia. A origem de uma proteína é definida determinando-se o ancestral comum mais recente (MRCA) do grupo de ortólogos ao qual pertence. O consenso das idades é determinado através da moda da distribuição dada pelos métodos. A classificação segue a árvore de espécies de referência do SwissTree (BOECKMANN et al., 2015). Utilizamos em nossas análises as estimativas para *Homo sapiens*. Dessa forma, o proteoma de humano é categorizado em oito idades: *Cellular_organisms*, *Euk+Bacteria*, *Euk_Archaea*, *Eukaryota*, *Opisthokonta*, *Eumetazoa*, *Vertebrata* e *Mammalia*. Todos estes clados são aninhados, com exceção de *Euk+Bacteria*, que representa genes presentes em eucariotos e bactérias, porém ausentes em arqueias. Quando analisamos somente proteínas humanas com estrutura no PDB, onde

há pouca amostragem, transformamos os oito clados em apenas quatro: unimos os clados *Cellular_organisms*, *Euk+Bacteria* e *Euk_Archaea* em um clado que denominamos *Euk_Bac_Arch*; unimos as proteínas de *Eumetazoa* com *Opisthokonta* em um clado chamado *Opistho_Eumeta*; e passamos os genes de *Mammalia* para *Vertebrata* (*Verteb_Mammalia*).

3.7. Integração e manipulação dos dados

Utilizamos neste trabalho dados de diversas fontes distintas. A manipulação de dados de sequências, a organização de *pipelines* para execução de ferramentas e a manipulação de arquivos de dados brutos (*raw data*) foram realizadas utilizando-se o MATLAB. Parte da integração entre identificadores foi obtida através da base de dados SIFTS (VELANKAR et al., 2013). Utilizamos o arquivo “pdb_chain_cath_uniprot.tsv” para relacionar identificadores do PDB com o UniProt e o arquivo “pdb_chain_taxonomy.tsv” para relacionar identificadores do PDB com o *NCBI Taxonomy*. Ambos os arquivos foram obtidos em agosto de 2016 (acessível em <https://www.ebi.ac.uk/pdbe/docs/sifts/quick.html>). As informações de interesse destes e outros arquivos foram transpostas em arquivos tabulados e armazenados em um banco de dados relacional (MySQL) para melhor manipulação. Informações de domínios taxonômicos foram obtidas a partir dos identificadores do NCBI utilizando a função “*classification*” do pacote “*taxize*” do R. A manipulação final dos dados, a apresentação em gráficos, e as análises estatísticas foram realizadas com R. Gráficos de explosão solar foram feitos com Microsoft Excel.

3.8. Análise de enriquecimento funcional

Sequências foram enriquecidas usando GOrilla (EDEN et al., 2009). Utilizamos o modo com duas listas de genes, sempre tendo como referência (*background*) a lista do proteoma humano completo. Utilizamos como parâmetro de corte para exibição dos termos enriquecidos, somente aqueles identificados com valor $p < 0,01$. Realizamos o enriquecimento de sequências humanas de acordo com as idades evolutivas e também analisamos sequências com baixa composição estrutural (sequências com menos de 20% de alfa-hélices e fitas beta). As listas dos termos foram então filtradas usando REVIGO (SUPEK et al., 2011). Utilizamos o SimRel como semântica de similaridade e corte de similaridade igual a 0.5. Os gráficos de mapas de árvore foram obtidos também com REVIGO.

4. Resultados

4.1. Análise dos métodos de predição de estrutura secundária

Neste projeto temos duas fontes de informação sobre estruturas secundárias em proteínas, DSSP e predição. O DSSP é um algoritmo que extrai a informação de estrutura secundária a partir das coordenadas atômicas das proteínas cristalizadas presentes no banco de dados PDB. O termo DSSP também se refere ao banco de dados com informações de estrutura secundária referente a todo o banco do PDB (KABSCH; SANDER, 1983). Predições de estrutura secundária utilizam técnicas de aprendizado de máquina para determinar a provável conformação estrutural secundária, sendo uma valiosa opção para ser aplicada ao estudo de proteomas completos. Utilizamos neste trabalho três ferramentas de predição cujos resultados representam o estado da arte em predição de estrutura secundária: PSIPRED versão 4.0 (JONES, 1999); DeepCNF versão 1.02 (WANG et al., 2016b); e SSpro versão 5.2 (MAGNAN; BALDI, 2014).

Definimos também um consenso das predições utilizando as informações dos três métodos determinada de acordo com (ALBRECHT et al., 2003). Dado que, para cada resíduo, existem três possibilidades de estrutura secundária: alfa-hélice (H), fita beta (E) e irregular (C). Determinamos o consenso do resíduo como a estrutura indicada pela maioria dos preditores. Sendo assim, casos de maioria absoluta entre os três métodos são indicados como 3:0 (resíduo com a mesma estrutura nos três métodos de predição). Casos de maioria simples são indicados como 2:1. E casos em que os três métodos apresentam predição para estruturas diferentes são indicados como 1:1:1, nessas situações o consenso é determinado como irregular (C). A Tabela 5 mostra as porcentagens de ocorrência para cada caso nos diferentes conjuntos de dados utilizados neste trabalho (ver Materiais e Métodos para mais detalhes). Notamos que são raros os casos em que não há consenso entre os três métodos (1:1:1), ocorrendo entre 0,49% e 2,05% dependendo do conjunto de dados utilizado. Como esperado, o valor mais baixo foi obtido para sequências de aminoácidos de estruturas do PDB, para as quais os métodos foram aprimorados. Sequências de proteomas também apresentaram uma boa concordância entre os métodos, com casos de maioria absoluta (3:0) ocorrendo em mais de 79% das vezes e casos de indeterminação (1:1:1) em torno de 0,5%. Já sequências simuladas por tradução de regiões não codificadoras ou por permutação de resíduos em proteínas verdadeiras apresentaram maior discordância entre as predições, com situações de 1:1:1 em mais de 1% dos casos e de maioria absoluta (3:0) inferiores.

Tabela 5 - Frequência do total de resíduos nos três tipos de situações para formação de consenso.

Conjunto de dados analisado	3:0	2:1	1:1:1
PDB	78,71%	20,80%	0,49%
<i>Hsa</i> - Proteoma	79,53%	19,94%	0,53%
<i>Hsa</i> - Proteoma permutado	68,26%	30,09%	1,65%
<i>Hsa</i> - Frameshift	73,69%	24,87%	1,44%
<i>Eco</i> - Proteoma	79,67%	19,81%	0,52%
<i>Eco</i> - Proteoma permutado	66,44%	31,67%	1,88%
<i>Eco</i> - Frameshift	67,70%	30,25%	2,05%
<i>Eco</i> - Intergenica	67,77%	30,24%	1,98%

3:0 representa posições com consenso entre os três métodos.

2:1 representa situações onde dois métodos concordam e um discorda sobre estrutura predita na posição. Nestes casos a estrutura do consenso é a indicada pela maioria dos métodos.

1:1:1 representa situações em que não há consenso, cada método prediz uma das três possibilidades (H, E, C). Nestes casos o consenso é determinado como resíduo irregular (C).

Para verificar o desempenho da predição de estruturas secundárias usualmente utilizam-se duas formas de medida: Q3, que é a porcentagem de resíduos classificados corretamente em cada estado; e SOV (*Segment Overlap*) cuja medida considera o contexto de sobreposição de segmentos estruturais (ZEMLA et al., 1999). Ambas as medidas são referentes a comparação entre estrutura observada (DSSP) e predita. Realizamos a predição com as três ferramentas isoladas ou utilizando o consenso para sequências do PDB e analisamos as medidas Q3 e SOV para cada método (Tabela 6). As medidas são calculadas para todos os pares de sequências (predição contra DSSP) e por fim calculamos a média para cada método. Como resultado, obtivemos valores considerados excelentes em todos os métodos de predição (MAGNAN; BALDI, 2014), especialmente com o SSpro, que apresentou escores de Q3 e SOV mais de 10% superiores aos demais preditores. PSIPRED e DeepCNF apresentaram resultados bastante semelhantes, com Q3 por volta de 83% e SOV em torno de 80%. Já o consenso apresentou valores entre o SSpro e os demais, com Q3 igual a 87,10% e SOV igual a 84,50%. A vantagem do SSpro neste quesito é explicada devido a utilização de modelos de estruturas resolvidas do PDB que aumentam a qualidade da predição para estas sequências. No entanto, foi observado que na ausência dessa informação seu desempenho fica limitado (WANG et al., 2016b).

Tabela 6 - Acurácia dos métodos de predição de estrutura secundária.

Método	Q3	SOV
PSIPRED	83,67	80,40
DeepCNF	82,92	80,17
SSpro	94,70	93,54
Consenso	87,10	84,50

Em seguida verificamos a composição estrutural (i.e., a porcentagem de resíduos de alfa-hélice, fita beta e irregular) de sequências de estruturas PDB e das suas respectivas

predições (três métodos e consenso). A Figura 5 mostra um gráfico de violino com as porcentagens de resíduos em alfa-hélice, fita beta e irregular de cada sequência nos conjuntos de dados. Com isso, é possível observar que todos os métodos de predição apresentam distribuições e medianas semelhantes entre si e com o PDB, principalmente para alfa-hélices e fitas beta. Testes estatísticos suportam essa informação. Aplicamos o teste de Kruskal-Wallis e contraste de Dunnett (valor $p < 0.001$) sobre as porcentagens de estrutura secundária calculadas para cada sequência nos diferentes conjuntos e notamos que predições com SSpro não apresentaram diferenças significativas em relação ao PDB nos três estados estruturais. Já o PSIPRED não apresentou diferenças com o PDB para distribuições de alfa-hélices enquanto que, para fitas beta, não houve diferenças significativas em relação às predições com DeepCNF e o consenso. A diferença de composição total entre alfa-hélices foi maior entre o DeepCNF e o PDB (quase 4%), já para fitas beta o método que apresentou maior diferença com o PDB foi o PSIPRED com cerca de 3% de diferença (Tabela 7). Para resíduos com estrutura irregular, os dados do PDB apresentaram composição total de 40,68%, enquanto o DeepCNF apresentou a maior diferença, com 46,88% (aproximadamente 6% de diferença). Já o SSpro obteve diferenças inferiores a 1% nas três estruturas, enquanto o consenso das predições obteve diferenças entre 1,5% a 4% aproximadamente.

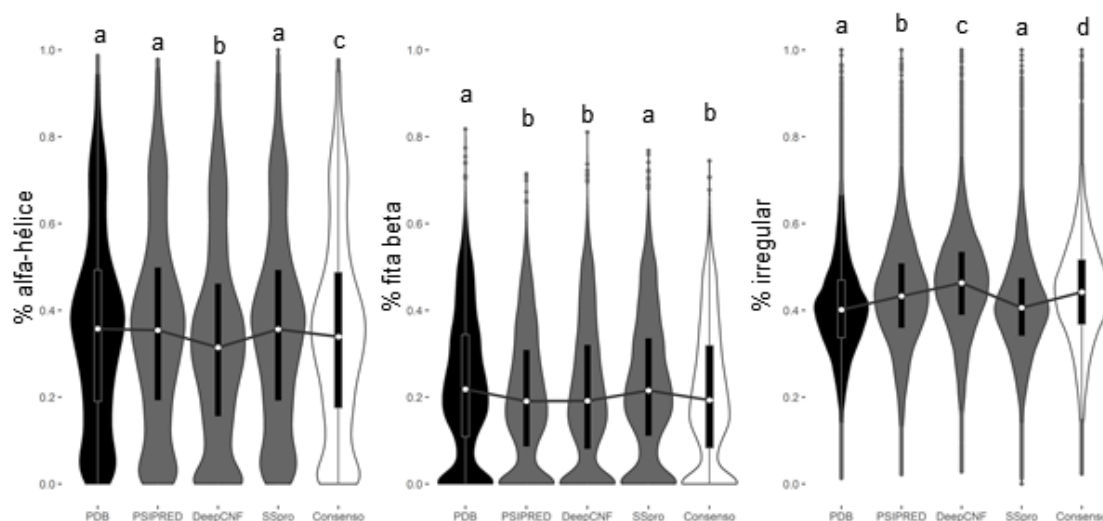


Figura 5 - Composição da estrutura secundária em estruturas do PDB e métodos de predição. Distribuição da porcentagem de alfa-hélice, fita beta e irregular para cada sequência. No gráfico, PDB refere-se às estruturas definidas experimentalmente extraídas via DSSP, já as predições são referentes às sequências primárias do PDB. Letras diferentes indicam grupos com diferenças estatisticamente significantes usando o teste de Kruskal-Wallis e contraste de Dunnett (valor $p < 0,001$).

Tabela 7 - Composição de estrutura secundária do PDB e métodos de predição.

Método	% de alfa-hélice*	% de fita beta*	% irregular*
PDB-DSSP	36,22%	23,09%	40,68%
PSIPRED	35,87%	20,08%	44,04%
DeepCNF	32,49%	20,63%	46,88%
SSpro	36,09%	22,65%	41,25%
Consenso	34,64%	20,50%	44,86%

*A composição de alfa-hélices, fitas beta e resíduos irregulares é calculada como a porcentagem de ocorrência do total de resíduos considerando todas as seqüências.

Verificamos também a distribuição dos tamanhos dos segmentos para cada método (Figura 6). Em geral o SSpro foi o método que mais se aproximou dos tamanhos encontrados no PDB. É esperado que segmentos com menos de três resíduos não ocorram em conformações de alfa-hélice (DEMUTH, 2005), observação confirmada na distribuição de seqüências de estruturas do PDB. No entanto, nas predições, segmentos curtos e resíduos isolados de alfa-hélice foram observados. O PSIPRED foi o único método a não apresentar resíduos isolados tanto de alfa-hélice como de fita beta. Contudo, notamos que o PDB possui alta frequência de ocorrência de resíduos isolados de fita beta, algo que somente o SSpro foi capaz de reproduzir. Os demais preditores apresentaram segmentos maiores, principalmente com o DeepCNF. Para segmentos com estrutura irregular, notamos que o PDB não apresenta frequências consideráveis de segmentos com mais de quinze resíduos. Predições do PSIPRED, DeepCNF e consenso apresentaram certa ocorrência de segmentos até cerca de 25 resíduos.

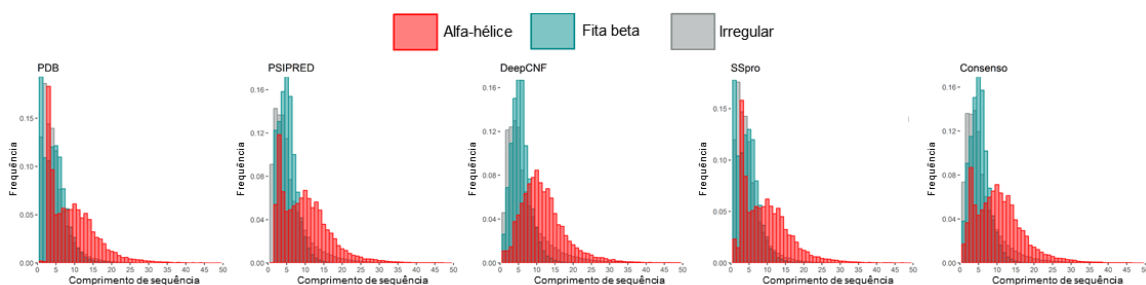


Figura 6 - Comprimentos de segmentos de estrutura secundária em estruturas do PDB e métodos de predição. Distribuição do comprimento de segmentos de alfa-hélice, fita beta e irregular por método. No gráfico, PDB refere-se às estruturas definidas experimentalmente extraídas via DSSP, já as predições são referentes às seqüências primárias do PDB.

Com o objetivo de verificar como se relacionam as medidas de composição estrutural entre si e entre cada método, calculamos a correlação de Pearson entre as distribuições de alfa-hélices, fitas beta e irregulares (Figura 7). As correlações entre as estruturas mostraram que distribuições de alfa-hélice possuem boa correlação com fita beta

(em torno de 0.85) e irregular (entre 0.63 e 0.79). Já fita beta não apresenta correlação significativa com a distribuição de estrutura irregular (menos de 0.30). Isso mostra que, apesar de serem medidas dependentes umas das outras, variações na composição de fitas beta e estruturas irregulares são geralmente explicadas por variações em alfa-hélices. Analisando as correlações entre métodos, observamos mais uma vez a vantagem do SSpro em relação às demais previsões, quando comparado com o PDB. Notamos também que as previsões com PSIPRED e DeepCNF são mais semelhantes entre si e que o consenso apresenta maior correlação com os mesmos. Além disso, vimos que distribuições de alfa-hélices possuem maior correlação entre o PDB e as previsões, com valores variando entre 0.96 (DeepCNF vs PDB) e 0.99 (SSpro vs PDB). Para fitas beta as correlações variaram entre 0.94 e 0.98 e para irregulares variaram entre 0.81 e 0.94.

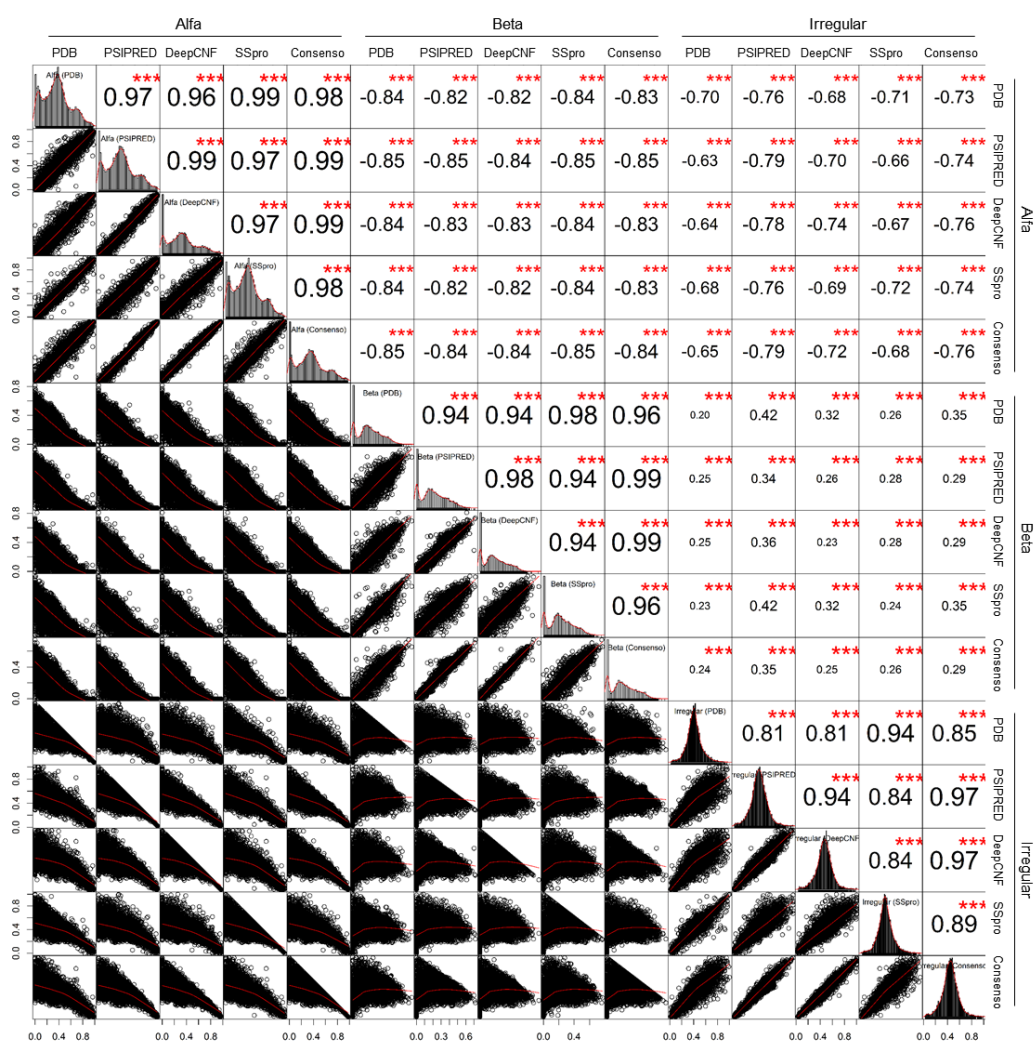


Figura 7 - Correlação entre distribuições de estrutura secundária de estruturas do PDB e métodos de predição. A figura mostra a comparação par a par entre distribuições de alfa-hélice, fita beta e irregular para estruturas do PDB (DSSP) e previsões usando SSpro, PSIPRED, DeepCNF e Consenso. Gráficos na matriz triangular inferior mostram a dispersão entre as distribuições em cada eixo. Gráficos na diagonal mostram histogramas com a distribuição de cada conformação por método. Gráficos na matriz triangular superior contêm os valores de correlação de Pearson e os asteriscos indicam a significância (* 0.05, ** 0.01, *** 0.001).

A fim de observar a extensão das diferenças entre os métodos de predição em situações além das observadas experimentalmente, comparamos as predições das sequências do proteoma humano completo. Para essas análises não utilizamos informações do PDB, somente as predições. A Figura 8b mostra a composição dos três estados estruturais para cada método. Notamos poucas diferenças entre os métodos na composição de alfa-hélice com o PSIPRED apresentando valores levemente superiores aos demais. Da mesma forma, as diferenças foram suaves para estruturas irregulares, com o DeepCNF apresentando valores levemente superiores. A maior discrepância foi observada na composição de fita beta, com o SSpro apresentando valores superiores aos outros métodos. Essa diferença é explicada pela ocorrência elevada de sequências com pouca ou nenhuma estrutura de fita beta nos métodos PSIPRED e DeepCNF (Figuras 8a e 8b). Notamos também que, da mesma forma que ocorreu com as predições do PDB (Figura 7), há maior correlação do consenso com o PSIPRED e o DeepCNF do que com o SSpro (Figura 9). No entanto, as distribuições de alfa e beta para os dados do proteoma humano se mostram menos correlacionadas do que quando comparadas com dados do PDB.

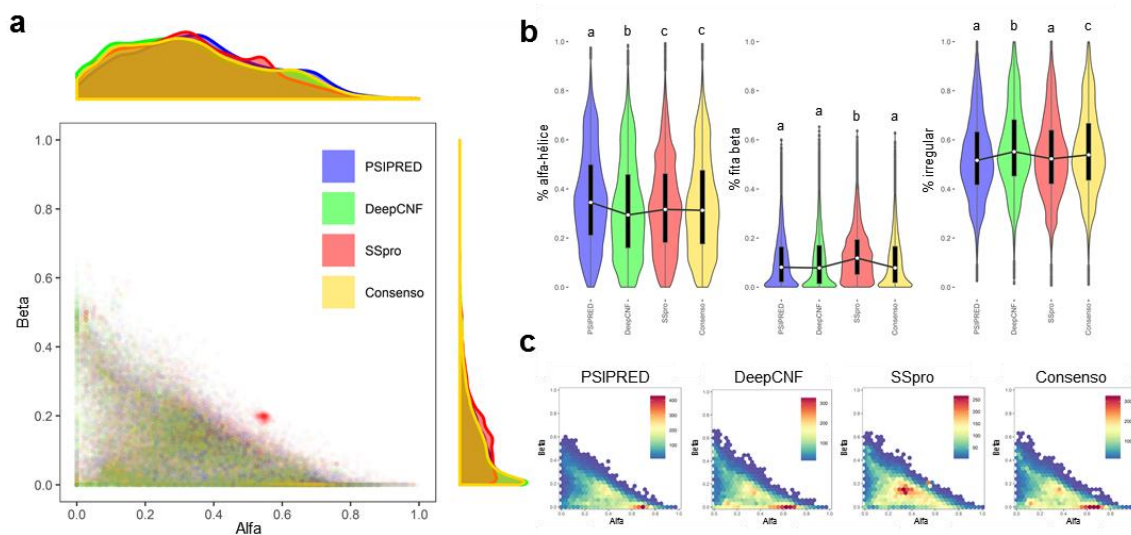


Figura 8 - Distribuição estrutural de sequências de *Homo sapiens* por método de predição. (a), gráfico de dispersão com a composição de alfa-hélices e fitas beta para cada método. Curvas de densidade para cada estrutura são exibidas nas laterais. (b), mostra composição de estrutura secundária (% de alfa-hélice, fita beta e irregular de cada sequência). Letras diferentes indicam grupos com diferenças estatisticamente significantes usando o teste de Kruskal-Wallis e contraste de Dunnett (valor $p < 0,001$). (c), gráficos de calor com a quantidade de sequências pela composição alfa-beta. Cores variando de azul para vermelho indicam de poucas para muitas sequências respectivamente. Cada gráfico apresenta escala de valores própria no canto superior direito.

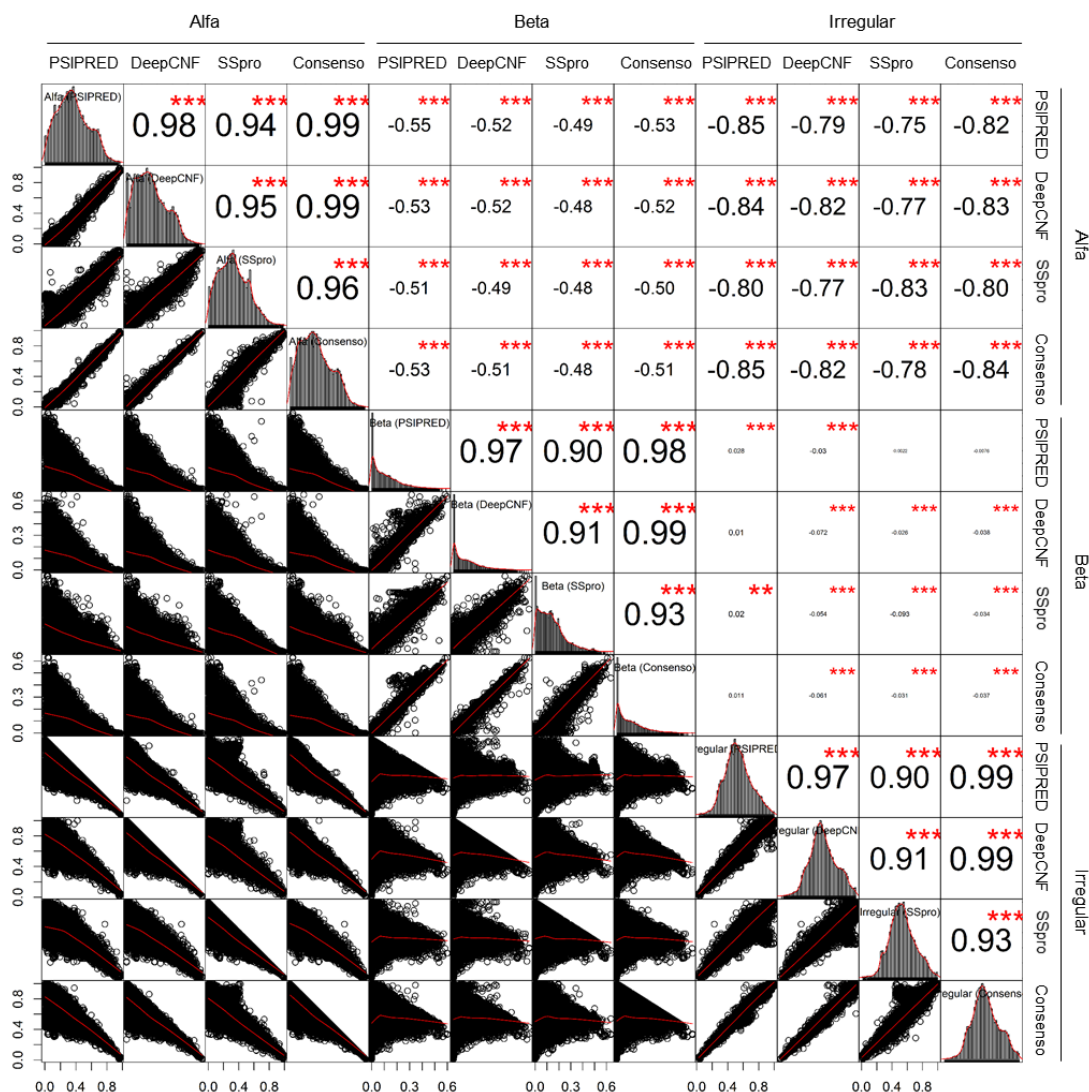


Figura 9 - Correlação entre distribuições de estrutura secundária de predições do proteoma de *Homo sapiens*. A figura mostra a comparação par a par entre distribuições de alfa-hélice, fita beta e irregular para dados de predições usando SSpro, PSIPRED, DeepCNF e Consenso. Gráficos na matriz triangular inferior mostram a dispersão entre as distribuições em cada eixo. Gráficos na diagonal mostram histogramas com a distribuição de cada conformação por método. Gráficos na matriz triangular superior contém os valores de correlação de Pearson e os asteriscos indicam a significância (* 0.05, ** 0.01, *** 0.001).

Outra situação que analisamos foi o comportamento das predições em sequências não verdadeiras. Inferimos a predição da estrutura secundária para sequências do proteoma humano com resíduos permutados aleatoriamente. Consideramos que as sequências derivadas deste processo não seriam funcionais, e dessa forma consideramos esses resultados como controle negativo. O viés que estaria operando seria o uso de aminoácidos, sem o peso da sua ordenação. Os métodos apresentaram resultados distintos, com composição estrutural compatível com observadas em proteínas reais (Figura 10). O método que apresentou maior discrepância com os demais foi o DeepCNF, com composição de alfa-hélices inferior aos demais. As distribuições de tamanhos de segmentos

também se mostraram semelhantes às encontradas em proteínas reais (Figura 10c). As correlações entre os métodos, apesar de inferiores aos observados nos dados do proteoma, foram relativamente altas, com valores variando entre 0.78 (distribuição de beta entre DeepCNF e SSpro) e 0.95 (distribuição de irregular entre PSIPRED e SSpro) (Figura 11). O consenso apresentou maior correlação com o PSIPRED. Comparando as previsões entre as sequências verdadeiras e falsas (proteoma e proteoma permutado, respectivamente), notamos uma boa correlação entre ambas em alfa-hélices e irregulares (Figura 12).

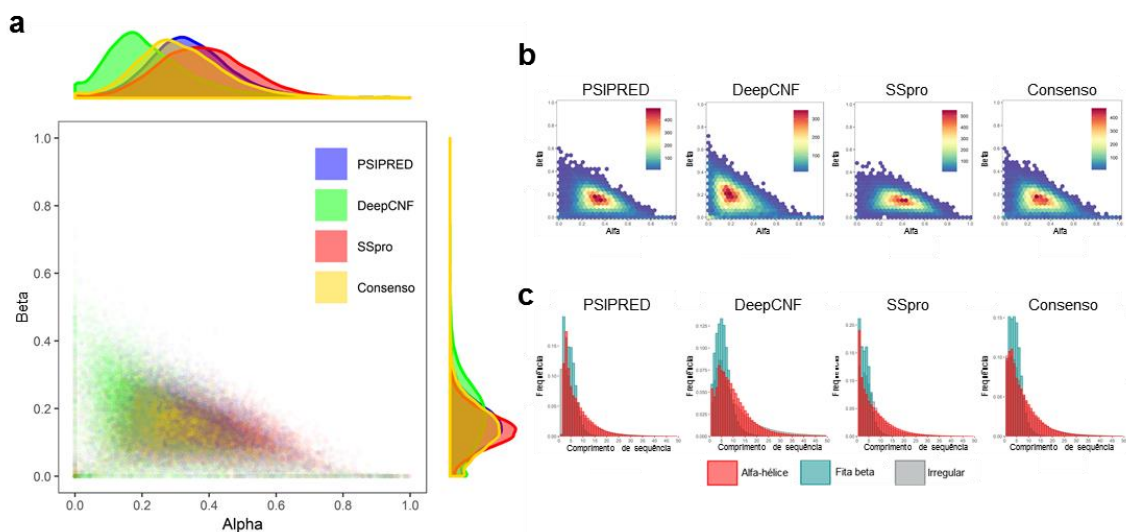


Figura 10 - Distribuição estrutural de sequências permutadas de *Homo sapiens* por método de predição. (a), gráfico de dispersão com a composição de alfa-hélices e fitas beta por preditor. Curvas de densidade para cada estrutura são exibidas nas laterais. (b), gráficos de calor com a quantidade de sequências pela composição alfa-beta. Cores variando de azul para vermelho indicam de poucas para muitas sequências respectivamente. Cada gráfico apresenta escala de valores própria no canto superior direito. (c), distribuição do comprimento de segmentos de alfa-hélice, fita beta e irregular por método.

Assim, apesar de encontrarmos diferenças significativas entre os métodos de predição, a abordagem de consenso aqui empregada fornece uma boa estimativa para as análises propostas neste trabalho. Mesmo nos casos em que observamos maiores diferenças, como no caso de proteínas permutadas, as situações de discordância (1:1:1) entre os métodos são desprezíveis e não afetam as conclusões do trabalho. Dessa forma, para as análises seguintes que utilizem dados de predição, utilizaremos o consenso das três ferramentas (exceto em casos indicados explicitamente). A grande concordância entre SSpro e a extração de dados do PDB poderia ser motivo de escolha deste método, mas poderia ser um risco utilizar só este método para proteomas completos e sequências não verdadeiras, como foi feito.

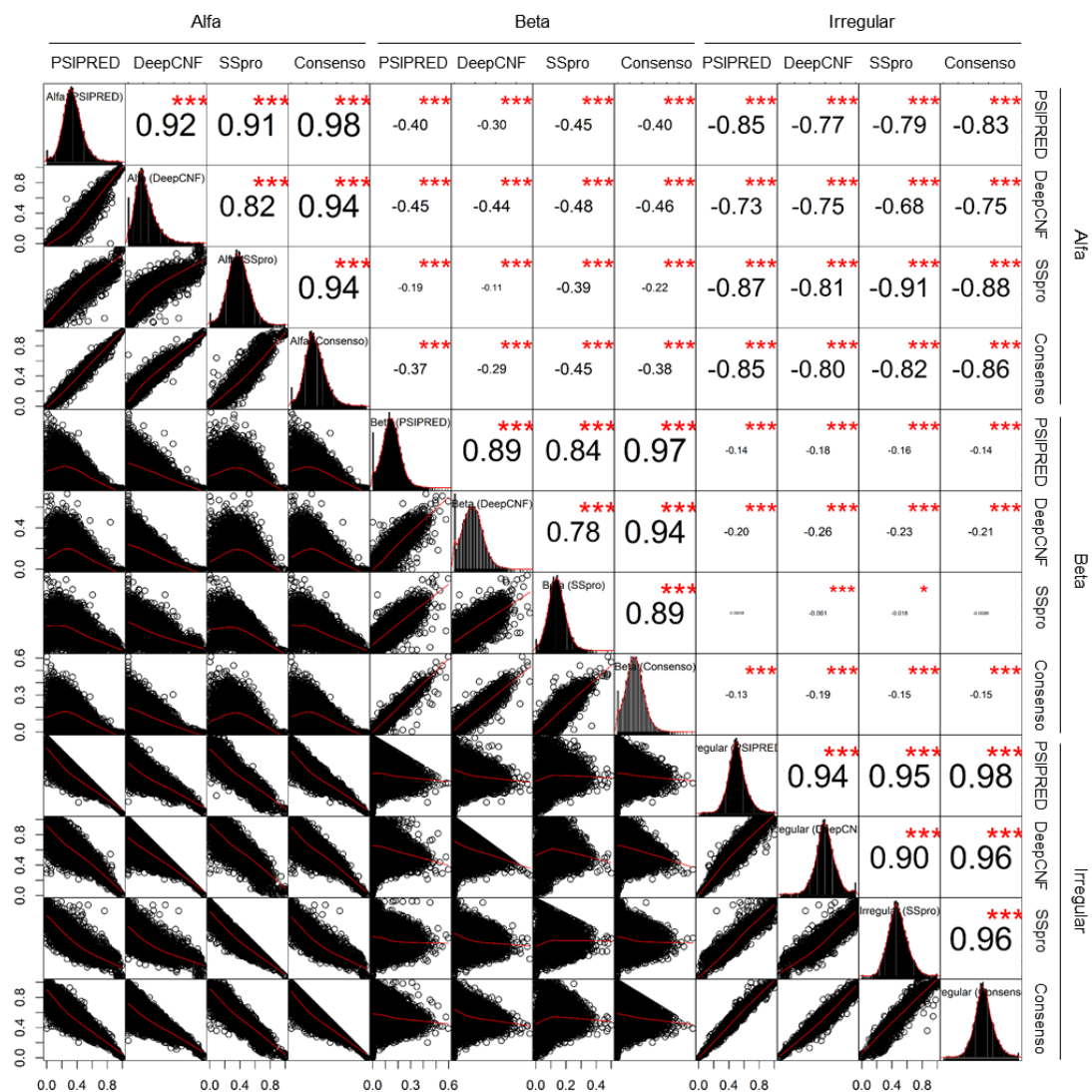


Figura 11 - Correlação entre distribuições de estrutura secundária de predições de seqüências permutadas do proteoma de *Homo sapiens*. A figura mostra a comparação par a par entre distribuições de alfa-hélice, fita beta e irregular para dados de predições usando SSpro, PSIPRED, DeepCNF e Consenso. Gráficos na matriz triangular inferior mostram a dispersão entre as distribuições em cada eixo. Gráficos na diagonal mostram histogramas com a distribuição de cada conformação por método. Gráficos na matriz triangular superior contêm os valores de correlação de Pearson e os asteriscos indicam a significância (* 0.05, ** 0.01, *** 0.001).

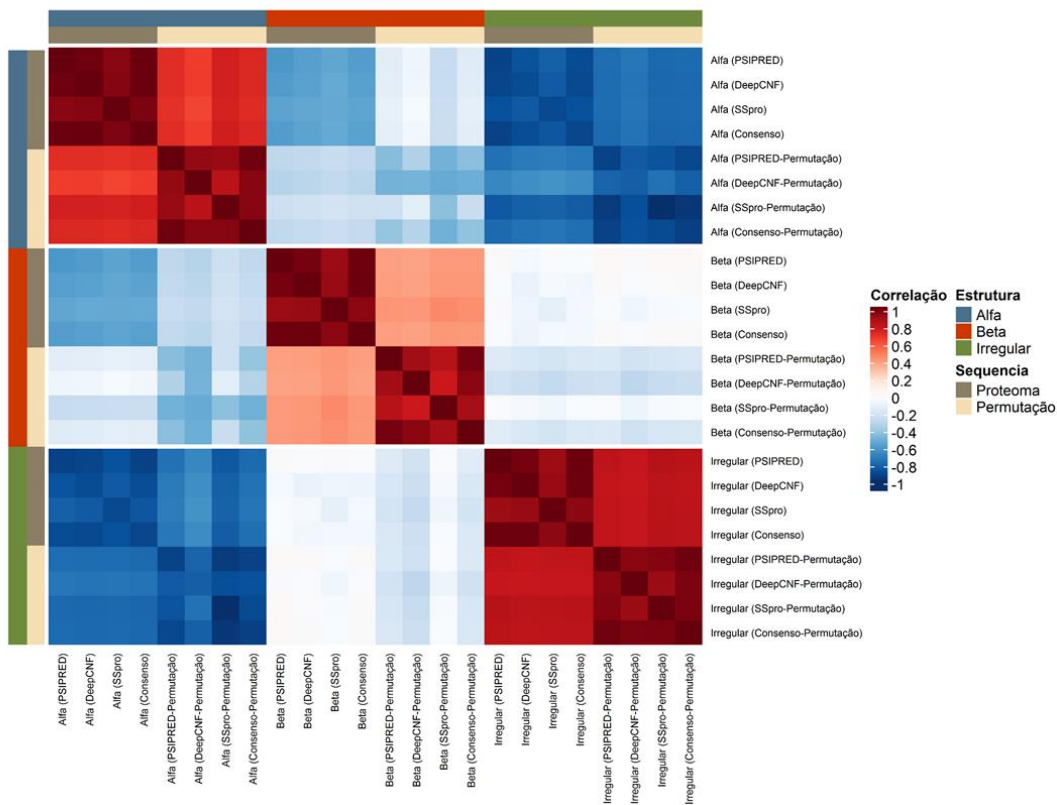


Figura 12 - Correlação entre distribuições de estrutura secundária entre predições de seqüências permutadas e não permutadas de *Homo sapiens*. O mapa de calor mostra a média da correlação de Pearson para cada estado de estrutura secundária para cada método de predição usando seqüências do proteoma humano permutadas e não permutadas. Cores vermelhas indicam alta correlação positiva, cores azuis indicam alta correlação negativa e cores brancas indicam baixa correlação.

4.2. Análise evolutiva com dados do PDB

Neste trabalho, estamos interessados em estudar mudanças estruturais em proteínas ao longo da evolução, mais especificamente com relação à estrutura secundária. Para tal, decidimos inicialmente utilizar informações de estruturas definidas experimentalmente disponibilizadas na base de dados do PDB. A extração da informação referente à estrutura secundária é realizada utilizando-se o algoritmo DSSP. Utilizamos a base de dados PDBFINDER2 (TOUW et al., 2015) como fonte destas informações. Esta base corrige alguns problemas de inconsistência encontrados em arquivos do PDB além de ser de fácil manuseio. Tendo em vista que o PDB também apresenta muita redundância, decidimos utilizar somente proteínas não redundantes. A determinação deste conjunto de dados foi obtida na base de dados de baixa redundância (valor- p 10^{-7}) do VAST (*Vector Alignment Search Tool*) (MADEJ et al., 2014). Mais detalhes sobre a obtenção dos dados estão disponíveis na seção de Materiais e métodos.

Inicialmente separamos as proteínas de acordo com o domínio taxonômico dos organismos aos quais pertencem. Assim, classificamos as proteínas em quatro grupos:

Archaea, *Bacteria*, *Eukaryota* e *Viruses*. Analisamos a composição estrutural (% de alfa-hélice, % de fita beta e % de irregular) dessas proteínas. Notamos que existe uma diferença significativa entre proteínas de procariotos (*Archaea* e *Bacteria*) e eucariotos (Figura 13). As proteínas presentes em eucariotos apresentam menor conteúdo de alfa-hélices e fitas beta (e por consequência maior conteúdo irregular) em comparação com o encontrado em procariotos. Do total de resíduos em todas as sequências, proteínas de eucariotos apresentaram cerca de 4,61% mais resíduos em estado irregular do que arqueias e 2,53% a mais do que bactérias (Tabela 8). Além disso, observamos maior diversidade de composições estruturais em proteínas de eucariotos e vírus (Figura 14). As proteínas de *Archaea* e *Bacteria*, em sua maioria, apresentam conteúdo de alfa-hélice e fita beta semelhante, em torno de 40% de alfa-hélices e 20% de fitas betas (Figura 14b). Investigamos também possíveis vieses com relação ao tamanho das sequências. Notamos que proteínas de eucariotos aparentemente são levemente menores do que as de procariotos (Figura 15a e Tabela 8), todavia há um grande viés nessa medida explicado pela baixa cobertura com o UniProt em boa parte das sequências de eucariotos resolvidas (Figura 16). Isso traz uma preocupação sobre as determinantes de conteúdo estrutural, já que aparentemente a porção resolvida de proteínas de eucariotos nem sempre compreende a proteína toda, mas provavelmente se concentra em um núcleo globular. Notamos também que proteínas menores tendem a apresentar menor conteúdo estrutural (em azul na Figura 15b). No entanto, existem casos que fogem à regra. Quanto aos tamanhos dos segmentos de cada classe estrutural, não notamos diferenças com relação ao domínio taxonômico ao qual pertence a proteína (Figura 17).

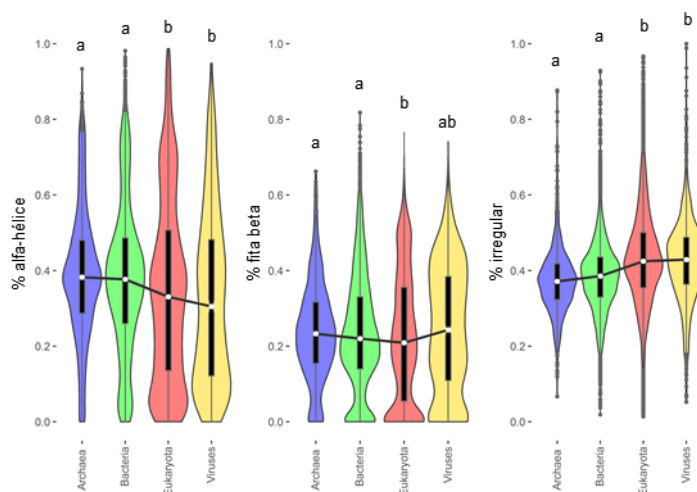


Figura 13 - Composição de estrutura secundária em estruturas do PDB por domínio taxonômico. Distribuição da porcentagem de alfa-hélice, fita beta e irregular de estruturas do PDB extraídas via DSSP. Letras diferentes indicam grupos com diferenças estatisticamente significantes usando o teste de Kruskal-Wallis e contraste de Dunnett (valor $p < 0,001$).

Tabela 8 - Sequências de estruturas do PDB por domínio taxonômico.

Domínio	N	Comprimento Médio (EP)	% de alfa-hélice*	% de fita beta*	% irregular*
<i>Archaea</i>	821	217,09 (4,589)	40,17%	22,48%	37,35%
<i>Bacteria</i>	7029	242,30 (1,926)	37,72%	22,86%	39,43%
<i>Eukaryota</i>	6349	209,00 (2,161)	35,99%	22,05%	41,96%
<i>Viruses</i>	902	235,40 (5,975)	29,70%	26,02%	44,27%

*A composição de alfa-hélices, fitas beta e de resíduos irregulares é calculada como a porcentagem de ocorrência do total de resíduos considerando todas as sequências por domínio taxonômico. N indica a quantidade de sequências. EP representa o erro padrão.

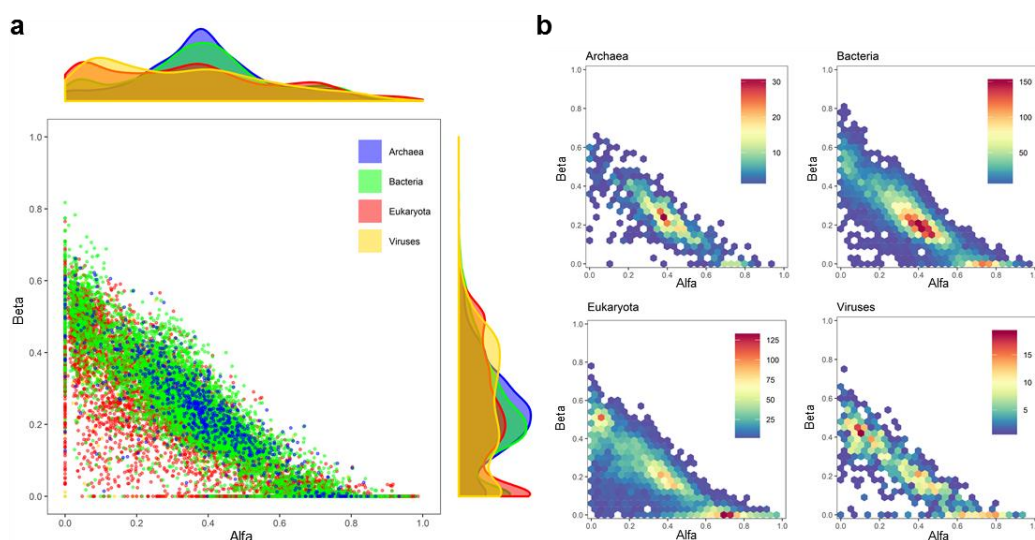


Figura 14 - Distribuição alfa-beta em estruturas do PDB por domínio taxonômico. (a), gráfico de dispersão com composição de alfa-hélices e fitas beta por domínio de ocorrência. Curvas de densidade para cada estrutura são exibidas nas laterais. (b), gráficos de calor com a quantidade de sequências pela composição de alfa hélices e fitas beta. Cores variando de azul para vermelho indicam de poucas para muitas sequências respectivamente. Cada gráfico apresenta escala de valores própria no canto superior direito.

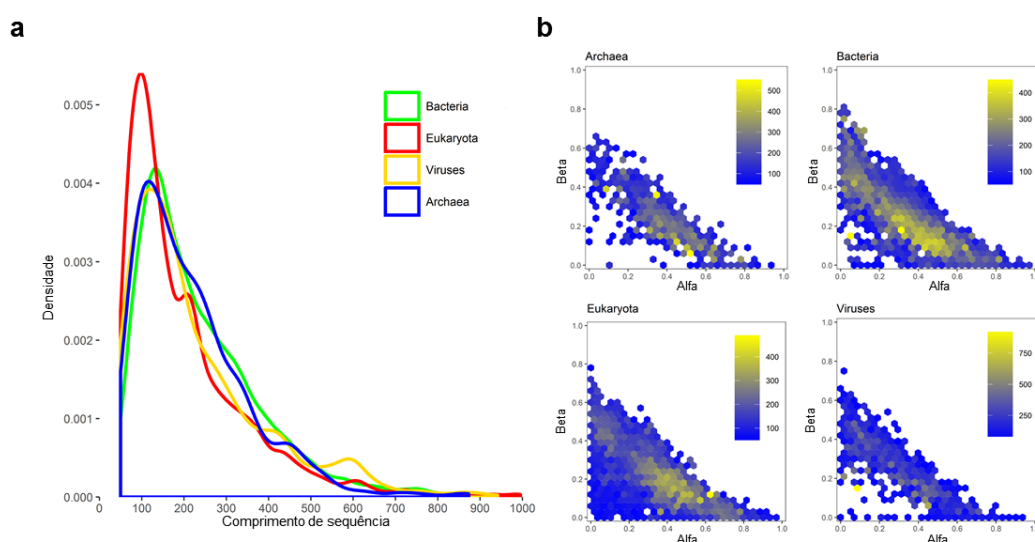


Figura 15 - Comprimento de estruturas do PDB por domínio taxonômico. (a), mostra a distribuição de comprimentos de sequências por domínio. (b), mostra mapas de calor com o comprimento médio de sequências para cada classe de composição alfa-beta. Cores variando de azul para amarelo indicam sequências curtas para longas respectivamente. Cada gráfico apresenta escala de valores própria no canto superior direito.

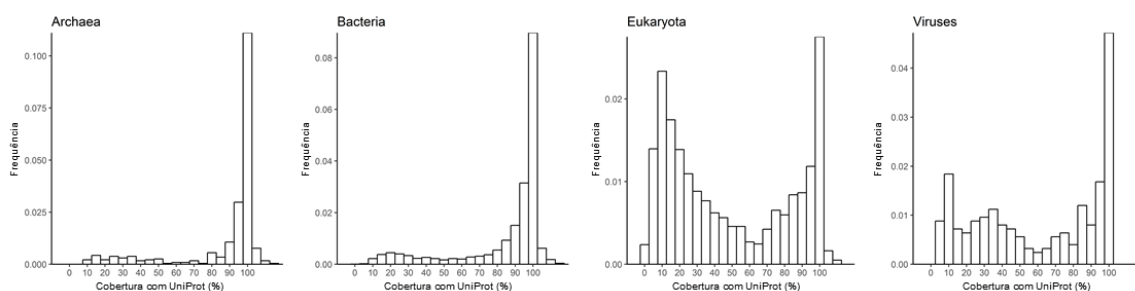


Figura 16 - Cobertura de estruturas do PDB em relação a seqüências do UniProt por domínio taxonômico. A cobertura é calculada dividindo o comprimento da seqüência do PDB pelo comprimento da seqüência do UniProt e multiplicado por 100. O mapeamento dos identificadores equivalentes nas bases de dados foi obtido pelos SIFTS (ver Materiais e Métodos).

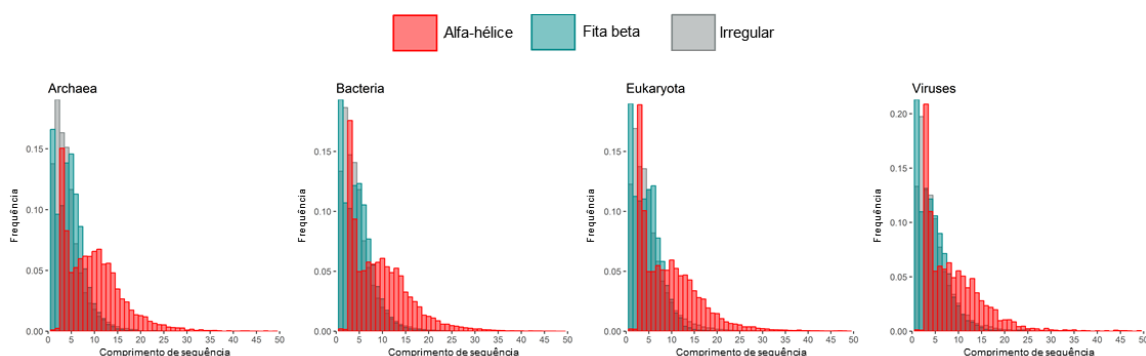


Figura 17 - Comprimentos de segmentos de estrutura secundária em estruturas do PDB por domínio taxonômico. Distribuição do comprimento de segmentos de alfa-hélice, fita beta e irregular por domínio de ocorrência.

Analisamos também informações de domínios estruturais por domínios taxonômicos. A base de dados CATH (SILLITOE et al., 2015) fornece uma classificação hierárquica de domínios em quatro níveis (Classe, Arquitetura, Topologia e Homologia). A Figura 18 mostra a classificação de classes e arquiteturas para proteínas dos quatro domínios taxonômicos (*Archaea*, *Bacteria*, *Eukaryota* e *Viruses*). A classificação apresenta resultados compatíveis com os resultados observados com estrutura secundária. Domínios de proteínas de *Archaea* e *Bacteria* apresentam maior ocorrência de classes com composição mista de alfa-hélice e fita beta (em azul). Enquanto os domínios de *Eukaryota* e *Viruses* possuem maior ocorrência de classes com composição majoritária de alfa-hélice ou fita beta. Além de apresentarem maior proporção de domínios com pouca estrutura secundária (em laranja). Essas informações suportam a maior diversidade estrutural encontrada em eucariotos e vírus.

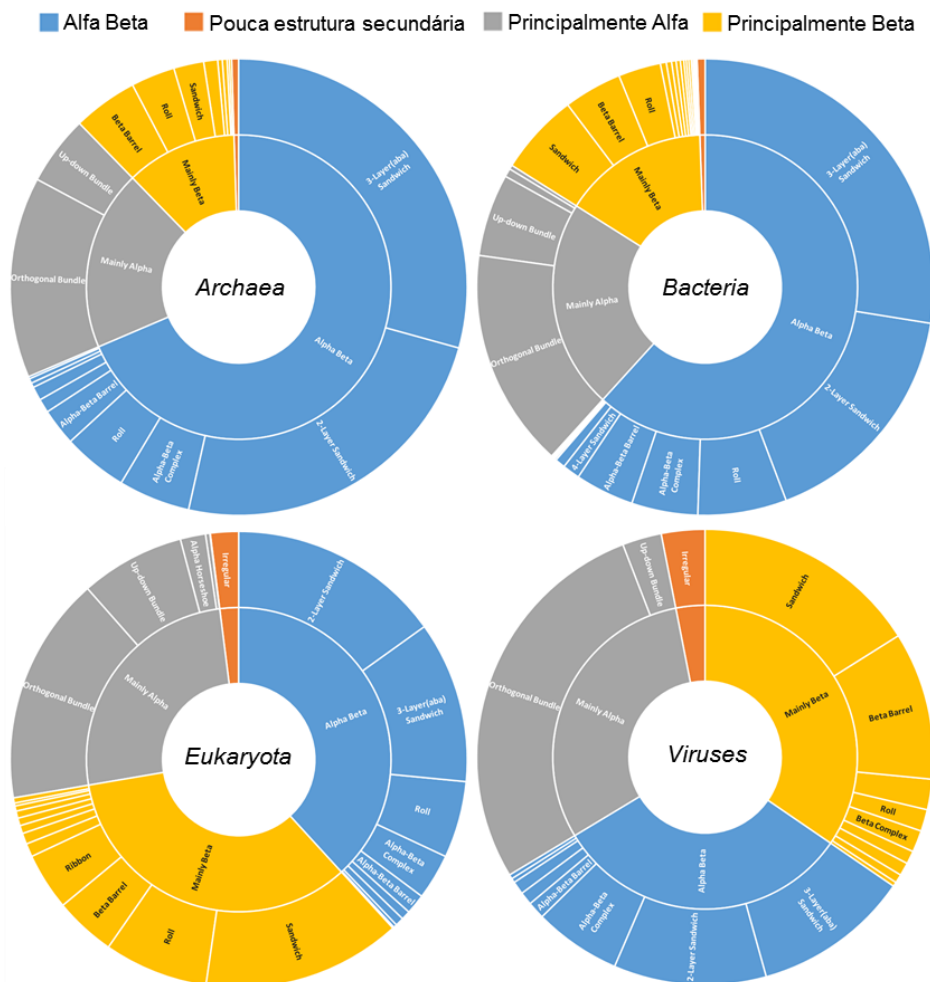


Figura 18 - Classificação de domínios CATH para seqüências do PDB por domínio taxonômico. Classificação de domínios estruturais em termos de classes (anel interno) e arquiteturas (anel externo) CATH para seqüências do PDB separadas por domínio taxonômico.

Com base nas diferenças observadas em proteínas de organismos em diferentes domínios taxonômicos, decidimos aprofundar a análise evolutiva. Para melhor classificar o período de origem das proteínas, selecionamos somente proteínas humanas presentes no PDB e classificamos sua origem de acordo com quatro clados da linhagem evolutiva: *Euk_Bac_Arch* (agrupando proteínas com origem no ancestral comum de eucariotos, bactérias e arqueias, i.e., organismos celulares), *Eukaryota*, *Opistho_Eumeta* (agrupando proteínas com origem nos clados *Opisthokonta* e *Eumetazoa*) e *Verteb_Mammalia* (agrupando proteínas com origem nos clados *Vertebrata* e *Mammalia*). Essa classificação foi determinada através de uma abordagem cladística com base em 13 métodos de inferência de ortologia (ver Materiais e Métodos para detalhes). A Figura 19a mostra o número de proteínas com idade classificada em cada clado. Analisando a composição estrutural entre as diferentes idades, identificamos um forte viés para redução da composição de alfa-hélices ao longo da evolução (Figura 19b). Proteínas com origem em

Euk_Bac_Arch apresentam 38,22% do total de resíduos em alfa-hélice enquanto proteínas com origem em *Verteb_Mammalia* apresentam somente 21,44% (Tabela 9). Com relação à composição de fitas beta, notamos que ocorre o oposto. Proteínas mais recentes apresentam maior proporção de fita beta (31,35% em *Verteb_Mammalia* contra 19,02% em *Euk_Bac_Arch*). Essa diferença ameniza o impacto na composição geral de estruturas irregulares, mas ainda assim notamos diferenças significativas entre proteínas recentes e antigas. Proteínas mais recentes apresentam maior conteúdo irregular do que comparado com as mais antigas. Notamos também maior diversidade estrutural em proteínas mais recentes, com conteúdo majoritário de alfa-hélice ou fita beta, de forma excludente (Figura 20).

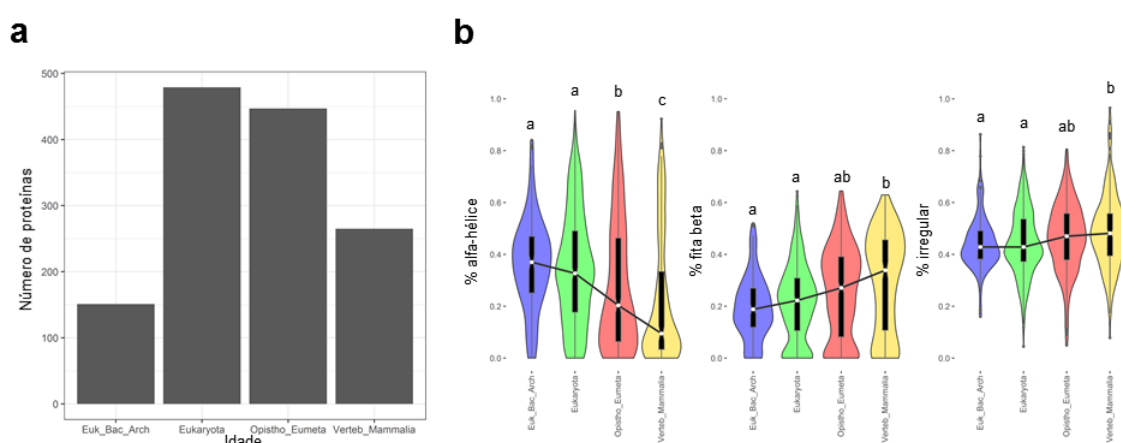


Figura 19 - Composição de estrutura secundária em estruturas do PDB de *Homo sapiens* por idade evolutiva. (a), mostra quantidades de proteínas por idade inferida (ver Materiais e Métodos para mais detalhes). (b), Composição de estrutura secundária (% de alfa-hélice, fita beta e irregular de cada sequência) por idade evolutiva. Letras diferentes indicam idades com diferenças estatisticamente significativas usando o teste de Kruskal-Wallis e contraste de Dunnett (valor $p < 0,001$).

Com relação aos tamanhos das sequências, vimos maiores diferenças entre proteínas com origem em *Euk_Bac_Arch* em relação às demais (Tabela 9 e Figura 21a). Essa diferença pode também ser explicada devido ao elevado número de sequências parciais cristalizadas, indicadas pela baixa cobertura com o UniProt (Figura 22). Com relação aos domínios, observamos a mesma tendência observada nas sequências de todo o PDB. Sequências antigas (origem em *Euk_Bac_Arch*) possuem em sua maioria domínios com composição mista de alfa-hélices e fitas beta (Figura 23). Sequências com origem em *Eukaryota* possuem uma distribuição semelhante para domínios com composição mista e composição majoritariamente alfa ou beta. Já sequências mais recentes começam a apresentar maior quantidade de domínios com composição de principalmente fita beta (em amarelo). Esses resultados refletem a composição estrutural observada por idade evolutiva e, pela primeira vez, descrevem nosso proteoma como uma mistura.

Tabela 9 - Estruturas do PDB de *Homo sapiens* por idade evolutiva.

Idade	N	Comprimento Médio (EP)	% de alfa-hélice*	% de fita beta*	% irregular*
<i>Euk_Bac_Arch</i>	151	250,91 (13,338)	38,22%	19,02%	42,76%
<i>Eukaryota</i>	479	181,01 (6,232)	38,23%	19,48%	42,29%
<i>Opistho_Eumeta</i>	447	149,26 (5,391)	29,03%	26,04%	44,93%
<i>Verteb_Mammalia</i>	265	149,52 (6,297)	21,44%	31,35%	47,21%

*A composição de alfa-hélices, fitas beta e de resíduos irregulares é calculada como a porcentagem de ocorrência do total de resíduos considerando todas as sequências por idade evolutiva. N indica a quantidade de sequências. EP representa o erro padrão.

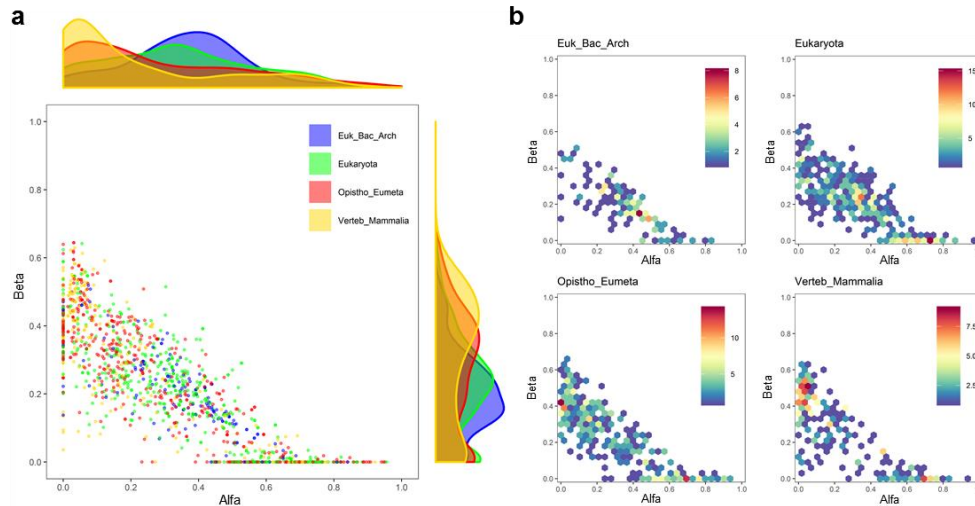


Figura 20 - Distribuição de alfa hélices e fitas beta em estruturas de *Homo sapiens* presentes no PDB por idade evolutiva. (a), gráfico de dispersão com a composição de alfa-hélices e fitas beta por idade evolutiva. Curvas de densidade para cada estrutura são exibidas nas laterais. (b), gráficos de calor com a quantidade de sequências pela composição alfa-beta. Cores variando de azul para vermelho indicam de poucas para muitas sequências respectivamente. Cada gráfico apresenta escala de valores própria no canto superior direito.

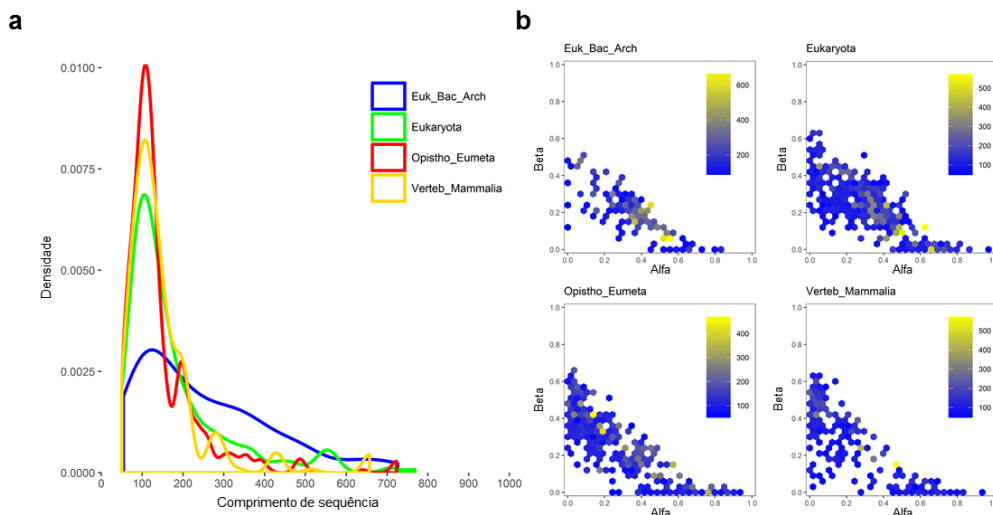


Figura 21 - Comprimento de estruturas PDB de *Homo sapiens* por idade evolutiva. (a), mostra a distribuição de densidades para comprimentos de sequências por idade evolutiva. (b), mostra mapas de calor com o comprimento médio de sequências pela composição alfa-beta. Cores variando de azul para amarelo indicam sequências curtas para longas respectivamente. Cada gráfico apresenta escala de valores própria no canto superior direito.

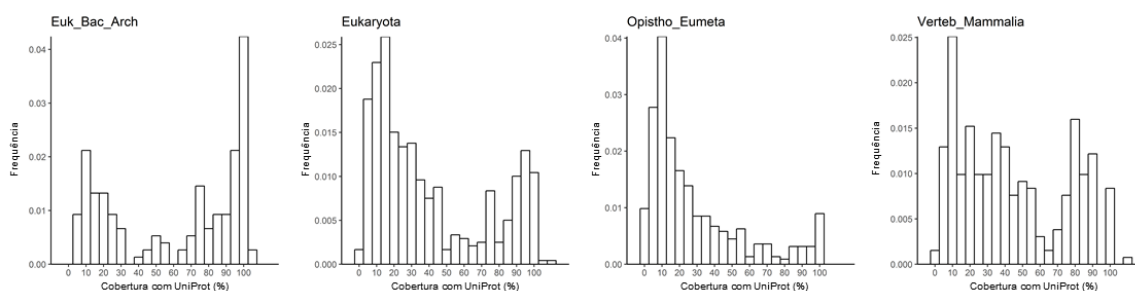


Figura 22 - Cobertura de estruturas do PDB de *Homo sapiens* em relação às sequências do UniProt por idade evolutiva. A cobertura é calculada dividindo o comprimento da sequência do PDB pelo comprimento da sequência do UniProt e multiplicando por 100. O mapeamento dos identificadores equivalentes nas bases de dados foi obtido pelos SIFTS (ver Materiais e Métodos).

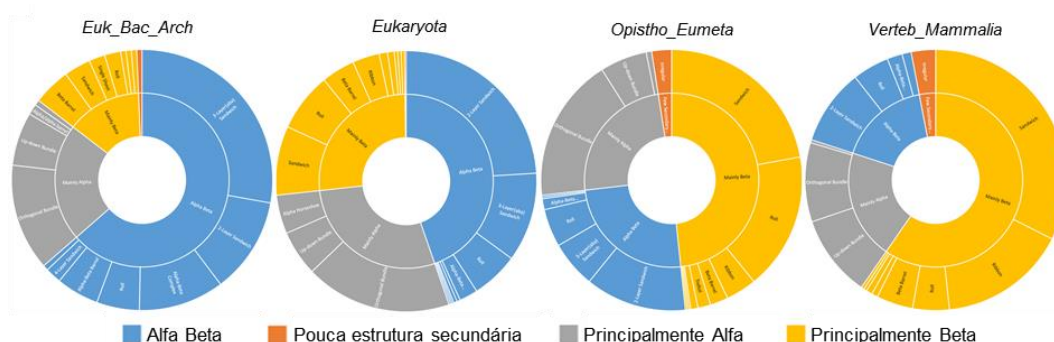


Figura 23 - Classificação de domínios CATH para sequências de *Homo sapiens* presentes no PDB por idade evolutiva. Classificação de domínios estruturais em termos de classes (anel interno) e arquiteturas (anel externo) CATH para sequências do PDB separadas por idade evolutiva.

Ainda com respeito às sequências do PDB, comparamos o uso de aminoácidos entre proteínas com origem antiga (*Euk_Bac_Arch*) e recente (*Verteb_Mammalia*) (Figura 24). A Tabela 10 mostra os aminoácidos ordenados pela diferença percentual entre *Euk_Bac_Arch* e *Verteb_Mammalia*. Vemos que alguns aminoácidos apresentam diferenças estatisticamente significativas (valor $p < 0,001$ com o teste Mann-Whitney-Wilcoxon) entre os dois grupos, principalmente no uso da cisteína. Sequências com origem em *Verteb_Mammalia* utilizam mais que o dobro da cisteína utilizada em *Euk_Bac_Arch*. Por outro lado, alanina, isoleucina, valina, aspartato e metionina tem uso mais de 14% superior em *Euk_Bac_Arch*.

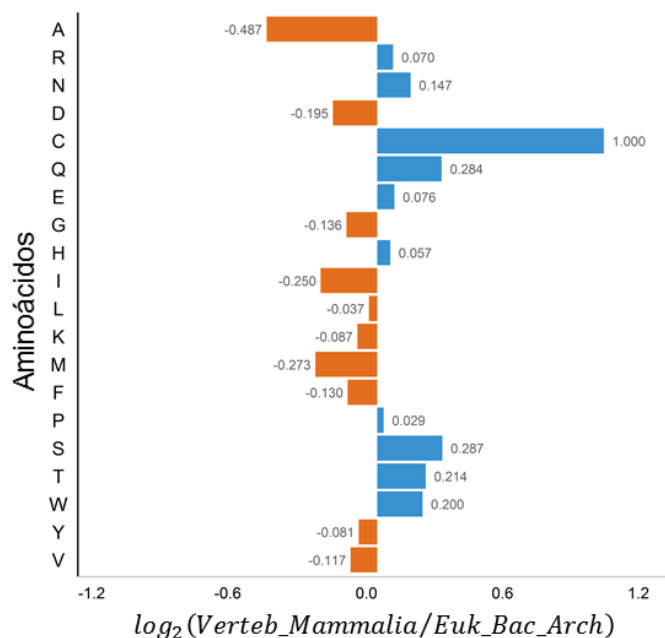


Figura 24 - Uso de aminoácidos em seqüências de *Homo sapiens* presentes no PDB com idade em *Euk_Bac_Arch* versus *Verteb_Mammalia*. Barras mostram a razão da porcentagem de uso aminoácidos de proteínas datadas em *Euk_Bac_Arch* por proteínas datadas em *Verteb_Mammalia*. A razão está em escala logarítmica (log₂). Valores negativos (em laranja) indicam maior ocorrência em *Euk_Bac_Arch*, valores positivos (em azul) indicam maior ocorrência em *Verteb_Mammalia*.

Tabela 10 - Uso de aminoácidos em seqüências de *Homo sapiens* presentes no PDB com idade em *Euk_Bac_Arch* e *Verteb_Mammalia*.

Abreviação	Símbolo	Frequência média (EP)*		Diferença percentual†	Significância	Complexidade‡	Códons
		<i>Euk_Bac_Arch</i>	<i>Verteb_Mammalia</i>				
A	Ala	7,35 (0,205)	5,29 (0,173)	28,08	**	4,76	GCN
I	Ile	5,39 (0,176)	4,47 (0,153)	16,99	**	16,04	ATT/ATC/ATA
V	Val	7,18 (0,183)	6,10 (0,164)	15,06	**	12,28	GTN
M	Met	2,22 (0,124)	1,89 (0,086)	14,73		64,68	ATG
D	Asp	5,46 (0,175)	4,68 (0,121)	14,36	**	32,72	GAT/GAC
K	Lys	6,47 (0,233)	5,94 (0,222)	8,24		30,14	AAA/AAG
G	Gly	7,41 (0,168)	6,90 (0,186)	6,95		1	GGN
F	Phe	3,94 (0,140)	3,72 (0,110)	5,58		44	TTT/TTC
L	Leu	9,31 (0,225)	8,96 (0,244)	3,72		16,04	CTN/TTA/TTG
E	Glu	6,95 (0,178)	7,01 (0,157)	-0,85		36,48	GAA/GAG
H	His	2,54 (0,114)	2,58 (0,112)	-1,72		58,7	CAT/CAC
P	Pro	5,05 (0,133)	5,20 (0,143)	-2,99		31,8	CCN
Y	Tyr	3,06 (0,127)	3,23 (0,118)	-5,49		57	TAT/TAC
R	Arg	5,20 (0,169)	5,52 (0,172)	-6,11		56,34	CGN/AGA/AGG
N	Asn	3,52 (0,133)	4,03 (0,135)	-14,29		33,72	AAT/AAC
Q	Gln	4,18 (0,144)	4,89 (0,140)	-17,08		37,48	CAA/CAG
T	Thr	5,06 (0,136)	5,93 (0,151)	-17,17	**	21,62	ACN
W	Trp	1,30 (0,074)	1,59 (0,079)	-21,54		73	TGG
S	Ser	6,55 (0,217)	8,02 (0,206)	-22,39	**	17,86	TCN/AGT/AGC
C	Cys	1,85 (0,128)	4,06 (0,239)	-119,49	**	57,16	TGT/TGC

*Frequência média (porcentagem e erro padrão, EP) de uso de aminoácidos em *Euk_Bac_Arch* e *Verteb_Mammalia*.

†Os aminoácidos estão listados da maior para menor diferença percentual entre genes de procaríotos contra genes de vertebrados.

‡Escores de complexidade representam o custo bioquímico de produção e estabilidade conformacional em uma proteína (DUFTON, 1997; WILLIFORD; DEMUTH, 2012).

** representa aminoácidos com diferença significativa de acordo com teste de Mann-Whitney-Wilcoxon (valor $p < 0.001$).

4.3. Análise evolutiva com dados de predição

Análises em larga escala utilizando dados do PDB apresentam diversas limitações. Uma delas é a amostragem, muitas proteínas não possuem estruturas determinadas experimentalmente. Dessa forma várias sequências não estão presentes no banco de dados, limitando análises globais de organismos. Outra limitação é a alta quantidade de cristais de somente partes de proteínas, observado pela alta ocorrência de sequências de estruturas do PDB com baixa cobertura com suas respectivas sequências completas no UniProt (Figura 16 e Figura 22). Ao tentar analisar o contexto evolutivo de organismos, como por exemplo, de *Homo sapiens*, tais limitações levantam um alerta para possíveis conclusões erradas com esses dados. Tendo em vista que para uma análise mais fidedigna seja necessário utilizar informações do proteoma completo e de proteínas inteiras, métodos computacionais para inferência de informações, onde não há dados experimentais, podem oferecer uma alternativa viável na solução destes problemas. Dessa forma, devido às limitações do PDB e à viabilidade do uso de preditores de estrutura secundária discutidos neste trabalho, utilizamos dados de predição do proteoma humano completo para as análises seguintes.

Para obter uma visão geral, analisamos a composição estrutural de proteomas completos espalhados ao longo da árvore da vida. Devido ao alto custo computacional envolvido, utilizamos para estas análises somente predições com o software SSpro. Utilizamos dados dos 66 proteomas de referência dados pelo consórcio *Quest for Orthologs* (QfO). Os proteomas de referência do QfO fornecem ampla cobertura da árvore da vida contendo organismos bem anotados e de interesse para a investigação biomédica e filogenética (ALTENHOFF et al., 2016; SONNHAMMER et al., 2014). A Figura 25 mostra os resultados da predição da estrutura secundária para os proteomas de referência. A árvore dos organismos de referência foi obtida do SwissTree (BOECKMANN et al., 2015). As barras laterais mostram as porcentagens de resíduos relativas ao total do proteoma para cada organismo. É possível notar a mudança na composição global à medida com que organismos se tornam mais complexos. Organismos procariotos apresentam, em sua maioria, conteúdo de alfa-hélices em torno de 40%, enquanto que eucariotos complexos apresentam cerca de 35%. Da mesma forma, notamos uma leve redução nos resíduos em fita beta. Realizamos também uma análise de componentes principais (PCA) para melhor visualizar as diferenças entre as composições estruturas nos organismos. A Figura 26 mostra clara distinção entre eucariotos e procariotos, com eucariotos concentrados na direção de maior conteúdo de estrutura secundária irregular. Dentro de *Eukaryota* podemos notar também diferenças entre grupos menores, como por exemplo, *Eumetazoa* e *Dikarya* (Figura 26b).

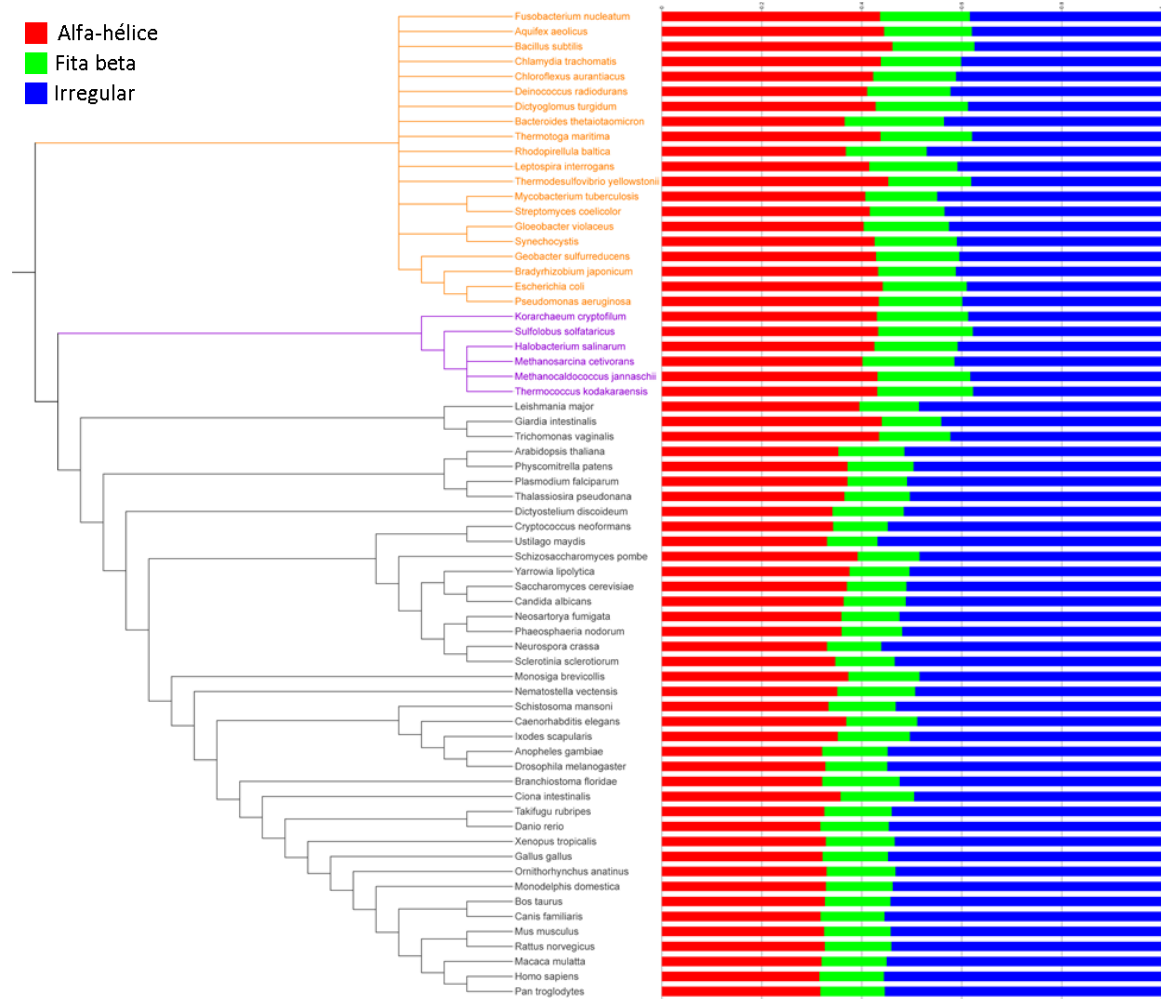


Figura 25 - Composição estrutural dos proteomas de referência. A árvore contém os 66 organismos de referência do SwissTree. Ramos na cor laranja representam bactérias, em roxo arqueias e em cinza eucariotos. As barras laterais representam a porcentagem do total de resíduos em conformação de estrutura secundária: alfa-hélices em vermelho, fitas beta em verde e de irregulares em azul. Dados de estrutura secundária são referentes à predição utilizando SSpro.

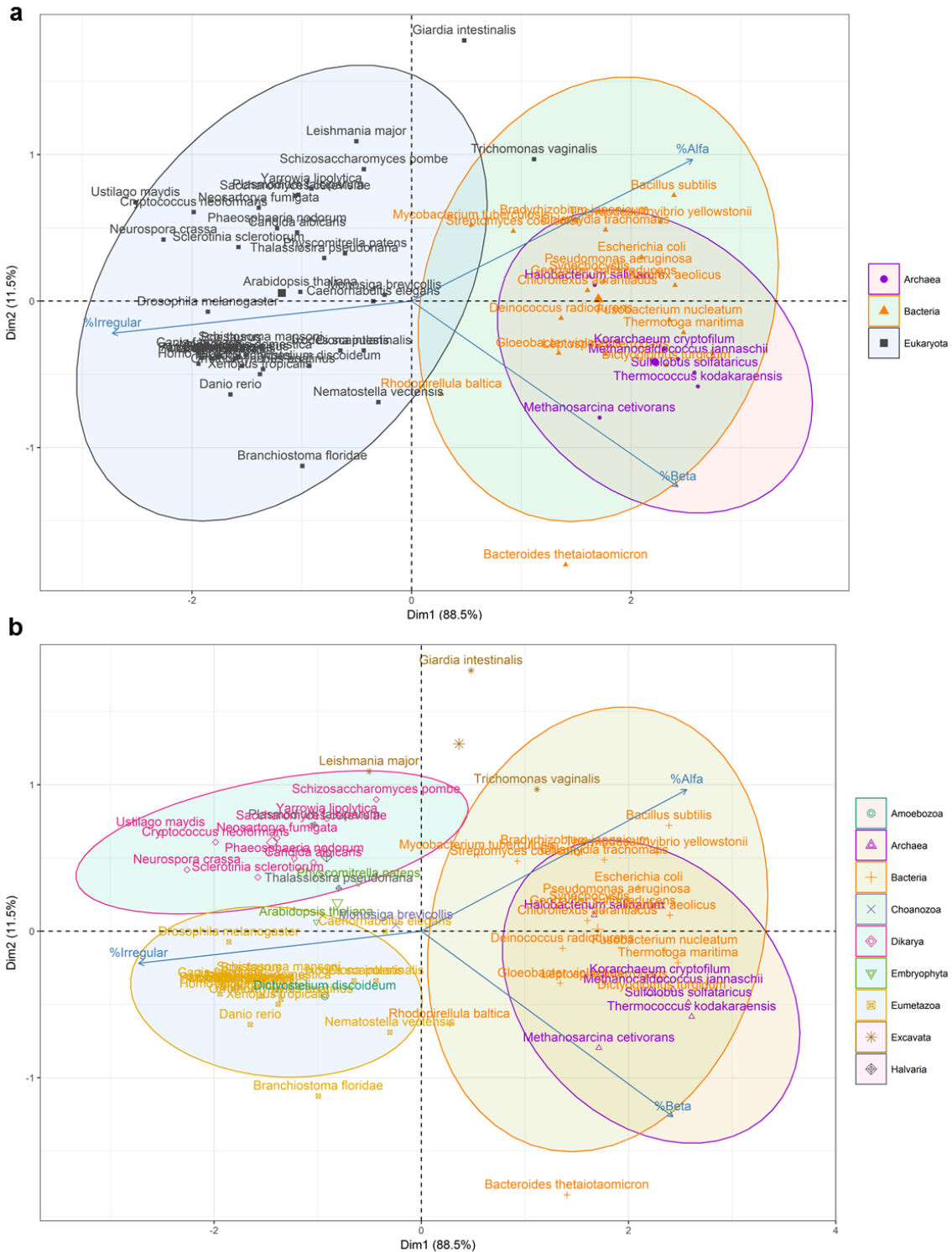


Figura 26 - Análise de componentes principais (PCA) da estrutura secundária dos proteomas de referência. Distribuição dos 66 organismos de referência de acordo com os componentes principais das porcentagens de resíduos nas três conformações de estrutura secundária (alfa-hélice, fita beta e irregular). **(a)**, mostra os organismos coloridos de acordo com o domínio taxonômico (*Archaea*, *Bacteria* e *Eukaryota*). **(b)**, mostra os organismos coloridos em grupos taxonômicos mais específicos.

4.3.1. Análise de *Homo sapiens*

Para melhor estudar as mudanças estruturais ao longo da evolução, analisamos especificamente o proteoma de *Homo sapiens*. Para tal, definimos as estruturas secundárias do proteoma humano através do consenso entre as três ferramentas de predição (ver Materiais e Métodos), e classificamos o proteoma de acordo com sua origem em clados específicos da linhagem humana. Utilizamos a classificação em oito grupos determinada por (LIEBESKIND; MCWHITE; MARCOTTE, 2016) baseada no consenso de 13 métodos de inferência de grupos de ortólogos. As proteínas foram classificadas em: *Cellular_organisms*, *Euk+Bac*, *Euk_Archaea*, *Eukaryota*, *Opisthokonta*, *Eumetazoa*, *Vertebrata* e *Mammalia*. A quantidade de proteínas classificadas em cada grupo é indicada na Figura 27a. Com isso, analisamos a composição estrutural, ou seja, a porcentagem de alfa-hélice, fita beta e irregular de cada proteína classificada de acordo com sua origem (Figura 27c). Dessa forma podemos notar que existe um decréscimo no conteúdo estrutural (alfa e beta) em proteínas mais recentes em comparação com as mais antigas. Diferenças significativas foram observadas para alfa-hélice na transição de *Eukaryota* para *Opisthokonta*, enquanto que para fita beta a maior diferença ocorre entre organismos celulares (*Cellular_organisms*, *Euk+Bac* e *Euk_Archaea*) e *Eukaryota*. Estas diferenças se somam a diferenças menores entre os demais clados chegando a um total de mais de 17% de diferença entre a porcentagem de resíduos com estruturas irregular encontradas em *Cellular_organisms* (com 45,61%) e *Mammalia* (com 63,47%) (Tabela 11). Outra característica interessante observada ao longo da evolução é o aumento de diversidade estrutural das proteínas mais recentes (Figura 27b). Notamos que antes de *Eukaryota* as proteínas apresentavam conteúdo em torno de 40% de alfa-hélice e 17% de fita beta. A partir daí, começam a surgir proteínas com conteúdo mais diversificado, muitas vezes com alto teor de alfa-hélice e pouca fita beta ou vice-versa. Há também a aparição de sequências com pouca estrutura de ambas, principalmente em *Opisthokonta*, que apresenta muitas proteínas com menos de 20% de alfa e beta.

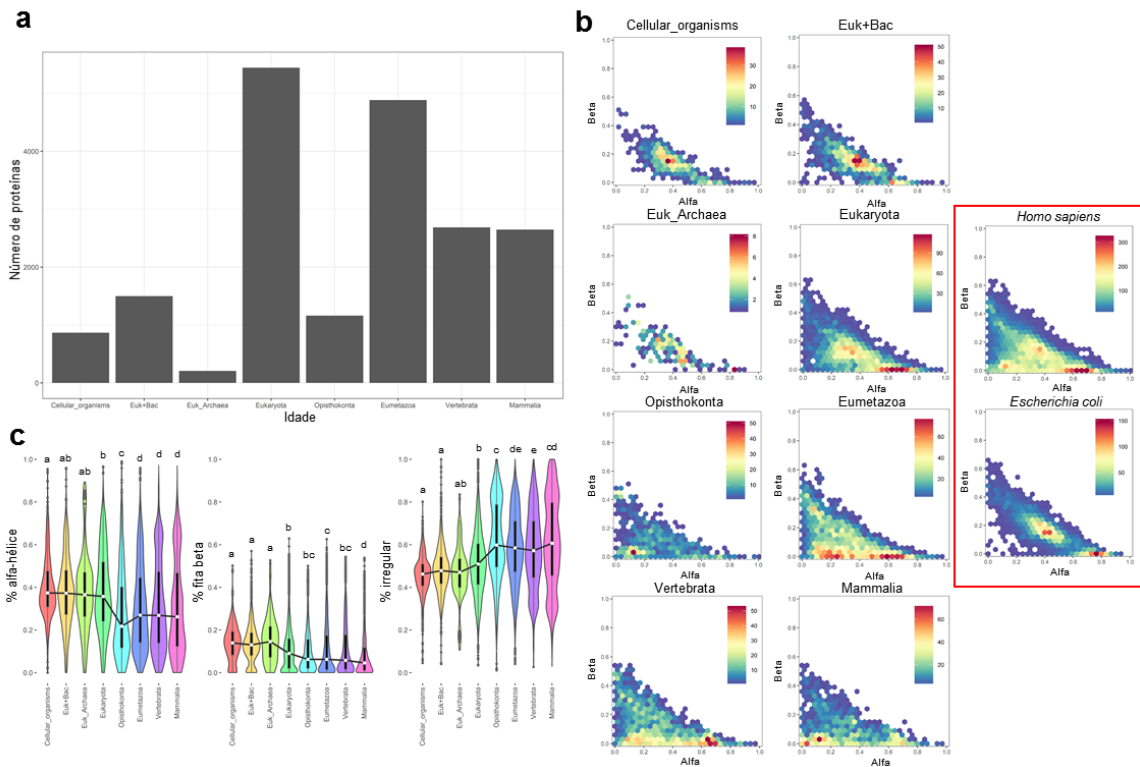


Figura 27 - Classificação por idade evolutiva do proteoma humano. Dados de estrutura secundária são referentes ao consenso da predição entre três métodos utilizados (ver Materiais e Métodos) **(a)**, mostra quantidades de proteínas por idade inferida. **(b)**, mostra gráficos de calor com a quantidade de seqüências pela composição de alfa-beta para idade. Em destaque à direita (em vermelho), gráficos para os proteomas de *Homo sapiens* e *Escherichia coli*. Cores variando de azul para vermelho indicam de poucas para muitas seqüências respectivamente. **(c)**, mostra composição de estrutura secundária (% de alfa-hélice, fita beta e irregular de cada seqüência). Letras diferentes indicam grupos com diferenças estatisticamente significantes usando o teste de Kruskal-Wallis e contraste de Dunnett (valor $p < 0,001$).

Tabela 11 - Seqüências de *Homo sapiens* por idade evolutiva

Idade	N	Comprimento Médio (EP)	% de alfa-hélice*	% de fita beta*	% irregular*
<i>Cellular_organisms</i>	867	541,13 (11,652)	41,75%	12,64%	45,61%
<i>Euk+Bac</i>	1495	526,10 (8,693)	37,70%	13,42%	48,89%
<i>Euk_Archaea</i>	208	418,20 (22,544)	41,10%	12,60%	46,29%
<i>Eukaryota</i>	5440	631,89 (7,595)	37,63%	10,11%	52,26%
<i>Opisthokonta</i>	1160	660,37 (15,766)	26,64%	9,88%	63,48%
<i>Eumetazoa</i>	4877	627,12 (8,368)	26,59%	12,13%	61,27%
<i>Vertebrata</i>	2684	508,90 (9,239)	27,49%	10,98%	61,53%
<i>Mammalia</i>	2647	368,27 (7,922)	29,18%	7,34%	63,47%

*A composição de alfa-hélices, fitas beta e de resíduos irregulares (referentes ao consenso da predição entre três métodos utilizados) é calculada como a porcentagem de ocorrência do total de resíduos considerando todas as seqüências por idade evolutiva.

N indica a quantidade de seqüências.

EP representa o erro padrão.

Outra característica que apresenta diferenças ao longo da evolução são os tamanhos das proteínas. Seqüências mais recentes tendem a apresentar comprimento menor (Figura 28a). Tal característica já foi descrita em outros estudos. (NEME; TAUTZ, 2013; YIN et al., 2016). Seqüências com origem em *Mammalia* apresentam em média 368

resíduos enquanto sequências de *Cellular_organisms* apresentam 541 resíduos (Tabela 11). Com relação aos tamanhos dos segmentos com estrutura secundária contínua, notamos também uma tendência para segmentos de alfa-hélice e fita beta menores nas proteínas mais recentes (Figura 28c). Isso está de acordo com a redução estrutural observada anteriormente. Já quanto a relação entre composição estrutural e comprimento de sequência, notamos que não há clara correlação (Figura 28b).

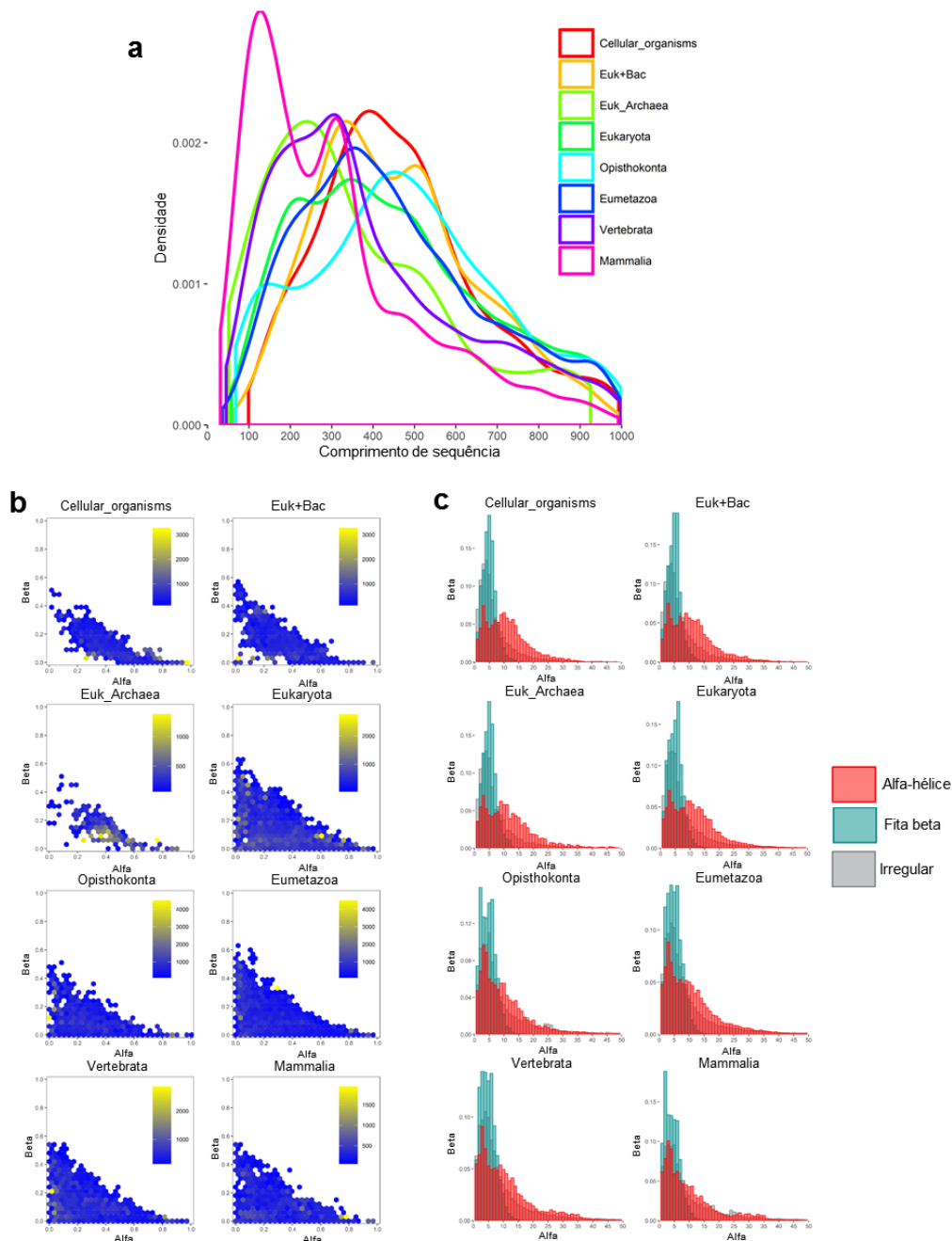


Figura 28 - Comprimento de seqüências do proteoma humano por idade evolutiva. Dados de estrutura secundária são referentes ao consenso da predição entre três métodos utilizados (ver Materiais e Métodos). **(a)**, mostra a distribuição de comprimentos de seqüências por idade evolutiva. **(b)**, mostra mapas de calor com o comprimento médio de seqüências pela composição alfa-beta. Cores variando de azul para amarelo indicam seqüências curtas para longas respectivamente. Cada gráfico apresenta escala de valores própria no canto superior direito. **(c)**, distribuição do comprimento de segmentos de alfa-hélice, fita beta e irregular em cada idade evolutiva.

4.3.1.1. Uso de aminoácidos

Um dos fatores principais que afetam a conformação de estrutura secundária é o repertório de aminoácidos e a sua frequência (“uso de aminoácidos”). Certos aminoácidos possuem maior probabilidade de apresentar certas conformações estruturais. Por exemplo, alanina, glutamato, lisina, leucina, metionina, glutamina e arginina tem maior propensão de formar hélices, enquanto cisteína, fenilalanina, isoleucina, treonina, valina, triptofano e tirosina possuem mais chances de formar folhas beta (COSTANTINI; COLONNA; FACCHIANO, 2006) (Tabela 1). Dessa forma, analisar mudanças no uso de aminoácidos durante a evolução pode também auxiliar o entendimento das mudanças observadas nas estruturas das proteínas.

Na Figura 29, mostramos a diferença no uso de aminoácidos entre proteínas de *Homo sapiens* com origem antiga (*Cellular_organisms*) e recente (*Mammalia*). A classificação quanto a origem das proteínas foi obtida através do consenso de diversos métodos de inferência de ortologia (ver Materiais e Métodos para mais detalhes). Comparamos esses dois grupos por representar os extremos das idades estabelecidas com esta metodologia. Notamos diferenças significativas entre vários resíduos (Tabela 12). Destacamos a maior ocorrência de resíduos de cisteína, prolina, serina em proteínas mais recentes quando comparado com mais antigas e, de isoleucina, aspartato e valina em mais antigas quando comparado com mais recentes. Principalmente a diferença no uso de cisteína, que supera 100%. Esta diferença foi também observada nas sequências de estruturas do PDB (Figura 29a e Tabela 10).

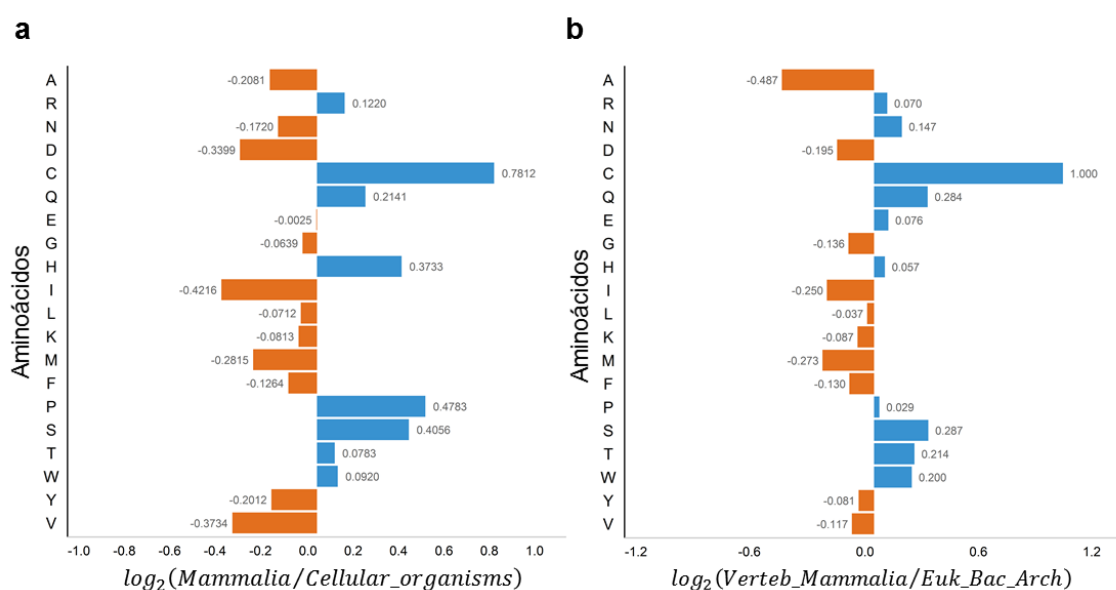


Figura 29 - Uso de aminoácidos em sequências do proteoma de *Homo sapiens* com idade em *Cellular_organisms* e *Mammalia*. (a), Barras mostram a razão da porcentagem de uso aminoácidos de proteínas datadas em *Mammalia* por proteínas datadas em *Cellular_organisms*. Valores negativos (em laranja) indicam maior ocorrência em *Cellular_organisms*, já valores positivos (em azul) indicam maior ocorrência em *Mammalia*. (b), reproduz a Figura 24 para comparação, com resultados obtidos com dados do PDB. A razão está em escala logarítmica (\log_2).

Tabela 12 - Uso de aminoácidos em sequências do proteoma de *Homo sapiens* com idade em *Cellular_organisms* e *Mammalia*.

Abreviação	Símbolo	Frequência média (EP)*		Diferença percentual†	Significância	Complexidade‡	Códons
		<i>Cellular organisms</i>	<i>Mammalia</i>				
I	Ile	5,41 (0,062)	3,99 (0,050)	25,41	**	16,04	ATT/ATC/ATA
D	Asp	4,83 (0,046)	3,57 (0,038)	22,83	**	32,72	GAT/GAC
V	Val	7,13 (0,059)	5,39 (0,044)	22,19	**	12,28	GTN
A	Ala	7,90 (0,075)	6,80 (0,066)	13,64	**	4,76	GCN
N	Asn	3,57 (0,045)	3,07 (0,035)	12,80	**	33,72	AAT/AAC
K	Lys	6,01 (0,079)	5,39 (0,068)	12,26	**	30,14	AAA/AAG
Y	Tyr	2,78 (0,037)	2,53 (0,045)	10,41	**	57	TAT/TAC
M	Met	2,55 (0,031)	2,33 (0,026)	8,03	**	64,68	ATG
F	Phe	4,04 (0,049)	3,70 (0,045)	5,77	**	44	TTT/TTC
E	Glu	6,65 (0,073)	6,20 (0,072)	4,06	**	36,48	GAA/GAG
G	Gly	7,23 (0,069)	6,82 (0,078)	3,95	**	1	GGN
L	Leu	10,34 (0,080)	10,17 (0,078)	2,28		16,04	CTN/TTA/TTG
T	Thr	5,14 (0,045)	5,30 (0,044)	-3,53		21,62	ACN
R	Arg	5,19 (0,058)	5,87 (0,061)	-5,25	**	56,34	CGN/AGA/AGG
W	Trp	1,23 (0,028)	1,41 (0,023)	-6,23		73	TGG
Q	Gln	4,07 (0,048)	4,76 (0,048)	-11,90	**	37,48	CAA/CAG
H	His	2,40 (0,032)	2,81 (0,036)	-16,50	**	58,7	CAT/CAC
S	Ser	6,68 (0,057)	8,84 (0,070)	-31,98	**	17,86	TCN/AGT/AGC
P	Pro	4,97 (0,052)	7,16 (0,087)	-37,80	**	31,8	CCN
C	Cys	1,87 (0,030)	3,86 (0,090)	-104,02	**	57,16	TGT/TGC

*Frequência média (porcentagem e erro padrão, EP) de uso de aminoácidos em *Cellular_organisms* e *Mammalia*.

†Os aminoácidos estão listados da maior para menor diferença percentual entre genes de organismos celulares contra genes de mamíferos.

‡Escores de complexidade representam o custo bioquímico de produção e estabilidade conformacional em uma proteína (DUFTON, 1997; WILLIFORD; DEMUTH, 2012).

** representa aminoácidos com diferença significativa de acordo com teste Mann-Whitney-Wilcoxon (valor $p < 0.001$).

4.3.1.2. Enriquecimento funcional

Para melhor identificar as proteínas relacionadas a cada idade evolutiva, realizamos uma análise de enriquecimento funcional de termos do *Gene Ontology* (ver Materiais e Métodos para detalhes). Na Figura 30, temos os termos de processos para as oito idades, determinadas para as proteínas humanas. Vemos que proteínas datadas com origem mais antiga (*Cellular_organisms*, *Euk+Bac* e *Euk_Archaea*) estão mais envolvidos com processos metabólicos basais enquanto proteínas recentes (*Vertebrata* e *Mammalia*) estão envolvidos com processos mais específicos, como resposta a estímulos.

Termos de funções mostram atividades catalíticas em proteínas procarióticas, atividades de quinases e GTPases em *Eukaryota*, ligação a sítios específicos de DNA em *Opisthokonta* e *Eumetazoa*, atividades de receptor de sinais e receptores olfatórios em *Vertebrata* e *Mammalia* (Figura 31).

Na Figura 32, temos os termos relacionados quanto a localização das proteínas (termos de componentes). Vemos que *Cellular_organisms* e *Euk+Bac* apresentam enriquecimento de proteínas na matriz mitocondrial, enquanto proteínas com origem em *Euk_Archaea* têm enriquecimento de proteínas do citosol. Em *Eukaryota* e *Opisthokonta*

vemos proteínas nucleares. E nas proteínas mais recentes vemos enriquecimento de proteínas de meio extracelular e de membrana.

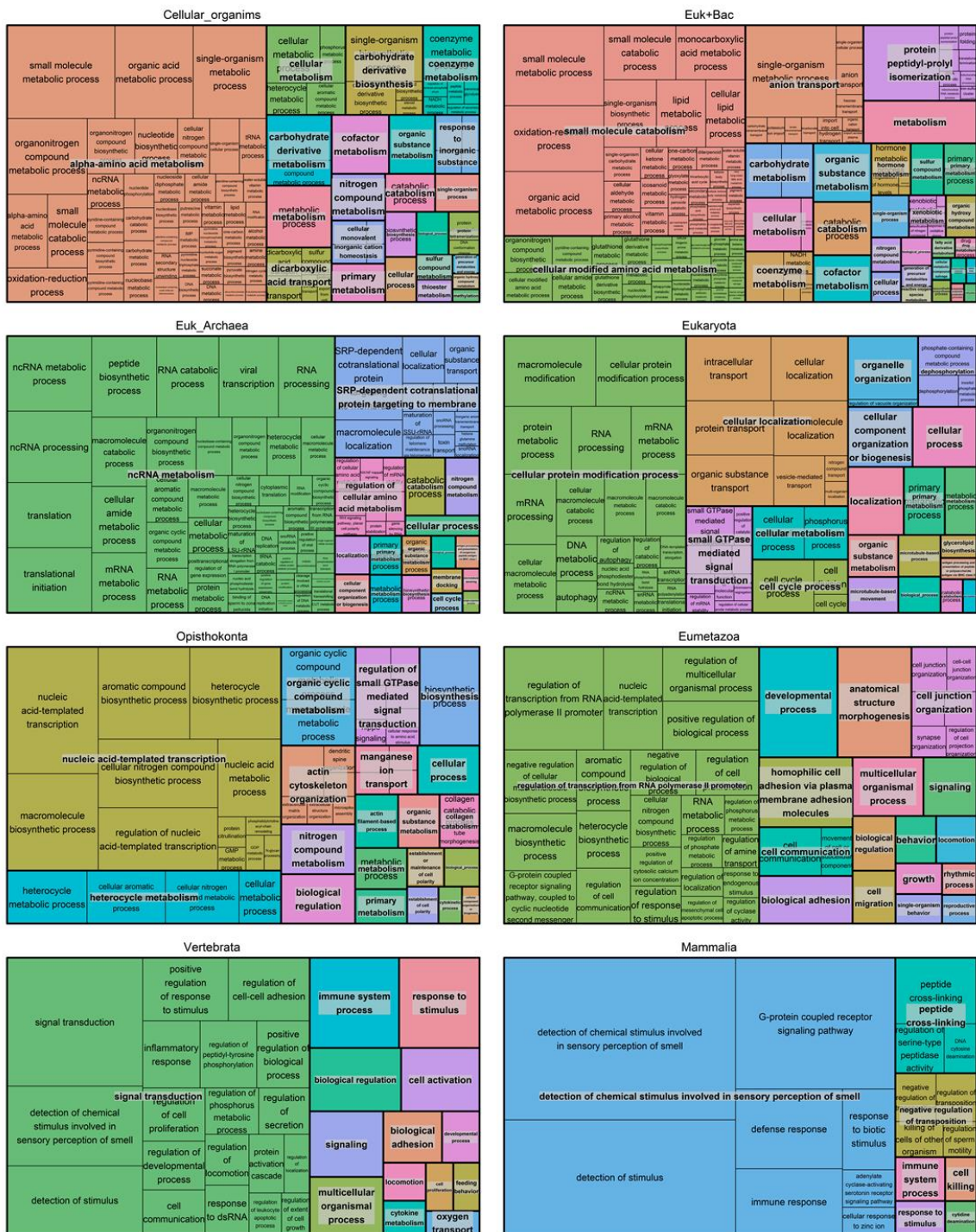


Figura 30 - Enriquecimento de termos de processos do Gene Ontology. A figura mostra o enriquecimento para termos de processos do Gene Ontology para cada idade evolutiva. Termos foram enriquecidos usando GOrrilla com valor $p < 0.01$. A lista dos termos foi então filtrada usando REVIGO usando a SimRel como semântica de similaridade e corte de similaridade de 0.5.

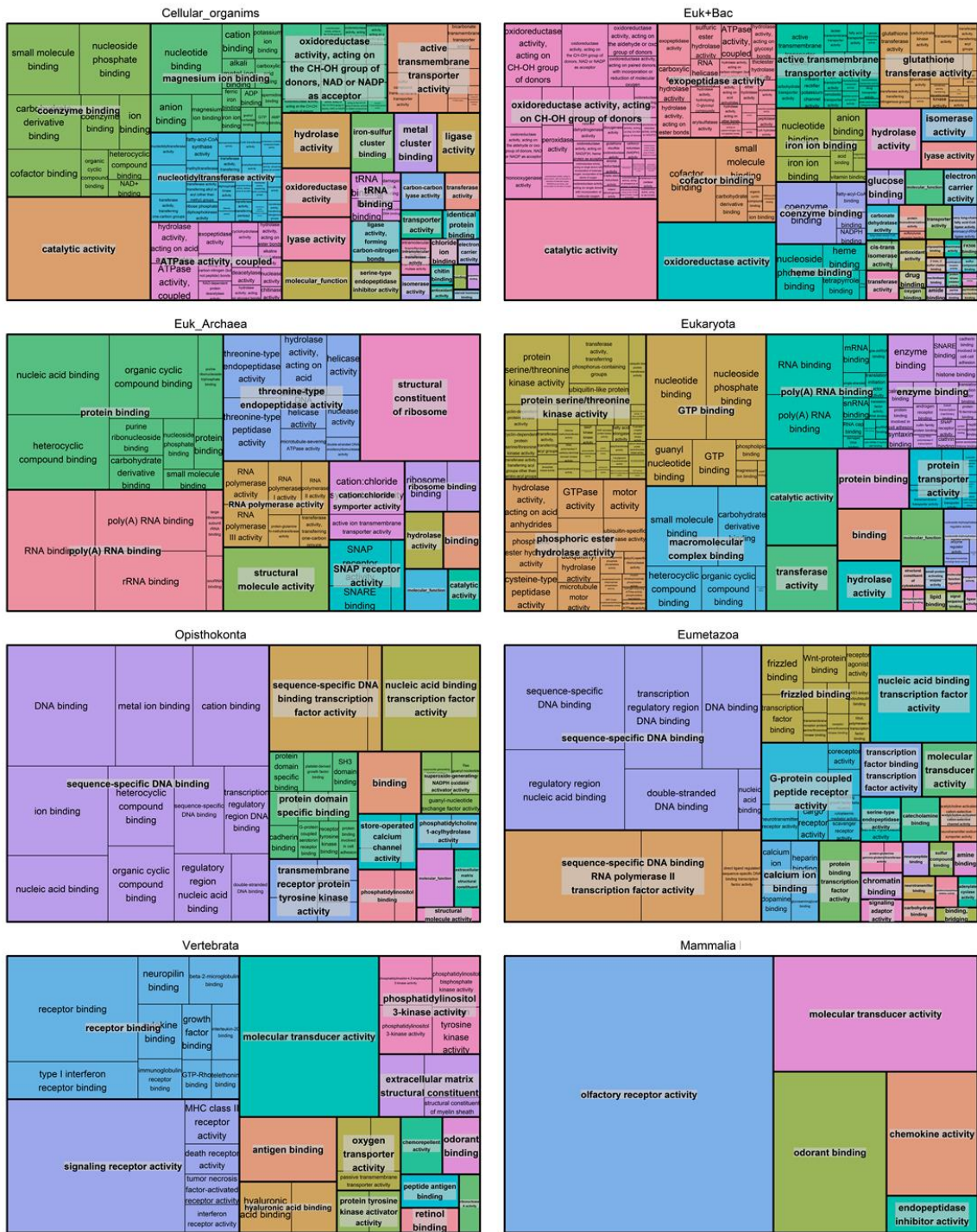


Figura 31 - Enriquecimento funcional de termos do *Gene Ontology*. A figura mostra o enriquecimento para termos funcionais do *Gene Ontology* para cada idade evolutiva. Termos foram enriquecidos usando GOrilla com valor $p < 0.01$. A lista dos termos foi então filtrada usando REVIGO usando a SimRel como semântica de similaridade e corte de similaridade de 0.5.



Figura 32 - Enriquecimento de termos de componentes do *Gene Ontology*. A figura mostra o enriquecimento para termos de componentes do *Gene Ontology* para cada idade evolutiva. Termos foram enriquecidos usando GOrilla com valor $p < 0.01$. A lista dos termos foi então filtrada usando REVIGO usando a SimRel como semântica de similaridade e corte de similaridade de 0.5.

Com o objetivo de identificar as proteínas de acordo com seu conteúdo estrutural, realizamos o enriquecimento daquelas que possuem pouca estrutura (menos de 20% de composição de alfa-hélice e fita beta) (Figura 33). Termos relacionados a essas proteínas são: funções de ligação de sítios específicos de DNA, fatores de transcrição e íons de zinco

(Figura 33a); processos de regulação transcrição, biossíntese e metabolismo (Figura 33b); e componentes nucleares e matriz extracelular (Figura 33c). Ao comparar estes termos com os termos enriquecidos por idade, vemos semelhança com os termos enriquecidos em *Opisthokonta*.

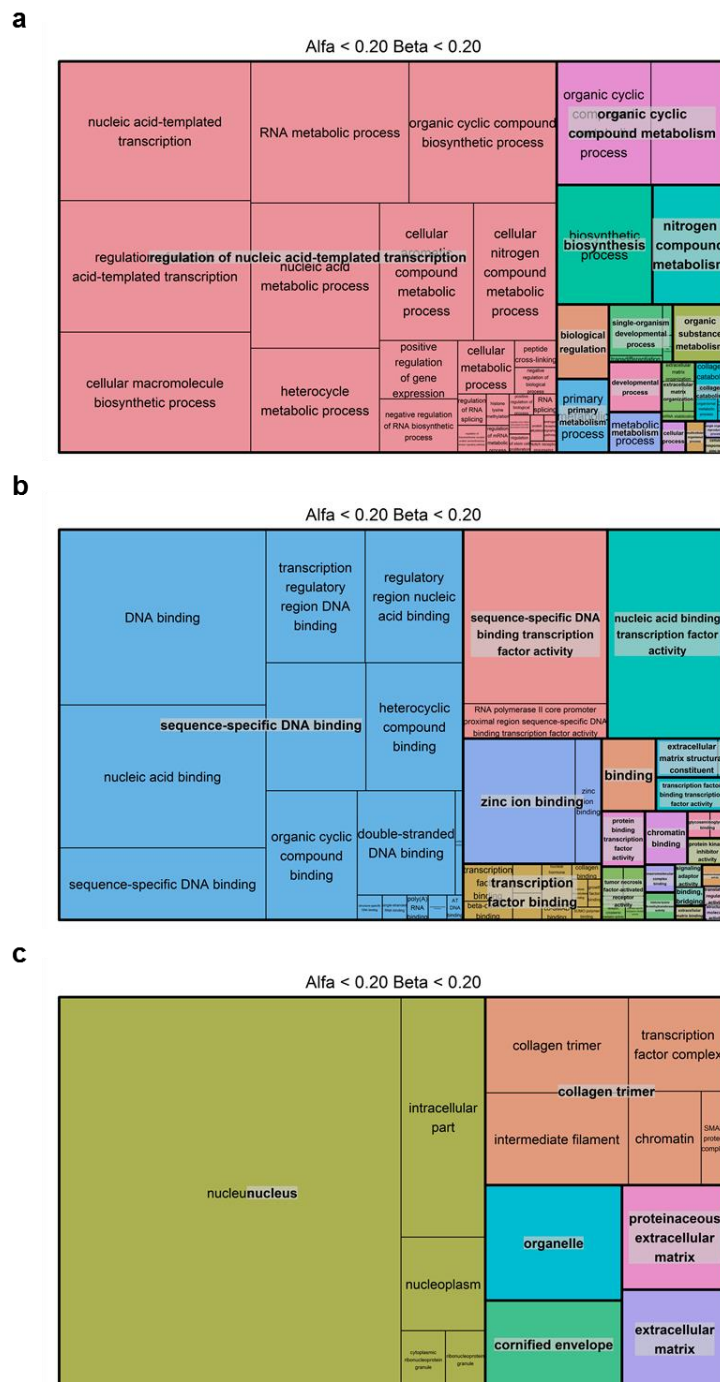


Figura 33 - Enriquecimento de termos do *Gene Ontology* para proteínas pouco estruturadas. A figura mostra o enriquecimento para termos do *Gene Ontology* para proteínas de *Homo sapiens* que possuem conteúdo de alfa-hélice e fita beta inferiores à 20%. **(a)**, mostra termos para processos. **(b)**, mostra termos para funções. **(c)**, mostra termos para componentes. Termos foram enriquecidos usando GOrilla com valor $p < 0.01$. A lista dos termos foi então filtrada usando REVIGO usando a SimRel como semântica de similaridade e corte de similaridade de 0.5.

4.3.1.3. Estruturas relacionadas a domínios

Analizamos também informações relacionadas a domínios estruturais. Utilizamos informações da base de dados Gene3D (LAM et al., 2016), que fornece predições de domínios para sequências do UniProt e Ensembl e é organizado de acordo com a classificação hierárquica do CATH (Classe, Arquitetura, Topologia e Homologia). A Figura 34 mostra a classificação de classes e arquiteturas para proteínas nas oito idades evolutivas (*Cellular_organisms*, *Euk+Bac*, *Euk_Archaea*, *Eukaryota*, *Opisthokonta*, *Eumetazoa*, *Vertebrata* e *Mammalia*). A classificação apresenta resultados compatíveis com os resultados observados com dados de estrutura secundária. Domínios de proteínas com origem antiga apresentam maior ocorrência de classes com composição mista de alfa-hélice e fita beta (em azul). Enquanto que, à medida que analisamos proteínas mais recentes (principalmente após *Opisthokonta*), vemos maior ocorrência de classes com composição majoritária de alfa-hélice ou fita beta. Além de apresentarem maior proporção de domínios com pouca estrutura secundária (em laranja). Essas informações suportam a maior diversidade estrutural encontrada nestas idades.

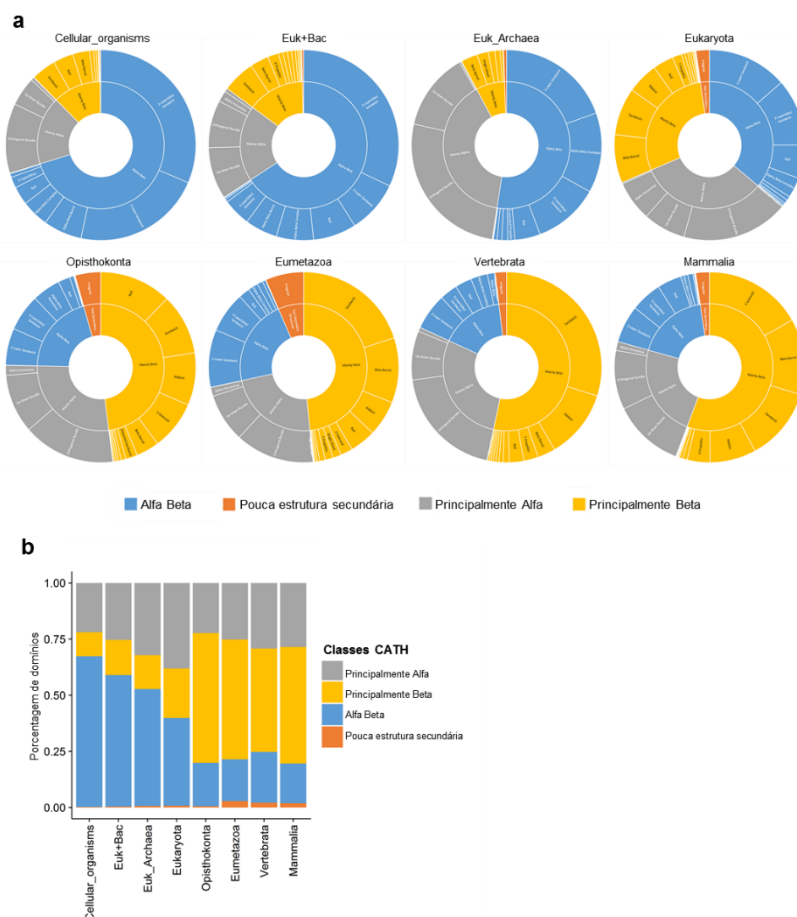


Figura 34 - Classificação de domínios CATH para sequências do proteoma de *Homo sapiens* preditas no Gene3D por idade evolutiva. (a), classificação de domínios estruturais em termos de classes (anel interno) e arquiteturas (anel externo) CATH para o proteoma humano separadas por idade evolutiva. (b), mostra as porcentagens de ocorrência dos domínios nas quatro classes para cada idade evolutiva (principalmente alfa (cinza), principalmente beta (amarelo), mistura alfa-beta (azul) e pouca estrutura secundária (laranja)).

Para verificar se a redução estrutural em proteínas recentes poderia ser explicada somente devido a regiões de domínio ou complementares a eles, separamos as sequências do proteoma humano de acordo com regiões intradomínio e extradomínio. Utilizamos as coordenadas dos domínios preditos no Gene3D para separar as regiões e medimos suas composições estruturais. Somente regiões com pelo menos 50 resíduos foram considerados. Com isso, observamos que domínios presentes em proteínas com origem em organismos celulares (*Cellular_organisms*, *Euk+Bac* e *Euk_Archaea*) têm maior cobertura de sequência, enquanto proteínas com origem mais recente apresentam maior variação em termos de cobertura (Figura 35). Além disso, vimos que proteínas recentes apresentam uma grande parcela de sequências sem domínios preditos (cerca de 50% em *Mammalia*).

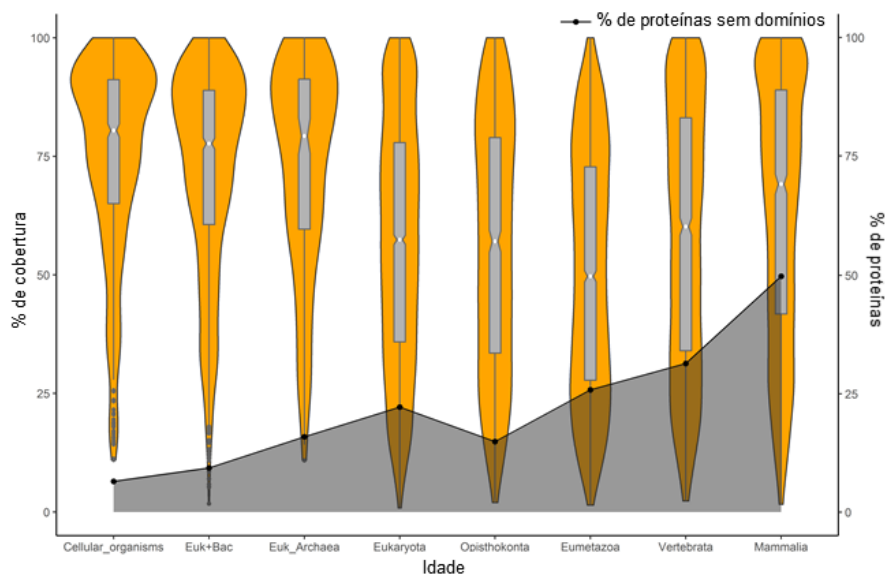


Figura 35 - Cobertura de domínios por idade evolutiva. Gráficos de violino mostram a distribuição da porcentagem de cobertura de sequência de regiões de domínios em relação à proteína completa. São consideradas somente proteínas com pelo menos um domínio mapeado no Gene3D. Em casos em que há mais de um domínio, são consideradas todas as regiões para cálculo da cobertura. A porcentagem de proteínas sem domínios mapeados é mostrada na curva preta sobreposta.

Ao compararmos a composição de estrutura secundária, notamos que regiões extradomínios apresentam menor conteúdo de alfa-hélices e fitas beta do que regiões intradomínios (Figura 36). Podemos notar também uma redução de alfa-hélices em proteínas mais recentes, independentemente da região ser de domínio ou não, principalmente entre *Eukaryota* e *Opisthokonta* (Figura 36b). Apesar dessa diferença ser mais evidente em regiões intradomínios (13,32% de diferença entre *Eukaryota* e *Opisthokonta*), percebemos também, de forma mais suave, em regiões extradomínios (9,50% de diferença) (Tabela 13). Com relação a mudanças na composição de fitas beta, vemos que há uma redução considerável em *Opisthokonta* e *Mammalia* para regiões

intradomínio (onde são mais frequentes), enquanto que, para regiões extradomínio, o conteúdo de fitas beta é muito baixo e não apresenta diferenças significativas quanto à idade evolutiva (Figura 36b).

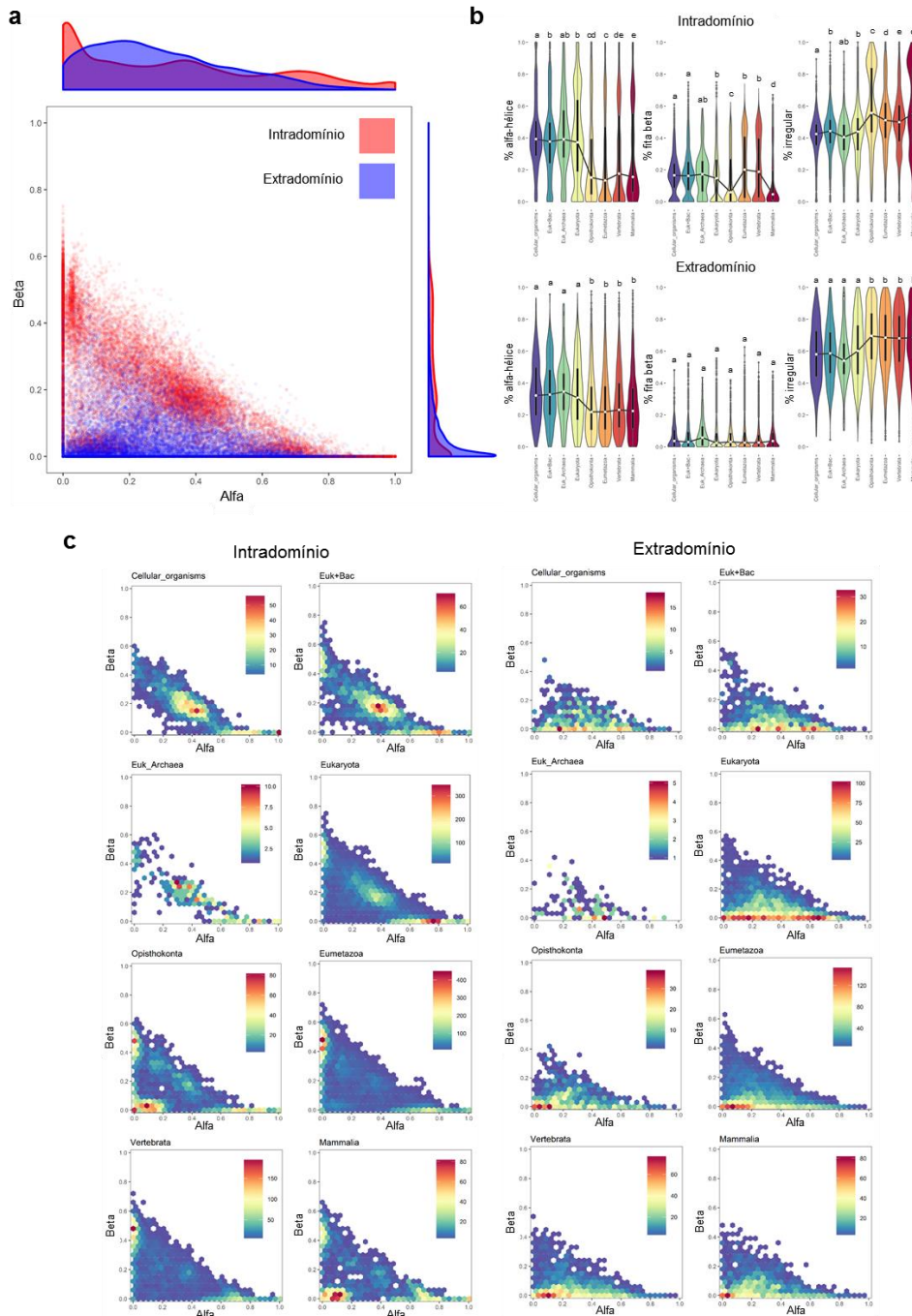


Figura 36 - Composição estrutural de domínios e regiões extradomínios. Dados de estrutura secundária são referentes ao consenso da predição entre três métodos utilizados (ver Materiais e Métodos). **(a)**, gráfico de dispersão com a composição de alfa-hélices e fitas beta de segmentos intradomínio (em vermelho) e extradomínio (em azul). Curvas de densidade para cada estrutura são exibidas nas laterais. **(b)**, composição de estrutura secundária por idade evolutiva para regiões de domínio e extradomínio. Letras diferentes indicam idades com diferenças estatisticamente significativas usando o teste de Kruskal-Wallis e contraste de Dunnett (valor $p < 0,001$). **(c)**, gráficos de calor com a quantidade de seqüências pela composição de alfa-beta para regiões de domínio e extradomínio. Cores variando de azul para vermelho indicam de poucas para muitas seqüências respectivamente.

Tabela 13 - Composição estrutural de regiões de domínios e extra-domínios por idade evolutiva.

Região	Idade	N	Comprimento Médio (EP)	% de alfa-hélice*	% de fita beta*	% irregular*
Intradomínio	<i>Cellular_organisms</i>	1874	155,80 (1,905)	42,64%	16,14%	41,22%
	<i>Euk+Bac</i>	2735	170,90 (1,746)	38,94%	16,95%	44,10%
	<i>Euk_Archaea</i>	311	135,97 (3,701)	42,72%	17,57%	39,71%
	<i>Eukaryota</i>	9466	140,18 (0,834)	41,68%	16,16%	42,16%
	<i>Opisthokonta</i>	2804	112,74 (1,093)	28,36%	15,42%	56,22%
	<i>Eumetazoa</i>	8943	117,64 (0,683)	28,92%	22,06%	49,02%
	<i>Vertebrata</i>	3682	133,03 (1,260)	33,24%	20,10%	46,67%
	<i>Mammalia</i>	2582	124,96 (1,504)	37,62%	11,62%	50,76%
	Total	32397	133,02 (0,427)	36,10%	17,75%	46,14%
Extradomínio	<i>Cellular_organisms</i>	582	279,42 (12,380)	40,57%	6,66%	52,77%
	<i>Euk+Bac</i>	1109	271,36 (8,941)	35,79%	8,16%	56,05%
	<i>Euk_Archaea</i>	131	321,32 (27,266)	39,87%	7,94%	52,19%
	<i>Eukaryota</i>	4673	435,23 (6,667)	35,01%	6,13%	58,86%
	<i>Opisthokonta</i>	1010	425,18 (12,885)	25,51%	5,78%	68,71%
	<i>Eumetazoa</i>	4503	428,36 (6,887)	25,38%	6,71%	67,91%
	<i>Vertebrata</i>	2227	378,00 (8,428)	24,21%	5,62%	70,17%
	<i>Mammalia</i>	2149	291,97 (8,121)	25,01%	5,08%	69,91%
	Total	16384	388,62 (3,333)	29,25%	6,23%	64,52%

*A composição de alfa-hélices, fitas beta e resíduos irregulares (referentes ao consenso da predição entre três métodos utilizados) é calculada como a porcentagem de ocorrência do total de resíduos considerando todas as sequências para regiões de domínios e extradomínios.

N indica a quantidade de sequências.

EP representa o erro padrão.

Diferenças no uso de aminoácidos também foram analisadas. Regiões intradomínio apresentaram maior uso, principalmente, de cisteína (C), tirosina (Y), isoleucina (I) e fenilalanina (F) em relação a regiões extradomínios. Por outro lado, resíduos de prolina (P) e serina (S) apresentaram uso superior em regiões extradomínios (Figura 37). Proteínas com diferentes idades evolutivas também apresentaram variações no uso de aminoácidos entre domínios e extradomínios. A Figura 38 mostra que sequências com origem antiga tem menor variação enquanto que em proteínas recentes, alguns aminoácidos específicos apresentaram maior diferença de uso, como a cisteína (mais usada em regiões intradomínio) e a prolina (mais usada em regiões extradomínio).

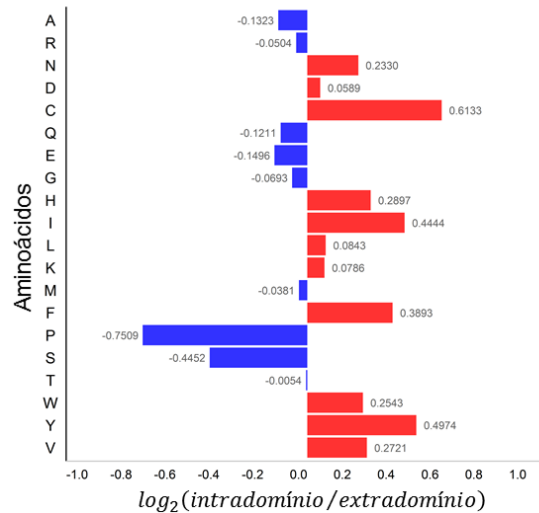


Figura 37 - Uso de aminoácidos em regiões de domínios e extradomínios. Barras mostram a razão da porcentagem de uso aminoácidos em intradomínios por regiões extradomínios. A razão está em escala logarítmica (log₂). Valores negativos (em azul) indicam maior ocorrência em extradomínios, valores positivos (em vermelho) indicam maior ocorrência em intradomínios.

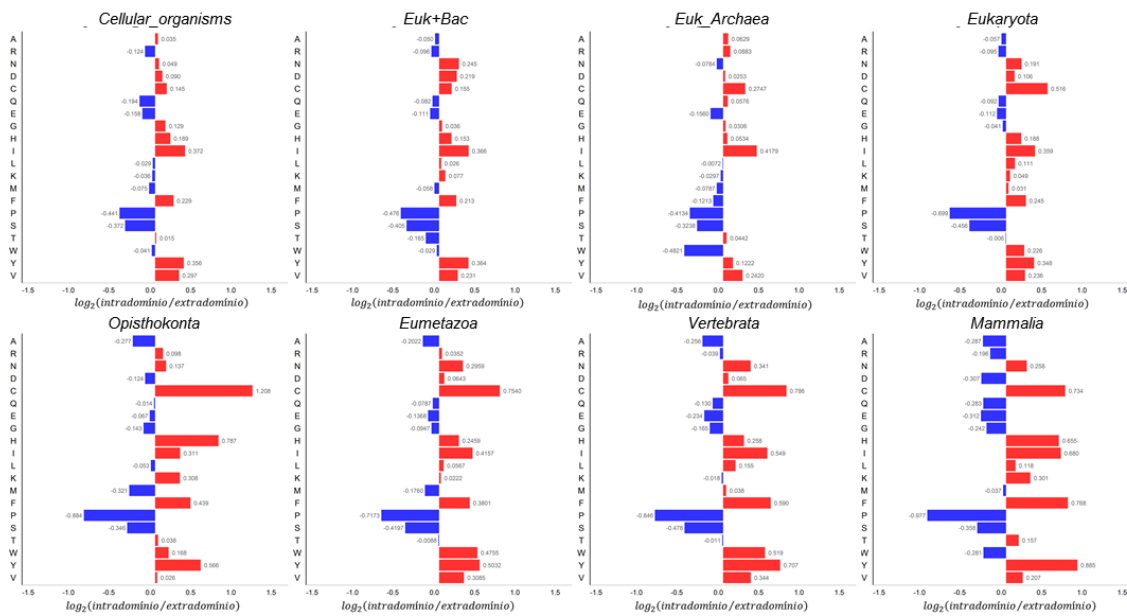


Figura 38 - Uso de aminoácidos em regiões de domínios e extradomínios por idade evolutiva. Barras mostram a razão da porcentagem de uso aminoácidos em intradomínios por regiões extradomínios. A razão está em escala logarítmica (log₂). Valores negativos (em azul) indicam maior ocorrência em extradomínios, valores positivos (em vermelho) indicam maior ocorrência em intradomínios.

4.3.2. Proteínas com origem *de novo* e ORFs de regiões não codificadoras

Genes com origem *de novo* são definidos como aqueles cuja origem decorre, ao menos em parte, a partir de sequências previamente não codificadoras (MCLYSAGHT; HURST, 2016). Investigamos as possíveis estruturas de proteínas indicadas como tendo essa

característica. Com base em uma lista de genes de primatas com provável origem *de novo* compilada por (MCLYSAGHT; GUERZONI, 2015) (Tabela 4), realizamos a predição consenso destas sequências e notamos que a maioria delas apresentou baixo conteúdo de estrutura de alfa-hélices e fitas beta (Figura 39).

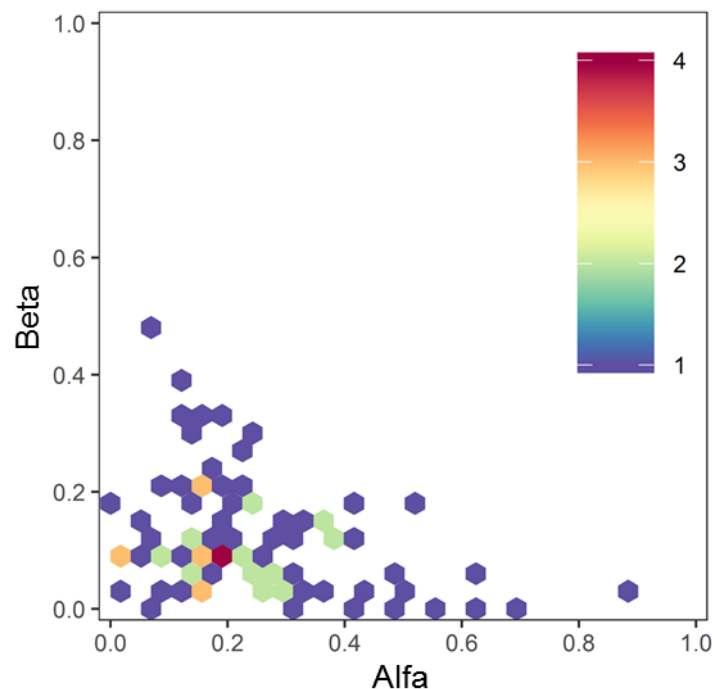


Figura 39 - Composição alfa-beta de sequências de primatas com origem de novo. Gráficos mostram a quantidade de sequências pela composição alfa-beta. Cores variando de azul para vermelho indicam de poucas para muitas sequências respectivamente.

Assim, a fim de investigar a viabilidade para origem de prováveis genes a partir de regiões não codificadoras, selecionamos ORFs de regiões de *frameshifts* e sequências antisense de proteínas de *Homo sapiens*. Selecionamos somente ORFs com mais de 50 resíduos a partir das sequências de cDNA. Realizamos a predição dessas sequências com as três ferramentas e calculamos o consenso (porcentagens das votações podem ser consultadas na Tabela 5). A Figura 40 mostra as distribuições das estruturas das ORFs preditas. Notamos que, em geral, o conteúdo estrutural das ORFs é inferior às estruturas observadas em proteínas reais. No entanto, observamos que ainda assim, ocorrem proteínas com composição estrutural próxima da esperada biologicamente. Vemos também que ORFs derivadas de sequências antisense de proteínas (linhas em tons avermelhados na Figura 40) apresentam maior similaridade com as distribuições do proteoma e do proteoma permutado. Já ORFs derivadas de mudanças na fase de leitura das sequências de cDNA (linha em tons azuis) apresentam distribuições mais semelhantes às encontradas

em proteínas *de novo*, principalmente a “Fita + Frame 2” (mudança de uma base na fase de leitura).

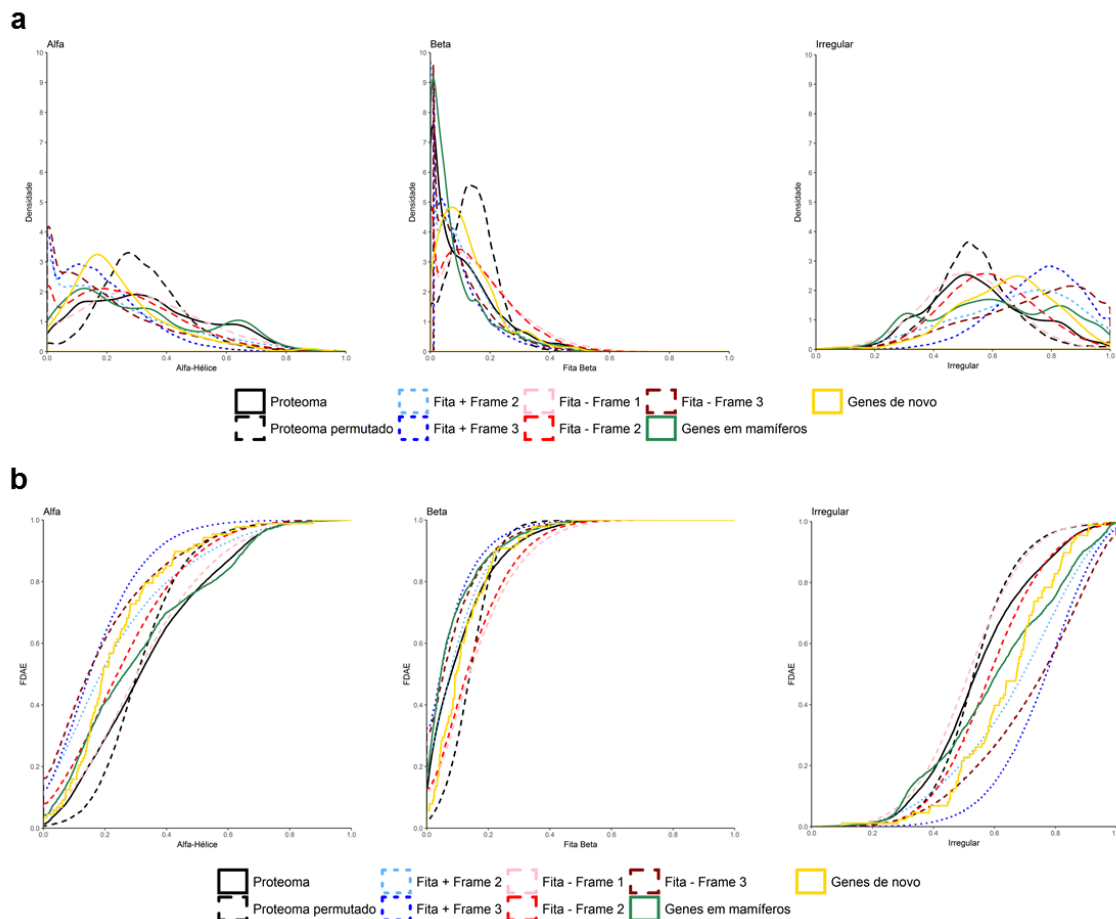


Figura 40 - Estrutura secundária em regiões não codificadoras. ORFs (>50aa) derivadas de *frameshifts* de seqüências senso e antisenso em *Homo sapiens*. **(a)**, distribuição de densidades para conteúdo de alfa-hélice, fita beta e irregular. **(b)**, mostra gráficos de função distribuição acumulada empírica (FDAE) para conteúdo de alfa-hélice, fita beta e irregular. Para fins de comparação, é mostrado o uso de aminoácidos do proteoma humano (linha sólida em preto), proteoma humano embaralhado (em preto tracejado), de proteínas com origem em mamíferos (em verde) e de genes com origem de novo (em amarelo).

Analizamos também diferenças quanto ao uso de aminoácidos nas seqüências *de novo* e seqüências derivadas de regiões não codificadoras (Figura 41). Comparamos o uso de cada grupo com o uso médio no proteoma completo. Notamos grandes diferenças entre as fases de leitura. ORFs derivadas de seqüências senso com mudança de leitura de duas bases (Fita + Frame 3), apresentam maiores discrepâncias. Vemos também que genes *de novo* apresentam as mesmas tendências de uso observadas em seqüências com origem a partir de *Mammalia*.

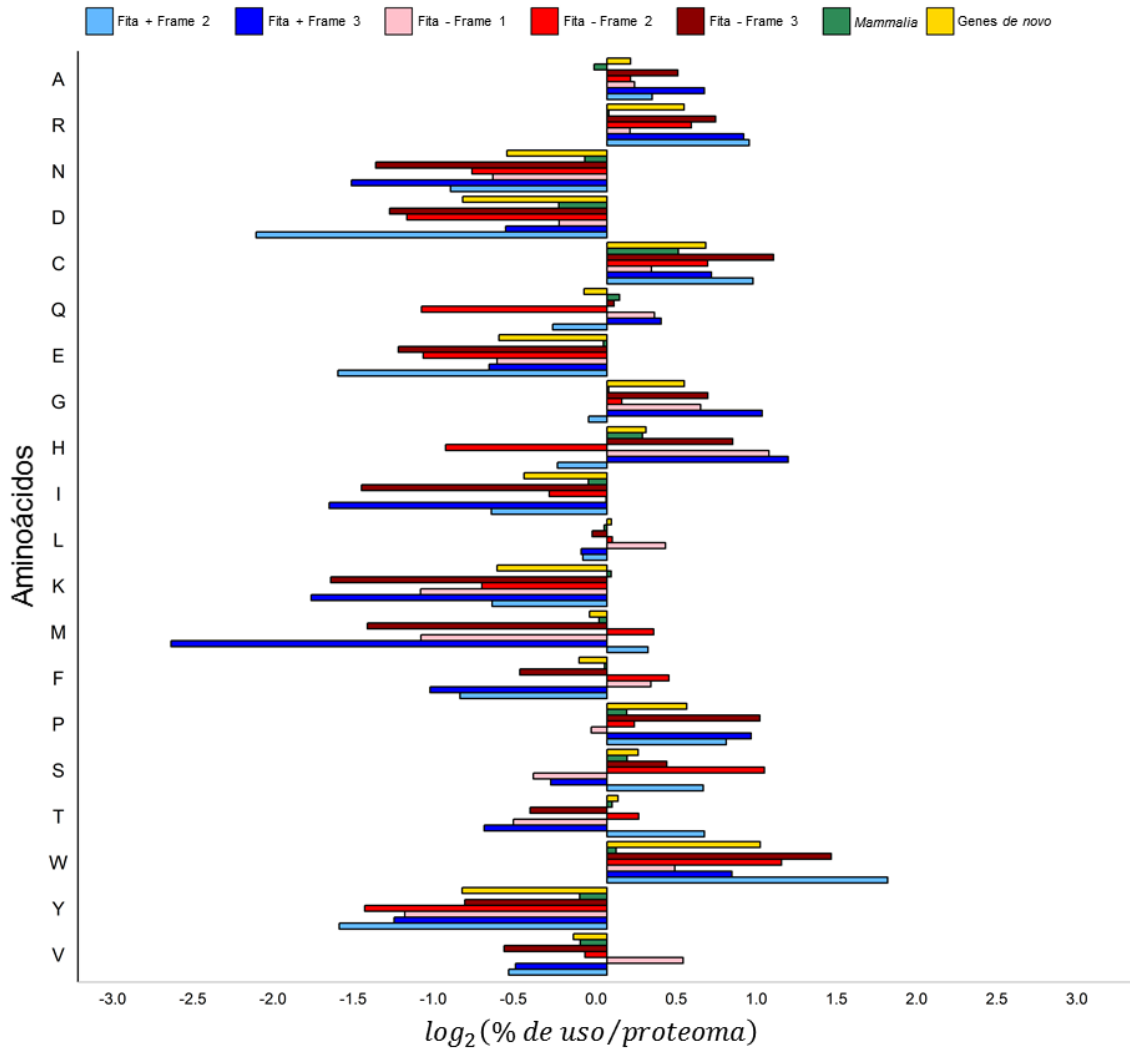


Figura 41 - Uso de aminoácidos em regiões não codificadoras. Barras mostram a razão da porcentagem de uso de aminoácidos para ORFs (>50aa) derivadas de *frameshifts* de sequências senso e antisenso em *Homo sapiens* versus a porcentagem de uso no proteoma completo. A razão está em escala logarítmica (\log_2). Para fins de comparação, é mostrado o uso de aminoácidos de proteínas com origem em mamíferos (em verde) e de proteínas com origem *de novo* (em amarelo).

5. Discussão

5.1. Limitações do PDB e predições de estrutura secundária

A utilização de estruturas definidas no PDB apresenta várias limitações (XIE; BOURNE, 2005), podendo conter fragmentos da proteína não resolvidos ou mutações na sequência (devido a artefatos, protocolos de cristalização ou à desordem natural). Estas pequenas variações podem ter efeitos em larga escala na estrutura e função das proteínas (BRUNK et al., 2016). Notamos vieses referentes à cobertura das estruturas do PDB (comparando com as sequências completas das proteínas) em relação ao domínio taxonômico pertencente. Em proteínas de procariotos observamos que as estruturas geralmente representam sequências completas, já em eucariotos, grande parte delas está descrita parcialmente (Figura 16). Isto se estende, em parte, quando analisamos as proteínas de acordo com suas idades evolutivas. Estruturas humanas com origem em organismos celulares apresentam em geral maior cobertura do que as sequências com origem em eucariotos e clados mais complexos (Figura 22). Tais vieses podem ser atribuídos devido às limitações associadas aos métodos utilizados para determinar as estruturas, tais como a propensão para resolução de proteínas globulares e solúveis dos métodos de cristalografia de raios X e as limitações quanto ao tamanho das proteínas dos métodos de ressonância nuclear magnética (XIE; BOURNE, 2005; YEE et al., 2005). Devido a essas razões, analisamos também dados derivados via métodos de predição de estrutura secundária. Dessa forma compensamos tanto o problema com estruturas de apenas fragmentos de proteínas, quanto o problema de amostragens insuficientes por espécies.

A acurácia dos métodos de predição de estrutura secundária tem se aproximado do limite teórico esperado (~88%) (YANG et al., 2016). Utilizamos para as principais análises o consenso de três métodos de predição de estrutura secundária (PSIPRED, DeepCNF e SSpro). Avaliamos a acurácia dos métodos e analisamos seus comportamentos para sequências com e sem estrutura definida. A acurácia só pode ser medida em casos que temos estruturas definidas. Logo, comparamos os métodos de predição para sequências do PDB. O SSpro se beneficia de informações de estrutura secundária experimentais encontradas em homólogos (MAGNAN; BALDI, 2014). Dessa forma seus resultados superam os demais métodos aqui utilizados (Tabela 6). Entretanto, a maioria das sequências de proteínas não possuem estruturas de homólogos depositada (YANG et al., 2016). Somado a isso, o próprio conjunto de proteínas do PDB não é representativo (WOOLFSON et al., 2015). Além disso, em casos de sequências com poucas informações

de homologia, a acurácia do SSpro fica reduzida (WANG et al., 2016b). Dessa forma, com a utilização do consenso dos três métodos, buscamos aprimorar a confiabilidade das predições em análises de proteomas completos.

5.2. Mudanças estruturais relacionadas à idade evolutiva dos genes

Diversas características estão relacionadas ao período de origem dos genes. Estudos anteriores mostraram que genes com origem mais recente são menores, apresentam menos íntrons, são expressos em menor intensidade (ELHAIK, 2005; WOLF et al., 2009), apresentam menos interações com outros genes (YIN et al., 2016; ZHANG et al., 2015), tendem a apresentar funções menos essenciais comparados a genes antigos (CHEN et al., 2012), possuem mais mutações para códons de terminação prematura (YANG et al., 2014), evoluem mais rápido (ALBÀ; CASTRESANA, 2005; WOLF et al., 2009) e experimentam pressão seletiva mais variável do que genes antigos (VISHNOI et al., 2010). Além disso, outros fatores também apresentam relação significativa com a idade evolutiva. Por exemplo, no estudo de Yin e colaboradores, uma análise de componentes principais foi realizada para identificar os fatores dominantes relacionados à idade dos genes (YIN et al., 2016). Foram considerados fatores como: composição gênica, tamanho do gene, pressão seletiva, nível de expressão, conectividade em interações proteína-proteína e metilação. Como resultados dessas análises, o conteúdo GC e as interações proteína-proteína foram os fatores dominantes.

Em nosso trabalho observamos a confirmação de estudos anteriores, expandimos outras análises com dados mais abrangentes e recentes e exploramos novos fatores com relação à macroevolução. Nossas análises têm como foco a questão da evolução estrutural de proteínas. Utilizamos para isso informações de estrutura secundária, seja ela proveniente de dados experimentais ou por métodos de predição. Nesse contexto, consideramos a estrutura secundária de cada resíduo definida em três estados: alfa-hélice (H), fita beta (E) e irregular (C) (ROST; SANDER, 1993a). Com isso, analisamos a porcentagem de resíduos em cada estado bem como os tamanhos dos segmentos em estados específicos para proteínas de diversos organismos. Ao investigar a estrutura secundária de proteínas com estrutura definida no PDB de diferentes domínios taxonômicos, observamos diferenças significativas entre eucariotos e procariotos. A composição de alfa-hélices e fitas beta de eucariotos é inferior à de procariotos, e suas proteínas apresentam maior diversidade em termos de composições (Figuras 13 e 14). Estes resultados são esperados, tendo em vista que boa parte das proteínas eucarióticas

apresentam estruturas diferentes das estruturas globulares típicas (ROST, 2002) e que regiões desordenadas ocorrem em maior abundância em proteomas de organismos mais complexos (ROMERO; OBRADOVIC; DUNKER, 2004; WARD et al., 2004).

Com o objetivo de obter uma visão abrangente acerca da evolução estrutural das proteínas, utilizamos métodos de predição de estrutura secundária para analisarmos a composição estrutural de proteínas de organismos em diversos ramos da árvore de vida. As mesmas tendências observadas com dados de estruturas definidas no PDB também foram observadas com predições de proteomas completos. A predição da estrutura secundária dos 66 proteomas de referência dados pelo consórcio *Quest for Orthologs* (QfO) nos mostrou diferenças na composição estrutural de diferentes grupos taxonômicos (Figuras 25 e 26). Por exemplo, dentro do grupo taxonômico *Excavata*, observamos semelhanças na composição estrutural dos organismos *Giardia intestinalis* e *Trichomonas vaginalis* com o observado em bactérias. Estes dois organismos possuem em comum apenas remanescentes mitocondriais (CARLTON et al., 2007; TOVAR et al., 2003). Já a *Leishmania major*, presente no mesmo grupo, que possui genoma mitocondrial (FLEGONTOV; STRELKOVA; KOLESNIKOV, 2006), apresenta semelhanças estruturais com *Dikarya* (Fungos). Essa característica pode estar envolvida nas diferenças encontradas. Mitocôndrias estão envolvidas em funções como fornecimento de energia, sinalização celular, regulação do metabolismo, controle do ciclo celular, desenvolvimento, respostas antivirais e morte celular (MCBRIDE; NEUSPIEL; WASIAK, 2006). Suas proteínas podem ser tanto de origem bacteriana quanto eucariótica. Em proteínas mitocondriais humanas, regiões desestruturadas são encontradas em torno de 10% das proteínas com origem bacteriana e em 20% das proteínas com origem eucariótica (ITO et al., 2012). Dessa forma, apesar dos repertórios proteicos variarem entre diferentes espécies, estudos futuros podem ajudar a elucidar o papel da mitocôndria nas composições de estrutura secundária observados no proteoma como um todo.

Classificamos também o proteoma humano de acordo com o período de origem estimado das proteínas e relacionamos esta informação com dados de estrutura secundária obtida por meio das predições. Com isso, observamos a redução em maior intensidade na composição de alfa-hélices e fitas beta nas transições para *Eukaryota* e *Opisthokonta* além do aumento na diversidade das composições encontradas em proteínas mais recentes (Figura 27). Em seguida, analisamos as diferenças estruturais entre regiões de domínios e extradomínios bem como diferenças no uso de aminoácidos para melhor entender os fatores envolvidos nessas mudanças.

A composição de estrutura secundária em regiões de domínios (definido aqui como, intradomínios) apresenta maior contribuição em termos de hélices e folhas quando comparadas com regiões extradomínios (LU et al., 2015). Essa diferença é mais evidente

em termos de fitas beta, onde regiões extradomínio apresentam conteúdo total inferior a 7% (Tabela 13). Regiões de domínio de proteínas com origem em organismos celulares (*Cellular_organisms*, *Euk+Bac* e *Euk_Archaea*) apresentam composições de estrutura secundária mista de hélices e folhas e abrangem quase a totalidade das sequências, enquanto que proteínas com origem a partir de *Eukaryota*, apresentam composições de maioria beta e menores porcentagens de cobertura (Figura 34 e 35). Somado a isso, há o aumento de sequências sem domínios mapeados nas proteínas mais recentes. Sabemos que regiões preditas como desordenadas são menos propensas a se dobrar em domínios globulares do que regiões não desordenadas (LAM et al., 2016). Logo, tais proteínas podem indicar ou proteínas desordenadas não cristalizáveis ou apresentarem conformações ainda não definidas.

Quanto aos aminoácidos, sabemos que seu uso varia entre espécies e tem papel importante na evolução das proteínas. Estudos anteriores mostraram haver uma tendência para ganho e perda de aminoácidos ao longo da evolução (JORDAN et al., 2005; LIU et al., 2015). Além disso, diferentes aminoácidos estão relacionados a regiões ordenadas e desordenadas. Aminoácidos podem ser classificados como promotores de ordem: cisteína (C), triptofano (W), isoleucina (I), tirosina (Y), fenilalanina (F), leucina (L), histidina (H), valina (V) e asparagina (N); ou promotores de desordem: prolina (P), ácido glutâmico (E), serina (S), glutamina (Q) e lisina (K) (THEILLET et al., 2013). As diferenças no uso de aminoácidos entre proteínas com origem antiga e recente, observadas em nossas análises, apresentaram algumas relações com o esperado relativo aos determinantes de ordem e desordem. Observamos que seis aminoácidos promotores de ordem (Y, I, F, V, L, N) e dois promotores de desordem (K, E), tem uso relativo superior em proteínas antigas (Tabela 12). Proteínas recentes apresentaram maior uso de C, P, S e H, enquanto proteínas antigas apresentam maior uso de I, D, V, A, N e K (diferença percentual superior a 12%, Tabela 12). O uso especialmente elevado de cisteínas em proteínas recentes é um resultado já esperado (JORDAN et al., 2005; LIU et al., 2015; MISETA; CSUTORA, 2000), entretanto é inesperado em termos da redução de estruturas secundárias observado, já que as cisteínas têm maior ocorrência em estruturas ordenadas (Figura 42). Entretanto, ao analisarmos o uso de aminoácidos de regiões de domínios e extradomínios encontramos relações que podem explicar os resultados observados. Regiões intradomínio, quando comparadas com regiões extradomínio, utilizam mais aminoácidos promotores de ordem, e apresentam maior conteúdo de estrutura secundária, principalmente de fitas beta (Figura 36 e 37). Comparando as preferências de uso de regiões intradomínio e extradomínio em proteínas com diferentes idades evolutivas (Figura 38), notamos que as diferenças se acentuam a partir de *Opisthokonta*. Tanto o aumento de cisteínas (promotor de ordem) em regiões intradomínio, como o aumento de prolinas (promotor de desordem) em regiões

extradomínio, sugerem que proteínas recentes são um misto de proteínas desordenadas e proteínas ordenadas. Fatores determinantes de proteínas estruturadas e desestruturadas são diferentes, regiões desordenadas apresentam alta compactação (em comparação com estruturas desnaturadas), cujo efeito está altamente correlacionado com a carga de rede e o conteúdo de prolinas (MARSH; FORMAN-KAY, 2010). Além disso, estudos sugerem que a inclusão da cisteína pode ter sido tardia na evolução, em comparação com os outros aminoácidos, e que a sua capacidade única de moldar estruturas de proteínas é provavelmente mais utilizada em organismos complexos (MISETA; CSUTORA, 2000), explicando as diferenças observadas a partir de *Opisthokonta*.

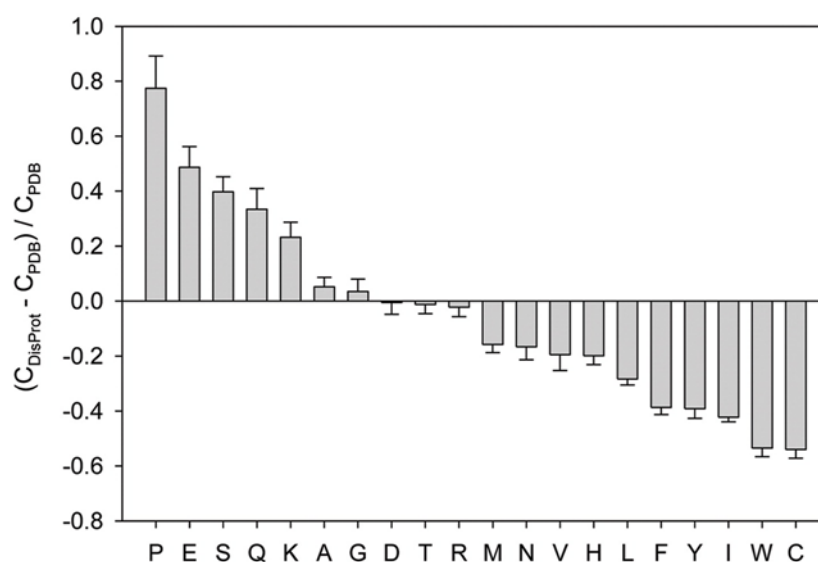


Figura 42 - Determinantes de aminoácidos que definem diferenças estruturais e funcionais entre as proteínas ordenadas e intrinsecamente desordenadas. Razão da diferença na composição de aminoácidos entre proteínas intrinsecamente desordenadas (proveniente do banco de dados DisProt) e proteínas completamente ordenadas (PDB). A diferença é calculada como $(C_{DisProt} - C_{PDB}) / C_{PDB}$, onde $C_{DisProt}$ é o conteúdo de aminoácidos na base de dados DisProt e C_{PDB} é o conteúdo de aminoácidos proteínas ordenadas do PDB. Os aminoácidos estão classificados em ordem decrescente de acordo com o seu potencial de promoção da desordem. Adaptado de (THEILLET et al., 2013).

O fato das diferenças mais acentuadas terem sido observadas em *Opisthokonta* nos leva a investigar os fatores macroevolutivos envolvidos nesse clado. A transição para multicelularidade ocorreu em diversos momentos e diferentes linhagens, dentre elas em *Opisthokonta*. As diferenças encontradas nas estruturas secundárias nesse período podem estar relacionadas a essas mudanças. Funções de ligação a sítios de DNA e atividades de fatores de transcrição observadas em *Opisthokonta* são também encontradas em proteínas com alto conteúdo de resíduos irregulares (Figuras 31 e 33). Regiões desordenadas são encontradas em abundância em proteínas com essas funções, caudas terminais e ligadores flexíveis, têm papel importante na afinidade e especificidade de processos de

reconhecimento de DNA (VUZMAN; LEVY, 2012). Além disso, as interações com o DNA podem promover a ordenação de regiões desordenadas (MARASCO; SCOGNAMIGLIO, 2015), cerca de 70% das proteínas humanas de ligação com DNA são compostas de caudas desordenadas, tendo papel importante na estabilidade de interações com DNA em fatores de transcrição e endonucleases (MARASCO; SCOGNAMIGLIO, 2015; POLETTO et al., 2013; TÓTH-PETRÓCZY et al., 2009). Fatores de transcrição (FT) também apresentam a prevalência de regiões desordenadas, com maior incidência em proteínas eucarióticas (LIU et al., 2006). A transição da vida unicelular para multicelular também tem papel importante na evolução dos FT, ocorrendo de forma convergente em pelo menos 26 grupos independentes (GROSBERG; STRATHMANN, 2007; PARFREY; LAHR, 2013). A multicelularidade complexa está associada ao enriquecimento do conjunto de ferramentas de FT (tanto em termos de abundância quanto de inovação) em linhagens com desenvolvimento embrionário complexo: plantas e animais (SEBÉ-PEDRÓS; DE MENDOZA, 2015). Dessa forma, o enriquecimento de proteínas relacionadas a essas funções em *Opisthokonta* poderiam explicar a redução estrutural observada nas estruturas secundárias dessas proteínas.

5.3. Origem *de novo* tem suporte com dados de estrutura secundária

Fatores estruturais podem também auxiliar o entendimento relativo à origem de novos genes. Analisamos as estruturas de genes *de novo* e notamos características que seguem o padrão esperado em proteínas recentes, tanto em termos de estrutura secundária (pouco conteúdo de alfa-hélices e fitas beta) como no uso de aminoácidos (Figura 39 e 41). Ao inferirmos as estruturas de proteínas originadas a partir de regiões não codificadoras (derivadas de *frameshifts* e antisense), notamos que tais regiões poderiam suportar conteúdos estruturais tais como as observadas em proteínas conhecidas (Figura 40). As composições estruturais dessas regiões seriam altamente irregulares, diferindo da composição esperada em sequências consideradas “falsas” (como as sequências permutadas analisadas neste trabalho). Sabemos que mesmo sequências randômicas apresentam menos “irregularidade” do que sequências naturais (YU et al., 2016). Além disso, o fato de proteínas intrinsecamente irregulares serem observadas em maior ocorrência em organismos mais complexos aponta para uma evolução direcionada (DEFORTE; UVERSKY, 2016). No entanto, os mecanismos de origem de genes *de novo* sugerem que genes funcionais podem ser obtidos a partir de ORFs já existentes com a simples aquisição de elementos regulatórios (SCHLÖTTERER, 2015). De fato, nossos

resultados apontam que apenas uma mutação resultando na expressão de uma ORF preexistente poderia resultar em uma proteína com conteúdo de estrutura similar ao encontrado em proteínas naturais.

6. Conclusão

No presente trabalho, utilizamos abordagens de bioinformática para investigar questões evolutivas com respeito a mudanças estruturais em proteínas e aspectos que permeiam a origem de novos genes. Utilizando dados de estrutura secundária, observamos diferenças na composição de alfa-hélices e fitas beta em proteínas de diferentes organismos. Notamos que organismos mais complexos, tendem a apresentar menor conteúdo estruturado, e que certos grupos taxonômicos específicos, apresentam grande variação na composição de elementos de estrutura secundária, que podem estar relacionados a fatores evolutivos. Além disso, observamos mudanças na composição de elementos de estrutura secundária em proteínas humanas ao longo da evolução. Proteínas mais recentes apresentam menor conteúdo de estrutura secundária e maior diversidade de composições, enquanto que proteínas antigas têm geralmente composições semelhantes. As mudanças mais acentuadas foram observadas em proteínas com origem a partir dos clados *Eukaryota* e *Opisthokonta*. Análises funcionais sugerem que estas proteínas estão relacionadas à funções de ligação ao DNA e a fatores de transcrição, nas quais é esperado que haja maior conteúdo desordenado. Questões macroevolutivas, como a origem da multicelularidade, podem estar associadas a essas mudanças.

Ao analisarmos fatores que podem afetar a composição estrutural das proteínas, vimos que diferenças no uso de aminoácidos em regiões de domínios e extradomínios podem explicar as mudanças observadas. Regiões de domínios apresentam maior conteúdo de alfa-hélices e fitas beta do que extradomínios. Proteínas antigas, com origem em organismos celulares, geralmente apresentam domínios que cobrem grande parte das sequências. Esses domínios, em sua maioria, apresentam composições de estrutura secundária mista de hélices e folhas. A partir de *Opisthokonta*, observamos uma mudança na composição dos domínios, que passam a ser compostos majoritariamente por folhas beta. Somado a isso, os domínios passam a ter menor cobertura de sequência e vemos um aumento de proteínas sem domínios. Além disso, notamos uma acentuação na diferença no uso de aminoácidos entre regiões intradomínios e extradomínios em proteínas recentes. Aminoácidos promotores de ordem (mais encontrados em intradomínios) apresentam ganhos em proteínas recentes, enquanto que aminoácidos promotores de desordem (mais encontrados em extradomínios) sofrem perdas. Essas mudanças estão relacionadas à maior diversidade de estruturas observadas em proteínas recentes, tanto em termos de proteínas ordenadas como desordenadas.

Por fim, investigamos se nossos dados suportam a teoria de origem de genes *de novo*. Analisamos a hipótese de genes se originarem a partir da simples tradução de ORFs

presentes em regiões não codificadoras. Predições realizadas em sequências de genes humanos, com mudanças na fase leitura e em sequências antisense, mostraram que estas possíveis proteínas poderiam obter a estrutura esperada em proteínas naturais. Observamos também que o viés de redução na composição de elementos de estrutura secundária, também ocorre em proteínas com origem *de novo*. Dessa forma, é possível que novos genes oriundos de regiões não codificadoras apresentem também uma estrutura mais desordenada com poucas conformações de alfa-hélices e fitas beta.

7. Produção científica

Artigos completos publicados em periódicos

1. LOPES, K. P.; CAMPOS-LABORIE, F. J.; **VIALLE, R. A.**; DE LAS RIVAS, J.; ORTEGA, J. M.. Evolutionary hallmarks of the human proteome: chasing the age and coregulation of protein-coding genes. BMC Genomics, 2016. Vol 17, Sup 8. DOI:10.1186/s12864-016-3062-y
2. *VELLOSO H., ***VIALLE R.A.**, ORTEGA J.M.. BOWS (bioinformatics open web services) to centralize bioinformatics tools in web services. BMC Res Notes. 2015;8: 206. DOI:10.1186/s13104-015-1190-0

Resumos publicados em anais de congressos

1. **VIALLE, R. A.**; ORTEGA, J. M.. Secondary structure changes according to evolutionary age. In: X-Meeting, 2016, Belo Horizonte. Abstract Book, 2016.
2. LOPES, K. P.; **VIALLE, R. A.**; ORTEGA, J. M.. Transcriptome meta-analysis reveals the human organs evolution. In: X-Meeting, 2016, Belo Horizonte. Abstract Book, 2016.
3. LAUX, M.; **VIALLE, R. A.**; ORTEGA, J. M.; GIANI, A.. Within and between gene variants: tracking for potential targets for populational linkage according to the metagenomic profile in a changing freshwater environment. In: X-Meeting, 2016, Belo Horizonte. Abstract Book, 2016.
4. LOPES, K. P.; CAMPOS-LABORIE, F. J.; **VIALLE, R. A.**; ORTEGA, J. M.; DE LAS RIVAS, J.. Global coexpression analysis of human protein coding genes. In: X-Meeting, 2016, Belo Horizonte. Abstract Book, 2016.
5. LOPES, K. P.; **VIALLE, R. A.**; ORTEGA, J. M.. Origin of genes obtained by transcriptomic data compared to KEGG Functional Hierarchies. In: X-Meeting 2015, 2015, São Paulo. Abstract Book, 2015.
6. LOPES, K. P.; **VIALLE, R. A.**; COSTA, V. R. M. ; ORTEGA, J. M.. Computational approach for amino acids visualization according to codon usage. In: ISCB-Latin America, 2014, Belo Horizonte. Abstract Book, 2014.
7. LAUX, M.; GIANI, A.; ORTEGA, J. M.; **VIALLE, R. A.**; FERNANDES, G. R.. Functional profile according to contribution and distribution of pathways in a cyanobacteria dominated reservoir over an annual scale. In: Metagenomics Workshop - III International Workshop on Environmental Microbiology, 2014, Belo Horizonte. Abstract Book, 2014.
8. COSTA, V. R. M.; **VIALLE, R. A.**; ORTEGA, J. M.. The origin of B-cell antigen receptor pathway. In: X-Meeting 2014, 2014, Belo Horizonte. Abstract Book, 2014.
9. **VIALLE, R. A.**; VELLOSO, HENRIQUE; ORTEGA, J. M.. Simple java command lines that request Bioinformatics Open Web Services (BOWS). In: X-Meeting 2014, 2014, Belo Horizonte. Abstract Book, 2014.
10. **VIALLE, R. A.**; ORTEGA, J. M.. Evidences of distinct wave of gene origins in recent evolutionary events. In: X-Meeting 2013, 2013, Recife. Abstract Book, 2013.
11. LANGOWSKI, E.; **VIALLE, R. A.**; ORTEGA, J. M.; BRAWERMAN, A.. Porting a desktop bioinformatics tool to the web - The benefits of using web services. In: X-Meeting 2013, 2013, Recife. Abstract Book, 2013.

Referências

- ABRUSÁN, G. Integration of new genes into cellular networks, and their structural maturation. *Genetics*, v. 195, n. 4, p. 1407–1417, dez. 2013.
- ALBÀ, M. M.; CASTRESANA, J. Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular biology and evolution*, v. 22, n. 3, p. 598–606, mar. 2005.
- ALBRECHT, M. et al. Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein engineering*, v. 16, n. 7, p. 459–462, jul. 2003.
- ALTENHOFF, A. M. et al. Standardized benchmarking in the quest for orthologs. *Nature methods*, v. 13, n. 5, p. 425–430, maio 2016.
- ALTENHOFF, A. M.; DESSIMOZ, C. Inferring orthology and paralogy. *Methods in molecular biology*, v. 855, p. 259–279, 2012.
- ALTSCHUL, S. F. et al. Basic local alignment search tool. *Journal of molecular biology*, v. 215, n. 3, p. 403–410, 5 out. 1990.
- ALTSCHUL, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, v. 25, n. 17, p. 3389–3402, 1 set. 1997.
- AMBROGELLY, A.; PALIOURA, S.; SÖLL, D. Natural expansion of the genetic code. *Nature chemical biology*, v. 3, n. 1, p. 29–35, 2007.
- ARNOLD, G. E. et al. Use of conditional probabilities for determining relationships between amino acid sequence and protein secondary structure. *Proteins*, v. 12, n. 4, p. 382–399, abr. 1992.
- AYDIN, Z.; ALTUNBASAK, Y.; BORODOVSKY, M. Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. *BMC bioinformatics*, v. 7, p. 178, 30 mar. 2006.
- BERG, J. M.; TYMOCZKO, J. L.; STRYER, L. *Biochemistry, Fifth Edition*. [s.l.] W. H. Freeman, 2002.
- BETRÁN, E.; LONG, M. Expansion of genome coding regions by acquisition of new genes. *Genetica*, v. 115, n. 1, p. 65–80, maio 2002.
- BLAIR HEDGES, S.; KUMAR, S. *The Timetree of Life*. [s.l.] OUP Oxford, 2009.

- BOECKMANN, B. et al. Quest for Orthologs Entails Quest for Tree of Life: In Search of the Gene Stream. *Genome biology and evolution*, v. 7, n. 7, p. 1988–1999, jul. 2015.
- BOHR, H. et al. Protein secondary structure and homology by neural networks. The alpha-helices in rhodopsin. *FEBS letters*, v. 241, n. 1-2, p. 223–228, 5 dez. 1988.
- BOKMA, F. Evolution as a Largely Autonomous Process. In: *Interdisciplinary Evolution Research*. [s.l.: s.n.]. p. 87–112.
- BRUNK, E. et al. Systems biology of the structural proteome. *BMC systems biology*, v. 10, p. 26, 11 mar. 2016.
- BUDD, A. Introduction to Genome Biology: Features, Processes, and Structures. In: *Methods in Molecular Biology*. [s.l.: s.n.]. p. 3–49.
- CAI, J. J.; PETROV, D. A. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome biology and evolution*, v. 2, p. 393–409, 12 jul. 2010.
- CAMPBELL-PLATT, G. *Food Science and Technology*. [s.l.] John Wiley & Sons, 2011.
- CARLTON, J. M. et al. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science*, v. 315, n. 5809, p. 207–212, 12 jan. 2007.
- CARVUNIS, A.-R. et al. Proto-genes and de novo gene birth. *Nature*, v. 487, n. 7407, p. 370–374, 19 jul. 2012.
- CHEN, W.-H. et al. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Molecular biology and evolution*, v. 29, n. 7, p. 1703–1706, jul. 2012.
- CHEN, X.; ZHANG, J. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS computational biology*, v. 8, n. 11, p. e1002784, 29 nov. 2012.
- CHOU, P. Y.; FASMAN, G. D. Prediction of protein conformation. *Biochemistry*, v. 13, n. 2, p. 222–245, 1974.
- CLOOS, P. A. C.; CHRISTGAU, S. Non-enzymatic covalent modifications of proteins: mechanisms, physiological consequences and clinical applications. *Matrix biology: journal of the International Society for Matrix Biology*, v. 21, n. 1, p. 39–52, jan. 2002.

COORDINATORS, N. R. Database Resources of the National Center for Biotechnology Information. *Nucleic acids research*, v. 45, n. D1, p. D12–D17, 2016.

COSTANTINI, S.; COLONNA, G.; FACCHIANO, A. M. Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochemical and biophysical research communications*, v. 342, n. 2, p. 441–451, 7 abr. 2006.

CRICK, F. Central Dogma of Molecular Biology. *Nature*, v. 227, n. 5258, p. 561–563, 1970.

DEFORTE, S.; UVERSKY, V. N. Order, Disorder, and Everything in Between. *Molecules*, v. 21, n. 8, 19 ago. 2016.

DEMUTH, D. Gregory A. Petsko and Dagmar Ringe, *Primers in Biology: Protein Structure and Function*, New Science Press, Ltd. (2004) 195 pages. *Reproductive toxicology*, v. 19, n. 4, p. 565–566, 2005.

DOMAZET-LOŠO, T. et al. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Molecular biology and evolution*, p. msw284, 2017.

DOMAZET-LOSO, T.; BRAJKOVIĆ, J.; TAUTZ, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in genetics: TIG*, v. 23, n. 11, p. 533–539, nov. 2007.

DOMAZET-LOSO, T.; TAUTZ, D. An ancient evolutionary origin of genes associated with human genetic diseases. *Molecular biology and evolution*, v. 25, n. 12, p. 2699–2707, dez. 2008.

DOMAZET-LOSO, T.; TAUTZ, D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC biology*, v. 8, p. 66, 21 maio 2010.

DOMAZET-LOŠO, T.; TAUTZ, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, v. 468, n. 7325, p. 815–818, 9 dez. 2010.

DOR, O.; ZHOU, Y. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins*, v. 66, n. 4, p. 838–845, 1 mar. 2007.

DUFTON, M. J. Genetic code synonym quotas and amino acid complexity: cutting the cost of proteins? *Journal of theoretical biology*, v. 187, n. 2, p. 165–173, 21 jul. 1997.

DYSON, H. J.; JANE DYSON, H.; WRIGHT, P. E. Intrinsically unstructured proteins and their functions. *Nature reviews. Molecular cell biology*, v. 6, n. 3, p. 197–208, 2005.

EDEN, E. et al. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics*, v. 10, p. 48, 3 fev. 2009.

ELHAIK, E. The “Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes” Is an Artifact of Increased Genetic Distance with Rate of Evolution and Time of Divergence. *Molecular biology and evolution*, v. 23, n. 1, p. 1–3, 2005.

FARLEY, A. R.; LINK, A. J. Chapter 40 Identification and Quantification of Protein Posttranslational Modifications. In: *Methods in Enzymology*. [s.l: s.n.]. p. 725–763.

FLEGONTOV, P. N.; STRELKOVA, M. V.; KOLESNIKOV, A. A. The *Leishmania major* maxicircle divergent region is variable in different isolates and cell types. *Molecular and biochemical parasitology*, v. 146, n. 2, p. 173–179, abr. 2006.

GABALDÓN, T.; KOONIN, E. V. Functional and evolutionary implications of gene orthology. *Nature reviews. Genetics*, v. 14, n. 5, p. 360–366, maio 2013.

GARNIER, J.; OSGUTHORPE, D. J.; ROBSON, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of molecular biology*, v. 120, n. 1, p. 97–120, 25 mar. 1978.

GROSBURG, R. K.; STRATHMANN, R. R. The Evolution of Multicellularity: A Minor Major Transition? *Annual review of ecology, evolution, and systematics*, v. 38, n. 1, p. 621–654, 2007.

GUTTERIDGE, A.; THORNTON, J. M. Understanding nature’s catalytic toolkit. *Trends in biochemical sciences*, v. 30, n. 11, p. 622–629, 2005.

HAYASHI, Y. et al. Can an arbitrary sequence evolve towards acquiring a biological function? *Journal of molecular evolution*, v. 56, n. 2, p. 162–168, fev. 2003.

HEFFERNAN, R. et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific reports*, v. 5, p. 11476, 22 jun. 2015.

HEMMRICH, G. et al. Molecular signatures of the three stem cell lineages in hydra and the emergence of stem cell function at the base of multicellularity. *Molecular biology and evolution*, v. 29, n. 11, p. 3267–3280, nov. 2012.

HOLLEY, L. H.; KARPLUS, M. Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences*, v. 86, n. 1, p. 152–156, 1989.

HUANG, C. R. L.; BURNS, K. H.; BOEKE, J. D. Active transposition in genomes. *Annual review of genetics*, v. 46, p. 651–675, 2012.

HUA, S.; SUN, Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of molecular biology*, v. 308, n. 2, p. 397–407, 27 abr. 2001.

ITO, M. et al. Intrinsically disordered proteins in human mitochondria. *Genes to cells: devoted to molecular & cellular mechanisms*, v. 17, n. 10, p. 817–825, out. 2012.

JENSEN, R. A. Orthologs and paralogs - we need to get it right. *Genome biology*, v. 2, n. 8, p. INTERACTIONS1002, 3 ago. 2001.

JIANG, N. et al. Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, v. 431, n. 7008, p. 569–573, 30 set. 2004.

JONES, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, v. 292, n. 2, p. 195–202, 1999.

JORDAN, I. K. et al. A universal trend of amino acid gain and loss in protein evolution. *Nature*, v. 433, n. 7026, p. 633–638, 10 fev. 2005.

KABAT, E. A.; WU, T. T. The Influence of Nearest-Neighbor Amino Acids on the Conformation of the Middle Amino Acid in Proteins: Comparison of Predicted and Experimental Determination of α -Sheets in Concanavalin A. *Proceedings of the National Academy of Sciences*, v. 70, n. 5, p. 1473–1477, 1973.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, v. 22, n. 12, p. 2577–2637, dez. 1983.

KAESSMANN, H. Origins, evolution, and phenotypic impact of new genes. *Genome research*, v. 20, n. 10, p. 1313–1326, out. 2010.

KAESSMANN, H.; VINCKENBOSCH, N.; LONG, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nature reviews. Genetics*, v. 10, n. 1, p. 19–31, jan. 2009.

KAHL, G. *The Dictionary of Genomics, Transcriptomics and Proteomics*. [s.l.: s.n.].

KNOWLES, D. G.; MCLYSAGHT, A. Recent de novo origin of human protein-coding genes. *Genome research*, v. 19, n. 10, p. 1752–1759, out. 2009.

KOONIN, E. V. Orthologs, paralogs, and evolutionary genomics. *Annual review of genetics*, v. 39, p. 309–338, 2005.

LAM, S. D. et al. Gene3D: expanding the utility of domain assignments. *Nucleic acids research*, v. 44, n. D1, p. D404–9, 4 jan. 2016.

LEVINE, M. T. et al. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences of the United States of America*, v. 103, n. 26, p. 9935–9939, 27 jun. 2006.

LI, C.-Y. et al. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS computational biology*, v. 6, n. 3, p. e1000734, 26 mar. 2010.

LIEBESKIND, B. J.; MCWHITE, C. D.; MARCOTTE, E. M. Towards Consensus Gene Ages. *Genome biology and evolution*, v. 8, n. 6, p. 1812–1823, 27 jun. 2016.

LIU, H. et al. Relationship between amino acid usage and amino acid evolution in primates. *Gene*, v. 557, n. 2, p. 182–187, 25 fev. 2015.

LIU, J. et al. Intrinsic Disorder in Transcription Factors†. *Biochemistry*, v. 45, n. 22, p. 6873–6888, 2006.

LIU, Y. et al. Comparison of probabilistic combination methods for protein secondary structure prediction. *Bioinformatics*, v. 20, n. 17, p. 3099–3107, 22 nov. 2004.

LOH, P. G.; SONG, H. Structural and mechanistic insights into translation termination. *Current opinion in structural biology*, v. 20, n. 1, p. 98–103, fev. 2010.

LONG, M. et al. The origin of new genes: glimpses from the young and old. *Nature reviews. Genetics*, v. 4, n. 11, p. 865–875, 2003.

LONG, M. et al. New gene evolution: little did we know. *Annual review of genetics*, v. 47, p. 307–333, 13 set. 2013.

LU, H.-C. et al. Anatomy of protein disorder, flexibility and disease-related mutations. *Frontiers in molecular biosciences*, v. 2, p. 47, 12 ago. 2015.

MADEJ, T. et al. MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic acids research*, v. 42, n. Database issue, p. D297–303, jan. 2014.

MAGNAN, C. N.; BALDI, P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, v. 30, n. 18, p. 2592–2597, 15 set. 2014.

MARASCO, D.; SCOGNAMIGLIO, P. L. Identification of inhibitors of biological interactions involving intrinsically disordered proteins. *International journal of molecular sciences*, v. 16, n. 4, p. 7394–7412, 2 abr. 2015.

MARSH, J. A.; FORMAN-KAY, J. D. Sequence determinants of compaction in intrinsically disordered proteins. *Biophysical journal*, v. 98, n. 10, p. 2383–2390, 19 maio 2010.

MCBRIDE, H. M.; NEUSPIEL, M.; WASIAK, S. Mitochondria: More Than Just a Powerhouse. *Current biology: CB*, v. 16, n. 14, p. R551–R560, 2006.

MCLYSAGHT, A.; GUERZONI, D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, v. 370, n. 1678, p. 20140332, 26 set. 2015.

MCLYSAGHT, A.; HURST, L. D. Open questions in the study of de novo genes: what, how and why. *Nature reviews. Genetics*, v. 17, n. 9, p. 567–578, set. 2016.

MISETA, A.; CSUTORA, P. Relationship between the occurrence of cysteine in proteins and the complexity of organisms. *Molecular biology and evolution*, v. 17, n. 8, p. 1232–1239, ago. 2000.

MITCHELL, E. M. et al. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *Journal of molecular biology*, v. 212, n. 1, p. 151–166, 5 mar. 1990.

MORGANTE, M. et al. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature genetics*, v. 37, n. 9, p. 997–1002, set. 2005.

MOYERS, B. A.; ZHANG, J. Phylostratigraphic bias creates spurious patterns of genome evolution. *Molecular biology and evolution*, v. 32, n. 1, p. 258–267, jan. 2015.

MOYERS, B. A.; ZHANG, J. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution. *Molecular biology and evolution*, v. 33, n. 5, p. 1245–1256, maio 2016.

MUGGLETON, S.; KING, R. D.; STERNBERG, M. J. Protein secondary structure prediction using logic-based machine learning. *Protein engineering*, v. 5, n. 7, p. 647–657, out. 1992.

NEME, R.; TAUTZ, D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC genomics*, v. 14, p. 117, 21 fev. 2013.

OHNO, S. *Evolution by Gene Duplication*. [s.l.: s.n.].

PARFREY, L. W.; LAHR, D. J. G. Multicellularity arose several times in the evolution of eukaryotes (response to DOI 10.1002/bies.201100187). *BioEssays: news and reviews in molecular, cellular and developmental biology*, v. 35, n. 4, p. 339–347, abr. 2013.

PAULING, L.; COREY, R. B.; BRANSON, H. R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, v. 37, n. 4, p. 205–211, abr. 1951.

PETSKO, G. A.; RINGE, D. *Protein Structure and Function*. [s.l.] New Science Press, 2004.

PHIZICKY, E. M.; HOPPER, A. K. tRNA biology charges to the front. *Genes & development*, v. 24, n. 17, p. 1832–1860, 1 set. 2010.

POLETTI, M. et al. Role of the unstructured N-terminal domain of the hAPE1 (human apurinic/aprimidinic endonuclease 1) in the modulation of its interaction with nucleic acids and NPM1 (nucleophosmin). *Biochemical Journal*, v. 452, n. 3, p. 545–557, 15 jun. 2013.

PRAT, Y. et al. Codon usage is associated with the evolutionary age of genes in metazoan genomes. *BMC evolutionary biology*, v. 9, p. 285, 8 dez. 2009.

PRINCE, V. E.; PICKETT, F. B. Splitting pairs: the diverging fates of duplicated genes. *Nature reviews. Genetics*, v. 3, n. 11, p. 827–837, nov. 2002.

RANZ, J. M. et al. Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS biology*, v. 5, n. 6, p. e152, jun. 2007.

REINHARDT, J. A. et al. De novo ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS genetics*, v. 9, n. 10, p. e1003860, 17 out. 2013.

ROMERO, P.; OBRADOVIC, Z.; DUNKER, A. K. Natively disordered proteins: functions and predictions. *Applied bioinformatics*, v. 3, n. 2-3, p. 105–113, 2004.

ROST, B. Review: Protein Secondary Structure Prediction Continues to Rise. *Journal of structural biology*, v. 134, n. 2-3, p. 204–218, 2001.

ROST, B. Did evolution leap to create the protein universe? *Current opinion in structural biology*, v. 12, n. 3, p. 409–416, jun. 2002.

ROST, B.; SANDER, C. Prediction of protein secondary structure at better than 70% accuracy. *Journal of molecular biology*, v. 232, n. 2, p. 584–599, 20 jul. 1993a.

ROST, B.; SANDER, C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, v. 90, n. 16, p. 7558–7562, 15 ago. 1993b.

RUIZ-ORERA, J. et al. Long non-coding RNAs as a source of new peptides. *eLife*, v. 3, p. e03523, 16 set. 2014.

SAMUSIK, N. et al. PBOV1 Is a Human De Novo Gene with Tumor-Specific Expression That Is Associated with a Positive Clinical Outcome of Cancer. *PloS one*, v. 8, n. 2, p. e56162, 2013.

SCHLÖTTERER, C. Genes from scratch – the evolutionary fate of de novo genes. *Trends in genetics: TIG*, v. 31, n. 4, p. 215–219, 2015.

SEBÉ-PEDRÓS, A.; DE MENDOZA, A. Transcription Factors and the Origin of Animal Multicellularity. In: *Advances in Marine Genomics*. [s.l.: s.n.]. p. 379–394.

SESTAK, M. S. et al. Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. *Frontiers in zoology*, v. 10, n. 1, p. 18, 12 abr. 2013.

ŠESTAK, M. S.; DOMAZET-LOŠO, T. Phylostratigraphic profiles in zebrafish uncover chordate origins of the vertebrate brain. *Molecular biology and evolution*, v. 32, n. 2, p. 299–312, fev. 2015.

SILLITOE, I. et al. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic acids research*, v. 43, n. Database issue, p. D376–81, jan. 2015.

SIMAKOV, O.; LARSSON, T. A.; ARENDT, D. Linking micro- and macro-evolution at the cell type level: a view from the lophotrochozoan *Platynereis dumerilii*. *Briefings in functional genomics*, v. 12, n. 5, p. 430–439, set. 2013.

SONNHAMMER, E. L. L. et al. Big data and other challenges in the quest for orthologs. *Bioinformatics*, v. 30, n. 21, p. 2993–2998, 1 nov. 2014.

SUENAGA, Y. et al. NCYM, a Cis-Antisense Gene of MYCN, Encodes a De Novo Evolved Protein That Inhibits GSK3 β Resulting in the Stabilization of MYCN in Human Neuroblastomas. *PLoS genetics*, v. 10, n. 1, p. e1003996, 2014.

SUPEK, F. et al. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one*, v. 6, n. 7, p. e21800, 18 jul. 2011.

SUZEK, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, v. 31, n. 6, p. 926–932, 15 mar. 2015.

THEILLET, F.-X. et al. The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins. *Intrinsically Disordered Proteins*, v. 1, n. 1, p. e24360, 29 jan. 2013.

TOLL-RIERA, M. et al. Origin of primate orphan genes: a comparative genomics approach. *Molecular biology and evolution*, v. 26, n. 3, p. 603–612, mar. 2009.

TÓTH-PETRÓCZY, A. et al. Disordered tails of homeodomains facilitate DNA recognition by providing a trade-off between folding and specific binding. *Journal of the American Chemical Society*, v. 131, n. 42, p. 15084–15085, 28 out. 2009.

TOUW, W. G. et al. A series of PDB-related databanks for everyday needs. *Nucleic acids research*, v. 43, n. Database issue, p. D364–8, jan. 2015.

TOVAR, J. et al. Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. *Nature*, v. 426, n. 6963, p. 172–176, 13 nov. 2003.

UVERSKY, V. N. Introduction to intrinsically disordered proteins (IDPs). *Chemical reviews*, v. 114, n. 13, p. 6557–6560, 9 jul. 2014.

VELANKAR, S. et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic acids research*, v. 41, n. Database issue, p. D483–9, jan. 2013.

VINCKENBOSCH, N.; DUPANLOUP, I.; KAESSMANN, H. Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, v. 103, n. 9, p. 3220–3225, 28 fev. 2006.

VISHNOI, A. et al. Young proteins experience more variable selection pressures than old proteins. *Genome research*, v. 20, n. 11, p. 1574–1581, nov. 2010.

VUZMAN, D.; LEVY, Y. Intrinsically disordered regions as affinity tuners in protein-DNA interactions. *Molecular bioSystems*, v. 8, n. 1, p. 47–57, jan. 2012.

WANG, J. et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, v. 516, n. 7531, p. 405–409, 18 dez. 2014.

WANG, S. et al. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Scientific reports*, v. 6, p. 18962, 11 jan. 2016a.

WANG, S. et al. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Scientific reports*, v. 6, p. 18962, 11 jan. 2016b.

WARD, J. J. et al. Secondary structure prediction with support vector machines. *Bioinformatics*, v. 19, n. 13, p. 1650–1655, 2003.

WARD, J. J. et al. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of molecular biology*, v. 337, n. 3, p. 635–645, 26 mar. 2004.

WILKIE, G. S.; DICKSON, K. S.; GRAY, N. K. Regulation of mRNA translation by 5'- and 3'-UTR-binding factors. *Trends in biochemical sciences*, v. 28, n. 4, p. 182–188, 2003.

WILLIAMS, R. W. et al. Secondary structure predictions and medium range interactions. *Biochimica et biophysica acta*, v. 916, n. 2, p. 200–204, 26 nov. 1987.

WILLIFORD, A.; DEMUTH, J. P. Gene expression levels are correlated with synonymous codon usage, amino acid composition, and gene architecture in the red flour beetle, *Tribolium castaneum*. *Molecular biology and evolution*, v. 29, n. 12, p. 3755–3766, dez. 2012.

WILSON, B. A.; MASEL, J. Putatively noncoding transcripts show extensive association with ribosomes. *Genome biology and evolution*, v. 3, p. 1245–1252, 26 set. 2011.

WOLF, Y. I. et al. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*, v. 106, n. 18, p. 7273–7280, 5 maio 2009.

WOOLFSON, D. N. et al. De novo protein design: how do we expand into the universe of possible protein structures? *Current opinion in structural biology*, v. 33, p. 16–26, ago. 2015.

WU, D.-D.; IRWIN, D. M.; ZHANG, Y.-P. De novo origin of human protein-coding genes. *PLoS genetics*, v. 7, n. 11, p. e1002379, nov. 2011.

XIE, C. et al. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS genetics*, v. 8, n. 9, p. e1002942, set. 2012.

XIE, L.; BOURNE, P. E. Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS computational biology*, v. 1, n. 3, p. e31, ago. 2005.

YANG, H. et al. Expression Profile and Gene Age Jointly Shaped the Genome-Wide Distribution of Premature Termination Codons in a *Drosophila melanogaster* Population. *Molecular biology and evolution*, v. 32, n. 1, p. 216–228, 2014.

YANG, Y. et al. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Briefings in bioinformatics*, p. bbw129, 2016.

YAO, X.-Q.; ZHU, H.; SHE, Z.-S. A dynamic Bayesian network approach to protein secondary structure prediction. *BMC bioinformatics*, v. 9, p. 49, 25 jan. 2008.

YEE, A. A. et al. NMR and X-ray crystallography, complementary tools in structural proteomics of small proteins. *Journal of the American Chemical Society*, v. 127, n. 47, p. 16512–16517, 30 nov. 2005.

YIN, H. et al. What Signatures Dominantly Associate with Gene Age? *Genome biology and evolution*, v. 8, n. 10, p. 3083–3089, 13 out. 2016.

YI, T.-M.; LANDER, E. S. Protein Secondary Structure Prediction Using Nearest-neighbor Methods. *Journal of molecular biology*, v. 232, n. 4, p. 1117–1129, 1993.

YU, J.-F. et al. Natural protein sequences are more intrinsically disordered than random sequences. *Cellular and molecular life sciences: CMLS*, v. 73, n. 15, p. 2949–2957, ago. 2016.

ZEMLA, A. et al. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, v. 34, n. 2, p. 220–223, 1 fev. 1999.

ZHANG, J. Evolution by gene duplication: an update. *Trends in ecology & evolution*, v. 18, n. 6, p. 292–298, jun. 2003.

ZHANG, W. et al. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome biology*, v. 16, p. 202, 1 out. 2015.

ZHAO, L. et al. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science*, v. 343, n. 6172, p. 769–772, 14 fev. 2014.

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.