UNIVERSIDADE FEDERAL DO PARANÁ

CAMILLA REGINATTO DE PIERRI

REPRESENTAÇÕES VETORIAIS DE PROTEOMAS: UM ESTUDO DE CASO COM SEQUÊNCIAS MITOCONDRIAIS

CURITIBA

2017

CAMILLA REGINATTO DE PIERRI

REPRESENTAÇÕES VETORIAIS DE PROTEOMAS: UM ESTUDO DE CASO COM SEQUÊNCIAS MITOCONDRIAIS

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, área de concentração em Bioinformática.

Orientador: Prof. Dr. Roberto Tadeu Raittz
Coorientador: Prof. Dr. Mauro Antônio Alves Castro

CURITIBA

2017

Pierri, Camilla Reginatto de
P623 Representações vetoriais de proteomas: um estudo de caso com sequências mitocondriais / Camilla Reginatto de Pierri. - Curitiba, 2017.
88 f.; il.

Orientador: Roberto Tadeu Raittz
Coorientador: Mauro Antônio Alves Castro
Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Curso de Pós-Graduação em Bioinformática.
Inclui Bibliografia.

1. Filogenia. 2. Proteoma Mitocondrial . 3. Bioinformática. I. Raittz, Roberto Tadeu. II. Castro, Mauro Antônio Alves. III. Título. IV. Universidade Federal do Paraná.

CDD 574.87342

TERMO DE APROVAÇÃO

CAMILLA REGINATTO DE PIERRI

REPRESENTAÇÕES VETORIAIS DE PROTEOMAS: UM ESTUDO DE CASO COM SEQUÊNCIAS MITOCONDRIAIS

Dissertação aprovada como requisito parcial para obtenção de grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

ROBERTO TADEU RAITTZ
Presidente/Programa de Pós-graduação em Bioinformática – UFPR

IRIS HASS
Avaliadora Externa/Departamento de Genética – UFPR

DIEVAL GUIZELINI
Avaliador Interno/Programa de Pós-graduação em Bioinformática – UFPR

FABIO DE OLIVEIRA PEDROSA
Avaliador Interno/Departamento de Bioquímica e
Programa de Pós-graduação em Bioinformática – UFPR

Curitiba, 15 de maio 2017.

Aos meus pais, pelo apoio incondicional em toda minha vida acadêmica e por não medirem esforços para eu conseguir alcançar meus objetivos.

À minha avó Doca, que mesmo em outro plano, me visita em sonho e me incentiva a estudar, pois segundo ela “saber não ocupa espaço”.

AGRADECIMENTOS

Finalmente é chegada a tão esperada conclusão do mestrado. Foram dois anos de dedicação, aprendizado, amizade, café e muitas, muitas risadas. Em toda minha caminhada acadêmica, eu nunca fiz parte de uma equipe que eu gostasse tanto e, sem nunca perder o bom humor, me acrescentou um conhecimento inimaginável. Sou muito grata por isso.

Grata em especial ao meu querido amigo e orientador Prof. Dr. Roberto Tadeu Raittz, que de uma maneira especial e única, me ensinou, auxiliou e me aconselhou sempre que precisei. Não posso deixar de citar aqui a notável equipe que fiz parte... meu coorientador Prof. Dr. Mauro Antônio Alves Castro, meus colegas Ricardo Voyceik, Aryel Marlus Repula de Oliveira, Bruno Thiago de Lima Nichio, Josué Oliveira Camargo, Letícia Graziela Costa Santos, Amanda Wilczek e Mariane Gonçalves.

Todos os professores do programa de Pós-graduação em Bioinformática merecem destaque, pois sem o conhecimento transmitido por eles, eu jamais teria chegado aqui. Minha querida amiga Suzana, secretária da Bioinformática, também foi muito importante, pois sempre esteve disponível a me auxiliar. Meus colegas da UFMG Prof. Dr. José Miguel Ortega, Dr. Tetsu Sakamoto e Me. Nilson Antônio da Rocha Coimbra, que durante meu estágio, me receberam muito bem e acrescentaram muito ao meu aprendizado.

Sempre tive o apoio incondicional dos meus pais. Não caberia aqui toda a minha gratidão e nem existem palavras para expressar o quão importante eles são em minha vida. São meus exemplos, meu espelho. Minha irmã também foi fundamental nesta minha caminhada, que sempre com alto astral contagiante, me animou quando precisei.

Sou muito agradecida à Deus por ter posto em meu caminho meu esposo e colega de mestrado Antonio Camilo da Silva Filho, que de maneira recíproca, a compreensão, colaboração, amizade e amor foi o que nos direcionou ao sucesso.

Por fim e tão importante quanto, gostaria de agradecer à CAPES pela concessão de minha bolsa de estudos, à Universidade Federal do Paraná e ao Programa de Pós-graduação em Bioinformática pela oportunidade de ampliar meus conhecimentos.

“It is our choices that show who we truly are, far more than our abilities”.

*(*Albus Dumbledore - *Harry Potter and the Chamber of Secrets*)

J.K Rowling

RESUMO

Grande parte dos estudos evolutivos para inferir ancestralidade são conduzidos utilizando apenas alguns genes mitocondriais. Filogenias baseadas em um único gene podem não conter informações suficientes para construir história evolutiva de determinados organismos. Utilizar genomas mitocondriais completos para análises evolutivas fornecem mais informações que a utilização de proteínas individuais, porém, dependendo do número de organismos, gera alto custo computacional. Não existem propostas até o momento referentes a filogenias derivadas de projeções em espaços vetoriais com redução de dimensão que representem o proteoma de organismos. A abordagem proposta neste trabalho, desenvolvida em ambiente MatLab® é inovadora e resolve problemas associados ao custo computacional na execução de filogenias. Utilizando os dados de 6.811 organismos depositados no RefSeq, realizamos inicialmente a clusterização dos dados, utilizando a ferramenta RAFTS3groups. Após tratamento dos dados, propomos uma estratégia de representação vetorial baseada em k-mers espaçados, utilizando janela deslizante com tamanho de 5 aminoácidos com 1 descontinuado. A partir disso, foi gerada uma matriz de co-ocorrência de 400x400 para cada organismo, representando o proteoma mitocondrial. Esta matriz foi disposta em um vetor de 160.000 atributos, o qual é utilizado para gerar representações vetoriais com redução de dimensão de 100, 400 e 800 coordenadas. Essas vetorizações são representadas em árvores filogenéticas e comparada com filogenia de alinhamento. Elaboramos um algoritmo baseado em UPGMA para realizar árvores filogenéticas e analisamos o proteoma mitocondrial dos Jakobidas e dos Homínidios. A estratégia de extração de atributos e representação vetorial do proteoma se mostrou eficiente para evidenciar relações de parentesco, sendo as filogenias vetorizadas correlacionadas com a filogenia de alinhamento.

Palavras-chave: Proteoma Mitocondrial, Filogenia, Bioinformática.

ABSTRACT

Much of the evolutionary studies to infer ancestry are conducted using only a few mitochondrial genes. Phylogenies based on a single gene may not contain enough information to construct evolutionary history of certain organisms. Using complete mitochondrial genomes for evolutionary analysis provides more information than the use of individual proteins, however, depending on the number of organisms, it generates a high computational cost. There are no proposals to date regarding phylogenies derived from projections in vector spaces with size reduction that represent the proteome of organisms. The approach proposed in this work, developed in MatLab® environment, is innovative and solves problems associated with computational cost in the execution of phylogenies. Using data from 6,811 organisms deposited in the RefSeq, we initially performed data clustering using the RAFTS3groups tool. After data processing, we propose a vector representation strategy based on spaced k-mers, using sliding window with size of 5 amino acids with 1 discontinued. From this, a co-occurrence matrix of 400x400 was generated for each organism, representing the mitochondrial proteome. This matrix was arranged in a vector of 160,000 attributes, which is used to generate vector representations with size reduction of 100, 400 and 800 coordinates. These vectorizations are represented in phylogenetic trees and compared with phylogeny of alignment. We developed an algorithm based on UPGMA to perform phylogenetic trees and analyzed the mitochondrial proteome of Jakobids and Hominids. The strategy of attribute extraction and vector representation of the proteome proved to be efficient to evidence kinship relations, with the phylogenies vectored correlated with the phylogeny of alignment.

Keywords: Mitochondrial Proteome, Phylogeny, Bioinformatics.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ATP	- Adenosina-trifosfato
BLAST	- *Basic Local Alignment Search Tool*
BLOSUM	- *BLOcks of amino acid SUbstitution Matrix*
COX	- Citocromo c oxidase subunidade
CYTB	- Citocromo b
DNA	- Ácido desoxirribonucleico
faa	- Arquivo fasta em formato de aminoácido
FFP	- *Feature Frequence Profile*
FTP	- *File Transfer Protocol*
GenBank	- *NIH genetic sequence database*
HMM	- Cadeia oculta de Markov
KEGG	- *Kyoto Encyclopedia of Genes and Genomes*
MATLAB	- *Matrix Laboratory*
mRNA	- RNA mensageiro
MRP	- *Matrix Representation with Parsimony analysis*
MtDNA	- DNA mitocondrial
NADH	- NADH-desidrogenase
NCBI	- *National Center for Biotechnology Information*
ORF	- *Open reading frame*
PAM	- *Point Accepted Mutation*
pb	- Pares de bases
Pdist	- Distância *pairwise*
PIR	- *Protein Information Resource*
RAFTS3	- *Rapid Alignment-Free Tool for Sequence Similarity Search*
RAFTS3groups	- *Rapid Alignment Free Tool for Sequences Similarity Search to Groups*
RefSeq	- *NCBI reference sequence*
RNA	- Ácido ribonucleico
RPL	- Proteína ribossomal subunidade grande
RPS	- Proteína ribossomal subunidade pequena
rRNA	- RNA ribossomal
RuBisCO	- Ribulose-1,5-bisfosfato carboxilase oxigenasse

tRNA	- RNA transportador
UniParc	- *UniProt Archive*
UniProt	- *Universal Protein Resource*
UniRef	- *UniProt Reference Clusters*
UPGMA	- *Unweighted Pair Group Method using Arithmetic averages*
W160k	- Vetor de 160.000 coordenadas
Wr100	- Vetor de 100 coordenadas
Wr400	- Vetor de 400 coordenadas
Wr800	- Vetor de 800 coordenadas

SUMÁRIO

1 INTRODUÇÃO

As mitocôndrias são organelas citoplasmáticas alongadas responsáveis principalmente pela produção de adenosina-trifosfato (ATP), presentes em células eucarióticas. Essas organelas possuem genoma próprio, podendo variar sua composição, estrutura e número de cópias dependendo do organismo (ZAHA; FERREIRA; PASSAGLIA, 2014).

Segundo a teoria endossimbiótica, as mitocôndrias evoluíram a partir de uma alfa-preteobactéria que foi fagocitada por uma célula de um organismo eucariótico primitivo, que com o decorrer do tempo, a relação de simbiose estabelecida entre organismo e hospedeiro, tornou-se irreversível (GRAY et al., 1999). O momento de aquisição das mitocôndrias durante o processo evolutivo dos eucariotos ainda é debate no meio científico (GRAY, 2015; PITTIS; GABALDÓN, 2016; DEGLI ESPOSTI, 2016).

Com o surgimento da Bioinformática, grandes volumes de dados de sequências biológicas puderam ser explorados com maior facilidade. Os estudos evolutivos possibilitaram a representação de várias classes de organismos bem como a inferência da história evolutiva, por meio da construção de árvores filogenéticas (MENG et al., 2015). Esses estudos em eucariotos são conduzidos utilizando preferencialmente o DNA mitocondrial (mtDNA) por conta de duas características importantes: a herança materna mitocondrial em mamíferos e a alta taxa de mutação quando comparada com o DNA nuclear (MORITZ et al., 1987).

Grande parte dos estudos evolutivos utilizam apenas alguns genes mitocondriais para verificar ancestralidade, como por exemplo o gene 16S, CYTB e COX1 (MEYER 1994; HEBERT et al., 2003). Porém, filogenias baseadas em um único gene podem não conter informações suficientes para construir uma história evolutiva de determinados organismos, tendo em vista as diferentes taxas de evolução e transferência horizontal de genes (OTU; SAYOOD, 2003). Utilizar genomas mitocondriais completos para análises evolutivas fornecem mais informações que a utilização de proteínas individuais (GRAY et al., 1999; BURKI, 2014; HUG et al., 2016).

Os principais métodos utilizados em filogenia são baseados em alinhamento múltiplo, o que gera uma alta complexidade computacional. Esses

métodos se tornam inviáveis quando existe a necessidade de análise de genomas completos (OTU; SAYOOD, 2003; SONG et al., 2013; VINGA, 2014; BRINDA; SYKULSKI; KUCHEROV, 2015).

O acúmulo de genomas mitocondriais depositados nos bancos de dados biológicos gera oportunidades para estudar os reinos eucarióticos ao longo da história evolutiva mitocondrial, e quando associados com métodos de mineração de dados, resulta em análises ricas em conteúdo (GRAY, 2015). A construção de novas ferramentas de bioinformática exige um grande conhecimento em linguagem de programação. A manipulação de dados genéticos utilizando o MatLab®, segundo as experiências obtidas pelo nosso grupo de estudo, otimiza resultados de qualquer análise, por conta do grande número de recursos algorítmicos disponíveis, além da possibilidade de trabalhar com matrizes e vetores.

Tendo como fundamento a inviabilidade do alinhamento múltiplo sequências grandes e no alto custo computacional que a utilização de genomas completos exige para este propósito, o objetivo deste estudo foi identificar analisar o proteoma mitocondrial por meio de técnicas de mineração de dados e desenvolver um método de representação de sequências proteicas que melhore reconstruções filogenéticas.

Para isso, utilizamos a ferramenta de clusterização RAFTS3groups, desenvolvida pelo grupo, para extrair informações relevantes sobre o proteoma mitocondrial, valendo-se de todos os genomas mitocondriais encontrados no banco de dados *NCBI Reference Sequences* (RefSeq). Definimos uma estratégia de representação vetorial dos proteomas, implementada em k-mers espaçados (BODEN et al., 2013; LEIMEISTER et al., 2014; HORWEGE et al., 2014). Com a construção dos vetores contendo a informação dos proteomas mitocondriais, filogenias foram inferidas utilizando o método UPGMA (MICHENER; SOKAL, 1957), como forma de validar a representação proposta neste trabalho.

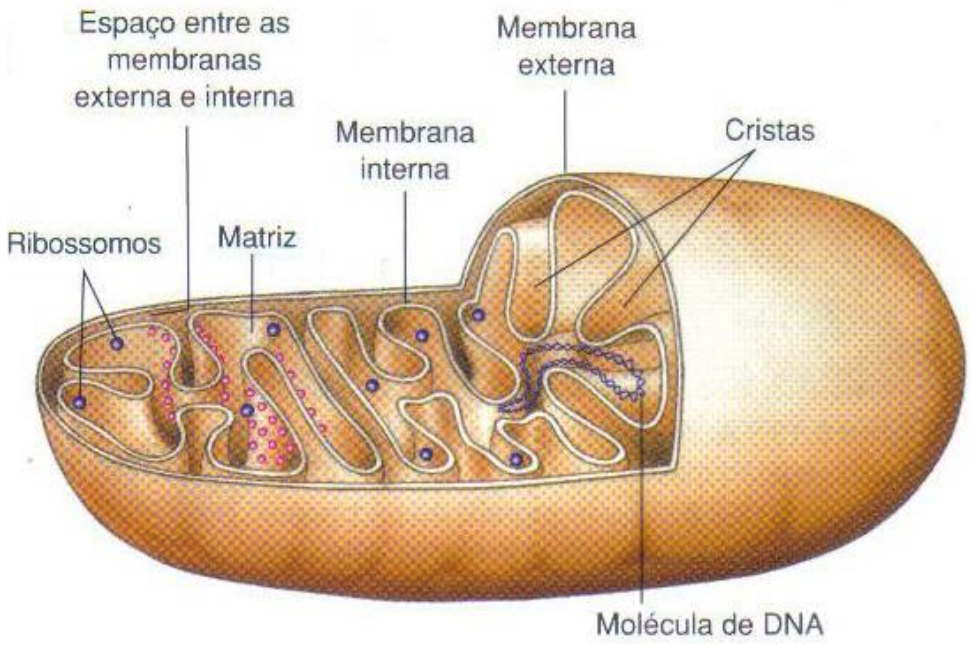
Quanto à manipulação de grandes volumes de dados com esforço computacional reduzido, não existe até o momento publicações referentes a filogenias derivadas de projeções em espaços vetoriais que representem o proteoma de organismos, tampouco filogenias com redução de dimensão de vetores. A abordagem proposta neste trabalho é inovadora e resolve problemas associados com falta de agilidade na execução de árvores filogenéticas.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 MITOCÔNDRIA

Mitocôndrias são organelas alongadas presentes no citoplasma de células eucarióticas. Grande parte das reações químicas do organismo ocorrem nas mitocôndrias, sendo a fosforilação oxidativa a principal função da organela, crucial para a funcionalidade da célula e do organismo (NELSON; COX, 2014). A figura abaixo (FIGURA 1) representa um esquema da estrutura mitocondrial.

FIGURA 1 - MITOCÔNDRIA E SEUS COMPONENTES PRINCIPAIS



FONTE: Adaptado de JUNQUEIRA; CARNEIRO, 2005.
NOTA: Descrição da figura – APÊNDICE 1.

Essas organelas possuem genoma próprio, que é diferente dos genomas nucleares em eucariotos. Neste genoma extranuclear, os genes são relacionados com a principal função da organela, que é a produção de ATP (ZAHA; FERREIRA; PASSAGLIA, 2014).

2.2.1 Cadeia transportadora de elétrons e fosforilação oxidativa

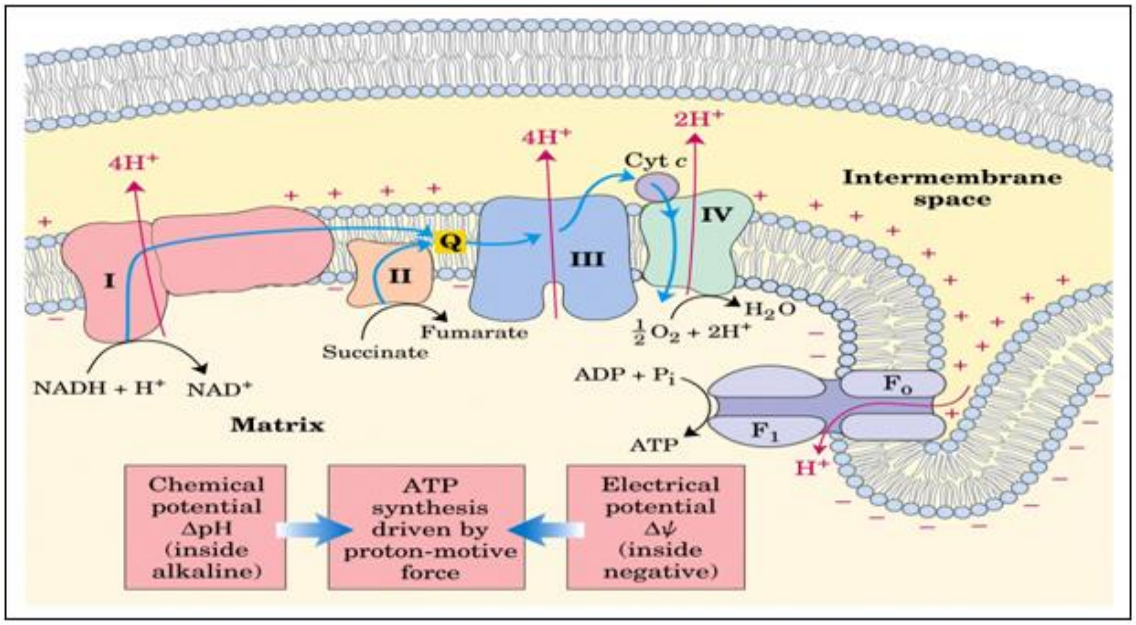
Para as células conseguirem realizar suas atividades necessitam de energia disponível. Essa energia é proveniente da quebra das ligações químicas dos carboidratos e das gorduras (JUNQUEIRA; CARNEIRO, 2005). A glicólise, a via inicial para a quebra de carboidratos, ocorre no citossol e produz pouco ATP, sendo insuficiente para manter os todos os processos metabólicos em grande parte dos organismos (HÜTTEMANN et al., 2007). Todas as atividades de decomposição de carboidratos e ácidos graxos convergem para a etapa final da fosforilação oxidativa.

A fosforilação oxidativa tem início com a entrada dos elétrons na cadeia transportadora de elétrons, denominada cadeia respiratória (NELSON; COX, 2014). Esta etapa fornece 15 vezes mais ATP que a glicólise (HÜTTEMANN et al., 2007). Para a energia ser utilizada, a produção deve ocorrer de forma gradual, por meio do transporte de elétrons através dos vários transportadores presentes na membrana interna mitocondrial, que compõem a cadeia transportadora de elétrons (COOPER, 2000).

Existem quatro complexos proteicos que servem como transportadores de elétrons, capazes de aceitar ou doar elétrons: Complexo I (NADH-desidrogenase), Complexo II (succinato desidrogenase), Complexo III (citocromo oxirredutase), e Complexo IV (citocromo c oxidase). Um quinto complexo proteico, denominado ATP sintetase, realiza a conversão de ATP (COOPER, 2000).

A energia liberada pelo transporte de elétrons através dos quatro primeiros complexos proteicos é utilizada para bombear prótons da matriz mitocondrial para o espaço intermembranoso, produzindo um gradiente de prótons. Esse gradiente auxilia na catálise rotacional gerada pela corrente de prótons através de uma subunidade do complexo V, levando à síntese de ATP (NELSON; COX, 2014). Este processo é representado na FIGURA 2.

FIGURA 2 - REPRESENTAÇÃO DA CADEIA TRANSPORTADORA DE ELÉTRONS E FOSFORILAÇÃO OXIDATIVA



FONTE: NELSON; COX (2000, p 748).
NOTA: Esquematização dos cinco complexos proteicos presentes na membrana interna mitocondrial.

Além da fosforilação oxidativa, as mitocôndrias ainda estão relacionadas com o ciclo de Krebs, processo de apoptose, envelhecimento celular e β-oxidação de ácidos graxos (NELSON; COX, 2014).

A β-oxidação de ácidos graxos, apoptose desregulada, bem como a falta de energia celular, desencadeiam mutações nos genes mitocondriais (HÜTTEMANN et al., 2007). Essas mutações estão relacionadas com diversas patologias, entre elas: perda auditiva, doença de Alzheimer e *diabetes mellitus* (ALCOLADO; THOMAS, 1995; GRAZINA et al., 2006; VIVERO et al., 2013).

2.2 CARACTERÍSTICAS DO GENOMA MITOCONDRIAL

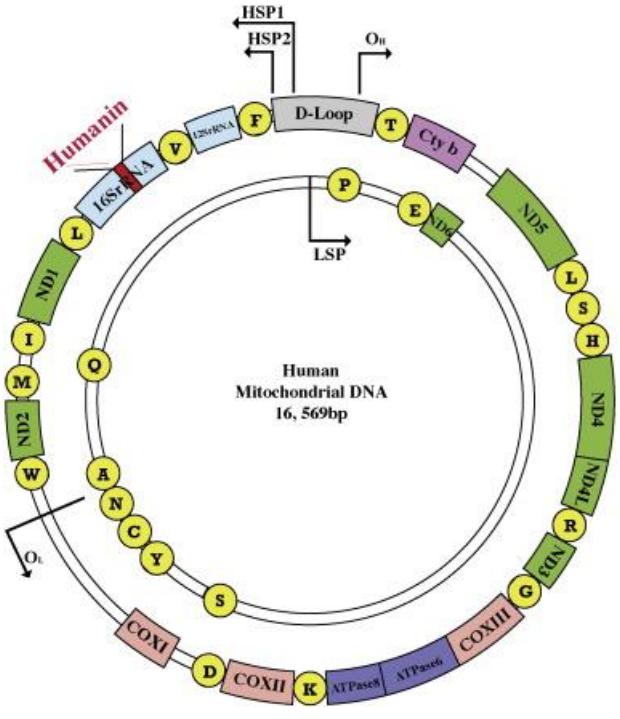
O genoma mitocondrial é uma molécula, na maioria das vezes, circular em DNA dupla fita. Cada mitocôndria possui aproximadamente cinco cópias deste DNA, variando em número de cópias, tamanho e composição gênica de espécie para espécie (NELSON; COX, 2014).

O conteúdo gênico do genoma mitocondrial de eucariotos não está diretamente relacionado com o tamanho do genoma, sendo variável o número de

proteínas e RNAs codificadores em cada espécie (SNUSTAD; SIMMONS, 2001). Diferentemente da maioria dos eucariotos, em alguns vegetais, fungos e leveduras, por exemplo, o mtDNA possui íntrons (GRAY, 2015).

Acreditava-se que o mtDNA de humanos apresentavam 37 genes, que codificavam para apenas 13 proteínas componentes da cadeia respiratória, além de 2 rRNAs de 12S e 16S e 22 tRNAs. Porém, um peptídeo denominado Humanin foi descoberto há alguns anos no genoma mitocondrial humano, localizado no rRNA 16S (HASHIMITO et al., 2001; GUO et al., 2003). Este peptídeo, que tem apenas 21 aminoácidos se traduzido dentro da mitocôndria e 24 aminoácidos se traduzido no citosol, apresentou efeitos neuroprotetores, sendo capaz de suprimir a morte dos neurônios em casos de doença de Alzheimer (XIAO et al., 2016). A FIGURA 3 apresenta o mapa do genoma mitocondrial humano, contendo 16.569 pb (pares de bases) e 14 proteínas codificadas.

FIGURA 3 - MAPA DO GENOMA MITOCONDRIAL HUMANO



FONTE: XIAO et al. (2016).
NOTA: Descrição da figura – APÊNDICE 2.

Apesar do pequeno número de proteínas codificadas pelo genoma mitocondrial, calcula-se que aproximadamente 1.500 proteínas, diferentes das proteínas de origem citoplasmática, auxiliam na manutenção e funcionamento das

mitocôndrias dos mamíferos. O genoma mitocondrial codifica uma pequena fração das proteínas totais sintetizadas pelo organismo de eucariotos. O menor genoma mitocondrial já identificado é o do *Plasmodium falciparum*, que codifica para apenas 3 proteínas, enquanto o maior genoma codifica para 72 proteínas, presentes na mitocôndria do *Andalucia goyodi*. O restante do genoma necessário para a manutenção da mitocôndria é codificado pelo núcleo e importadas do citossol para a mitocôndria (KANNAN; ROGOZIN; KOONIN, 2014).

O código genético do mtDNA é ligeiramente diferente do código genético universal. Existem alguns códons específicos de inicialização e parada, além de códons que representam aminoácidos diferentes dos códons do código universal, dependendo do organismo. Por exemplo, no genoma mitocondrial, o códon UGA dará origem ao aminoácido triptofano, ao passo que, no genoma universal, o códon UGA representa um códon terminal e o triptofano é representado pelo códon UGG. O aminoácido metionina de códon AUG, sinaliza o início de uma proteína no código usual, no mitocondrial pode ser sinalizado também pelo códon AUA. O códon AUA no código universal representa uma isoleucina (NCBI, 2017). Sendo assim, caso os genes mitocondriais sejam transferidos para o genoma nuclear, serão traduzidos com erros, pois os ribossomos do citosol utilizam o código genético padrão (KANNAN; ROGOZIN; KOONIN, 2014).

Os erros de tradução podem ser corrigidos por meio da edição do RNA, processo bastante comum que foi descoberto primeiramente nas mitocôndrias. O processo de edição consiste na substituição ou inserção pós-transcrição de bases, sendo a substituição mais frequente de citosina (C) para uracila (U). O processo de inserção de bases é extremamente raro (MOREIRA et al., 2016).

A análise do mtDNA é empregada principalmente para estudos evolutivos, populacionais e de parentesco, pois a principal característica da mitocôndria em metazoários é a sua herança materna. As mitocôndrias são originadas do óvulo, pois o gameta masculino tem de citoplasma com número reduzido de mitocôndrias (JUNQUEIRA; CARNEIRO, 2005). Dessa forma, apenas alterações mitocondriais originárias do gameta feminino são passadas para a prole independente do sexo (BORGES-OSORIO; ROBINSON, 2013).

Com o processo evolutivo, as mitocôndrias perderam parte de seu genoma que foram incorporadas ao núcleo das células hospedeiras, por meio de transferência horizontal, sendo um processo onde um determinado organismo

transfere seu material genético para outro organismo que não é seu descendente (ZAHA; FERREIRA; PASSAGLIA, 2014). No entanto, o mecanismo de transferência de herança é complicado, pois alguns genes mitocondriais, para serem expressos, dependem da interação com genes presentes no núcleo (BORGES-OSORIO; ROBINSON, 2013).

2.2.1 Teoria endossimbiótica

Segundo a teoria endossimbiótica existem evidências que as mitocôndrias surgiram a partir de uma alfa-proteobactéria fagocitada por uma célula primitiva anaeróbia e, por meio do englobamento da bactéria e estabelecimento de simbiose entre bactéria e hospedeiro, com o passar do tempo, essa relação tornou-se irreversível (JUNQUEIRA; CARNEIRO, 2012).

O período de aquisição da mitocôndria no decorrer da história evolutiva dos eucariotos ainda é motivo de debate no meio científico (PITTIS; GABALDÓN, 2016; DEGLI ESPOSTI, 2016). Existem duas teorias em relação ao hospedeiro endossimbionte: A primeira hipótese postula que o hospedeiro era uma arqueia com características eucarióticas já desenvolvidas, incluindo a capacidade fagocitária; A segunda hipótese afirma que o hospedeiro arqueano desenvolveu suas características eucarióticas após a fagocitose (KANNAN; ROGOZIN; KOONIN, 2014).

Recentemente Pittis e Gabaldón (2016) encontraram evidências da aquisição tardia das mitocôndrias, ou seja, o hospedeiro já possuía desenvolvimento do sistema de endomembranas e características eucarióticas, não sendo um organismo tão primitivo, como a segunda hipótese prega. Porém, para o meio científico, este fato ainda não está consolidado, tendo em vista as divergências de opiniões entre pesquisadores (GRAY, 2015; PITTIS, GABALDÓN, 2016; DEGLI ESPOSTI, 2016).

2.3 OBTENÇÃO DE GENOMAS

Diversas são as informações relevantes que podem ser extraídas do mtDNA. A interpretação dos dados de genoma mitocondrial é fácil quando se utilizam softwares atuais de bioinformática. Para realizar a obtenção de qualquer genoma, sendo ele mitocondrial ou não, é necessário passar pelas etapas de sequênciamento, anotação e montagem genômica.

2.3.1 Sequênciamento genômico

A primeira etapa para a obtenção de um genoma é o sequênciamento. Anteriormente aos avanços tecnológicos, os genomas eram obtidos pelo método de Sanger, que consistia em obter pequenos trechos de DNA, conforme o tamanho do vetor no qual era clonado. Nas técnicas utilizadas atualmente não há mais a necessidade da fase de clonagem (ZAHA; FERREIRA; PASSAGLIA, 2014).

Algumas tecnologias de sequenciamento de DNA de última geração são capazes de fornecer milhares de pares de bases por corrida, informações que são muito maiores que as geradas pelo sequenciamento de Sanger. A exemplo disso, temos a plataforma 454 (Roche), Solexa (Illumina), SOLiD System (Applied Biosystems) e tSMS (Helicos) que são utilizadas amplamente. Esses equipamentos quando comparados com o sequênciamento de Sanger são rápidos e possuem menores custos (CARVALHO; SILVA, 2010).

2.3.2 Montagem genômica

A montagem genômica, é realizada por meio de agrupamento, onde sequências do DNA são compiladas levando em consideração a similaridade dos fragmentos, denominados contigs, que são gerados no processo de sequenciamento. Utilizam-se programas de Bioinformática para realizar o alinhamento dos contigs, a fim de reconstruir o genoma original (TROY et al., 2001).

Existem programas que realizam a montagem genômica, utilizando diferentes tipos de algoritmos para obter a ordem das sequências do DNA. Mesmo vários programas disponíveis, ainda existem problemas que podem ocorrer durante a montagem como agrupamentos repetidos, regiões com baixa qualidade de sequenciamento, etc (ZAHA; FERREIRA; PASSAGLIA, 2014).

Os métodos baseados na identificação de sequências homólogas de outros genomas são confiáveis. As informações obtidas por meio da anotação são utilizadas para mineração dos dados em banco de dados públicos, comparando com genes já existentes, a fim de finalizar a montagem de forma correta. Deve-se considerar a utilização de ferramentas de bioinformática de forma integrada, de modo que favoreça e facilite a verificação de dados importantes (TROY et al., 2001).

2.3.3 Anotação genômica

A pesquisa por similaridade de sequências é o próximo passo, denominada anotação genômica. Segundo Stein (2001), a anotação genômica possui três categorias que se destacam por serem as mais importantes: Anotação em nível de nucleotídeos, ou seja, quais são os nucleotídeos presentes nas sequências; anotação em nível de proteínas, que determina quais proteínas estão presentes nos genes; e anotação em nível de processo, definindo qual será a função das proteínas no organismo.

De acordo com Yandell e Ence (2012), anotações de genoma equivocadas podem arruinar outras montagens que se baseiam nela, tendo em vista que futuras pesquisas serão baseados nessas anotações. Portanto, esta etapa deve ser de alta qualidade, para que se possa identificar corretamente os principais genes e seus produtos, evidenciando as características do genoma em particular (STEIN, 2001).

2.4 BANCOS DE DADOS DE GENOMAS MITOCONDRIAIS

Como consequência do número crescente de análises genéticas que geram inúmeros dados, foram criados bancos de dados biológicos públicos, os quais armazenam sequencias genéticas. Estes repositórios contêm arquivos de sequências genéticas em formato único (fasta) ou múltiplo (multifasta). As informações das sequências genéticas são disponibilizadas juntamente com os cabeçalhos, os quais contêm informações importantes, tais como: identificação do organismo e do gene/proteína e o número de acesso à sequência (MOHAMMED et al., 2012).

Com as técnicas avançadas de sequenciamento, o número de submissões de genomas completos nas bases de dados públicos nos últimos anos aumentou muito (O'LEARY et al., 2016). Existem hoje, vários bancos de dados biológicos com informações mitocondriais, porém poucos são específicos à esta organela.

O NCBI é uma divisão da Biblioteca Nacional de Medicina localizada nos Estados Unidos na *National Institutes of Health (NIH)* que fornece publicamente dados de sequências biológicas sem custo, além da incorporação inúmeros recursos (PRUITT et al., 2005). O GenBank, RefSeq e *Organelles Genome Resources* são alguns dos bancos de dados do NCBI que possuem genomas mitocondriais.

O GenBank é um banco de dados de sequências nucleotídicas e anotações bibliográficas de apoio. É um banco abrangente (BENSON et al., 2013), que possui uma grande importância para a comunidade científica, devido sua extensão de dados e estatísticas sobre diversos aspectos de genomas (SMITH, 2016). Porém, mesmo sendo o banco mais utilizado e por não ser curado, o GenBank necessita de uma maior mineração de dados quando referido à alguns tipos de genomas, como o mtDNA (YAO et al., 2009; SMITH, 2016).

O RefSeq (*Reference Sequences*) é um banco de dados de sequências de nucleotídeos e proteínas com recurso correspondente e anotação bibliográfica. Este banco compreende dados selecionados do GenBank, que passam pelo processo de curadoria. As anotações do RefSeq possuem várias fontes disponíveis no GenBank, incluindo a submissão original e pesquisadores que desenvolveram a anotação genômica, além da análise computacional desenvolvida pelo grupo do NCBI (PRUITT et al., 2005). É um banco de dados

não redundantes, que fornece dados melhorados de sequências representativas redundantes do GenBank (O'LEARY et al., 2016).

O NCBI fornece a maior parte das anotações do RefSeq para o *Organelle Genome Resources,* outro banco de dados integrado onde é possível realizar pesquisas de sequências biológicas de organelas. A anotação de genomas mitocondriais de mamíferos é complementada com a curadoria manual (O'LEARY et al., 2016).

Um dos bancos de dados de genoma de organelas mais utilizados é o GOBASE (O'BRIEN et al., 2003), porém, este banco encontra-se desatualizado. As informações contidas neste banco de dados foram fundamentais para o campo da genômica comparativa, facilitando pesquisas e interpretação de dados genéticos mitocondriais (SMITH, 2016). No GOBASE é disponibilizado dados de sequências mitocondriais originários do GenBank e uma série de informações taxonômicas, mapas genéticos e estruturas secundárias de RNAs (O'BRIEN et al., 2009).

2.5 ANÁLISE COMPARATIVA DE SEQUÊNCIAS

Para estudar a evolução molecular de organismos por meio de filogenia, as sequências genômicas precisam ser comparadas. Para isso, é necessário definir uma medida quantitativa de similaridade entre as sequências analisadas (SALEMI; VANDAMME, 2003). Existem algumas maneiras de se comparar sequências: Por meio de alinhamento ou por meio de análise de frequências de palavras (LEIMEISTER; MORGENSTERN, 2014).

2.5.1 Alinhamento de sequências

O alinhamento de sequências é uma ferramenta básica em bioinformática, utilizada para a identificação de correspondências entre pares de resíduos. Qualquer semelhança que corresponda aos resíduos, mantendo a ordem dos mesmos dentro de uma sequência, é denominado alinhamento (LESK, 2008).

Os métodos de alinhamento de sequências são classificados em: Alinhamento global (ao longo de toda a extensão da sequência) e Alinhamento local (apenas uma fração da extensão das sequências). Estes métodos de alinhamento podem ser realizados em pares ou de maneira múltipla (ZAHA; FERREIRA; PASSAGLIA, 2014).

Basicamente, para ocorrer o alinhamento, cada resíduo da sequência é comparada com os resíduos das outras sequências e, quando existe identidade entre elas, um valor é dado para cada posição em uma matriz de pontuação. Os *gaps* (lacunas) são penalizados. O melhor alinhamento é aquele onde houver a maximização da pontuação do escore (SALEMI; VANDAMME, 2003).

As matrizes de pontuação foram desenvolvidas para serem usadas de maneira que respeite as regras das mutações. As matrizes PAM (Point Accepted Mutation) e BLOSUM (BLOcks of amino acid SUbstitution Matrix) são as mais utilizadas pelas ferramentas de alinhamento (SALEMI; VANDAMME, 2003), e consistem em sistemas de pontuação para correspondências para inserções, deleções, desajustes e deleções que influenciam no alinhamento das sequências genéticas (MOUNT, 2008). Dentre as ferramentas que utilizam essas matrizes de pontuação para o alinhamento, destacam-se o ClustalW (THOMPSON; HIGGINS; GIBSON, 1994) e o MUSCLE (EDGAR, 2004).

Existem muitos algoritmos que fazem alinhamento de sequências: O Needleman-Wunsch, que realiza alinhamento global (NEEDLEMAN; WUNSCH, 1970); o Smith-Waterman, que realiza alinhamento local (SMITH; WATERMAN, 1981); e o BLAST, que realiza busca por similaridade no banco de dados do NCBI (ALTSCHUL et al., 1990). Genericamente, as soluções algorítmicas na comparação de sequências por meio de alinhamento mostram-se satisfatórias, porém a carga computacional aumenta de acordo com o comprimento das sequências. Assim, o uso de alinhamento para grandes dados torna-se inviável (VINGA; ALMEIDA, 2003).

Para solucionar os problemas na análise de sequências biológicas em grande escala, utiliza-se metodologias que são independentes de alinhamento. Pesquisas por similaridade entre genes, “todos-contra-todos” por exemplo, é um meio rápido de comparação de sequências (ARUNACHALAM et al., 2010; MAHMOOD et al., 2012; VIALLE et al., 2016).

2.5.2 K-mers

A maioria das filogenias livre de alinhamento são baseadas em k-mer, ou seja, na ocorrência de palavras exatas de comprimento fixo “k” em um conjunto de sequências de DNA (VINGA; ALMEIDA, 2003). O comprimento destas palavras pode variar, de acordo por exemplo com o conjunto de dados a ser mapeado (SIMS et al., 2009). Este tipo de análise comparativa inicia-se pelo mapeamento das sequências em vetores, transformando uma sequência em um objeto. Os vetores representam a sequência original, com resolução determinada pelo comprimento da palavra. A partir disso, uma medida de distância é aplicada entre espaços dos vetores, as quais refletem em uma matriz de distância par-a-par (*pairwise*). Várias medidas de distância podem ser aplicadas. A medida mais utilizada é a euclidiana, que é definida pela soma da raiz quadrada da diferença entre as coordenadas (BODEN et al., 2013).

Alguns estudos sugerem a utilização de k-mers espaçados, que se baseiam em um único padrão fixo P de correspondência sem se preocupar com posições. As palavras espaçadas são representadas em um vetor binário, a partir das frequências relativas, onde o cálculo da distância dos pares é aplicado (BODEN et al., 2013; LEIMEISTER et al., 2014; HORWEGE et al., 2014).

2.6 MÉTODOS DE INFERÊNCIA FILOGENÉTICA

Cada organismo é resultante de um processo evolutivo. É possível estabelecer graus de semelhança e parentesco através de árvores filogenéticas, baseado nas manifestações de genoma, transcriptoma e proteoma (PATWARDHAN et al., 2014).

Uma filogenia é um modelo de história genealógica representada por uma árvore onde os comprimentos dos ramos são parâmetros desconhecidos. Se levarmos em consideração que a taxa de substituição de resíduos nas sequencias genéticas é constante ao longo do tempo ou entre as linhagens, dizemos que possui relógio molecular, resultando em uma árvore ultramétrica onde todas as unidades taxonômicas são equidistantes da raiz. Para espécies distantes, o relógio molecular não deve ser utilizado (YANG; RANNALA, 2012).

Uma árvore não enraizada possui os parâmetros de cumprimento dos ramos conhecidos, porém não existe indicação do nó que está representado o ancestral comum. Uma estratégia usada para criar raiz em uma árvore sem raiz é incluir um grupo externo na análise, devendo ser distante das espécies a serem investigadas. Isto leva a raiz a ser estabelecida ao longo do ramo que contém o grupo externo (YANG; RANNALA, 2012).

De acordo com Jun e colaboradores (2010), as ferramentas que realizam filogenia de genomas completos são inseridas em duas categorias: Filogenias baseadas em alinhamento e filogenias livre de alinhamento. Uma ferramenta interessante para comparar sequências e realizar filogenias, que utiliza k-mers de tamanhos calculados, é a FFP (*Feature Frequence Profile),* proposta por Sims e colaboradores (2009). Esta ferramenta é baseada em metodologia livre de alinhamento que, por meio da análise da taxa de substituição entre os aminoácidos, estima o tamanho dos k-mers.

À exemplo de algoritmos em filogenia, a utilização de matrizes para a representação de sequências genéticas, proposta inicialmente na década de 90 por Baum e Ragam (1992 apud BINIDA-EMODS; SANDERSON, 2001), é uma opção que facilita a análise de genomas completos. Tal abordagem é denominada *Matrix Representation with Parsimony Analysis* (MRP) e produz como resultado uma “superárvore” (*supertree*), baseada em um consenso entre as árvores geradas.

Para realizar a reconstrução da história evolutiva dos organismos, usa-se de métodos matemáticos. A precisão da inferência é altamente dependente da qualidade dos dados e do método de reconstrução adotado. Os três métodos matemáticos existentes são: Métodos baseados em parcimônia, métodos baseados em probabilidades e métodos baseados em distância (DELSUC; BRINKMANN; PHILIPPE, 2005).

2.6.1 Métodos baseados em parcimônia

O método de máxima parcimônia visa diminuir o número de alterações em uma filogenia atribuindo caracteres aos nós internos da árvore. O comprimento do caractere equivale ao mínimo de alterações necessárias para aquele local, sendo

que a pontuação da árvore é igual ao somatório do comprimento dos caracteres de todos os locais. Dessa forma, a máxima parcimônia minimiza a pontuação da árvore (YANG; RANNALA, 2012).

Este método possui uma séria desvantagem. A pontuação da árvore é baseada completamente no número de mutações entre todas as reconstruções, e é pouco provável que o número de mutações seja igual em todas as árvores (HOLDER; LEWIS, 2003).

2.6.2 Métodos baseados em probabilidades

Existem dois conceitos probabilísticos na reconstrução de filogenias: a máxima verossimilhança e a inferência bayesiana (YANG; RANNALA, 2012).

No método de máxima verossimilhança, a árvore será aquela cujo valor de verossimilhança seja maior, ou seja, calcula-se a probabilidade de um conjunto de sequências genéticas observadas dado algum modelo probabilístico assumido. Avalia-se todas as hipóteses e seleciona-se aquela que maximiza a probabilidade de gerar os dados observados (SULLIVAN, 2005).

Uma maneira de avaliar a confiança da árvore filogenética obtida é por meio do *bootstrap*. O *bootstrap* disponibiliza uma medida de quais partes da árvore é fracamente sustentada. Grupos com baixo bootstrap devem ser analisados cautelosamente (HOLDER; LEWIS, 2003).

A inferência bayesiana combina um método utilizado anteriormente (uma filogenia prévia) com a probabilidade da árvore de produzir árvores posteriores, selecionando a melhor estimativa de filogenia (árvore de maior probabilidade posterior (RANNALA; YANG, 1996). Dessa forma, a análise bayesiana do método prévio produz uma árvore avaliando diretamente o modelo de substituição, comprimento dos ramos entre outras variáveis (DOUADY et al., 2003).

2.6.3 Métodos baseados em distância

Nos métodos baseados em distância, cada par de sequências é calculado, resultando em uma matriz de distância que é usada para a reconstrução da

árvore, podendo chegar a uma filogenia muito bem resolvida (YANG; RANNALA, 2012). Desde que contenham informações filogenéticas suficientes, esses métodos são bons marcadores, pois são menos propensos a interferências, como a semelhança decorrente da convergência evolutiva (DELSUC; BRINKMANN; PHILIPPE, 2005).

O alinhamento de sequências não é necessário após o cálculo da distância ser realizado. Alguns dos métodos aplicados a matriz de distância são: *Neighbor joining*, UPGMA e o Mínima evolução. O método de mínima evolução utiliza a soma do comprimento dos ramos da árvore, portanto árvores mais curtas são provavelmente mais corretas que árvores mais longas. O método *Neighbor joining* inicia-se com base na distância de táxons, formando grupos até que a árvore seja totalmente resolvida (YANG; RANNALA, 2012).

O UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*), assim como no *Neighbor joining*, utiliza métodos baseados em clusterização, sendo um método simples em que a cada estágio de agrupamento se forma um novo nó na árvore. A árvore construída pelo método UPGMA é aditiva, assumindo taxa de evolução constante, onde todos organismos estão igualmente distantes da raiz (RIZZO; ROUCHKA, 2007).

O grande problema dos métodos de distância é que as diferenças observadas entre as sequências não são reflexos exatos das distâncias evolutivas entre elas, além de que, se houver múltiplas substituições em um mesmo local, a verdadeira distância pode ser camuflada, aparentando equivocadamente que as sequências são próximas umas das outras (HOLDER; LEWIS, 2003).

2.7 FILOGENIA MITOCONDRIAL

Existem alguns conceitos que são indispensáveis para o entendimento da filogenia de um modo geral, entre eles estão a Homologia e a Similaridade. As características que são derivadas de um ancestral comum são ditas homólogas. Outras características que aparentam ser similares podem ter surgido de modo independente, por evolução convergente. À exemplo dessas semelhanças que não possuem o mesmo ancestral comum, podemos citar asas de abelhas e asas de águia (LESK, 2008).

Outros conceitos essenciais para a filogenia são as definições de genes ortólogos e genes parálogos. Genes ortólogos são aqueles que surgiram a partir de um conjunto de duplicatas, ou seja, são duas cópias do mesmo gene do mesmo loco. Genes parálogos são dois genes em locus diferentes, produzidos a partir de uma duplicação que possuem funções diferentes (RIDLEY, 2006).

Importante destacar que a inferência filogenética deve ser baseada em genes ortólogos e não em genes parálogos. O problema é que é fácil de confundir estas definições, pois durante a evolução podem ocorrer perdas de genes (RIDLEY, 2006). Saber diferenciar estas definições exige conhecimento detalhado dos caminhos evolutivos que levaram à divergência de funções biológicas (JENSEN, 2001).

2.7.1 A árvore da vida

A filogenia estabelece uma topologia de relações baseadas em classificações, de acordo com um conjunto de características dos organismos. A nomenclatura biológica se apoia na idéia de que os organismos vivos são classificados de acordo com uma hierarquia, sendo ela, Reino, Filo, Classe, Ordem, Família, Gênero e Espécie, respectivamente (LESK, 2008). O mundo vivo, de acordo com Margulis e Schwartz (1988), é classificado em cinco reinos: Monera, Protista, Plantae, Fungi e Animalia.

Há quase três décadas, tendo como base o rRNA, Woese, Kandler e Wheelis (1990), sugeriram que a divisão dos organismos em 5 reinos vivos não seria filogeneticamente correta, dado as diferenças profundas entre eubacterias, archeobacterias e eucariotos. Os autores então propuseram que as archeas deveriam ser definidas em um nível taxonômico mais elevado.

As árvores baseadas em rRNA são muito populares desde a década de 1990 (WOESE; KANDLER; WHEELIS, 1990; BROWN; DOLITTLE, 1997). Porém, de acordo com Forterre (2015) elas fornecem um falso cenário da história da vida onde os ramos mais longos das árvores são representados por organismos basais (Archeozoa) ausentes de mitocôndrias.

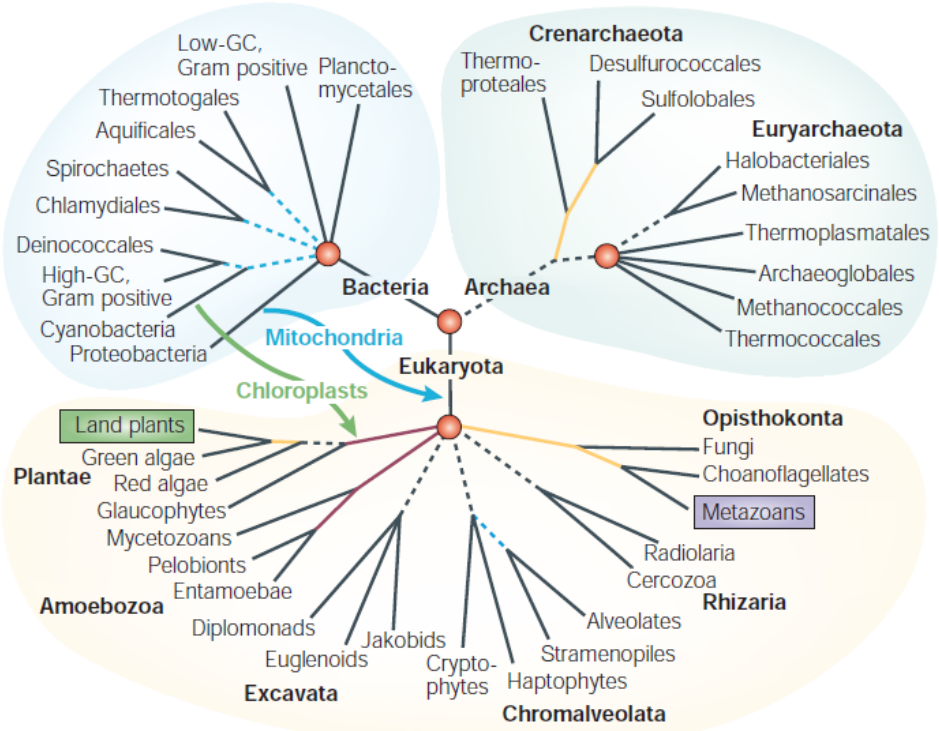
Desde a definição das Archeas como táxon superior (WOESE; KANDLER; WHEELIS, 1990), pesquisadores têm utilizado a designação “árvore da vida” para

a representação da história evolutiva que vincula os três maiores domínios entre os seres vivos: Eubacteria, Archaea e Eukaria. A utilização de genes presentes no genoma mitocondrial tem direcionado muitas reconstruções filogenéticas (DELSUC; BRINKMANN; PHILIPPE, 2005; BURKI, 2014).

Quanto aos eucariotos, devido à falta de representantes de várias linhagens em estudos conduzidos (DELSUC; BRINKMANN; PHILIPPE, 2005), a divisão do mundo eucariótico é representada por seis grupos principais: Opisthokonta, Amoebozoa, Plantae, Chromalveolata, Rhizaria e Excavata (SIMPSON; ROGER, 2004). Burki e colaboradores (2007) propuseram o termo SAR para a junção do grupo Rhizaria e duas clades do grupo Chromoalveotas, sendo este termo bem aceito atualmente.

Conforme a árvore da vida proposta por Delsuc, Brinkmann e Philippe (2005) (FIGURA 4), a subdivisão dos eucariotos (anterior ao agrupamento SAR) e o evento de fusão de uma espécie bacteriana e uma archaea que originou os eucariotos está representado. Essa fusão gera uma estrutura anelar na raiz da árvore da vida.

FIGURA 4 - REPRESENTAÇÃO DA ÁRVORE DA VIDA



FONTE: DELSUC; BRINKMANN; PHILIPPE, 2005.
NOTA: No fundo azul, representação de bactérias e archeas que deram origem aos eucariotos. As subdivisões dos eucariotos Opisthokonta, Amoebozoa, Plantae, Chromalveolata, Rhizaria e Excavata são representados pelo fundo rosa.

Os organismos mais primitivos que têm mitocôndrias, segundo estudos evolutivos, são os Jakobidas, do supergrupo Excavata. Árvores enraizadas entre Excavatas e todos os outros organismos eucarióticos foram propostas inicialmente por Stechmann e Cavalier-Smith (2003).

2.7.2 Marcadores mitocondriais

Existem alguns genes específicos que servem como marcadores mitocondriais, devido ao fato de estarem presentes no genoma de grande parte dos organismos vivos. Esses genes mitocondriais estão relacionados com a produção de energia, como é o caso dos genes codificantes das proteínas Citocromo b, da Citocromo Oxidase, e de rRNAs (RIDLEY, 2006).

O CYTB é um gene codificador de uma proteína universal (Citocromo b) que não afeta-se por múltiplas substituições. É uma proteína necessária na cadeia respiratória mitocondrial e presente também em bactérias, sendo utilizado principalmente em análises filogenéticas para a diferenciação de espécies (MEYER, 1994). O gene CYTB se mostrou eficaz na recuperação de informações importantes em uma variedade de níveis taxonômicos, porém sua utilidade é dependente da linhagem e sua precisão diminui conforme aumenta a profundidade da evolução (PATWARDHAN et al., 2014).

O COX1, que codifica para a subunidade 1 da proteína Citocromo oxidase, é outro gene também utilizado na identificação molecular em nível de espécie, por possuir uma região do mtDNA denominada *DNA barcoding*, uma parte do mtDNA padronizada no genoma usada para identificar diferentes espécies (HEBERT et al., 2003).

Em uma pesquisa realizada por Tobe e colaboradores (2010), compararam-se os genes CYTB e COX1 para a reconstrução filogenética e identificação de espécies de mamíferos. Concluiu-se que o gene CYTB demonstra ter uma taxa de falso positivo menor que a metade de COX1 e um valor preditor positivo mais alto, sendo superior na identificação de mamíferos.

Alguns genes de rRNAs também são adequados para reconstruções filogenéticas, sendo frequentemente usados para vários grupos de organismos. Os genes mitocondriais 12S e 16S possuem substituições mais funcionalmente

neutras, permitindo sua utilização como marcador (KUZNETSOVA et al., 2002). Esses genes são especialmente úteis para resolver problemas filogenéticos na faixa dos 10 a 100 milhões de anos (RIDLEY, 2006).

2.8 MINERAÇÃO DE DADOS

A bioinformática é uma área que combina várias ciências, como a bioestatística, as ciências biológicas, bioquímica, biologia molecular, genética, entre outras ciências. O grande objetivo desta área é fazer com que as análises de processos biológicos, interpretação dos dados e inferência biológica seja facilitado, e a mineração de dados fornece abordagens para essas investigações moleculares (MOORE, 2007).

Mineração de dados é o processo pelo qual é possível reconhecer padrões e modelos em grandes volumes de dados. Para a bioinformática, a manipulação de informações biológica cria a oportunidade para o desenvolvimento de métodos em mineração de dados que desempenhem um papel na compreensão dos dados biológicos (ZAKI et al., 2007).

Alguns métodos de aprendizado de máquina e de mineração de dados para o reconhecimento de padrões em um grande volume de dados biológicos têm a vantagem de reproduzir com exatidão a complexidade implícita da organização dos sistemas biológicos (MOORE, 2007). Existem duas maneiras de categorizar o reconhecimento de padrões, por meio da aprendizagem supervisionada, onde as classes são definidas, e aprendizagem não supervisionada, onde as classes são aprendidas de acordo com a similaridade dos padrões (RUSSELL, NORVIG, 2004).

2.8.1 Clusterização

A Clusterização é uma técnica de aprendizagem não supervisionada que tem por objetivo agrupar automaticamente as ‘”n” ocorrências da base de dados em “k” grupos (JACKSON, 2002), sendo dividida em dois métodos: Métodos hierárquicos e Métodos de particionamento (FARLEY; RAFTERY, 1998).

Os Métodos hierárquicos funcionam por etapas, produzindo sequência de partições, sendo que cada uma corresponde a um número de agrupamentos. Esse método pode ser aglomerativo, onde os grupos são fundidos, ou divisivo, onde os grupos são divididos em cada fase. Já os Métodos de particionamento movem informações de um grupo para outro, a partir de uma partição inicial, onde o número de agrupamentos deve ser especificado antes. O k-means é método de particionamento mais utilizado (FARLEY; RAFTERY, 1998).

O algoritmo de clusterização k-means calcula distância euclidiana dos centróides dos *k* grupos de cada classe, repetidas vezes, até determinar um agrupamento estável. O número de clusters definida previamente é dada por “k”. Este método possui sensibilidade à ruídos, sendo que valores muito altos podem distorcem a distribuição dos dados (DONI, 2004).

2.8.2 Lógica Fuzzy

A Lógica Difusa, ou Lógica Fuzzy é utilizada para classificar tudo o que não é absolutamente preciso, sendo uma ferramenta capaz de captar informações escritas em linguagem normal e converte-las em números, atribuindo graus de pertinência ao conjunto de dados. Diferente da Lógica Booleana que atribui valores apenas verdadeiros e falsos, a Lógica fuzzy classifica os dados em valores que variam de 0 à 1; Logo, uma meia verdade seria representada pelo valor 0.5 (SILVA, 2005).

Esta forma de classificação e reconhecimento de dados é frequentemente utilizada na Bioinformática, principalmente para agrupar dados genéticos, atribuir graus de associação de clusters aos genes, prever localizações subcelulares de proteínas, estudar as diferenças entre nucleotídeos, mapear padrões de sequências, entre outras utilidades (TORRES; NIETO, 2006).

Um algoritmo válido em clusterização baseado em Lógica Fuzzy é o fuzzy c-means. Este algoritmo é uma generalização dos métodos de particionamento, permitindo visualizar o grau de associação de cada elemento em cada grupo. Esta técnica proporciona uma grande vantagem pois fornece informações reais sobre as estruturas dos dados, associando graus de incerteza aos elementos de cada grupo (DONI, 2004).

2.8.3 Média móvel

A média móvel é um método baseado em padrões de comportamento que tem como objetivo distinguir ruídos. A técnica assume que os valores extremos apresentam aleatoriedade e, por meio de suavização de pontos extremos, identifica padrões básicos (MORETTIN; TOLOI, 2006).

Para realizar a estimativa de densidade utilizando média móvel é necessário medir quais pontos são difundidos em uma dada vizinhança, sendo indispensável especificar exatamente o tamanho desta vizinhança qual deve abranger pontos suficientes para assegurar uma estimativa significativa. Para identificar os vizinhos próximos de um ponto de consulta, utiliza-se uma métrica de distância, como por exemplo a distância euclidiana (RUSSELL, NORVIG, 2004).

2.9 ESPAÇO VETORIAL

Este trabalho foi inteiramente desenvolvido em ambiente MatLab®. O MatLab® é um software pago utilizado principalmente para análise e exploração de dados por meio de cálculos com matrizes. Este software oferece pacotes direcionados ao estudo de diversas áreas, entre elas a Bioinformática e a Estatística (SANTOS, 2010).

Os comandos do MatLab® são próximos da forma como escrevemos expressões algébricas, o que torna o seu uso simplificado (SANTOS, 2010). Para utilizar este software é necessário que o usuário tenha conhecimento de álgebra linear.

O objetivo da álgebra linear é estudar o comportamento de operações definidas sobre conjuntos, tratando especificamente de espaços vetoriais (PELLEGRINI, 2016). Na matemática define-se “vetor” como uma sequência ordenada de valores. Os vetores são frequentemente interpretados como segmentos de retas orientados em um espaço euclidiano *n*-dimensional, sendo a adição vetorial e a multiplicação escalar as duas operações fundamentais sobre vetores. Os elementos de um vetor são representados por barras ou flechas sobre os nomes: $\vec{v}$ (RUSSELL, NORVIG, 2004).

Uma matriz é um arranjo retangular de valores organizados em linhas e colunas (RUSSELL, NORVIG, 2004), conforme representado a seguir na matriz A, onde o primeiro índice de $A_{m,n}$ significa linha e o segundo significa coluna.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & a_{i3} & \dots & a_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}$$

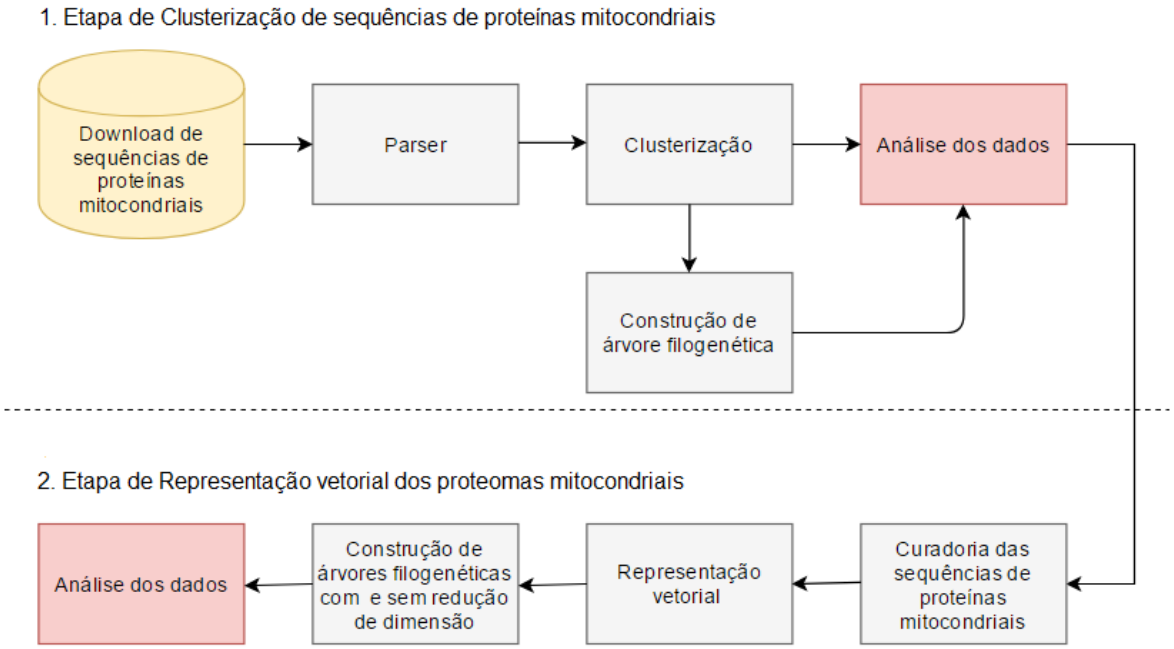
Um vetor pode ser escrito em notação matricial, sob a forma de uma matriz linha (com uma única linha) ou uma matriz coluna (com uma única coluna). Essa forma de representação de vetores pode ser justificada pelo fato de que as operações matriciais produzem resultados iguais as operações vetoriais (SANTOS, 2010). Em um espaço vetorial pode-se somar e multiplicar elementos de forma escalar, ou seja, em elementos de um outro conjunto que não seja do próprio espaço (PELLEGRINI, 2016).

3 MÉTODOS

Esta seção destina-se a descrição dos métodos utilizados para a realização deste trabalho, apoiando-se nos conceitos abordados da fundamentação teórica. O trabalho todo foi desenvolvido no ambiente MatLab®, utilizando a biblioteca de funções desenvolvidas pelo grupo, além das funções fornecidas pelo próprio software disponíveis na Toolbox de Bioinformática.

Segue abaixo um fluxograma que representa os processos de execução deste trabalho (FIGURA 5).

FIGURA 5 - FLUXOGRAMA DOS PROCESSOS DE EXECUÇÃO DO TRABALHO



Fonte: O Autor (2017).

A pesquisa foi dividida em duas etapas: A primeira voltada à clusterização (agrupamento) das sequencias de proteínas mitocondriais e construção de árvore filogenética, realizado com o intuito de estudar as proteínas mitocondriais de um modo geral. A representação vetorial e a construção de árvores filogenéticas com e sem redução de dimensão ocorreu na segunda etapa.

Todos os resultados obtidos neste trabalho estão disponíveis para download em https://sourceforge.net/projects/mitochondrialtrees/.

3.1 OBTENÇÃO DAS SEQUÊNCIAS DE PROTEÍNAS MITOCONDRIAIS

As sequências mitocondriais foram obtidas do banco de dados RefSeq, via FTP, disponível em ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/. Foram utilizados todas as proteínas que estavam disponíveis em formato fasta de sequências de aminoácidos (.faa).

3.1.1 Curadoria dos dados

Para a etapa de clusterização foi realizado um Parser (extração de características importantes) dos cabeçalhos das sequências, obtendo as informações com o nome da proteína e o organismo. No multifasta de entrada, contendo todas as proteínas do RefSeq, foram inseridas as siglas “GN” antes da definição de proteína e “OS” antes da definição de organismo (APÊNDICE 3).

Para a representação vetorial, a curadoria dos dados foi realizada concatenando as sequências das proteínas mitocondriais, com o propósito de representar de cada organismo de forma única.

Segundo Kannan, Rogozin e Koonin (2014), o organismo que possui o menor genoma mitocondrial é o *Plasmodium falciparum* (filo Apicomplexa), que codifica para 3 proteínas. O *Andalucia goyodi* (filo Excavata) possui o maior número de proteínas mitocondriais, codificando para 72 proteínas. Esta afirmação foi base para o critério de exclusão deste estudo, sendo que todos os organismos que têm um número menor de resíduos de aminoácidos que filo Apicomplexa e maior que o filo Excavata, foram excluídos das análises deste trabalho.

3.2 CLUSTERIZAÇÃO DAS SEQUÊNCIAS DE PROTEÍNAS MITOCONDRIAIS

Para a clusterização das sequências de proteínas mitocondriais, utilizamos a ferramenta RAFTS3groups *(Rapid Alignment Free Tool for Sequences Similarity Search to Groups*), desenvolvida pelo grupo da Bioinformática da UFPR. O algoritmo de clusterização desta ferramenta é livre de alinhamento, baseado na busca por similaridade de sequências. A pesquisa por similaridade é realizada por

k-mers, obtidos por meio de uma janela deslizante e utilizados para comparar com as todas as sequências, de acordo com uma matriz de co-ocorrência de aminoácidos (COIMBRA, 2015; NICHIO, 2016).

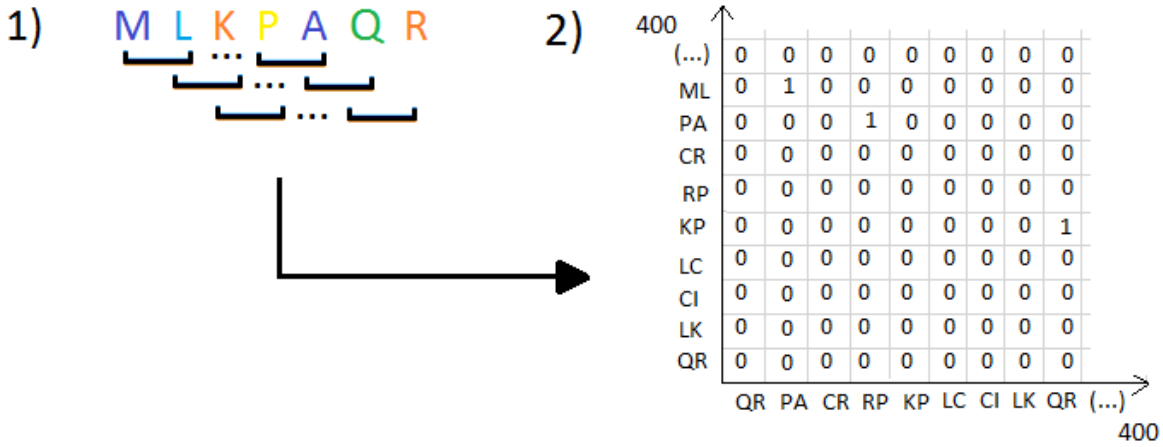
Para a formação dos clusters, uma sequência de proteína mitocondrial foi seleciona aleatoriamente pela ferramenta RAFTS3groups, e comparada todas contra todas. Para esta análise, definimos o limiar de corte em 50% de similaridade. O algoritmo de clusterização encontra-se no ANEXO 1 e ANEXO 2. Esta etapa foi realizada em sistema operacional Windows 7, com processador Intel Core i5, com 8,00 GB de memória RAM.

3.3 REPRESENTAÇÕES VETORIAIS DOS PROTEOMAS MITOCONDRIAIS

A representação vetorial dos proteomas mitocondriais neste trabalho é baseada em k-mers espaçados (BODEN et al., 2013; LEIMEISTER et al., 2014; HORWEGE et al., 2014).

Inicialmente, para cada sequência mitocondrial dos organismos, as ocorrências de todas as combinações possíveis dos 20 aminoácidos existentes, são representados por coordenadas em um mapa. Este mapa, que é uma matriz de 400x400 (Ms400), acumula os resultados das combinações de aminoácidos de cada sequência, resultantes de uma janela deslizante contendo 5 mers com 1 espaçamento. As combinações são realizadas em dipeptídeos (FIGURA 6). Esta matriz resultante é representada em um vetor binário de 160.000 atributos (W160k), que representa os proteomas mitocondriais originais, podendo ser representado de maneiras diferentes. Os elementos foram ordenados por índice, para possibilitar a manipulação dos dados de maneira ordenada. Para esta pesquisa, o vetor foi construído concatenando as colunas da matriz randomizada.

FIGURA 6 - ESQUEMA DA EXTRAÇÃO DE ATRIBUTOS DAS PROTEÍNAS MITOCONDRIAIS



FONTE: O Autor (2017)

LEGENDA: 1) Método de k-mers espaçados, com janela deslizante de 5 mers com espaçamento de tamanho 1, aplicadas para cada sequência mitocondrial. Utiliza-se o primeiro e o segundo aminoácido, descontinua-se o terceiro e utiliza-se o quarto e o quinto. 2) Matriz binária contendo as ocorrências de cada combinação de dipeptídeos, sendo 400 combinações possíveis em cada coordenada. A ocorrência de uma combinação é representada por 1 e a não ocorrência por 0.

A partir da representação vetorial binária com as 160.000 possíveis ocorrências, foi realizada uma representação baseada em vetor de 100, 400 e de 800 dimensões.

Para a construção do vetor de 100 dimensões (Wr100), foi utilizado o vetor principal que contém os proteomas mitocondriais originais. A representação para cada uma das 100 coordenadas do vetor contém 1.600 dimensões, resultantes da soma das ocorrências do vetor principal. A construção do vetor de 800 (Wr800) e 400 (Wr400) dimensões é realizada de maneira similar, porém, para cada coordenada do vetor, são representados 200 dimensões para o de 800 e 400 dimensões para o de 400. O algoritmo utilizado para a construção dos vetores encontra-se no APÊNDICE 4.

Após a construção dos vetores, a matriz Ms400 foi utilizada para gerar representações gráficas dos proteomas, utilizando a função “mmvfuzzy2d” (desenvolvida pelo grupo), baseada na média móvel Fuzzy. Esta função tem a aplicação de plotar imagens de acordo com um mapa de coordenadas, representado aqui pela matriz Ms400, onde os picos de maior ocorrência são suavizados e interpretados em um gradiente de cores. A função “mmvfuzzy2d” consta no ANEXO 3.

3.4 INFERÊNCIA FILOGENÉTICA

As árvores foram construídas com base nos vetores W160k (sem redução de dimensão), Wr100, Wr400 e Wr800. O alinhamento das sequências foi feito utilizando algoritmo baseado em Smith-Waterman (SMITH; WATERMAN, 1981), aplicando a distância *pairwise*. A metodologia utilizada para a inferência filogenética foi o UPGMA (MICHENER; SOKAL, 1957), tanto na filogenia dos vetores quanto na do alinhamento. O algoritmo utilizado foi o “FILOMAT” (APÊNDICE 5).

As etapas de representação vetorial e construção de árvores filogenéticas foram realizadas em sistema operacional GNU/Linux Ubuntu 16.04.1 LTS, com processador Intel Xeon 40 núcleos, com 256 GB de memória RAM.

A visualização e manipulação das árvores filogenéticas foram realizadas no software Dendroscope 3 (HUSON; SCORNAVACCA, 2012). As árvores resultantes foram comparadas entre si e confrontadas com filogenias realizadas por diversos autores. Foram realizados os cálculos estatísticos de Pearson e Spearman entre as distâncias *pairwise* para avaliar a correlação.

4 RESULTADOS E DISCUSSÃO

Nesta seção encontram-se os resultados na ordem em que foram gerados seguidos de suas respectivas discussões. Inicialmente, apresentamos uma breve análise da clusterização das proteínas mitocondriais, elaborada no primeiro momento deste trabalho. Posteriormente, segue as análises detalhadas da extração de atributos e inferência filogenética.

4.1 CLUSTERIZAÇÃO DE PROTEÍNAS MITOCONDRIAIS

Foi clusterizado todo o conteúdo de sequências de proteínas mitocondriais presente no RefSeq. Foram 98.933 proteínas clusterizadas, representando 6.811 organismos. Após o agrupamento, essas proteínas foram dispostas em 8.431 clusters, sendo 2.159 destes com no mínimo 2 sequências de proteínas por cluster. O agrupamento levou 37 minutos para ser concluído, tendo o arquivo fasta de entrada 38.8 MB de tamanho.

Através dos agrupamentos, obteve-se um panorama de todas as proteínas codificadas pelo genoma mitocondrial existentes nos diversos organismos disponíveis no banco de dados. Para isso, utilizamos a ferramenta de clusterização RAFTS3groups, por meio da função “agruparafts” (COIMBRA, 2015). Não foram testados outros clusterizadores pois, segundo resultados recentes em nosso grupo de pesquisa, o RAFTS3groups se mostrou eficiente na detecção de grupos ortólogos em comparação com ferramentas já consolidadas, como o UCLUST E CD-HIT, obtendo um desempenho superior (COIMBRA, 2015; NICHIO, 2016).

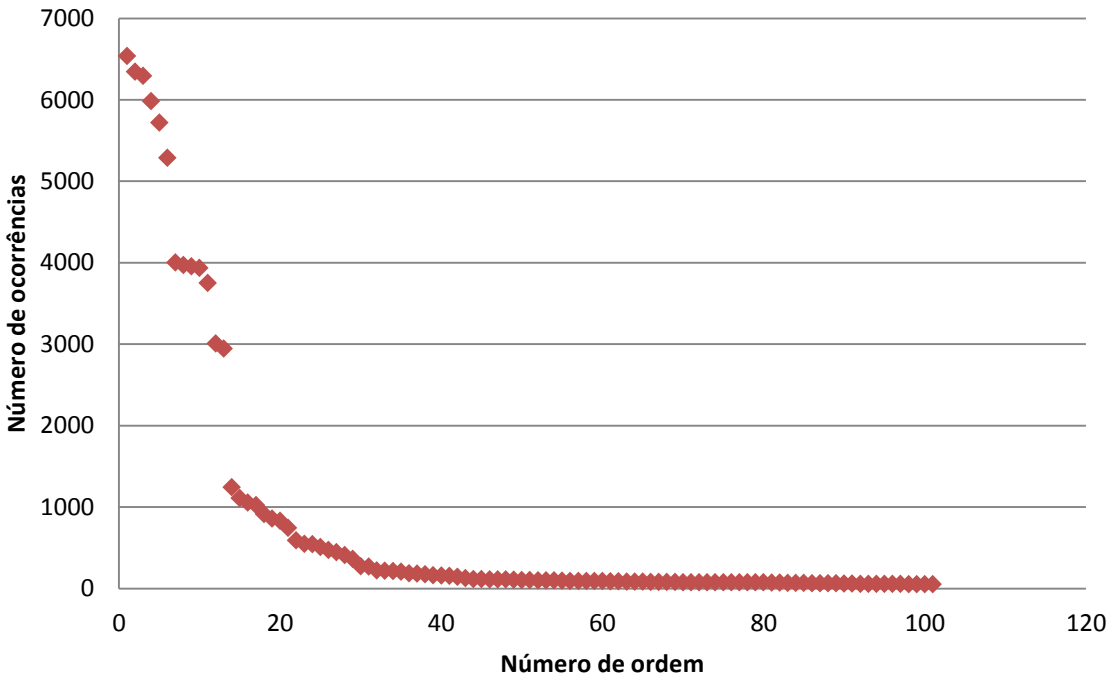
Definimos o limiar de corte em 50% de similaridade. De acordo com Pearson (2013), o limiar de 50% de similaridade, em um banco de dados com menos de 7 milhões de sequências de proteínas, é significativo para inferir homologia em proteínas.

Nos agrupamentos, todas as proteínas fundamentais responsáveis pelo processo de produção de ATP foram identificadas. Outras proteínas menos comuns à maioria dos organismos e também fundamentais à produção energética também foram constatadas e algumas proteínas hipotéticas.

4.1.1 Identificação dos clusters com maiores ocorrências e análise de seus produtos

Em primeira análise, para facilitar a interpretação e visualização do conteúdo dos clusters, selecionamos todos os clusters que tiveram em seu conteúdo 50 ou mais ocorrências. Com essa seleção foi possível identificar 101 clusters, os quais são representados no gráfico abaixo (GRÁFICO 1).

GRÁFICO 1 - NÚMERO DE OCORRÊNCIAS POR CLUSTER



FONTE: O Autor (2017).
NOTA: Os pontos em vermelho representam os clusters que possuem o número de ocorrências maior ou igual à 50, de acordo com o número de ordem de cada cluster. Neste gráfico estão representados 101 clusters.

Conforme constatado, alguns clusters, especificamente 13, se destacam por possuírem maior conteúdo de sequências em relação aos demais. Estes 13 clusters, de fato, contém as sequências das principais proteínas codificadas pela mitocôndria, responsáveis pela função de produção energética.

As proteínas codificadas pelo mtDNA presentes nos complexos proteicos, foram identificadas nos clusters e listadas na TABELA 1, onde consta o número de ordem do cluster, complexo proteico pertencente, produto do cluster contendo o nome da proteína, abreviação utilizada para designar a proteína e o número de

ocorrência por cluster. A identificação destas proteínas bem como suas respectivas funções, são bem fundamentadas na literatura (BOORE, 1999; TAANMAN, 1999; HÜTTEMANN et al., 2007; KANNAN; ROGOZIN; KOONIN, 2014; KÜHLBRANDT, 2015).

TABELA 1 – PRODUTO DOS CLUSTERS MAIS SIGNIFICATIVOS, DE ACORDO COM O NÚMERO DE OCORRÊNCIAS.

Número de ordem	Complexo proteico	Produto do Cluster	Abreviação	Ocorrências (sequências)
1	IV	Citocromo oxidase subunit. 1	COX 1	6.535
2	III	Citocromo b	CYTB	6.345
3	IV	Citocromo oxidase subunit. 3	COX 3	6.291
4	IV	Citocromo oxidase subunit.2I	COX 2	5.981
5	I	NADH desidrogenase subunit. 1	NADH 1	5.717
6	I	NADH desidrogenase subunit. 3	NADH 3	5.282
7	I	NADH desidrogenase subunit. 4	NADH 4	4.001
8	I	NADH desidrogenase subunit. 4L	NADH 4L	3.969
9	I	NADH desidrogenase subunit. 2	NADH 2	3.954
10	I	NADH desidrogenase subunit. 5	NADH 5	3.933
11	V	ATP sintetase subunit. 6	ATP F06	3.751
12	V	ATP sintetase subunit. 8	ATP F08	3.007
13	I	NADH desidrogenase subunit. 6	NADH 6	2.946

FONTE: O Autor (2017)

Considerando o número de ocorrências nos quatro primeiros clusters (COX1, CYTB, COX3 e COX2, respectivamente), percebe-se que o número de sequências proteicas presentes em cada um é próximo ao número total de organismos (6.811). Os genes Citocromo oxidase e citocromo b são considerados marcadores mitocondriais pelo fato de estarem presentes em praticamente todos os organismos vivos, e por essa razão, diversos estudos filogenéticos foram conduzidos utilizando estes genes (MEYER, 1994; HEBERT et al., 2003; TOBE et al., 2010). No estudo realizado por Kannan, Rogozin e Koonin (2014), estes genes apresentaram menor propensão à perdas por transferência.

Os genes que codificam para as proteínas listadas acima são altamente conservados nos metazoários (GISSI et al., 2008), e poucas são as perdas destes genes. Segundo Adams e Palmer (2003), a incorporação de genes mitocondriais ao núcleo, em animais, ocorreu com o gene ATP8 em nematóides e em moluscos. Em oposição, o gene ATP9 (cluster de ordem 23), comum em plantas e fungos,

está presente no genoma nuclear em animais, porém, é componente do mtDNA de esponjas (LIU et al., 2009; PETT et al., 2011).

4.1.2 Produto dos Clusters menores

Dos 8.431 agrupamentos, apenas 101 clusters contém mais de 50 ocorrências, como foi analisado anteriormente. Nesta pesquisa, a existência de grande número de clusters com um pequeno número de ocorrências ocorreu principalmente pela presença de proteínas menos comuns entre os organismos, além de proteínas hipotéticas, sequências parciais de proteínas ou anotações provisórias.

Proteínas agrupadas equivocadamente ocorreram raras vezes quando avaliadas em relação ao número de sequências agrupadas. Esses agrupamentos inconsistentes não foram quantificados, pois o objetivo da clusterização neste estudo foi tomar conhecimento da existência da variedade de proteínas mitocôndrias entre todos os organismos presentes no banco de dados. Todas as proteínas encontradas estão listadas na QUADRO1.

QUADRO 1 - RELAÇÃO DE PROTEÍNAS ENCONTRADAS NA CLUSTERIZAÇÃO

Síntese de ATP	
Complexo I	NADH 1, 2, 3, 4, 4L, 5, 6, 7, 8, 9, 10, 11
Complexo II	Sdh 2, 3, 4
Complexo III	CYTB
Complexo IV	COX1, 2, 3
Complexo V	ATP1, 3, 4, 6, 8, 9
Transcrição	
Polimerases	DNA, RNA
Fator sigma	
Proteínas de tradução	
Proteínas ribossomais Subunidade pequena (small)	RPS1, 2, 3, 4, 7, 8, 10, 11, 12, 13, 14, 19, VAR1
Proteínas ribossomais Subunidade grande (large)	RPL1, 2, 5, 6, 10, 11, 14, 16, 19, 20, 27, 31, 32, 34
Fator de alongamento	Tu
Maturação	
Proteínas de transporte	ccmA, ccmB, ccmC, ccmF secY, TatA, TatC
Proteínas maturases	mtK, mtR
Endonucleases	LAGLIDADG, GIY-YIG
Transcriptase reversa	
Reparo	
Proteína de reparo	MutS
Outras	
*Fotorrespiração	RuBisCO

FONTE: O Autor (2017).
NOTA: Descrição do Quadro – APÊNDICE 6.

A análise do conteúdo foi feita por meio de busca textual, utilizando a variável “RSM.reprfas” dos agrupamentos, que contém 2.159 representantes. Nesta variável estão presentes os centroides de cada cluster com mais de 2 ocorrências, obtidas por k-means. Primeiramente, observou-se que uma mesma proteína/gene pode ter nomenclaturas diferentes, o que dificulta as análises. O inconveniente da não utilização de um padrão para designar os genes e proteínas, também foi abordada por Gissi e colaboradores (2008). Dos 2.159 representantes analisados, foram identificadas 67 proteínas diferentes, além de diversas orfs e proteínas hipotéticas. Os resultados obtidos foram comparados com os de GRAY (2012) e Kannan, Rogozin e Koonin (2014). O produto dos genes mitocondriais listadas pelos autores foram igualmente encontradas nesta pesquisa, com exceção das proteínas COX11 e COX15. Estas proteínas são codificadas pelo genoma nuclear, entretanto está presente no genoma mitocondrial leveduras (GLERUM et al., 1997; TZAGOLLOFF et al., 1999).

As proteínas ribossomais RPL35 e RPL36 também foram identificadas no arquivo de entrada, presente nos organismos *Andalucia gogoyi*, *Tsukubamonas*

globosa, *Malawimonas jakobiformis* e *Vitis vinífera,* porém, não estão presentes nos clusters analisados, pois houve somente uma ocorrência.

Com ressalva das proteínas envolvidas na fosforilação oxidativa, existe bastante variação na composição do proteôma mitocondrial entre os organismos. No domínio eucariótico, os protistas possuem a maior multiplicidade em aspectos evolutivos. O genoma mitocondrial destes organismos foi utilizado por diversos autores em estudos evolutivos (DELSUC; BRINKMANN; PHILIPPE, 2005; KANNAN; ROGOZIN; KOONIN; 2014; GRAY, 2015; PITTIS; GABALDÓN, 2016; DEGLI ESPOSTI, 2016), pelo fato de alguns representantes desse táxon possuírem praticamente todas proteínas mitocondriais identificadas até o momento.

Nos dados extraídos do RefSeq, os protistas *Jakoba libera*, *Reclinomonas americana* e *Andalucia godoyi* possuem respectivamente 84, 67 e 72 sequências de proteínas mitocondriais. Nos clusters gerados, o *Reclinomonas americana* apresentou o maior número de ocorrências, estando presente em 50 clusters diferentes, seguido de *Andalucia godoyi*, presente em 35 clusters e *Jakoba libera*, presente em 32. O número de sequências de proteínas é diferente do número de ocorrências em clusters pelo fato da variável analisada (“RSM160317.Cnew.contg2”) apresentar clusters com mais de 2 ocorrências. Isto significa que as demais sequências foram agrupadas individualmente.

As plantas também possuem grande número de proteínas diferentes codificadas pelo mtDNA. A transferência de genes funcionais é um processo contínuo em plantas, o que contribui para a evolução. Os genes mitocondriais de plantas codificam aproximadamente 40 proteínas conhecidas, das quais a maioria são proteínas ribossomais e da cadeia respiratória (LIU et al., 2009).

4.1.3 Análise do arquivo de entrada

Para realizar análise do conteúdo do proteoma mitocondrial de qualquer organismo é necessário que haja uma lista de proteínas completas já preditas para tornar possível comparações entre sequências. Para isso, é fundamental que o genoma a ser analisado esteja livre de contaminação por outros genomas não mitocondriais. Perante a dificuldade de obtenção de genomas mitocondriais

completos e validados de forma robusta, é justificável que haja poucos proteomas mitocondriais totalmente completos depositados em base de dados até o momento (GRAY, 2015).

De acordo com Gissi e colaboradores (2008), existem muitos erros de anotação nos dados genomas mitocondriais depositados no Banco de Dados de Organelas do NCBI. A maioria dos erros constatados pelos autores incluem a definição errada do gene/proteína, sequenciamentos parciais, além da falta de correções nas anotações já depositadas. Segundo O'Leary et al. (2016), os únicos dados de sequências mitocondriais que possuem cura manual são as sequências de mamíferos.

Neste trabalho, a falta de padrão na designação dos genes e proteínas prolongou a análise dos dados, uma vez que nomes diferenciados para uma mesma proteína geram dúvidas e requer busca na literatura bem como embasamento teórico consistente.

Outro aspecto que requer destaque são os dados parciais e anotações provisórias depositados no banco. Em busca textual no arquivo utilizado para a clusterização ("RSM160317.fallgaa"), a palavra "*partia*l" ocorre 1.466 vezes. Levando em consideração que o total é de 98.933 sequências, parece relativamente baixo esse número. Porém, se analisarmos o fato que existe dados mal anotados no banco, é de se esperar que o número de proteínas parciais seja maior.

Ao todo, 2.870 proteínas hipotéticas foram encontradas no banco de dados. A maior parte delas é pertencente a plantas, fungos e protozoários. Porém, em alguns casos, quando realizado BLAST, observa-se que as proteínas hipotéticas são na realidade trechos de proteínas existentes, o que pode ser resultante de anotações inadequadas ou erros de sequenciamento.

Dados anotados equivocadamente faz com que os resultados de pesquisas sejam duvidosos, e por meio da análise do arquivo de entrada, ficou evidente que o banco de dados RefSeq precisa ser revisado. Este fato também foi abordado na pesquisa de diversos autores (GISSI et al., 2008; YAO et al., 2009; BERNT et al., 2013; GRAY, 2015; SMITH, 2016).

Existe a necessidade de maior caracterização de aspectos relacionados com genomas mitocondriais, como estruturas cromossômicas, mecanismos de tradução e transcrição e genética de populações referente a eucariotos (SMITH,

2016). Para isso, é necessário um banco de dados confiável para a extração dos dados de mtDNA, pois as sequências de referência utilizadas em pesquisas biomédicas são importantes para a elaboração e reprodutibilidade de pesquisas biomédicas (O'LEARY et al., 2016).

4.1.4 Árvore Filogenética dos Clusters

A filogenia resultante dos clusters abrangeu os 6.811 organismos obtidos do RefSeq e foi realizada em aproximadamente 15 minutos. Na árvore gerada, observamos que os metazoários ficaram agrupados coerentemente nos ramos analisados, sendo muito parecida à proposta por Delsuc, Brinkmann e Philippe (2005).

Os Jakobidas ficaram agrupados de maneira similar aos autores Hampl et al. (2009), Burguer et al., (2013) e Kannan; Rogozin; Koonin (2014), o que demostra que a clusterização foi uma ferramenta eficiente para estudar e agrupar o conteúdo proteico do mtDNA.

Grande parte dos organismos agruparam na árvore filogenética respeitando a classificação dos subgrupos eucariotos (SIMPSON; ROGER, 2004; BURKI et al., 2007). Ramos muito longos em comparação aos demais ramos, nessa representação filogenética, caracterizaram organismos mal agrupados. Estes agrupamentos ocorreram devido à erros na sequencia de proteica original, o que confirmou os resultados obtidos na análise do arquivo de entrada.

De modo geral, o resultado desta filogenia foi satisfatório. Em comparação com as demais filogenias propostas nesse trabalho por representação vetorial, concluímos que a filogenia resultante da clusterização possui relevância menor, mas pode igualmente ser utilizada para representação do proteoma mitocondrial dos eucariotos.

4.2 REPRESENTAÇÃO VETORIAL DE PROTEÍNAS MITOCONDRIAIS

Na análise realizada pela clusterização, verificamos que o conjunto de dados de estudo possui algumas sequências proteicas parciais. Por conta disso,

optamos por realizar um “corte” nos dados, selecionando alguns organismos para serem excluídos do grupo de estudo.

Optamos por retirar da representação vetorial todos os organismos considerados extremos, de acordo com a referência de Kannan, Rogozin e Koonin (2014), citada na metodologia deste trabalho. Ao todo foram 14 organismos excluídos, representados na TABELA 2. Deste modo, utilizamos sequências proteicas mitocondriais de 6.797 organismos, as quais foram concatenadas, representadas sob a forma de proteoma.

TABELA 2 – ORGANISMOS EXCLUÍDOS DA PESQUISA, DE ACORDO COM O CRITÉRIO DE EXCLUSÃO.

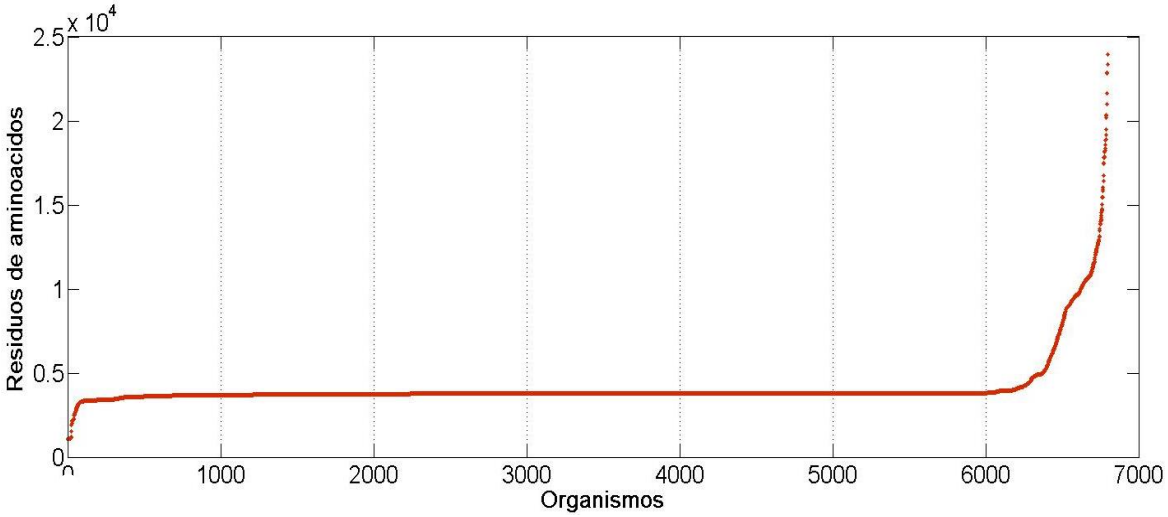
ID*	**Organismo**	**Número de Proteínas**	**Tamanho da sequência (Resíduos de aminoácidos)**
1686	*Cryphonestria parasítica*	1	73
6782	*Zea mays*	3	175
2371	*Fusarium oxysporum f. sp. matthiolae*	1	541
2372	*Fusarium oxysporum f. sp. Raphanin*	1	541
3244	*Leishmania tarentolae* *	2	557
6493	*Trichoderma harzianum*	1	621
6343	*Theileria equi* *	3	1.046
5629	*Salvia miltiorrhiza*	138	26.051
4116	*Nicotiana tabacum*	156	28.428
789	*Beta Vulgaris subsp. Vulgaris*	140	30.404
6783	*Zea Mays subsp. Mays*	163	33.765
788	*Beta vulgaris subsp. Marítima*	150	34.365
787	*Beta macrocarpa*	156	34.875
1102	*Capsicum annuum*	193	38.597

FONTE: O Autor (2017).
NOTA: A coluna ID corresponde ao identificador do organismo no arquivo fasta concatenado. * Os organismos *Leishmania tarentolae* e *Theileria equi*, mesmo sendo representantes do filo Apicomplexa, foram excluídos por possuirem número de resíduos de aminoácidos inferior ao *Plasmodium spp.*

4.2.1 Conjunto de dados

Para avaliar o conjunto de dados como um todo, plotamos um gráfico a partir de todos os proteomas do estudo, onde é possível analisar o tamanho dessas sequências expressas por resíduos de aminoácidos. No gráfico disposto, cada organismo representa um ponto (GRÁFICO 2). A variável utilizada para gerar o gráfico foi a “RSM16037.FsOrg”.

GRÁFICO 2 – TAMANHO DOS PROTEOMAS DOS 6.797 ORGANISMOS



FONTE: O Autor (2017).

Existe uma diferença grande entre o menor e o maior proteoma mitocondrial. O menor proteoma é o do *Plasmodium chibaudi chibaudi* (ID 4881), com 1.055 resíduos de aminoácidos, e o maior do *Jakoba libera* (ID 3075), com 24.648 resíduos de aminoácidos.

O conteúdo de genoma mitocondrial, bem como a ordem dos genes e o tamanho em resíduos de aminoácidos pode variar bastante de um organismo para outro. De acordo com Rand (1993), os organismos ectotérmicos (aqueles que necessitam do calor do ambiente para executar a regulação da temperatura do corpo) em relação aos endotérmicos (com capacidade de regular a temperatura corporal a partir de mecanismos internos), por exemplo, a variação do tamanho do genoma ocorre por pressões mutacionais.

Confirmamos também que a ordem dos genes nos genomas dos animais tende a ser conservada entre vertebrados para os 37 genes e as regiões não codificadoras de proteínas (rRNA), porém em animais invertebrados, plantas, fungos e protistas a ordem pode não ser a mesma, variando de espécie para espécie. Este fato também é abordado por Fox (2012), Gray (2012) e Satoh et al. (2016) em seus estudos.

4.2.2 Espaço Vetorial

A representação de sequências proteicas auxilia na análise de dados genéticos. Recentemente, Asgari e Mofrad (2015) propuseram um método de

extração de atributos, por meio de redes neuronais, utilizando dados do Swiss-prot. De acordo com os autores, o método se mostrou consistente na identificação de entropia presente em sequências, podendo ser utilizadas em várias aplicações na área de reconhecimento de padrões.

Muitas são as informações que podem ser obtidas através de uma sequência de proteína, e a representação vetorial é um método eficaz para considerar diversos aspectos referentes a sequências genéticas. À exemplo de alguns estudos, a identificação de domínios proteicos (TEICHERT; PORTO, 2006; ASGARI, MOFRAD, 2015), relações evolutivas entre espécies (DELSUC; BRINKMANN; PHILIPPE, 2005), além de perfis de hidrofobicidade e propriedades estruturais (BASTOLLA et al., 2005), são exemplos de aspectos que podem ser evidenciados por meio de representações vetoriais.

Mapear sequências genéticas em espaços vetoriais está se tornando uma área de pesquisa promissora (VINGA, 2014). A representação por k-mers espaçados tem norteado diversos estudos recentes (BODEN et al., 2013; LEIMEISTER et al., 2014; HORWEGE et al., 2014; NOÉ; MARTIN, 2014; BRINDA; SYKULSKI; KUCHEROV, 2015), e quando comparado com k-mers contíguo, se mostrou ser mais eficiente na estimativa de distancias filogenéticas e, consequentemente, melhor para a reconstrução de filogenias (BRINDA; SYKULSKI; KUCHEROV, 2015).

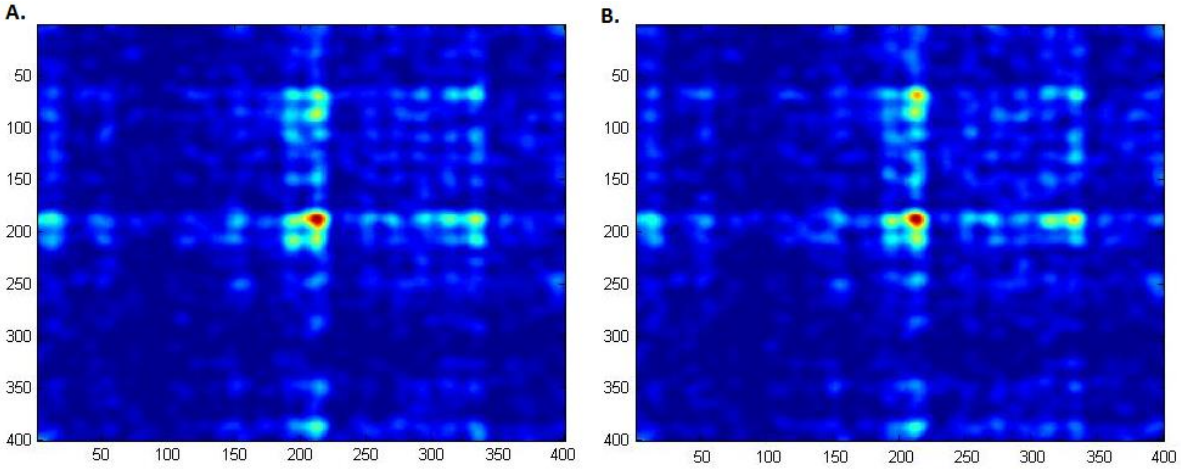
Neste trabalho, a construção dos vetores levou poucos minutos para ser concluída.

4.2.3 Representação dos organismos, de acordo com a matriz de ocorrência.

Após a criação das matrizes e vetores que representam o proteoma mitocondrial de cada organismo presente no conjunto de dados, utilizamos a matriz Ms400, que contém as ocorrências de aminoácidos, para gerar representações gráfica dos proteomas mitocondriais. Para isso, utilizamos a função “mmvfuzzy2d” (ANEXO 3). Esta função emprega média móvel Fuzzy para verificar e suavizar os picos de maior ocorrência entre as combinações de aminoácidos presentes na matriz Ms400. A figura abaixo (FIGURA 7), elaborada

considerando 10 pontos na vizinhança, mostra as poucas diferenças existentes entre os proteomas do *Homo sapiens* e *Pan paniscus*.

FIGURA 7 – MAPA DO PROTEOMA MITOCONDRIAL DE *Homo sapiens* X *Pan paniscus*

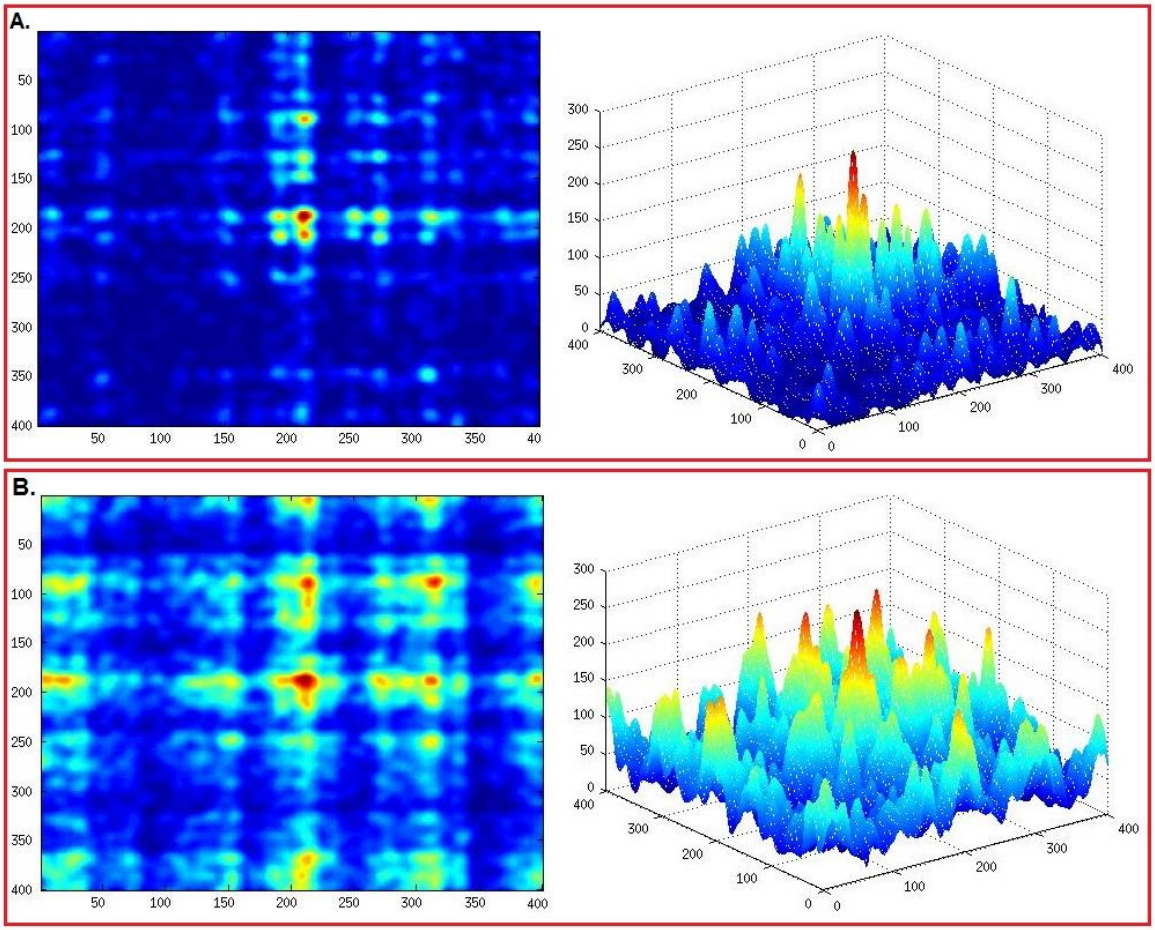


FONTE: O Autor (2017).
LEGENDA: **A.** *Homo sapiens*. **B.** *Pan paniscus*.
NOTA: Representação dos proteomas mitocondriais, segundo a função “mmvfuzzy2d”. Os picos em tons avermelhados refletem as áreas suavizadas no proteoma onde existe maior ocorrência de combinações de aminoácidos na vizinhança.

Os proteomas mitocondriais de ambos organismos possuem 3.789 resíduos de aminoácidos. Estes proteomas são destacados aqui pois, segundo estudos conduzidos nas últimas décadas, macacos e humanos possuem ascendência comum, sendo o bonobo (*Pan paniscus*) e o chimpanzé (*Pan troglodytes*) os mais semelhantes ao homem moderno (BRADLEY, 2008; BEGUN, 2010; DAS et al., 2014).

Para demonstrar que as diferenças entre proteomas são facilmente identificáveis adotando esta forma de representação da matriz Ms400, exemplificamos utilizando os proteomas de *Drosophila melanogaster* (3.750 resíduos de aminoácidos) e *Arabidopsis thaliana* (22.863 resíduos de aminoácidos), evidenciando a execução em duas e três dimensões (FIGURA 8).

FIGURA 8 – MAPAS DOS PROTEOMAS MITOCONDRIAIS DE *Drosophila melanogaster* E *Arabidopsis thaliana*



FONTE: O Autor (2017).
LEGENDA: **A.** *Drosophila melanogaster*. **B.** *Arabidopsis thaliana*.
NOTA: Ambos proteomas mitocondriais estão representados em duas e três dimensões. Os picos em tons avermelhados refletem as áreas suavizadas nos proteomas onde existe maior ocorrência de combinações de aminoácidos. As imagens foram elaboradas utilizando 10 pontos na vizinhança.

Nas FIGURA 7 e FIGURA 8 é possível visualizar os locais onde há maior ocorrências de combinações de aminoácidos. Na matriz Ms400, os locais onde ocorre *match*, é representado pelo número 1, e onde ocorre *mismatch* pelo número 0. A média se desloca de acordo com a cobertura estabelecida na vizinhança, que pode ser ajustada. Conforme aumenta a concentração de ocorrências, a cor vai variando em uma escala de azul escuro (menos ocorrências) até o vermelho escuro (mais ocorrências).

Medindo a diferença entre as distâncias euclidiana dos vetores dos organismos representados na FIGURA7 e na FIGURA 8, obtivemos 31,84 de diferença entre o *Homo sapiens* e o *Pan paniscus* e 141,05 entre a *Drosophila*

melanogaster e a *Arabidopsis thaliana*. Essas medidas são justificáveis se compararmos com a distância entre o *Homo sapiens* e o *Homo neanderthalis* (espécie humana extinta), que é de 16.

Os organismos da FIGURA 7 ficaram muito próximos nas árvores filogenéticas construídas, ao passo que os organismos da FIGURA 8 ficaram bem distantes. Estes fatos foram abordados nos resultados das filogenias.

4.3 RECONSTRUÇÃO FILOGENÉTICA

Quando realizamos buscas por artigos científicos utilizando a palavra-chave “livre de alinhamento” (*alignment-free*), nos deparamos com um grande número de publicações relacionadas ao termo, tanto na divulgação de novas abordagens quanto da utilização de abordagens já existentes. Com novas tecnologias sendo inseridas no estudo do conteúdo genético dos seres vivos, a classificação “livre de alinhamento” acaba se tornando um pouco ampla demais, tendo em vista que a filogenia e comparação de sequências biológicas podem ser feitas utilizando vários algoritmos diferentes. As vantagens das metodologias livres de alinhamento, quando se compara genomas completos, são abordadas na literatura por Otu e Sayood (2003), Sims et al. (2009), Delsuc, Brinkmann e Philippe (2005), Burki (2014) e Lemeister et al. (2014).

Uma das ferramentas baseada em metodologia livre de alinhamento que norteou estudos relevantes (WU et al., 2009; JUN et al., 2010; CHEN et al., 2016; ZHANG et al., 2017) é a FFP (SIMS et al., 2009. Recentemente, Zhang e colaboradores (2017) realizaram reconstrução filogenética de 3.905 genomas completos de vírus utilizando esta metodologia. Como o tamanho dos genomas virais varia consideravelmente, os autores optaram por calcular os tamanhos dos k-mers dividindo o grupo de dados em quartis, de acordo com o tamanho dos genomas. Em comparação, no método proposto aqui, não há necessidade de dividir o conjunto total de dados (proteomas) em conjuntos menores. Utilizamos k-mers espaçados de tamanho fixo para toda a extensão dos proteomas, obtendo resultados igualmente satisfatórios, com menor custo computacional.

Optamos por utilizar o ‘UPGMA’ para a inferência das árvores filogenéticas neste trabalho, pois não estimamos o tempo de divergência entre as espécies, e

sim avaliamos a coerência do método de representação vetorial proposto aqui. Nas análises das árvores filogenéticas com e sem redução de dimensão, utilizamos dentre os 6.797 organismos, os Hominídios e os Jakobidas como exemplo.

Expomos parte dos resultados de nossas reconstruções filogenéticas utilizando os Jakobidas com exemplo, motivados pelo grande conteúdo proteico do genoma mitocondrial, fato que foi abordado e reconfirmado com a clusterização neste trabalho. Os Hominídios foram utilizados como exemplo devido ao genoma mitocondrial de mamíferos ser bem conservado.

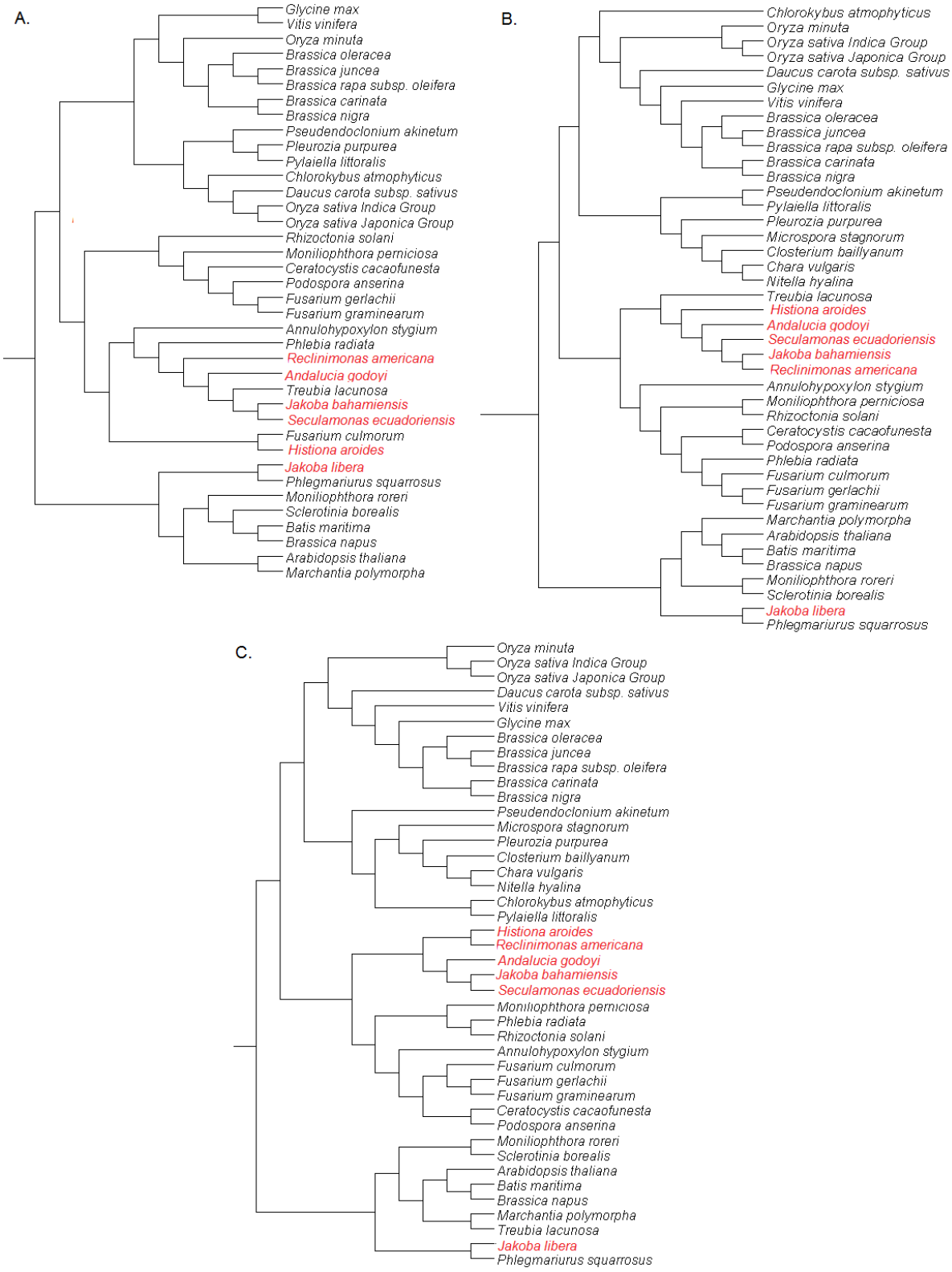
4.3.1 Representação dos Jakobidas

A utilização de genomas completos pode afetar a distribuição de organismos em árvores construídas por metodologias baseadas em alinhamento pois, segundo Forterre (2015), a composição genômica global é influenciada pela transferência horizontal de genes. A falta de padrão na representação proteínas em alguns organismos pode ocasionar representações filogenéticas fracas, como no caso das plantas e fungos, que apresentam maior variabilidade na ordem dos genes em relação ao mtDNA animal. Recombinações, repetições e regiões intergênicas também podem resultar em agrupamentos filogenéticos equivocados (AGUILETA et al., 2014).

Uma das vantagens da metodologia proposta neste trabalho é que a ordem de ocorrência das sequencias de proteínas nos proteomas, bem como áreas de repetições, não afeta a construção das árvores filogenéticas, tornando a distribuição dos organismos nas árvores coerente.

A FIGURA 9 mostra a representação dos organismos Jakobidas nas árvores filogenéticas Wr100, Wr400 e Wr800. Estes organismos aparecem muito entre sí, juntamente com outros organismos contemplados na figura abaixo.

FIGURA 9 – COMPARAÇÃO ENTRE OS JAKOBIDAS NAS ÁRVORES Wr100, Wr400 e Wr800.

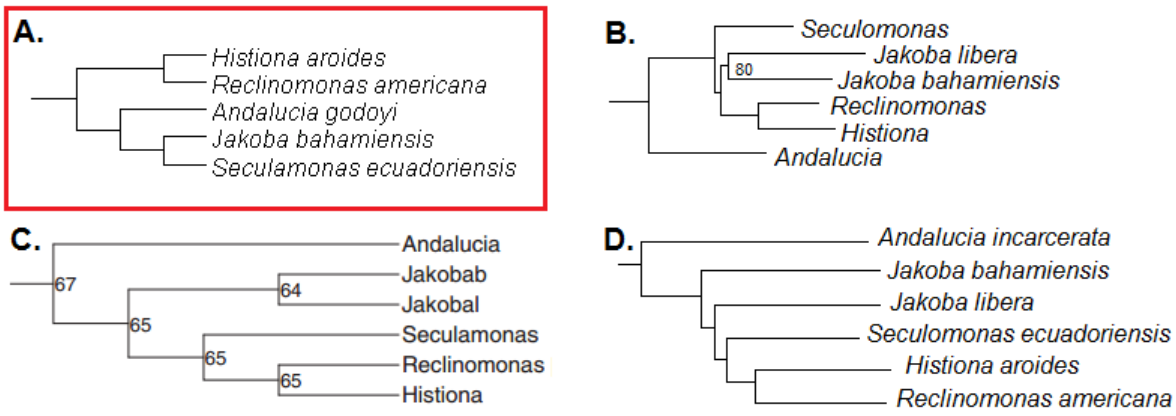


FONTE: O autor (2017).
LEGENDA: **A.** Árvore com 100 atributos. **B.** Árvore com 400 atributos. **C.** Árvore com 800 atributos.
NOTA: Os jakobidas estão representados em vermelho.

Nos nós das árvores Wr100, Wr400 e Wr800 que estão representados os Jakobidas, não houve diferenças significativas. Os jakobidas encontram-se próximos entre sí. Grande parte dos organismos representados nesse nó são plantas, e sabe-se que as plantas possuem um grande conjunto de proteínas codificadas pelo mtDNA, sendo a maioria produto subunidades ribossomais, além de regiões intrônicas que contribuem para aumentar o mtDNA (AGUILETA et al., 2014).

O organismo *Jakoba libera*, apontado neste trabalho como o de maior proteoma mitocondrial, em relação ao número de resíduos de aminoácidos, está representado na FIGURA 9 em um nó separado do organismo *Jakoba bahamiensis*. Nas filogenias mitocondriais de Hampl et al. (2009), Burguer et al., (2013) e Kannan; Rogozin; Koonin (2014), estes organismos estão representados no mesmo nó, como pode ser visto na comparação realizada na FIGURA 10.

FIGURA 10 – REPRESENTAÇÃO DAS RELAÇÕES DE PARENTESCO ENTRE OS JAKOBIDAS NA ÁRVORE Wr800 EM COMPARAÇÃO COM ÁRVORES CONSOLIDADAS.



FONTE: Adaptado de Hampl et al. (2009), Burguer et al. (2013) e Kannan; Rogozin; Koonin (2014).
LEGENDA: **A.** Árvore Wr800 **B.** Filogenia proposta por Burguer et al. (2013). **C.** Filogenia proposta por Kannan; Rogozin; Konnin (2014). **D.** Filogenia proposta por Hampl et al. (2009).

Nas filogenias acima (FIGURA 10), o organismo *Andalucia godoyi* é apontado como sendo o mais ancestral, convergindo para as outras espécies de Jakobidas. Utilizamos a árvore Wr800 (no quadrado vermelho) como comparação, pois ela agrupou quase todos os Jakobidas no mesmo clado, com exceção de um organismo.

O *Jakoba libera* é o único dos Jakobidas representados aqui que possui o mtDNA linear, além de um maior número de repetições em relação aos outros Jakobidas. De acordo com Burguer et al. (2013), maiores caracterizações bioquímica do mtDNA desse organismo são necessárias para realizar conclusões a respeito de parentesco. O mtDNA de Jakobidas sofreu muitas perdas de genes no decorrer de processos evolutivos (GRAY et al., 1998). A diversidade de proteínas codificadas pelo mtDNA nesses organismos é abordada detalhadamente no estudo de Burguer et al. (2013).

Para avaliar a filogenia proposta de outra maneira, utilizando um conjunto de organismos com proteínas mitocondriais totalmente evidenciadas, optamos por representar os primatas Hominídios, que possuem uma história evolutiva bem esclarecida na literatura. Em contraste aos Jakobidas e as plantas, os mamíferos possuem as regiões mtDNA bem conservadas (GISSI et al., 2008), conforme constatado nos resultados da clusterização.

4.3.2 Representação dos Hominídeos

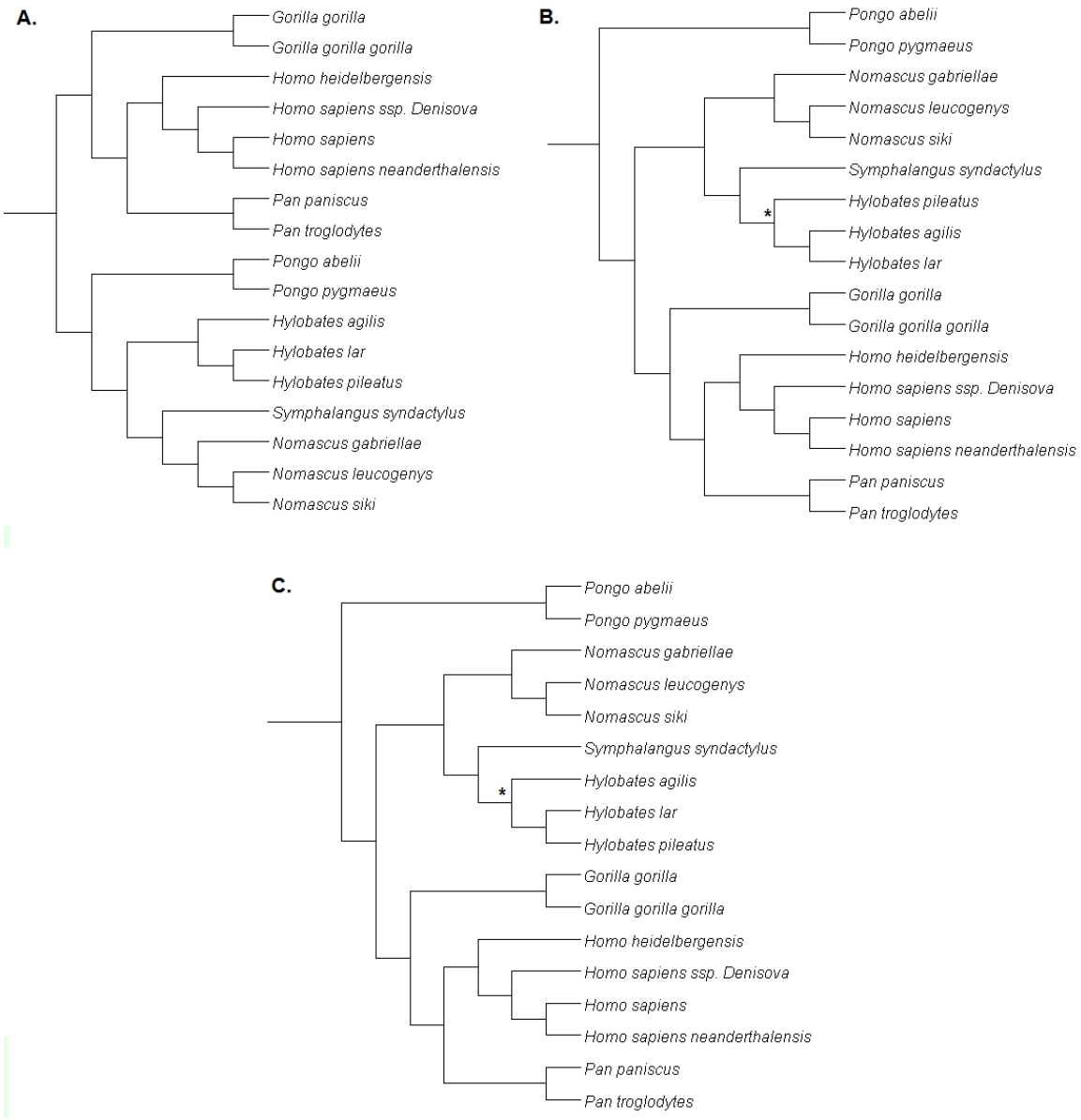
Árvores filogenéticas evidenciando o parentesco dos seres humanos com outros mamíferos são bem populares (CHATTERJEE et al., 2009; CAMPBELL; LAPOINTE, 2010; FINSTERMEIER et al., 2013; DAS et al., 2014). É comum encontrarmos em livros e certos artigos científicos imagens de humanos associadas aos macacos, porém a história evolutiva dos *Homo sapiens* é mais complexa que uma simples associação por características fenotípicas.

Segundo a história evolutiva, os hominídeos divergiram em pequenos macacos (Hyalobatidae) e grandes símios, representados pelos primatas mais próximos ao homem. O Nomascus foi o primeiro representante do grupo dos pequenos macacos a divergir, seguido do Symphalangus e Hylobates. Os grandes símios orangotangos (Pongo) divergiram dos grandes macacos africanos e dos seres humanos, enquanto o Gorilla se separou do Homo e do Pan. Por fim, os chimpanzés e os seres humanos se divergiram (FINSTERMEIER et al., 2013).

Utilizando o proteoma do *Homo sapiens* como exemplo, realizamos comparações entre os nós da árvore onde constam os hominídios, a fim de

confirmar as evidências evolutivas propostas pela maioria dos autores. A FIGURA 11 apresenta o nó nas árvores geradas com 100, 400 e 800 atributos.

FIGURA 11 - COMPARAÇÃO ENTRE OS NÓS DOS HOMINÍDEOS NAS ÁRVORES Wr100, Wr400 e Wr800.

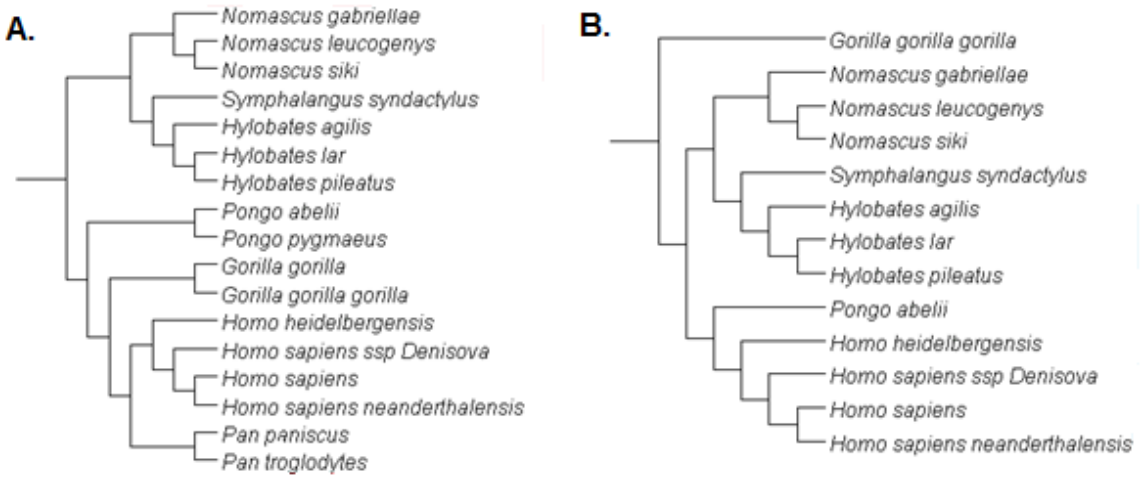


FONTE: O Autor (2017).
LEGENDA: **A.** Árvore gerada com 100 atributos. **B.** Árvore gerada com 400 atirbutos. **C.** Árvore gerada com 800 atributos. * Nó em que houve diferenças entre B e C.

Note que, se associarmos com a história evolutiva dos primatas, de acordo Chatterjee et al. (2009), Campbell e Lapointe (2010), Finstermeier et al. (2013), Das et al. (2014), com as árvores de 400 e 800 atributos são bem coerentes e praticamente iguais, sendo a única diferença entre as duas é o nó em que se

encontra o Hylobates. A árvore de 100 atributos é menos sensível. Realizamos também a reconstrução filogenética sem redução de dimensão, utilizando o vetor W160k, proveniente da matriz de Ms400 e comparamos com o método de alinhamento (FIGURA 12).

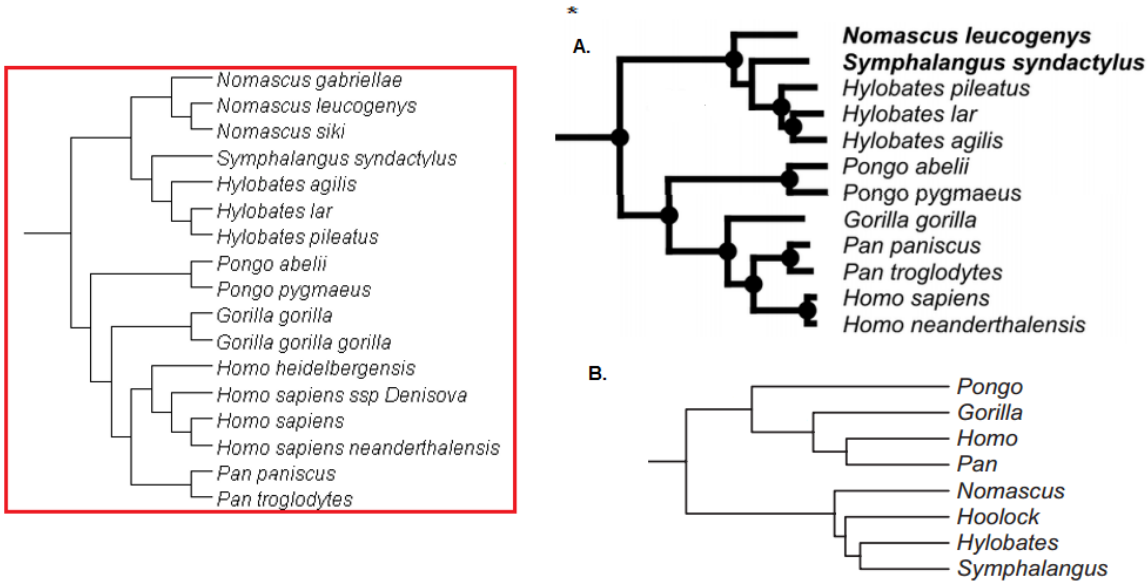
FIGURA 12 - FILOGENIA W160k X FILOGENIA DE ALINHAMENTO



FONTE: O autor (2017).
LEGENDA: **A.** árvore gerada sem redução de dimensão, com 160k. **B.** árvore gerada com alinhamento, utilizando s função “Selfalign2”.

A árvore gerada a partir do vetor W160k, é muito semelhante à representação de Chatterjee et al. (2009) e Finstermeier et al. (2013), conforme visualizada na FIGURA 13.

FIGURA 13 – COMPARAÇÃO ENTRE A FILOGENIA DE 160K COM A LITERATURA



FONTE: Adaptado de Finstermeier et al. (2013), Chatterjee et al. (2009).
LEGENDA: Em vermelho filogenia proposta sem redução de dimensão, utilizando vetor de 160k. **A.** Filogenia proposta por Finstermeier et al. (2013). **B.** Filogenia proposta por Chatterjee et al. (2009). * O organismo Hoolock não está representado em nossa filogenia.

Houve uma certa dificuldade para encontrar estudos filogenéticos que pudessem ser utilizados como referência, tendo que vista que as árvores geradas por meio de nossa metodologia são diferentes das metodologias utilizadas nos consensos. As árvores utilizadas como comparação foram obtidas por meio de alinhamentos concatenados.

A utilização de *supertrees* para reconstruir filogenias com grande número de sequências genômicas, usada por Delsuc, Brinkmann e Philippe, (2005), Binida-Emonds et al. (2007), Chatterjee et al. (2009), Campbell e Lapointe (2010) e Burki (2007), é eficiente para demonstrar relações de parentesco e evolutivas, porém demanda tempo. Em comparação, o método proposto aqui com redução de dimensão é extremamente ágil, produzindo árvores filogenéticas em segundos, com a mesma eficiência na representação de relações de parentesco.

A correlação de Pearson entre a distância *pairwise* das árvores de 160K e de sequências é de aproximadamente 70%. Obtivemos resultados melhores com a correlação de Spearman, com aproximadamente 85%. Isso demonstra que a árvore W160k é correlata à árvore gerada por alinhamento de sequências. As correlações de Pearson e Spearman entre as distâncias de todas as árvores é apresentada QUADRO 2.

QUADRO 2 – CORRELAÇÃO DE PEARSON E SPEARMAN ENTRE A DISTÂNCIA *PAIRWISE* DAS FILOGENIAS

	SelfAlign	Wr100	Wr400	Wr800	W160k
SelfAlign	100	37,39	39,28	41,34	68,01
Wr100	81,24	100	99,83	99,52	84,34
Wr400	83,53	93,99	100	99,88	85,98
Wr800	84,37	92,31	97,54	100	87,64
W160k	84,97	86,96	90,69	92,52	100

FONTE: O autor (2017)
LEGENDA: Em cinza: correlação de Pearson. Em laranja: correlação de Spearman.

Analisando a filogenia como um todo, não somente a representação dos Jakobidas ou a dos Hominídeos, as árvores resultantes do vetor Wr800 e W160k estão com os ramos mais bem resolvidos em relação às outras árvores com reduções maiores. Isto é refletido principalmente nos trechos das árvores onde o conjunto de dados é mais ordenado, como no caso dos animais.

5. CONCLUSÃO

Analisando a etapa de clusterização como um todo, obteve-se uma visão geral do comportamento do mtDNA nos diversos organismos, o que foi satisfatório. O agrupamento realizado pelo RAFTS3groups se mostrou rápido e eficiente. Praticamente todas as proteínas codificadas pelo genoma mitocondrial puderam ser evidenciadas e isto ajudou na construção da filogenia.

De modo geral, as três árvores com redução de dimensão (W100, W400 e W800) são válidas e representam a história evolutiva. A filogenia W160k se mostrou tão eficiente quanto a filogenia resultante de alinhamento. Os nós que possuem grupos com grande número de representantes são bem consistentes quando comparados com reconstruções de outros autores ao passo que, os grupos que possuem um número pequeno de representantes, podem estar instavelmente representados na árvore.

A princípio, podemos afirmar que nosso método de representação vetorial é tão eficiente quanto o alinhamento para representar os organismos nas árvores filogenéticas. Com isso, o que pode-se questionar é se as metodologias utilizadas em grande parte das filogenias até o momento, são realmente a melhor maneira de evidenciar processos evolutivos e expor todos os aspectos presentes em um proteoma, uma vez que utilizar somente partes conservadas de alinhamentos concatenados gera perda de informação.

Como perspectivas futuras, pretendemos realizar mais testes utilizando nosso método em um conjunto de dados bem estruturado, com o proteoma mitocondrial de todos os organismos ordenados igualmente. Com isso, acreditamos que as representações filogenéticas ficarão mais próximas ainda da real distribuição dos organismos nas árvores, com potencial para ser um método melhor que o alinhamento de sequencias.

REFERÊNCIAS

ADAMS, K. L.; PALMER, J. D. Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. **Molecular Phylogenetics and Evolution**, v. 29, n. 3, p. 380–395, 2003.

AGUILETA, G.; DE VIENNE, D. M.; ROSS, O. N.; et al. High variability of mitochondrial gene order among fungi. **Genome Biology and Evolution**, v. 6, n. 2, p. 451–465, 2014.

ALBERT, G.; JUN, S.-R.; SIMS, G. E.; et al. Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. **Proceedings of the National Academy of Sciences of the United States of America**, v. 106, n. 31, p. 12826–31, 2009. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2722272&tool=pmcentrez&rendertype=abstract>. .

ALCOLADO, J. C.; THOMAS, A. W. Maternally inherited diabetes mellitus: the role of mitochondrial DNA defects. **Diabet. Med.**, v. 12, p. 102–108, 1995.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403–10, 1990. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0022283605803602>. .

APWEILER, R.; BAIROCH, A.; WU, C. H.; et al. UniProt: the Universal Protein knowledgebase. **Nucleic acids research**, v. 32, n. Database issue, p. D115-9, 2004. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308865&tool=pmcentrez&rendertype=abstract>. .

ARUNACHALAM, M.; JAYASURYA, K.; TOMANCAK, P.; OHLER, U. An alignment-free method to identify candidate orthologous enhancers in multiple Drosophila genomes. **Bioinformatics**, v. 26, n. 17, p. 2109–2115, 2010.

ASGARI, E.; MOFRAD, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. **PLoS ONE**, v. 10, n. 11, p. 1–15, 2015.

BAIROCH, A.; APWEILER, R.; WU, C. H.; et al. The Universal Protein Resource (UniProt). **Nucleic Acids Research**, v. 33, n. DATABASE ISS., p. 154–159, 2005.

BARRIENTOS, A.; ZAMBRANO, A.; TZAGOLOFF, A. Mss51p and Cox14p jointly regulate mitochondrial Cox1p expression in Saccharomyces cerevisiae. **The EMBO journal**, v. 23, n. 17, p. 3472–3482, 2004.

BASTOLLA, U.; PORTO, M.; ROMAN, H. E. H.; M; VENDRUSCOLO, M. Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. **Proteins:**, v. 58, n. 1, p. 22–30, 2004. Disponível em: <http://arxiv.org/abs/q-bio/0406003%5Cnhttp://www3.interscience.wiley.com/journal/1096/abstract>.

BEGUN, D. R. Miocene Hominids and the Origins of the African Apes and Humans. **Annual Review of Anthropology**, v. 39, p. 67–84, 2010.

BENSON, D. A.; CAVANAUGH, M.; CLARK, K.; et al. GenBank. **Nucleic Acids Research**, v. 41, n. D1, p. 36–42, 2013.

BERNT, M.; BRABAND, A.; MIDDENDORF, M.; et al. Bioinformatics methods for the comparative analysis of metazoan mitochondrial genome sequences. **Molecular Phylogenetics and Evolution**, v. 69, n. 2, p. 320–327, 2013. Elsevier Inc. Disponível em: <http://dx.doi.org/10.1016/j.ympev.2012.09.019>. .

BININDA-EMONDS, O. R. P.; CARDILLO, M.; JONES, K. E.; et al. The delayed rise of present-day mammals. **Nature**, v. 446, n. March, p. 507–512, 2007. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/17392779>. .

BININDA-EMONDS, O. R. P.; SANDERSON, M. J.; BIOLOGY, S.; AUG, N.; SANDERSON, J. Assessment of the Accuracy of Matrix Representation with Parsimony Analysis Supertree Construction Assessment of the Accuracy of Matrix Representation with Parsimony Analysis Supertree Construction. , v. 50, n. 4, p. 565–579, 2001.

BODEN, M.; SCHÖNEICH, M.; HORWEGE, S. Alignment-free sequence comparison with spaced k-mers. **German Conference on Bioinformatics 2013**, v. 34, p. 24–34, 2013.

BORGES-OSÓRIO, M. R.; ROBINSON, W. M. **Genética Humana**, 3 ed. Artmed: Porto Alegre, 2013.

BOORE, J. L. Animal mitochondrial genomes. **Nucleic Acids Res**, v. 27, n. 8, p. 1767–1780, 1999. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/10101183>. .

BOURENS, M.; BARRIENTOS, A. A *CMC1* -knockout reveals translation-independent control of human mitochondrial complex IV biogenesis. **EMBO reports**, v. 18, n. 3, p. 477–494, 2017. Disponível em: <http://embor.embopress.org/lookup/doi/10.15252/embr.201643103>. .

BRADLEY, B. J. Reconstructing phylogenies and phenotypes: A molecular view of human evolution. **Journal of Anatomy**, v. 212, n. 4, p. 337–353, 2008.

BRANDON, M. C.; LOTT, M. T.; NGUYEN, K. C.; et al. MITOMAP: A human mitochondrial genome database - 2004 update. **Nucleic Acids Research**, v. 33, n. DATABASE ISS., p. 2004–2006, 2005.

BROWN, J. R.; DOOLITTLE, W. F. Archaea and the prokaryote-to-eukaryote transition. **Microbiology and molecular biology reviews : MMBR**, v. 61, n. 4, p. 456–502, 1997. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/10664676>.

BRINDA, K.; SYKULSKI, M.; KUCHEROV, G. Spaced seeds improve k-mer-based metagenomic classification. **Bioinformatics**, v. 31, n. 22, p. 3584–3592, 2015.

BURGER, G.; GRAY, M. W.; FORGET, L.; LANG, B. F. Strikingly bacteria-like and gene-richmitochondrial genomes throughout jakobid protists. **Genome Biology and Evolution**, v. 5, n. 2, p. 418–438, 2013.

BURKI, F. The eukaryotic tree of life from a global phylogenomic perspective. **Cold Spring Harbor Perspectives in Biology**, v. 6, n. 5, p. 1–17, 2014.

BURKI, F.; SHALCHIAN-TABRIZI, K.; MINGE, M.; et al. Phylogenomics reshuffles the eukaryotic supergroups. **PLoS ONE**, v. 2, n. 8, p. 1–6, 2007.

CALVO, S. E.; CLAUSER, K. R.; MOOTHA, V. K. MitoCarta2.0: An updated inventory of mammalian mitochondrial proteins. **Nucleic Acids Research**, v. 44, n. D1, p. D1251–D1257, 2016.

CAMPBELL, V.; LAPOINTE, F. J. An application of supertree methods to mammalian mitogenomic sequences. **Evolutionary Bioinformatics**, v. 2010, n. 6, p. 57–71, 2010.

CARVALHO, M. C. D. C. G. DE; SILVA, D. C. G. DA. Sequênciamento de DNA de nova geração e suas aplicações na genômica de plantas. **Ciência Rural**, v. 40, n. 3, p. 735–744, 2010.

CHATTERJEE, H. J.; HO, S. Y. W.; BARNES, I.; GROVES, C. Estimating the phylogeny and divergence times of primates using a supermatrix approach. **BMC evolutionary biology**, v. 19, p. 1–19, 2009. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2774700&tool=pmcentrez&rendertype=abstract>. .

CHEN, S.; DENG, L.-Y.; BOWMAN, D.; et al. Phylogenetic tree construction using trinucleotide usage profile (TUP). **BMC Bioinformatics**, v. 17, n. S13, p. 381, 2016. BMC Bioinformatics. Disponível em: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1222-3>. .

COIMBRA, N. A. R. Metodologia computacional para estudo de genes com vizinhança conectada: análise do cluster nif. Dissertação de mestrado (Pós-graduação em Bioinformática) - Universidade Federal do Paraná, Curitiba, 2015.

COOPER, G. M. **The cell: A Molecular Approach**, 2 ed. Boston University: Sunderland, 2000.

COTTER, D.; GUDA, P.; FAHY, E.; SUBRAMANIAM, S. MitoProteome: mitochondrial protein sequence database and annotation system. **Nucleic Acids Res.**, v. 32, n. Database issue, p. D463–D467, 2004. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/14681458>. .

DAMAS, J.; CARNEIRO, J.; AMORIM, A.; PEREIRA, F. MitoBreak: The mitochondrial DNA breakpoints database. **Nucleic Acids Research**, v. 42, n. D1, p. 1261–1268, 2014.

DAS, R.; HERGENROTHER, S. D.; SOTO-CALDERÓN, I. D.; et al. Complete mitochondrial genome sequence of the eastern gorilla (Gorilla beringei) and

implications for african ape biogeography. **Journal of Heredity**, v. 105, n. 6, p. 752–761, 2014.

DELSUC, F.; BRINKMANN, H.; PHILIPPE, H. Phylogenomics and the reconstruction of the tree of life. **Nature reviews. Genetics**, v. 6, n. 5, p. 361–375, 2005.

DEGLI ESPOSTI, M. Late Mitochondrial Acquisition, Really? **Genome Biology and Evolution**, v. 8, n. 6, p. 2031–2035, 2016. Disponível em: <http://gbe.oxfordjournals.org/lookup/doi/10.1093/gbe/evw130>. .

DENNERLEIN, S.; REHLING, P. Human mitochondrial COX1 assembly into cytochrome c oxidase at a glance. **Journal of cell science**, , n. February, p. 1–5, 2015. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/25663696>. .

DONI, M. Análise de Cluster: Métodos Hierárquicos e de Particionamento. Trabalho de Graduação (Bacharelado em Sistemas de Informação) – Faculdade de Computação e Informática, Universidade Presbiteriana Mackenzie, 2004.

DOUADY, C. J.; DELSUC, F.; BOUCHER, Y.; DOOLITTLE, W. F.; DOUZERY, E. J. P. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. **Molecular Biology and Evolution**, v. 20, n. 2, p. 248–254, 2003.

EDGAR, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792–1797, 2004.

FINSTERMEIER, K.; ZINNER, D.; BRAMEIER, M.; et al. A Mitogenomic Phylogeny of Living Primates. **PLoS ONE**, v. 8, n. 7, p. 1–10, 2013.

FORTERRE, P. The universal tree of life: An update. **Frontiers in Microbiology**, v. 6, n. JUN, p. 1–18, 2015.

FOX, T.D. Mitochondrial Protein Synthesis, Import, and Assembly. Genetics, v.192, n. 4, p.1203-1234, 2012.

FRALEY, C.; RAFTERY, A E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. **The Computer Journal**, v. 41, n. 8, p. 578–588, 1998. Disponível em: <http://comjnl.oxfordjournals.org.proxy.lib.uiowa.edu/content/41/8/578>. .

GISSI, C.; IANNELLI, F.; PESOLE, G. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. **Heredity**, v. 101, p. 301–32062, 2008.

GLERUM, D. M.; MUROFF, I.; JIN, C.; TZAGOLOFF, A. COX15 codes for a mitochondrial protein essential for the assembly of yeast cytochrome oxidase [In Process Citation]. **J Biol Chem**, v. 272, n. 30, p. 19088–19094, 1997.

GRAY, M. W. Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. **Proceedings of the National Academy of Sciences of the United States of America**, v. 112, n. 33, p. 10133–8, 2015.

Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4547279&tool=pmcentrez&rendertype=abstract>. .

GRAY, M. W.; GRAY, M. W.; BURGER, G.; LANG, B. F. Mitochondrial Evolution. , v. 1476, n. 1999, p. 1–16, 2008.

GRAY, M. W.; LANG, B. F.; CEDERGREN, R.; et al. Genome structure and gene content in protist mitochondrial DNAs. **Nucleic Acids Research**, v. 26, n. 4, p. 865–878, 1998.

GRAZINA, M.; PRATAS, J.; SILVA, F.; et al. Genetic basis of Alzheimer's dementia: Role of mtDNA mutations. **Genes, Brain and Behavior**, v. 5, n. SUPPL. 2, p. 92–107, 2006.

GUO, B.; ZHAI, D.; CABEZAS, E.; WELSH, K.; NOURAINI, S.; SATTERTHWAIT, A.C.; REED, J.C. Humanin peptide suppresses apoptosis by interfering with Bax activation. **Nature**, v.423, n. 6938, p.456–461, 2003.

HAMPL, V.; HUG, L.; LEIGH, J. W.; et al. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". **Proceedings of the National Academy of Sciences of the United States of America**, v. 106, n. 10, p. 3859–64, 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/19237557%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2656170>. .

HEBERT, P. D. N.; RATNASINGHAM, S.; WAARD, J. Barcoding animal life : cytochrome c oxidase subunit 1 divergences among closely related species Barcoding animal life : cytochrome c oxidase subunit 1 divergences among closely related species. **Proc. R. Soc. Lond. B**, v. 270, n. figure 1, p. S96–S99, 2003.

HOLDER, M.; LEWIS, P. O. Phylogeny estimation: traditional and Bayesian approaches. **Nature Reviews Genetics**, v. 4, n. 4, p. 275–284, 2003.

HORWEGE, S.; LINDNER, S.; BODEN, M.; et al. Spaced words and kmacs: Fast alignment-free sequence comparison based on inexact word matches. **Nucleic Acids Research**, v. 42, n. W1, p. 7–11, 2014.

HUG, L. A.; BAKER, B. J.; ANANTHARAMAN, K.; et al. A new view of the tree of life. **Nature Microbiology**, v. 1, n. 5, p. 16048, 2016. Disponível em: <http://www.nature.com/articles/nmicrobiol201648>. .

HUSON, D. H.; SCORNAVACCA, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. **Systematic Biology**, v. 61, n. 6, p. 1061–1067, 2012.

HÜTTEMANN, M.; LEE, I.; SAMAVATI, L.; YU, H.; DOAN, J. W. Regulation of mitochondrial oxidative phosphorylation through cell signaling. **Biochimica et Biophysica Acta (BBA) - Molecular Cell Research**, v. 1773, n. 12, p. 1701–1720, 2007. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167488907002364>. .

INGMAN, M.; GYLLENSTEN, U. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. **Nucleic acids research**, v. 34, n. Database issue, p. D749–D751, 2006.

JACKSON, J. Communications of the Association for Information Systems Data Mining; A Conceptual Overview DATA MINING: A CONCEPTUAL OVERVIEW. **Communications of the Association for Information Systems**, v. 8, n. 8, p. 267–296, 2002. Disponível em: <http://aisel.aisnet.org/cais%5Cnhttp://aisel.aisnet.org/cais/vol8/iss1/19>. .

JENSEN, R. A. Orthologs and paralogs - we need to get it right. **Genome biology**, v. 2, n. 8, p. interactions1002.1-interactions1002.3, 2001.

JUN, S.-R.; SIMS, G. E.; WU, G. A; KIM, S.-H. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. **Proceedings of the National Academy of Sciences of the United States of America**, v. 107, n. 1, p. 133–138, 2010. Disponível em: <http://www.pnas.org/content/107/1/133.long>. .

JUNQUEIRA, L. C. U.; CARNEIRO, J. **Biologia celular e molecular**, 8 ed. Guanabara Koogan: Rio de Janiero, 2005.

KANEHISA, M.; GOTO, S.; SATO, Y.; et al. Data, information, knowledge and principle: Back to metabolism in KEGG. **Nucleic Acids Research**, v. 42, n. D1, p. 199–205, 2014.

KANNAN, S.; ROGOZIN, I. B.; KOONIN, E. V. MitoCOGs: clusters of orthologous genes from mitochondria and implications for the evolution of eukaryotes. **BMC Evolutionary Biology**, v. 14, n. 1, p. 237, 2014. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4256733&tool=pmcentrez&rendertype=abstract>. .

KÜHLBRANDT, W. Structure and function of mitochondrial membrane protein complexes. **BMC biology**, v. 13, n. 1, p. 89, 2015. BMC Biology. Disponível em: <http://bmcbiol.biomedcentral.com/articles/10.1186/s12915-015-0201-x>. .

KUZNETSOVA, M. V.; KHOLODOVA, M. V.; LUSCHEKINA, A. A. Phylogenetic Analysis of Sequences of the 12S and 16S rRNA Mitochondrial Genes in the Family Bovidae: New Evidence. **Russian Journal of Genetics**, v. 38, n. 8, p. 942–950, 2002.

LANG, B. F.; SEIF, E.; GRAY, M. W.; O'KELLY, C. J.; BURGER, G. A comparative genomics approach to the evolution of eukaryotes and their mitochondria. **The Journal of Eukaryotic Microbiology**, v. 46, n. 4, p. 320–326, 1999. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/10461380>. .

LEIMEISTER, C. A.; BODEN, M.; HORWEGE, S.; LINDNER, S.; MORGENSTERN, B. Fast alignment-free sequence comparison using spaced-word frequencies. **Bioinformatics**, v. 30, n. 14, p. 1991–1999, 2014.

LEIMEISTER, C. A.; MORGENSTERN, B. Kmacs: The k-mismatch average common substring approach to alignment-free sequence comparison. **Bioinformatics**, v. 30, n. 14, p. 2000–2008, 2014.

LESK, A. M. **Introdução à Bioinformática**, 2 ed. Porto Alegre, RS: Artmed, 2008. 348 p.

LIU, S.-L.; ZHUANG, Y.; ZHANG, P.; ADAMS, K. L. Comparative analysis of structural diversity and sequence evolution in plant mitochondrial genes transferred to the nucleus. **Molecular biology and evolution**, v. 26, p. 875–891, 2009.

MAHMOOD, K.; WEBB, G. I.; SONG, J.; WHISSTOCK, J. C.; KONAGURTHU, A. S. Efficient large-scale protein sequence comparison and gene matching to identify orthologs and co-orthologs. **Nucleic Acids Research**, v. 40, n. 6, 2012.

MARGULIS, L. Biodiversity: molecular biological domains, symbiosis and kingdom origins. **BioSystems**, v. 27, n. 1, p. 39–51, 1992.

MENG, Z.; DONG, H.; LI, J.; CHEN, Z. Darwintree: A Molecular Data Analysis and Application Environment for Phylogenetic Study. **Data Science …**, p. 1–10, 2015. Disponível em: <http://datascience.codata.org/article/10.5334/dsj-2015-010/>. .

MOORE, J. H. Bioinformatics. J**ournal of Cellular Physiology**, v. 213, n. 2, p. 365-369, 2007. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/jcp.21218/full>. .

MOREIRA, S. VALACH, M.; AOULAD-AISSA, M.; OTTO, C.; BURGER, G.; Novel modes of RNA editing in mitochondria. **Nucleic Acids Res**, v.44 , n.10, p.: 4907-4919, 2016.

MORITZ, C.; DOWLING, T.; BROWN, W. M. Evolution of animal mitochondrial DNA: relevance for population biology and systematics. **Annual Review of Ecology and Systematics**, v. 18, n. 1987, p. 269–292, 1987. Disponível em: <http://www.jstor.org/stable/10.2307/2097133>.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de séries temporais**. 2. ed. São Paulo: Edgard Blücher, 2006.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of Molecular Biology**, v. 48, n. 3, p. 443–453, 1970.

NELSON, D.; COX, M. M. **Princípios de Bioquímica de Lehninger**. 6 ed. Artmed: Porto Algere, 2014.

NICHIO, B. T. L. Rapid Alignment Free Tool for Sequences Similarity Search to Groups (RAFTS3groups) – A fast clustering software for big data and consistent orthologs proteins finder. Dissertação de mestrado (Pós-graduação em Bioinformática) - Universidade Federal do Paraná, Curitiba, 2016.

NOÉ, L.; MARTIN, D. E. K. A Coverage Criterion for Spaced Seeds and Its Applications to Support Vector Machine String Kernels and *k* -Mer Distances. **Journal of Computational Biology**, v. 21, n. 12, p. 947–963, 2014. Disponível em: <http://online.liebertpub.com/doi/abs/10.1089/cmb.2014.0173>. .

O'BRIEN, E. A.; BADIDI, E.; BARBASIEWICZ, A.; et al. GOBASE - A database of mitochondrial and chloroplast information. **Nucleic Acids Research**, v. 31, n. 1, p. 176–178, 2003.

O'BRIEN, E. A.; ZHANG, Y.; WANG, E.; et al. GOBASE: An organelle genome database. **Nucleic Acids Research**, v. 37, n. SUPPL. 1, p. 946–950, 2009.

O'KELLY, C. J. The Jakobid Flagellates: Structural Features of Jakoba, Reclinomonas and Histiona and Implications for the Early Diversification of Eukaryotes. **Journal of Eukaryotic Microbiology**, v. 40, n. 5, p. 627–636, 1993.

O'LEARY, N. A.; WRIGHT, M. W.; BRISTER, J. R.; et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. **Nucleic Acids Research**, v. 44, n. D1, p. D733–D745, 2016.

OTU, H. H.; SAYOOD, K. A new sequence distance measure for phylogenetic tree construction. **Bioinformatics**, v. 19, n. 16, p. 2122–2130, 2003.

PATWARDHAN, A.; RAY, S.; ROY, A. Molecular Markers in Phylogenetic Studies-A Review. **Journal of Phylogenetics & Evolutionary Biology**, v. 2, n. 2, p. 1–9, 2014. Disponível em: <http://esciencecentral.org/journals/molecular-markers-in-phylogenetic-studiesa-review-2329-9002-2-131.php?aid=30965>. .

PEARSON, W. R. An introduction to sequence similarity ("homology") searching. **Current Protocols in Bioinformatics**, , n. SUPPL.42, p. 1–8, 2013.
PETT, W.; RYAN, J. F.; PANG, K.; et al. leidyi : Insights from mtDNA and the Nuclear Genome. **Fossils**, v. 22, n. 4, p. 130–142, 2012.

PELLEGRINI, J.C. **Álgebra Linear** . 223 p., 2016. Disponível em: http://aleph0.info/cursos/al/notas/al.pdf

PITTIS, A. A.; GABALDÓN, T. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. **Nature**, v. 531, n. 7592, p. 101–4, 2016. Nature Publishing Group. Disponível em: <http://www.nature.com/doifinder/10.1038/nature16941%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26840490>. .

PRUITT, K. D.; TATUSOVA, T.; MAGLOTT, D. R. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. **Nucleic Acids Research**, v. 35, n. SUPPL. 1, p. 61–65, 2007.

RANNALA, B.; YANG, Z. H. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. **Journal of Molecular Evolution**, v. 43, n. 3, p. 304–311, 1996.

RIDLEY, M. **Evolução**, 3 ed. Artmed: Porto Alegre, 2006. 752p.

RIZZO, J.; ROUCHKA, E. C. Review of Phylogenetic Tree Construction. **University of Louisville Bioinformatics Laboratory Technical Report Series**, 2007.

SALEMI, M.; VANDAMME, A. **The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny.** Cambridge University Press, Cambridge, UK, 2003. 466p.

SANTOS, R.J. **Álgebra Linear e Aplicações**. Universidade Federal de Minas Gerais, 2010.

SILVA, R.A. C. **Inteligência artificial aplicada à ambientes de Engenharia de Software: Uma visão gera**l. Universidade Federal de Viçosa, 2005.

SIMPSON, A. G. B.; ROGER, A. J.; METAZOA, A. The real " kingdoms " of eukaryotes. **Current Biology**, p. 693–696, 2004.

SIMS, G. E.; JUN, S.-R.; WU, G. A.; KIM, S.-H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. **Proceedings of the National Academy of Sciences of the United States of America**, v. 106, n. 8, p. 2677–2682, 2009.

SMITH, A. C.; BLACKSHAW, J. A.; ROBINSON, A. J. MitoMiner: A data warehouse for mitochondrial proteomics data. **Nucleic Acids Research**, v. 40, n. D1, p. 1160–1167, 2012.

SMITH, D. R. The past, present and future of mitochondrial genomics: Have we sequenced enough mtDNAs? **Briefings in Functional Genomics**, v. 15, n. 1, p. 47–54, 2016.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **Journal of molecular biology**, v. 147, n. 1, p. 195–7, 1981. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/7265238>. .

SNUSTAD, P.; SIMMONS, M. Fundamentos de Genética. 2. ed. Rio de Janeiro: Editora Guanabara, 2001.

SONG, K.; REN, J.; REINERT, G.; et al. New developments of alignment-free sequence comparison: Measures, statistics and next-generation sequencing. **Briefings in Bioinformatics**, v. 15, n. 3, p. 343–353, 2014.

STECHMANN, A.; CAVALIER-SMITH, T. The root of the eukaryote tree pinpointed. **Current Biology**, v. 13, n. 17, p. 665–666, 2003.

STEIN, L. Genome annotation: from sequence to biology. **Nature reviews. Genetics**, v. 2, n. 7, p. 493–503, 2001. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/11433356>. .

SULLIVAN, J. Maximum-likelihood methods for phylogeny estimation. **Methods in Enzymology**, v. 395, n. 2002, p. 757–779, 2005.

TAANMAN, J.-W. The mitochondrial genome: structure, transcription, translation and replication. **Biochimica et Biophysica Acta (BBA) - Bioenergetics**, v. 1410, p. 103–123, 1999. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0005272898001613>. .

TEICHERT, F.; PORTO, M. Vectorial representation of single- and multi-domain protein folds. **European Physical Journal B**, v. 54, n. 1, p. 131–136, 2006.

THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Research**, v. 22, n. 22, p. 4673–4680, 1994.

TIAN, Y.; SMITH, D. R. Recovering complete mitochondrial genome sequences from RNA-Seq: A case study of Polytomella non-photosynthetic green algae. **Molecular Phylogenetics and Evolution**, v. 98, p. 57–62, 2016. Elsevier Inc. Disponível em: <http://dx.doi.org/10.1016/j.ympev.2016.01.017>. .

TOBE, S. S.; KITCHENER, A. C.; LINACRE, A. M. T. Reconstructing mammalian phylogenies: A detailed comparison of the cytochrome b and cytochrome oxidase subunit i mitochondrial genes. **PLoS ONE**, v. 5, n. 11, 2010.

TOREN, D.; BARZILAY, T.; TACUTU, R.; et al. MitoAge: A database for comparative analysis of mitochondrial DNA, with a special focus on animal longevity. **Nucleic Acids Research**, v. 44, n. D1, p. D1262–D1265, 2016.

TORRES, A.; NIETO, J.J. Fuzzy Logic in Medicine and Bioinformatics. **Journal of Biomedicine and Biotechnology**.;v. 2006, n. 91908, 2006.

TROY, C. S.; MACHUGH, D. E.; BAILEY, J. F.; et al. Genetic evidence for Near-Eastern origins of European cattle. **Nature**, v. 410, n. 6832, p. 1088–1091, 2001.

TZAGOLOFF, A.; CAPITANIO, N.; NOBREGA, M. P.; GATTI, D. Cytochrome oxidase assembly in yeast requires the product of COX11, a homolog of the P. denitrificans protein encoded by ORF3. **The EMBO journal**, v. 9, n. 9, p. 2759–64, 1990. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=551984&tool=pmcentrez&rendertype=abstract>. .

VIALLE, R. A.; PEDROSA, F. D. O.; WEISS, V. A. RAFTS3 : Rapid Alignment-Free Tool for Sequence Similarity Search. , p. 1–32, 2016.

VINGA, S. Editorial: Alignment-free methods in computational biology. **Briefings in Bioinformatics**, v. 15, n. 3, p. 341–342, 2014.

VINGA, S.; ALMEIDA, J. Alignment-free sequence comparison - A review. **Bioinformatics**, v. 19, n. 4, p. 513–523, 2003.

VIVERO, R. J.; OUYANG, X.; KIM, Y. G. et al. Audiologic and genetic features of the A3243G mtDNA mutation. **Genetic testing and molecular biomarkers**, v. 17, n. 5, p. 383–9, 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23477312%5Cnhttp://www.ncbi.nlm.nih.gov/

pubmed/23477312>. .

WOESE, C. R.; KANDLER, O.; WHEELIS, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. **Proceedings of the National Academy of Sciences of the United States of America**, v. 87, n. 12, p. 4576–4579, 1990.

XIAO J.; KIM S.J.; COHEN P.; YEN K. Humanin: Functional Interfaces with IGF-I. **Growth hormone & IGF research**.; v. 29, p. 21–7, 2016.

YANDELL, M.; ENCE, D. A beginner ' s guide to eukaryotic genome annotation. **Nature Publishing Group**, v. 13, n. 5, p. 329–342, 2012. Nature Publishing Group. Disponível em: <http://dx.doi.org/10.1038/nrg3174>. .

YANG, Z.; RANNALA, B. Molecular phylogenetics: principles and practice. **Nature Reviews Genetics**, v. 13, n. 5, p. 303–314, 2012. Nature Publishing Group. Disponível em: <http://www.nature.com/doifinder/10.1038/nrg3186>. .

YAO, Y. G.; SALAS, A.; LOGAN, I.; BANDELT, H. J. mtDNA Data Mining in GenBank Needs Surveying. **American Journal of Human Genetics**, v. 85, n. 6, p. 929–933, 2009.

ZAHA, A.; FERREIRA, H.; PASSAGLIA L. Biologia Molecular Básica. 5. ed. Porto Alegre; Artmed 2014.

ZAKI, M. J.; KARYPIS, G.; YANG, J. Data Mining in Bioinformatics (BIOKDD). **Algorithms for molecular biology : AMB**, v. 2, p. 4, 2007.

ZHANG, Q.; JUN, S.-R.; LEUZE, M.; USSERY, D.; NOOKAEW, I. Viral Phylogenomics Using an Alignment-Free Method: A Three-Step Approach to Determine Optimal Length of k-mer. **Scientific Reports**, v. 7, n. December 2016, p. 40712, 2017. Nature Publishing Group. Disponível em: <http://www.nature.com/articles/srep40712>. .

APÊNDICE 1 – DESCRIÇÃO DA FIGURA 1

Os ribossomos contêm o rRNA que participa na síntese de proteínas. O espaço intermembranoso possui enzimas que auxiliam no acúmulo prótons, além de conter várias enzimas. A matriz mitocondrial contém enzimas que metabolizam o piruvato e ácido graxo, além de tRNA, mRNA e rRNA. A membrana interna contém os componentes da cadeia transportadora de elétrons. A membrana externa é permeável e auxilia a entrada de moléculas, além de conter enzimas de degradação de lipídios. As cristas auxiliam no aumento da produção de ATP, aumentando a superfície da membrana interna. A molécula de DNA é compacta e pequena, podendo ser linear ou circular dependendo do organismo.

APÊNDICE 2 – DESCRIÇÃO DA FIGURA 2

Em verde estão os genes que codificam para subunidades do complexo I (NADH1 – NADH6 e NADH4L). Em rosa, as subunidades do complexo IV (COX1-COX3). Em roxo está representado o complexo III (CYTB). Em azul escuro o complexo V, compreendendo as subunidades da ATP sintetase (ATPase 6 e 8). Em azul claro estão os dois rRNAs (12S e 16S), e junto ao rRNA 16S o peptídeo Humanin em vermelho. Os 22 tRNAs estão representados por círculos amarelos. A região de controle não codificada (D-loop), fundamental para o início da replicação e transcrição, está representada em cinza.

APÊNDICE 3 – PARSER DOS DADOS ANTERIOR À CLUSTERIZAÇÃO

```
Z = char(allf.Header);
n = length(Z(:,1));
VOrg = cell(n,1);
for iaux=1:n
    ui = strfind(Z(iaux,:),'OS= [');
    uf = strfind(Z(iaux,:),']');
    VOrg{iaux} = Z(iaux,5+ui:uf-1);
end
OrgN = unique(char(VOrg),'rows');
[a OrgI] = ismember(VOrg,OrgN,'rows');
```

APÊNDICE 4 – SCRIPT DE CONSTRUÇÃO DOS VETORES

```
%Script unify fasta Orgs
F = struct2cell(RSM160317.fas);
C = F(2,:)';
lxcat = mat2celllines(repmat('*****',length(C(:,1)),1)); % Finaliza
com lixo para acertar emendas
Z = cellfun(@(x,y) [x y],C,lxcat,'UniformOutput',false)';
n = length(RSM160317.OrgN);
clear OrgFs
for ii=1:n
    OrgFs(ii).Header = RSM160317.OrgN(ii,:);
    OrgFs(ii).Sequence = [Z{RSM160317.OrgI==ii}];
    ii
end
% Vetoriza
Ms400 = cell(n,1);
W160k = zeros(n,160000);
for ii=1:n
    Ms400{ii} = seq2mat(OrgFs(ii).Sequence);
    W160k(ii,:) = matinline(Ms400{ii});
    ii
end
% Reorganiza pela nome de ocorrencias dos dots no mitocondrial
%
[xx rord] = sort(rand(1,160000));
Wr160k = RSM160317.W160k(:,rord);
Wr100 = totWbyk(Wr160k,nkinds(160000,100));
%
%
```

APÊNDICE 5 – FUNÇÃO FILOMAT

```
function [mret dist] = filomat(M,names, varargin)
% Gera arvore filogenetica da variavel RSM
%dist = pdist(M,@Ccosdist);
if ~isempty(varargin)
    dist = varargin{1};
else
    dist = pdist(M);
end
%dist =  dist2;   %dist1.^(1-0.01*dist2);
tree = seqlinkage(dist,'UPGMA',names);
%tree = seqlinkage(dist,'complete',names);
phytreetool(tree)
mret = tree;
```

APÊNDICE 6 – DESCRIÇÃO DO QUADRO 1

Os genomas de plantas, leveduras e protistas codificam para grande parte destas proteínas. * A proteína RuBisCO, presente no cluster de ordem 771, responsável pela fotorrespiração, é derivada de cloroplasto, e segundo os dados obtidos no RefSeq, está presente no proteôma de *Salvia miltiorrhiza*, *Capsicum annuum*, *Brassica nigra* e *Vitis vinífera*. A origem desta proteína no proteoma mitocondrial é incerta.

ANEXO 1 – FUNÇÃO RAFTS3groups

```
function mret = rafts3group(fsall, varargin)
%%
% Agrupa sequências de uma variavel fasta fsall
% com corte de selfscore 0.5
%%%
%tipo = 1; %Sequência em NT
%tipo = 2; %Sequência em aa
tipo = 2;
try hasfield(fsall,'rdfs')
    dball = fsall; %Recebe banco já formatado;
    xinfile = ischar(dball.rdfs);
    if ~xinfile
        xfas = varargin{1}; % Se o banco é baseado em variavel de
memoria passa-la em varargin
    end
catch % Precisa formatar banco
    if ~isempty(varargin)
        tipo = varargin{1};
    end
    tic;
    if tipo==1
        dball = formatdb(nt2aafas(fsall)); %Formata Rafts3
        xfas = nt2aafas(fsall);
    else
        fsall
        dball = formatdb(fsall); %Formata Rafts3
        xfas = fsall;
    end
end
% Banco formatado a partir de arquivo gravado (não fasta em memoria)
xinfile = ischar(dball.rdfs);
%
if xinfile
    n = length(dball.rdid);
else
    n = length(xfas);
end
%
grps = zeros(n,1);
cont = 1;
for i=1:n
    if mod(i,100)==0
        disp([i cont]);
    end
    if ~grps(i)
        if xinfile
            ifas = readfastadirectp2(dball.rdfs,i,dball.rdid);
        else
            ifas = xfas(i);
        end
        q = rafts3(ifas,dball,50);
        u = q.scores;
        igr =
u(u(:,2)>0.50,1:2);%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
        lgrs = grps(igr(:,1));
        ugp = unique(lgrs(find(lgrs)))';
        %disp(ugp);
        if sum(ugp)==0
```

```
            grps(igr(:,1)) = cont;
            cont = cont+1;
        else
            grps(igr(:,1)) = mode(ugp);
        end
    end
end
%
z = contaocorr(grps);
zu = z(z(:,2)>1,:);
[xx ii] = sort(zu(:,2),'descend');
toc;
mret.igrp = grps;
mret.contall = z;
mret.contg2 = zu(ii,:);
```

ANEXO 2 – FUNÇÃO AGRUPARAFTS

```
function mret = agruparafst(file)
%   função de agrupamento de genes orthologos usando o rafsts3groups
%   copyright (c) 2014  federal university of parana
%
%   laboratório de bioinformática - sept
%   pós-graduação em bioinformática
%   universidade federal do paraná
%   rua dr. alcides vieira arcoverde, 1225, jardim das américas
%   curitiba - pr
%   cep 81520-260
%   brasil
%
%   nilson coimbra
%   nilson.coimbra@ufpr.com
%load dbstruct
%carrega dbstruct do sila
system('mkdir Clusters')
output = [pwd '/Clusters'];
allf = fastaread(file);
b = tic;
rftgroups = rafts3group(allf,2);
e = toc;
save rftsgroups.mat, rftgroups;
posis = (int64(rftgroups.igrp));
%indíces das sequências agrupadas
fileid = fopen([output '/Orthologs_Clustered_report.tab'],'w');
fprintf(fileid,['Cluster n°\t Produto\tQuantidade\n']);
for i=1:length(rftgroups.contg2)
%verifica os grupos com mais de duas sequências
    % agrupa usando o rafts
    clu = allf(posis(:)==(rftgroups.contg2(i)));
    % anota sequencidas do grupo identificado e adiciona informação no
    % report
    qtd = int2str(length(clu));
    %anota = sila(dbstruct,clu);
    %name = anota.annotation(1).annotatedFasta.Header;
    %barra = strfind(name,'|');
    %colchete = strfind(name,'[');
    %name = name(barra(4)+1:colchete(1)-1);
    rs = [int2str(i)  '\t' qtd];
%formata escrita
    fprintf(fileid,[rs '\n' ]);
    % Formata o cabeçalho do grupo identificado
    for j=1:length(clu)
        clu(j).Header = ['>Cluster_' int2str(i) '_' clu(j).Header];
    end

    fastawrite([ output '/Cluster_' int2str(i) '.fasta'], clu);
end
fclose(fileid);
cd([pwd '/Clusters'])
system('type Cluster* > allClusters.fasta');
disp(e);
end
```

ANEXO 3 – FUNÇÃO “mmvfuzzy2d”

```
function mret = mmvfuzzy2d(M0, vz)
%c0 = [1 1; 50 30; 100 50; 21 47];
fg = figure;
xlines = length(M0(:,1));
%T = prod(size(M0));
M1 = M0;
nloo = 1;  %Recursões
xx = 100;   %Linhas por iteração - deve ser divisor do tamanho da
matriz - melhorar
xhedge = 1;    %0.5;% Distanciamento de cores
%
for iloo=1:nloo  % Recursões
    image(normmaxmin(M1)*255)
    M2 = [M1 M1(:,end)*0];
    M3 = M2;
    xinds = (find(M1==M1));
    nvz = vz;
    istep = min(xx*length(xinds)/xlines,length(xinds));    %
Generalizar/melhorar
    for iter=1:istep:length(xinds) %Fatias
        ixinds = iter:(iter+istep-1);
        try
            c0 = vints2ij(xinds(ixinds),xlines); %Colocar dimensao do
y (nlins)
        catch
            'Usar número de linhas total multiplo de xx'
            exit
        end
        M = zeros(2*nvz+1,2*nvz+1);
        I = find(M==M);
        ijs = ints2ij(length(I),2*nvz+1);
        c = [nvz+1 nvz+1];
        Ids = normavect(ijs - repmat(c,length(I),1));
        Iin = Ids<=nvz;
        %M(I(Iin)) = 1;
        %********Descarta a borda
        I = I(Iin);
        Ids = Ids(Iin);
        ijs = ijs(Iin,:);
        %********End
        L = length(I);
        %
        I1 = find(M1==M1);
        L1 = length(I1);
        %CL = length(M1(:,1));
        %LL = length(M1(1,:));
        %
        nc = length(c0(:,1));
        %
        Ifz = repmat((Ids/nvz)',nc,1);
        %
        % Funcional para 1 ponto:    ids1 = ij2ints((ijs + repmat(c0-
[nvz nvz],length(I),1)),length(M1(1,:)));
        a = (repmat(ijs,nc,1));
        b2 = c0-repmat([(nvz+1) (nvz+1)],nc,1);
        %b = vints2ij(mat2vet(repmat(ij2ints(c0-repmat([(nvz)
(nvz)],nc,1),CL),1,L))',CL);
```

```
        b = [(mat2vet(repmat(b2(:,1),1,L))')
(mat2vet(repmat(b2(:,2),1,L))')];
        ab = a + b;
        Msz = size(M1);
        %
        ids1 = ij2ints(ab,length(M1(:,1)));
        %ii = find(ids1>0 & ids1<=length(I1));
        %I0 = repmat(I,nc,1);
        %
        Idss = vet2mat(ids1,L);
        Inot = vet2mat(min(ab')<=0 | (ab(:,1)>Msz(2))' |
(ab(:,2)>Msz(1))',L);
        Iin = repmat(Iin',length(Idss(:,1)),1);
        %
        Idss(logical(Inot)) = L1+1;
        Idss(logical(Inot)) = L1+1;
        %
        Ina = (M3==-1);
        Ina = Ina(Idss);  %Tratamento -1
        %
        %M2(Idss) = Ifz.^3; %Iin;    %Ifz.*M2(Idss);
        %M2(Idss) = Ifz.*M2(Idss);
%        U = sum(M3(Idss).*((1-Ifz).^xhedge),2)./sum(~Inot,2);
        U = sum(M3(Idss).*(((1-
Ifz).*double(~Ina)).^xhedge),2)./sum(~(Inot | Ina),2);
        %M2(xinds(1:length(U))) = U;
        M2(xinds(ixinds)) = U;
        figure(fg)
        image(M2*255)
    end
    M1 = M2(:,1:end-1);
end
%
%M1 = normmaxmin(M1);                %Normaliza
M1 = normmaxmin(M1)*255; %Aplica Hedge
mret = M1;
figure
mesh(M1)
mret = imrotate(mret,-90)';
figure(fg)
imagesc(mret)
```