Universidade Federal do Rio Grande do Sul
Instituto de Biociências
Curso de Graduação em Biotecnologia

Maiara Kolbe Musskopf

Filogenia de proteínas da subfamília DNAJ/HSP40 e estudo da organização dos membros das subclasses A, B e C

Porto Alegre
2018

Maiara Kolbe Musskopf

Filogenia de proteínas da subfamília DNAJ/HSP40 e estudo da organização dos membros das subclasses A, B e C

Trabalho de conclusão de curso de graduação apresentado ao Instituto de Biociências da Universidade Federal do Rio Grande do Sul como requisito parcial para a obtenção do título de Bacharela em Biotecnologia.

Área de habilitação: Bioinformática

Orientador: Prof. Dr. Hugo Verli

Porto Alegre
2018

Resumo

A família de chaperonas moleculares DNAJ/HSP40 é uma das maiores e mais complexas subfamílias de *Heat Shock Proteins* (HSPs) descobertas até hoje, e seus membros são encontrados desde bactérias até humanos. A assinatura desta subfamília é o domínio-J, o qual é uma região de aproximadamente 70 aminoácidos bastante conservada evolutivamente entre diversos organismos e é fundamental para a maquinaria HSP70 das células. Apesar de compartilharem o domínio-J, estas proteínas são divididas em 3 subclasses (A, B ou C) de acordo com um sistema de classificação baseado na presença ou ausência de outras regiões, levando ao agrupamento de proteínas funcionalmente similares no caso das subclasses A e B, mas muito heterogêneas no caso da subclasse C. De forma interessante, a quantidade de proteínas DNAJC em relação às proteínas DNAJA e DNAJB é significativamente maior, representando 78.34% da família DNAJ proveniente do banco de dados utilizado neste trabalho, enquanto as subclasses A e B representam 13.66% e 8.00%, respectivamente. Esta proporção destoante e a considerável heterogeneidade desta família instigam questionamentos referentes à relação entre as subclasses de DNAJ em diferentes organismos. Apesar dos estudos evolutivos voltados para a família DNAJ explorarem a diversidade de alguns dos seus membros em uma ou mais espécies, análises filogenéticas abrangendo as três subclasses ao longo da evolução ainda precisam ser exploradas. Desta forma, o presente estudo visa analisar a organização das proteínas DNAJ das subclasses A, B e C de diversos organismos através de uma abordagem filogenética utilizando um banco de dados de HSPs manualmente curado. Portanto, diversas reconstruções filogenéticas foram geradas a partir do alinhamento de toda a extensão da proteína ou da região correspondente ao domínio-J, de acordo com o conjunto de dados. Foi possível observar que um pequeno grupo de proteínas da subclasse C é pontualmente distinto dos outros membros da família DNAJ em relação à estrutura primária do domínio-J, permanecendo agrupado nas diferentes árvores. Além disso, a filogenia das 3 subclasses reunidas demonstrou uma possível divisão evolutiva da maioria dos membros das subclasses A e B em relação à maioria dos membros da subclasse C.

Plavras-chave: DNAJ/HSP40, chaperonas moleculares, heterogeneidade, filogenia

Abstract

The DNAJ/HSP40 family of molecular chaperones is one of the largest and more complex subfamilies of Heat Shock Proteins (HSPs) known until today, and its members are found from bacteria to humans. The signature of this subfamily is the J-domain, which is a 70 amino acid region evolutionarily conserved througout many organisms and is fundamental for the celular HSP70 machinery. Although these proteins share the J-domain, they are divided in 3 subclasses (A, B or C) according to a classification system based on the presence or absence of other regions, leading to the grouping of functionally similar proteins in the case of subclasses A and B, but very hetereogenous in the case of subclass C. Interestingly, the amount of DNAJC compared to DNAJA and DNAJB is significantly higher, representing 78.34% of the DNAJ family present in the database employed on this work, while subclasses A and B represent 13.66% e 8.00%, respectively. This disparate proportion and the high heterogeneity of this family instigate questionings regarding the relationship among the subclasses from different organisms in the evolutionary scale. Although evolutionary studies aimed at DNAJ family explore the diversity of some members in one or more species, phylogenetic analyses encompassing the three subclasses must be explored. Thus, the present study aims to analyze the organization of DNAJ proteins from subclasses A, B and C from multiple organisms through a phylogenetic approach using a manually cured HSPs database. Therefore, several phylogenetic reconstructions were performed from alignments corresponding to either the whole protein extension or the J-domain region, according to data set. It was observed that a small group of subclass C proteins is strictly distinct from the other members of DNAJ family regarding the primary structure of J-domain, remaining grouped in the different trees. Moreover, the phylogeny of all 3 subclasses demonstrated a possible evolutionary split of most members from subclasses A and B with respect to the majority of subclass C members.

Key-words: DNAJ/HSP40, molecular chaperones, heterogeneity, phylogeny

Lista de Figuras

1 Modo de ação da maquinaria HSP70, baseado em estudos de re-enovelamento *in vitro* de proteínas desnaturadas. A proteína DNAJ se liga ao substrato (1) e interage com a proteína HSP70 através do domínio-J (2). O substrato interage transientemente com o sítio de ligação peptídica da HSP70 e, junto com a proteína DNAJ, estimula a hidrólise do ATP e causa uma mudança conformacional na HSP70 que estabiliza a interação com o substrato. A proteína DNAJ deixa o complexo (3) e a proteína NEF, a qual possui uma maior afinidade pelo complexo HSP70-ADP, se liga à proteína HSP70 (4). O ADP então se dissocia do complexo através da distorção do domínio de ligação de ADP da HSP70 (5) e, posteriormente, o ATP se liga ao complexo (6). Por fim, o substrato é liberado devido à sua baixa afinidade com HSP70-ATP (7), e se o estado nativo não é alcançado, a proteína DNAJ liga-se novamente às regiões hidrofóbicas expostas do substrato, reiniciando o ciclo (Kampinga & Craig, 2010). 11

2 Estrutura do domínio-J e organização das subclasses da família DNAJ. (A) Representação da estrutura tridimensional do domínio-J da *hs*DNAJB1 (PDB: 1HDJ) composta por quatro α-hélices e uma região de *loop* contendo o motivo HPD. (B) Proteínas da subclasse A possuem um domínio-J N-terminal, seguido de uma região G/F, um domínio dedo-de-zinco e dois domínios C-terminais. A subclasse B é bastante similar à subclasse A, exceto pela ausência do domínio dedo-de-zinco. Por outro lado, a subclasse C possui um domínio-J que não está restrito à porção N-terminal, podendo conter qualquer outra composição de regiões fora do domínio-J (Musskopf *et al.* 2018). 12

3 Porcentagem de conservação dos aminoácidos-consenso do domínio-J correspondentes à α-hélice I (A), α-hélice II (B), região de *loop* (C), α-hélice III (D) e α-hélice IV (E) (Hennessy *et al*, 2000). 14

4 Comparação das estruturas tridimensionais de TIM14 (*Saccharomyces cerevisiae*), TIM16 (*Saccharomyces cerevisiae*) e DnaJ (*Escherichia coli*) com o motivo HPD de TIM14 e DnaJ em amarelo e o motivo DKE de TIM16 em azul (Groll *et al.*, 2006). 15

5 Comparação do número de genes codificando proteínas HSP70 e DNAJ em diferentes organismos (Craig & Marszalek, 2017). 16

6 Fluxograma das etapas de processamento dos dados deste trabalho. 20

7 Representação esquemática do processo de recuperação e curadoria de dados implementado no banco de dados HSPIR (Kumar *et al*, 2012). 21

8 Escores do alinhamento de cada proteína (tipos I, II, III e IV) contra o domínio-J (PF00226) 25
9 Escores do alinhamento de cada proteína (tipos I, II, III e IV) contra o domínio *J-like* de TIM16 (PF03656) 26
10 Árvores filogenéticas das proteínas do tipo I (DNAJA) a partir do alinhamento múltiplo de (A) toda a extensão da proteína ou apenas do (B) domínio-J de proteínas clusterizadas com um valor de *cut-off* de 40%. . . . 28
11 Árvores filogenéticas das proteínas do tipo II (DNAJB) a partir do alinhamento múltiplo de (A) toda a extensão da proteína ou apenas do (B) domínio-J de proteínas clusterizadas com um valor de *cut-off* de 40%. . . . 29
12 Árvore filogenética das proteínas dos tipos I (DNAJA) e II (DNAJB) a partir do alinhamento múltiplo tanto da proteína completa (A) quanto apenas do domínio-J (B) utilizando as sequências clusterizadas com um valor de *cut-off* de 40%. 30
13 Árvore filogenética das proteínas do tipo III e IV (DNAJC) a partir do alinhamento múltiplo apenas do domínio-J utilizando as sequências clusterizadas com um valor de *cut-off* de 40%. 30
14 Árvore filogenética de todas as proteínas (DNAJA, DNAJB e DNAJC) a partir do alinhamento múltiplo apenas do domínio-J utilizando as sequências clusterizadas com um valor de *cut-off* de 40%. As linhas tracejadas indicam o clado que concentra as proteínas das subclasses A e B. 31
15 Árvore filogenética das proteínas da subclasse A (tipo I) a partir do alinhamento múltiplo de toda a extensão da proteína utilizando as sequências clusterizadas com um valor de *cut-off* de 80%. 32
16 Árvore filogenética das proteínas da subclasse B (tipo II) a partir do alinhamento múltiplo de toda a extensão da proteína utilizando as sequências clusterizadas com um valor de *cut-off* de 80%. 33
17 Árvore filogenética das proteínas das subclasses A e B (tipos I e II) a partir do alinhamento múltiplo de toda a extensão da proteína utilizando as sequências clusterizadas com um valor de *cut-off* de 80%. 33
18 Árvore filogenética das proteínas das subclasses A, B e C (tipos I, II, III e IV) a partir do alinhamento múltiplo da região correspondente ao domínio-J utilizando as sequências clusterizadas com um valor de *cut-off* de 80%. As linhas tracejadas indicam o clado que concentra as proteínas das subclasses A e B. 34

Lista de Tabelas

1 Quantidade de proteínas após a respectiva etapa de curagem. *Retirada de proteínas contendo caracteres não-padrão: X, B e Z. 24

Lista de Abreviaturas e Siglas

DNAJA	proteína DNAJ pertencente à subclasse A
DNAJB	proteína DNAJ pertencente à subclasse B
DNAJC	proteína DNAJ pertencente à subclasse C
HSP	*Heat Shock Protein* ou proteína de choque térmico
hs	*Homo sapiens*
NEF	*Nucleotide Exchange Factor*
motivo HPD	motivo histidina-prolina-ácido aspártico
região G/F	região rica nos aminoácidos glicina e fenilalanina
CTDI	domínio C-terminal I
CTDII	domínio C-terminal II
CXXCXGXG	motivo encontrado no domínio dedo-de-zinco
motivo DKE	motivo ácido aspártico-lisina-ácido glutâmico
HMM	*Hidden Markov Model*
L-INS-i	algoritmo de alinhamento local do *software* MAFFT

Sumário

Lista de Figuras **4**

Lista de Tabelas **6**

Lista de Abreviaturas e Siglas **7**

1 Introdução **10**
1.1 HSP40 . . . 10
1.2 Maquinaria HSP70 . . . 10
1.3 Classificação . . . 11
1.3.1 DNAJA . . . 12
1.3.2 DNAJB . . . 12
1.3.3 DNAJC . . . 13
1.4 Diversidade das DNAJs . . . 15
1.5 Evolução da Família DNAJ . . . 16

2 Justificativa **18**

3 Objetivos **19**
3.1 Objetivo geral . . . 19
3.2 Objetivos Específicos . . . 19

4 Metodologia **20**
4.1 Coleta dos dados . . . 20
4.1.1 Plataforma HSPIR . . . 20
4.2 Curagem e clusterização dos dados . . . 21
4.3 *Hidden Markov Model* (HMM) . . . 22
4.3.1 Alinhamento contra perfis HMM . . . 22
4.4 Alinhamento múltiplo de sequências . . . 22
4.5 Curagem do alinhamento múltiplo . . . 23
4.6 Filogenia . . . 23
4.7 Visualização e manipulação da árvore . . . 23

5 Resultados **24**
5.1 Alinhamentos contra os perfis HMM . . . 24
5.2 Filogenia . . . 27
5.2.1 Clusterização dos dados com um *cut-off* de 80% . . . 32

6 Discussão **35**

7 Conclusões e Perspectivas **38**

Referências **39**

Apêndices **46**

A
Diferentes táxons após clusterização com *cut-off* de identidade de 40% **46**

B
Diferentes táxons após clusterização com *cut-off* de identidade de 80% **46**

C
Alinhamento múltiplo de proteínas dos tipos I ou II **47**

D
Alinhamento múltiplo de proteínas dos tipos III e IV **48**

E
Alinhamento múltiplo de proteínas dos tipos I, II, III e IV **49**

F
Politomias presentes nas árvores provenientes do alinhamento de toda a extensão da proteína **50**

G
Politomias presentes nas árvores provenientes do alinhamento do domínio-J **51**

1 Introdução

1.1 HSP40

A subfamília HSP40, também conhecida como DNAJ, é uma subdivisão do grande grupo de *Heat Shock Proteins* (HSPs), o qual contém proteínas com função de chaperonas moleculares que conduzem o sistema de controle de qualidade proteico das células [1] [2]. Atualmente, as HSPs são divididas em seis subfamílias (*small* HSPs, HSP40, HSP60, HSP70, HSP90 e HSP100), e apesar de terem sido descobertas e adquirido fama por possuirem uma expressão relacionada ao choque térmico e outros tipo de estresse [3], subsequentes descobertas levaram ao estudo mais abrangente dos papéis das HSPs como chaperonas moleculares [4].

Hoje já se sabe que estas proteínas desempenham os mais variados papéis, até mesmo em células que não estão sob uma condição de estresse. Suas funções variam desde assistir ao enovelamento e montagem de novos polipeptídeos [5], re-enovelamento de proteínas maduras [6], translocação de proteínas [7], marcação de proteínas para degradação [8] e remodelamento de complexos proteicos [9] [10].

Proteínas da subfamília DNAJ são encontradas desde espécies de bactérias até humanos, e podem ser expressas nos mais variados tecidos e compartimentos subcelulares de organismos mais complexos. Além disso, o tamanho da proteína pode variar desde algumas centenas de aminoácidos (ex: 116 em *hs*DNAJC19 até milhares (ex: 2243 em *hs*DNAJC13), demonstrando o alto nível de diversidade desta subfamília de HSPs [11].

1.2 Maquinaria HSP70

As proteínas da família DNAJ são amplamente conhecidas por atuarem como co-chaperonas da maquinaria HSP70, um componente central do sistema de controle de qualidade proteico que garante a homeostase celular através de ciclos de (re)-enovelamento de proteínas [10]. No modelo canônico deste processo, as proteínas DNAJ são responsáveis por reconhecer e interagir com o substrato-alvo e facilitar a sua transferência para a proteína HSP70, a qual necessita da proteína DNAJ como estimuladora da sua atividade ATPase. Após a seleção do substrato, o complexo DNAJ-substrato induz a hidrólise de ATP que leva à dissociação da DNAJ e formação do complexo de alta afinidade HSP70-substrato. Por sua vez, este novo complexo recruta outra família de co-chaperonas conhecida como *Nucleotide Exchange Factors* (NEFs), as quais são responsáveis por promover a transição ADP-ATP no centro catalítico da proteína HSP70, levando ao desprendimento do substrato [12]. Se, ao final de um ciclo, o substrato não atingiu seu estado nativo de enovelamento, ele pode re-entrar no processo de modo iterativo até que o estado nativo seja alcançado (Figura 1). Entretanto, se após um determinado número de iterações o enovelamento não acontecer, o substrato geralmente é enviado para degradação [13].

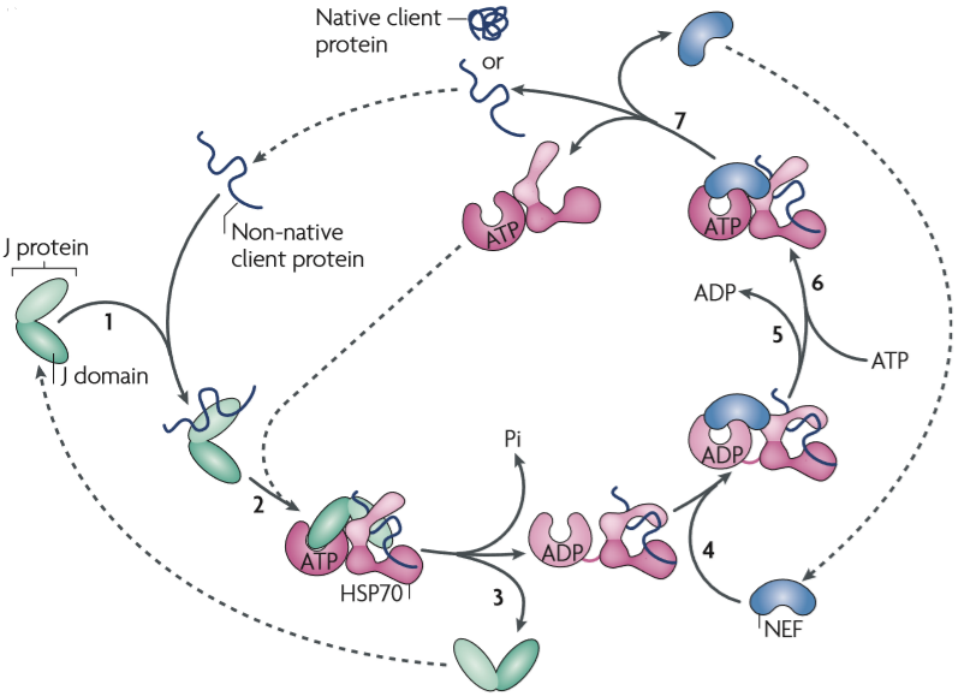


Figura 1: Modo de ação da maquinaria HSP70, baseado em estudos de re-enovelamento *in vitro* de proteínas desnaturadas. A proteína DNAJ se liga ao substrato (1) e interage com a proteína HSP70 através do domínio-J (2). O substrato interage transientemente com o sítio de ligação peptídica da HSP70 e, junto com a proteína DNAJ, estimula a hidrólise do ATP e causa uma mudança conformacional na HSP70 que estabiliza a interação com o substrato. A proteína DNAJ deixa o complexo (3) e a proteína NEF, a qual possui uma maior afinidade pelo complexo HSP70-ADP, se liga à proteína HSP70 (4). O ADP então se dissocia do complexo através da distorção do domínio de ligação de ADP da HSP70 (5) e, posteriormente, o ATP se liga ao complexo (6). Por fim, o substrato é liberado devido à sua baixa afinidade com HSP70-ATP (7), e se o estado nativo não é alcançado, a proteína DNAJ liga-se novamente às regiões hidrofóbicas expostas do substrato, reiniciando o ciclo (Kampinga & Craig, 2010).

1.3 Classificação

Todas as DNAJs compartilham o domínio-J, uma vez que a sua presença é o critério de inclusão da proteína nesta família. Este domínio possui aproximadamente 70 aminoácidos e contém uma região bem conservada conhecida como motivo histidina-prolina-ácido aspártico (HPD). O motivo HPD se encontra num *loop* entre as duas principais α-hélices do domínio e é através dele que ocorre a interação DNAJ-HSP70 mencionada anteriormente [14]. Além disso, a estrutura do domínio-J é conservada não apenas entre os membros de uma mesma espécie, mas através de diversos organismos da escala evolutiva, demonstrando uma notável similaridade tridimensional [15]. Apesar de haver baixa similaridade entre as diversas DNAJs fora do domínio-J, estas proteínas são atualmente divididas em 3 subclasses (A, B e C), de acordo com a presença de outras regiões ao longo da proteína (Figura 2) [16].

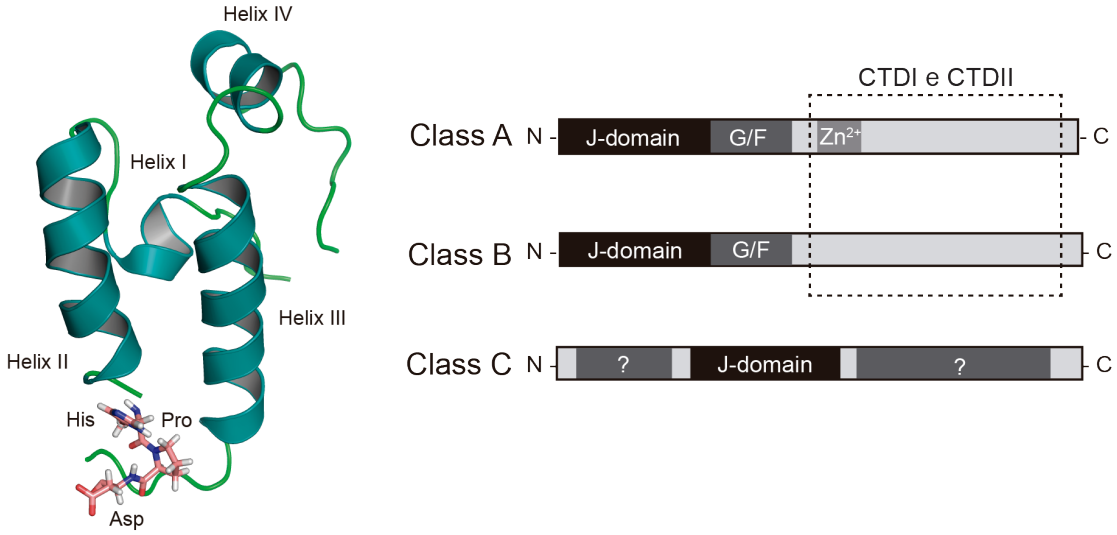


Figura 2: Estrutura do domínio-J e organização das subclasses da família DNAJ. (A) Representação da estrutura tridimensional do domínio-J da *hs*DNAJB1 (PDB: 1HDJ) composta por quatro α-hélices e uma região de *loop* contendo o motivo HPD. (B) Proteínas da subclasse A possuem um domínio-J N-terminal, seguido de uma região G/F, um domínio dedo-de-zinco e dois domínios C-terminais. A subclasse B é bastante similar à subclasse A, exceto pela ausência do domínio dedo-de-zinco. Por outro lado, a subclasse C possui um domínio-J que não está restrito à porção N-terminal, podendo conter qualquer outra composição de regiões fora do domínio-J (Musskopf *et al.* 2018).

1.3.1 DNAJA

Os membros da classe A (ou tipo I) são os homólogos mais próximos da proteína DNAJ em *E. coli* e são compostos por um domínio-J N-terminal, seguido de uma região de 50-100 aminoácidos rica em glicina e fenilalanina (G/F) e dois domínios C-terminais (CTDI e CTDII) que contribuem para a formação de um bolso hidrofóbico que se associa a substratos através do domínio dedo-de-zinco formado por quatro motivos CXXCXGXG no CTDI. A porção C-terminal mais extrema se dobra em uma estrutura de β-folha e forma o domínio de dimerização que aumenta a afinidade da DNAJA pelos seus substratos [17] [18] [10] (Figura 2B).

1.3.2 DNAJB

As proteínas da subclasse B (ou tipo II) também contêm um domínio-J N-terminal e uma região G/F, mas não possuem uma região C-terminal contendo o domínio dedo-de-zinco. Apesar dos tipos I e II serem distintos em sua composição de regiões conservadas, ambos contêm porções C-terminais com um domínio de dimerização que se ligam a peptídeos e funcionam de forma similar na ligação de substratos não-nativos (Figura 2B) [11, 17, 18, 19, 20]. Além disso, as subclasses A e B possuem domínios-J mais conservados quando comparados aos domínios-J das proteínas da subclasse C (Figura 3) [18, 19, 21, 22]. Estudos recentes também demonstram que as proteínas das subclasses A

e B podem interagir fisicamente para aumentar a eficiência da atividade de desagregação de proteínas, comparada à eficiência de desagregação quando estas proteínas agem separadamente [23]. Aparentemente, a formação destes complexos passou a ocorrer na transição evolutiva de procariotos-eucariotos e pode ter levado a consequências funcionais relacionadas a mudanças fisiológicas específicas, demonstrando a adição de um nível de flexibilidade funcional ao sistema de controle de qualidade proteico celular [24] (Figura 1B).

1.3.3 DNAJC

Ao contrário das subclasses A e B, a subclasse C não possui uma região G/F e nem um domínio dedo-de-zinco, portanto o único domínio em comum é o domínio-J, o qual não é restrito à porção N-terminal da proteína. Alguns autores têm adotado a subdivisão desta subclasse em tipos III e IV devido à descoberta de um grupo de proteínas que possuem substituições de aminoácidos no motivo HPD, mas que mantém uma significativa similaridade estrutural com o domínio-J, sendo assim reconhecidas como proteínas *J-like* [18, 20]. O exemplo clássico de proteína *J-like* refere-se à proteína TIM16 (tipo IV) componente do complexo translocase (TIM23), localizado na membrana mitocondrial interna, para mediar a translocação de polipeptídeos precursores (pré-proteínas) para a matriz mitocondrial [25, 26, 27]. Este processo é ATP-dependente, pois envolve ciclos de ligação e liberação dos polipeptídeos importados pela proteína HSP70 mitocondrial (mtHSP70), a qual depende da proteína TIM14 (tipo III) para estimular sua atividade ATPase. Curiosamente, a proteína TIM16 forma um subcomplexo com a proteína TIM14 que é essencial para o recrutamento de TIM14 ao complexo translocase TIM23, e a estrutura cristalográfica deste subcomplexo mostra que ambas proteínas possuem um dobramento virtualmente idêntico, mas superfícies completamente distintas que as permitem desenvolver funções diferentes [25]. Essencialmente, o domínio *J-like* de TIM16 contém as mesmas características do domínio-J, mas possui um motivo DKE no lugar do motivo HPD e não é capaz de estimular a atividade ATPase da proteína HSP70. Apesar de "inativo", é através deste domínio que ocorre a interação estável com TIM14 e serve para posicionar esta proteína corretamente e garantir sua atividade ótima [25, 26, 27]. Ainda não existe um consenso a respeito da subdivisão da subclasse C entre tipos III e IV e até mesmo a respeito da inclusão de proteínas *J-like* na família DNAJ, mas neste trabalho será adotada a classificação no formato DNAJA (tipo I), DNAJB (tipo II) e DNAJC (tipos III e IV). A Figura 4 evidencia a similaridade das estruturas tridimensionais do domínio-J de TIM14 (tipo III) e DnaJ (tipo I) e do domínio *J-like* de TIM16 (tipo IV).

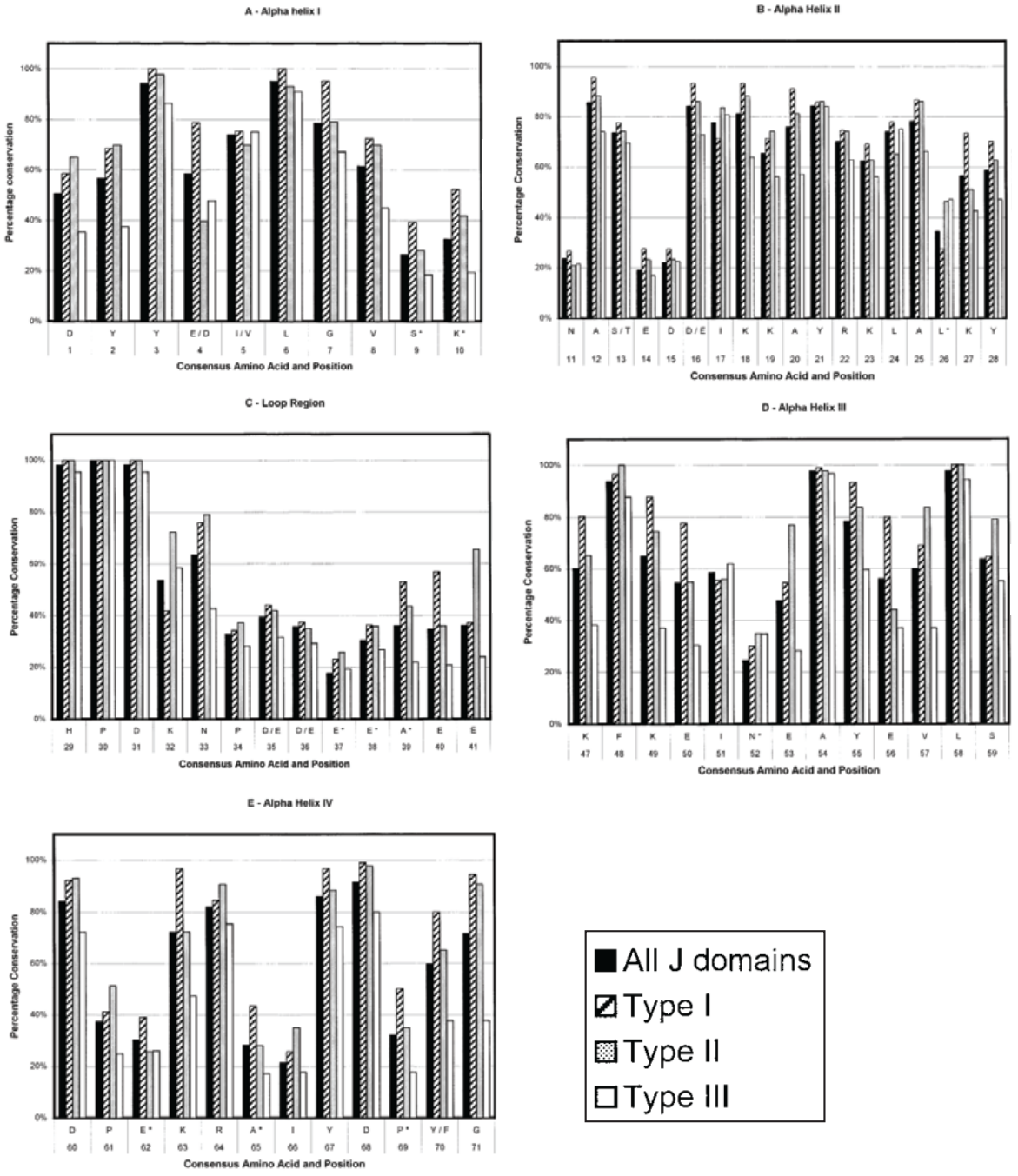


Figura 3: Porcentagem de conservação dos aminoácidos-consenso do domínio-J correspondentes à α-hélice I (A), α-hélice II (B), região de *loop* (C), α-hélice III (D) e α-hélice IV (E) (Hennessy *et al*, 2000).

A subclasse C é o subgrupo de proteínas DNAJ mais heterogêneo em nível de estrutura primária, secundária e funcional, mesmo que ainda existam diversas questões não respondidas a respeito de como estas proteínas exercem suas funções de chaperonas moleculares. Sabe-se que muitas destas proteínas servem como "concentradoras" do domínio-J em determinados compartimentos celulares de modo a recrutar proteínas HSP70 nestes locais

sem haver a formação do complexo DNAJ-substrato [10]. Comparadas com proteínas das subclasses A e B, a subclasse C possui uma diversa distribuição subcelular, e geralmente observa-se um número restrito de ligantes para cada DNAJC, enquanto que as DNAJAs e DNAJBs interagem mais promiscuamente com outras proteínas [9, 28].

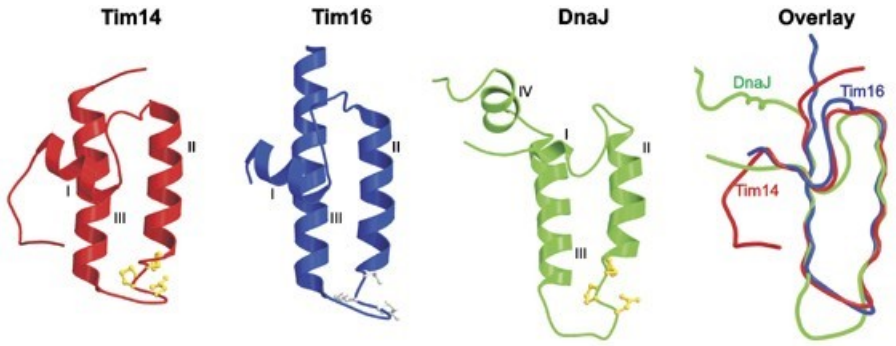


Figura 4: Comparação das estruturas tridimensionais de TIM14 (*Saccharomyces cerevisiae*), TIM16 (*Saccharomyces cerevisiae*) e DnaJ (*Escherichia coli*) com o motivo HPD de TIM14 e DnaJ em amarelo e o motivo DKE de TIM16 em azul (Groll *et al.*, 2006).

1.4 Diversidade das DNAJs

O atual sistema de classificação das proteínas pertencentes à família DNAJ não explora a interação DNAJ-substrato, que é uma característica crucial dentro do contexto das DNAJs como proteínas selecionadoras de clientes para a maquinaria HSP70 [10]. Grande parte de tal diversidade deve-se a regiões externas ao domínio-J, as quais podem variar entre os membros da família DNAJ e podem inclusive carecer de qualquer relação em nível de sequência [9]. Estes diversificados domínios de ligação estão associados a diferentes modos de interação com o substrato, resultando em diferentes especificidades de ligação que variam de DNAJs mais seletivas (ligante específico) a mais promíscuas (ligante não-específico). Dois exemplos de ligantes específicos são as proteínas auxilina e Hsc20, as quais são responsáveis pelo transporte transmembrana de moléculas provenientes de vesículas de clatrina [29] e transferência de grupamentos ferro-enxofre (Fe-S) para proteínas receptoras [30], respectivamente. Os ligantes não-específicos geralmente são encontrados tanto em procariotos quanto em diferentes compartimentos celulares de eucariotos, como no caso das DNAJs do retículo endoplasmático *hs*DNAJB9,*hs* DNAJC10 e *hs*DNAJC3 que ligam diversos substratos não-enovelados ou mal-enovelados [31]. Um caso mais intermediário compreende a DNAJB6 e seu homólogo DNAJB8, os quais possuem uma região rica em serina e treonina (S/T) que interage com fragmentos de poliglutamina (poliQ) de diversas proteínas para suprimir sua agregação, mas que não é necessária para mediar a interação com os outros substratos destas DNAJs [32].

Outra observação importante sobre esta família de chaperonas moleculares é que o número de genes que codificam DNAJs varia evolutivamente entre organismos, estando po-

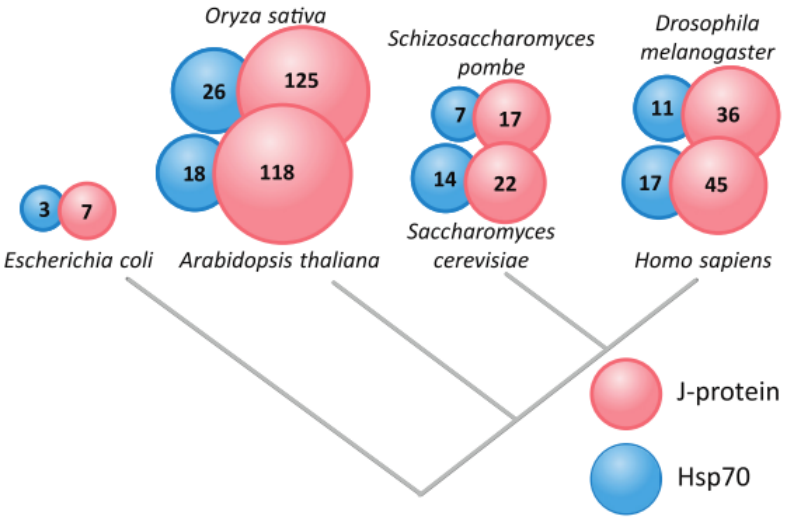


Figura 5: Comparação do número de genes codificando proteínas HSP70 e DNAJ em diferentes organismos (Craig & Marszalek, 2017).

sitivamente correlacionado com a complexidade genômica da espécie em questão [33]. Isto de certa forma implica que estas proteínas são o produto de eventos evolutivos dinâmicos como, por exemplo, duplicação gênica e rearranjos. Entretanto, a origem de tal divergência genética entre as DNAJs é difícil de ser explicada, e uma das hipóteses é o ganho evolutivo do domínio-J em proteínas específicas de certos processos celulares que permitiu a participação da maquinaria HSP70 de forma mais direta. A Figura 5 mostra a proporção do número de genes codificadores de proteínas HSP70 e DNAJ em 7 organismos diferentes, o que corrobora com a hipótese de que o aumento do número de proteínas DNAJ, junto com a sua diversificação na escala evolutiva, está relacionado com a diversificação da maquinaria HSP70. De maneira geral, a diversidade estrutural e funcional das DNAJs apresenta desafios para as comparações entre espécies e para generalizações, fazendo-se necessárias análises evolutivas e estruturais mais profundas que possam levar a um sistema de classificação mais informativo desta família de chaperonas moleculares [9].

1.5 Evolução da Família DNAJ

O interesse por estudos evolutivos da família DNAJ não é novidade e tem sido aplicado tanto para melhor compreender esta família de chaperonas moleculares isoladamente quanto para estudar a organização das subfamílias de HSPs dentro do amplo contexto do controle de qualidade proteico. Diversos alinhamentos de sequências correspondentes a diferentes regiões de proteínas DNAJ e também reconstruções filogenéticas de um ou mais organismos já foram gerados com os mais variados objetivos, entre eles: (i) comparação da estrutura primária e/ou estudo de processos evolutivos associados à diversidade da família DNAJ [20, 21, 22, 23, 34, 35, 36, 37, 38, 39], (ii) análise da sequência nucleotídica como uma ferramenta de identificação de espécies (marcador filogenético) [40, 41, 42, 43]

e (iii) estudo de mecanismos do controle de qualidade proteico envolvendo membros da família DNAJ e também de outras subfamílias de HSPs [44, 45, 46, 47, 48].

A maioria destes trabalhos envolvendo uma abordagem filogenética reiteram a heterogeneidade desta família de chaperonas moleculares, a qual pode ser averiguada no alinhamento das sequências primárias e também na comparação das respectivas reconstruções filogenéticas com as de outros grupos ou famílias de proteínas. Um trabalho recente nesta área avaliou a organização de proteínas das subclasses A e B de diversos organismos da escala evolutiva através do alinhamento dos domínios-J e região C-terminal. Junto com evidências estruturais e de imuno-histoquímica, este estudo demonstrou o surgimento de uma assinatura específica de eucariotos na formação de complexos entre DNAJAs e DNAJBs canônicas envolvidos no sistema de desagregação de proteínas [24]. Entretanto, é desconhecida qualquer reconstrução filogenética envolvendo as 3 subclasses de proteínas DNAJ de organismos representativos da escala evolutiva.

2 Justificativa

A subfamília DNAJ/HSP40 de HSPs é determinada pela assinatura do domínio-J, o qual é responsável por estimular a atividade ATPase no centro da maquinaria HSP70 [10]. Entretanto, as regiões fora do domínio-J também exercem funções importantes que podem estar envolvidas com o reconhecimento do substrato a ser enviado para o sistema de controle de qualidade proteico, interação com proteínas amilóides envolvidas em doenças neurodegenerativas, regiões transmembrana para posicionamento subcelular, associação com o ribossomo para garantir o correto enovelamento de proteínas recém sintetizadas e outras regiões que garantem a diversidade desta família de proteínas [11, 18, 49]. De acordo com o atual sistema de classificação da família DNAJ, os membros das subclasses A e B possuem certo grau de similaridade em relação às regiões fora do domínio-J, enquanto que os membros da subclasse C são extremamente heterogêneos [21]. Especula-se que esta heterogeneidade tenha sido de extrema importância evolutiva, permitindo tanto a diversificação da maquinaria HSP70 quanto o desenvolvimento de mecanismos de controle de qualidade proteico mais complexos [9], e alguns trabalhos têm demonstrado isso através de abordagens filogenéticas que exploram a organização e dispersão de algumas proteínas da família DNAJ em um ou mais organismos [24, 45, 50]. Apesar dos estudos e interesse nesta área serem recorrentes, é desconhecido algum trabalho que explore a relação entre as subclasses A, B e C reunidas em diferentes organismos da escala evolutiva, instigando o questionamento de como diferentes proteínas DNAJ se dispersam filogeneticamente e quais fatores determinam a sua organização.

3 Objetivos

3.1 Objetivo geral

Realização de um estudo filogenético que caracterize as relações evolutivas entre as três subclasses da família DNAJ e que possa apontar para uma organização mais informativa destas proteínas.

3.2 Objetivos Específicos

- Observar a divisão das subclasses A e B em reconstruções filogenéticas apenas destas duas subclasses e a divisão das subclasses A, B e C em reconstruções filogenéticas das 3 subclasses reunidas.

- Analisar as diferenças entre reconstruções filogenéticas realizadas a partir do alinhamento apenas do domínio-J ou a partir da sequência completa da proteína.

- Analisar o agrupamento de proteínas de um mesmo reino nas diferentes reconstruções filogenéticas.

4 Metodologia

Os métodos empregados neste trabalho envolveram a coleta e curagem das sequências de aminoácidos das proteínas DNAJ, seleção do domínio-J, alinhamento múltiplo das proteínas em questão e, finalmente, a construção das respectivas árvores filogenéticas. Todas estas etapas estão resumidas no fluxograma da Figura 6.

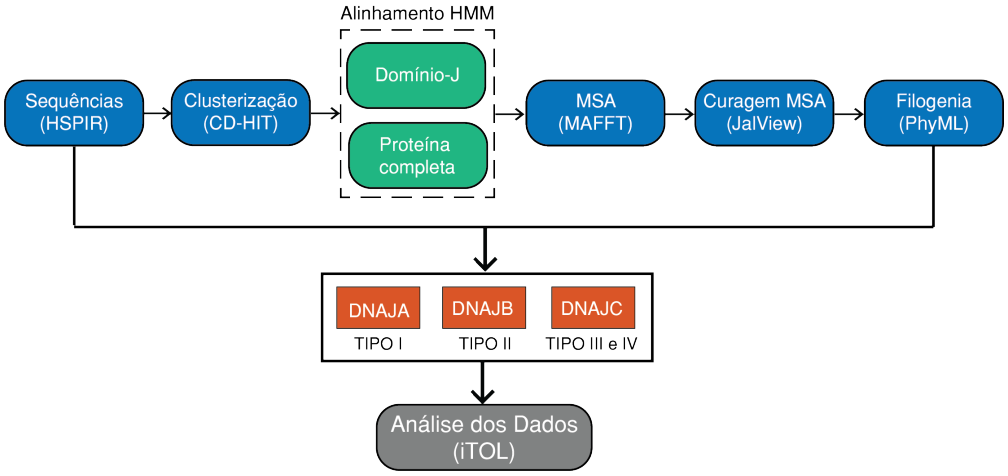


Figura 6: Fluxograma das etapas de processamento dos dados deste trabalho.

4.1 Coleta dos dados

As proteínas DNAJ analisadas no presente estudo foram coletadas na plataforma HSPIR [51] descrita abaixo, totalizando 3901 proteínas.

4.1.1 Plataforma HSPIR

A plataforma *Heat Shock Protein Information Resource* (HSPIR; `http://pdslab.biochem.iisc.ernet.in/hspir/`) contém um banco de dados online manualmente curado que fornece informações sobre os seis grandes subgrupos de HSPs: HSP20, HSP40, HSP70, HSP60, HSP90 e HSP100 [51]. Dentre os dados disponíveis para cada proteína nesta plataforma, constam: (i) nomes da proteína; (ii) nomes do gene correspondente; (iii) informação taxonômica; (iv) classificação da família e subtipo; (v) informações da sequência nucleotídica e proteica, e outras informações de acordo com a disponibilidade de dados. Desde a última atualização (20 de Janeiro de 2014), a plataforma contém cerca de 9.900 registros de proteínas, abrangendo 277 genomas que variam de procariotos a eucariotos e incluem a maioria dos organismos modelo. Os registros são verificados semanalmente por atualizações de *scripts* automatizados que mantêm os conteúdos atualizados, além de existir uma equipe de curadores dedicada que revisa estas atualizações e as insere no banco de dados. Todo o processo de curadoria está esquematizado na Figura 7.

Esta plataforma possui uma ferramenta de identificação e classificação de sequências desconhecidas em uma determinada família HSP. Nesta ferramenta é definido um perfil HMM (Hidden Markov Model) para cada família HSP, o qual é criado a partir de um conjunto de "sequências base" validadas (*seed sequences*). A sequência desconhecida é verificada em relação às bibliotecas de perfis HSP pré definidas usando a função *hmmscan* do HMMER (`http://hmmer.org/`). Para o conjunto de dados inicial, foram recuperadas todas as estruturas 3D de HSPs disponíveis no PDB e extraídos os domínios conservados de cada família HSP (i.e. domínio-J para HSP40 e domínio α-cristalina para sHSP). No caso de proteínas da família HSP60 e HSP70, as quais são conservadas em toda a sua extensão, foram utilizadas as sequências completas posteriormente alinhadas para criar o perfil HMM. Para suportar o conjunto de dados inicial, foi recuperado um conjunto de sequências estruturalmente bem anotadas do UniprotKB (`https://www.uniprot.org/help/uniprotkb`) e executado o mesmo procedimento acima para extrair as regiões conservadas, e estas sequências filtradas foram então alinhadas com o perfil inicial para gerar um novo perfil [51].

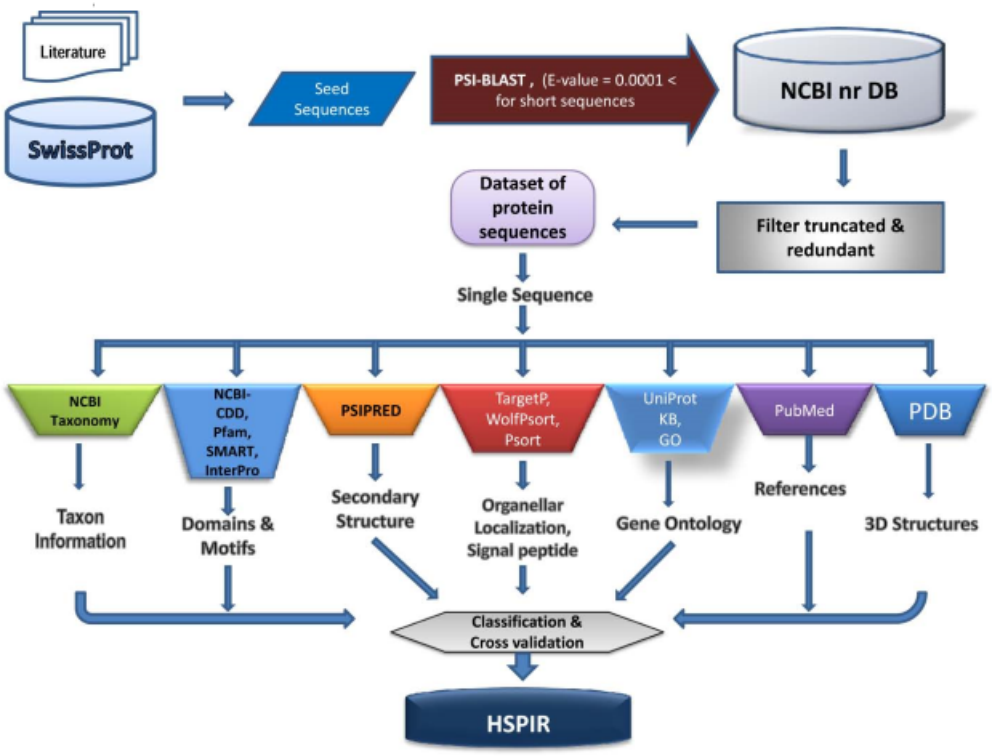


Figura 7: Representação esquemática do processo de recuperação e curadoria de dados implementado no banco de dados HSPIR (Kumar *et al*, 2012).

4.2 Curagem e clusterização dos dados

Primeiramente foram excluidas as sequências contendo os caracteres não-padrão 'B', 'X' ou 'Z'. Depois deta etapa, os dados foram clusterizados através do software *CD-HIT*

Suite: Biological Sequence Clustering and Comparison (`http://weizhongli-lab.org`) [52] utilizando um valor de *cutoff* de 40% de identidade de acordo com o trabalho descrito em [53] ou de 80% de acordo com a metodologia descrita em [24].

4.3 *Hidden Markov Model* (HMM)

Resumidamente, um *Hidden Markov Model* (HMM) é um modelo estatístico que pode ser usado para descrever a evolução de eventos observáveis que dependem de fatores internos, os quais não podem ser diretamente observados. Estes eventos observáveis são chamados de "símbolos", enquanto que o fator não observável é chamado de "estado". Desta forma, um HMM consiste de dois processos estocásticos: um processo invisível de estados ocultos (*hidden*) e um processo visível de símbolos observáveis. Os estados ocultos formam uma cadeia de Markov, e a probabilidade de distribuição do símbolo observado depende de cada estado [54]. Um HMM pode ser útil para a modelagem de sequências biológicas como DNA e proteínas, uma vez que tais sequências consistem de pequenas estruturas com diferentes funções que, frequentemente, exibem diferentes propriedades estatísticas. Para uma proteína não caracterizada, por exemplo, é interessante aplicar este modelo estatístico para fazer a predição de domínios proteicos (correspondem a um ou mais estados em um HMM) e da localização destes na sequência peptídica (observações) [54, 55].

4.3.1 Alinhamento contra perfis HMM

Para a seleção de domínios específicos de cada proteína foram realizados alinhamentos par-a-par contra dois perfis HMM correspondentes ao domínio-J (PF00226) e ao domínio PAM16 (*J-like domain*; PF03656), retirados da plataforma PFAM (`https://pfam.xfam.org/`) [56]. Os alinhamentos foram realizados através da função *hmmprofalign* do programa MATLAB R2018a©, a qual se baseia em um alinhamento local e retorna a sequência alinhada junto com um escore no formato *log-odd*. Para ambos os perfis HMM foram computados os histogramas dos escores de acordo com os subtipos I, II, III e IV.

4.4 Alinhamento múltiplo de sequências

Foram realizados alinhamentos separados para as proteínas das subclasses A, B e C, assim como alinhamentos contendo as proteínas das 3 subclasses reunidas. Cada alinhamento foi realizado através do programa MAFFT v7. 397 [57], o qual foi utilizado localmente. O algoritmo L-INS-i foi escolhido para todos os cálculos por ser um método de refinamento iterativo que gera um alinhamento local (Smith-Waterman) mais acurado. Para cada alinhamento múltiplo foi selecionado um número máximo de iterações igual a 1000.

4.5 Curagem do alinhamento múltiplo

Foi utilizado o *software* Jalview 2.10.4b1 [58] tanto para visualizar quanto para curar os alinhamentos provenientes do MAFFT. O escore de qualidade do Jalview é inversamente proporcional ao custo médio de todos os pares de mutação observados em uma coluna particular do alinhamento. Desra forma, um alto valor de escore para uma coluna sugere que não existem mutações ou que a maioria das mutações observadas são favoráveis. Para determinar o valor de cada coluna do alinhamento é somada, para todas as mutações, a razão entre os dois escores de um par de mutação e o escore de cada resíduo conservado de acordo com a matriz de substituição BLOSUM62. Este valor é normalizado para cada coluna e então plotado em uma escala de 0 a 1. Os valores de *cutoff* escolhidos foram de 40% para os alinhamentos das subclasses A e B (Apêndice A) e de 30% para os alinhamentos da subclasse C (Apêndice B) e das três subclasses reunidas (Apêndice C) [59, 58]. A escolha destes valores foi realizada de maneira empírica, de forma a minimizar a perda de colunas do alinhamento.

4.6 Filogenia

Foi escolhido o *software* PhyML 3.0 [60, 61] para realizar as reconstruções filogenéticas, o qual se baseia no princípio de máxima verossimilhança (*maximum-likelihood*). O modelo de substituição LG é o padrão do PhyML aplicado para sequências de aminoácidos [60, 62, 63], e o teste estatístico aLRT (*approximate likelihood ratio test*) foi escolhido para testar os ramos [64]. A escolha do teste aLRT baseou-se principalmente pelo fato deste ser muito mais rápido que o método por *bootstraps*, uma vez que o programa é rodado apenas uma vez, enquanto que pelo método de *bootstrap* o programa é executado um número pré-definido de vezes, determinado pela quantidade de replicatas solicitada. Além disso, a aplicação de ambos os testes para um mesmo conjunto de dados demonstrou que os resultados concordam, apontando para a vantagem da utilização do teste estatístico aLRT quando o conjunto de dados é extenso ou quando os recursos computacionais são limitados [60].

4.7 Visualização e manipulação da árvore

Cada árvore resultante da reconstrução filogenética foi salva em formato Newick e visualizada através da ferramenta online *iTOL: Interactive Tree Of Life* (`https://itol.embl.de/`) [65]. Neste trabalho, todas as árvores estão representadas no formato circular com um enraizamento por ponto médio matemático [66] padrão do iTOL e algumas árvores estão representadas no formato não enraizado [65]. Além disto, estão representados apenas ramos com um valor limite igual ou maior que 0.75, de acordo com a recomendação do *software* PhyML 3.0 [60].

5 Resultados

Após a curagem das 3901 sequências provenientes da plataforma HSPIR e posterior clusterização utilizando um valor de *cut-off* de 40%, foi realizado o alinhamento de 1199 proteínas DNAJ contra dois perfis HMM para definir as regiões correspondentes ao domínio-J e ao domínio TIM16. Dentre estas proteínas, foram realizados os alinhamentos múltiplos locais da sequência completa e do domínio-J de 63 proteínas da subclasse A e 55 proteínas da subclasse B separadas ou reunidas. Por sua vez, as 1081 proteínas da subclasse C (1061 do tipo III e 20 do tipo IV) tiveram apenas os domínios-J alinhados. Após a curagem dos alinhamentos, foram geradas oito reconstruções filogenéticas (Figuras 10 a 14). As 3901 proteínas também foram clusterizadas utilizando um valor de *cut-off* de 80%, e as sequências completas de 323 proteínas da subclasse A e 198 proteínas da subclasse B foram alinhadas de forma separada ou reunida, gerando três outras reconstruções filogenéticas. Os resultados destas etapas estão descritos em detalhes a seguir.

Tabela 1: Quantidade de proteínas após a respectiva etapa de curagem. *Retirada de proteínas contendo caracteres não-padrão: X, B e Z.

Tipo	HSPIR original	HSPIR *	CD-HIT 40%	CD-HIT 80%
I	533	530	63	323
II	312	309	55	198
III	2968	2932	1061	2126
IV	88	88	20	55
Total	3901	3859	1199	2702

5.1 Alinhamentos contra os perfis HMM

As sequências de aminoácidos de todas as 1199 proteínas resultantes após a etapa de curagem e clusterização do banco de dados com um valor de *cut-off* de 40% foram submetidas aos alinhamentos par-a-par contra os perfis HMM do domínio-J (PF00226) e do domínio *J-like* de TIM16 (PF03656). Os escores de cada alinhamento foram agrupados por tipo (I, II, III ou IV) para cada domínio de interesse em histogramas representados na mesma escala, no eixo horizontal. Para o alinhamento contra a sequência consenso do domínio-J, as proteínas das subclasses A (tipo I) e B (tipo II) obtiveram escores altos e similares, enquanto que as proteínas da subclasse C (tipos III e IV) obtiveram escores mais baixos e distintos quando comparados entre os tipos III e IV e com as subclasses A e B. Entretanto, apesar de obter uma média menor do que os tipos I e II, o tipo III ainda obteve escores consideravelmente maiores do que o tipo IV (Figura 8). Por outro lado, os resultados do alinhamento contra a sequência consenso do domínio *J-like* mostram valores

de escore baixos e similares para os tipos I, II e III e valores altos e contrastantes para o tipo IV (Figura 9).

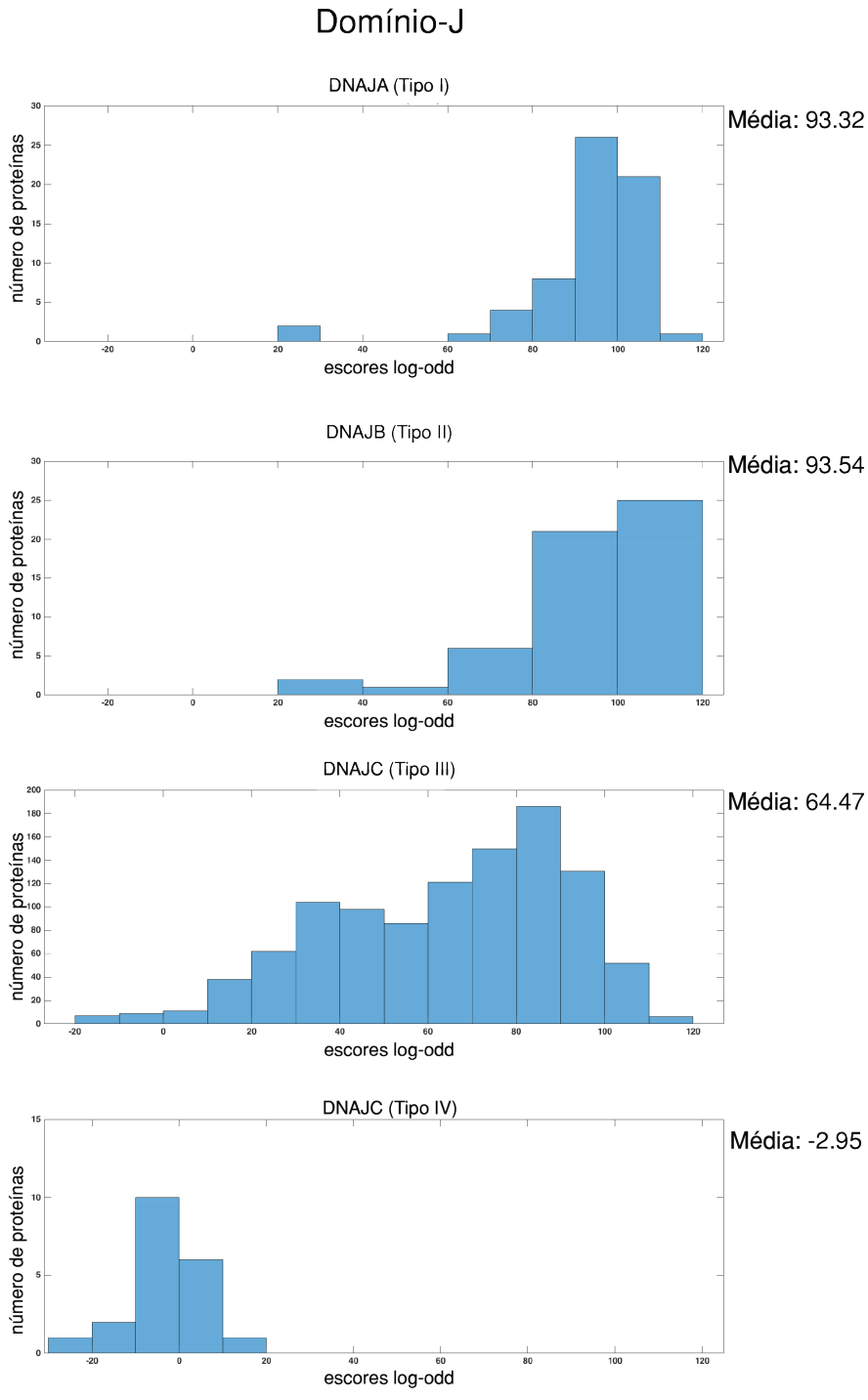


Figura 8: Escores do alinhamento de cada proteína (tipos I, II, III e IV) contra o domínio-J (PF00226)

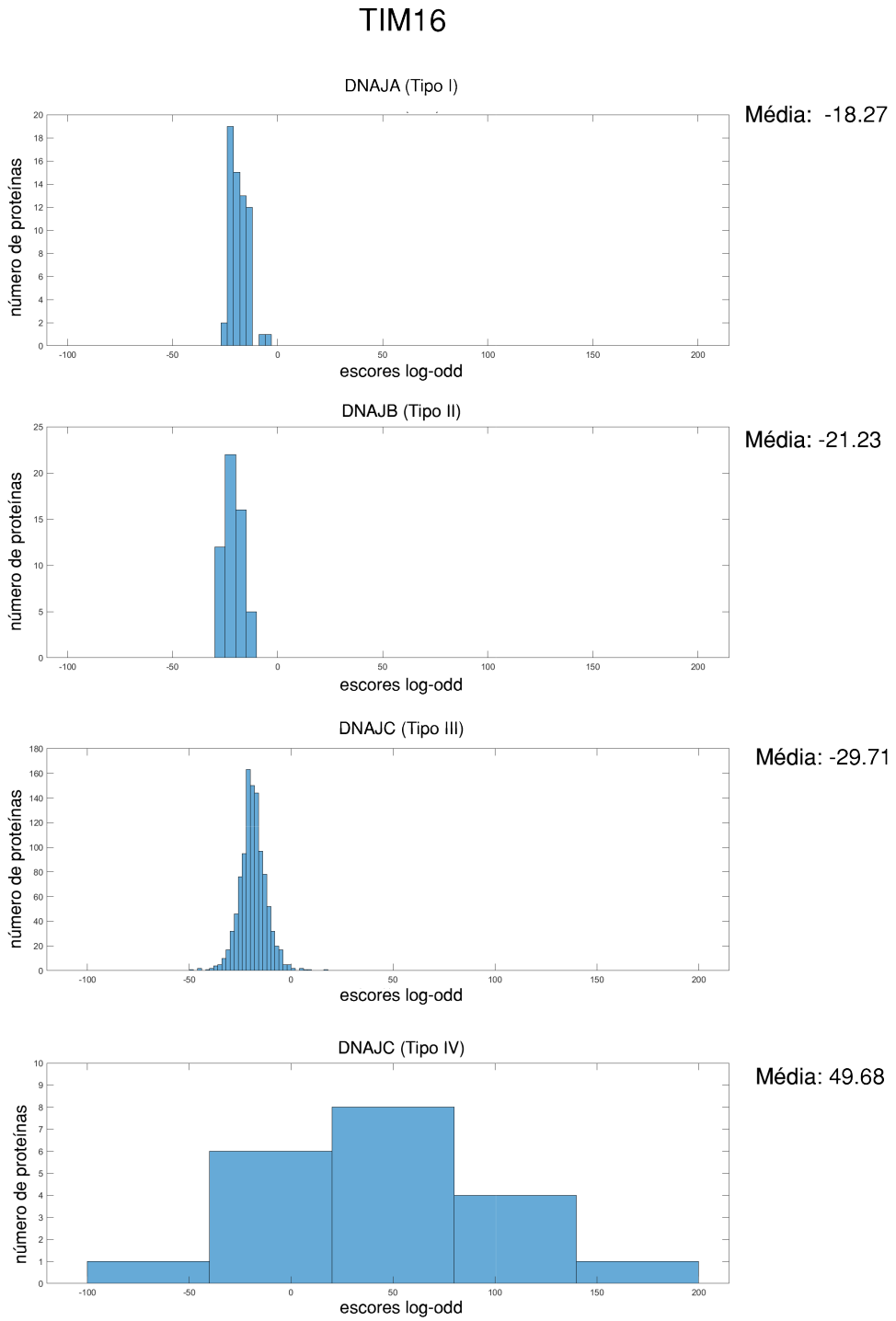


Figura 9: Escores do alinhamento de cada proteína (tipos I, II, III e IV) contra o domínio *J-like* de TIM16 (PF03656)

5.2 Filogenia

Em todas as árvores geradas neste trabalho é possível observar a organização dos diferentes reinos e, quando aplicado, os diferentes tipos de DNAJ (I, II, III e IV). Devido à quantidade de ramos da maioria das árvores geradas neste trabalho, estão representados apenas os valores de suporte das árvores das subclasses A e B após a clusterização com um *cut-off* de 40%.

De forma a avaliar a capacidade da metodologia aplicada neste trabalho de conseguir agrupar proteínas de um mesmo reino ou mais próximas evolutivamente, foram realizadas as reconstruções filogenéticas das subclasses A (tipo I) e B (tipo II) separadamente a partir do alinhamento apenas do domínio-J ou da proteína completa. As Figuras 10 e 11 mostram as árvores das subclasses A e B, respectivamente. Além disso, para avaliar a eficiência no agrupamento de subclasses, foi realizada a reconstrução filogenética das subclasses A e B agrupadas a partir do alinhamento apenas do domínio-J ou da proteína completa (Figura 12).

Considerando que os tipos III e IV possuem diferenças pontuais na estrutura do domínio-J/*J-like* mas estão agrupados na subclasse C, foi realizada a reconstrução filogenética deste dois tipos agrupados a partir do alinhamento referente ao domínio-J para verificar se existe uma discriminação entre os tipos III e IV (Figura 13).

Semelhantemente às proteínas da subclasse C, foi realizada a reconstrução filogenética das 3 subclasses agrupadas a partir do alinhamento das sequências correspondentes ao domínio-J (Figura 14). Nesta árvore, os ramos de linha tracejada representam o grande clado no qual as proteínas das subclasses A e B se concentraram, demonstrando uma possível divisão evolutiva dos domínios-J da maioria das DNAJAs e DNAJBs em relação aos domínios-J da maioria das DNAJCs que precisa ser explorada.

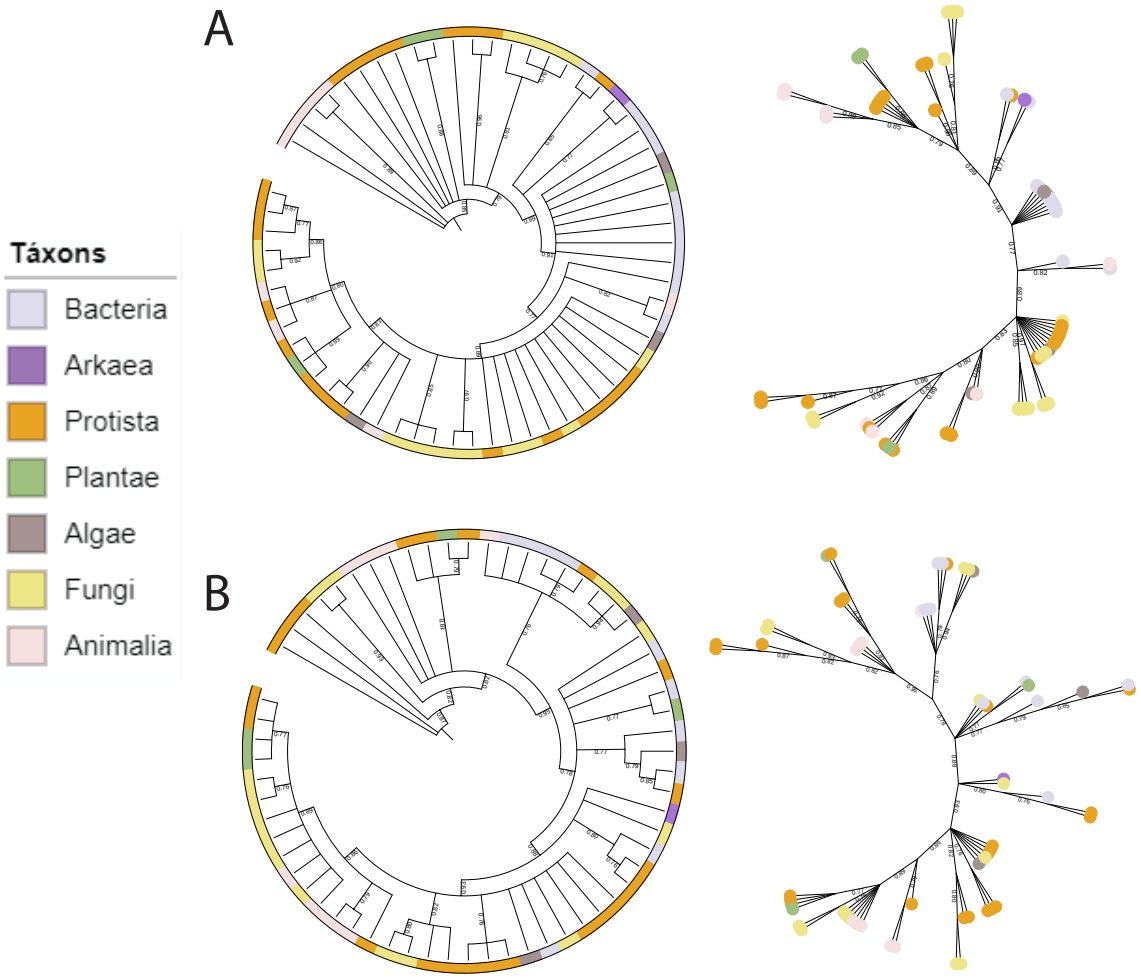


Figura 10: Árvores filogenéticas das proteínas do tipo I (DNAJA) a partir do alinhamento múltiplo de (A) toda a extensão da proteína ou apenas do (B) domínio-J de proteínas clusterizadas com um valor de *cut-off* de 40%.

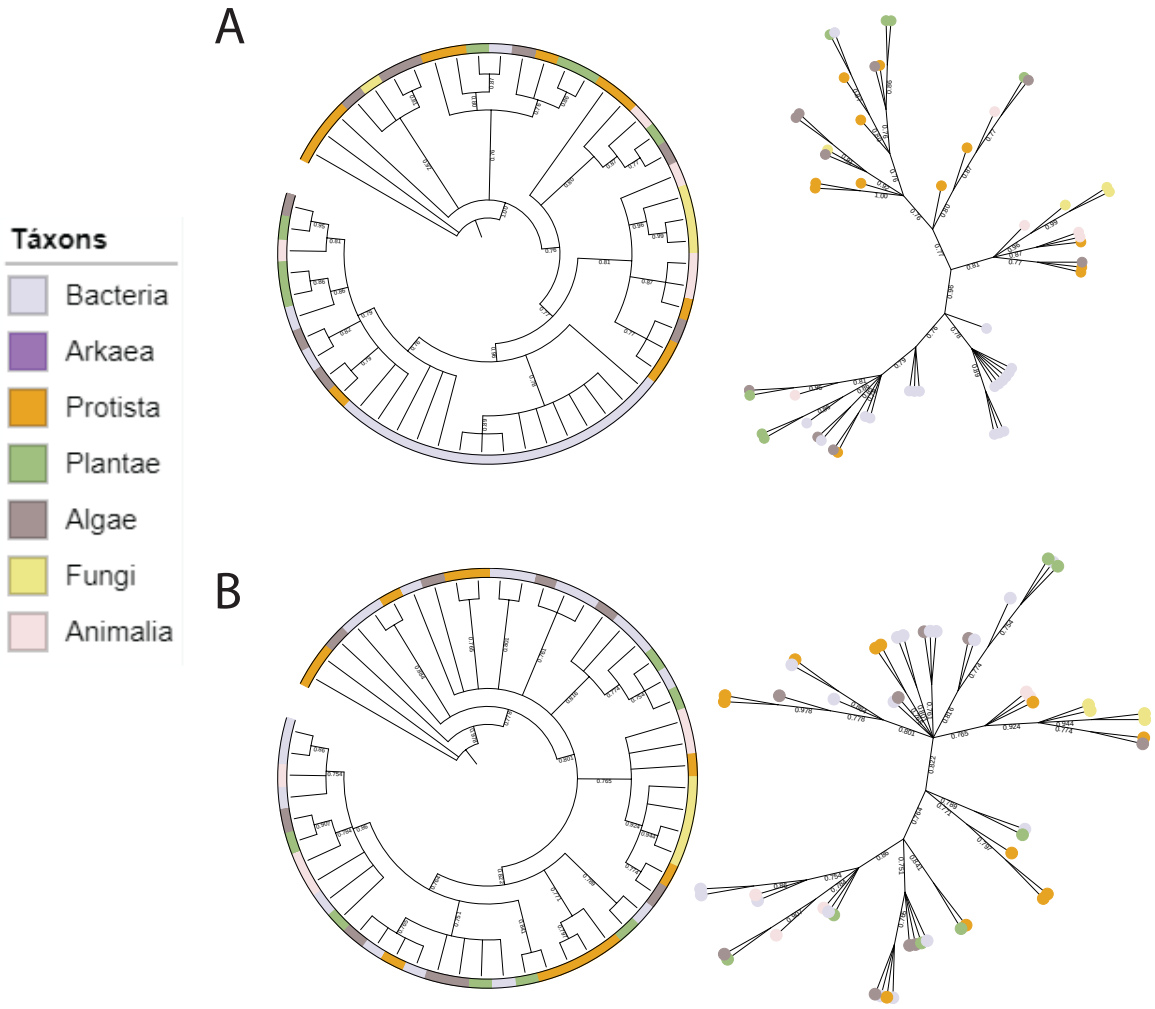


Figura 11: Árvores filogenéticas das proteínas do tipo II (DNAJB) a partir do alinhamento múltiplo de (A) toda a extensão da proteína ou apenas do (B) domínio-J de proteínas clusterizadas com um valor de *cut-off* de 40%.

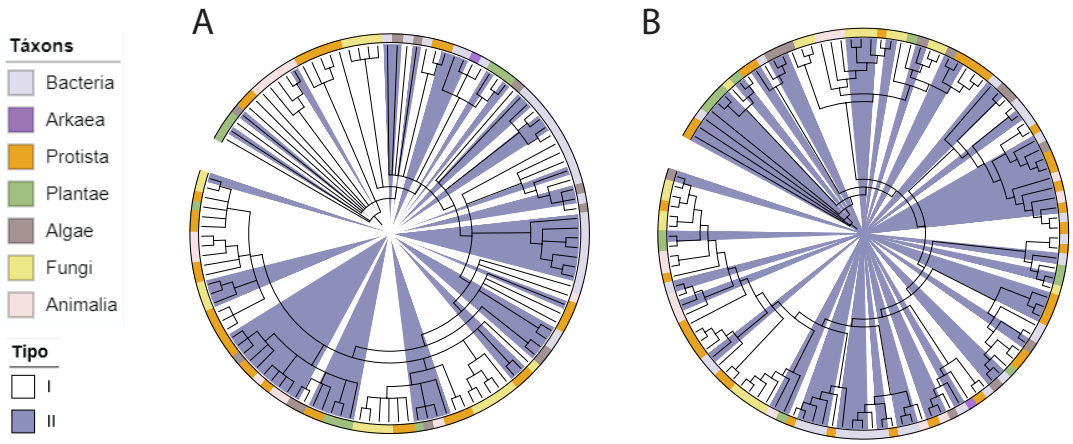


Figura 12: Árvore filogenética das proteínas dos tipos I (DNAJA) e II (DNAJB) a partir do alinhamento múltiplo tanto da proteína completa (A) quanto apenas do domínio-J (B) utilizando as sequências clusterizadas com um valor de *cut-off* de 40%.

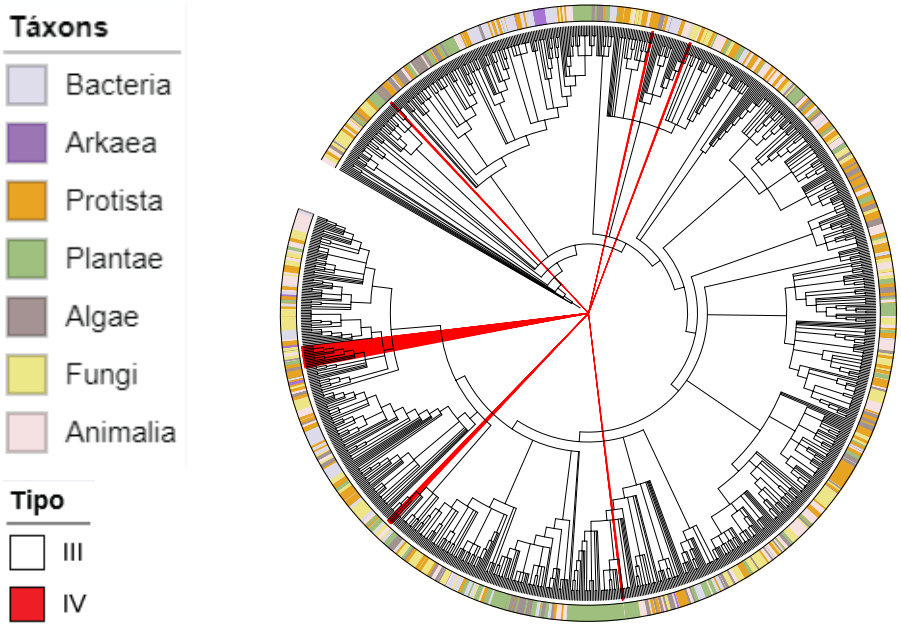


Figura 13: Árvore filogenética das proteínas do tipo III e IV (DNAJC) a partir do alinhamento múltiplo apenas do domínio-J utilizando as sequências clusterizadas com um valor de *cut-off* de 40%.

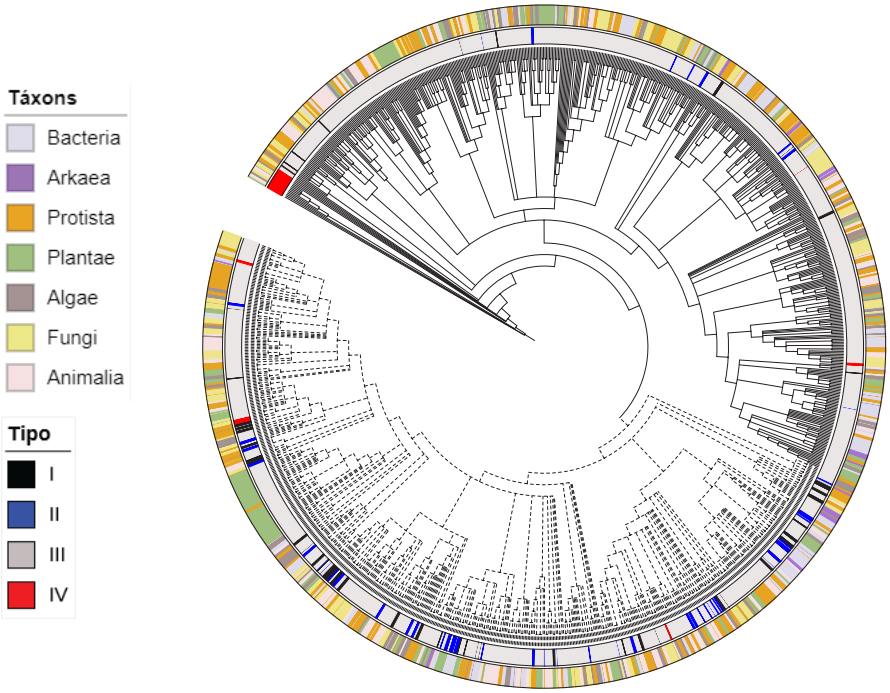


Figura 14: Árvore filogenética de todas as proteínas (DNAJA, DNAJB e DNAJC) a partir do alinhamento múltiplo apenas do domínio-J utilizando as sequências clusterizadas com um valor de *cut-off* de 40%. As linhas tracejadas indicam o clado que concentra as proteínas das subclasses A e B.

5.2.1 Clusterização dos dados com um *cut-off* de 80%

Após a análise dos dados resultantes da clusterização utilizando um valor de *cut-off* de 40%, foram realizadas novas reconstruções filogenéticas posteriores à clusterização das 3901 proteínas utilizando um valor de *cut-off* de 80%, com o objetivo de comparar os agrupamentos por reino e por subclasse com as árvores anteriores. Neste caso, toda a extensão da proteína foi considerada para os alinhamentos múltiplos, e não apenas o domínio-J. Novamente, as reconstruções filogenéticas das subclasses A e B foram realizadas para cada subclasse separadamente ou para as duas subclasses agrupadas para verificar a capacidade da metodologia aplicada em agrupar proteínas pelos respectivos reinos e/ou pelas respectivas subclasses. As Figuras 15 e 16 representam as árvores das subclasses A e B, respectivamente, enquanto a Figura 17 representa a árvore das subclasses A e B agrupadas, todas provenientes do alinhamento de toda a extensão da proteína. Por outro lado, a Figura 18 representa a árvore proveniente do alinhamento dos domínios-J de todas as subclasses após esta nova clusterização. Além de novamente demonstrar a concentração de proteínas das subclasses A e B em um grande clado, a árvore proveniente da nova clusterização dos dados acentuou esta divisão. Apesar de não garantir a robustez da reconstrução filogenética, este resultado aponta para a consistência do dado obtido em relação à separação das subclasses.

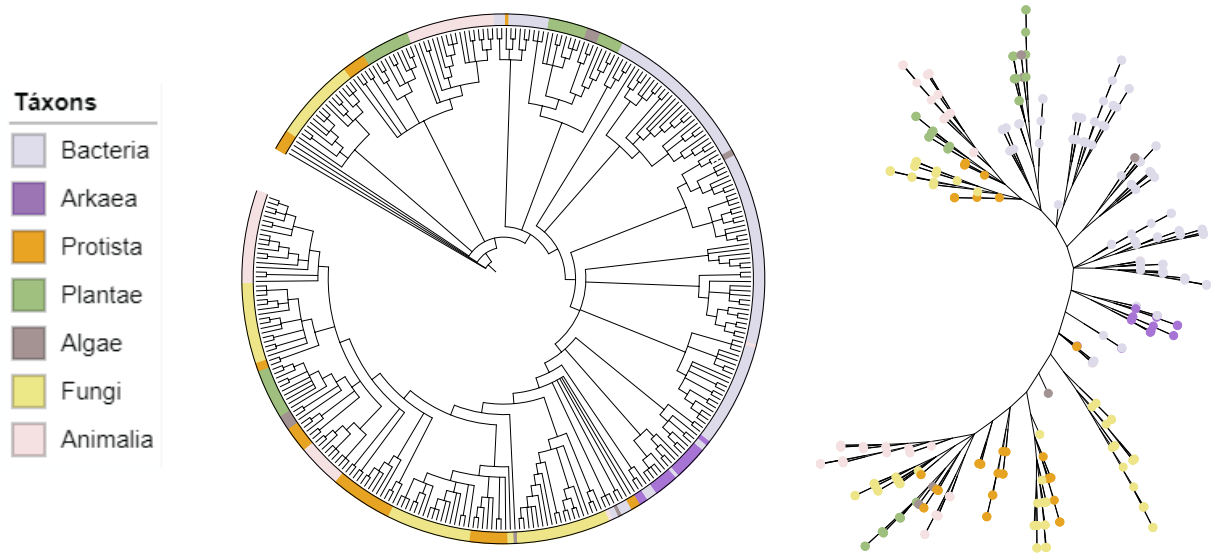


Figura 15: Árvore filogenética das proteínas da subclasse A (tipo I) a partir do alinhamento múltiplo de toda a extensão da proteína utilizando as sequências clusterizadas com um valor de *cut-off* de 80%.

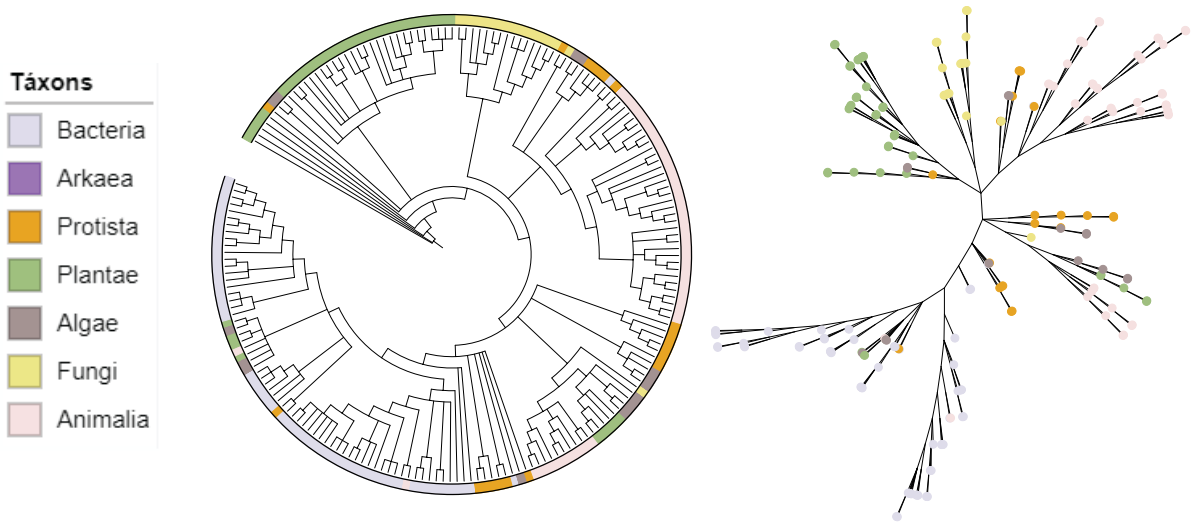


Figura 16: Árvore filogenética das proteínas da subclasse B (tipo II) a partir do alinhamento múltiplo de toda a extensão da proteína utilizando as sequências clusterizadas com um valor de *cut-off* de 80%.

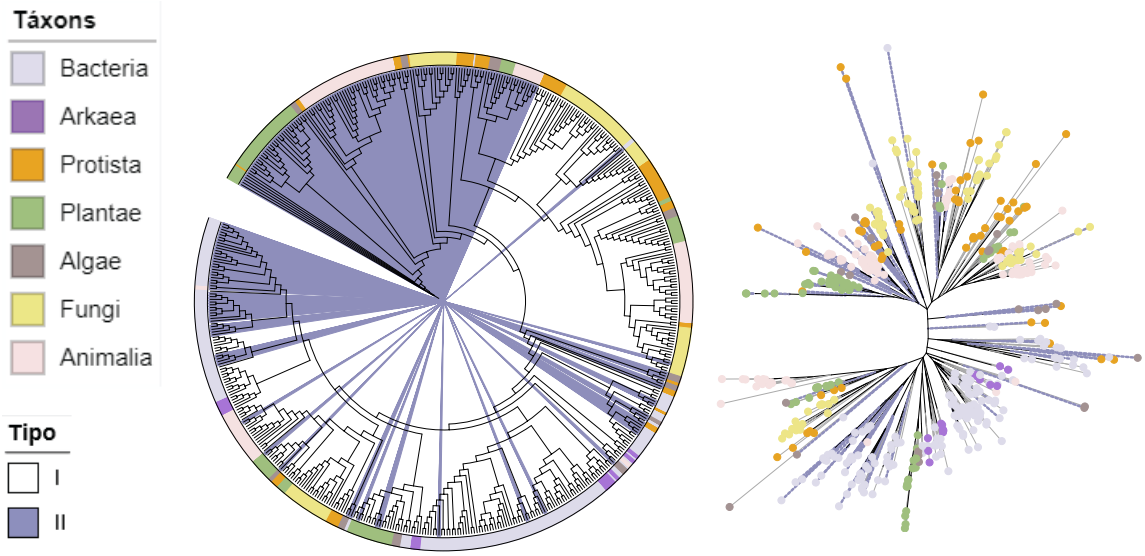


Figura 17: Árvore filogenética das proteínas das subclasses A e B (tipos I e II) a partir do alinhamento múltiplo de toda a extensão da proteína utilizando as sequências clusterizadas com um valor de *cut-off* de 80%.

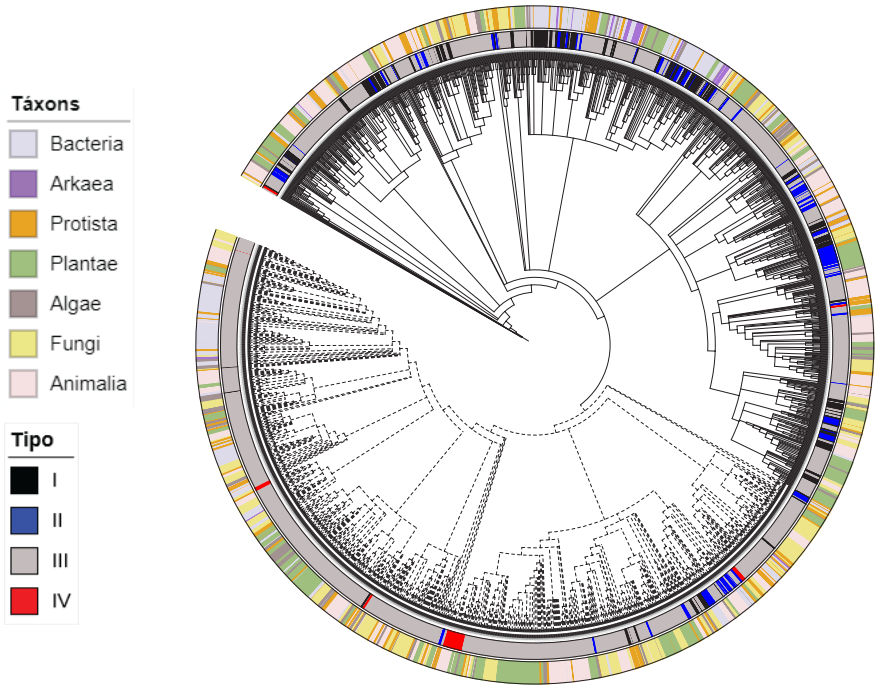


Figura 18: Árvore filogenética das proteínas das subclasses A, B e C (tipos I, II, III e IV) a partir do alinhamento múltiplo da região correspondente ao domínio-J utilizando as sequências clusterizadas com um valor de *cut-off* de 80%. As linhas tracejadas indicam o clado que concentra as proteínas das subclasses A e B.

6 Discussão

Uma vez que diversas reconstruções filogenéticas deste trabalho foram baseadas no alinhamento apenas dos domínios-J, uma etapa essencial da metodologia foi a realização de alinhamentos par-a-par de cada proteína contra o perfil HMM do domínio-J (PF00226). Conforme os histogramas apresentados na Figura 8, a maioria das proteínas das subclasses A (tipo I) e B (tipo II) obtiveram altos valores de escore, refletindo uma alta similaridade com a sequência consenso do domínio-J. No caso das proteínas da subclasse C, os histogramas foram divididos entre tipo III e tipo IV, e enquanto os membros do tipo III obtiveram escores mais dispersos com uma média de escore menor em relação às subclasses A e B, os membros do tipo IV se concentraram na porção mais inferior do histograma, refletindo muito pouca similaridade com a sequência consenso do domínio-J. Este resultado concorda com dados da literatura que demonstram o menor índice de conservação do domínio-J entre membros da subclasse C, comparado com a conservação dos domínios-J das subclasses A e B. Além disso, os histogramas refletem a clara divisão do tipo IV em relação aos tipos I, II e III devido à baixa, ou praticamente nula, conservação da estrutura primária do domínio-J. Semelhantemente ao alinhamento contra a sequência consenso do domínio-J, foram realizados alinhamentos par-a-par de cada proteína contra o perfil HMM do domínio *J-like* de TIM16 (PF03656). Apesar de dispersos, a maioria dos membros do tipo IV obtiveram escores medianos e altos que indicam considerável similaridade com o domínio TIM16, enquanto que todas as outras proteínas correspondentes aos tipos I, II e III obtiveram valores de escore baixos (Figura 9). Este resultado, junto com os histogramas do domínio-J, demonstram diferenças pontuais entre proteínas dos tipos III e IV, as quais são atualmente agrupadas na subclasse C.

As reconstruções filogenéticas das subclasses A (tipo I) e B (tipo II) foram utilizadas como parâmetros para analisar a qualidade e a robustez dos dados filogenéticos gerados, uma vez que os números de proteínas são bem menores e estas apresentam um grau maior de conservação tanto do domínio-J quanto de outras regiões. Comparando-se as reconstruções filogenéticas separadas da subclasse A (Figura 10) e da subclasse B (Figura 11), evidencia-se que as árvores geradas a partir do alinhamento de toda a extensão da proteína possuem uma divisão mais clara dos diferentes reinos quando comparadas às árvores geradas a partir do alinhamento do domínio-J. Semelhantemente, a reconstrução filogenética das subclasses A e B agrupadas (Figura 12) resultante do alinhamento de toda a extensão da proteína também resultou em uma melhor divisão das subclasses (tipos I e II) quando comparada à reconstrução filogenética resultante do alinhamento do domínio-J. Isto provavelmente resulta do fato de o alinhamento da proteína inteira ser mais informativo devido à presença de um número maior de resíduos e suas respectivas combinações. Além disto, o fato do alinhamento pelo domínio-J não apresentar uma separação clara das subclasses A e B de certa forma reforça a conservação deste domínio

entre diferentes subclasses.

Para a reconstrução filogenética da subclasse C (tipos III e IV) foi considerado apenas o domínio-J para realizar o alinhamento, e de acordo com a árvore da Figura 13 não há uma divisão clara dos reinos, mas existe uma maior concentração de proteínas do tipo IV em um pequeno clado. A reconstrução filogenética das 3 subclasses agrupadas (tipos I, II, III e IV) também foi realizada considerando-se apenas o alinhamento do domínio-J e novamente não apresenta uma divisão clara dos reinos (Figura 14). Entretanto, a árvore resultante sugere a concentração de proteínas das subclasses A e B em um grande clado, enquanto que a maioria das proteínas da subclasse C se encontram no clado-irmão. Além disso, as proteínas do tipo IV novamente se concentraram em um pequeno clado pertencente ao clado maior junto com a maioria das proteínas do tipo III. Isto sustenta os dados presentes nos histogramas provenientes do alinhamento par-a-par contra o perfil HMM do domínio-J, mais uma vez demonstrando a divergência deste grupo em relação aos tipos I, II e III. Estes resultados apontam para a necessidade de uma análise filogenética apenas dos domínios-J correspondentes aos tipos I, II e III, uma vez que proteínas do tipo IV não demonstram conservação da estrutura primária do domínio-J.

Uma vez que todos as árvores resultantes da clusterização utilizando um *cut-off* de 40% não apresentaram uma distinção clara dos reinos, decidiu-se testar a clusterização das 3901 proteínas com um valor *cut-off* de 80%. Ao invés de realizar o alinhamento dos domínios-J, estas novas reconstruções filogenéticas foram baseadas no alinhamento das proteínas completas, e novamente as árvores das subclasses A e B serviram como parâmetro para analisar a qualidade do resultado filogenético. Tanto a árvore da subclasse A (Figura 15) quanto da subclasse B (Figura 16) separadas ou reunidas apresentou uma melhor divisão dos reinos. Além disso, a árvore resultante do alinhamento das subclasses A e B agrupadas também apresentou uma melhor divisão das subclasses, demonstrando que a clusterização com um *cut-off* de 80% está relacionado com uma reconstrução filogenética mais informativa.

Todas as árvores geradas neste trabalho apresentam politomias dispersas e de diferentes tamanhos, as quais estão relacionadas com a incapacidade de resolução de bifurcações para o conjunto de dados em questão [67]. As politomias "simples" (*soft*) são as mais comuns e ocorrem quando a informação filogenética é insuficiente, não permitindo a resolução de uma bifurcação pois não se sabe qual sequência de eventos binários é mais relevante. Isto significa que estas politomias podem ser resolvidas com o melhoramento dos dados da reconstrução filogenética ou pelo fornecimento de mais dados. Por outro lado, as politomias "complicadas" (*hard*) representam três ou mais eventos de especiação simultânea a partir de um mesmo ancestral, gerando ramos equidistantes entre si [68] [69]. A maioria das politomias encontradas neste trabalho são claramente resultado do conjunto de dados utilizado, podendo ser consideradas "simples". Os Apêndices D e E apresentam a comparação das árvores das subclasses A e B reunidas a partir do alinha-

mento da proteína completa ou do domínio-J, respectivamente, e a comparação entre a distribuição de todos os ramos (letra A) e os ramos com escore acima de 0.75 (letra B). Muitas bifurcações apresentam um valor de escore abaixo de 0.75 e isso demonstra que, ou o conjunto de dados tanto a partir da sequência completa quanto do domínio-J não é suficiente para resultar em uma árvore completamente resolvida, ou as etapas da metodologia estão levando à redundância de alguns dados.

7 Conclusões e Perspectivas

De maneira geral, o presente trabalho reiterou a heterogeneidade das proteínas da família DNAJ através de uma abordagem filogenética abrangendo as subclasses A, B e C de diferentes organismos da escala evolutiva. Mais especificamente, demonstrou-se filogeneticamente que proteínas do tipo IV são pontualmente distintas das 3 subclasses e talvez devessem passar por uma re-classificação, principalmente considerando trabalhos na literatura demonstrando que estas proteínas não são capazes de estimular a atividade ATPase de proteínas HSP70.

Considerando a presença de politomias nas reconstruções filogenéticas e a tentativa de alcançar árvores melhor resolvidas, as etapas de clusterização dos dados e curadoria dos alinhamentos precisam ser revisadas como uma alternativa de melhoramento das reconstruções filogenéticas. Entretanto, é necessário considerar que as politomias das árvores provenientes do alinhamento do domínio-J possam ser resultado do alto nível de conservação desta região, levando à incapacidade de resolução de algumas bifurcações.

Quanto à organização das 3 subclasses em uma mesma árvore, a concentração de proteínas das subclasses A e B em um grande clado é um dado interessante que deve ser explorado de forma a avaliar quais características estão envolvidas neste agrupamento. Análises prospectivas podem explorar características como quantidade de ligantes através do *software* STRING (`https://string-db.org`) e predição de localização subcelular.

Ainda não foram geradas reconstruções filogenéticas com as 2702 proteínas totais (subclasses A, B e C) e/ou com as 2181 proteínas da subclasse C provenientes da clusterização com um *cut-off* de 80%. De acordo com os resultados das árvores das subclasses A e B após esta clusterização, espera-se que estas reconstruções filogenéticas sejam mais informativas e apresentem uma melhor subdivisão dos reinos. Além disso, faz-se necessária a comparação do árvore representada na Figura 14 com a nova árvore das três subclasses para confirmar a tendência de concentração das subclasses A e B em um grande clado ou para averiguar a formação de uma nova dispersão.

Referências

[1] BALCHIN, D.; HAYER-HARTL, M.; HARTL, F. U. In vivo aspects of protein folding and quality control. *Science (New York, N.Y.)*, v. 353, n. 6294, p. aac4354, 2016.

[2] MUSSKOPF, M. K.; MATTOS, E. P. D.; BERGINK, S. HSP40 / DNAJ Chaperones. *eLS © 2018, John Wiley & Sons, Ltd.*, p. 1–11, 2018.

[3] RITOSSA, F. Discovery of the heat shock response. *Cell Stress & Chaperones*, v. 1, n. 2, p. 97, 1996.

[4] ELLIS, R. J. Assembly chaperones: a perspective. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, v. 368, n. 1617, p. 20110398, 2013.

[5] JOAZEIRO, C. A. Ribosomal Stalling During Translation: Providing Substrates for Ribosome-Associated Protein Quality Control. *Annual Review of Cell and Developmental Biology*, v. 33, n. 1, p. annurev–cellbio–111315–125249, 2017.

[6] HARTL, F. U.; HAYER-HARTL, M. Converging concepts of protein folding in vitro and in vivo. *Nature Structural & Molecular Biology*, v. 16, n. 6, p. 574–581, 2009.

[7] AGARRABERES, F. A.; DICE, J. F. A molecular chaperone complex at the lysosomal membrane is required for protein translocation. *Journal of cell science*, v. 114, n. Pt 13, p. 2491–2499, 2001.

[8] GOLDBERG, A. L. Protein degradation and protection against misfolded or damaged proteins. *Nature*, London, v. 426, n. 6968, p. 895–899, 2003.

[9] CRAIG, E. A.; MARSZALEK, J. How Do J-Proteins Get Hsp70 to Do So Many Different Things? *Trends in Biochemical Sciences*, v. 42, n. 5, p. 355–368, 2017.

[10] KAMPINGA, H. H.; CRAIG, E. A. The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nature reviews. Molecular cell biology*, v. 11, n. 8, p. 579–92, aug 2010.

[11] QIU, X. B.; SHAO, Y. M.; MIAO, S.; WANG, L. The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. *Cellular and Molecular Life Sciences*, v. 63, n. 22, p. 2560–2570, 2006.

[12] BREHMER, D.; RÜDIGER, S.; GÄSSLER, C. S.; KLOSTERMEIER, D.; PACKSCHIES, L.; REINSTEIN, J.; MAYER, M. P.; BUKAU, B. Tuning of chaperone activity of Hsp70 proteins by modulation of nucleotide exchange. *Nature structural biology*, v. 8, n. 5, p. 427–432, 2001.

[13] TSAI, J.; DOUGLAS, M. G. A conserved HPD sequence of the J-domain is necessary for YDJ1 stimulation of Hsp70 ATPase activity at a site distinct from substrate binding. *Journal of Biological Chemistry*, v. 271, n. 16, p. 9347–9354, 1996.

[14] JIANG, J.; MAES, E. G.; TAYLOR, A. B.; WANG, L.; HINCK, A. P.; LAFER, E. M.; SOUSA, R. Structural Basis of J Cochaperone Binding and Regulation of Hsp70. *Molecular Cell*, v. 28, n. 3, p. 422–433, 2007.

[15] QIAN, Y. Q.; PATEL, D.; HARTL, F. U.; MCCOLL, D. J. Nuclear magnetic resonance solution structure of the human Hsp40 (HDJ- 1) J-domain. *J Mol Biol*, v. 260, n. 2, p. 224–235, 1996.

[16] BORK, P.; SANDER, C.; VALENCIA, A.; BUKAU, B. A module of the DnaJ heat shock proteins found in malaria parasites. *Trends in Biochemical Sciences*, v. 17, n. 4, p. 129, 1992.

[17] LI, J.; QIAN, X.; SHA, B. The Crystal Structure of the Yeast Hsp40 Ydj1 Complexed with Its Peptide Substrate. *Structure*, v. 11, n. 12, p. 1475–1483, 2003.

[18] WALSH, P.; BURSAĆ, D.; LAW, Y. C.; CYR, D.; LITHGOW, T. The J-protein family: Modulating protein assembly, disassembly and translocation. *EMBO Reports*, v. 5, n. 6, p. 567–571, 2004.

[19] LI, J.; QIAN, X.; SHA, B. Heat shock protein 40: structural studies and their functional implications. *Protein and peptide letters*, v. 16, n. 6, p. 606–612, 2009.

[20] RAJAN, V. B. V.; D'SILVA, P. Arabidopsis thaliana J-class heat shock proteins: Cellular stress sensors. *Functional and Integrative Genomics*, v. 9, n. 4, p. 433–446, 2009.

[21] HENNESSY, F.; CHEETHAM, M. E.; DIRR, H. W.; BLATCH, G. L. Analysis of the levels of conservation of the J domain among the various types of DnaJ-like proteins. *Cell stress & chaperones*, v. 5, n. 4, p. 347–58, 2000.

[22] VOS, M. J.; HAGEMAN, J.; CARRA, S.; KAMPINGA, H. H. Structural and functional diversities between members of the human HSPB, HSPH, HSPA, and DNAJ Chaperone Families. *HSPB, HSPH, HSPA, and DNAJ chaperone families. Biochemistry*, v. 47, p. 7001–7011, 2008.

[23] NILLEGODA, N. B.; KIRSTEIN, J.; SZLACHCIC, A.; BERYNSKYY, M.; STANK, A.; STENGEL, F.; ARNSBURG, K.; GAO, X.; SCIOR, A.; AEBERSOLD, R.; GUILBRIDE, D. L.; WADE, R. C.; MORIMOTO, R. I.; MAYER, M. P.; BUKAU, B. Crucial HSP70 co-chaperone complex unlocks metazoan protein disaggregation. *Nature*, London, v. 524, n. 7564, p. 247–251, aug 2015.

[24] NILLEGODA, N. B.; STANK, A.; MALINVERNI, D.; ALBERTS, N.; SZLACHCIC, A.; BARDUCCI, A.; De Los Rios, P.; WADE, R. C.; BUKAU, B. Evolution of an intricate J-protein network driving protein disaggregation in eukaryotes. *eLife*, v. 6, p. e24560, 2017.

[25] MOKRANJAC, D.; BOURENKOV, G.; HELL, K.; NEUPERT, W.; GROLL, M. Structure and function of Tim14 and Tim16, the J and J-like components of the mitochondrial protein import motor. *The EMBO journal*, v. 25, n. 19, p. 4675–85, oct 2006.

[26] D'SILVA, P. R.; SCHILKE, B.; WALTER, W.; CRAIG, E. A. Role of Pam16's degenerate J domain in protein import across the mitochondrial inner membrane. *Proceedings of the National Academy of Sciences of the United States of America*, Washington, v. 102, n. 35, p. 12419–24, 2005.

[27] LI, Y.; DUDEK, J.; GUIARD, B.; PFANNER, N.; REHLING, P.; VOOS, W. The presequence translocase-associated protein import motor of mitochondria: Pam16 functions in an antagonistic manner to Pam18. *Journal of Biological Chemistry*, v. 279, n. 36, p. 38047–38054, 2004.

[28] DEKKER, S. L.; KAMPINGA, H. H.; BERGINK, S. DNAJs: more than substrate delivery to HSPA. *Frontiers in molecular biosciences*, v. 2, n. June, p. 35, 2015.

[29] FOTIN, A.; CHENG, Y.; GRIGORIEFF, N.; WALZ, T.; HARRISON, S. C.; KIRCHHAUSEN, T. Structure of an auxilin-bound clathrin coat and its implications for the mechanism of uncoating. *Nature*, London, v. 432, n. 7017, p. 649–53, dec 2004.

[30] LILL, R.; DUTKIEWICZ, R.; FREIBERT, S. A.; HEIDENREICH, T.; MASCARENHAS, J.; NETZ, D. J.; PAUL, V. D.; PIERIK, A. J.; RICHTER, N.; STÜMPFIG, M.; SRINIVASAN, V.; STEHLING, O.; MÜHLENHOFF, U. The role of mitochondria and the CIA machinery in the maturation of cytosolic and nuclear iron-sulfur proteins. *European Journal of Cell Biology*, v. 94, n. 7-9, p. 280–291, 2015.

[31] BEHNKE, J.; MANN, M. J.; SCRUGGS, F. L.; FEIGE, M. J.; HENDERSHOT, L. M. Members of the Hsp70 Family Recognize Distinct Types of Sequences to Execute ER Quality Control. *Molecular Cell*, v. 63, n. 5, p. 739–752, 2016.

[32] KAKKAR, V.; MANSSON, C.; DE MATTOS, E. P.; BERGINK, S.; VAN DER ZWAAG, M.; VAN WAARDE, M. A. W. H.; KLOOSTERHUIS, N. J.; MELKI, R.; VAN CRUCHTEN, R. T. P.; AL-KARADAGHI, S.; AROSIO, P.; DOBSON, C. M.; KNOWLES, T. P. J.; BATES, G. P.; VAN DEURSEN, J. M.; LINSE, S.; VAN DE SLUIS, B.; EMANUELSSON, C.; KAMPINGA, H. H. The S/T-Rich Motif in the

DNAJB6 Chaperone Delays Polyglutamine Aggregation and the Onset of Disease in a Mouse Model. *Molecular Cell*, v. 62, n. 2, p. 272–283, 2016.

[33] BORNBERG-BAUER, E.; ALBÀ, M. M. Dynamics and adaptive benefits of modular protein evolution. *Current Opinion in Structural Biology*, v. 23, n. 3, p. 459–466, 2013.

[34] BUSTARD, K.; GUPTA, R. S. The sequences of heat shock protein 40 (DNAJ) homologs provide evidence for a close evolutionary relationship between the deinococcus-thermus group and cyanobacteria. *Journal of Molecular Evolution*, v. 45, n. 2, p. 193–205, 1997.

[35] OHTSUKA, K.; HATA, M. Mammalian HSP40/DNAJ homologs: cloning of novel cDNAs and a proposal for their classification and nomenclature. *Cell stress & chaperones*, v. 5, n. 2, p. 98–112, 2000.

[36] NYDAM, M. L.; HOANG, T. A.; SHANLEY, K. M.; De Tomaso, A. W. Molecular evolution of a polymorphic HSP40-like protein encoded in the histocompatibility locus of an invertebrate chordate. *Developmental and Comparative Immunology*, v. 41, n. 2, p. 128–136, 2013.

[37] SARKAR, N. K.; THAPAR, U.; KUNDNANI, P.; PANWAR, P.; GROVER, A. Functional relevance of J-protein family of rice (Oryza sativa). *Cell Stress and Chaperones*, v. 18, n. 3, p. 321–331, 2013.

[38] CHEN, D. H.; HUANG, Y.; LIU, C.; RUAN, Y.; SHEN, W. H. Functional conservation and divergence of J-domain-containing ZUO1/ZRF orthologs throughout evolution. *Planta*, v. 239, n. 6, p. 1159–1173, 2014.

[39] LI, Y.; BU, C.; LI, T.; WANG, S.; JIANG, F.; YI, Y.; YANG, H.; ZHANG, Z. Cloning and analysis of DnaJ family members in the silkworm, Bombyx mori. *Gene*, v. 576, n. 1 Pt 1, p. 88–98, jan 2016.

[40] MORITA, Y.; MARUYAMA, S.; KABEYA, H.; NAGAI, A.; KOZAWA, K.; KATO, M.; NAKAJIMA, T.; MIKAMI, T.; KATSUBE, Y.; KIMURA, H. Genetic diversity of the dnaJ gene in the Mycobacterium avium complex. *Journal of Medical Microbiology*, v. 53, n. 8, p. 813–817, 2004.

[41] NHUNG, P. H.; SHAH, M. M.; OHKUSU, K.; NODA, M.; HATA, H.; SUN, X. S.; IIHARA, H.; GOTO, K.; MASAKI, T.; MIYASAKA, J.; EZAKI, T. The dnaJ gene as a novel phylogenetic marker for identification of Vibrio species. *Systematic and Applied Microbiology*, v. 30, n. 4, p. 309–315, 2007.

[42] HONG NHUNG, P.; OHKUSU, K.; MISHIMA, N.; NODA, M.; MONIR SHAH, M.; SUN, X.; HAYASHI, M.; EZAKI, T. Phylogeny and species identification of the family Enterobacteriaceae based on dnaJ sequences. *Diagnostic Microbiology and Infectious Disease*, v. 58, n. 2, p. 153–161, 2007.

[43] ALEXANDRE, A.; LARANJO, M.; YOUNG, J. P. W.; OLIVEIRA, S. DnaJ is a useful phylogenetic marker for alphaproteobacteria. *International Journal of Systematic and Evolutionary Microbiology*, v. 58, n. 12, p. 2839–2849, 2008.

[44] MACARIO, ALBERTO J L, M. M.; CONWAY DE MACARIO, E. Wadsworth center, new york state department of health, division of molecular medicine; and 2 department of biomedical sciences, school of public health, the university at albany (suny), albany, new york, usa, 3 current address: Johannes gutenberg universi. v. 70, p. 1318–1332, 2004.

[45] HAGEMAN, J.; KAMPINGA, H. H. Computational analysis of the human HSPH/HSPA/DNAJ family and cloning of a human HSPH/HSPA/DNAJ expression library. *Cell Stress and Chaperones*, v. 14, n. 1, p. 1–21, 2009.

[46] HAGEMAN, J.; VAN WAARDE, M. A. W. H.; ZYLICZ, A.; WALERYCH, D.; KAMPINGA, H. H. The diverse members of the mammalian HSP70 machine show distinct chaperone-like activities. *The Biochemical journal*, v. 435, n. 1, p. 127–142, 2011.

[47] PETITJEAN, C.; MOREIRA, D.; LÓPEZ-GARCÍA, P.; BROCHIER-ARMANET, C. Horizontal gene transfer of a chloroplast DnaJ-Fer protein to Thaumarchaeota and the evolutionary history of the DnaK chaperone system in Archaea. *BMC evolutionary biology*, v. 12, p. 226, 2012.

[48] RAJARAM, H.; CHAURASIA, A. K.; APTE, S. K. Cyanobacterial heat-shock response: Role and regulation of molecular chaperones. *Microbiology (United Kingdom)*, v. 160, n. PART 4, p. 647–658, 2014.

[49] CRAIG, E. A.; HUANG, P.; ARON, R.; ANDREW, A. The diverse roles of J-proteins, the obligate Hsp70 co-chaperone. *Reviews of Physiology, Biochemistry and Pharmacology*, p. 1–21, 2006.

[50] VERMA, A. K.; DIWAN, D.; RAUT, S.; DOBRIYAL, N.; BROWN, R. E.; GOWDA, V.; HINES, J. K.; SAHI, C. Evolutionary Conservation and Emerging Functional Diversity of the Cytosolic Hsp70:J Protein Chaperone Network of Arabidopsis thaliana. *G3 (Bethesda, Md.) Genes—Genomes—Genetics*, v. 7, n. June, p. g3.117.042291, apr 2017.

[51] RATHEESH KUMAR, R.; NAGARAJAN, N. S.; ARUNRAJ, S. P.; SINHA, D.; VEEDIN RAJAN, V. B.; ESTHAKI, V. K.; D'SILVA, P. HSPIR: A manually annotated heat shock protein information resource. *Bioinformatics*, v. 28, n. 21, p. 2853–2855, 2012.

[52] HUANG, Y.; NIU, B.; GAO, Y.; FU, L.; LI, W. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, v. 26, n. 5, p. 680–682, 2010.

[53] ZHANG, L.; ZHANG, C.; GAO, R.; YANG, R. JPPRED: Prediction of Types of J-Proteins from Imbalanced Data Using an Ensemble Learning Method. *BioMed Research International*, v. 2015, p. 1–12, 2015.

[54] YOON, B.-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, v. 10, n. 6, p. 402–415, 2009.

[55] PACHTER, L.; ALEXANDERSSON, M.; CAWLEY, S. Applications of Generalized Pair Hidden Markov Models to Alignment and Gene Finding Problems. *Journal of Computational Biology*, v. 9, n. 2, p. 389–399, 2002.

[56] FINN, R. D.; COGGILL, P.; EBERHARDT, R. Y.; EDDY, S. R.; MISTRY, J.; MITCHELL, A. L.; POTTER, S. C.; PUNTA, M.; QURESHI, M.; SANGRADOR-VEGAS, A.; SALAZAR, G. A.; TATE, J.; BATEMAN, A. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, v. 44, n. D1, p. D279–D285, 2016.

[57] KATOH, K.; ROZEWICKI, J.; YAMADA, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, , n. May, p. 1–7, 2017.

[58] WATERHOUSE, A. M.; PROCTER, J. B.; MARTIN, D. M.; CLAMP, M.; BARTON, G. J. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, v. 25, n. 9, p. 1189–1191, 2009.

[59] CLAMP, M.; CUFF, J.; SEARLE, S. M.; BARTON, G. J. The Jalview Java alignment editor. *Bioinformatics*, v. 20, n. 3, p. 426–427, 2004.

[60] GUINDON, S.; DUFAYARD, J.-F.; LEFORT, V.; ANISIMOVA, M.; HORDIJK, W.; GASCUEL, O. New Algorithms and Mehtods to Estimate Maximum-Likelihood Phylogenies: Asessing the Performance of PhyML 2.0. *Systematic Biology*, v. 59, n. 3, p. 307–321, 2010.

[61] GUINDON, S.; LETHIEC, F.; DUROUX, P.; GASCUEL, O. PHYML Online - A web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Research*, v. 33, n. SUPPL. 2, p. 557–559, 2005.

[62] LE, S. Q.; GASCUEL, O. An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, v. 25, n. 7, p. 1307–1320, 2008.

[63] LEFORT, V.; LONGUEVILLE, J. E.; GASCUEL, O. SMS: Smart Model Selection in PhyML. *Molecular biology and evolution*, v. 34, n. 9, p. 2422–2424, 2017.

[64] ANISIMOVA, M.; GASCUEL, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology*, v. 55, n. 4, p. 539–552, 2006.

[65] LETUNIC, I.; BORK, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research*, v. 44, n. W1, p. W242–W245, 2016.

[66] HESS, P. N.; DE MORAES RUSSO, C. A. An empirical test of the midpoint rooting method. *Biological Journal of the Linnean Society*, v. 92, n. 4, p. 669–674, dec 2007.

[67] DAVIDSON, R.; SULLIVANT, S. Distance-Based Phylogenetic Methods Around a Polytomy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, v. 11, n. 2, p. 325–335, 2014.

[68] PURVIS, A.; GARLAND, T. Polytomies in comparative analyses of continuous characters. *Systematic Biology*, v. 42, n. 4, p. 569–575, 1993.

[69] WALSH, H. E.; KIDD, M. G.; MOUM, T.; FRIESEN, V. L. Polytomies and the Power of Phylogenetic Inference. *Evolution*, v. 53, n. 3, p. 932, 1999.

Apêndices

A

Diferentes táxons após clusterização com *cut-off* de identidade de 40%

Quantidade de proteínas de cada subtipo nos diferentes táxons após clusterização com *cut-off* de identidade de 40%.

Tipo	Bacteria	Arkaea	Algae	Protista	Fungi	Plantae	Animalia
I	10	3	1	22	15	4	8
II	18	0	9	12	4	7	5
III	133	17	131	246	178	182	174
IV	0	0	3	4	6	2	5
Total	161	20	144	284	203	195	192

B

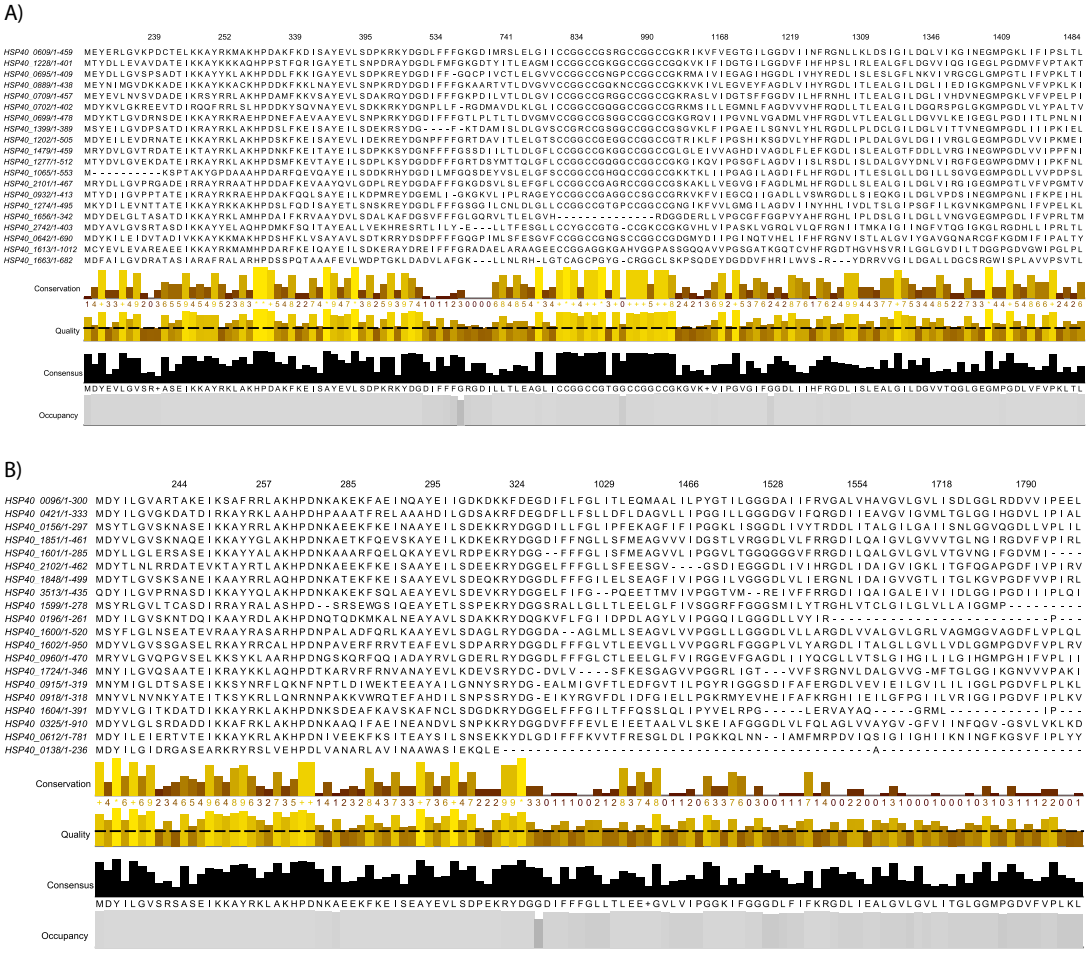
Diferentes táxons após clusterização com *cut-off* de identidade de 80%

Quantidade de proteínas de cada subtipo nos diferentes reinos após clusterização com *cut-off* de identidade de 80%.

Tipo	Eubacteria	Arkaea	Algae	Protista	Fungi	Plantae	Animalia
I	103	14	9	40	70	33	48
II	57	0	15	21	17	41	47
III	241	23	193	297	402	469	501
IV	0	0	5	6	16	5	20
Total	401	37	222	364	505	548	616

C

Alinhamento múltiplo de proteínas dos tipos I ou II



Alinhamento múltiplo de toda a extensão da proteína do (A) tipo I ou (B) tipo II contendo apenas as colunas com valor de qualidade acima de 40% (representação de apenas 20 proteínas)

D

Alinhamento múltiplo de proteínas dos tipos III e IV

Alinhamento múltiplo das sequências correspondentes ao domínio-J das proteínas do tipo I e IV contendo apenas as colunas com valor de qualidade acima de 30% (representação de apenas 40 proteínas)

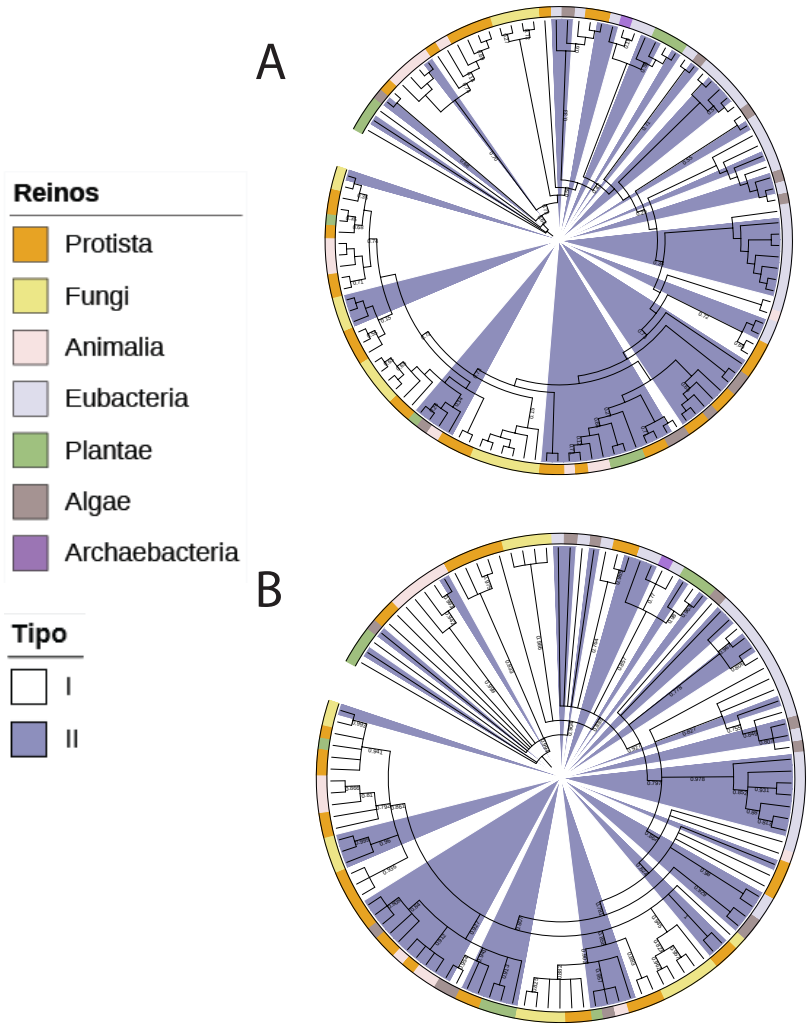
E

Alinhamento múltiplo de proteínas dos tipos I, II, III e IV

Alinhamento múltiplo das sequências correspondentes ao domínio-J de todas as proteínas (tipos I, II, III e IV) contendo apenas as colunas com valor de qualidade acima de 30% (representação de apenas 40 proteínas)

F

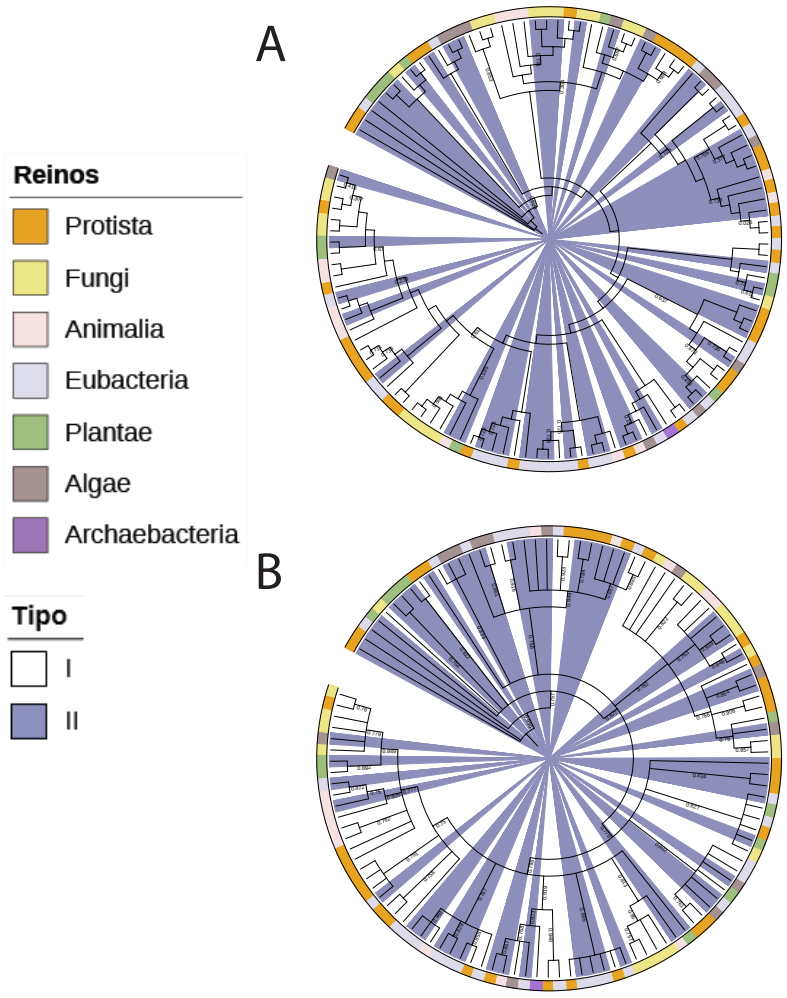
Politomias presentes nas árvores provenientes do alinhamento de toda a extensão da proteína



Comparação das árvores das subclasses A e B após clusterização com um *cut-off* de 40% e alinhamento de toda a extensão da proteína. A) Representação dos ramos originais e indicação dos escores abaixo de 0.75. B) Representação dos ramos apenas com escores acima de 0.75 e indicação dos respectivos valores.

G

Politomias presentes nas árvores provenientes do alinhamento do domínio-J



Comparação das árvores das subclasses A e B após clusterização com um *cut-off* de 40% e alinhamento do domínio-J. A) Representação dos ramos originais e indicação dos escores abaixo de 0.75. B) Representação dos ramos apenas com escores acima de 0.75 e indicação dos respectivos valores.