**Instituto de Química**

**Igor Santos Duarte Costa**

# Análise de conservação de fragmentos intragênicos antimicrobianos em proteínas humanas.

**Trabalho de conclusão de curso**

**Brasília – DF**

**2º/2019**

**Instituto de Química**

**Igor Santos Duarte Costa**

# Análise de conservação de fragmentos intragênicos antimicrobianos em proteínas humanas.

*Trabalho de conclusão do curso de Bacharelado em Química apresentado ao instituto de Química da Universidade de Brasília como requisito parcial para a obtenção do título de bacharel em Química*

**Orientador:Prof. Dr. Guilherme Dotto Brand**

**Co-orientador: Prof. Dr. Antônio Francisco Araújo**

**2º/2019**

## SUMÁRIO

## LISTA DE ABREVIATURAS E ACRÔNIMOS

AMP Antimicrobial Peptides (Peptídeos antimicrobianos)
ASA Acessible Surface Area (Área de superfície acessível)
IAP Intragenic Antimicrobial Peptides (Peptídeos antimicrobianos Intragênicos)
MSA Multiple Sequences Alignment (Alinhamento de múltiplas sequências)
PDB Protein Data Bank (Banco de Dados de proteína)
RSA Relative Solvent Accessibility (Área relativa de acessibilidade ao solvente)
WCN Weighted Contact Number (Número de contato ponderado)

## LISTA DE TABELAS

## ÍNDICE DE FIGURAS

**Resumo**

Proteínas podem ser estudadas sob variadas perspectivas, sendo uma delas a busca por módulos protéicos menores com estrutura/função característica. No presente trabalho foi feita a investigação da frequência e conservação de sequências internas de proteínas humanas com características catiônicas e anfifílicas, as quais, uma vez retiradas do contexto protéico parental e sintetizadas como entes individuais, apresentam potencial atividade antimicrobiana, também chamados de IAPs (*Intragenic Antimicrobial Peptides*). Tais sequências foram identificadas com o auxílio do software Kamal, o qual percorreu o proteoma humano padrão em busca de alfas hélices anfifílicas de carga líquida superior a +2. Além disso, o site PantherDB foi utilizado para acessar as classes e funções biológicas das proteínas que apresentam potenciais IAPs como parte de sua estrutura. Para tentar identificar a ação de seleção purificadora nos segmentos caracterizados como potenciais IAPs, o site Consurf foi utilizado para a construção de alinhamentos múltiplos, sendo estes posteriormente submetidos ao algoritmo RATE4SITE. A estrutura terciária das proteínas parentais foi extraída do PDB e a área de acessibilidade relativa ao solvente foi calculada para cada aminoácido com ajuda do site PDBEPisa. Os dados demonstram que segmentos catiônicos anfifílicos com características compatíveis com IAPs são mais frequentes em proteínas humanas do que na mesma amostra aleatorizada. Segundo o site PantherDB, proteínas humanas contendo potenciais IAPs são majoritariamente de duas classes: proteínas transmembrana e proteínas relacionadas ao metabolismo de fosfato, especialmente as que atuam na vicinidade de membranas plasmáticas. Para 29 das 30 proteínas avaliadas até o momento, não foi encontrado um grau de conservação dos resíduos de aminoácidos superior ao esperado por sua acessibilidade ao solvente para o fragmento catiônico anfifílico em relação aos resíduos remanescentes da mesma proteína. A única exceção foi a proteína exportina 1, em que o fragmento catiônico anfifílico está de fato em uma região de alta conservação e parece ter grande relevância estrutural/funcional para a proteína. O presente trabalho constitui um primeiro esforço de investigação da natureza e função de segmentos catiônicos anfifílicos em proteínas e na inferência de sua relevância para as proteínas que os contém.

**Palavras-chave**: antimicrobiano, bioativo, bioinformática, catiônico, conservação, encriptado fragmentos, peptídeos, proteínas, proteoma, RATE4SITE.

**ABSTRACT**

Proteins can be studied from several perspectives; Those macromolecules can be thought as composed by smaller proteins modules with structure/function. In this work we investigated the frequency and conservation of internal human proteins sequences with cationic and amphiphilic characteristics. Once removed from their context in the parental protein and synthesized as individuals those fragments presents potential antimicrobial activity giving their name IAPs (Intragenic Antimicrobial Peptides). Those sequences were identified by a software named Kamal that scanned a reference human proteome for amphiphilic helices with a net charge greater than +2. In addition the PnahterDB website was used to access the biological classes and functions of proteins containing putative IAPs. To identify a purified selection in the segments characterized as IAPs, the Consurf site was used for the construction of multiple sequences alignments, wich were later used by the RATE4SITE algorithm. The tertiary 3D structures of parental proteins were extracted from the PDB database and the acessible surface area (ASA) was calculated for each amino acid residue using the PDBEPISA website. Data demonstrated that putative IAPs are more frequent in human proteins when compared with random virtual proteins. According to the PantherDB website, the human proteins containing putative IAPs are from two big groups: transmembrane proteins and phosphate metabolism related proteins, especially those who act close to membranes. For 29 of the 30 proteins evaluated so far, we have not found a higher conservation score than expected from the ASA for the amino acids residues from the putative IAP. The only exception was the exportin 1 wich putative IAP is in fact in a highly conserved region and appears to have a big structural/functional relevance for the parental protein. This work represents a first research effort on the investigation of nature and relevance for cationic amphiphilic segments and how they are important for their parental protein.

**Keywords**: antimicrobial, bioactive, bioinformatics, cationic, conservation, encrypted, fragments, human proteome, peptides, proteins, RATE4SITE.

## 1. Introdução

Proteínas, quando submetidas à hidrólise parcial, podem liberar uma ampla gama de fragmentos, dentre eles, fragmentos com atividades biológicas diversas não manifestas na proteína parental (Meisel, 2004). Por vezes estes fragmentos possuem atividade antimicrobiana, como o caso de um peptídeo derivado da porção C-terminal da Interleucina-8 (Björstad et al., 2005). A investigação destes peptídeos bioativos encriptados em proteínas tem sido feita, tradicionalmente, a partir de uma abordagem experimental, com a obtenção de matrizes protéicas, submissão à hidrólise parcial, e caracterização experimental dos produtos hidrolíticos com posterior investigação das potenciais atividades biológicas. Desta maneira foram caracterizados peptídeos antimicrobianos, quimiotáticos, hipotensores e opioides, em matrizes como sangue e matriz extracelular, entre outros. Recentemente foram introduzidas ferramentas de busca *in sílico* que demonstraram que segmentos protéicos com potenciais atividades biológicas encriptados em proteínas maduras são mais frequentes do que previamente pensado. Proteínas dos mais variados organismos apresentando diferentes estruturas e funções passaram a ser consideradas fontes de peptídeos bioativos encriptados. O software Kamal (http://alelobag.cenargen.embrapa.br/Kamal/), aplicado a proteínas humanas, encontrou aproximadamente 1700 segmentos protéicos com propriedades físico-químicas compatíveis com peptídeos antimicrobianos, os chamados *Intragenic Antimicrobial Peptides* (IAPs) (Brand *et al.*, 2019). Contudo, pouco se sabe sobre a natureza e o papel biológico destes segmentos catiônicos anfifílicos em suas proteínas parentais. Além do mais, pouco se sabe se estes fragmentos são efetivamente liberados em algum momento entre a síntese e a degradação destas proteínas. Este trabalho tem como objetivo a utilização de ferramentas bioinformáticas como uma primeira aproximação na tentativa de compreender a natureza de tais segmentos no proteoma de *Homo sapiens* e sua relevância às proteínas que os contém.

# 2 Fundamentação teórica

## 2.1 PROTEÍNAS E AMINOÁCIDOS

Proteínas são macromoléculas com diversas funções essenciais aos seres vivos, sendo elas a maneira de expressão da informação genética (COX; NELSON, 2014). A química da vida é praticamente toda regulada por proteínas (enzimas), além de estarem nesta classificação, também, diversas moléculas que apresentam papel estrutural essencial à vida. Outras proteínas fundamentais podem atuar como anticorpos e hormônios.

Das mais simples às mais complexas, todas as proteínas são construídas da mesma matéria prima: aminoácidos. Essas são moléculas com um grupo carboxila e um grupo amino ligados ao mesmo átomo de carbono, diferindo por suas cadeias laterais, totalizando 20 aminoácidos comuns (Tabela 1).

Tabela 1 - Os 20 tipos comuns de aminoácidos (Cox ; Nelson,2014).

| Aminoácido | Abreviação/ símbolo | $M_r$* | Valores de p$K_a$ | | | pI | Índice de hidropatia† | Ocorrência em proteínas (%)‡ |
|---|---|---|---|---|---|---|---|---|
| | | | p$K_1$ (—COOH) | p$K_2$ ($—NH_3^+$) | p$K_R$ (grupo R) | | | |
| **Grupos R alifáticos, apolares** | | | | | | | | |
| Glicina | Gly G | 75 | 2,34 | 9,60 | | 5,97 | –0,4 | 7,2 |
| Alanina | Ala A | 89 | 2,34 | 9,69 | | 6,01 | 1,8 | 7,8 |
| Prolina | Pro P | 115 | 1,99 | 10,96 | | 6,48 | –1,6 | 5,2 |
| Valina | Val V | 117 | 2,32 | 9,62 | | 5,97 | 4,2 | 6,6 |
| Leucina | Leu L | 131 | 2,36 | 9,60 | | 5,98 | 3,8 | 9,1 |
| Isoleucina | Ile I | 131 | 2,36 | 9,68 | | 6,02 | 4,5 | 5,3 |
| Metionina | Met M | 149 | 2,28 | 9,21 | | 5,74 | 1,9 | 2,3 |
| **Grupos R aromáticos** | | | | | | | | |
| Fenilalanina | Phe F | 165 | 1,83 | 9,13 | | 5,48 | 2,8 | 3,9 |
| Tirosina | Tyr Y | 181 | 2,20 | 9,11 | 10,07 | 5,66 | –1,3 | 3,2 |
| Triptofano | Trp W | 204 | 2,38 | 9,39 | | 5,89 | –0,9 | 1,4 |
| **Grupos R polares, não carregados** | | | | | | | | |
| Serina | Ser S | 105 | 2,21 | 9,15 | | 5,68 | –0,8 | 6,8 |
| Treonina | Thr T | 119 | 2,11 | 9,62 | | 5,87 | –0,7 | 5,9 |
| Cisteína¶ | Cys C | 121 | 1,96 | 10,28 | 8,18 | 5,07 | 2,5 | 1,9 |
| Asparagina | Asn N | 132 | 2,02 | 8,80 | | 5,41 | –3,5 | 4,3 |
| Glutamina | Gln Q | 146 | 2,17 | 9,13 | | 5,65 | –3,5 | 4,2 |
| **Grupos R carregados positivamente** | | | | | | | | |
| Lisina | Lys K | 146 | 2,18 | 8,95 | 10,53 | 9,74 | –3,9 | 5,9 |
| Histidina | His H | 155 | 1,82 | 9,17 | 6,00 | 7,59 | –3,2 | 2,3 |
| Arginina | Arg R | 174 | 2,17 | 9,04 | 12,48 | 10,76 | –4,5 | 5,1 |
| **Grupos R carregados negativamente** | | | | | | | | |
| Aspartato | Asp D | 133 | 1,88 | 9,60 | 3,65 | 2,77 | –3,5 | 5,3 |
| Glutamato | Glu E | 147 | 2,19 | 9,67 | 4,25 | 3,22 | –3,5 | 6,3 |

*Os valores de $M_r$ refletem as estruturas como mostradas na Figura 3-5. Os elementos da água ($M_r$ 18) são removidos quando o aminoácido é incorporado a um polipeptídeo.

As propriedades físico-químicas variam entre os resíduos de aminoácidos, por exemplo, carga, solubilidade, reatividade e potencial para fazer ligação de

hidrogênio (FENNEMA, 1996). Eles são usualmente classificados como alifáticos apolares, aromáticos, polares não carregados, carregados positivamente e carregados negativamente.

Todos os aminoácidos possuem carbonos quirais (carbono alfa ao carbono carbonílico), exceto a glicina, podendo existir em duas formas enantioméricas, embora a maioria dos aminoácidos encontrados na natureza sejam L-aminoácidos (classificação enantiomérica) (Cox ; Nelson,2014). Também possuem caráter anfótero (podem agir como ácido e base ao mesmo tempo), já que o grupo amino é básico enquanto o grupo carboxílico é ácido, podendo existir em três estados de ionização em função do pH do meio – neutro, ácido (positivo) e básico (negativo) – e, asssim, todos os aminoácidos possuem no mínimo dois pKas, um para o grupo amino e outro para o grupo carboxílico, sendo que alguns aminoácidos possuem ainda um terceiro pKa referente a protonação da cadeia lateral. O pH em que um aminoácido apresenta carga líquida igual a zero é chamado de ponto isoelétrico.

Fazendo-se a junção das propriedades físico-químicas individuais dos aminoácidos, a natureza molda estruturas macromoleculares essenciais à vida. Por meio de ligações peptídicas (entre o grupo carboxílico e a amina, liberando água), os aminoácidos (agora chamados de resíduos de aminoácidos) dão origem a estruturas polipeptídicas, como as proteínas (Figura 1).

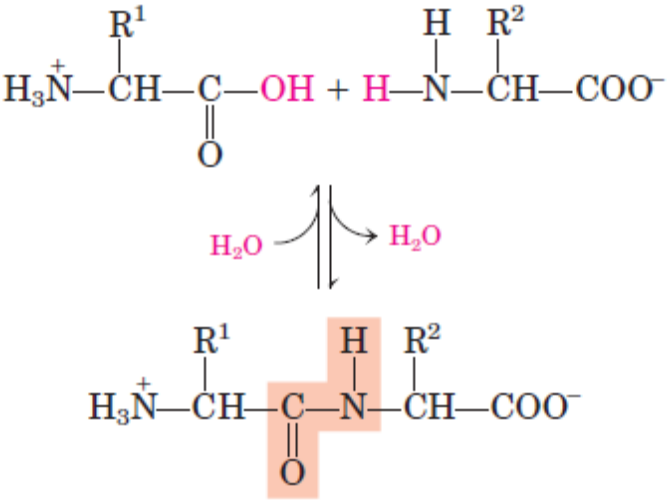


*Figura 1 - Ligação peptídica sendo formada por reação de condensação (Cox ; Nelson,2014).*

O estudo de proteínas é mais complexo do que parece no primeiro momento, pois cada proteína possui muitos graus de liberdade, podendo dobrar de diversas formas diferentes. Uma ligação peptídica é planar, o que pode ser explicado pela doação de densidade eletrônica advinda do par de elétrons livre do nitrogênio para

o carbono carbonílico e, consequentemente, para o átomo de oxigênio, ilustrado pela estrutura de ressonância (Figura 2). Tal fenômeno aumenta a energia necessária para rotacionar esta ligação, lhe conferindo planaridade (FENNEMA, 1996).

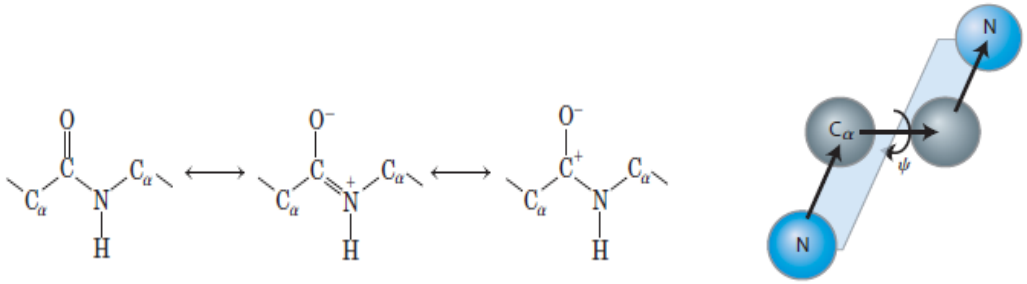


*Figura 2 – Estrutura de ressonância que explica geometria da ligação peptídica e os planos e ângulos do torção formados em ligações peptídicas (Cox ; Nelson,2014).*

Sendo assim, peptídeos poderiam se manifestar como diastereoisômeros *cis* ou *trans*, apresentando preferencialmente a conformação *trans* devido a maior estabilidade do isomêro, explicada pelo maior afastamento de grupos volumosos (FENNEMA, 1996). Os peptídeos, portanto, em um primeiro momento, formam planos com dois ângulos de rotação – um em torno da ligação entre o nitrogênio e o carbono-α, chamado de ângulo ϕ, e outro em torno da ligação entre o carbono-α e o carbono carbonílico, chamado de ângulo ψ. Isto não é o suficiente para descrever a estrutura de uma proteína, porém, pois a mesma é muito mais complexa que uma sequência linear de aminoácidos (chamada de estrutura primária).

Os ângulos de dobramento de peptídeos não podem assumir quaisquer valores, pois muitos deles são proibidos devido ao impedimento estérico. Esses ângulos assumem valores específicos de acordo com o dobramento da proteína no espaço, gerando assim uma estrutura secundária, sendo as mais comuns alfa-hélices (Figura 3) e folhas-beta (Figura 4) (Cox ; Nelson,2014), embora existam ainda outras estruturas. (FENNEMA, 1996).

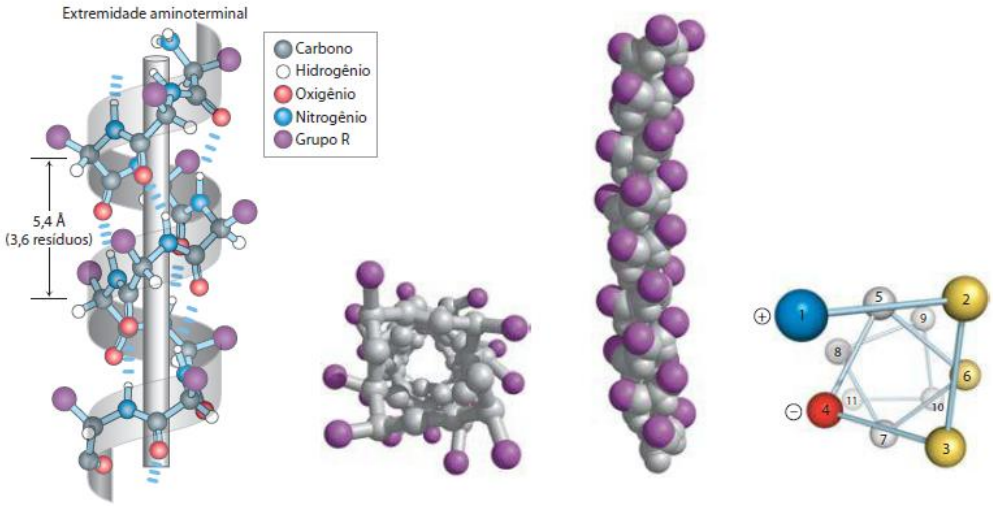


*Figura 3 - Diversas representações da estrutura secundária em alfa hélice (Cox ; Nelson,2014)*

Estruturas em alfa hélices são estabilizadas por ligação de hidrogênio entre os grupos N-H e o C=O do esqueleto carbônico, as quais estão paralelas ao eixo principal da hélice. Proteínas com estrutura primária P-N-P-P-N-N-P (em que P é polar e N não polar) costumam formar alfa hélices em meio aquoso rapidamente (FENNEMA, 1996), porém, naturalmente, a estrutura é dependente do tipo de resíduo de aminoácido ali presente. Por exemplo, resíduos de glicina (possui uma alta liberdade conformacional) e prolina (N não está ligado com H para fazer ligação de hidrogênio) quase não estão presentes em alfa hélices, por exemplo (Cox ; Nelson,2014).

Alfa hélices são usualmente representadas por rodas helicoidais (Figura 3), as quais indicam a posição de cada resíduo de aminoácido (sendo uma planificação da visão panorâmica da alfa hélice) (FENNEMA, 1996). Muitas vezes as hélices possuem resíduos hidrofílicos e hidrofóbicos muito bem separados, gerando características anfifílicas e podem ser caracterizadas por um ângulo polar (Ω) (UEMATSU; MATSUZAKI, 2000) – o ângulo que separa efetivamente os resíduos de aminoácidos hidrofílicos dos hidrofóbicos na roda helicoidal. O ângulo polar é uma medida da proporção entre as faces polar e apolar e caso uma face seja exclusivamente polar enquanto a outra é exclusivamente apolar, este ângulo seria de 180º (Figura 4) (YEAMAN; YOUNT, 2003).

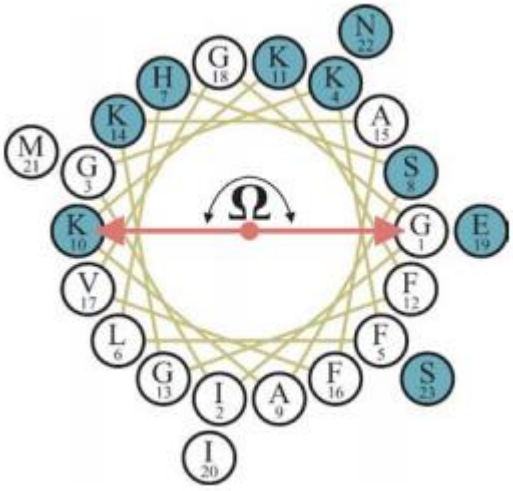


*Figura 4 - Roda helicoidal e a representação do ângulo polar (MURZYN; PASENKIEWIECZ-GIERULA, 2003).*

Folhas Beta são estruturas mais lineares, nas quais as interações por ligação de hidrogênio estão entre duas fitas diferentes do esqueleto peptídico, as quais podem estar dispostas de forma paralela ou antiparalela (Cox ; Nelson,2014).

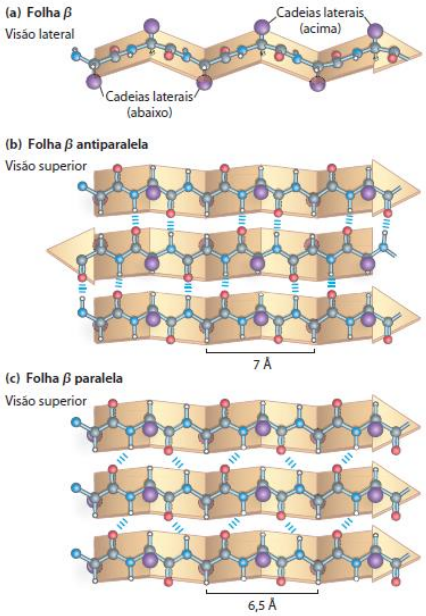


*Figura 5 - Representações de folhas Beta (Cox ; Nelson,2014)*

Além das estruturas secundárias, proteínas também apresentam estrutura terciária. Ela refere-se ao arranjo espacial tridimensional da proteína, levando em conta diferentes estruturas secundárias que ali podem estar presentes. Portanto é a proteína de uma forma mais global, possibilitando, inclusive, ver interações entre resíduos de aminoácidos em estruturas secundárias diferentes ali presentes (Cox ; Nelson,2014). Quando proteínas possuem mais de uma cadeia peptídica, formando

diversas subunidades, estas apresentam também uma estrutura quaternária, que diz respeito a como tais unidades estão dispostas no espaço.

### 2.2 CONSERVAÇÃO

Analisar o encadeamento de aminoácidos é, também, analisar a evolução dos seres vivos, afinal muitas proteínas devem ser tão primordiais que são encontradas, ao menos de forma similar, em diversos deles como a DNA polimerase. É importante considerar que divergências podem ocorrer entre proteínas que vieram de um ancestral comum, também chamadas de proteínas homólogas (PEARSON, 2013), no curso da evolução, devido a mutações que podem causar a inserção de novos resíduos de aminoácido, sua substituição, ou mesmo deleção dos resíduos de aminoácidos antigos (CARRILLO; LIPMAN, 2005).

Parece certo, porém, que resíduos de aminoácidos essenciais à função de determinada proteína devem ser conservados ao longo do processo evolutivo, ao contrário dos resíduos menos fundamentais (ECHAVE; SPIELMAN; WILKE, 2016). Entretanto, tal conservação nem sempre é rígida: muitas vezes uma proteína pode manter sua função permitindo apenas que alguns resíduos de aminoácido específicos sejam encontrados em determinada posição (seleção purificadora) (MASSINGHAM; GOLDMAN, 2005), sendo que essas trocas podem ocorrer entre aminoácidos com propriedades físico-químicas semelhantes tais quais carga e tamanho, gerando uma mutação não tão prejudicial.

Quando uma mutação no DNA acarreta na substituição de um resíduo de aminoácido conservado ao longo do processo evolutivo, dois fenômenos bem distintos podem ocorrer: a proteína pode perder sua função, ou mantê-la (talvez até com uma mudança em sua capacidade ligante) (MASSINGHAM; GOLDMAN, 2005). Mutações que não acarretam mudanças nas proteínas de maneira geral são chamadas de mutações neutras.

Seleção purificadora diz respeito à pressão evolutiva sofrida por determinados fragmentos em proteínas para permanecerem imutáveis ao longo do tempo (MASSINGHAM; GOLDMAN, 2005). Resíduos de aminoácidos que sofreram seleção purificadora são mais conservados que os demais (Figura 6). Na figura abaixo, que apresenta um alinhamento múltiplo de sequencias (MSA), é possível ver claramente que a Valina (V) na terceira posição é bem conservada em proteínas

homólogas de espécies diferentes, tal qual a Cisteina (C) na sexta posição e a Glicina (G) na sétima.


*Figura 6 - Exemplo de alinhamento de múltiplas sequências (MSA) em que é possível identificar resíduos de aminoácidos que sofreram seleção purificadora (CHENNA et al., 2003). As proteínas são encontradas em bactérias (1fdr e 2pia), javali (1ndh) e em espinafre (1nfc).*

A construção de um MSA é uma das ferramentas mais úteis à compreensão do grau de conservação de resíduos de aminoácidos em proteínas específicas ao longo do processo evolutivo. Agrupando-se todas as proteínas homólogas, forma-se uma família de proteínas. Sua identificação é dada por modelos computacionais baseados em diversos bancos de dados (como por exemplo o PDB). Para tanto, se fazem necessários critérios de correspondência, sob os quais são comparadas diversas sequências. Pode-se, então, produzir um *score* de conservação, atribuindo valores numéricos ao grau de conservação de cada resíduo. Às vezes o algoritmo pode achar correspondências separadas por um intervalo vazio em determinada sequência. Para resolver tal problema o código cria espaços vazios e busca minimizar o número de espaços tal qual maximizar o número resíduos alinhados (Figura 7).

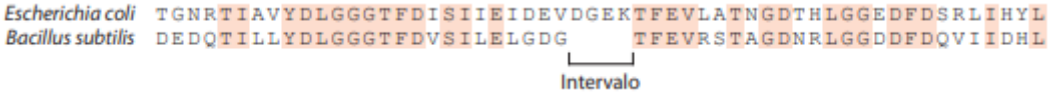


*Figura 7 - Exemplo de quando uma lacuna deve ser criada para que seja possível fazer um alinhamento efetivo (Cox ; Nelson,2014).*

Um bom algoritmo também deve identificar quando uma substituição entre dois resíduos de aminoácidos com propriedades semelhantes ocorre – ou seja, uma mutação conservativa – atribuindo-lhes uma maior pontuação que a uma substituição não conservativa. É também uma prática comum embaralhar uma sequência aleatória e refazer o alinhamento, provando-se, assim – quando o alinhamento original traz números de *score* de conservação muito mais significativos que o aleatório – que o código de fato está investigando uma família.

Proteínas, sendo macroestruturas, podem manifestar-se em diversas conformações diferentes, porém suas atividades estão intimamente ligadas a uma conformação nativa (uma estrutura em três dimensões específica) e, na maioria das

vezes, sítios específicos. Em tais sítios encontra-se um grupo de resíduos de aminoácido que realiza, efetivamente, sua função. Deve haver, então, um balanço entre restrições estruturais e funcionais no decorrer do processo evolutivo (ECHAVE; SPIELMAN; WILKE, 2016).

### 2.2.2 Fatores estruturais

Desde os anos 60 é sabido que resíduos de aminoácidos em sítios internos tendem a ser melhor conservados quando comparados a resíduos de sítios mais expostos, sendo uma das explicações possíveis a não permissão de resíduos de aminoácidos polares na porção interna da proteína, a qual é apolar – menos mutações são permitidas na porção interna da proteína, portanto (ECHAVE; SPIELMAN; WILKE, 2016). Para quantificar tal ideia, surgiu o conceito de acessibilidade ao solvente – área de superfície acessível (ASA), ou, também, área relativa de acessibilidade ao solvente (RSA) quando normalizada. Trata-se de uma medida bastante robusta, já sendo bem definido que a mesma não sofre influência de outras variáveis como estrutura secundárias, formação de ligação de hidrogênio, hidrofobicidade ou tamanho (ECHAVE; SPIELMAN; WILKE, 2016). Também já foi demonstrado que a taxa de variação de um sítio cresce linearmente com o aumento da RSA (ECHAVE; SPIELMAN; WILKE, 2016).

Há também medidas que levam em consideração a densidade de resíduos de aminoácidos por região, ou seja, quantos vizinhos entram em contato com determinado resíduo de aminoácido (densidade de empacotamento), porém, em geral, elas não são capazes de gerar modelos preditivos tão bons quanto as RSAs, embora há casos específicos em que a medida WCN (número de contato ponderado – leva em consideração todos os resíduos de aminoácidos da proteína e os da pesos de acordo com o quadrado distância entre os mesmos e o resíduo de aminoácido analisado) leva a resultados melhores que a RSA (ECHAVE; SPIELMAN; WILKE, 2016).

Há ainda medidas considerando a flexibilidade das proteínas, afinal são elas grandes polímeros constantemente em movimento que sofrem pequenas mudanças conformacionais a todo momento. Como uma proteína só exerce sua função a partir de uma estrutura nativa (ou ativa), embora existam exceções, é esperado que regiões com alta flexibilidade sejam menos importantes e, portanto, apresentem maiores taxas mutacionais que regiões rígidas (ECHAVE; SPIELMAN; WILKE,

2016). Muitas outras variáveis estão relacionadas a restrições estruturais, porém ainda não apresentam correlação tão forte quanto as três citadas.

### 2.2.3 Fatores funcionais

Uma proteína precisa de sua conformação ativa para desempenhar sua função, mas apenas o arranjo espacial correto não garante seu funcionamento. Proteínas que não são meramente estruturais possuem sítios ativos tão fundamentais que devem ser conservados, e, mesmo os resíduos de aminoácidos que não estão em tal sítio, mas interagem com o mesmo (direta ou indiretamente), comportam poucas mutações, pois possuem a capacidade de desabilitar a proteína (ECHAVE; SPIELMAN; WILKE, 2016). Tais resíduos precisam de mais que as medidas de RSA para descrevê-los.

Todavia, nem sempre resíduos de aminoácidos que são fundamentais e participam ativamente da função da proteína são altamente conservados. Uma das formas de se identificar um sítio ativo é encontrar resíduos de aminoácidos que sofreram seleção positiva, ou seja, resíduos que sofreram mutações recentemente, mas que sofrem pressão para permanecerem imutáveis desde então (MASSINGHAM; GOLDMAN, 2005).

No decorrer do processo evolutivo, é comum que uma determinada espécie mude seus hábitos como também seu habitat. Tais mudanças, além de outros grandes eventos como a transmissão de determinada doença, por exemplo (VOIGHT et al., 2006), podem ser decisivas para determinar a importância de certas características. Uma seleção positiva acontece em função da transmissão de características que, a partir de determinado evento, tornam-se fundamentais ou, ao menos, de grande importância. Ela é uma força que age em função de novos genes em detrimento de genes antigos.

A seleção positiva é o contrário da seleção purificadora (MASSINGHAM; GOLDMAN, 2005), sendo que uma diz sobre novas mutações que são bem-vindas a população, enquanto a outra diz sobre sequências de aminoácidos que não devem sofrer mutações em nenhuma hipótese. Fragmentos que sofreram seleção positiva possuem uma variância muito maior que os demais e, geralmente, estão presentes apenas em populações que presenciaram determinado evento, sendo ele responsável por tal fenômeno. Enquanto a seleção purificadora é dominada por

mutações conservativas, a maior evidência de seleção positiva são muitas mutações não-conservativas ocorrendo em determinados sítios da proteína (MASSINGHAM; GOLDMAN, 2005).

Em suma, fatores funcionais complementam modelos estruturais: eles explicam porque determinados resíduos sofreram mais ou menos mutações que o esperado e um bom modelo preditivo sempre deve levar em conta os dois.

## 2.3 AVALIANDO A CONSERVAÇÃO UTILIZANDO O CONSURF

O estudo da conservação de resíduos é fundamental à compreensão da história evolutiva de proteínas e, para tanto, se faz necessário o uso de algum algoritmo computacional como os implementados no site consurf (CELNIKER et al., 2013). Seu input pode ser a estrutura primária de uma proteína ou uma estrutura 3D e, embora sua metodologia seja baseada apenas na sequência, o algoritmo extrai a sequência a partir da estrutura como consta no banco de dados *protein data bank* (PDB).

### 2.3.1 Busca por homólogos

Sequências podem ser ditas homólogas quando são bastantes similares, embora o limiar dependa de diversos fatores de cálculo. Proteínas homólogas, em sua definição, são as que possuem ancestral comum. Analisar homólogos não é uma tarefa simples, pois, muitas das vezes, algoritmos não indicam sequências que são claramente próximas em termos de função e distância evolutiva, enquanto podem identificar não homólogos como sendo pertencentes a uma mesma família (PEARSON, 2013). Algoritmos de busca por homólogos tendem a minimizar falsos positivos, embora não sejam efetivos no controle de falsos negativos (PEARSON, 2013).

Por meio de modelos estatísticos, é possível identificar excessos de similaridade e, portanto, homologia. Para tanto, é preciso antes falar de uma similaridade esperada por mero acaso, proporcionada pelas longas base de dados existentes de proteínas, sendo já observado que sequências aleatórias possuem *scores* de similaridade próximos a sequências não relacionadas (PEARSON, 2013). O método BLAST calcula tal probabilidade para um único par de alinhamento, sendo necessário que, em termos práticos, o algoritmo retorne a esperança $E(b) \leq p(b)D$

após várias interações (PEARSON, 2013). D representa o número de sequências presentes no *banco de dados*, E a esperança e P a probabilidade, ambas dependentes do número de bits.

Uma vez que a esperança dependente da quantidade de dados a ser considerada, é de se esperar que a análise de conservação encontre inconsistências. Dois pares ditos homólogos podem ser identificados quando comparados em bancos de dados pequenos, porém serem um falso positivo em bancos de dados maiores, pois não passaram no novo critério mais rigoroso. Métodos HMM são usados para contornar tal problema (PEARSON, 2013).

É importante considerar que duas sequências homólogas não necessariamente advenham de proteínas homólogas. Os algoritmos de busca utilizados pelo consurf tratam de alinhamentos locais que indicam regiões. Como os *scores* são planejados para detectar similaridades longínquas, é comum que regiões homólogas sejam extrapoladas para vizinhos não homólogos, sobretudo em métodos interativos como o PSI-BLAST e o CSI-BLAST (PEARSON, 2013).

Alinhamentos são feitos com base em uma matriz de pontuação (método de programação dinâmica), a qual define valores para *GAPs* (quando há uma lacuna – há um resíduo a mais em uma das sequências). Métodos de alinhamento local não penalizam GAPs iniciais), *Matchs* (quando os dois resíduos são semelhantes) e *Mismatches* (quando os resíduos são distintos). Cada método possui uma matriz de pontuação diferente, além de diversas particularidades estatísticas que visam contornar os problemas discutidos.

O consurf permite escolher o número de interações, além do valor de corte para a esperança e o banco de dados desejado.

### 2.3.2 Alinhamento de múltiplas sequências

O aumento na dificuldade de alinhamento de mais de um par de sequências protéicas é evidente. Quando uma matriz de pontuação é definida, com ela surge um custo de alinhamento (baseado em pontuações negativas – GAPs e mutações não conservativas). Em um alinhamento de múltiplas sequências (MSA) a melhor maneira de otimização seria minimizar custo de pares associados aos galhos de uma árvore evolutiva, na qual as folhas são sequências (LIPMAN; ALTSCHUL; KECECIOGLUT, 1989). Sendo utilizado também a minimização da soma dos custos de todos os pares de sequências possíveis.

Uma proposta bastante aceita para melhorar a velocidade da programação dinâmica é já usar sequências homólogas no processo de MSA (KATOH et al., 2002). O consurf permite a utilização de quatro métodos: MAFFT (algoritmo baseado na transformada rápida de fourier), PRANK (algoritmo que busca identificar o evento evolutivo que leva às diferenças e, com isso, definir as pontuações) (LO, 2014), MUSCLE (algoritmo baseado em distâncias estatísticas) (EDGAR, 2004) e CLUSTAW (algoritmo de alinhamento baseado em palavras) (CHENNA et al., 2003).

O consurf delimita uma lista de sequências de número específico, padronizado em 150 (sendo possível alterar caso desejado). Tal lista vale-se de dois limiares baseado na %ID (quantos por cento um homólogo é igual a outro), sendo um máximo (para evitar redundância na análise) e um mínimo (para evitar falsos positivos). Também há a opção de escolher as proteínas mais próximas entre si, ou realizar uma amostragem aleatória a partir de todos os homólogos encontrados.

Tendo o MSA pronto, a última etapa realizada pelo consurf é encaminhá-lo para o algoritmo da RATE4SITE (MAYROSE et al., 2004), o qual busca criar uma nova escala de *score* de conservação, desta vez, diretamente para os sítios da proteína. É impossível separar uma correlação de conservação com sítios específicos de uma árvore filogenética da proteína, porém não é trivial, mesmo após ter um MSA em mão (MAYROSE et al., 2004).

O RATE4SITE o faz utilizando uma estimativa Bayesiana ou pelo princípio da máxima verossimilhança (ML) e costuma ter sucesso em identificar sítios ativos e mesmo resíduos de aminoácidos essenciais que não são tão óbvios. Entre os dois métodos disponíveis, a simulação Bayesiana fornece melhores resultados (MAYROSE et al., 2004).

O próprio score de conservação advindo do RATE4SITE pode ser bem elucidativo, como demonstrado por Toporik e colaboradores (Toporik *et al.*, 2014). Em uma forma de análise de perfil evolutivo de proteínas, o grupo mostrou que, além de regiões funcionais importantes, o grau de conservação da estrutura primária pode ser usado como método de identificação de porções ativas em proteínas secretadas, a partir da identificação de regiões de alta conservação da proteína, como exemplificado para a Insulina (Figura 8). O peptídeo sinal é uma porção exposta da molécula responsável por sinalizar que esta molécula deve ser

secretada. Também é possível ver que o propeptídeo entre as cadeias A e B não apresenta alta conservação, algo que faz sentido já que o mesmo não faz parte da proteína madura, composta somente das cadeias A e B, não apresentando função conhecida.

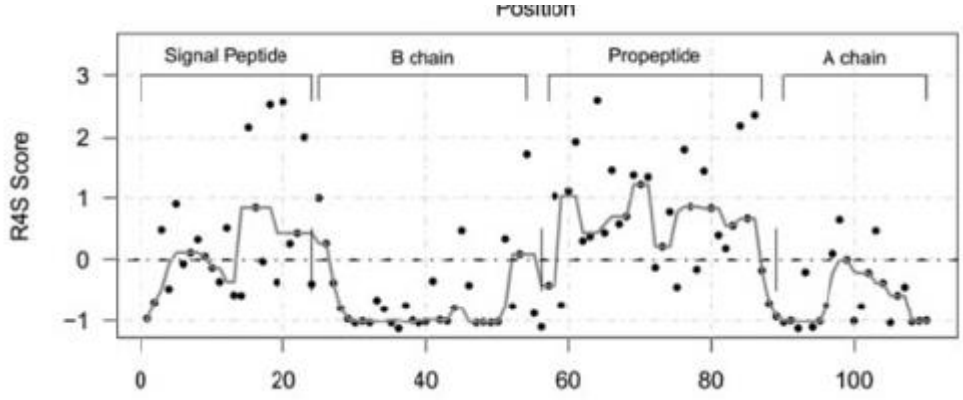


*Figura 8 - Demonstração sobre como peptídeos excretados são mais bem conservados que demais domínios na insulina (Toporik et al., 2014). A insulina madura é formada das cadeias A e B, as quais estão em regiões altamente conservadas.*

### 2.4 FRAGMENTOS ENCRIPTADOS EM PROTEÍNAS E OS IAPS

Vários experimentos demonstram que é possível submeter uma proteína a hidrólise parcial e com isso mobilizar fragmentos bioativos diversos da proteína parental (MÖLLER et al., 2008). Esses fragmentos protéicos são ditos encriptados em proteínas e há uma larga gama de exemplos. Fragmentos com propriedades imunomoduladoras, citomoduladoras, inibidora de ECA, opióides, entre outras, já foram encontrados em proteínas advindas do leite (MEISEL, 2004), uma matriz muito estudada como fonte de peptídeos encriptados.

Peptídeos encriptados podem ser liberados *in vivo* por meio de hidrólise por enzimas digestivas como tripsina ou de enzimas microbianas durante o processamento do alimento ou no processo de amadurecimento (MÖLLER et al., 2008). Quando os mesmos possuem função antimicrobiana (possuem a capacidade de inibir o crescimento de microrganismos), são chamados de peptídeos antimicrobianos intragênicos ou IAPs (BRAND et al., 2012). Já foram descritos alguns peptídeos antimicrobianos obtidos a partir da hidrólise de proteínas humanas, tal qual a interleucina-8 (BJÖRSTAD et al., 2005), uma proteína relevante na resposta imune a microrganismos. O mesmo artigo especula que o peptídeo IL-8 é produzido em condições fisiológicas e representa uma segunda instância de combate a microrganismos, inibindo diretamente seu crescimento. Assim, a

publicação especula que tal segmento possui valor adaptativo ao organismo humano.

Recentemente, alguns artigos que buscam peptídeos intragênicos antimicrobianos, hipotetizaram que tal fenômeno é mais amplo que previamente especulado (RAMADA et al., 2017). A título de exemplo, em uma varredura em 160 mil proteínas humanas, foram encontrados aproximadamente 1700 segmentes protéicos com características físico-químicas compatíveis com IAPs (Brand *et al.*, 2019). Contudo, dada a amplitude do fenômeno, torna-se necessário investigar se há algum indício de que tais segmentos existem, em algum momento entre a síntese e degradação de suas proteínas parentais, como agentes independentes, e, portanto, desempenham alguma função biológica outra daquela apresentada pela proteína parental. Uma das maneiras de se fazer tal investigação é identificar se há alguma pressão para manutenção da identidade dos resíduos de aminoácidos pertencentes a esses segmentos frente a outros da mesma proteína, ou seja, se estes estão submetidos a seleção purificadora.

Para investigar os IAPs encriptados em proteínas, um software chamado Kamal fora desenvolvido no Laboratório de Espectrometria de Massa da Embrapa Recursos Genéticos e Biotecnologia (BRAND et al., 2012). O software foi utilizado em publicações prova-de-conceito (BRAND et al., 2012) (RAMADA et al., 2017 (http://alelobag.cenargen.embrapa.br/Kamal/) (Brand *et al.*, 2019) .Este varre largos bancos de dados de proteínas executando uma digestão *in silico*, ou sua proteólise simulada, em que ocorre a varredura por sequências de tamanho determinado. O algoritmo assume que os fragmentos encontrados estão orientados em alfa hélices com determinado ângulo polar, configurado pelo usuário, e, simplesmente, permite dados resíduos, também escolhidos pelo usuário, de aminoácidos na porção polar e outros na porção apolar da hélice.

Um input no Kamal consiste na opção de escolha de resíduos de aminoácidos permitidos em determinadas posições, baseadas no ângulo polar de tal sequência, além do tamanho mínimo e máximo para peptídeo encriptado.

Outra questão a ser avaliada é a ancestralidade comum dos IAPs. Peptídeos antimicrobianos (AMPs) já foram observados em praticamente todos os organismos, mesmo em procariotos que, provavelmente, os usam para eliminar outros procariotos de nicho semelhante (RAMADA et al., 2016). Há indícios, portanto, que os mesmos tenham sido utilizados como base para a construção de

proteínas maiores, gerando, assim, uma gama de proteínas contendo IAPs encriptados, semelhante a determinados domínios evolutivos interagentes com moléculas de RNA especulados como "*building blocks*" para a construção de proteínas maiores e mais complexas (SÖDING; LUPAS, 2003). Soding e Lupas discutem sobre como proteínas podem ser vistas como combinações entre proteínas menores, mesmo não homólogas, sendo uma boa explicação acerca de padrões encontrados em estruturação mesmo em um universo tão denso de possibilidades. Evidências apontam que muitas das estruturas mais complexas e menos comuns decorrem de estruturas bem conhecidas, sendo que a natureza parece carregar o que já foi bem sucedido para o futuro (SÖDING; LUPAS, 2003).

Todavia, em muitos casos, já foi demonstrado que os IAPs são liberados por proteólise e participam ativamente do sistema imune, mesmo estando em proteínas sem quaisquer atividades antimicrobianas aparentes como o DPC6026 liberado da caseína (HAYES et al., 2006). IAPs parecem ser mais uma instância da economia energética observada recorrentemente na natureza, trazendo sequências de aminoácidos que ao mesmo tempo são estruturais para um determinado peptídeo e, quando necessário, desempenham funções próprias.

### 2.4.2 Anfifilicidade e os IAPs

Normalmente associadas a lipídeos, moléculas anfifílicas são compostos que possuem grandes cadeias apolares e um agrupamento polar, o que lhes confere características únicas, sendo bastante explorados em novas tecnologias de polímeros (FENG et al., 2017) (PERACCHIA et al., 2002). Tais substâncias conseguem agrupar-se com facilidade, formando micelas em condições adequadas, além de sempre apresentarem atividade em superfície – tendem a acumular na interface entre água-ar – e terem um papel fundamental na formação de membranas celulares. Não se pode imaginar a atividade celular, como vista na contemporaneidade, sem a regulação por moléculas anfifílicas, e, portanto, é racional considerar as mesmas como fundamentais no curso da evolução (DEAMER, 1986).

Estruturas em alfa-hélice anfifílicas carregadas positivamente parecem ser, muitas vezes, responsáveis pela atividade antimicrobiana de peptídeos, sendo comumente encontradas em muitos peptídeos com tal atividade (UEMATSU;

MATSUZAKI, 2000). São esses peptídeos que interagem com membranas carregadas negativamente e permeabilizam as células gerando, assim, suas características antimicrobianas (outros mecanismos também são discutidos na literatura). Peptídeos anfifílicos também são teorizadas como intermediários globais de dobramento de proteínas globulares, mesmo em sequências estruturadas em folhas beta, que possuem altos potenciais para formação de alfas hélices anfifílicas (LEE; PARKER, 2011). Vários argumentos sustentam essa hipótese como o rápido tempo de formação de alfa hélices e a sua compactação, capaz de gerar um glóbulo compacto que se difundiria com mais facilidade (CHEN et al., 2010).

Embora o potencial para a formação de alfa hélices anfifílicas não difira tanto de uma lista aleatória de proteínas e várias proteínas virtuais aleatorizadas (LEE; PARKER, 2011), é fundamental que o estudo seja feito com uma amostragem mais real (como um proteoma) e sob critérios que vão além do momento hidrofóbico. O KAMAL já provou ser bastante útil e dar resultados condizentes, e, sob uma nova ótica, é possível dar novos ares à investigação de uma questão bastante complexa.

## 3 OBJETIVOS

O objetivo geral do presente trabalho é investigar, por meio de ferramentas bioinformáticas, a ocorrência e a conservação de peptídeos intragênicos antimicrobianos no proteoma humano. Mais especificamente, ele se propõe a:

1. Determinar se existem mais potenciais IAPs no proteoma humano quando comparado a mesma amostra aleatorizada, visando identificar se tais fragmentos não teriam sido encontrados por mero acaso e, assim, ter uma primeira evidência de que sofreram algum tipo de pressão evolutiva para sua seleção.

2. Verificar se há algum padrão, em termos de funcionalidade, para as proteínas que possuem potenciais IAPs, buscando investigar se a presença destes está confinada a famílias de proteínas específicas, ou se trata-se de fenômeno mais amplo.

3. Por fim, determinar se de fato os fragmentos com potencial atividade antimicrobiana sofreram seleção purificadora, valendo-se de um modelo que leve em consideração fatores estruturais. Caso a hipótese seja verdadeira, a conservação de resíduos de aminoácidos presentes nos fragmentos com possível atividade antimicrobiana desviará do previsto pela acessibilidade ao solvente.

## 4 METODOLOGIA

As seguintes etapas foram realizadas (Figura 9):

1. Busca por IAPs no proteoma humano.
2. Aleatorização do proteoma humano.
3. Investigação da funcionalidade das proteínas fonte de IAPs
4. Construção de um alinhamento múltiplo das proteínas parentais.
5. Estudo da conservação dos IAPs em relação a suas proteínas-fonte.

### 4.1 BUSCA POR IAPS NO PROTEOMA HUMANO UTILIZANDO O SOFTWARE KAMAL

O software Kamal v2.0 foi utilizado para encontrar fragmentos protéicos por meio de critérios desenvolvidos em publicações anteriores (Ramada *et al.*, 2017)(Brand *et al.*, 2019). Como modelo, foi utilizado o proteoma humano referência, baixado do EMBL-EBI (https://www.ebi.ac.uk/reference_proteomes) na data de 23/11/2018. Nas posições em que os resíduos de aminoácidos são apolares foram permitidos apenas: Alanina, Fenilalanina, Isoleucina, Leucina, Metionina, Valina e Triptofano. Nas outras posições (polares) foram permitidos: Alanina, Cisteína, Ácido aspártico, Ácido glutâmico, Glicina, Histidina, Lisina, Metionina, Asparagina, Glutamina, Arginina, Serina, Treonina e Tirosina. A varredura foi feita em três ângulos polares (128º, 160º e 192º), visando ampliar as diversas possibilidades de se encontrar alfa hélices anfifílicas, e os fragmentos foram limitados entre 16 e 24 resíduos de aminoácidos. Após a varredura, foram considerados somente potenciais IAPS com carga líquida superior a +2 para as próximas etapas.

#### 3.1.1 Aleatorização do proteoma humano

Um script foi escrito em python 3.7 para aleatorizar completamente o proteoma humano referência de acordo com o seguinte algoritmo:

*1. Proteínas foram divididas em vários pequenos fragmentos de acordo com o formato do arquivo (sem critério específico);*

*2. Estes fragmentos foram aleatorizados, tanto na ordem dos resíduos de aminoácidos, quanto na ordem global em que apareceriam entre si.*

*3. Os novos fragmentos foram associados com os próximos dez fragmentos da nova ordem global, formando novos fragmentos maiores.*

*4. Novamente, esses foram aleatorizados tanto na ordem global, quanto na ordem de resíduos de aminoácidos interna.*

*5. O processo foi repetido por mais nove vezes, gerando dez proteomas completamente aleatorizados.*

Foram gerados 10 proteomas humano referência aleatorizados. Após a aleatorização, o Kamal foi executado utilizando os mesmos critérios descritos acima.

### 4.2 Determinação das famílias as quais pertencem as proteínas fonte de IAPs (PantherDB)

Após a desduplicação das proteínas humanas fontes de IAPs, a análise da função das proteínas contendo no mínimo 1 IAP foi feita com um software de distribuição livre chamado PantherDB (http://www.pantherdb.org) (Mi *et al.*, 2019), o qual compara uma lista de proteínas fornecidas pelo usuário com um banco de dados próprio. Panther permitiu o rápido uso do proteoma humano de referência, realizando automaticamente testes estatísticos de super-representação – os quais identificaram que grupos estavam mais presentes nas proteínas parentais, sendo uma excelente ferramenta para o estudo de enriquecimento seletivo em dada amostra de proteínas. Foi feito um teste exato de Fisher, com correção de cálculo de falsos positivos.

### 4.3 Construção de alinhamentos múltiplos e determinação da conservação de resíduos

As proteínas, então, foram levadas para o algoritmo do consurf (http://consurf.tau.ac.il/), obtendo-se os *scores* de conservação de cada resíduo de aminoácido retirado da comparação de proteínas homólogas de diversas espécies. As configurações do consurf foram todas padrão, utilizando-se o método HMMER para a busca de homólogos, o método mais novo implementado ao site. Para o alinhamento de múltiplas sequências, será utilizado o MAFFT-L-INS-i. O único parâmetro modificado foi a %ID mínima para os homólogos, que será posta em 50 ao invés de 35, visando evitar que falsos positivos atrapalhem a análise. Os scores de conservação do algoritmo RATE4SITE são dados automaticamente pelo consurf.

Os scores de conservação foram analisados em função das posições dos resíduos de aminoácido visando identificar se os potenciais IAPs estariam em região de alta conservação.

### 4.4 DETERMINAÇÃO DA ASA E RSA USANDO PDBEPISA

As medidas de ASA foram calculadas pelo algoritmo do site PDBe Pisa (https://www.ebi.ac.uk/pdbe/pisa/) (KRISSINEL; HENRICK, 2007) e foram convertidas para RSA através de um script em Python 3.7 (dividindo-se todos os valores pelo maior). Uma comparação foi feita alinhando-se os RSA divididos em três categorias (0-33% / 33%-66% / 66%-100%) e os *scores* de conservação através de uma análise estatística e, por fim, foi checado se os IAPs se encontrarão fora da margem prevista estatisticamente.

# 5 Resultados

## 5.1 Frequência de potenciais IAPs no proteoma humano

É de fundamental importância analisar se os critérios de busca utilizados pelo software Kamal não são tão permissivos que ocorreriam por mero acaso em bancos de dados constituídos por grandes números de sequencias protéicas. Para tanto, o proteoma humano referência foi submetido ao software Kamal assim como um proteoma aleatorizado, conforme descrito na seção de metodologia. Foi encontrada uma diferença estatisticamente significativa no número de segmentos que atendem aos critérios avaliados entre o proteoma humano padrão e sua aleatorização, conforme demonstrado na figura 9.

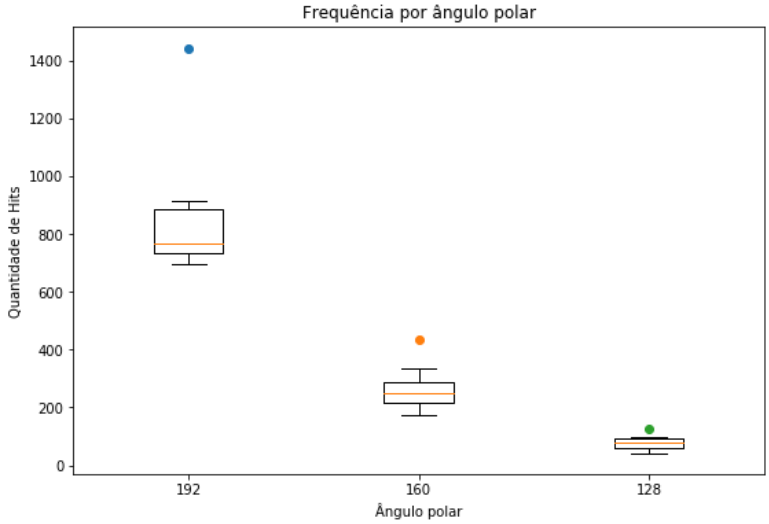


*Figura 9 - - Quantidade de potenciais IAPs por ângulo polar no proteoma humano x aleatorizações. No boxplot estão compiladas 10 amostras aleatórias e os pontos coloridos representam o valor real encontrado no proteoma humano de referência.*

É notável, também, que a diferença diminui com a diminuição do ângulo polar, porém não deixando de ser significativa. Tal resultado é esperado já que nos critérios escolhidos para o uso do software Kamal (Figura 10), há uma maior variabilidade de resíduos de aminoácidos permitidos na face hidrofílica, e, portanto, com sua redução (diminuição do ângulo polar), há uma significativa redução no número de fragmentos que se adequam ao critério.

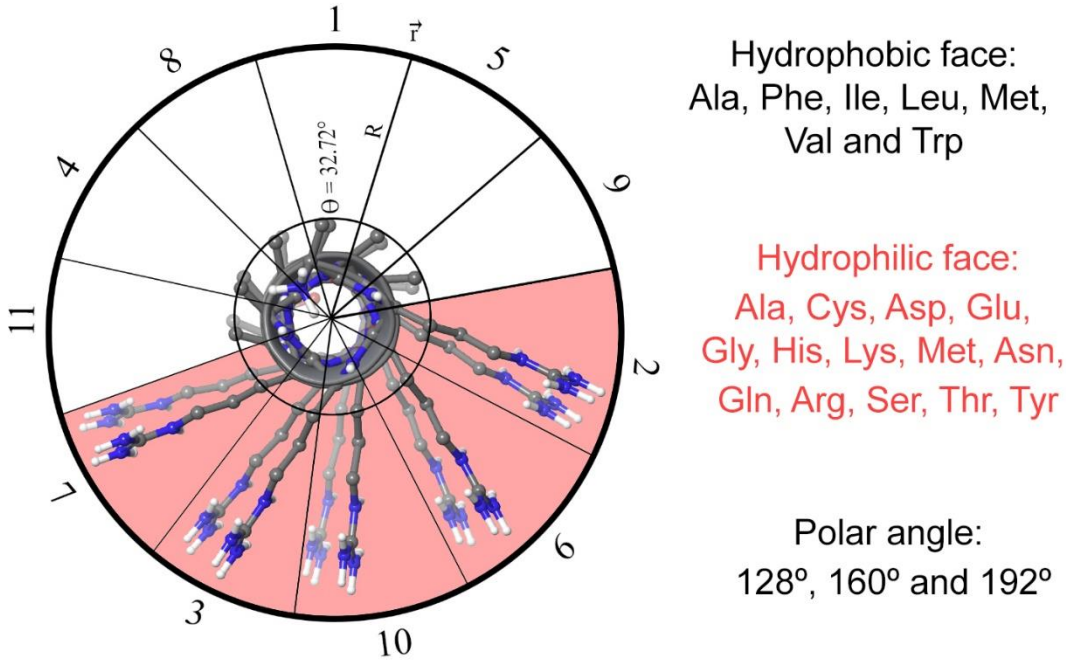


*Figura 10 - Ilustração dos critérios utilizados pelo Kamal (http://alelobag.cenargen.embrapa.br/Kamal/). Nela está descrito que resíduos de aminoácidos foram permitidos na face hidrofóbica, e quais foram permitidos na face hidrofílica, além dos três ângulos polares utilizados.*

É possível, então, concluir que fragmentos catiônicos e anfifílicos são mais comuns em proteínas codificadas pelo proteoma humano do que em amostras aleatórias. Assim, mesmo que possam ser encontrados ao acaso, há um enriquecimento deste padrão no proteoma humano, talvez por alguma propriedade funcional que segmentos protéicos com esse padrão possam conferir às proteínas que os contém.

### 5.2 FAMÍLIAS DAS PROTEÍNAS PARENTAIS

Como um primeiro esforço para entender o motivo da frequência de segmentos catiônicos anfifílicos em proteínas presentes no proteoma humano, ou os potenciais IAPs, pode-se analisar suas classes e funções biológicas. Deve-se questionar se estes fragmentos estão localizados em proteínas específicas, que desempenham certa função, ou se são gerais e estão distribuídos por diversos tipos de proteínas diferentes.

As proteínas parentais foram extraídas pelo próprio Kamal, rastreando-se a origem dos potenciais IAPs. Ao todo, após a colapsação dos resultados de todos os três ângulos polares e a desduplicação da lista foram encontradas 458 proteínas parentais, sendo 451 catalogadas pela database da ferramenta bioinformática

PantherDB. Uma amostra relativamente pequena frente as aproximadamente 21000 proteínas analisadas, presente no proteoma de referência.

### 5.2.1 IAPs estão relacionados com proteínas transmembrana e metabolismo de fosfato

Analisando-se as classes advindas do Gene Ontology através da ferramenta bioinformática Panther, é possível ver como as proteínas com potenciais IAPs diferem-se do esperado para o proteoma humano, a partir de um teste de super-representação (Figura 11):

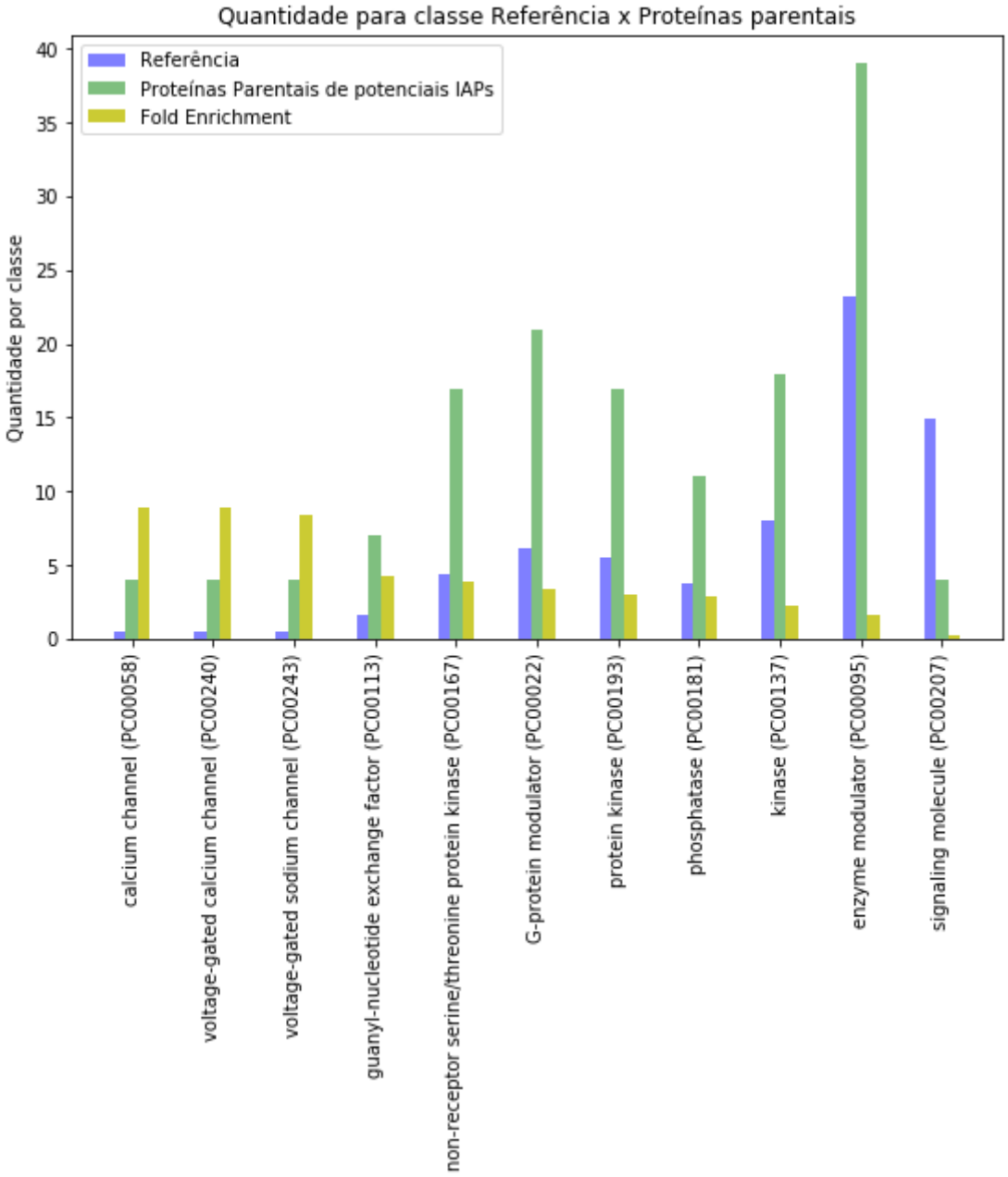


*Figura 11 - Classes de proteínas parentais com diferença significativa para com o proteoma humano. Em azul estão os valores esperados em uma amostragem aleatória do proteoma humano e em verde os valores encontrados analisando-se as proteínas parentais de potenciais IAPs. A barra amarela representa a razão entre as duas outras barras.*

Todas as diferenças representadas graficamente possuem um valor p menor que 0.05, e são consideradas, portanto, estatisticamente significativas.

Dentre as classes encontradas é evidente que dois grupos se destacam: proteínas transmembrana e proteínas envolvidas no metabolismo do grupo fosfato (Tabela 2). As classes são condizentes entre si, não diferindo muito em suas funções.

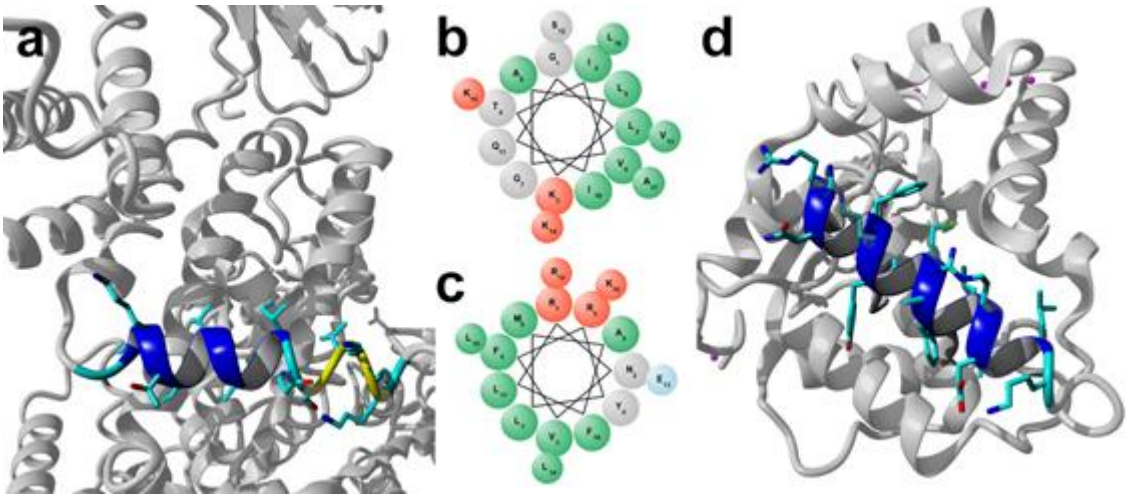


*Figura 12 - a) Detalhe da estrutura da proteína transmembrana Sodium channel protein type 5 subunit alpha (modelo da proteína 6J8E2, com 100% de identidade para a região sob avaliação). Colorida está o segmento anfifílico e catiônico desta, correspondente ao segmento citoplasmático entre as hélices S4 e S5. b) IAP Q14524 (235-251), estrutura primária GLKTIVGALIQSVKKLA. c) IAP Q14524 (833-848), estrutura primária RNVFRYLMAFLRELLK. d) Detalhe da estrutura da proteína 3QIS inositol 5-phosphatase OCRL, uma proteína da classe das fosfatases.*

Potenciais IAPs são encontradas em duas classes de proteínas de canal (de sódio e de cálcio), função conhecida por ser realizada por proteínas transmembrana. Além disso, existem, além de proteínas do tipo quinase de treonina e serina, as quais clivam fosfato e o transferem para as outras, fosfatases cuja função é clivar éteres de fosfato. Um último grupo de grande relevância são as proteínas que modulam a atividade das proteínas G, as quais são responsáveis pela clivagem de grupo fosfato da molécula guanosina trifosfato (GTP) transformando-a em uma guanosina diofosfato (GDP) (Vögler *et al.*, 2008).

É de se se esperar que proteínas transmembrana abriguem segmentos catiônicos anfifílicos. Contudo, vale ressaltar que os segmentos encontrados estão presentes tanto em domínios citosólicos quanto extracelulares e não são parte direta de hélices transmembrana, como ilustrado pelo fragmento da Q14524 (235-251) de estrutura primária GLKTIVGALIQSVKKLA, constituinte do fragmento que liga as hélices S4 e S5 da proteína de canal de sódio de subunidade alfa (SCN5A) (Darbar *et al.*, 2008).

As proteínas G usualmente são responsáveis pela transdução de sinais e sua estrutura é frequentemente descrita como a junção de três domínios protéicos, os quais são separados mediante estímulo adequado, gerando um peptídeo $\alpha$ e um complexo com os outros dois ($\beta$ e $\gamma$), e podem, inclusive, regular canais protéicos, além de serem reconhecidos por diversas outras funções no organismo, inclusive à de quinase (outro subgrupo encontrado como bastante pronunciado) (Venetia

Zachariou, Ronald S. Duman, 2012). Em suma, as proteínas G estão relacionadas com os dois grandes grupos encontrados.

*Tabela 2 - Dados obtidos pelas classes do Panther sintetizados. É possível ver claramente a presença de grandes grupos específicos: proteínas que interagem com membrana e proteínas relacionadas com metabolismo de fosfato. As moduladoras de proteínas G inserem-se em ambos os casos.*

| Grupo | Class | Definition | Uniprot Ids | SubClass | Definition |
|---|---|---|---|---|---|
| Proteínas que interagem com membrana | Calcium channel | A transmembrane ion channel whose selective permeability to calcium is sensitive to the transmembrane potential difference. | Q14524<br>P35499<br>Q15858<br>O43497 | | |
| | voltage-gated sodium channel | A transmembrane ion channel whose selective permeability to sodium is sensitive to the transmembrane potential difference. | Q14524<br>P35499<br>Q15858<br>O43497 | | |
| Proteínas que participam da quebra do grupo fosfato | kinase | An enzyme that catalyzes the transfer of a phosphate from ATP to a second substrate (EC2.7). | O15111<br>P45983<br>Q13164<br>Q9Y4A5<br>P27361<br>Q9Y3S1<br>Q86XP1<br>Q96PF2<br>O15264<br>Q9H093<br>Q9UHD2<br>Q9UKI8<br>Q86UE8<br>Q96PN8<br>P78362<br>Q96Q15<br>O14920<br>Q13315 | non-receptor serine/threonine protein kinase | A soluble protein catalyzing transfer of phosphate from ATP to serine or threonine residue. |
| | phosphatases | An enzyme that hydrolyzes phosphomonoesters. | Q86XP1<br>P30154<br>Q86WG5<br>O15297<br>Q01968<br>O95861<br>Q9Y5P8<br>P32019<br>Q06190<br>Q9Y6X5<br>Q9ULR3<br>O95248 | | |
| Proteínas relacionadas a proteínas G | G-protein modulator | A protein that directly interacts with a G-protein and affects its activity. | Q9BZ29<br>Q13905<br>Q0VAM2<br>O75038<br>Q8IV61<br>Q07889<br>Q07890 | guanyl-nucleotide exchange factor | An enzyme that catalyzes the exchange of GDP and GTP in a G-protein. |
| | | | Q5TG30<br>Q13563<br>Q13563<br>O60347<br>Q2PPJ7<br>Q8TCX5<br>Q9BRR9<br>Q5R372<br>Q9Y4I1<br>Q9Y4I1<br>Q96H55<br>Q9H0H5<br>Q9Y3P9<br>Q9UJF2<br>O94832 | | |

Os resultados também reforçam a ideia de que segmentos anfifílicos catiônicos são relevantes a interação com grupamentos fosfato - talvez os fragmentos catiônicos estabilizem a carga negativa presente no grupo fosfato e

consigam imobilizá-lo em uma boa orientação em proteínas fosfatases. Há diversas evidências sobre a modulação de proteínas G a partir de peptídeos antimicrobianos – os quais possuem as mesmas características que os IAPs (Pundir e Kulka, 2010). Também já é conhecido que peptídeos antimicrobianos são capazes, de fato, de se ligar com o moléculas iônicas como o próprio ATP (Hilpert *et al.*, 2010). Isto indica que os fragmentos podem competir com os substratos das proteínas G, dando origem a sua atividade moduladora.

Outra explicação para esta regulação é o fato de as interações das proteínas G sempre se darem na vizinhança de membranas, sendo possível que fragmentos catiônicos anfifílicos nessas proteínas moduladoras compitam por sítios na membrana, liberando a forma solúvel de alguma subunidade da proteína G e, portanto, modificando sua eficiência.

A única classe pouco pronunciada encontrada são as moléculas sinalizadoras, as quais transmitem sinais entre células, basicamente por modificar um receptor ligando-se com o mesmo. Este é um resultado bastante interessante e ao mesmo tempo inesperado ao considerar que proteínas sinalizadoras são secretadas e seriam alternativas lógicas para o abrigo de moléculas catiônicas e anfifílicas encriptadas.

### 5.2.2 Funções biológicas das proteínas parentais

Analisando-se as funções biológicas das proteínas parentais (figura 13), novamente estão presentes funções relacionadas às proteínas transmembrana (como transporte de íons) e fosfatases (como processos biológicos contendo fosfato), além de um grande destaque que são funções relacionadas com docking vesicular. Todas as proteínas do proteoma humano catalogadas pelo Gene Ontology como responsáveis pelo processo de priming de grânulos vesiculares, as proteínas homólogas unc-13 A, B and C (Q8NB66, O14795 e Q9UPW8, respectivamente) (Thomas C. Sudhof, 2011) contém segmentos catiônicos anfifílicos compatíveis com IAPs. Outras moléculas relacionadas com exocitose, mais especificamente com docking vesicular, também estão presentes, como os componentes 6 e 6B do complexo exocisto: Q8TAG9 e Q9Y2D4. Estes processos envolvem a fusão de membranas em neurônios pré-sinápticos e pode-se hipotetizar que fragmentos catiônicos anfifílicos destas proteínas são importantes no processo

de priming de vesículas sinápticas. É possível notar também como a função de transdução de sinais está bem pronunciada nas proteínas parentais, independente de moléculas essencialmente sinalizadoras não estarem presentes nas classes. Uma provável explicação é justamente a alta presença de proteínas reguladoras de proteínas G, as quais, como já citado, tem como uma das funções principais a sinalização em processos cerebrais (Venetia Zachariou, Ronald S. Duman, 2012).

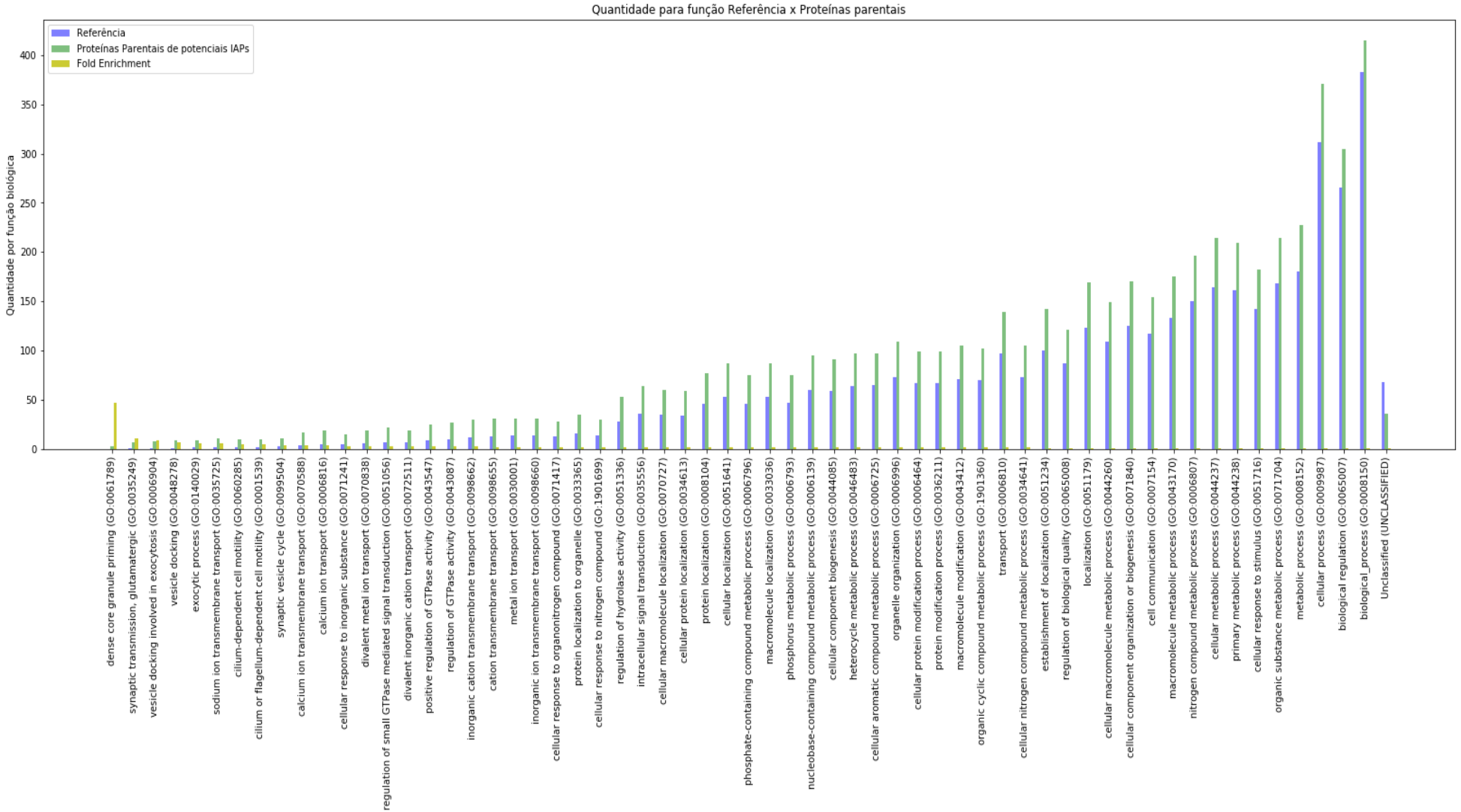


*Figura 13 - Funções biológicas de proteínas parentais com diferença significativa para com o proteoma humano. Em azul estão os valores esperados em uma amostragem aleatória do proteoma humano e em verde os valores encontrados analisando-se as proteínas parentais de potenciais IAPs. A barra amarela representa a razão entre as duas outras barras.*

## 5.3 ANÁLISE DE CONSERVAÇÃO DOS SEGMENTOS COMPATÍVEIS COM IAPS EM RELAÇÃO À PROTEÍNA PARENTAL

A análise de conservação de conservação dos potenciais IAPs mostrou-se complicada por estar condicionada a dados experimentais de estrutura tridimensional das proteínas que incluem os fragmentos desejados, não sendo possível realizá-la com muitas proteínas. Das 458 proteínas, apenas 172 possuem estrutura catalogada no banco de dados PDB e destas poucas possuíam o potencial IAP dentro do domínio cristalizado. Das 100 proteínas exploradas até o momento, os IAPs estavam presentes em cerca de 30.

A análise foi feita calculando-se o score de conservação pelo algoritmo RATE4SITE através da ferramenta consurf e os RSA foram obtidos através da normalização dos ASA gerados pelo pdbepisa. Dois gráficos foram feitos: um de posição pelo score de conservação, com o objetivo de encontrar posições mais bem conservadas que as demais e outro, um boxplot entre intervalos de RSA (cerca de 33% em cada) e scores de conservação, visando verificar se um modelo capaz de prever a conservação apenas pela exposição do resíduo de aminoácido é o suficiente para explicar a conservação dos resíduos de aminoácidos presentes no potencial IAP. Em ambos os casos, o IAP fora colorido para diferenciar-se dos demais pontos.

### 5.3.3 Os resíduos de aminoácidos dos IAPs são, em geral, tão conservados quanto outros resíduos de mesmo RSA dentro de cada proteína.

De maneira geral não foram encontrados fragmentos com um grau de conservação maior que o esperado pela previsão do RSA, não se tendo uma evidência inicial de que eles desempenham um papel diferenciado dos demais segmentos na estrutura/função da proteína parental. Assim, pode-se especular que resíduos de aminoácidos destes segmentos não sofrem pressões seletivas adicionais em relação a outros da mesma proteína, as quais seriam passíveis de especulação em uma eventual hidrólise seguida da efetuação de uma função de valor adaptativo ao organismo. Contudo, é importante considerar que não foram obtidos dados o suficiente para fazer uma amostragem significativa e diversa dentro das proteínas encontradas pelo Kamal, não sendo possível afirmar que o padrão visto irá repetir-se ou extrapolar os resultados.

Algumas proteínas analisadas foram a tRNA wybutosine-sintetizadora (A2RUC4), a Importin-13 (O94829) e a quinase ativada por mitógeno (P45983), as quais demonstram que o RSA é o suficiente para prever conservação dos IAPs (Figura 14):

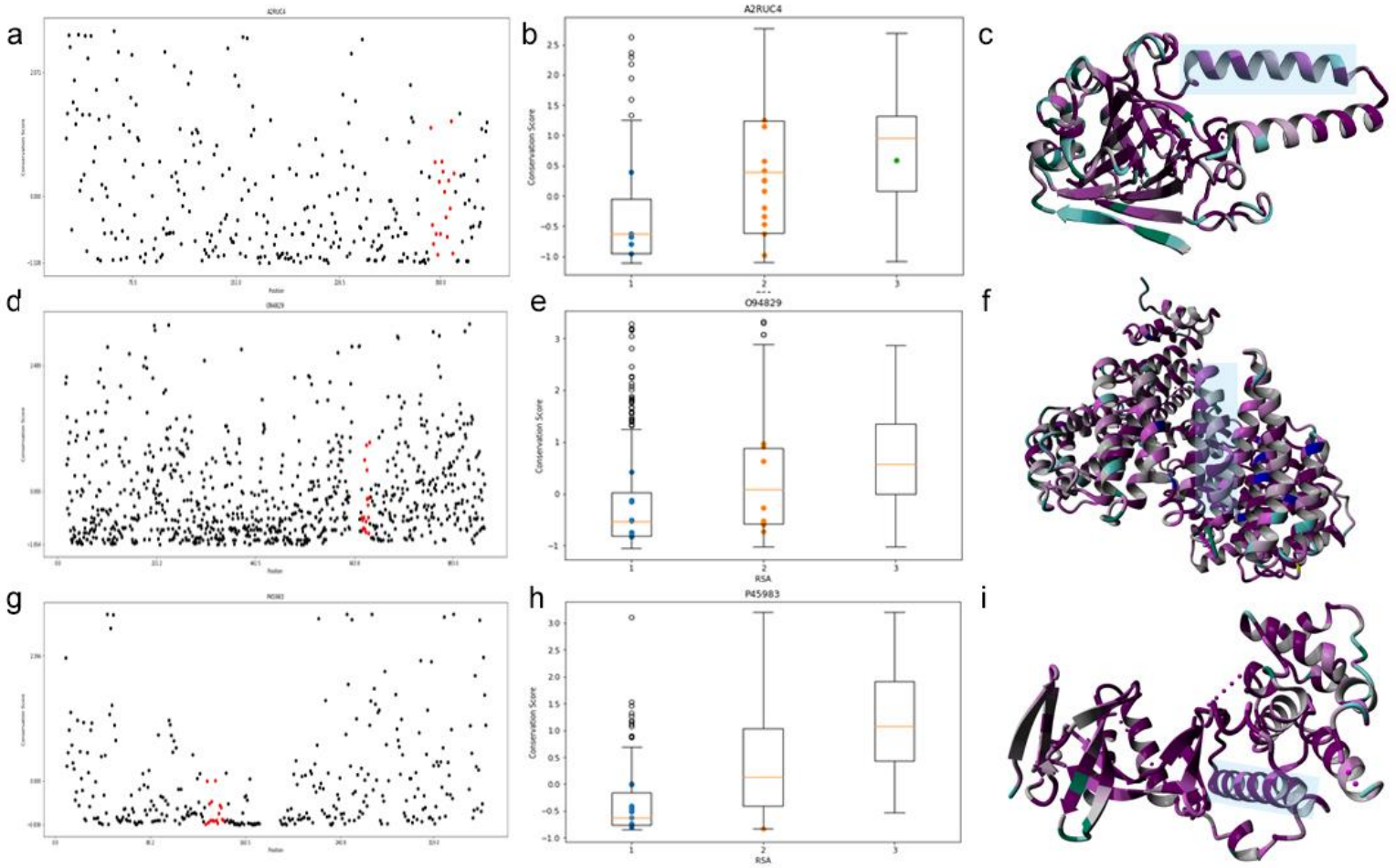


*Figura 14 -a), d) e g) Análises de perfil evolutivo (score de conservação x posição) nas proteínas A2RUC4, Q9429 e P45983. O fragmento potencial IAP está colorido;; b) e) e g) Análises de score de conservação em função da acessibilidade ao solvente das proteínas A2RUC4, Q9429 e P45983. O fragmento potencial IAP está colorido; c) e) e f) fragmentos potenciais IAPs AASRAAQILDRALKTLA, VLQQVFQLIQKVLSKWLN e RMSYLLYQMLCGIKHL destacados na estrutura tridimensional de suas proteínas parentais A2RUC4, Q9429 e P45983 respectivamente.*

Considerando haver certa variabilidade nos critérios utilizados, é de fundamental importância também notar que diferentes fragmentos bastante diversos ainda possuiriam a mesma característica físico-química (alfas hélices anfifílicas), sendo possível que eles de fato tenham funcionalidade na proteína parental, mas ainda assim não sejam bem conservados.

### 5.3.2 A proteína Exportin-1 (O14980)

Dentre todas as proteínas estudadas, a que mais chamou atenção foi a proteína Exportin-1 (O14980) por constituir uma exceção. Essa está presente na membrana nuclear e regula a exportação de proteínas nucleares através de sinais de exportação (Fornerod *et al.*, 1997), além de desempenhar um papel fundamental em infecção viral e ser atacada pela leptomycin B (Hakata, Yamada e Shida, 2003). O potenciai IAP WKFLKTVVNKLFEFMH está em uma região altamente conservada (figura 15) (Central conserved region), bem próxima do resíduo de aminoácido Cys529 – o qual é modificado pela leptomycin B, causando a desativação da proteína (Nobuaki Kudo et al 1999).

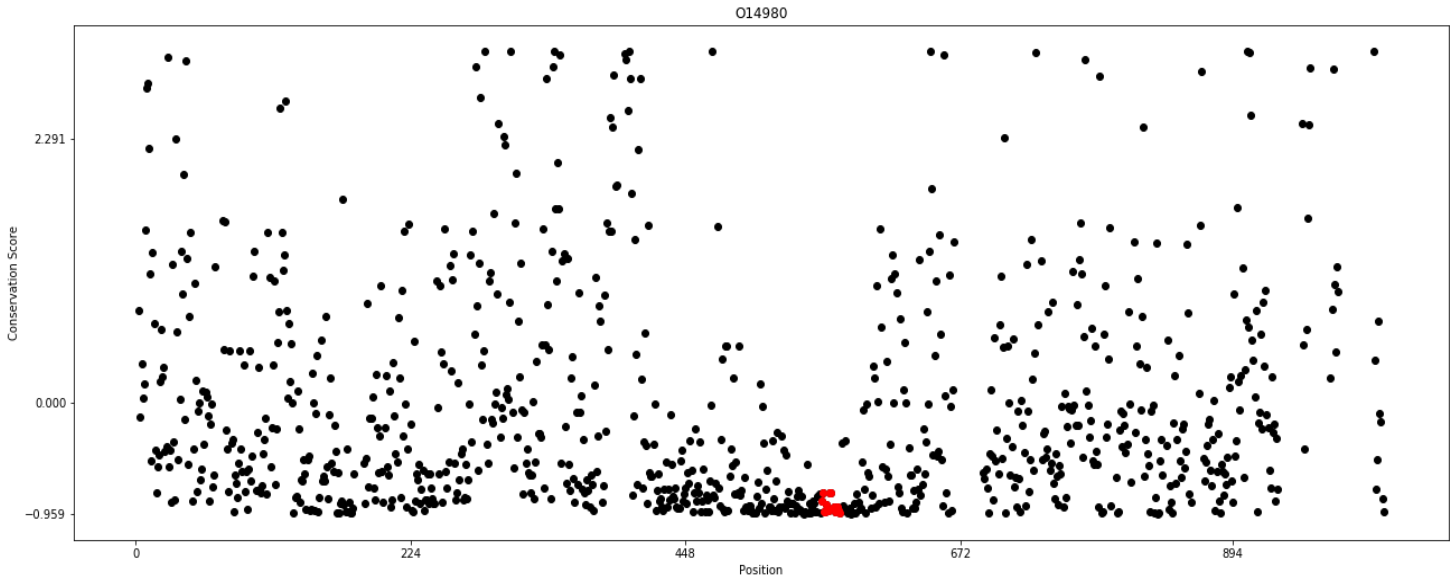


*Figura 15 - Análise de perfil evolutivo (score de conservação x posição) na proteína O14980. O fragmento potencial IAP está destacado em vermelho e encontra-se em uma região de alta de conservação da proteína.*

Com isto, há boas evidências de que o fragmento encontrado pelo Kamal é fundamental para o funcionamento da proteína, embora não existam relatos na literatura. Analisando a conservação do mesmo segmento em função do RSA, é possível ver que os resíduos de aminoácidos que compõem este fragmento são

melhor conservados que os demais mesmo após a correção pelo seu RSA (Figura 16).

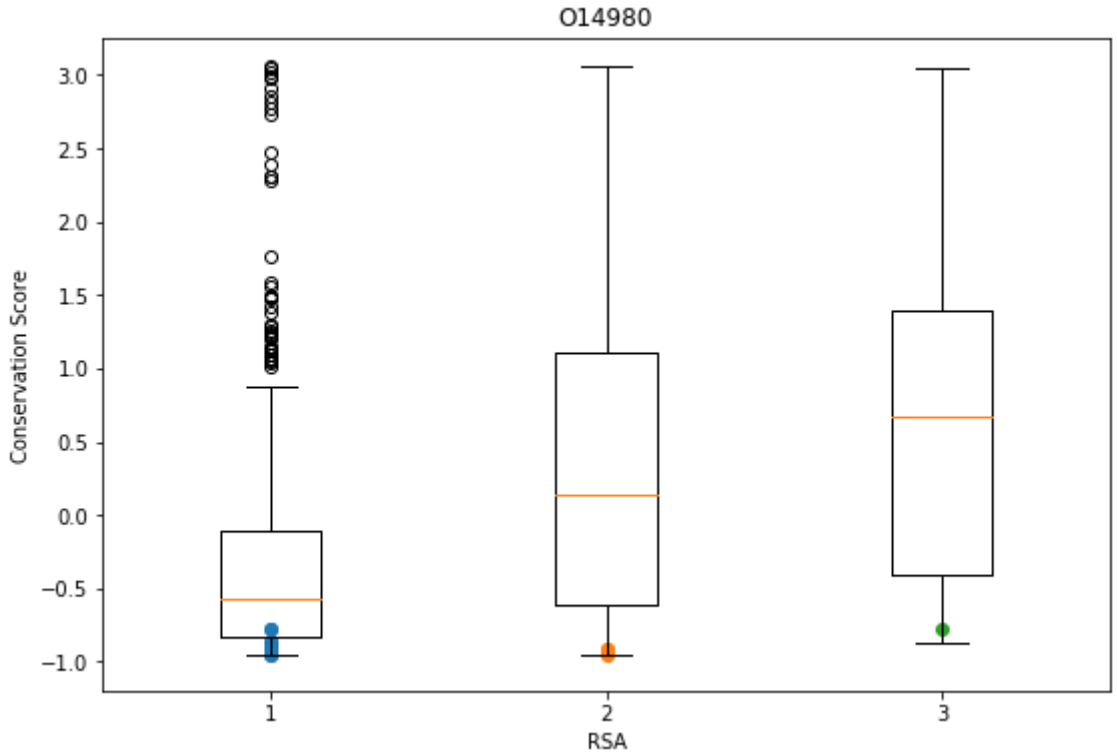


*Figura 16 - Análise de RSA por score de conservação da proteína parental O14980. É notável que o fragmento potencial possui resíduos de aminoácidos expostos, e ainda assim encontram-se bem conservados. Há um bom indicativo de que o fragmento potencial IAP, portanto, possui alguma função relevante à proteína parental..*

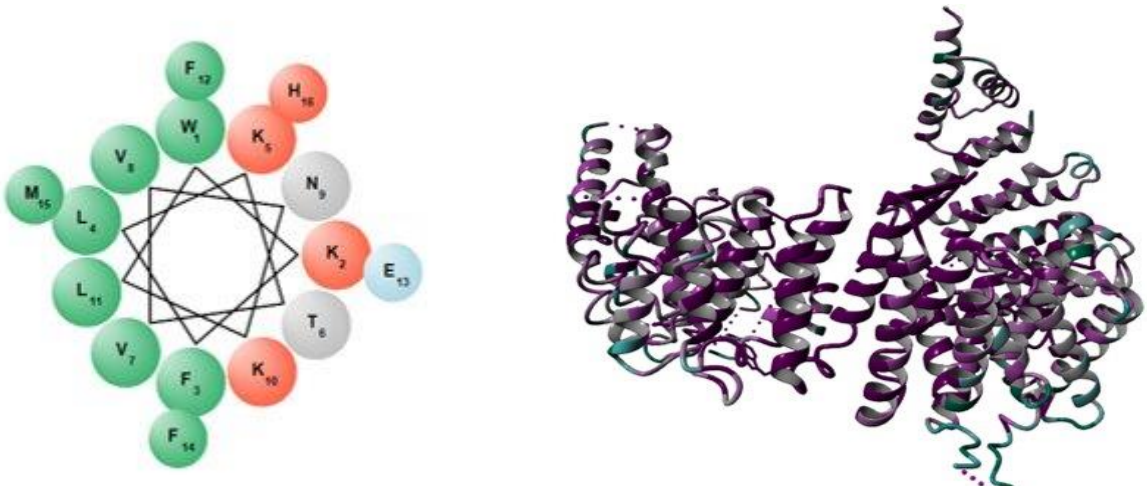

*Figura 17- a) Fragmento potencial IAP encontrado em representação de roda helicoidal e b) Estrutura da proteína parental O14980 colorida baseando-se no score de conservação advindo do RATE4SITE.*

## 6 CONCLUSÃO

Os resultados adquiridos até o momento mostraram que o presente trabalho possui potencial de avanço, tendo caráter preliminar e abrindo caminho para diversas outras perguntas, embora não se tenha sido possível responder todos os questionamentos inicialmente propostos. A pesquisa indicou que proteínas associadas a interação com membranas, como canais iônicos e proteínas envolvidas no amadurecimento e liberação de vesículas sinápticas, são ricas em segmentos catiônicos e anfifílicos compatíveis com IAPs. Além disso, moléculas envolvidas no metabolismo do fosfato, como moduladores de proteínas G também apresentam segmentos com essas características.

Também foi possível observar que o grau de conservação de resíduos destes segmentos não é maior que o previsto por sua acessibilidade ao solvente, entretanto, existem exceções, como a proteína Exportin-1. Isto levanta o questionamento acerca de quantas outras não existiriam e mesmo como a O14980 se comporta. Seria o fragmento de fato liberado? E, caso seja, será que o mesmo possui atividade antimicrobiana?

Por fim, os resultados aparentemente não corroboram com a hipótese inicial do projeto - a de que peptídeos antimicrobianos teriam sido usado como bloco de construção para peptídeos maiores durante a evolução e manteriam sua função - pois, para tanto, seria esperado nenhuma correlação entre as funcionalidades das proteínas parentais e muitos acertos pelo Kamal e, portanto, também não sustentam a hipótese de Lim (Lee e Parker, 2011) sobre hélices anfifílicas serem intermediários globais de dobramento, a qual poderia ser pronunciada pelo mesmo resultado. Não é possível afirmar muito, pois é possível que os critérios utilizados pelo Kamal tenham sido conservadores demais, ou que o algoritmo ainda não esteja sensível o suficiente, não conseguindo achar todos os fragmentos com hélices anfifílicas.

## 7 REFERÊNCIAS BIBLIOGRÁFICAS


BJÖRSTAD, Å. et al. Interleukin-8-derived peptide has antibacterial activity. **Antimicrobial Agents and Chemotherapy**, v. 49, n. 9, p. 3889–3895, 2005.

BRAND, G. D. et al. Intragenic antimicrobial peptides ( IAPs ) from human proteins with potent antimicrobial and anti-inflammatory activity. **PLos ONE,** p. 1–20, 2019.

BRAND, G. D. et al. Probing Protein Sequences as Sources for Encrypted Antimicrobial Peptides. **PLoS ONE**, v. 7, n. 9, 2012.

CARRILLO, H.; LIPMAN, D. The Multiple Sequence Alignment Problem in Biology. **SIAM Journal on Applied Mathematics**, v. 48, n. 5, p. 1073–1082, 2005.

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

CHEN, E. et al. Short-Lived α-Helical Intermediates in the Folding of β-Sheet Proteins. **Biochemistry**, v. 49, n. 26, p. 5609–5619, 2010.

CHENNA, R. et al. Multiple sequence alignment with the Clustal series of programs. **Nucleic Acids Research**, v. 31, n. 13, p. 3497–3500, 2003.

DARBAR, D. et al. Cardiac Sodium Channel ( SCN5A ) Variants Associated with Atrial Fibrillation. Circulation, v. 117, n. 15, 2008.

David L. Nelson, Michael M. Cox. **Princípios de bioquímica de Lehninger**. 6ª edição, artmed, 2014.

DEAMER, D. W. Role of amphiphilic compounds in the evolution of membrane structure on the early earth. **Origins of Life and Evolution of the Biosphere**, v. 17, n. 1, p. 3–25, 1986.

ECHAVE, J.; SPIELMAN, S. J.; WILKE, C. O. Causes of evolutionary rate variation among protein sites. **Nature Reviews Genetics**, v. 17, n. 2, p. 109–121, 2016.

EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic acids research**, v. 32, n. 5, p. 1792–7, 2004.

FENG, H. et al. Block copolymers: Synthesis, self-assembly, and applications. **Polymers**, v. 9, n. 10, 2017.

FENNEMA, O. R. **Food Chemistry.** 3ª edição, MARCEL DEKKER, 1996.

FORNEROD, M. et al. CRM1 Is an Export Receptor for Leucine-Rich. v. 90, p. 1051–1060, 1997.

HAKATA, Y.; YAMADA, M.; SHIDA, H. A Multifunctional Domain in Human CRM1 ( Exportin 1 ) Mediates RanBP3 Binding and Multimerization of Human T-Cell Leukemia Virus Type 1 Rex Protein. v. 23, n. 23, p. 8751–8761, 2003.

HAYES, M. et al. Casein-derived antimicrobial peptides generated by Lactobacillus acidophilus DPC6026. **Applied and Environmental Microbiology**, v. 72, n. 3, p. 2260–2264, 2006.

HILPERT, K. et al. Short Cationic Antimicrobial Peptides Interact with ATP □. v. 54, n. 10, p. 4480–4483, 2010.

KATOH, K. et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. **Nucleic acids research**, v. 30, n. 14, p. 3059–66, 2002

KRISSINEL, E.; HENRICK, K. Inference of Macromolecular Assemblies from Crystalline State.

**Journal of Molecular Biology**, v. 372, n. 3, p. 774–797, 2007.

LEE, S. Y. R.; PARKER, W. Amphiphilic α-helical potential: A putative folding motif adding few constraints to protein evolution. **Journal of Molecular Evolution**, v. 73, n. 3–4, p. 166–180, 2011.

LIPMAN, D. J.; ALTSCHUL, S. F.; KECECIOGLUT, J. D. A tool for multiple sequence alignment (proteins/structure/evolution/dynamic programming). **Proc. Nail. Acad. Sci. USA**, v. 86, n. June, p. 4412–4415, 1989.

MARCELO HENRIQUE SOLLER RAMADA, CARLOS BLOCH JR, G. D. B. Explorando genomas : a busca por peptídeos antimicrobianos intragênicos Explorando genomas : a busca por peptídeos antimicrobianos intragênicos. 2016.

MASSINGHAM, T.; GOLDMAN, N. Detecting amino acid sites under positive selection and purifying selection. **Genetics**, v. 169, n. 3, p. 1753–1762, 2005.

MAYROSE, I. et al. Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. **Molecular Biology and Evolution**, v. 21, n. 9, p. 1781–1791, 2004.

MEISEL, H. Multifunctional peptides encrypted in milk proteins. **BioFactors**, v. 21, n. 1–4, p. 55–61, 2004.

Mi, H. et al. PANTHER version 14: more genomes , a new PANTHER GO-slim and improvements in enrichment analysis tools. v. 47, n. November 2018, p. 419–426, 2019.

MÖLLER, N. P. et al. Bioactive peptides and proteins from foods: Indication for health effects. **European Journal of Nutrition**, v. 47, n. 4, p. 171–182, 2008.

MURZYN, K.; PASENKIEWIECZ-GIERULA, M. Construction of a toroidal model for the magainin pore. **Journal of Molecular Modeling**, v. 9, n. 4, p. 217–224, 2003.

NOBUAKI KUDO, NOBUAKI MATSUMORI, HIROSHI TAOKA, DAISUKE FUJIWARA, ERWIN P. SCHREINER, BARBARA WOLFF, MINORU YOSHIDA, S. H. Leptomycin B inactivates CRM1 ¯ exportin 1 by covalent modification at a cysteine residue in the. Cell Biology, v. 96, n. August, p. 9112–9117, 1999.

PEARSON, W. R. An Introduction to Sequence and Series (“Homology”) Searching. **Curr Protoc bioinformatics**, v. 1, n. 10, p. 1286–1292, 2013.

PERACCHIA, M. T. et al. Synthesis of a Novel Poly(MePEG cyanoacrylate- co -alkyl cyanoacrylate) Amphiphilic Copolymer for Nanoparticle Technology ‡ . **Macromolecules**, v. 30, n. 4, p. 846–851, 2002.

PUNDIR, P.; KULKA, M. The role of G protein-coupled receptors in mast cell activation by antimicrobial peptides : is there a connection?, **Immunology and Cell Biology**, v. 88, n. 6, p. 632–640, 2010.

RAMADA, M. H. S. et al. Encrypted Antimicrobial Peptides from Plant Proteins. **Scientific Reports**, v. 7, n. 1, p. 1–14, 2017.

SÖDING, J.; LUPAS, A. N. More than the sum of their parts: On the evolution of proteins from peptides. **BioEssays**, v. 25, n. 9, p. 837–846, 2003.

Sudhof, T. C.; Rizo, J. Synaptic Vesicle Exocytosis. **Cold Spring Harbor Laboratory Press,** 2011.

TOPORIK, A. et al. Computational identification of natural peptides based on analysis of molecular evolution. BIOINFORMATICS, v. 30, n. 15, p. 2137–2141, 2014.

UEMATSU, N.; MATSUZAKI, K. Polar angle as a determinant of amphipathic α-helix-lipid interactions: A model peptide study. **Biophysical Journal**, v. 79, n. 4, p. 2075–2083, 2000.

VENETIA ZACHARIOU, RONALD S. DUMAN, E. J. N. Basic Neurochemistry.

VOIGHT, B. F. et al. A map of recent positive selection in the human genome. **PLoS biology**, v. 4, n. 3, p. e72, 2006.

VÖGLER, O. et al. Membrane interactions of G proteins and other related proteins. **Biochimica et Biophysica**, v. 1778, p. 1640–1652, 2008.

YEAMAN, M. R.; YOUNT, N. Y. Mechanisms of Antimicrobial Peptide Action and Resistance. **Pharmacological Reviews**, v. 55, n. 1, p. 27–55, 2003.